САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Ци Дунфан

Инвестиционная привлекательность и экологическая безопасность в Китае и Юго-Восточной Азии: эмпирические модели и анализ данных

Научная специальность – 1.2.2. Математическое моделирование, численные методы и комплексы программ

Диссертация на соискание ученой степени кандидата технических наук

Перевод с английского языка

Научный руководитель:

доктор технических наук, профессор

Буре Владимир Мансурович

Санкт-Петербург

2024

Оглавление

Введение

Актуальность темы диссертации

Тема исследования, затронутая в данной диссертации, имеет значительную актуальность как с академической, так и с практической точек зрения. Путем исследования взаимосвязи между различными статистическими моделями и исследованиями в области экономики и экологии, с фокусом на Китае и Юго-Восточной Азии, данное исследование рассматривает важные вопросы, имеющие значение для экономического развития и экологической устойчивости в регионе. Анализ привлекательности инвестиций Китая и ASEAN5 с использованием множественной линейной регрессии имеет большое значение в свете быстрого роста и увеличивающегося значения этих экономик. Понимание факторов, влияющих на привлекательность инвестиций, предоставляет ценные практические результаты для политиков, предприятий и инвесторов, стремящихся эффективно распределять ресурсы. Использование последовательной регрессии также усиливает анализ путем идентификации наиболее важных факторов среди широкого диапазона потенциальных переменных, помогая заинтересованным сторонам оптимизировать процесс принятия решений.

Сегментация Китая на четыре отдельные группы на основе привлекательности инвестиций позволяет проводить более тонкий анализ и оценку региональных различий. Понимание уникальных характеристик и проблем, с которыми сталкиваются каждая из этих групп, может служить основой для разработки целевых стратегий экономического развития и сохранения окружающей среды. Политики могут использовать полученные результаты для разработки целевых политик, принимающих во внимание региональные различия и оптимизирующих распределение ресурсов. Анализ факторов, определяющих качество воздуха с использованием последовательной регрессии, освещает актуальную экологическую проблему загрязнения воздуха. Выявление ключевых влияющих факторов способствует принятию обоснованных решений в области управления

окружающей средой.

Более того, применение передовых моделей нейронных сетей, таких как ANN, RNN, LSTM, GRU, BiRNN, BiLSTM и BiGRU, а также ансамблевых моделей, таких как XGBoost, LightGBM и CatBoost, повышает точность прогнозирования качества воздуха. Эти методы позволяют улавливать сложные временные закономерности и создают основу для разработки систем прогнозирования, помогающих заинтересованным сторонам принимать обоснованные решения для решения проблем загрязнения воздуха. Применение методов объяснимого искусственного интеллекта, таких как значения SHAP, повышает прозрачность и интерпретируемость черных ящиковых моделей. Понимание факторов, вносящих вклад в предсказания моделей, крайне важно для построения доверия, получения полезной информации и принятия обоснованных решений на основе выводов этих моделей. Данный анализ соответствует растущему требованию к объяснимости и отслеживаемости систем искусственного интеллекта, отвечая на опасения заинтересованных сторон в отношении использования сложных моделей машинного обучения в реальных приложениях.

В заключение, тема исследования, затронутая в данной диссертации, имеет значительную актуальность как с академической, так и практической точек зрения. Путем исследования взаимосвязи между статистическими моделями и исследованиями в области экономики и экологии в Китае и Юго-Восточной Азии, данное исследование предоставляет ценные практические результаты по привлекательности инвестиций, анализу качества воздуха, региональным различиям, прогнозированию временных рядов, ансамблевому моделированию и объяснительности черных ящиковых моделей. Полученные результаты способствуют продвижению научных знаний, предлагают практические рекомендации для принятия решений и облегчают разработку политики на основе доказательств в области экономики и экологии.

Обзор литературы

Исследование, проведенное в данной диссертации, строится на значительных вкладах ученых и исследователей в области анализа инвестиций и экологической науки. Продвижение в понимании привлекательности инвестиций, анализа качества воздуха, региональных различий, прогнозирования временных рядов, выбора моделей и интерпретируемости сложных моделей стали основой для

данного исследования.

Привлекательность инвестиций играет важную роль в процессах принятия решений инвесторами, способствуя снижению рисков, максимизации доходности и эффективному распределению капитала. Оценивая различные факторы, такие как состояние рынка, конкуренция, законодательство и экономическая стабильность, инвесторы могут принимать обоснованные решения, соответствующие их финансовым целям. Моделирование привлекательности инвестиций улучшает оценку, предлагая всесторонний подход. Эти модели включают как количественные, так и качественные переменные, позволяя проводить прогнозирование, анализ сценариев и сравнительную оценку. Такие подходы снижают зависимость от субъективных суждений и повышают точность инвестиционных решений. Признание значимости привлекательности инвестиций и использование методов моделирования способствует оптимизации результативности портфеля и стимулирует устойчивый экономический рост.

Линейная регрессия и пошаговая регрессионный анализ широко используемые методы для моделирования привлекательности инвестиций. Эти подходы предоставляют простоту, интерпретируемость и возможность статистического тестирования, что позволяет систематический анализ факторов, влияющих на привлекательность инвестиций. Линейная регрессия позволяет включать количественные данные и облегчает определение связей между независимыми и зависимыми переменными. Кроме того, проверка гипотез дает представление о статистической значимости наблюдаемых связей, что подтверждает надежность полученных выводов. Пошаговая регрессионный анализ расширяет линейную регрессию, автоматически выбирая значимые независимые переменные, улучшая интерпретируемость и вычислительную эффективность.

Во второй главе применяется кластерный анализ для выявления отдельных групп или кластеров на основе сходства и различий между выбранными переменными. Этот анализ раскрывает скрытые закономерности и структуры внутри набора данных, предоставляя ценную информацию о взаимосвязях и характеристиках рассматриваемых инвестиционных объектов.

Статья [1] направлена на изучение методологической поддержки для оценки привлекательности инвестиций инновационных компаний, удовлетворяющих информационным потребностям заинтересованных сторон. Авторы используют методы анализа и синтеза для определения и структурирования концеп-

ции привлекательности инвестиций и ее аналитических характеристик. Статья [2] направлена на выявление и проверку теоретических характеристик города, привлекательного для инвестиций, путем анализа экономической литературы и опроса предпринимателей. Основные факторы, влияющие на выбор города для инвестиций, включают доступность квалифицированной рабочей силы, трудовые затраты, цены на ресурсы и конкуренцию на рынке. В этих статьях [3, 4, 5, 6] рассмотрены различные аспекты привлекательности инвестиций, начиная от анализа производительности и рейтингов конкретных стран до изучения инвестиционных стратегий и решения проблем и перспектив повышения привлекательности инвестиций.

Ученые, использующие статистические методы для анализа привлекательности инвестиций, могут столкнуться с ограничениями, включая слишком упрощенный подход, недостаток учета контекстуальных факторов и встроенные предположения. Эти ограничения могут привести к неточным прогнозам и неполному пониманию факторов, определяющих привлекательность инвестиций. Однако пошаговый регрессионный анализ предлагает преимущества и инновации для преодоления этих ограничений. Путем автоматического выбора соответствующих переменных он преодолевает слишком упрощенный подход и позволяет выявлять нелинейные связи и контекстуальные факторы. Гибкость модели позволяет итеративное усовершенствование, адаптацию к изменяющимся рыночным условиям и учет новых тенденций. Благодаря этим преимуществам, пошаговый регрессионный анализ предоставляет новые практические результаты относительно факторов, определяющих привлекательность инвестиций, что способствует более глубокому пониманию процесса принятия решений. Этот подход соответствует академической строгости, логическому рассуждению и стандартам научных исследований, повыш

Область экологических наук значительно продвинулась в анализе качества воздуха и его последствиях для общественного здоровья и окружающей среды. Были проведены исследования, направленные на изучение пространственно-временных закономерностей загрязнения воздуха в Китае с использованием данных индекса качества воздуха (AQI), что позволило выявить высокие уровни загрязнения по всей стране и определить PM2,5, PM10 и O3 как основные загрязнители [7]. Прогнозирование качества воздуха, особенно параметров загрязнения, стало важным для принятия решений в этой области.

Исследователи разработали модели для прогнозирования ежедневного AQI, адаптируя методы, аналогичные методам, используемым Агентством по охране окружающей среды США (USEPA) для индийских стандартов [8]. Исследование [9] не выявило значительных различий в AQI в выходные дни по сравнению с рабочими днями, рассматривая все дни одинаково в моделях. Кроме того, была подчеркнута взаимосвязь между загрязнением воздуха и изменением климата, требующая скоординированных действий, учитывающих их взаимосвязь [10].

Исследования выявили превалирование природных факторов над социоэкономическими влияющих на загрязнение воздуха, с взаимодействием между факторами, приводящими к нелинейно усиленным или двусторонним эффектам [11, 12, 13, 14, 15]. Эти результаты имеют значительное значение для разработки политики по смягчению загрязнения воздуха в Китае.

Предыдущие исследования проводились в различных аспектах загрязнителей воздуха, таких как атмосферная пыль, озон, угарный газ, диоксид серы и диоксид азота, с целью понять источники, закономерности распространения и воздействия на здоровье [16, 17, 18, 19, 20]. Статистические методы, включая анализ регрессии, прогнозирование временных рядов и алгоритмы машинного обучения, были применены для моделирования и прогнозирования качества воздуха. Эти исследования ([21, 22, 23, 24]) углубили наше понимание региональных неравенств, выявили ключевые факторы, влияющие на развитие, и предложили практические рекомендации для лиц, принимающих решения в этой области.

Эти достижения в понимании закономерностей загрязнения воздуха и прогнозировании способствуют улучшению стратегий управления и разработке политики. Они освещают сложные взаимодействия между факторами воздействия и дают представление о источниках, распространении и воздействии загрязнителей воздуха, позволяя более эффективно бороться с ними.

В последние годы произошли значительные прогрессивные шаги в прогнозировании временных рядов при использовании искусственных нейронных сетей (ANN). Исследователи изучили различные архитектуры, такие как рекуррентные нейронные сети (RNN), долгосрочную память (LSTM), блокирующие рекуррентные блоки (GRU) и их варианты для моделирования и прогнозирования зависящих от времени данных. Эти модели продемонстрировали свою эффек-

тивность в улавливании сложных временных закономерностей и прогнозировании будущих значений [25, 26, 27, 28]. Для достижения более высокой точности прогнозирования временных рядов была предложена гибридная методология, предложенная Чжаном и др.[29], которая комбинирует авторегрессионное интегрированное скользящее среднее (ARIMA) и модели ANN.

Применение АНС также оказалось успешным в экологических науках. Палани и др. [30] продемонстрировали использование моделей ANN для прогнозирования показателей качества воды в прибрежных водах Сингапура, точно симулируя уровни солености, температуры, растворенного кислорода и хлорофилла-а.

Исследования [31, 32, 33] предлагают исчерпывающие руководства по RNN и LSTM, объясняя вывод уравнений, решение проблем обучения и представление улучшенных версий моделей LSTM. Эти ресурсы предоставляют ценные идеи для исследователей, стремящихся реализовать расширенные модели LSTM.

Обзор от De Gooijer и Hyndman [34] охватывает 25 лет исследований в области прогнозирования временных рядов, с фокусом на статьях, опубликованных в журналах, управляемых Международным институтом прогнозирования. Обзор подчеркивает значительный вклад, выявляет области для дальнейшего развития и предлагает направления будущих исследований.

Лим и др. [35] представляют обширный обзор прогнозирования временных рядов, акцентируя внимание на все большем использовании глубоких нейронных сетей. Они обсуждают распространенные архитектуры глубокого обучения, включая прямые нейронные сети, рекуррентные нейронные сети (Эльман, LSTM, GRU, двунаправленные) и сверточные нейронные сети. Также исследуются практические аспекты, такие как настройка гиперпараметров и выбор фреймворков.

В контексте прогнозирования финансовых временных рядов Sezer и др. [36] представляют обзор исследований, использующих модели глубокого обучения. Категоризируя реализации по области (индексы, форекс, товары) и выбору моделей глубокого обучения (CNNs, DBNs, LSTM), этот обзор предоставляет представление о потенциале и ограничениях использования моделей глубокого обучения в финансовом прогнозировании.

Применение АНС распространяется на фармацевтические науки, как подчеркнуто Agatonovic-Kustrin и др. [37]. АНС имитируют информационную об-

работку человеческого мозга, что позволяет применять их в классификации, прогнозировании и моделировании, поддерживая разработку лекарств и клиническую фармакологию.

Более того, Лим и др. [38] представляют обзор архитектур глубокого обучения в прогнозировании временных рядов, обсуждают проектирование кодировщиков и декодировщиков для прогнозирования на один шаг вперед и на несколько периодов в будущее. Исследуется интеграция статистических моделей с нейронными сетями в гибридных моделях, а также потенциальные преимущества глубокого обучения при принятии решений с использованием временных рядов. Эти результаты способствуют пониманию и применению методов глубокого обучения в прогнозировании данных, зависящих от времени.

Эти статьи [39, 40, 41] продемонстрировали значительные достижения в прогнозировании временных рядов благодаря применению моделей градиентного усиления, таких как XGBoost, LightGBM и CatBoost. Они показали свою превосходность по сравнению с традиционными статистическими методами в области прогнозирования качества воздуха. Эти модели эффективно улавливают сложные взаимосвязи, нелинейные закономерности и обрабатывают категориальные признаки. Внедрение этих передовых моделей градиентного усиления повышает точность и надежность прогнозов качества воздуха, что поддерживает информированное принятие решений политиками и заинтересованными сторонами.

Исследование Sagi и др. [42] решает потребность в интерпретируемых моделях машинного обучения и предлагает метод преобразования моделей GBDT в интерпретируемые деревья решений без потери прогностической производительности. Исследование Ramraj и др. [43] сравнивает точность и скорость работы XGBoost с традиционным градиентным бустингом в многопоточном однопользовательском режиме, демонстрируя превосходство времени обучения и производительности XGBoost. Работа [44, 45, 46, 47] представляет XGBoost, масштабируемую систему градиентного усиления деревьев, широко используемую в задачах машинного обучения, с новыми алгоритмами для работы со разреженными данными и приближенным построением деревьев. Аналогично, статья [48] предлагает методы Gradient-based One-Side Sampling (GOSS) и Exclusive Feature Bundling (EFB) для повышения эффективности и масштабируемости алгоритмов градиентного усиления деревьев (GBDT), что приводит к

разработке LightGBM.

Другие соответствующие статьи, такие как те, посвященные CatBoost и его применению [49, 50, 51, 52], а также исследования алгоритмов деревьев решений [53, 54], способствуют более широкому пониманию моделей градиентного усиления в машинном обучении. Статья [55] фокусируется на прогнозировании концентрации PM2.5 на почасовой основе в Китае с использованием алгоритма XGBoost. В этом исследовании оценивается производительность XGBoost путем сравнения наблюдаемых и прогнозируемых концентраций PM2.5, что демонстрирует его превосходство над другими методами анализа данных. В статье Ju et al. [56] предлагается модель, объединяющая сверточную нейронную сеть (CNN) и алгоритм LightGBM для прогнозирования ветровой энергии на краткосрочный период. Этот подход использует преимущества обоих моделей для достижения улучшенной точности прогнозирования ветровой энергии.

Статья Dorogush et al. [57] представляет CatBoost, фреймворк градиентного усиления с поддержкой категориальных признаков. В ней освещаются преимущества CatBoost в работе с наборами данных, содержащими категориальные признаки, что позволяет добиться улучшенной производительности по сравнению с традиционными моделями градиентного усиления.

Эти исследования дополняют область прогнозирования временных рядов, исследуя применение алгоритмов XGBoost, LightGBM и CatBoost в различных областях, включая прогнозирование качества воздуха и прогнозирование ветровой энергии, демонстрируя их эффективность и превосходство над другими методами.

В рамках интерпретируемого искусственного интеллекта значительный вклад вносит анализ значений SHAP. Исследования Марсилио и др. [58], Мэнга и др. [59] и Мохтари и др. [60] исследуют применение значений SHAP в качестве механизма выбора признаков и интерпретируемых решений в различных областях.

Эти исследования подчеркивают значимость значений SHAP в интерпретации моделей в различных областях, включая управление процессами в очистных сооружениях сточных вод [61] и объяснения машинного обучения [62, 63, 64]. В литературе, посвященной значениям SHAP, существуют ключевые работы, такие как работы Винтера [65] и Рота [66], а также исследования, исследующие вариации и алгоритмические подходы [67, 68, 69]. Однако необходимо признать сложности, связанные с использованием значений SHAP, о которых

говорят статьи Кумара и др. [70], где рассматриваются потенциальные ограничения и проблемы при использовании значений SHAP для измерения важности признаков. Эти дополнительные статьи дополняют понимание и применение значений Шепли в различных областях, включая кооперативную игровую теорию, машинное обучение и объяснения моделей.

Кроме того, исследования в области кооперативной игровой теории подчеркивают значимость значений Шепли. Исследования Литтлчайлда [71], Калаи и Самета [72], Харта и Мас-Колелля [73], Роземберцки и др. [74] и Меррик и др. [75] способствуют пониманию и применению значений Шепли при интерпретации прогнозов модели, важности признаков и сценариев кооперативных игр.

Интерпретируемый искусственный интеллект (XAI) стал ключевой областью исследований, стремящейся обеспечить прозрачность и интерпретируемость сложных моделей машинного обучения. Исследователи изучали различные концепции, методологии и вызовы в стремлении к ответственному искусственному интеллекту[76, 77, 78, 79]. Программа DARPA XAI [80] и статьи, такие как Арриета и др. [81], Дас и др. [82] и Ван и др. [83], значительно способствуют пониманию и развитию XAI. В них обсуждаются таксономии, возможности, вызовы и подходы к достижению интерпретируемости в различных областях, таких как анализ медицинских изображений, системы клинической поддержки принятия решений и пользовательский опыт.

Кроме того, исследования Адади и Беррада [84], Тжоа и др. [85] и Лангер и др. [86] акцентируют внимание на точках зрения заинтересованных сторон, междисциплинарных исследованиях XAI и важности объяснений, ориентированных на пользователя, в XAI. Антониади и др. [87] и Лиао и др. [88] рассматривают вызовы и возможности применения XAI в системах клинической поддержки принятия решений и принципах дизайна, ориентированных на человека.

Обзор Всираса и Гейста [89] исследует аргументационный XAI, тогда как Саид и др. [90] и Ван и др. [91] предлагают систематические оценки и сравнения методов XAI. Волф и др. [92] фокусируются на сценарийно-ориентированном проектировании XAI, а Паэс и др. [93] вводят прагматическое направление в исследованиях XAI.

Более того, Мин и др. [94] представляют обширный обзор XAI, охватывая

различные методы и применения. Шлегель и др. [77] и Рожат и др. [95] обсуждают методы XAI для анализа временных рядов, в то время как Мачлев и др. [96] исследуют техники XAI в энергетических и электроэнергетических системах. Кроме того, статья Кенни и др. [97] исследует пост-фактум объяснения на основе примеров и их влияние на пользовательские исследования.

Эти статьи в совокупности вносят свой вклад в понимание и развитие XAI, предлагая представления о его методологиях, вызовах и возможностях в различных областях.

В заключение, достижения в области анализа инвестиций и экологических наук были значительными и многогранными. Исследователи внесли заметный вклад, изучая ключевые факторы, влияющие на привлекательность инвестиций, анализируя индекс качества воздуха, изучая региональные неравенства, совершенствуя техники прогнозирования временных рядов и повышая интерпретируемость моделей. Настоящая диссертация основывается на этом фундаменте, применяя различные статистические методы, модели машинного обучения и методы объяснения AI, продвигая наше понимание этих областей и предоставляя практические применения для инвесторов, политиков и исследователей.

Цели и задачи диссертации

Основная цель данной диссертации заключается в внесении вклада в развитие знаний в области экономики и экологии, с акцентом на Китае и странах АСЕАН5. Целью является предоставление практических инсайтов, которые могут служить основой для принятия решений, способствовать экономическому развитию, содействовать экологической устойчивости и помочь в формулировке политики. Для достижения этой всесторонней цели были определены следующие подцели и соответствующие задачи:

1. Оценка привлекательности инвестиций: Применение множественной линейной регрессии для оценки привлекательности инвестиций в Китае и странах АСЕАН5. Использование пошаговой регрессии для определения наиболее значимых факторов, влияющих на привлекательность инвестиций. Предоставление инсайтов политикам, бизнесу и инвесторам для принятия обоснованных решений по распределению ресурсов и стратегиям инвестирования.

2. Кластерный анализ и региональный анализ: Применение методов кластерного анализа для деления Китая на отдельные региональные группы в зависимости от характеристик. Проведение отдельного регрессионного анализа для каждой группы с целью выявления факторов, влияющих на экономическое развитие и экологические условия в каждом регионе. Помощь в разработке целевых политик и стратегий распределения ресурсов, учитывающих уникальные вызовы и возможности в каждом регионе.

3. Анализ индекса качества воздуха: Применение пошаговой регрессии для анализа индекса качества воздуха. Определение ключевых детерминант и факторов, влияющих на уровни загрязнения воздуха. Помощь политикам и органам охраны окружающей среды в разработке целевых мероприятий и политик для улучшения качества воздуха.

4. Прогнозирование временных рядов качества воздуха: Применение передовых моделей нейронных сетей, таких как ANN, RNN, LSTM, GRU, BiRNN, BiLSTM и BiGRU, для симуляции и прогнозирования качества воздуха в контексте временных рядов. Повышение точности прогнозирования путем учета сложных временных закономерностей. Обеспечение принятия предупредительных решений и целевых мероприятий для контроля и управления загрязнением воздуха.

5. Ансамблевое моделирование для прогнозирования качества воздуха: Применение ансамблевых моделей, включая XGBoost, LightGBM и CatBoost, для симуляции и прогнозирования качества воздуха в течение определенного времени. Сравнительный анализ эффективности этих моделей для прогнозирования качества воздуха. Помощь заинтересованным сторонам в выборе подходящих методов для точных и достоверных прогнозов.

6. Анализ черных ящиковых моделей с использованием значимости SHAP: Анализ черных ящиковых моделей с использованием значений SHAP в Explainable AI. Исследование факторов, вносящих вклад в предсказания моделей и оценка их влияния. Повышение прозрачности и интерпретируемости модели для поддержки ответственного принятия решений на основе выводов моделей.

Путем выполнения этих основных задач диссертация стремится достичь сво-

ей главной цели - продвижения знаний, предоставления практических инсайтов и поддержки обоснованного принятия решений в экономической и экологической областях, особенно в отношении привлекательности инвестиций и качества воздуха в Китае и странах АСЕАН5.

Научная новизна

Данная диссертация представляет собой новаторский и всесторонний подход к анализу привлекательности инвестиций и качества воздуха, вносящий вклад в области анализа инвестиций и экологической науки. Она предлагает научную новизну в нескольких аспектах путем интеграции множества аналитических методов и их инновационного применения. Путем объединения множественной линейной регрессии, кластерного анализа и различных моделей машинного обучения, данное исследование обеспечивает целостное понимание привлекательности инвестиций и индекса качества воздуха, расширяя наше представление об этих сложных явлениях. Применение пошаговой регрессии добавляет новую размерность в исследования привлекательности инвестиций, выявляя наиболее значимые факторы, влияющие на привлекательность инвестиций в Китае и странах АСЕАН5. Более того, использование кластерного анализа позволяет разделить Китай на отдельные группы для регионального анализа, предоставляя ценные инсайты в разнообразные факторы, влияющие на привлекательность инвестиций в различных регионах. В отношении прогнозирования качества воздуха, данная диссертация исследует прогнозирование временных рядов с использованием различных моделей нейронных сетей, таких как ANN, RNN, LSTM, GRU, BiRNN, BiLSTM и BiGRU, внося новые идеи в существующую литературу. Дополнительно, сравнительный анализ моделей XGBoost, LightGBM и CatBoost предлагает ценные выводы о их пригодности для задач прогнозирования качества воздуха. Наконец, данное исследование использует методы объяснимого искусственного интеллекта, а именно анализ значимости SHAP, для интерпретации черных ящиковых моделей, что обеспечивает прозрачность и понимание. В целом, данная диссертация представляет собой научную новизну благодаря интеграции нескольких аналитических методов, инновационному применению моделей регрессии, исследованию нейронных сетей, сравнительному анализу моделей усиления и использованию методов объяснимого искусственного интеллекта, продвигая наше понимание привлекательности инвести-

ций и качества воздуха, и предоставляя значимые выводы для принятия решений в соответствующих областях.

Структура статьи и организация глав

Эта статья состоит из пяти глав, каждая из которых исследует различные аспекты инвестиционной привлекательности и оценки качества воздуха с использованием различных статистических методов.

Глава 1: Анализ инвестиционной привлекательности с использованием множественной линейной регрессии: всестороннее исследование ключевых факторов для устойчивого экономического развития. В этой главе используется метод множественной линейной регрессии для моделирования инвестиционной привлекательности. Кроме того, применяется метод шаговой регрессии для выявления ключевых факторов, влияющих на инвестиционную привлекательность. Глава представляет всесторонний анализ этих факторов.

Глава 2: Исследование инвестиционной привлекательности: глубинный анализ с использованием метода кластерного анализа. Эта глава более детально исследует инвестиционную привлекательность с использованием метода кластерного анализа. На основе этого метода вся территория Китая делится на четыре отдельные группы, что позволяет проводить моделирование и анализ данных в каждой группе. Глава представляет углубленное исследование инвестиционной привлекательности внутри этих кластеров.

Глава 3: Анализ и моделирование индекса качества воздуха с использованием шаговой регрессии: исследование тенденций и оценка соответствия. Эта глава фокусируется на использовании шаговой регрессии для анализа и моделирования индекса качества воздуха (AQI). Она предоставляет обзор значимости изучения качества воздуха, объясняет методологию шаговой регрессии и описывает процесс выбора влиятельных переменных. В главе также рассматривается исследование тенденций и оценка адекватности модели.

Глава 4: Методы глубокого обучения для системы оценки качества воздуха. В этой главе представлены методы глубокого обучения, включая модели, такие как ANN, RNN, LSTM, GRU, BiRNN, BiLSTM и BiGRU. С использованием этих моделей проводятся экспериментальные симуляции качества воздуха, в частности PM2.5. Производится оценка производительности каждой модели с различных точек зрения, предоставляя понимание качества и эффективности

каждой из семи моделей.

Глава 5: Методы ансамблевого обучения для системы оценки качества воздуха. Эта глава фокусируется на методах ансамблевого обучения, а именно XGBoost, LightGBM и CatBoost. Кроме того, вводится интерпретация результатов через объяснения на основе значимости SHAP (SHapley Additive exPlanations). С помощью трех упомянутых моделей проводятся симуляции и прогнозирование данных о качестве воздуха (PM2.5). Более того, в главе производится анализ влияющих факторов на основе подхода SHAP, что позволяет интерпретировать исходно непрозрачную модель.

Структурируя статью по этим пяти главам, исследование охватывает всестороннее исследование инвестиционной привлекательности и оценки качества воздуха. Использование нескольких статистических подходов позволяет провести всеобъемлющий анализ различных факторов и методик, способствуя всестороннему изучению этих важных областей.

Теоретическая и практическая значимость

Теоретическое значение:

С теоретической точки зрения, данное исследование вносит значительный вклад в уже существующие знания в нескольких областях. Во-первых, в анализе привлекательности инвестиций, данное исследование расширяет наше понимание факторов, определяющих инвестиции в Китае и странах АСЕАН5. Исследуя переменные, такие как доход на душу населения, основные фонды, строительная деятельность и ВВП на душу населения, данное исследование предоставляет инсайты в сложную динамику, влияющую на решения об инвестициях. Эти результаты способствуют теоретическому пониманию регионального развития и являются основой для будущих исследований стратегий привлечения инвестиций.

Во-вторых, исследование изменений качества воздуха предлагает ценные инсайты в факторы, влияющие на уровень загрязнения воздуха. Определение значимых переменных, таких как SO2 и NO2, путем анализа регрессии, продвигает наше понимание экологических детерминант, влияющих на качество воздуха. Полученные результаты предоставляют теоретические предпосылки для понимания динамики загрязнения воздуха и способствуют развитию знаний в области управления окружающей средой и общественного здравоохранения.

Кроме того, оценка различных моделей прогнозирования для прогнозирования качества воздуха углубляет наше теоретическое понимание их эффективности и эффективности. Сравнивая модели, такие как BiRNN и LightGBM, данное исследование вносит свой вклад в область прогнозирования, предоставляя эмпирические данные о производительности этих моделей в учете сложности данных по качеству воздуха. Эти выводы углубляют наше понимание возможностей и ограничений различных алгоритмов машинного обучения, дополняя теоретическую основу исследований прогнозирования качества воздуха.

Практическое значение:

Практическая значимость данного исследования заключается в его последствиях для политиков, бизнеса и инвесторов. Полученные результаты предлагают ценные практические рекомендации, которые могут служить основой для принятия решений и указывать на действия в реальных ситуациях.

Во-первых, анализ привлекательности инвестиций предоставляет практическую помощь политикам, стремящимся привлечь инвестиции и способствовать экономическому росту. Определение ключевых факторов, таких как доход на душу населения и строительная деятельность, позволяет политикам адаптировать свои стратегии, чтобы создать привлекательную инвестиционную среду. Это знание может помогать эффективно направлять ресурсы и реализовывать целевые политики, способствующие развитию бизнеса и улучшению региональных экономик.

Во-вторых, исследование изменений качества воздуха имеет практическое значение для управления окружающей средой и принятия политических решений. Понимание факторов, влияющих на уровни загрязнения воздуха, таких как SO2 и NO2, позволяет политикам разрабатывать обоснованные меры и нормативные акты. Целенаправленное воздействие на эти конкретные загрязнители позволяет проводить более эффективные мероприятия по защите от загрязнения воздуха, повышению качества воздуха и обеспечению общественного здоровья.

Кроме того, оценка методов прогнозирования качества воздуха имеет практическое значение для заинтересованных сторон, занимающихся контролем и управлением загрязнением воздуха. Выявление моделей с более высокой точностью и эффективностью, таких как BiRNN и LightGBM, предоставляет практическую помощь при выборе наиболее подходящего подхода для прогнози-

рования качества воздуха. Это дает возможность заинтересованным сторонам принимать своевременные решения, предпринимать проактивные меры по снижению загрязнения и эффективно распределять ресурсы.

В целом, практическое значение данного исследования заключается в его способности информировать политиков, бизнес-сектор и инвесторов о эффективных стратегиях и мерах для улучшения привлекательности инвестиций, качества воздуха и стимулирования устойчивого развития. Полученные результаты исследования могут служить основой для принятия решений, оптимизации распределения ресурсов и достижения положительных экологических и экономических результатов в реальных ситуациях.

Положения, выносимые на защиту

1. Проведено статистическое моделирование для оценки инвестиционной привлекательности регионов Китая и стран АСЕАН. Факторы, влияющие на инвестиционную привлекательность в регионах Китая и странах АСЕАН, были проанализированы с использованием пошаговой регрессии. Был выполнен кластерный анализ для изучения вариаций инвестиционной привлекательности в различных регионах Китая.

2. Для изучения динамики качества воздуха в регионах Китая было проведено статистическое моделирование. Факторы, влияющие на качество воздуха, были выявлены с помощью анализа пошаговой регрессии.

3. Сравнительная оценка алгоритмов глубокого обучения (ANN, RNN, BiRNN, LSTM, BiLSTM, GRU, BiGRU) для прогнозирования временных рядов качества воздуха.

4. Сравнительная оценка алгоритмов ансамблевого обучения (LightGBM, Cat Boost, XGBoost) для прогнозирования временных рядов качества воздуха. Факторный анализ черных ящиковых моделей с использованием значений SHAP в рамках технологии объяснимого искусственного интеллекта (XAI).

Основные научные результаты

К **основным научным результатам**, достигнутым в ходе диссертационного исследования, следует отнести:

1. Исследование сфокусировано на изучении инвестиционной привлекательности. Сначала был применен метод множественной регрессионного анализа для построения моделей оценки инвестиционной привлекательности ([98], стр. 3). Затем был использован метод пошаговой регрессии для выявления и включения наиболее влиятельных факторов в каждую модель ([98], стр. 4). Для всесторонней оценки воздействия выявленных факторов на инвестиционную привлекательность был проведен окончательный анализ результатов каждой модели ([98], стр. 5). На основе полученных данных авторы предложили список наиболее значимых определяющих факторов, влияющих на инвестиционную привлекательность ([98], стр. 6). Подробные результаты данного исследования задокументированы в публикации авторов [98]. Исследователи активно участвовали в различных этапах исследования, включая сбор данных, разработку моделей, анализ результатов, обзор литературы, интерпретацию результатов и написание статьи.

2. Построив на основе базового исследования, цитируемого в [99], расширенный набор данных был использован для тщательного классификационного анализа инвестиционной привлекательности. Сначала регионы подвергаемые изучению были разделены на четыре отдельные группы в зависимости от уровня инвестиционной привлекательности ([99], стр. 4). Затем были проведены сравнения инвестиционной привлекательности в каждой группе для оценки вариаций и тонкостей ([99], стр. 5). Для каждой группы затем были разработаны индивидуальные модели, после чего были проведены подробные анализы для ясного выявления конкретных тенденций и идей ([99], стр. 8–11). Результаты исследования показали, что стратегии привлечения инвестиций существенно различаются в различных регионах и группах ([99], стр. 12). Обширные выводы, вытекающие из этого исследования, задокументированы в публикации авторов [99]. Исследователи активно участвовали во всех аспектах исследовательского процесса, включая сбор данных, обзор соответствующей литературы, интерпретацию результатов и написание статьи.

3. Основная цель данного исследования - выявить и оценить ключевые факторы, влияющие на зависимую переменную. Исследовательские выводы проиллюстрированы на примере кейс-стади о состоянии качества воздуха

в Китае ([100], стр. 2). В частности, анализируются такие аспекты, как характеристики модели, оценка качества, проверка параметров и диагностика остатков ([100], стр. 3–4). Далее, для улучшения предсказательной точности модели была использована методика обратного исключения пошаговой регрессии для создания окончательной уточненной модели, продемонстрированной на примере ситуации с качеством воздуха в Китае ([100], стр. 6). Параллельно было проведено подробное изучение каждого аналитического шага для выявления результатов по динамике качества воздуха в Китае ([100], стр. 7). Установленная модель затем использовалась для прогнозирования годового индекса качества воздуха (AQI) для 31 столичных городов провинций Китая за период с 2013 по 2019 год. Полученные прогнозы подтверждены фактическими данными о качестве воздуха в Китае ([100], стр. 8). Эти обширные результаты были опубликованы в работе автора [100]. Исследователь активно участвовал в получении этих результатов, включая сбор материалов, анализ данных литературы и интерпретацию результатов.

4. Основной целью данного исследования является оценка долгосрочных тенденций качества воздуха в Китае путем применения мультиномиальных логистических методов регрессии на основе индекса качества воздуха (AQI) и комплексного индекса качества воздуха (AQCI). Были разработаны две отдельные модели, каждая из которых использует различные зависимые переменные - AQI и AQCI - при сохранении постоянных контрольных переменных, таких как валовой внутренний продукт (ВВП) и основные загрязнители ([101], стр. 4). В частности, основные загрязнители, рассматриваемые в анализе, связаны с одним или несколькими из шести факторов загрязнения: O3, PM2.5, PM10, NO2, SO2 и CO ([101], стр. 6). Обеспечение качества и достоверности моделей имеет первостепенное значение и является важной составляющей аналитического процесса ([101], стр. 7). Результаты, опубликованные в соответствующем исследовании [101], разъясняются с использованием аутентичных данных о качестве воздуха, полученных из Китая. Автор активно участвовал в получении этих результатов, включая сбор материалов, обзор литературы и интерпретацию результатов.

5. Данный документ проводит глубокий анализ моделей прогнозирования

временных рядов, используемых для прогнозирования качества воздуха. Исследование сосредоточено на выявлении и оценке прогностических моделей, подходящих для анализа окружающей среды, охватывая такие важные алгоритмические фреймворки, как нейронные сети и ансамблевые модели ([102], стр. 4–6). Эффективность и производительность этих моделей оцениваются с использованием ключевых метрик, включая среднюю абсолютную ошибку (MAE), корень из среднеквадратичной ошибки (RMSE) и значения R-квадрат ([102], стр. 7–9). Результаты свидетельствуют о том, что нейронные сети и ансамблевые модели обладают надежными способностями для надежного прогнозирования данных временных рядов качества воздуха ([102], стр. 12). Эти результаты были формально опубликованы автором [102]. Исследователь активно участвовал на различных этапах исследовательского процесса, включая сбор материалов, анализ данных литературы, интерпретацию результатов и составление рукописи.

6. Данная научная статья представляет собой всестороннее исследование Shap Time, методологии объяснимого искусственного интеллекта (XAI), основанной на значениях Шэпли и специально разработанной для улучшения интерпретируемости и эффективности прогнозирования временных рядов, погружаясь в сложные временные динамики. Основные инновации, связанные с ShapTime, заключаются в его способности предоставлять стабильные объяснения, отражая тем самым внутреннюю значимость времени и делая его более подходящим для прогнозирования временных рядов по сравнению с обычными техниками XAI ([103], стр. 8). Кроме того, исследование вводит прагматическую рамку применения в XAI, где разъясненные результаты служат направляющими принципами для повышения точности прогнозирования, выделяясь тем самым из предыдущих исследований, которые использовали результаты XAI исключительно как демонстрации новизны ([103], стр. 10). Особо следует отметить, что на пяти различных реальных наборах данных ShapTime продемонстрировал заметное улучшение средней производительности в моделях Boosting, основанных на RNN и Bi-RNN, что привело к увеличению на 18%, 20% и 35% соответственно ([103], стр. 13). Эти исследовательские результаты были формально задокументированы и распространены в упомянутой публикации [103], при

этом автор активно участвовал в тщательном процессе сбора материалов, анализа литературы и интерпретации результатов, подчеркивая практический подход к получению знаний.

Глава 1

Анализ инвестиционной привлекательности с использованием множественной линейной регрессии: всестороннее исследование выявления ключевых факторов для устойчивого экономического развития

Обсуждение, представленное в данной главе, опубликовано в статье [98].

1.1 Группы объектов исследования

1. **Регионы Китая** (Исключены провинции с низкой инвестиционной привлекательностью)

   - Пекин
   - Тяньцзинь
   - Хэбэй
   - Шаньси
   - Внутренняя Монголия
   - Ляонин
   - Цзилинь
   - Хэйлунцзян
   - Шанхай

- Цзянсу
- Чжэцзян
- Аньхой
- Фуцзянь
- Цзянси
- Шаньдун
- Хэнань
- Хубэй
- Хунань
- Гуандун
- Гуанси
- Чунцин
- Сычуань
- Гуйчжоу
- Юньнань
- Шаньси
- Синьцзян-Уйгурский автономный район

2. **ASEAN-5**

- Индонезия
- Малайзия
- Сингапур
- Таиланд
- Филиппины

1.2 Источники данных и методы

1.2.1 Источник и сбор данных

Данные об инвестиционной привлекательности регионов Китая: Исходные данные об инвестиционной привлекательности различных регионов Китая со-

браны из Национального статистического годового отчета Китая. Этот авторитетный источник информации, выпущенный Национальным бюро статистики Китая, предоставляет подробную статистическую информацию о экономике, населении и обществе всех провинций, автономных регионов и муниципалитетов, непосредственно подчиненных центральному правительству. Собрав данные за период с 2008 по 2017 годы, мы получаем десятилетний промежуток времени для всестороннего анализа динамики изменений в инвестиционной привлекательности различных регионов Китая.

Данные об инвестиционной привлекательности стран АСЕАН: Данные об инвестиционной привлекательности стран АСЕАН получены из базы данных Всемирного банка. Всемирный банк, международная организация, предлагает широкий спектр экономических, социальных и экологических данных по всему миру. Доступ к этой базе данных позволяет получить данные об инвестиционной привлекательности стран АСЕАН с 1998 по 2014 год. С помощью 16-летних данных мы можем проанализировать инвестиционную привлекательность стран АСЕАН за последние два десятилетия.

1.2.2 Очистка данных и обработка выбросов

После сбора данных необходимо провести очистку данных и обработку выбросов, чтобы обеспечить точность и надежность данных, а также удалить возможные выбросы, способные исказить результаты анализа.

Очистка данных включает операции, такие как проверка данных, обработка пропущенных значений и преобразование формата данных. Тщательное изучение собранных данных обеспечивает целостность и последовательность данных. Если выявлены пропущенные значения, принимаются соответствующие меры для их заполнения или удаления. Кроме того, в зависимости от требований исследования, данные могут потребоваться преобразовать из исходного формата в числовые переменные, пригодные для анализа множественной линейной регрессии.

Обработка выбросов направлена на исключение экстремальных наблюдений, значительно отклоняющихся от нормального диапазона, обеспечивая надежность результатов анализа.

Путем сбора данных из авторитетных источников, таких как Национальный статистический годовой отчет Китая и база данных Всемирного банка, а также

выполнения очистки данных и обработки выбросов можно получить надежные данные об инвестиционной привлекательности Китая и пяти стран АСЕАН. Это обеспечивает логическую и академическую основу для последующего анализа множественной линейной регрессии.

1.2.3 Метод множественной линейной регрессии

Метод множественной линейной регрессии является неотъемлемым инструментом в различных научных дисциплинах, включая экономику, финансы, социальные науки и экологию. Он основан на принципах математической статистики и используется для анализа взаимосвязи между независимыми и зависимыми переменными, установления линейного уравнения для выражения этой взаимосвязи, а также для прогнозирования и объяснения.

Предполагая линейную связь между независимыми и зависимыми переменными, модель множественной линейной регрессии формулирует линейное уравнение, состоящее из коэффициентов регрессии и члена пересечения. Эти коэффициенты представляют величину и направление влияния независимых переменных на зависимую переменную, в то время как член пересечения обозначает значение зависимой переменной при нулевом значении всех независимых переменных. Лучшие оценки коэффициентов регрессии определяются путем минимизации суммы квадратов остатков, которые характеризуют расхождение между наблюдаемыми значениями и значениями, предсказанными линейным уравнением.

Эффективность и надежность моделей множественной линейной регрессии могут быть оценены с использованием статистических показателей, причем коэффициент детерминации (R^2) является одним из наиболее часто используемых метрик. Коэффициент детерминации R^2 количественно характеризует долю дисперсии зависимой переменной, объясненную моделью. Более высокое значение R^2 указывает на лучшую аппроксимацию данных моделью. Кроме того, скорректированный коэффициент детерминации может учесть влияние числа независимых переменных на R^2, а F-статистика может оценить значимость регрессионной модели.

Перед проведением анализа множественной линейной регрессии необходимо учитывать несколько предварительных требований и предположений. Во-первых, предполагается линейная связь, что ожидается одинаковый прирост

или убывание между независимыми и зависимыми переменными. Во-вторых, требуется независимость независимых переменных без мультиколлинеарности, когда одну независимую переменную можно предсказать на основе других. Кроме того, ошибка модели должна соответствовать определенным предположениям, таким как независимость и одинаковая распределенность ошибок, а также постоянная дисперсия.

Для обеспечения выполнения этих требований и предположений модели проводятся диагностические проверки. Они включают анализ графиков остатков, оценку нормальности и независимости остатков и другие.

В заключение, метод множественной линейной регрессии является логическим и академическим инструментом, используемым для моделирования и анализа взаимосвязи между независимыми и зависимыми переменными. Применяя этот подход, исследователи получают представление об влиянии и толковании переменных, что способствует прогнозированию и выводам.

1.2.4 Метод множественной шаговой регрессии

Метод множественной шаговой регрессии широко используется в исследованиях для изучения и построения моделей множественной линейной регрессии. В отличие от традиционной множественной линейной регрессии, этот подход поэтапно исключает независимые переменные для создания более краткой и информативной модели.

Основная цель этого метода - выявить независимые переменные, значительно влияющие на зависимую переменную, и исключить те, которые имеют слабую или незначительную объяснительную силу. Итерационный процесс удаления независимых переменных следует заранее определенным критериям исключения. Обычно эти критерии основаны на уровнях статистической значимости, таких как предопределенный порог p-значения для удаления. Также можно использовать другие показатели, такие как максимально скорректированный коэффициент детерминации или незначительная F-статистика.

Метод множественной шаговой регрессии имеет несколько преимуществ. Во-первых, он помогает выявить влиятельные переменные из множества потенциальных предикторов, что приводит к более лаконичной модели с улучшенными объяснительными возможностями. Это помогает устранить избыточную информацию и повысить интерпретируемость и прогностическую способность модели.

Во-вторых, метод эффективно решает проблемы мультиколлинеарности, вызванные высокой корреляцией между независимыми переменными. Во время процесса последовательного удаления обычно исключаются сильно коррелированные переменные, тем самым минимизируется негативное влияние коллинеарности на результаты модели. Более того, множественная шаговая регрессия учитывает малые объемы выборки, ограничивая количество независимых переменных, что уменьшает риск переобучения.

Необходимо проявлять осторожность при использовании методов множественной шаговой регрессии. Тщательно следует подходить к выбору критериев удаления, учитывая предметную область и цели исследования. Чрезмерное удаление может привести к исключению важных независимых переменных, а сохранение незначительных переменных может внести шум в модель. Кроме того, поскольку метод множественной шаговой регрессии основан на данных, результаты могут зависеть от конкретной выборки.

В заключение, метод множественной шаговой регрессии предоставляет возможность создания компактных и информативных моделей множественной линейной регрессии. Последовательное удаление независимых переменных позволяет выбрать подмножество, которое демонстрирует наиболее значимые и существенные взаимосвязи с зависимой переменной, повышая объяснительную силу и прогностическую способность модели. Однако при использовании этого метода необходимо внимательно подбирать критерии удаления и осуществлять тщательную проверку и оценку для обеспечения достоверности и валидности модели.

1.3 Эксперименты и результаты

1.3.1 Построение и анализ моделей инвестиционной привлекательности регионов Китая.

Данные об инвестиционной привлекательности Китая показывают растущую тенденцию с 2008 по 2018 год (см. Рисунок 1).

Таблица 1 представляет объем инвестиций в основные средства Китая (в миллиардах юаней) за период с 2008 по 2018 годы.

Анализ охватывает период с 2008 по 2017 годы. Для каждого года строится регрессионная модель, отражающая уровень инвестиций (y) на основе набора

Year	Объем инвестиций в основной капитал(100 млн юаней)
2008	172828
2009	224599
2010	251684
2011	311485
2012	374695
2013	446294
2014	512021
2015	562000
2016	606466
2017	641238
2018	645675

Таблица 1.1: Объем инвестиций в основной капитал Китая

объясняющих переменных $(x_k, k = 1, .., 10)$. Регрессионная модель строится для всей страны, рассматривая регионы как отдельные наблюдения.

Рассматриваемые факторы:

- $\hat{y_t}$ - логарифм оценки объема инвестиций в текущем году;
- $x_{1,t}$ - потребление электроэнергии;
- $x_{2,t}$ - среднедушевой доход в текущем году;
- $x_{3,t}$ - задолженность по кредитам, предоставленным юридическим лицам кредитными организациями в текущем году;
- $x_{4,t}$ - стоимость основных средств в текущем году;
- $x_{5,t}$ - расходы на научно-исследовательскую работу в текущем году;
- $x_{6,t}$ - объем деятельности в секторе "Строительство"в текущем году;
- $x_{7,t}$ - количество предприятий и организаций в текущем году;
- $x_{8,t}$ - оборот розничной торговли в текущем году;
- $x_{9,t}$ - ВВП на душу населения в текущем году;
- $x_{10,t}$ - уровень безработицы в текущем году (в процентах);

Уравнение представляет используемую регрессионную модель для логарифма оценки объема инвестиций в текущем году $\hat{y}_t$ на основе объясняющих переменных $(x_{1,t}, x_{2,t}, \cdots, x_{10,t})$.

$$\begin{aligned}\hat{y}_t = b_0 + b_1 x_{1,t} + b_2 x_{2,t} + b_3 x_{3,t} + b_4 x_{4,t} + b_5 x_{5,t} + b_6 x_{6,t} + b_7 x_{7,t} \\ + b_8 x_{8,t} + b_9 x_{9,t} + b_{10} x_{10,t}\end{aligned} \tag{1.1}$$

где t представляет номер года, с $t = 1, ..., 10$.

Для каждого года с номером $t(t = 1, ..., 10)$ имеется информация о каждом отдельном регионе. Всего существует 26 регионов с высоким уровнем инвестиций.

$y_{1,t}, \ldots, y_{26,t}$ представляют уровни инвестиций в регионах.

$x_{i,1,t}, \ldots, x_{i,10,t}$; где $i = 1, \ldots, 26$ и i обозначает номер региона.

На основе информации для года t, представленной на предыдущем слайде, с использованием метода наименьших квадратов строится регрессионная модель (Уравнение 1):

$$\sum_{i=1}^{26}(y_{i,t} - b_0 - b_1 x_{i,1,t} - b_2 x_{i,2,t} - .. - b_{10} x_{i,10,t})^2 \to \min_{b_0,..,b_{10}} \tag{1.2}$$

Для выполнения расчетов был применен анализ множественной линейной регрессии. С использованием инструмента "Регрессия"в дополнении "Анализ данных"в MS Excel мы проведем регрессионный анализ на доступных значениях столбцов вектора Y и X.

Модель множественной регрессии для 2017 года. Оценки коэффициентов, вычисленные с использованием инструмента регрессии в дополнении анализа данных MS Excel, представлены для наблюдаемых значений в Таблице 1.2 (стандартные ошибки коэффициентов указаны в скобках). Оценки параметров для уравнения регрессии в 2017 году следующие:

b10(c.o.)	b9(c.o.)	b8(c.o.)	b7(c.o.)	b6(c.o.)	b5(c.o.)
90.12045361	0.757670787	-0.074390939	0.002579218	0.678321663	-0.001846309
(1645.726812)	(0.426035564)	(8.61877992)	(0.008987614)	(0.302921691)	(0.001176357)

Таблица 1.2: Оценки параметров для уравнения регрессии в 2017 году

Модель записывается следующим образом:

b4(c.o.)	b3(c.o.)	b2(c.o)	b1(c.o.)	b0(c.o.)
0.760781591	-5.183897966	-0.399105119	-1.873479417	9312.257203
(0.457066773)	(3.137349872)	(0.15744744)	(1.834540235)	(8232.963495)

$$\begin{aligned}\hat{y}_t = {} & 9312.2572 - 1.8734 \cdot x_{1,t} - 0.3991 \cdot x_{2,t} - 5.1838 \cdot x_{3,t} + 0.7607 \cdot x_{4,t} \\ & - 0.00184 \cdot x_{5,t} + 0.6783 \cdot x_{6,t} + 0.00257 \cdot x_{7,t} - 0.07439 \cdot x_{8,t} \\ & + 0.7576 \cdot x_{9,t} + 90.1204 \cdot x_{10,t}\end{aligned} \tag{1.3}$$

Далее мы проверим, что общая форма изучаемых функциональных связей успешно определена. Это включает оценку качества и адекватности построенной модели на основе эмпирических данных.

Основными критериями оценки качества множественной регрессии являются коэффициент детерминации R^2 , множественный коэффициент корреляции R и скорректированный коэффициент детерминации R^2_{adj}.

Затем строятся модели для определения значений R^2, R и R^2_{adj} , как показано в Таблице 1.3.

R	0.967293678
R^2	0.93565706
R^2_{adj}	0.892761767

Таблица 1.3

$R^2 = 0.935$, что указывает на то, что примерно 94% вариации в изучаемой переменной y может быть "объяснено"комбинацией факторов, включенных в модель. Учитывая штраф, накладываемый за большое число объясняющих переменных, $R^2_{\text{adj}} = 0.892$, что означает, что уравнение регрессии объясняет 89% дисперсии зависимой переменной (в пределах наблюдаемых значений y). Множественный коэффициент корреляции $R = 0.976$, который очень близок к 1, указывает на сильную связь между y и набором предикторных переменных $(x_1, \cdots, x_{10})$.

Классический подход к проверке адекватности эконометрических моделей - это F-тест, который включает оценку значимости уравнения регрессии на основе критерия Фишера и сравнение этой оценки с критическим значением теста. Критическое значение зависит от доверительного интервала и степеней свободы k и $n - k$. Если вычисленное значение превышает критическое значение, то модель считается пригодной для анализа на выбранном уровне значимости.

Для нашей модели значение $F = 21.8125$, $n = 26$, $k = 10$. Давайте установим уровень значимости $\alpha = 0.05$, как это обычно делается в статистических исследованиях. Мы рассчитываем критическое значение F_{crit} как $F_{(0.05;10;15)} = 2.543$. Поскольку $F > F_{\text{crit}}$, модель является статистически значимой на уровне значимости $\alpha = 0.05$. Также, используя отчет Excel, мы можем оценить значимость уравнения путем сравнения p-значения F с α. Если $p - \text{value}(F) < \alpha$, то регрессионная модель считается статистически значимой на заданном уровне значимости. Для анализируемой модели $p - \text{value}(F) = 3.72 \times 10^{-7} (< 0.05)$.

Таким образом, уравнение регрессии является статистически значимым на уровне значимости $\alpha = 0.05$, что подтверждает результаты F-теста.

Оценка качества параметров уравнения регрессии. Мы сравниваем абсолютные значения вычисленной t-статистики коэффициентов полученной регрессионной модели с табличными значениями. Оценка параметра считается значимой, если абсолютное вычисленное (наблюдаемое) значение превышает табличное значение.

Для заданного уровня значимости $\alpha = 0.05$ и степеней свободы $df = 15$, теоретическое (табличное) значение t-статистики равно $t_{\text{table}} = 2.131$.

В Таблице 1.4 представлены вычисленные значения t-статистики для соответствующих коэффициентов регрессионной модели. Ясно, что только оценки для α_2 и α_6 являются статистически значимыми на уровне значимости 5%. В этой таблице также приведены p-значения для наблюдаемых t-статистик. P-значение (t) представляет собой критическое значение уровня значимости α для текущей величины t-статистики. Если p-значение меньше указанного α, то коэффициент считается статистически значимым. С точки зрения p-значения, оценка значимости параметров в построенном уравнении регрессии приводит к тем же результатам, что и t-тест.

	t Stat	P-value
b0	1.182053995	0.254454547
b1	-1.094048678	0.29013139
b2	-3.914915737	0.001234375
b3	-1.721193368	0.104490736
b4	1.746797082	0.099837148

Таблица 1.4: t-статистики параметров уравнения регрессии для 2017 года.

	t Stat	P-value
b5	-1.76043987	0.097432206
b6	2.345780882	0.032203352
b7	0.320095135	0.753038323
b9	1.867411275	0.080274135
b10	0.05782103	0.954607156

Таблица 1.5: t-статистики параметров уравнения регрессии для 2017 года.

Таким образом, построенная модель демонстрирует высокое качество аппроксимации. Однако из 11 оценок параметров только 2 оказались значимыми. Вероятно, это связано с наличием множественных линейных факторов в модели. Учитывая относительно большое количество объясняющих переменных и относительно небольшой размер выборки, разумно предположить, что некоторые показатели могут дублировать друг друга. Поэтому необходимо скорректировать набор факторов, используя метод последовательной регрессии. Мы последовательно исключим переменные, соответствующие статистически незначимым оценкам коэффициентов, пока все они не станут значимыми на заданном уровне значимости α.

Для "просеивающего"показателя мы будем рассматривать p-значение. Из таблицы 1.5 определим наибольшее p-значение. Коэффициент b_8 для переменной x_8 имеет наименьшее абсолютное значение t-статистики и, следовательно, наибольшее p-значение. Поэтому мы исключим переменную x_8 из модели. В результате матрица исходных данных примет вид:

$X = (x_{1i}, x_{2i}, \cdots, x_{7i}, x_{9i}, x_{10i})^\top,\ i = 1, \cdots, 26.$

Для $y = (y_1, \cdots, y_{26})^\top$ и нового X с использованием MS Excel пересчитаем оценки параметров уравнения регрессии.

Коэффициент детерминации и коэффициент множественной корреляции остаются неизменными, в то время как R^2_{adj} увеличивается, что указывает на то, что удаленная объясняющая переменная является статистически незначимой. Обновленная модель теперь имеет скорректированный коэффициент детерминации $R^2_{adj} = 0.899$, что указывает на достаточно высокий уровень качества. Уровень значимости вычисленного значения критерия Фишера ниже по сравнению с предыдущим уравнением. С $P - \text{value}(F) = 6.97669 \cdot 10^{-8}$ преобразованная регрессионная модель считается статистически значимой при $\alpha = 0.05$.

Оценки коэффициентов:

b10(c.o.)	b9(c.eo.)	b7(c.o.)	b6(c.o.)	b5(c.o.)
91.62080512	0.75698959	0.002606575	0.678749549	-0.001850222
(1584.558496)	(0.405368437)	(0.008143127)	(0.289349084)	(0.001051)

Таблица 1.6: Оценки параметров уравнения регрессии для 2017 года без x_8

Оценка качества оценок на основе коэффициентов нового уравнения. Наблюдаемые значения t-статистики и соответствующие им p-значения представлены

b4(c.o.)	b3(c.o.)	b2(c.o.)	b1(c.o.)	b0(c.o.)
0.760061396	-5.18729215	-0.400113466	-1.869145659	9322.593323
(0.435117166)	(3.013776515)	(0.102202319)	(1.708466631)	(7886.774517)

в Таблице 1.5.

Учитывая постоянный член, имеется 10 коэффициентов. Исходя из оценочных значений и в соответствии с результатами t-теста, только 2 из них являются значимыми: b_2 и b_6. P-значение достигает максимального значения при $t = 0.0578$, рассчитанного для оценки b_{10}. Здесь b_{10} представляет собой коэффициент для объясняющей переменной x_{10}. В результате x_{10} исключается из набора факторов. Таким образом,

$X = (x_{1i}, x_{2i}, \cdots, x_{7i}, x_{9i})^\top,\ i = 1, \cdots, 26.$

Статистика регрессии модели после удаления фактора x_{10} отличается лишь незначительно по сравнению со статистикой предыдущей модели (см. Таблицу 1.7):

R	0.9672865
R^2	0.935643
R^2_{adj}	0.905357

Таблица 1.7

Наблюдаемая F-статистика ($F = 37.1309$) превышает значения, соответствующие предыдущим уравнениям, с $P - \text{value}(F) = 1.95853 \cdot 10^{-9}$, что указывает на высокий уровень качества аппроксимации.

Оценки параметров модели (см. Таблицу 1.8) в основном являются статистически незначимыми. Оценка коэффициента b_7 для объясняющей переменной x_7 имеет наибольшее значение $P - \text{value}(F) = 0.7458$. Таким образом, мы исключаем фактор x_7, поскольку он не является статистически значимым в анализируемой модели. Его отсутствие статистической значимости дополнительно подтверждается качественной оценкой значений R^2 и R уравнений после исключения, которые остаются практически на том же уровне. Скорректированный R^2_{adj} увеличивается в связи с уменьшением количества объясняющих переменных ($R^2_{\text{adj}} = 0.910$). Значение $P - \text{value}(F)$ уменьшается, и доверительный интервал, при котором модель является статистически значимой на уровне значимости $\alpha = 0.05$, расширяется.

Путем исключения наименее значимых объясняющих переменных x_1 и x_4 из

		c.o.	t Stat	P-value
b0	10034.49788	3991.129234	2.51420019	0.021658637
b1	-1.610506826	1.429250709	-1.12681898	0.274617446
b2	-0.402520397	0.094263852	-4.270145875	0.000460606
b3	-5.433737956	2.681372429	-2.026476403	0.057795086
b4	0.734552101	0.397333472	1.848704307	0.0809958
b5	-0.001708463	0.000903599	-1.8907318	0.074869
b6	0.743294813	0.198865735	3.737671606	0.001506765
b9	0.769269896	0.376296307	2.044319548	0.055832772

Таблица 1.8: Оценки параметров уравнения регрессии для 2017 года без x_8,x_{10} ,x_7

моделей, построенных в следующих двух шагах, мы получаем модель, в которой все оценки параметров являются статистически значимыми при $\alpha = 0,05$. Значение $P - \text{value(F)} = 2.8909 \cdot 10^{-10}$ указывает на то, что модель является статистически значимой.

Модель имеет следующий вид:

$$\begin{aligned}\hat{y}_t \underset{(t-Stat)}{=} & \underset{(2.8720)}{10712.1937} - \underset{(-5.2899)}{0.4501} \cdot x_{2,t} - \underset{(-2.4570)}{6.7041} \cdot x_{3,t} - \underset{(-3.1611)}{0.002625} \cdot x_{5,t} \\ & + \underset{(3.6163)}{0.7599} \cdot x_{6,t} + \underset{(4.9829)}{1.2368} \cdot x_{9,t}\end{aligned} \tag{1.4}$$

Общее качество принятого уравнения регрессии достаточно высоко, как подтверждают значения показателей R^2, R^2_{adj} и R, указанные в Таблице 1.9.

Таблица 1.9: Dynamic Quality Metrics during Variable Selection in 2017

Model	**Multiple R**	**R-Squared**	**Norm. R-Squared**	**F-Value**	**p-value(F)**	AIC	p(RNT)	p(HT)
1,2,3,4,5,6,7,**8**,9,10	0.967	0.935	0.892	21.81	3.7254E-07	519.3	0.8756	0.3478
1,2,3,4,5,6,7,9,**10**	0.967	0.935	0.899	25.85	6.9766E-08	517.3	0.8771	0.3478
1,2,3,4,5,6,**7**,9	0.967	0.935	0.905	30.89	1.1947E-08	515.3	0.8831	0.3479
1,2,3,4,5,6,9	0.967	0.935	0.910	37.13	1.9585E-09	513.5	0.8059	0.3757
2,3,**4**,5,6,9	0.964	0.930	0.908	42.50	5.1728E-10	**513.3**	0.6676	0.8337
2,3,5,6,9	0.958	0.919	0.899	45.69	2.8909E-10	515.1	0.46	0.6470

Модель с наименьшим значением AIC:

$$\begin{aligned}\hat{y}_t = & 8484.5941 - 0.375 \cdot x_{2,t} - 5.4173 \cdot x_{3,t} + 0.6973 \cdot x_{4,t} \\ & - 0.0019 \cdot x_{5,t} + 0.7458 \cdot x_{6,t} + 0.7238 \cdot x_{9,t}\end{aligned} \tag{1.5}$$

Представленная модель (4) характеризует 89,9% вариации исходных значений y_i, где $i = 1, \cdots, 26$ представляет номер наблюдения (региона), с учетом

штрафов за каждую новую объясняющую переменную. Коэффициент множественной корреляции равен $R = 0,958$, что указывает на сильную линейную связь между зависимой переменной и рассматриваемым набором факторов. Модель оказалась статистически значимой на основе критерия Фишера при уровне значимости $\alpha = 0,05$ с $F = 45.6923$. Для вычисленного значения F-статистики $P - \text{value}(F) = 2.89 \cdot 10^{-10}$.

Диаграммы теоретических значений y_i и фактических значений y для номеров наблюдений $i = 1, \cdots, 26$ (см. Рис. 1.1) являются убедительным доказательством хорошего качества аппроксимации между исходными значениями y и изучаемой моделью.

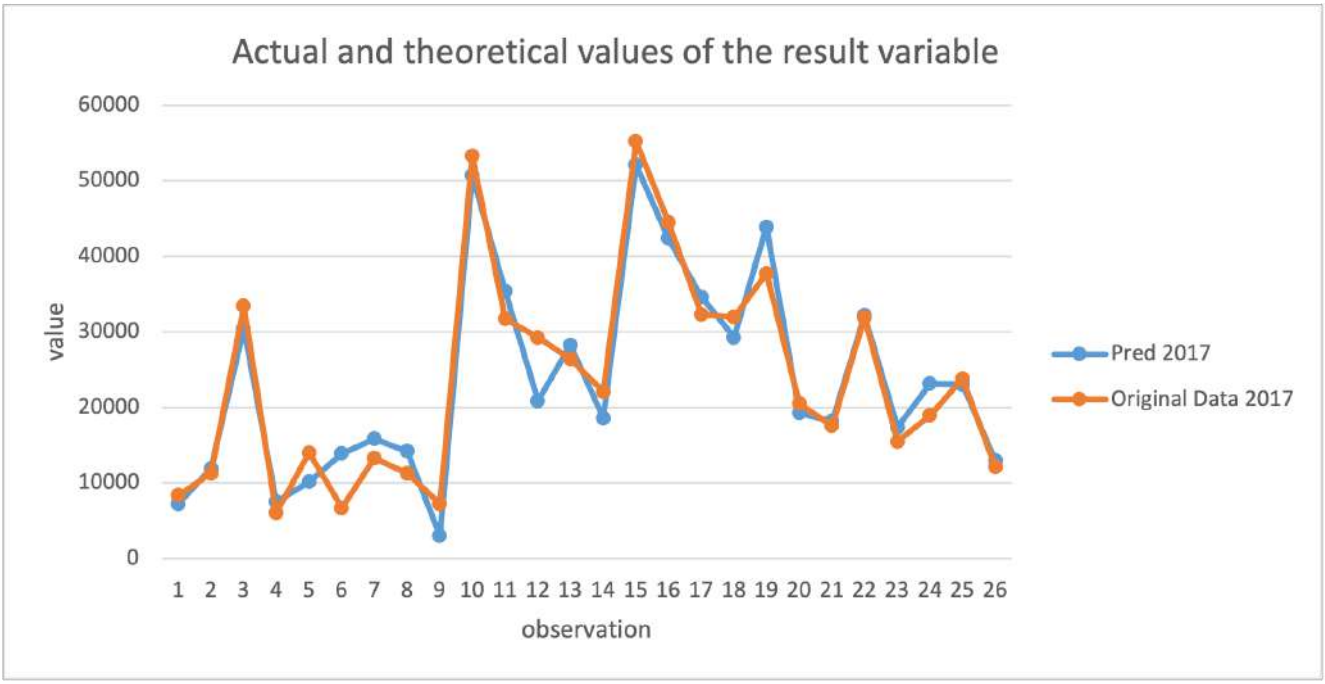


Рис. 1.1

С помощью критерия Дарбина-Уотсона можно определить, наличие автокорреляции в остатках или нет. Оцененная статистика d-критерия (DW) имеет следующий вид:

$$DW = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

Суть критерия заключается в сравнении эмпирического значения DW с табличными нижним (d_L) и верхним (d_U) статистиками, которые зависят от размера выборки, уровня значимости и количества объясняющих переменных в модели. Если $DW > 2$, мы рассчитываем скорректированное значение $DW' =$

$4 - DW$.

В соответствии с условиями регрессионного анализа ($\alpha = 0,05, n = 26, k = 15$), табличные значения составляют $d_l = 0,256$ и $d_U = 3,179$. Отсутствие автокорреляции можно считать подтвержденным, если выполняются оба условия: $DW > d_U$ и $4 - d_U > DW$. В прикладном статистическом анализе данных обычно считается, что при показателях Durbin-Watson от 1.5 до 2.5 регрессионная модель может считаться адекватной. В нашем уравнении $DW' = 2,596$, и наблюдаемое значение попадает в "зону неопределенности где $d_L < DW' < d_U$.

Давайте вычислим среднюю относительную ошибку аппроксимации $E_{\text{rel.}}$, чтобы оценить точность регрессионной модели, с использованием следующей формулы:

$$E_{\text{rel.}} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{\varepsilon}_i|}{\ln y_i} \cdot 100\%.$$

Для построенного уравнения регрессии это значение составляет $E_{\text{rel.}} = 1,75\%$, что указывает на крайне небольшую ошибку модели (около 2%).

Таким образом, разработанная модель регрессии в форме уравнения (1.4), учитывая ее высокие показатели качества и минимальную ошибку аппроксимации, демонстрирует статистически значимое влияние набора объясняющих факторов на значение зависимой переменной. Она также подходит для дальнейшего анализа вероятностных зависимостей между lny и переменными x_2, x_3, x_5, x_6, x_9 в рамках 2017 года.

Множественная регрессионная модель для 2016 года

На основе данных за 2016 год, предоставленных Национальным бюро статистики (Китай), была составлена вычислительная таблица для определения параметров уравнения регрессии, которое принимает следующую форму:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}, x_{10i}), \quad i = 1..26. \tag{1.6}$$

где: i представляет номер региона, на который идет исследование;
y_i - логарифм оценки объема инвестиций в текущем году в i-м регионе;
x_{1i} - потребление электроэнергии в i-м регионе;
x_{2i} - среднедушевой доход в текущем году в i-м регионе;

	Coefficients	c.o.
b0	3262.169924	4832.230643
b1	-0.492576567	1.746298859
b2	-0.284674564	0.137878968
b3	-2.647672589	2.925816588
b4	1.36612506	0.335356386
b5	-0.000111808	7.46915E-05
b6	8.62503E-05	3.41244E-05
b7	0.002335237	0.008644545
b8	-0.475518489	7.629315583
b9	-0.040903966	0.284922846
b10	915.815172	1211.023343

Таблица 1.10: Оценки параметров уравнения регрессии для 2016 года.

x_{3i} - долг по кредитам, предоставленным юридическим лицам кредитными организациями в текущем году в i-м регионе;
x_{4i} - стоимость основных фондов в текущем году в i-м регионе;
x_{5i} - расходы на научные исследования в текущем году в i-м регионе;
x_{6i} - объем деятельности в сфере "Строительство"в текущемi-м регионе;
x_{8i} - оборот розничной торговли в текущем году в i-м регионе;
x_{9i} - ВВП на душу населения в текущем году в i-м регионе;
x_{10i} - уровень безработицы в текущем году (в процентах) в i-м регионе;

Оценки коэффициентов рассчитываются с использованием инструмента регрессии во встроенном анализе данных MS Excel для наблюдаемых значений $y = (y_1, ..., y_26)^\top$ и $X = (x_{1i}, x_{2i}, \cdots, x_{10i})^\top$, где $i = 1, \cdots, 26$. Эти оценки представлены в виде, показанном в Таблице 1.10:

Общие критерии качества модели указывают на высокий уровень ее аппроксимационных возможностей. Коэффициент множественного детерминации составляет $R^2 = 0,937$, а после корректировки на штраф, накладываемый количеством объясняющих переменных, скорректированный коэффициент детерминации становится $R^2_{adj} = 0,906$. Кроме того, коэффициент множественной корреляции приближается к единице с $R = 0,968$. Наблюдаемое значение F-статистики равно $F = 30.066$. При заданном уровне значимости $\alpha = 0,05$ критическое значение равно $F_{0,05;10;15} = 2,544$. Таким образом, модель является статистически значимой и объясняет почти всю дисперсию зависимой пере-

менной. Однако, несмотря на общее хорошее качество регрессионной модели, оценки параметров для объясняющих переменных не считаются статистически значимыми для выбранного уровня α. (См. Таблицу 1.11)

	t Stat	P-value
b0	0.675085724	0.507351564
b1	-0.282068882	0.780788875
b2	-2.064669959	0.052164857
b3	-0.904934574	0.376272554
b4	4.073651542	0.000592215
b5	-1.496936962	0.150025462
b6	2.527527177	0.02001929
b7	0.270139899	0.789820503
b8	-0.062327804	0.950920479
b9	-0.143561551	0.887283334
b10	0.756232469	0.458325579

Таблица 1.11: t-статистики для 2016 года

Давайте продолжим методологию последовательного исключения незначимых факторов из уравнения регрессии. На основе полученных результатов на втором этапе строится модель без x_8 ($b_8 = 0$). Новая модель имеет практически такое же качество, как и предыдущая, но она статистически более значима согласно F-тесту Фишера.

Динамика изменения показателей R^2,R, R^2_{adj} и F при последовательном исключении незначимых факторов из множественной регрессионной модели представлена в Таблице 1.12.

Explanatory variables	R	R^2	$R_{\text{adj}}{}^2$	$F_{\text{obs.}}$	p-value(F)	AIC	p(RNT)	p(HT)
x1,x2,x3,x4,x5,x6,x7,***x*9**,x10	0.9592	0.9201	0.8751	20.4764	3.74E-07	519.6	0.6831	0.4306
***x*1**,x2,x3,x4,x5,x6,x7,x10	0.9591	0.9199	0.8823	24.4319	7.27E-08	517.6	0.6685	0.4313
x2,x3,x4,x5,x6,***x*7**,x10	0.9589	0.9196	0.8883	29.4284	1.31E-08	515.7	0.6585	0.4396
x2,x3,x4,x5,x6,***x*10**	0.9589	0.9195	0.8940	36.1754	2.08E-09	511.8	0.4813	0.4706
x2,***x*3**,x4,x5,x6	0.9577	0.9172	0.8965	44.3232	3.80E-10	**510.6**	0.5182	0.6497
x2,x4,***x*5**,x6	0.9533	0.9088	0.8914	52.3272	1.26E-10	511.1	0.4665	0.7968
x2,x4,x6	0.94950	0.9015	0.8881	67.1576	3.10E-11	511.1	0.5968	0.4702

Таблица 1.12: Динамика показателей качества во время процесса выбора объясняющих переменных в 2016 году

Модель с наименьшим значением AIC:

$$\begin{aligned}\hat{y}_t =& 7356.1238 - 0.2947 \cdot x_{2,t} - 3.3135 \cdot x_{3,t} + 1.2733 \cdot x_{4,t} \\ & - 0.0001 \cdot x_{5,t} + 0.00009029 \cdot x_{6,t}\end{aligned} \tag{1.7}$$

Исключение переменных x_8, x_9, x_1, x_7 приводит к модели с максимальным значением R^2_{adj}, указывающим на их незначительное влияние на полученный показатель. Однако проверка качества оценок коэффициентов для отдельных переменных с использованием t-теста Стьюдента не позволяет рассматривать модель как высокочувствительную, так как p-значение для t-статистики b_3 значительно сильнее выбранного уровня значимости α.

После исключения x_3, x_5 получается регрессионная модель, которая включает только те факторы, чьи коэффициенты являются статистически значимыми при $\alpha = 0,05$. Модель также обладает удовлетворительным качеством аппроксимации (см. Таблицу 1.12) и имеет следующий вид:

Исключение переменных x_8, x_9, x_1, x_7 приводит к модели с максимальным значением R^2_{adj}, указывающим на их незначительное влияние на полученный показатель. Однако, тестирование качества оценок коэффициентов для отдельных переменных с использованием t-теста Стьюдента не позволяет считать модель высоко значимой, так как p-значение для t-статистики b_3 значительно превышает выбранное значение α.

После исключения переменных x_3, x_5 получается регрессионная модель, которая включает только те факторы, коэффициенты которых являются статистически значимыми при $\alpha = 0,05$. Модель также обладает удовлетворительным качеством аппроксимации (см. Таблицу 1.12) и имеет следующий вид:

$$\underset{(t-Stat)}{\hat{y}_t =} \underset{(3.5024)}{9857.72} - \underset{(-4.2072)}{0.38741} \cdot x_{2,t} - \underset{(8.3246)}{1.1236} \cdot x_{4,t} + \underset{(2.9978)}{0.000059547} \cdot x_{6,t} \tag{1.8}$$

Графики значений y_i и $\hat{y}_i$ для номеров наблюдений $i, i = 1, \cdots, 26$ отображены на рисунке 1.2.

$E_{\text{rel.}} = 1,49\%$, следовательно, способность аппроксимации построенной модели относительно высокая, позволяющая практически полностью восстановить исходные данные.

Связь между остатками соседних наблюдений является слабой, что указывает на независимость остатков регрессии. Таким образом, модель, представленная на рисунке 1.2, построенная на основе данных за 2016 год, обладает

Рис. 1.2

достаточной способностью к аппроксимации. Оценки качества подтверждают правильность ее спецификации, а анализ остатков подтверждает ее адекватность.

Наблюдаемое значение статистики Дарбина-Уотсона равно $DW = 2,075$. При $\alpha = 0,05, n = 26$ и $k = 15$, табличные значения составляют $d_L = 0,256$ и $d_U = 3,179$. Поскольку $d_L < DW < d_U$, невозможно определить наличие или отсутствие автокорреляции.

Таким образом, модель, представленная на рисунке 1.2, построенная на основе данных за 2016 год, обладает достаточной способностью к аппроксимации. Оценки качества указывают на правильность ее спецификации, а анализ остатков подтверждает ее адекватность.

Множественная регрессионная модель для лет 2008-2015

В соответствии с вышеупомянутой процедурой, используемой для разработки регрессионных моделей для 2017 года, аналогичный подход применяется для построения моделей индивидуально для каждого года в период с 2008 по 2015 годы.

Множественная регрессионная модель для 2008 года

В отношении конкретного года 2008 после тщательного анализа и обработки доступных данных получается следующая регрессионная модель:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.0111)}{1687.51} - \underset{(-2.7983)}{0.1026} \cdot x_{2,t} - \underset{(24.2854)}{0.9663} \cdot x_{4,t} \tag{1.9}$$

В дальнейшем (см. Рис. 1.3) представлены графики, иллюстрирующие взаимосвязь между наблюдаемыми (y_i) и регрессионными $(\hat{y}_i)$ значениями полученного показателя по отношению к номеру наблюдения $i, i = 1, \cdots, 26$.

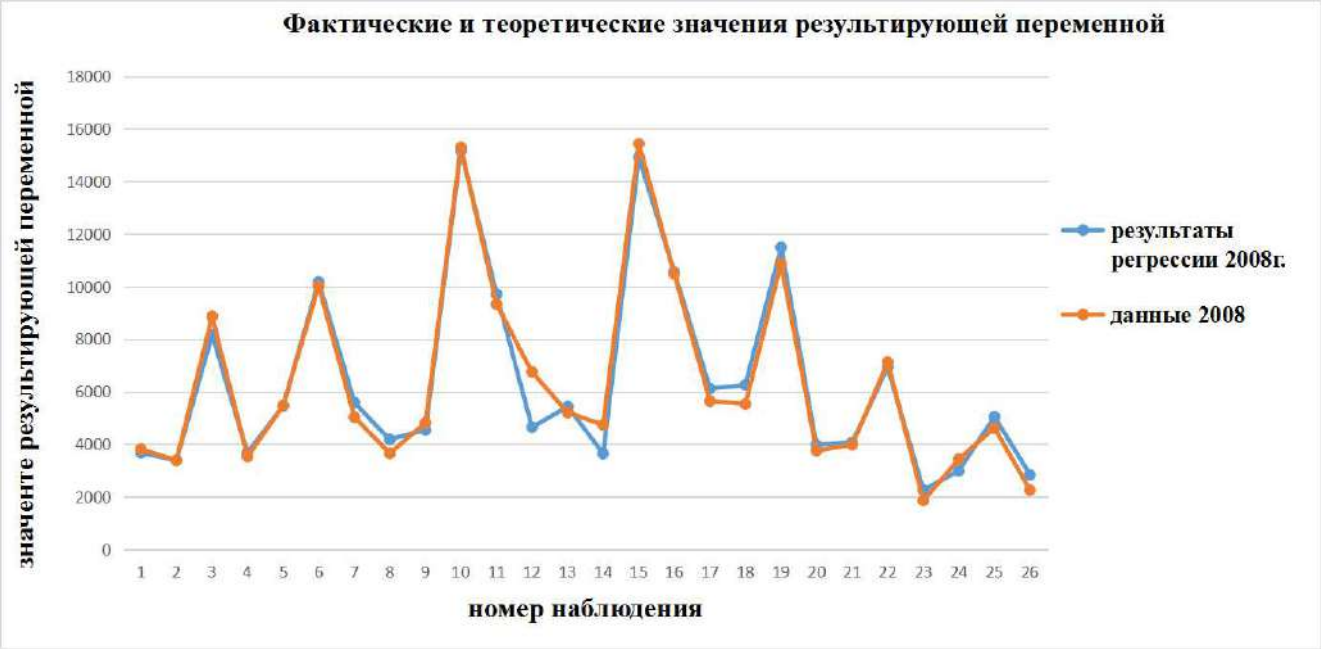


Рис. 1.3

Таблица 1.13 представляет изменения показателей качества регрессии при выборе факторов. В последней строке приведены значения R^2,R,R^2_{adj} и F для модели (1.9).

С приближенным значением, близким к 1, величина R^2_{adj} указывает на сильную связь между y и x_2. Кроме того, учитывая x_4, мы также можем заключить, что модель эффективно аппроксимирует наблюдаемые значения полученной переменной, о чем свидетельствует $R = 0,9819$. Средняя относительная ошибка аппроксимации составляет $E_{\text{rel.}} = 0,86\%$. Общая статистическая значимость регрессионной модели на уровне значимости 5% подтверждается тестом Фишера: наблюдаемая статистика $F = 309,1741$ значительно превышает критическое значение $F_{(0,05;10;26-10-1)} = 2,543$.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9856	0.9715	0.9525	51.1401	9.45E-10	428.6	0.3739	0.2070
x1, x2, x3, x4, x5, x6, x8, x9, **x10**	0.9856	0.9714	0.9554	60.5808	1.16E-10	426.6	0.5827	0.2082
x1, x2, x3, x4, x5, x6, x8, **x9**	0.9856	0.9714	0.9581	72.3808	1.31E-11	424.6	0.5616	0.2093
x1, x2, x3, x4, x5, x6, x8	0.9855	0.9713	0.9601	87.0548	1.40E-12	422.8	0.5572	0.2486
x2, x3, x4,**x5**, x6, x8	0.9855	0.9712	0.9621	107.0218	1.29E-13	420.8	0.2571	0.2946
x2, **x3**, x4, x6, x8	0.9853	0.9708	0.9636	133.4335	1.19E-14	419.1	0.3136	0.2674
x2, x4, x6, **x8**	0.9839	0.9681	0.9620	159.3198	2.17E-15	419.5	0.2263	0.3083
x2, x4, **x6**	0.9831	0.9665	0.9620	212.0222	2.22E-16	418.7	0.3676	0.4269
x2, x4	0.9819	0.9641	0.9610	309.1741	2.39E-17	**418.6**	0.2075	0.3233

Таблица 1.13: Динамика показателей качества в процессе выбора объясняющих переменных в 2008 году

Множественная регрессионная модель для 2009 года

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(4.2544)}{4642.4701} - \underset{(-3.9761)}{0.2680} \cdot x_{2,t} + \underset{(11.1866)}{0.9244} \cdot x_{4,t} + \underset{(2.3127)}{0.000039604} \cdot x_{6,t} \qquad (1.10)$$

Значения R^2,R,R^2_{adj} и F для каждой модели, построенной с последовательным исключением статистически незначимых переменных, представлены в Таблице 1.14. На Рисунке 1.4 изображены различия значений зависимой переменной между наблюдаемыми и рассчитанными значениями с использованием регрессионной модели 1.10. Значение $E_{\text{rel.}} = 1,31\%$ указывает на то, что средняя ошибка аппроксимации изученных данных моделью составляет 1%.

На основе значений R^2,R,R^2_{adj} качество построенной модели (1.10) можно считать хорошим. Статистическая значимость на уровне значимости 5% была подтверждена результатом F-теста. Так как $F = 102.59$ (в Таблице 1.14), а $F_{(0,05;10;26-10-1)} = 2,543$, неравенство выполняется:

$$F > F_{(0,05;10;26-10-1)}$$

Наблюдаемое значение статистики Дарбина-Уотсона составляет $DW = 1,9732$. При $\alpha = 0,05, n = 26, k = 15$, табличные значения равны $d_L = 0,256$ и $d_U = 3,179$. Поскольку $d_L < DW < d_U$, невозможно определить наличие или отсутствие автокорреляции.

Множественная регрессионная модель для 2010 года

Теперь перейдем к множественной регрессионной модели для 2010 года. После последовательного исключения регрессоров было получено следующее урав-

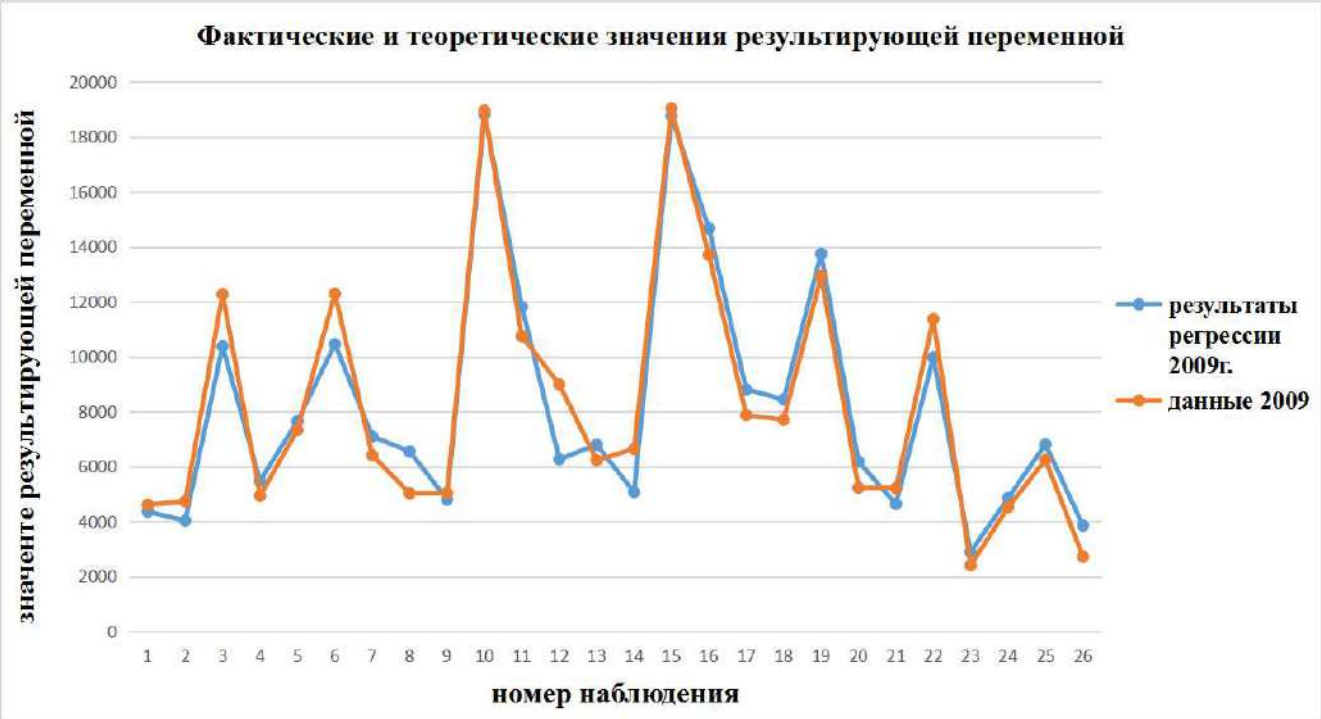


Рис. 1.4

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, **x3**, x4, x5, x6, x7, x8, x9, x10	0.9692	0.9394	0.8991	23.2695	2.39E-07	459.0	0.2118	0.2706
x1, x2, x4, x5, x6, **x7**, x8, x9, x10	0.9691	0.9393	0.9051	27.5183	4.42E-08	457.0	0.2020	0.2127
x1, x2, x4, x5, x6, **x8**, x9, x10	0.9691	0.9392	0.9107	32.8815	7.34E-09	455.0	0.2065	0.2755
x1, x2, x4, x5, x6, x9, **x10**	0.9687	0.9385	0.9146	39.2845	1.23E-09	453.4	0.2635	0.2137
x1, x2, x4, x5, x6, x9	0.9684	0.9379	0.9183	47.8724	1.82E-10	451.6	0.2518	0.2697
x2, x4, x5, x6,**x9**	0.9674	0.9359	0.9359	58.4359	3.02E-11	450.4	0.2692	0.3069
x2, x4, **x5**, x6	0.9668	0.9347	0.9222	75.1750	3.88E-12	448.9	0.2086	0.4119
x2, x4,x6	0.9661	0.9332	0.9241	102.5999	4.36E-13	**447.5**	0.2433	0.5307

Таблица 1.14: Динамика показателей качества в процессе выбора объясняющих переменных в 2009 году

нение регрессии:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.2979)}{4682.4156} - \underset{(-2.9976)}{0.2274} \cdot x_{2,t} + \underset{(14.3482)}{1.067} \cdot x_{4,t} \tag{1.11}$$

Результаты регрессии графически изображены на Рисунке 1.5.

Полный обзор построенных моделей на каждом этапе выбора переменных представлен в Таблице 1.15.

Модель с наименьшим значением AIC:

$$\hat{y}_t = 6030.6164 - 0.3115 \cdot x_{2,t} + 0.9447 \cdot x_{4,t} + 0.00003703 \cdot x_{6,t} \tag{1.12}$$

Рис. 1.5

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, x7, **x8**, x9, x10	0.9609	0.9234	0.8723	18.0894	1.31×10^{-06}	475.0	0.2175	0.2643
x1, x2, x3, x4, x5, x6, x7, x9, **x10**	0.9608	0.9232	0.8801	21.3856	2.74×10^{-07}	473.1	0.1922	0.2715
x1, x2, x3, x4, x5, x6, **x7**, x9	0.9605	0.9226	0.8862	25.3526	5.48×10^{-08}	471.3	0.1979	0.2965
x1, x2, **x3**, x4, x5, x6, x9	0.9599	0.9214	0.8909	30.1653	1.07×10^{-08}	469.7	0.1805	0.3549
x1, x2, x4, **x5**, x6, x9	0.9594	0.9204	0.8953	36.6491	1.86×10^{-09}	468.0	0.2333	0.4098
x1, x2, x4, x6, **x9**	0.9584	0.9186	0.8982	45.1438	3.22×10^{-10}	466.6	0.2021	0.5352
x1, x2, x4, x6	0.9575	0.9169	0.9011	57.9834	4.76×10^{-11}	465.2	0.2187	0.6726
x2, x4, **x6**	0.9567	0.9153	0.9037	79.2873	5.94×10^{-12}	**463.7**	0.2266	0.4168
x2, x4	0.9489	0.9004	0.8917	103.9760	3.02×10^{-12}	465.9	0.3722	0.5357

Таблица 1.15: Динамика показателей качества при выборе объясняющих переменных в 2010 году

Качественные характеристики для Таблицы 1.15 представлены в последней строке таблицы. Как видно, значения коэффициента множественной корреляции, коэффициента детерминации и скорректированного коэффициента детерминации чрезвычайно высокие (>0.85), что указывает на хорошее качество построенной модели.

Значение p-значения для наблюдаемой F-статистики равно $P - \text{value}(F) = 5.9492 \cdot 10^{-12}$, что демонстрирует статистическую значимость модели при $\alpha = 0.05$. Ошибка аппроксимации незначительна, с $E_{\text{rel.}} = 1,38\%$.

Для проверки предположения о независимости остатков регрессии вычисляется линейный коэффициент автокорреляции.

Множественная регрессионная модель для 2011 года

Для 2011 года была получена следующая множественная регрессионная модель:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.8588)}{6043.8067} - \underset{(-3.8596)}{0.2865} \cdot x_{2,t} + \underset{(11.2310)}{0.8883} \cdot x_{4,t} + \underset{(2.6593)}{0.000039970} \cdot x_{6,t} \tag{1.13}$$

Средняя относительная ошибка аппроксимации составляет $E_{\text{rel.}} = 1,16\%$. Модель точно восстанавливает наблюдаемые значения y, что подтверждается графиками, показанными на Рисунке 1.6.

Рис. 1.6

В соответствии с данными в Таблице 1.16, конечное уравнение регрессии демонстрирует хорошие оценки качества. Гипотеза о том, что все коэффициенты модели равны нулю, отвергается с помощью F-теста. Наблюдаемое значение статистики Фишера значительно превышает табличное значение для данного α ($F_{\text{obs}} = 102.2490, F_{(0,05;10;26-10-1)} = 2,543$), что указывает на статистическую значимость уравнения 1.13.

Множественная регрессионная модель для 2012 года

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(4.1382)}{8342.0697} - \underset{(-4.0912)}{0.34} \cdot x_{2,t} + \underset{(10.0552)}{0.8786} \cdot x_{4,t} + \underset{(3.5651)}{0.00005317} \cdot x_{6,t} \tag{1.14}$$

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, **x3**, x4, x5, x6, x7, x8, x9, x10	0.9692	0.9394	0.8991	23.2695	2.39E-07	459.0	0.2118	0.2706
x1, x2, x4, x5, x6, **x7**, x8, x9, x10	0.9691	0.9393	0.9051	27.5183	4.42E-08	457.0	0.2020	0.2127
x1, x2, x4, x5, x6, **x8**, x9, x10	0.9691	0.9392	0.9107	32.8815	7.34E-09	455.0	0.2065	0.2755
x1, x2, x4, x5, x6, x9, **x10**	0.9687	0.9385	0.9146	39.2845	1.23E-09	453.4	0.2635	0.2137
x1, x2, x4, x5, x6, x9	0.9684	0.9379	0.9183	47.8724	1.82E-10	451.6	0.2518	0.2697
x2, x4, x5, x6, **x9**	0.9674	0.9359	0.9359	58.4359	3.02E-11	450.4	0.2692	0.3069
x2, x4, **x5**, x6	0.9668	0.9347	0.9222	75.1750	3.88E-12	448.9	0.2086	0.4119
x2, x4, x6	0.9661	0.9332	0.9241	102.5999	4.36E-13	**447.5**	0.2433	0.5307

Таблица 1.16: Динамика показателей качества при выборе объясняющих переменных в 2011 году

В Таблице 1.17 представлены значения показателей качества для моделей, построенных на каждом шаге последовательного удаления статистически незначимых параметров. Конечная модель, в которой все параметры являются значимыми на уровне 5%, считается значимой и обладает отличными оценками.

Коэффициент множественной корреляции равен $R = 0.9606$, коэффициент детерминации составляет $R^2 = 0.92287$, а скорректированный коэффициент детерминации с учетом числа факторов равен $R^2_{adj} = 0.91235$. Таким образом, модель (1.14) объясняет примерно 90% вариации наблюдаемых значений зависимой переменной $y_i (i = 1, \cdots, 26)$. Адекватность модели при $\alpha = 0,05$ подтверждается низким p-значением $P - \text{value}(F) = 2.1425 \cdot 10^{-12}$ (для наблюдаемого значения статистики Фишера $F_{\text{obs}} = 87.7465$).

Рис. 1.7

График, иллюстрирующий взаимосвязь между исходными значениями y_i и номером наблюдения $i (i = 1, \cdots, 26)$, а также график, показывающий пред-

сказанные значения модели (1.14) как функцию $i(i = 1, \cdots, 26)$, практически совпадают (см. Рисунок 1.7). Полученная средняя относительная ошибка 1.1% демонстрирует высокое качество аппроксимации.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9715	0.9438	0.9064	25.2225	1.38E-07	480.9	0.4570	0.6975
x1, x2, x3, x4, x5, x6, x8, x9, x10	0.9714	0.9437	0.9121	29.8452	2.43E-08	479.0	0.4082	0.7102
x2, x3, x4, x5, x6,x8, x9, **x10**	0.9713	0.9435	0.9169	35.5240	4.01E-09	477.1	0.4411	0.7444
x2, x3, x4, x5, x6,x8, **x9**	0.9711	0.94309	0.9209	42.6169	6.23E-10	475.3	0.4280	0.7197
x2, x3, x4, **x5**, x6,x8	0.9707	0.94235	0.9241	51.7688	9.15E-11	473.6	0.3685	0.5454
x2, **x3**, x4, x6,x8	0.9691	0.9391	0.9239	61.7467	1.81E-11	**473.0**	0.6762	0.8025
x2,x4, x6,**x8**	0.9623	0.9261	0.912	65.8038	1.41E-11	476.1	0.4179	0.7474
x2, x4, x6	0.9606	0.92287	0.91235	87.7465	2.14E-12	475.2	0.3052	0.6742

Таблица 1.17: Динамика показателей качества при выборе объясняющих переменных в 2012 году

Модель с наименьшим значением AIC:

$$\begin{aligned}\hat{y}_t =& 11830 - 0.4771 \cdot x_{2,t} - 6.157 \cdot x_{3,t} + 0.8173 \cdot x_{4,t} + 0.00006969 \cdot x_{6,t} \\ &+ 10.4689 \cdot x_{8,t}\end{aligned} \tag{1.15}$$

Множественная регрессионная модель для 2013 года

Множественная регрессионная модель для 2013 года, после последовательного исключения статистически незначимых параметров (на основе t-теста Стьюдента с $\alpha = 0,05$), принимает следующую форму:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{4.49705)}{10791.7171} - \underset{(-4.4098)}{0.3947} \cdot x_{2,t} + \underset{(9.919)}{0.9138} \cdot x_{4,t} + \underset{(4.0522)}{0.00005939} \cdot x_{6,t} \tag{1.16}$$

Качество построенных моделей, соответствующих каждому этапу последовательного исключения незначимых факторов, можно оценить, проанализировав показатели в Таблице 1.18. Значения R , R^2 , и R^2_{adj} для конечной модели 1.16, представленные в последней строке, указывают на ее хорошее качество. При $\alpha = 0,05$ уравнение регрессии считается статистически значимым ($F_{\text{obs}} = 87.7465, F_{(0,05;10;26-10-1)} = 2,543$).

Ошибка аппроксимации модели является незначительной, средняя относительная ошибка составляет $E_{\text{rel.}} = 1,008\%$. Графики, показанные на Рисунке 1.8, подтверждают это.

Рис. 1.8

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9715	0.9438	0.9064	25.2225	1.38E-07	488.3	0.2052	0.2595
x1, x2, x3, x4, x5, x6, x8, x9, x10	0.9714	0.9437	0.9121	29.8452	2.43E-08	486.4	0.2015	0.3511
x2, x3, x4, x5, x6,x8, x9, **x10**	0.9713	0.9435	0.9169	35.5240	4.01E-09	485.4	0.2361	0.6711
x2, x3, x4, x5, x6,x8, **x9**	0.9711	0.94309	0.9209	42.6169	6.22E-10	484.6	0.3212	0.6277
x2, x3, x4, **x5**, x6,x8	0.9707	0.94235	0.9241	51.7688	9.15E-11	484.9	0.4562	0.3289
x2, **x3**, x4, x6,x8	0.9691	0.9391	0.9239	61.7467	1.81E-11	484.9	0.6160	0.7984
x2,x4, x6,**x8**	0.9623	0.9261	0.912	65.8038	1.41E-11	485.5	0.3986	0.2646
x2, x4, x6	0.9606	0.92287	0.91235	87.7465	2.14E-12	**484.2**	0.5032	0.3547

Таблица 1.18: Динамика показателей качества при выборе объясняющих переменных в 2013 году

Тест на автокорреляцию остатков с использованием теста Дарбина-Уотсона не дает определенного ответа. Поскольку $DW = 2,035$, при $\alpha = 0,05, n = 26, k = 15$, где $d_L = 0,256$ и $d_U = 3,179$ являются нижним и верхним табличными критическими значениями соответственно, наблюдаемая статистика попадает в "зону неопределенности".

Множественная регрессионная модель для 2014 года

Множественная регрессионная модель, созданная на основе данных за 2014 год, содержит самое большое количество регрессоров и имеет следующий вид:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.7025)}{7537.9438} - \underset{(-3.9412)}{0.3349} \cdot x_{2,t} + \underset{(9.3550)}{0.9605} \cdot x_{4,t} + \underset{(3.7005)}{0.00005521} \cdot x_{6,t} \qquad (1.17)$$

Таблица 1.19 представляет значения характеристик, позволяющих оценить качество каждой модели в общем виде, построенной на основе соответствующих наборов объясняющих переменных, сформированных путем последовательного исключения незначимых переменных.

Модель 1.17 обладает относительно высоким скорректированным коэффициентом детерминации $R^2_{adj} = 0.904$, даже при большом количестве предикторных переменных. Коэффициент множественной корреляции составляет $R^2 = 0.9155$, указывая на сильную связь между зависимой переменной и набором регрессоров. Наблюдаемая статистика Фишера соответствует p-значению $P-\text{value}(F) = 5.8 \cdot 10^{-12}$. Модель является статистически значимой на уровне значимости 5%.

Рис. 1.9

Удовлетворительное качество аппроксимации исходных данных моделью подтверждается графиками, показанными на Рисунке 1.9. Средняя относительная ошибка аппроксимации составляет $E_{\text{rel.}} = 1,06\%$.

Модель с наименьшим значением AIC:

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 12700 - 0.2101 \cdot x_{2,t} - 5.4055 \cdot x_{3,t} + 0.5423 \cdot x_{4,t} + 0.00009806 \cdot x_{6,t} \\ & - 0.0231 \cdot x_{7,t} + 0.5635 \cdot x_{9,t} - 1704.98 \cdot x_{10,t}.\end{aligned} \tag{1.18}$$

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9705	0.9419	0.9082	24.3349	1.76E-07	498.1	0.4011	0.3937
x2, x3, x4, **x5**, x6, x7, x8, x9, x10	0.9701	0.9412	0.9082	28.4854	3.43E-08	496.4	0.6827	0.3929
x2, x3, x4, x6, x7, **x8**, x9, x10	0.9687	0.9384	0.9095	32.4221	8.19E-09	494.7	0.7097	0.3543
x2, x3, x4, x6, x7, x9, **x10**	0.9686	0.9382	0.9142	39.0596	1.28E-09	**493.7**	0.2604	0.2153
x2, **x3**, x4, x6, x7, x9	0.9658	0.9329	0.9117	44.0432	3.79E-10	493.8	0.3784	0.5661
x2,x4, x6, x7, **x9**	0.9626	0.9267	0.9084	50.6182	1.13E-10	494.1	0.3487	0.5764
x2,x4, x6, **x7**	0.9589	0.9194	0.9041	59.9660	3.46E-11	494.6	0.6049	0.4713
x2, x4, x6	0.9568	0.9155	0.9040	79.4885	5.8E-12	493.8	0.9644	0.8195

Таблица 1.19: Динамика показателей качества при выборе объясняющих переменных в 2014 году

Множественная регрессионная модель для 2015 года

Рассмотрим построение модели на основе данных за 2015 год. В таблице 1.20 представлен выбор наиболее значимых факторов в уравнении регрессии, отражающий связь между значениями показателей качества регрессии и набором экзогенных переменных.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9634	0.9281	0.8802	19.3829	8.27E-07	509.9	0.6264	0.2581
x1, x2, x3, x4, x5, x6, x8, x9, **x10**	0.9634	0.9281	0.88775	22.9691	1.64E-07	508.0	0.6017	0.2601
x1, x2, x3, x4, x5, x6, x8, x9	0.9633	0.9281	0.8942	27.4313	2.99E-08	506.0	0.6241	0.2720
x2, x3, x4, x5, x6, **x8**, x9	0.9632	0.9279	0.8998	33.1041	5.04E-09	504.0	0.6910	0.3098
x2, **x3**, x4, x5, x6, x9	0.9622	0.9259	0.9026	39.6144	9.54E-10	502.7	0.7131	0.3841
x2, x4, **x5**, x6,x9	0.96139	0.9242	0.9053	48.8209	1.58E-10	501.3	0.6327	0.3244
x2, x4, x6,**x9**	0.9598	0.9213	0.90635	61.4932	2.71E-11	500.3	0.8492	0.3608
x2, x4, x6	0.9595	0.9207	0.9099	85.2255	2.88E-12	**498.5**	0.8464	0.3921

Таблица 1.20: Динамика показателей качества при выборе объясняющих переменных в 2015 году

На последнем этапе была получена модель, в которой все параметры являются статистически значимыми на заданном уровне $\alpha = 0,05$, согласно значимости оценок факторов на основе t-теста Стьюдента. Она принимает следующую форму:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.8623)}{8714.3617} - \underset{(-4.2869)}{0.3407} \cdot x_{2,t} + \underset{(9.6562)}{1.0344} \cdot x_{4,t} + \underset{(4.0281)}{0.00006283} \cdot x_{6,t} \qquad (1.19)$$

Для оценки точности аппроксимации полученной модели регрессии были построены графики, отображающие начальные значения полученного показателя для номера наблюдения $i (i = 1, \cdots, 26)$, а также значения полученного показателя, рассчитанные с использованием метода 1.19, также для номера на-

блюдения $i(i = 1, \cdots, 26)$ (см. Рисунок 1.10). Очевидно, что графики практически идентичны. Средняя относительная ошибка аппроксимации составляет $E_{\text{rel.}} = 1,22\%$, что указывает на то, что ошибка в предсказанных значениях lny составляет примерно 1%.

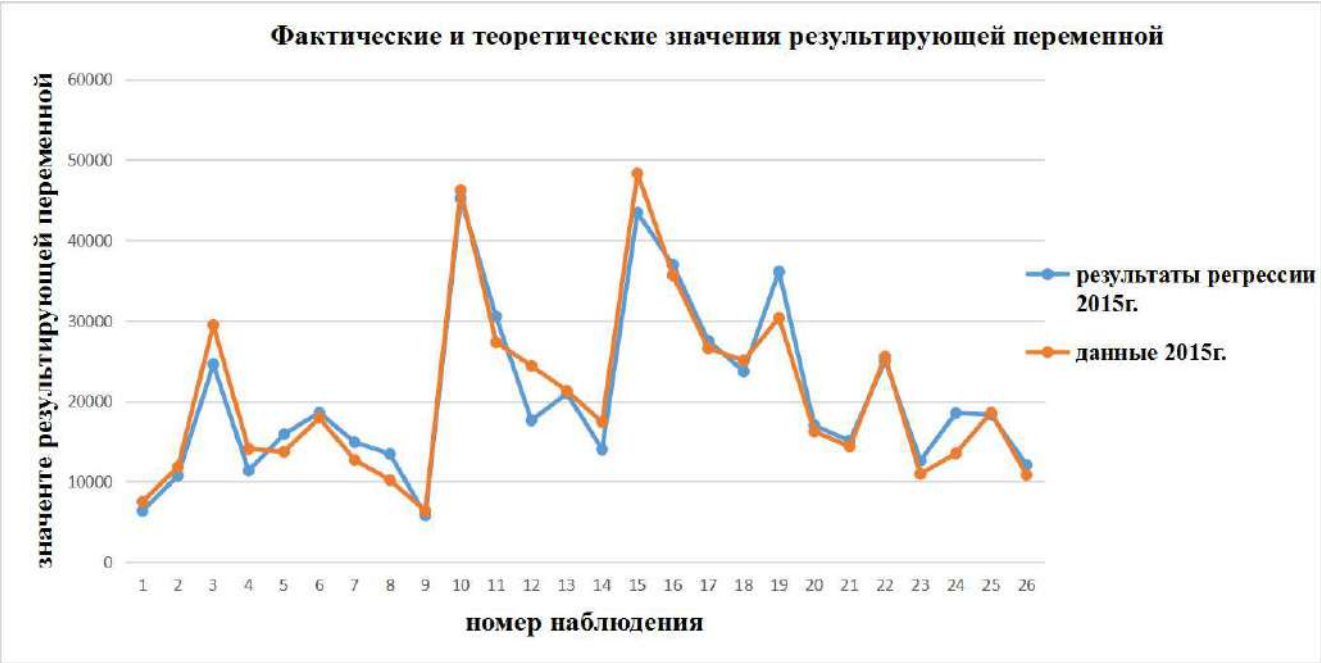


Рис. 1.10

Скорректированный коэффициент детерминации $R^2_{adj} = 0.9099$ также указывает на хорошую способность модели к аппроксимации и подтверждает наличие функциональной связи между зависимой переменной (lny) и объясняющими переменными (x_2, x_4, x_6). Величина коэффициента множественной корреляции близка к 1, что указывает на очень сильную связь. Адекватность модели, или другими словами, проверка гипотезы о том, что все параметры равны нулю, была проведена с использованием F-теста с регрессорами. Было установлено, что уравнение (1.19) является статистически значимым в общем на заданном уровне $\alpha = 0,05 (F = 85.2255)$. Наблюдаемое значение статистики Дарбина-Уотсона равно $DW = 1,814$, что указывает на отсутствие автокорреляции остатков. Таким образом, множественная регрессионная модель (1.19) для 2015 года является адекватной, хорошего качества и подходит для экономического анализа.

year \ country	Indonesia	Malaysia	Singapore	Thailand	Philippines
1998	-0.241	0.163	7.314	7.315	2.287
1999	-1.866	3.895	16.578	6.103	1.247
2000	-4.55	3.788	15.515	3.366	1.487
2001	-2.977	0.5539	17.007	5.067	0.76
2002	0.1451	3.193	6.157	3.342	1.769
2003	-0.5969	3.219	17.051	5.232	0.492
2004	1.896	4.376	24.39	5.86	0.592
2005	8.336	3.925	18.09	8.216	1.664
2006	4.914	7.691	36.924	8.917	2.707
2007	6.928	9.071	47.733	8.634	2.919
2008	9.318	7.573	12.201	8.562	1.34
2009	4.877	0.1146	23.821	6.411	2.065
2010	15.292	10.886	55.076	14.747	1.07
2011	20.565	15.119	49.156	2.474	2.007
2012	21.201	8.896	55.31	12.899	3.215
2013	23.282	11.296	64.39	15.936	3.737
2014	25.121	10.619	68.698	4.975	5.74

Таблица 1.21: Индекс привлекательности инвестиций (в триллионах долларов) для ASEAN-5

1.3.2 Построение и анализ моделей привлекательности инвестиций в регионах ASEAN-5

АСЕАН является организацией, созданной 31 июля 1961 года, под названием Ассоциация Юго-Восточной Азии (ASA), группа, состоящая из Таиланда, Филиппин и Малайзии.

Множественные регрессионные модели для Индонезии

На основе данных Всемирного банка по Индонезии с 1998 по 2014 годы строится таблица для определения параметров уравнения регрессии в следующем формате:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), i = 1 \dots 17. \tag{1.20}$$

Где: i - индекс года;

y_i- логарифм оценки объема инвестиций в году i;

x_{1i} - потребление электроэнергии в году i;

x_{2i} - уровень дохода на душу населения в году i;

x_{3i} - задолженность по займам, предоставленным юридическим лицам в году i;

x_{4i} - стоимость основных средств в году i;

x_{5i} - объем строительной деятельности в году i;

x_{6i} - число предприятий и организаций в году i;

x_{7i} - оборот розничной торговли в году i;

x_{8i} - ВВП на душу населения в году i;

x_{9i} - уровень безработицы в году i (in percentage).

Оценки коэффициентов рассчитываются с использованием метода регрессионного анализа в инструменте добавления данных MS Excel для наблюдаемых значений $y = (y_1, ..., y_{17})^\top$ and $X = (x_{1i}, x_{2i}, \cdots, x_{9i})^\top$, $i = 1, \cdots, 17$. Оценки коэффициентов принимают вид, показанный в таблице 1.12:

b0(c.o.)	208.8437	(90.64741)
b1(c.o.)	0.082245	(0.046266)
b2(c.o.)	-0.00285	(0.005001)
b3(c.o.)	0.103778	(0.10592)
b4(c.o.)	-0.28036	(0.497253)

b5(c.o.)	-2.84121	(2.320046)
b6(c.o.)	0.04912	(0.043409)
b7(c.o.)	0.012576	(0.093517)
b8(c.o.)	0.012724	(0.003133)
b9(c.o.)	-2.40566	(0.553048)

Таблица 1.22: Параметры уравнения регрессии для Индонезии.

Общие показатели качества модели указывают на высокий уровень ее способности к аппроксимации. Коэффициент множественного детерминации равен $R^2 = 0,9889$, а после корректировки с учетом штрафа за большое количество объясняющих переменных скорректированный коэффициент детерминации составляет $R^2_{adj} = 0,9746$. Множественный коэффициент корреляции приближается к единице с $R = 0,9944$(см. Таблицу 1.23).

R	0,9944
R^2	0,9889
R^2_{adj}	0,9746

Таблица 1.23

Рассчитанное значение F-статистики равно $F = 69,3286$. При заданном уровне значимости $\alpha = 0,05$ критическое значение составляет $F_{0,05;9;7} = 3,68$. Следо-

вательно, модель является статистически значимой и практически объясняет большую часть общей дисперсии зависимой переменной. Однако, несмотря на общее хорошее качество регрессионной модели, оценки параметров для объясняющих переменных не считаются статистически значимыми на выбранном уровне α (см. Таблицу 1.24).

	t Stat	P-value
b0	2.303912	0.054675016
b1	1.777675	0.118697712
b2	-0.57053	0.586160124
b3	0.97978	0.359835289
b4	-0.56381	0.590481302

	t Stat	P-value
b5	-1.22463	0.260321937
b6	1.131562	0.295089445
b7	0.134478	0.896809381
b8	4.06166	0.004799517
b9	-4.34982	0.003355596

Таблица 1.24: t-статистика для Индонезии

Перейдем к методологии последовательного исключения статистически незначимых факторов из уравнения регрессии. Согласно полученным результатам, на втором этапе строится модель без x_7 ($b_7 = 0$). Новая модель имеет практически такое же качество, как и предыдущая, но статистически более значима согласно F-тесту Фишера.

Динамика изменения показателей R^2,R, R^2_{adj} и F при последовательном исключении незначимых факторов из множественной регрессионной модели представлена в таблице 1.25.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9944	0.9888	0.9777	88.9045	5.21E-07	66.49	0.4568	0.3919
x1, x3, x4, x5, x6, x8, x9	0.9941	0.9883	0.9793	109.1355	5.61E-08	65.28	0.6183	0.4863
x1, x3, x5, x6, x8, x9	0.9935	0.9871	0.97936	127.5862	7.33E-09	**65.00**	0.6306	0.5709
x1, x5, x6, x8, x9	0.9924	0.9849	0.9781	144.4038	1.2E-09	65.49	0.4901	0.3775
x1, x5, x8, x9	0.9903	0.9807	0.9743	152.8212	3.51E-10	67.83	0.8697	0.4648

Таблица 1.25: Динамика показателей качества при выборе объясняющих переменных в Индонезии

Исключение переменных x_7, x_2, x_4 приводит к модели с максимальным значением R^2_{adj}, указывающим на их незначительное влияние на зависимую переменную. Оценка качества коэффициентов для отдельных переменных с использованием t-теста не позволяет считать модель статистически значимой, поскольку p-значение для t-статистики b_3 значительно выше выбранного уровня α(см. Таблицу 1.24).

После исключения x_3, x_6 получается регрессионная модель, в которую входят только те факторы, чьи коэффициенты являются статистически значимыми при $\alpha = 0,05$. Модель также демонстрирует хорошее качество аппроксимации (см. Таблицу 1.25) и имеет следующий вид:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(4.7593)}{8714.3617} + \underset{(6.1031)}{0.1190} \cdot x_{1,t} - \underset{(-5.4129)}{4.7414} \cdot x_{5,t} + \underset{(5.2747)}{0.01066} \cdot x_{8,t} - \underset{(-4.3683)}{2.1467} \cdot x_{9,t} \tag{1.21}$$

Графики значений y_i и $\hat{y}_i$ для наблюдений $i, i = 1, \cdots, 17$ показаны на рисунке 1.2.

Относительная ошибка $E_{\text{rel.}} = 3,31\%$, указывающая на высокую способность модели к аппроксимации, так как она почти полностью восстанавливает исходные данные.

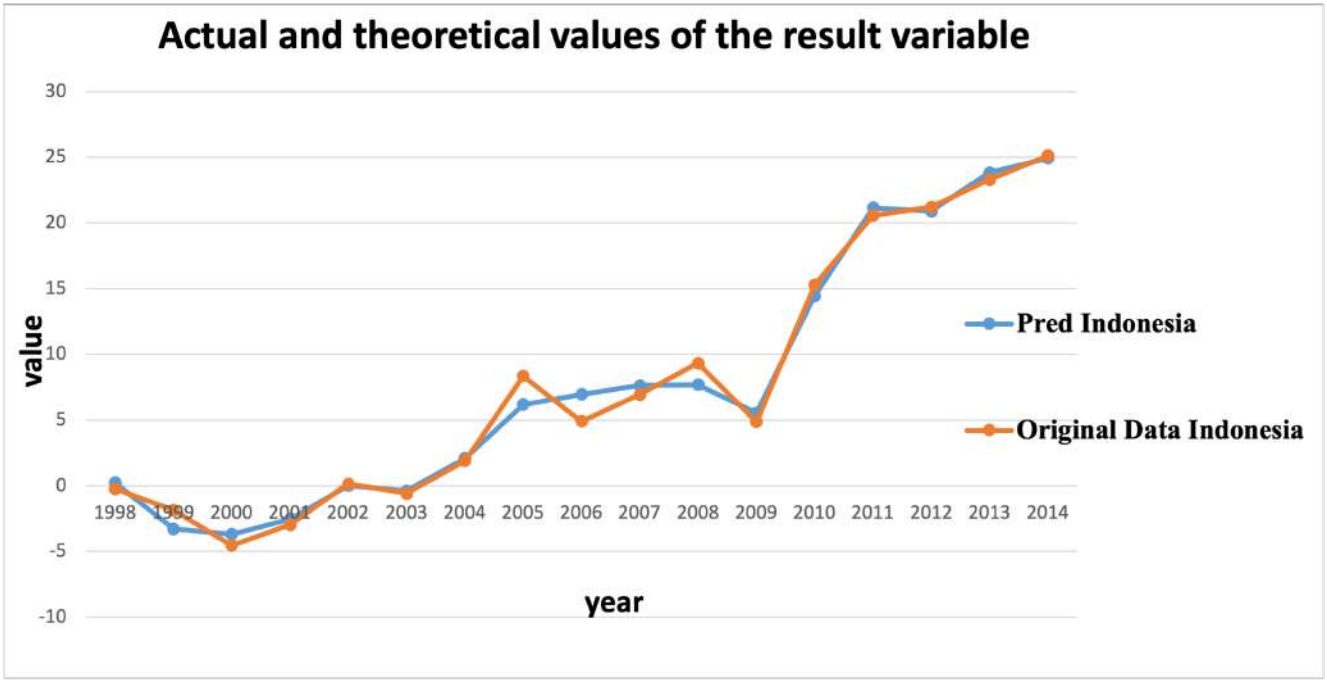


Рис. 1.11

Рассчитанное значение статистики Дарбина-Уотсона составляет $DW = 2,920$. При $\alpha = 0,05, n = 17, k = 7$, табличные значения равны $d_L = 0,451$ и $d_U = 2,537$. Поскольку $4 - d_U < DW < 4 - d_L$, невозможно определить наличие или отсутствие автокорреляции.

Модель с наименьшим значением AIC:

$$\begin{aligned}\hat{y}_t =& 250.447 + 0.0828 \cdot x_{1,t} + 0.0499 \cdot x_{3,t} - 3.803 \cdot x5, t \\ &+ 0.048 \cdot x6, t + 0.0107 \cdot x_{8,t} - 2.4341 \cdot x_{9,t}\end{aligned} \tag{1.22}$$

Множественные регрессионные модели для Индонезии

В соответствии с проведенным исследованием в Малайзии на 17 объектах (1998-2014 годы) был зафиксирован следующий набор показателей, характеризующих состояние и тенденции социально-экономического развития:

$$\begin{aligned}y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \\ i = 1..17.\end{aligned} \tag{1.23}$$

Где: i - индекс года;
y_i- логарифм оценки объема инвестиций в году i;
x_{1i} - потребление электроэнергии в году i;
x_{2i} - уровень дохода на душу населения в году i;
x_{3i} - задолженность по займам, предоставленным юридическим лицам в году i;
x_{4i} - стоимость основных средств в году i;
x_{5i} - объем строительной деятельности в году i;
x_{6i} - число предприятий и организаций в году i;
x_{7i} - оборот розничной торговли в году i;
x_{8i} - ВВП на душу населения в году i;
x_{9i} - уровень безработицы в году i (in percentage).

Давайте обсудим построение модели на основе данных для Малайзии. Выбор наиболее значимых факторов для уравнения регрессии представлен в таблице 1.26, которая иллюстрирует зависимость качественных характеристик регрессии от множества экзогенных переменных.

На последнем этапе была построена модель, в которой все параметры являются статистически значимыми на выбранном уровне $\alpha = 0,05$, на основе значимости оценок факторов с использованием t-теста Стьюдента. Она имеет следующий вид:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(-1.4541)}{-1.9291} - \underset{(-2.8484)}{0.0038179} \cdot x_{2,t} + \underset{(3.7713)}{0.0048033} \cdot x_{8,t} \tag{1.24}$$

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x4, x5, x6, **x7**, x8, x9	0.9574	0.9168	0.8336	11.0195	0.001364	72.06	0.3843	0.2903
x1, x2, x4, x5, x6, **x8**, x9	0.9572	0.9163	0.8512	14.0769	0.0003445	70.17	0.3948	0.2744
x1, x2, x5, x6, x8, **x9**	0.9557	0.9134	0.8615	17.5982	8.76E-05	**68.73**	0.2305	0.2689
x1, x2, x5, x6, x8	0.9333	0.8712	0.8127	14.885	0.0001407	73.49	0.3962	0.4037
x2, x5, **x6**, x8	0.9171	0.8411	0.7882	15.8888	9.706E-05	75.12	0.2522	0.2793
x2, **x5**, x8	0.9055	0.82	0.7784	19.74364	4.025E-05	75.28	0.3789	0.4228
x2, x8	0.8898	0.7917	0.762	26.6206	1.69E-05	75.77	0.4002	0.2107

Таблица 1.26: Динамика показателей качества при выборе объясняющих переменных в Малайзии

Скорректированный коэффициент детерминации составляет $R^2_{adj} = 0.762$, что также указывает на хорошую способность модели к аппроксимации и подтверждает наличие функциональной связи между зависимой переменной (y) и объясняющими переменными (x_2, x_8). Значение множественного коэффициента корреляции близко к 1, что указывает на сильную корреляцию.

Адекватность модели, или другими словами, проверка гипотезы о том, что все параметры равны нулю с использованием F-теста с регрессорами, показала, что уравнение 1.24 статистически значимо в целом на выбранном уровне $\alpha = 0,05$ level ($F_{\text{tab}} = 3,23$, $F_{\text{obs}} = 26.6206$).

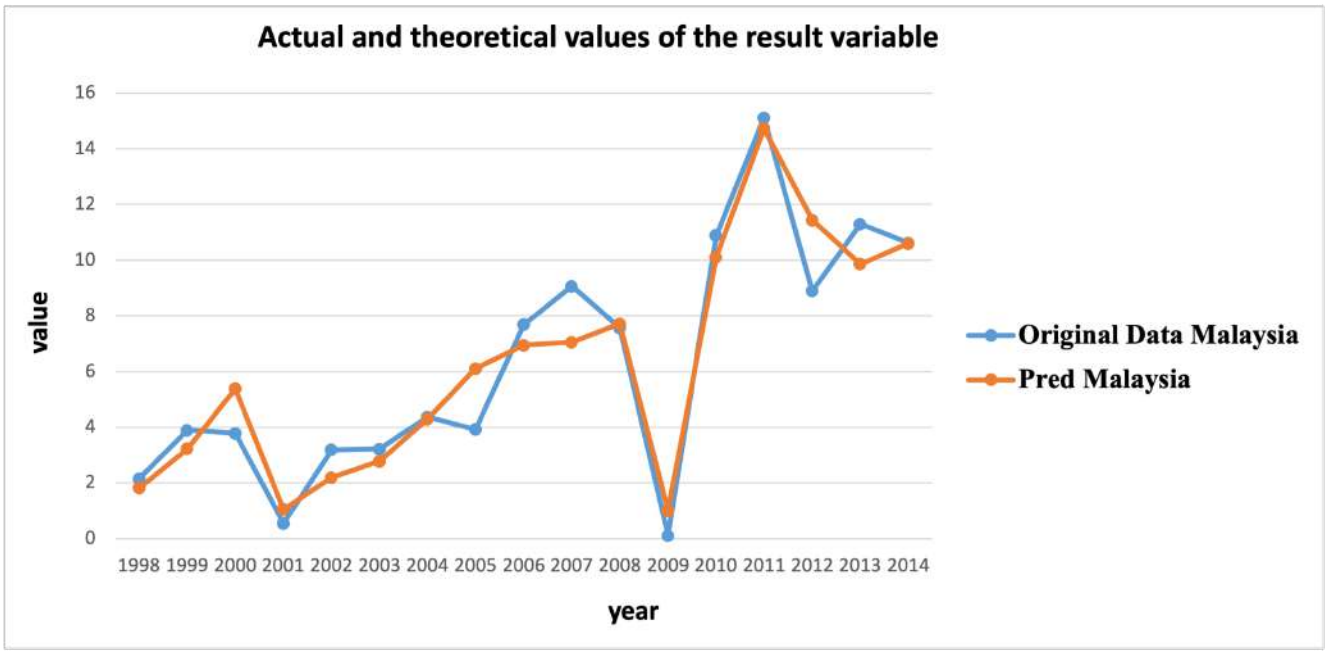


Рис. 1.12

Наблюдаемое значение статистики Дарбина-Уотсона составляет $DW = 2,414$. При $\alpha = 0,05, n = 17, k = 8$, табличные значения равны $d_L = 0,356$ и $d_U =$

2, 757. Поскольку $d_L < DW < d_U$, невозможно определить наличие или отсутствие автокорреляции.

Таким образом, модель 1.24, построенная на основе данных для Малайзии, обладает достаточной способностью аппроксимации. Оценки качества указывают на правильность спецификации, а анализ остатков подтверждает ее адекватность.

Модель с наименьшим значением AIC:

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 75.9333 + 0.0115 \cdot x_{1,t} - 0.0051 \cdot x_{2,t} - 3.9452 \cdot x_{5,t} \\ &+ 0.066 \cdot x_{6,t} + 0.0079 \cdot x_{8,t} + 1.2757 \cdot x_{9,t}\end{aligned} \tag{1.25}$$

Множественные регрессионные модели для Сингапура

Аналогично описанному выше процессу построения регрессионной модели для Индонезии и Малайзии, теперь мы построим модель для Сингапура.

$$\begin{aligned}y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}),& \\ i = 1..17.&\end{aligned} \tag{1.26}$$

Где: i - индекс года;

y_i- логарифм оценки объема инвестиций в году i;

x_{1i} - потребление электроэнергии в году i;

x_{2i} - уровень дохода на душу населения в году i;

x_{4i} - стоимость основных средств в году i;

x_{5i} - объем строительной деятельности в году i;

x_{6i} - число предприятий и организаций в году i;

x_{7i} - оборот розничной торговли в году i;

x_{8i} - ВВП на душу населения в году i;

x_{9i} - уровень безработицы в году i (in percentage).

Конечная модель регрессии, полученная после обработки данных в Сингапуре, имеет следующий вид:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(-4.5665)}{-140.6215} + \underset{(8.6277)}{0.03392} \cdot x_{1,t} - \underset{(-5.2671)}{0.372139} \cdot x_{7,t} \tag{1.27}$$

Отношение между наблюдаемыми (y_i) и предсказанными $(\hat{y}_i)$ значениями

зависимой переменной для каждого наблюдения $i, i = 1, \cdots, 17.$, показано ниже (см. Рисунок 1.13).

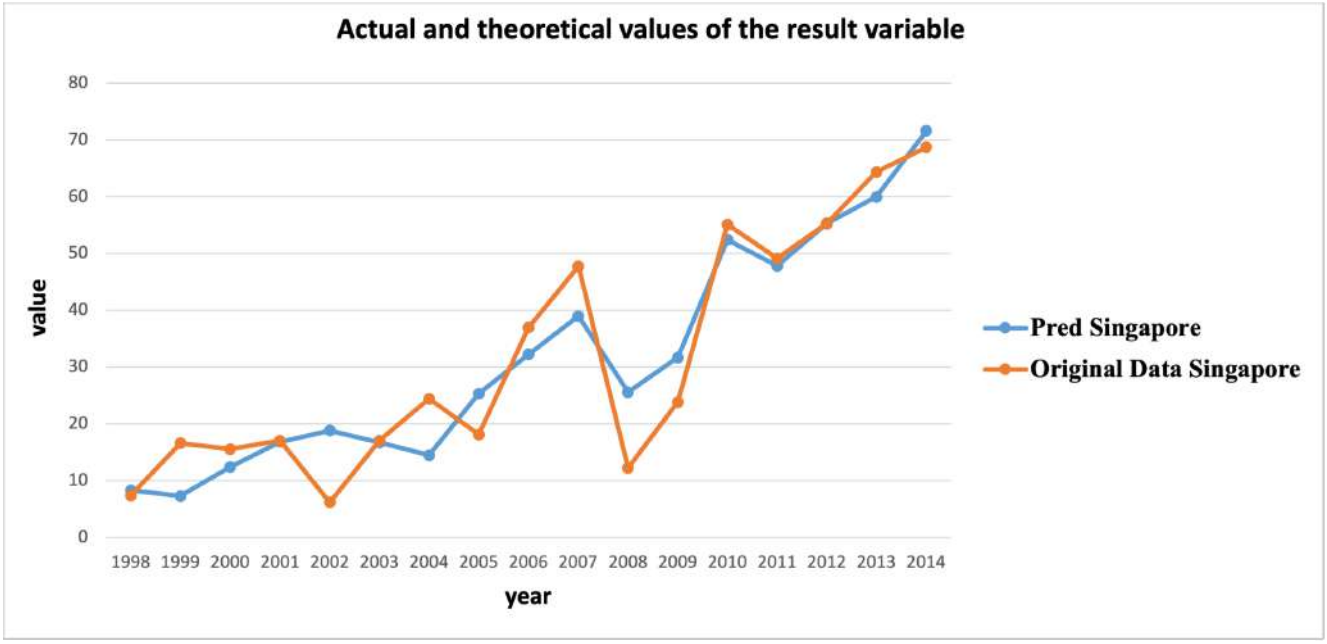


Рис. 1.13: The change in qualitative characteristics of the regression as factors were selected

Таблица 1.27 иллюстрирует изменения качественных характеристик регрессии при выборе факторов. В последней строке таблицы приведены значения R, R^2, R^2_{adj} и F_{obs} для модели из таблицы 1.27.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x4, x6, x7, x8, x9	0.9421	0.8876	0.8002	10.158	0.001216	129.4	0.5945	0.2602
x1, x2, x6, x7, x8, x9	0.942	0.8873	0.8198	13.132	0.0003125	127.4	0.5441	0.4628
x1, x2, x7, x8, x9	0.9411	0.8858	0.8339	17.0734	7.40E-05	125.7	0.6155	0.3796
x1, x7, x8, x9	0.9393	0.8824	0.8432	22.513	1.66E-05	124.2	0.7161	0.2220
x1, x7, x9	0.938	0.8799	0.8522	31.7596	2.98E-06	**122.5**	0.5762	0.2571
x1, x7	0.9214	0.849	0.8275	39.3794	1.78E-06	124.4	0.3717	0.2281

Таблица 1.27: Динамика показателей качества при выборе объясняющих переменных в Сингапуре.

Близкое к единице значение R^2_{adj} указывает на сильную связь между y и x_1, x_7.. Кроме того, модель адекватно аппроксимирует наблюдаемые значения зависимой переменной, что подтверждается значением $R = 0.9214$. На общую статистическую значимость регрессионной модели на уровне значимости 5% указывает тест Фишера: наблюдаемая статистика $F = 39.3794$ значительно превышает табличное значение $F_{\text{tab}} = 3, 29$.

Модель с наименьшим значением AIC:

$$\hat{y}_t = -309.2364 + 0.0277 \cdot x_{1,t} - 0.2632 \cdot x_{7,t} + 3.0053 \cdot x_{9,t} \tag{1.28}$$

Множественные регрессионные модели для Таиланда

Следуя описанному выше подходу к построению регрессионной модели в Индонезии, Малайзии и Сингапуре, теперь мы перейдем к построению регрессионной модели в Таиланде.

$$y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \quad i = 1..17. \tag{1.29}$$

Где: i - индекс года;
y_i- логарифм оценки объема инвестиций в году i;
x_{1i} - потребление электроэнергии в году i;
x_{2i} - уровень дохода на душу населения в году i;
x_{3i} - задолженность по займам, предоставленным юридическим лицам в году i;
x_{4i} - стоимость основных средств в году i;
x_{5i} - объем строительной деятельности в году i;
x_{6i} - число предприятий и организаций в году i;
x_{7i} - оборот розничной торговли в году i;
x_{8i} - ВВП на душу населения в году i;
x_{9i} - уровень безработицы в году i (in percentage).

Однако на начальном этапе анализа регрессии со всеми факторами полученные значения R^2 и R^2_{adj} были очень низкими (далеко от 1) (см. Таблицу 1.28). Кроме того, p-значение (F) значительно превышало уровень значимости $\alpha = 0,05$ (см. Таблицу 1.29).

R	0.599
R^2	0.5186
R^2_{adj}	0.474

Таблица 1.28

Эти обстоятельства не позволяют нам построить регрессионную модель для Таиланда.

F-obs	F	p-value(F)
3.68	1.374583452	0.345016996

Таблица 1.29

Множественные регрессионные модели для Филиппин

На основе данных, предоставленных Всемирным банком для Филиппин, мы построим вычислительную таблицу для определения параметров уравнения регрессии в следующем формате:

$$y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \quad i = 1..17. \tag{1.30}$$

Где: i - индекс года;
y_i- логарифм оценки объема инвестиций в году i;
x_{1i} - потребление электроэнергии в году i;
x_{2i} - уровень дохода на душу населения в году i;
x_{3i} - задолженность по займам, предоставленным юридическим лицам в году i;
x_{4i} - стоимость основных средств в году i;
x_{5i} - объем строительной деятельности в году i;
x_{6i} - число предприятий и организаций в году i;
x_{7i} - оборот розничной торговли в году i;
x_{8i} - ВВП на душу населения в году i;
x_{9i} - уровень безработицы в году i (in percentage).

Динамика показателей R^2,R, R^2_{adj} и F при последовательном исключении незначимых факторов из множественной регрессионной модели представлена в таблице 1.30.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9665	0.9341	0.8495	11.0402	0.002268	30.13	0.2083	0.2672
x1, x2, x3, **x4**, x5, x6, x8, x9	0.9663	0.9339	0.8678	14.1291	0.0005677	28.34	0.2668	0.2935
x1, x2, x3, x5, x6, **x8**, x9	0.9602	0.9221	0.8615	15.2224	0.0002524	26.85	0.2879	0.3679
x1, x2, x3, x5, x6, **x9**	0.9319	0.8685	0.7896	11.0091	0.0006546	26.35	0.2819	0.3125
x1, x2, x3, x5, x6	0.9309	0.8666	0.8059	14.2942	0.0001696	25.77	0.2221	0.3256
x2, x3, x5, x6	0.9272	0.8597	0.8129	18.3863	4.69E-05	**25.34**	0.3795	0.2248

Таблица 1.30: Динамика показателей качества при выборе объясняющих переменных на Филиппинах.

Исключение переменной x_7 приводит к модели с максимальным значением $R^2_{adj} = 0.8678$, указывающим на их незначительное влияние на зависимую переменную. Оценка коэффициентов для отдельных переменных с использованием t-теста не позволяет считать модель статистически значимой, поскольку p-значение для t-статистики b_4 значительно превышает выбранный уровень $\alpha(0.6027 > 0.05)$.

После исключения x_4, x_8, x_9 была получена регрессионная модель, включающая только те факторы, коэффициенты которых статистически значимы при $\alpha = 0,05$. Модель также демонстрирует хорошее качество аппроксимации (см. Таблицу 1.30) и имеет следующий вид:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(-4.6988)}{-335.3628} - \underset{(-2.6266)}{0.003292} \cdot x_{2,t} - \underset{(-3.9649)}{0.1683} \cdot x_{3,t} + \underset{(4.9556)}{6.9177} \cdot x_{5,t} + \underset{(2.20622)}{0.1421} \cdot x_{6,t} \tag{1.31}$$

Графики значений y_i и $\hat{y}_i$ для номеров наблюдения $i, i = 1, \cdots, 17$ показаны на рисунке 1.14.

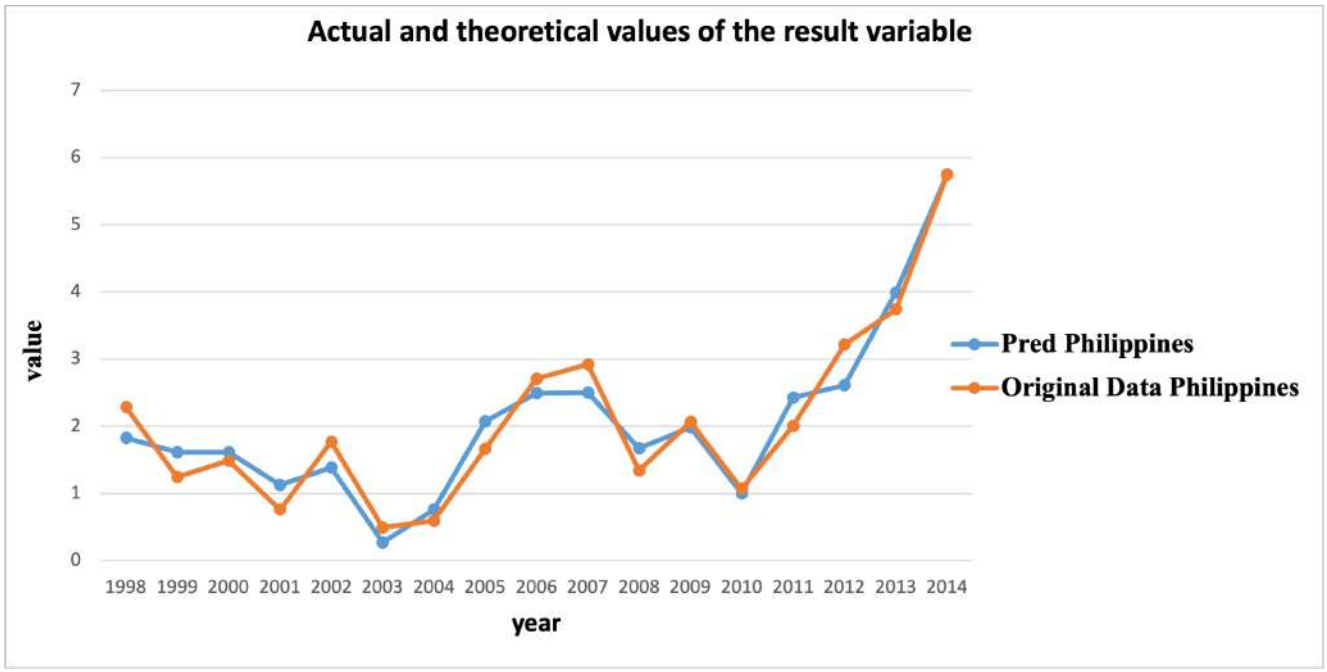


Рис. 1.14

Относительная ошибка $E_{\text{rel.}} = 9,44\%$, что указывает на достаточно высокую способность модели к аппроксимации, поскольку она успешно восстанавливает практически все исходные данные.

Рассчитанное значение статистики Дарбина-Уотсона составляет $DW = 2,473$. При $\alpha = 0,05, n = 17, k = 7$ критические значения равны $d_L = 0,451$ и $d_U = 2,537$. Поскольку $d_L < DW < d_U$, невозможно определить наличие или отсутствие автокорреляции.

Таким образом, модель 1.31, построенная на основе данных Филиппин, обладает достаточной способностью аппроксимации. Оценки качества указывают на правильность спецификации, и анализ остатков подтверждает ее адекватность.

1.3.3 Обсуждение и анализ результатов для Китая

Полученные данные из китайского статистического годового отчета учитывают уровень инфляции (все значения записаны по сопоставимым ценам).

Интерпретация полученных результатов является важным аспектом прикладного статистического исследования. После обработки собранных данных математическая модель, связывающая ежегодные инвестиции в регион с различными факторами, представляющими различные аспекты его социально экономического потенциала, должна быть интерпретирована в экономическом плане.

В статистическом анализе привлекательности инвестиций в Китае была построена множественная регрессионная модель на основе метода последовательного исключения с использованием данных с 2008 по 2017 годы. Было разработано десять уравнений регрессии для 26 регионов Китая с наибольшими инвестициями в 2018 году. Путем сравнения изменений социально-экономических показателей за последнее десятилетие и учитывая их накопленный эффект, можно наблюдать значительное влияние на среднюю стоимость инвестиций. Результаты представлены ниже (см. рисунок 1.15):

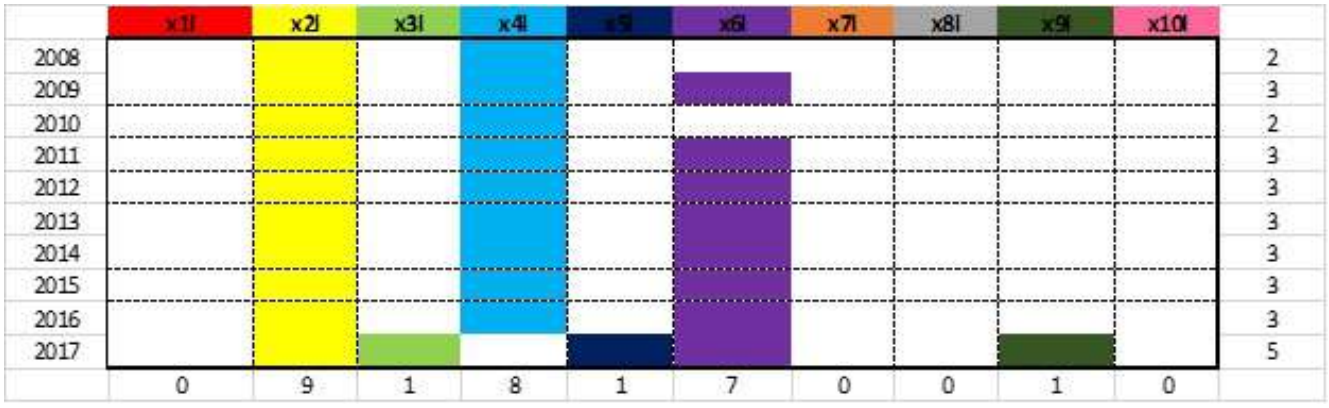


Рис. 1.15

Факторы соответствуют столбцам таблицы (для наглядности каждому фак-

тору назначен определенный цвет ячейки), а строки представляют собой наборы экзогенных переменных в конечной регрессионной модели для каждого соответствующего года. Внизу таблицы указано количество включений k-го фактора ($k = 1, ..., 10$) в регрессоры конечных моделей, построенных для лет 2008-2017. Число справа от таблицы представляет собой количество объясняющих переменных в конечной модели для соответствующего года.

Что касается фактора x_2, он включается во все модели, что указывает на его постоянное влияние на формирование региональных инвестиций с 2008 по 2017 годы. Однако факторы x_1, x_7, x_8, x_{10} никогда не включаются в уравнение регрессии, что свидетельствует о том, что эти показатели имеют минимальное влияние на объем инвестиций в период с 2008 по 2017 годы.

Фактор x_2 - уровень дохода на душу населения - включен во все модели. Это указывает на то, что уровень дохода на душу населения является важным фактором, влияющим на привлекательность инвестиций и оказывающим положительное влияние на привлекательность инвестиций.

Следующая по частоте включения переменная - x_4- стоимость основных средств. Логично, что стоимость основных средств является одним из важных факторов, влияющих на привлекательность инвестиций страны и играющих значительную роль в ее общей социально-экономической сфере. Однако в 2017 году темпы роста инвестиций замедлились, увеличившись на 7.5% за первые три квартала и сократившись на 0.7 процентных пункта по сравнению с годовыми значениями. Инвестиционные операции демонстрируют новые признаки "стабильности"с положительными изменениями нескольких показателей, таких как источники финансирования, инвестиции в гражданскую промышленность и инвестиции в производство оборудования. В настоящее время экономическая выгода от инвестиций значительно снизилась из-за сокращения государственных и инфраструктурных инвестиций, ограничений финансирования финансовыми расходами, замедления рынка недвижимости, продолжительного снижения отраслевых инвестиций и других проблем.

Фактор x_6 - объем строительных работ - также имеет значение в регрессионных переменных. Очень вероятно, что исследуемые регионы испытали рост в строительном секторе за рассматриваемый период. Как известно, многие считают инвестиции в недвижимость одними из самых надежных. Отсутствие фактора x_6 в модели для 2008 и 2010 годов может быть связано с пиковой точкой

кризиса 2008 года и высокими экономическими рисками.

В целом эти результаты позволяют сделать вывод о том, что нестабильная глобальная экономическая ситуация существенно влияет на формирование инвестиционных предпосылок на микро и среднем уровнях страны. Одновременно важна текущая стадия промышленного цикла. Особую значимость имеют экономические показатели риска в предкризисный и посткризисный периоды. Однако при условии стабильных экспортных поставок, существенного роста прибыли промышленности, активного спроса на инвестиции в научные и технологические реформы и ускоренного развития новых отдельных инвестиций объем привлекаемых инвестиций должен продолжать возрастать.

1.3.4 Обсуждение и анализ результатов для ASEAN-5

Прикладное статистическое исследование направлено на интерпретацию полученных результатов. После обработки данных необходимо объяснить, с экономической точки зрения, математическую модель, описывающую взаимосвязь между инвестициями в страну и различными факторами, отражающими различные аспекты ее социально-экономического потенциала.

На основе статистического анализа привлекательности инвестиций в АСЕАН была разработана множественная регрессионная модель, используя метод последовательного исключения с использованием данных с 1998 по 2014 годы. Было построено пять уравнений регрессии для пяти стран АСЕАН. За последние 18 лет многие социально-экономические показатели изменились, и их накопленный эффект значительно повлиял на среднюю стоимость инвестиций.

	x1i	x2i	x3i	x4i	x5i	x6i	x7i	x8i	x9i	
Indonesia										3
Malaysia										2
Singapore										3
Thailand										0
Philippines										4
	2	2	1	0	1	1	1	3	1	

Рис. 1.16

Для пяти стран АСЕАН каждый фактор имеет различную значимость из-за различий в экономических системах и национальных политиках. Однако x_8 - ВВП на душу населения - является важным фактором для АСЕАН в целом.

Что касается Таиланда, в последние годы он привлекает значительное количество иностранных туристов, и объем иностранных инвестиций ежегодно

растет. В настоящее время Таиланд служит инвестиционным центром в рамках АСЕАН, преимущественно ориентирующимся на инвестиции в инфраструктуру и цифровую экономику. Поэтому он может не подходить для данной модели.

Страны АСЕАН, пострадавшие от текущего глобального экономического спада, ускоряют процесс внутренней экономической интеграции для преодоления вызовов замедляющегося экономического роста. Необходимо укрепить экономические связи и сотрудничество с миром и больше полагаться на глобальные потоки инвестиций. Потенциал инвестиционного рынка АСЕАН будет продолжать расширяться. Инвестиции в АСЕАН помогают странам использовать обширные возможности региона в сфере энергетики, рабочей силы и других уникальных ресурсов, способствуя улучшению и модернизации внутренних промышленных структур и устойчивому экономическому развитию.

Уровень притока иностранных прямых инвестиций (ИПИ) в страны АСЕАН положительно коррелирует с такими факторами, как размер их экономик, уровнями природных ресурсов и технологий. Отрицательная корреляция может быть наблюдаема с внутренними конфликтами в странах АСЕАН, свободой торговли, свободой инвестиций, финансовой свободой и влиянием двусторонних соглашений об инвестициях на эффективность инвестиций. Однако влияние стабильности правительства в АСЕАН и усовершенствования инфраструктуры может значительно повысить эффективность инвестиций. Уровень экономического развития, уровень исследований и разработок, приток иностранного капитала и иностранные инвестиции оказывают положительное влияние, в то время как монополистические финансовые системы и торговые барьеры оказывают отрицательное влияние на ИПИ. Налогообложение также играет значительную роль в определении уровня иностранных прямых инвестиций.

1.4 Заключение к главе 1

В заключение данной главы было проведено множественное линейное регрессионное исследование, направленное на изучение привлекательности инвестиций в Китае и странах АСЕАН-5. С помощью модели последовательного исключения были выявлены наиболее влиятельные факторы, оказывающие влияние на привлекательность инвестиций для каждой страны.

Результаты показали, что уровень дохода на душу населения (x_2) непрерывно

влияет на региональные инвестиции с 2008 по 2017 годы, указывая на его положительное воздействие на привлекательность инвестиций. С другой стороны, факторы, такие как размер населения (x_1), стабильность правительства (x_7), ВВП на душу населения (x_8) и валютные резервы (x_4), оказывают минимальное влияние на объем инвестиций в период исследования.

Еще одним важным фактором, выявленным в ходе анализа, является стоимость основных средств (x_4), которая значительно влияет на привлекательность инвестиций страны и играет важную роль в ее социально-экономической структуре. Однако в 2017 году темпы роста инвестиций замедлились из-за различных вызовов, включая консолидацию бюджета, сокращение финансовых расходов, замедление на рынке недвижимости и снижение отраслевых инвестиций.

Объем строительных работ (x_6) также является значимой переменной в регрессионных моделях, указывая на важность роста строительного сектора для привлекательности инвестиций. Отсутствие x_6 в моделях для 2008 и 2010 годов можно объяснить экономическими рисками, связанными с глобальным финансовым кризисом в эти годы.

В целом, полученные результаты подчеркивают значительное влияние нестабильной глобальной экономической ситуации на инвестиционные предпосылки на микро и макроуровнях страны. Текущая стадия промышленного цикла и экономические риски играют важную роль, особенно в предкризисный и посткризисный периоды. Однако при условии стабильных экспортных поставок, существенного роста промышленной прибыли, активного спроса на инвестиции в научные и технологические реформы и ускоренного развития новых инвестиций можно ожидать, что объем привлекаемых инвестиций будет продолжать расти.

Для стран АСЕАН-5 каждый фактор имеет различную значимость из-за различий в экономических системах и национальных политиках. Однако ВВП на душу населения (x_8) является важным фактором для всего региона АСЕАН.

В отношении Таиланда в последние годы он привлек значительное количество иностранных туристов, и объем иностранных инвестиций ежегодно растет. В настоящее время Таиланд является инвестиционным центром в рамках АСЕАН со стратегией, сосредоточенной преимущественно на инвестициях в инфраструктуру и цифровую экономику. Поэтому он может не соответствовать конкретной модели, использованной в данном исследовании.

Страны АСЕАН, пострадавшие от текущего глобального экономического спада, активно ускоряют процесс внутренней экономической интеграции для решения вызовов замедляющегося экономического роста. Необходимо укрепить экономические связи и сотрудничество с миром, а также больше полагаться на глобальные потоки инвестиций. Потенциал инвестиционного рынка АСЕАН ожидается будет продолжать расширяться, позволяя странам использовать обширные возможности региона в области энергии, рабочей силы и других уникальных ресурсов для улучшения и модернизации структур внутренней промышленности и обеспечения устойчивого экономического развития.

Кроме того, анализ показывает, что приток иностранных прямых инвестиций (ИПИ) в страны АСЕАН положительно коррелирует с такими факторами, как размер их экономик, уровни природных ресурсов и технологии. Отрицательная корреляция наблюдается с внутренними конфликтами, свободой торговли, свободой инвестиций, финансовой свободой и влиянием двусторонних соглашений об инвестициях на эффективность инвестиций. Стабильность правительства в рамках АСЕАН и улучшение инфраструктуры могут значительно повысить эффективность инвестиций. Кроме того, уровень экономического развития, возможности для исследований и разработок, приток иностранного капитала и иностранные инвестиции оказывают положительное воздействие, в то время как монополистические финансовые системы и торговые барьеры оказывают отрицательное влияние на ИПИ.

В заключение, на основе полученных результатов данной главы можно предложить несколько рекомендаций для дальнейшего усиления привлекательности инвестиций в Китае и странах АСЕАН-5:

Усиление реформ политики: Правительствам следует сосредоточиться на внедрении комплексных реформ для улучшения инвестиционной среды. Это включает упрощение регулирования, сокращение бюрократии, повышение прозрачности и обеспечение честной конкуренции. Четкие и исполнимые правила и регуляции обеспечат инвесторам уверенность и стимулируют приток иностранных прямых инвестиций.

Продвижение экономического разнообразия: Страны должны стремиться к экономическому разнообразию с целью снижения зависимости от одного сектора или отрасли. Это позволит создать более стабильную и устойчивую экономику, привлекать более широкий спектр инвестиций. Правительства должны

выявлять и поддерживать развивающиеся отрасли с высоким потенциалом роста через целевые политики, стимулы и развитие инфраструктуры.

Улучшение инфраструктуры: Улучшение инфраструктуры является ключевым моментом для привлечения инвестиций. Правительства должны приоритизировать инфраструктурные проекты, способствующие торговле, связности и логистике. Развитие надежных транспортных сетей, модернизация портов и расширение телекоммуникационной инфраструктуры создаст благоприятную среду для бизнеса и способствует инвестиционным возможностям.

Формирование исследовательских и разработческих возможностей: Инвестиции в программы и инновации крайне важны для привлечения инвестиций высокой стоимости. Правительства должны выделять ресурсы на развитие отечественных технологий и поощрять сотрудничество между академией, промышленностью и исследовательскими учреждениями. Предоставление стимулов компаниям для инвестиций в исследовательскую деятельность способствует технологическим прорывам и привлекает инвестиции на основе знаний.

Укрепление региональной интеграции: Глубокая экономическая интеграция в рамках АСЕАН и укрепление сотрудничества с внешними партнерами может создать более привлекательную инвестиционную среду. Гармонизация регулирования, снижение торговых барьеров и продвижение межрегиональных экономических кооперационных механизмов расширят доступ на рынок и увеличат потоки инвестиций в регионе. Участие в региональных торговых соглашениях и активное участие в глобальных цепочках создания стоимости повысит конкурентоспособность и привлечет иностранные инвестиции.

Улучшение финансовой системы: Создание надежной и прозрачной финансовой системы является важным фактором привлекательности инвестиций. Правительства должны работать над развитием эффективных банковских и финансовых рынков, обеспечивать доступ к доступному финансированию для бизнеса. Продвижение либерализации финансовых рынков, укрепление защиты прав инвесторов и развитие корпоративного управления внедрит доверие в инвесторов и поощрит долгосрочные инвестиции.

Инвестиции в человеческий капитал: Развитие образованной и квалифицированной рабочей силы является критическим для привлечения инвестиций, требующих специализированных знаний и навыков. Правительства должны инвестировать в образовательные программы и профессиональную подготов-

ку, чтобы соответствовать потребностям отраслей. Кроме того, стоит поощрять предпринимательство и создавать благоприятную среду для инноваций, что позволит привлекать инвестиции в высокорастущие сектора.

Улучшение условий для ведения бизнеса: Правительства должны продолжать работы по улучшению условий для ведения бизнеса, упрощать административные процедуры, сокращать бюрократические преграды и обеспечивать эффективные государственные услуги. Создание специализированных агентств по привлечению инвестиций может предоставить инвесторам универсальные услуги, помощь на всех этапах инвестиционного процесса и решение возникающих проблем.

Укрепление регионального сотрудничества: Китай и страны АСЕАН-5 должны тесно сотрудничать в рамках инициатив по привлечению инвестиций. Обмен передовыми практиками, обмен информацией и создание условий для кроссграничных инвестиций улучшат общий инвестиционный климат в регионе. Совместные маркетинговые мероприятия, инвестиционные форумы и бизнес совмещения помогут привлечь потенциальных инвесторов и развить экономическое сотрудничество.

Экологическая устойчивость: Уделять особое внимание экологической устойчивости и поддержке зеленых инвестиций может повысить привлекательность страны или региона. Реализация политик, приоритизирующих использование возобновляемых источников энергии, устойчивое развитие и экологически дружественные практики, привлечет социально ответственных инвесторов и способствует долгосрочному экономическому росту.

Реализация этих рекомендаций поможет создать более благоприятную инвестиционную среду, привлечь разнообразные инвестиции и способствовать устойчивому экономическому развитию в Китае и странах АСЕАН-5. Важно постоянно оценивать и адаптировать политику в соответствии с изменяющимися потребностями и динамикой глобального инвестиционного ландшафта.

Глава 2

Исследование привлекательности инвестиций: глубинный анализ с использованием метода кластерного анализа

В данной главе представлен всесторонний обзор принципов, предположений и сценариев применения метода кластерного анализа. Кроме того, продемонстрировано применение кластерного анализа с использованием языка R для изучения привлекательности инвестиций в различных регионах Китая. Анализ включает создание тепловых карт для визуализации результатов и соответствующие алгоритмы. Процедуры сбора и предварительной обработки данных, используемые в данном исследовании, согласуются с описанными в разделе 2 главы 1.

Результаты, представленные в этой главе, опубликованы в статьях [99].

2.1 Данные и методы

Методология

Кластерный анализ - это широко используемый метод надзорного обучения, направленный на разбиение сравнимых наблюдений на взаимно исключающие кластеры. Путем распознавания внутренних закономерностей и структур в данных он обеспечивает исследовательский анализ и раскрывает скрытые взаимосвязи или особенности.

Основная цель кластерного анализа заключается в определении сходства между выборками путем использования метрик сходства или расстояния, а

затем разделения выборок на отдельные кластеры в зависимости от степени сходства. Этот процесс включает два важных шага: измерение расстояния и алгоритм кластеризации.

Измерение расстояния предполагает оценку сходства или различия между выборками путем использования метрик расстояния, таких как евклидово расстояние, манхэттенское расстояние и метрика Минковского. Выбор соответствующих метрик зависит от конкретных требований задачи.

Вот некоторые распространенные математические формулы и метрики для расчета сходства или расстояния между точками данных:

Евклидово расстояние: Евклидово расстояние является наиболее часто используемой метрикой расстояния, которая вычисляет прямое расстояние между двумя N-мерными точками данных. Для двух точек данных $x = (x_1, x_2, \cdots, x_N)$ и $y = (y_1, y_2, \cdots, y_N)$ евклидово расстояние может быть представлено как:

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Манхэттенское расстояние: Манхэттенское расстояние является еще одной распространенной метрикой расстояния, которая вычисляет прямое расстояние между двумя N-мерными точками данных, но учитывает только вертикальное и горизонтальное перемещение по осям координат. Для двух точек данных $x = (x_1, x_2, \cdots, x_N)$ и $y = (y_1, y_2, \cdots, y_N))$ манхэттенское расстояние может быть представлено как:

$$d_{manhattan}(x, y) = \sum_{i=1}^{N}|x_i - y_i|$$

Метрика Минковского: Метрика Минковского является обобщенной формой евклидова расстояния и манхэттенского расстояния. Для двух точек данных $x = (x_1, x_2, \cdots, x_N)$ и $y = (y_1, y_2, \cdots, y_N)$ метрика Минковского может быть представлена как:

$$d_{minkowski}(x, y) = \left(\sum_{i=1}^{N}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

где параметр p определяет форму метрики расстояния. При $p = 1$ метрика Минковского эквивалентна манхэттенскому расстоянию, а при $p = 2$ она эквивалентна евклидовому расстоянию.

Кластерные алгоритмы, с другой стороны, создают отдельные кластеры на основе рассчитанных расстояний. Обычно используемые алгоритмы включают в себя K-средних кластеризацию, иерархическую кластеризацию и плотностную кластеризацию. K-средние подходят для ситуаций, когда число кластеров заранее определено, тогда как иерархическая кластеризация создает кластеры с иерархической структурой.

Кластерный анализ служит мощным методом анализа данных путем выявления скрытых закономерностей и структур, что позволяет категоризировать похожие наблюдения в отдельные кластеры. Знакомство с принципами и методами кластерного анализа повышает способность извлекать ценные выводы из данных и оказывает помощь в решении проблемных задач.

2.2 Эксперименты и результаты

2.2.1 Основа для построения модели

Для обеспечения достоверности результатов необходимо установить определенные критерии для эконометрической модели, используемой в данном статистическом исследовании, с акцентом на необходимости наличия количественных факторов. Основная цель данного исследования заключается в разработке моделей, способных оценивать эффективность инвестиций в различных регионах с учетом географических особенностей. Предполагается, что инвестиционная активность подвержена влиянию определенных инвестиционных условий, что указывает на необходимость использования показателя, отражающего объем инвестиций в основной капитал в регионе, в качестве зависимой переменной.

Необходимо признать, что объем инвестиций в основной капитал включает различные экономические аспекты, которые подвержены широкому спектру социально-экономических характеристик. Поэтому первый шаг состоит в выявлении экзогенных факторов, которые должны охватывать финансовые, физические, географические, юридические, социокультурные и экологические аспекты. Таким образом, рассматриваются следующие факторы: уровень доходов на душу населения, стоимость основных активов, связанная с строительством деятельность, ВНП на душу населения и уровень безработицы.

Для построения модели необходимо провести дополнительное исследование на основе следующих предположений. Во-первых, предполагается, что уровень

инвестиционной активности в регионе определяется объемом инвестиций. Во-вторых, предполагается, что привлекательность инвестиций в регион в основном зависит от финансового климата. Цель данного исследования - разработать эконометрическую модель, которая может оценить объем инвестиций в основной капитал в регионе, учитывая линейную связь между наблюдаемыми результатами. Предполагается, что объем инвестиций в основной капитал зависит от нескольких социо-экономических показателей, которые могут быть представлены в виде функции. Таким образом, мы стремимся построить функцию в следующей форме:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \cdots, 31. \tag{2.1}$$

В уравнении i- регион; y_i - оценка объема инвестиций в текущем году; x_{1i} - средний доход на душу населения в текущем году; x_{2i} - стоимость основных активов в текущем году; x_{3i} - объем работ по виду деятельности "Строительство"в текущем году; x_{4i} - ВНП на душу населения в текущем году; x_{5i} - уровень безработицы в текущем году (в процентах).

Использование логарифмов в построении эконометрической модели

В области экономики распространенной практикой является использование логарифмов при построении эконометрических моделей, так как считается, что это улучшает статистические свойства получаемых оценок. Неиспользование логарифмов может привести к гетероскедастичности ошибок, что может снизить эффективность оценок метода наименьших квадратов. В результате становится сложно делать точные статистические выводы о качестве оценок. Путем применения логарифмов дисперсия стабилизируется, что приводит к постоянной дисперсии ошибок независимо от изменений значения независимой переменной. Поэтому необходимо взять натуральный логарифм зависимой переменной относительно основания e, чтобы обеспечить точность последующего статистического анализа. Новообразованная трансформированная функция может быть выражена следующим образом:

$$lny_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \cdots, 31.$$

Все вычисления для данного исследования проводились с использованием статистической среды обработки данных "RStudio".

2.2.2 Кластерный анализ регионов Китая: Выявление закономерностей и взаимосвязей

Перед построением модели оценки был проведен иерархический кластерный анализ набора данных, учитывая его ограниченное количество наблюдений. В качестве алгоритма кластеризации был выбран метод Уорда. Этот метод основан на принципе минимизации суммы квадратов различий внутри каждой группы при одновременном максимизации суммы квадратов различий между группами. В контексте данного исследования метод Уорда оказывается особенно подходящим для классификации показателей.

Основная концепция, лежащая в основе метода Уорда, заключается в начальной обработке каждого из n выборок или переменных как отдельного кластера. Затем кластеры последовательно объединяются таким образом, чтобы увеличение суммы квадратов отклонений, обозначаемое как S, было минимальным при каждом объединении. Объединяются два кластера, которые вносят наименьший вклад в увеличение S, и этот процесс продолжается до тех пор, пока все выборки или переменные не будут присвоены одному конечному кластеру.

Применяя метод Уорда для иерархического анализа кластеров, мы стремимся выявить значимые закономерности и взаимосвязи между переменными в нашем исследовании. Эта процедура позволяет эффективно группировать похожие наблюдения и выявлять внутреннюю структуру данных. Применяя данную методологию, мы сможем получить ценные представления о характеристиках региона Китая, находящегося под исследованием.

В процессе разделения набора данных на k отдельных классов, обозначенных как G_1, G_2, G_3 и так далее до G_k, мы измеряем сумму квадратов отклонений в каждом классе. Это может быть представлено уравнением:

$$S_t = \sum_{i=1}^{n_t} (X_{it} - \bar{X}_t)^T (X_{it} - \bar{X}_t),$$

Здесь X_{it} представляет i-й образец в классе G_t (m-мерный вектор), n_t обозначает количество образцов в G_t, а X_t обозначает центроид или центр тяжести для класса G_t. При объединении двух классов, скажем, G_p и G_q, в новый объединенный класс G_r, мы рассчитываем три суммы квадратов: Sp, Sq и Sr. Увеличение суммы квадратов, обозначаемое как D^2_{pq}, определяется как $D^2_{pq} = Sr - Sp - Sq$.

Величина D^2_{pq} дает представление о разумности объединения двух классов.

Если G_p и G_q тесно связаны, то D^2_{pq} будет меньше, что указывает на более обоснованную классификацию. Напротив, если D^2_{pq} больше, это указывает на неразумную классификацию. Таким образом, рассматривая сумму квадратов отклонений, полученных при объединении двух классов, как квадрат расстояния, мы получаем формулу для расстояния:

$$D^2_{pq} = \frac{n_p n_q}{n_r} \left(X_p - \bar{X}_q\right)^T \left(X_p - \bar{X}_q\right).$$

Кроме того, мы можем использовать рекуррентную формулу для расчета квадрата расстояния при объединении класса G_k с классом G_r:

$$D^2_{kr} = \frac{n_k + n_p}{n_r + n_k} D^2_{kp} + \frac{n_k + n_q}{n_r + n_k} D^2_{kq} - \frac{n_k}{n_r + n_k} D^2_{pq}.$$

Метод Уорда принимает квадрат евклидового расстояния в качестве статистической меры классификации. Для любых двух образцов i и j евклидово квадратное расстояние определяется как:

$$\begin{aligned} d^2_{ij} = &\left(X_{i1} - X_{j1}\right)^2 + \left(X_{i2} - X_{j2}\right)^2 + \cdots + \\ &+ \left(X_{i\ \mathrm{m}} - X_{j\ \mathrm{m}}\right)^2 = \sum_{n=1}^{\mathrm{m}} \left(X_{i\mathrm{n}} - X_{j\mathrm{n}}\right)^2. \end{aligned} \tag{2.2}$$

Применяя эти формулы и уравнения, метод Уорда позволяет проводить иерархический кластерный анализ на основе квадратов расстояний между образцами, облегчая выявление значимых кластеров в наборе данных.

В уравнении 2.2, X_{in} и X_{jn} представляют значение n-й переменной для i-го образца и j-го образца соответственно. Для снижения влияния масштаба переменных на измерение расстояний между образцами в кластерном анализе обычной практикой является стандартизация переменных. Путем стандартизации переменных мы преобразуем их в стандартизированные значения, которые затем используются для проведения кластерного анализа.

Стандартизация включает вычитание среднего значения каждой переменной из ее индивидуальных наблюдений и деление на стандартное отклонение этой переменной. Этот процесс гарантирует, что все переменные имеют сопоставимый масштаб, что позволяет справедливо и значимо сравнивать их в процессе кластеризации. Используя стандартизированные значения, влияние переменных с большими масштабами не затмевает тех с меньшими масштабами, поскольку они находятся на подобном уровне положения.

Процедура стандартизации повышает надежность и интерпретируемость результатов кластерного анализа. Она позволяет точно выявлять закономерности и взаимосвязи между образцами на основе относительных расстояний между ними, а не абсолютных значений переменных. Такой подход гарантирует, что алгоритм кластеризации фокусируется на внутренней структуре и сходствах между образцами, а не искажается различиями в масштабах переменных.

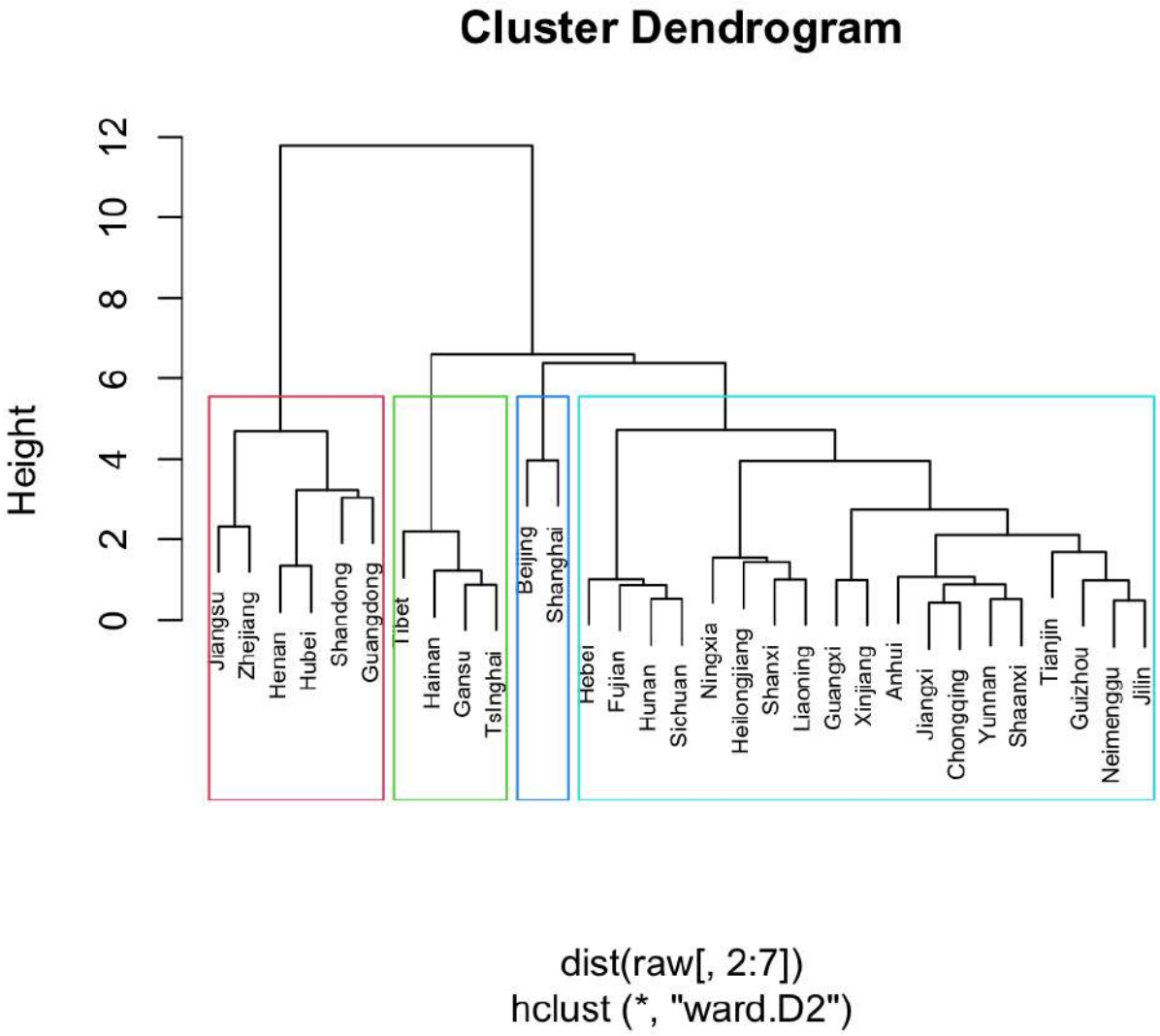


Рис. 2.1: Cluster Dendrogram

Метод реализован на языке R следующим образом:

library("factoextra")

d < - dist(my data, method = "euclidean")

res.hc < - hclust(d, method = "ward.D2").

Результаты вычислений наглядно представлены на рисунке 2.1, где показано дерево разбиения. В данном анализе использовался набор данных из надежных

Таблица 2.1: Описательная статистика для кластера 1

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	2 693	20 854	10 087	35 478	2.5	10.36
1st Qu.	21 067	24 835	11 398	46 357	2.625	10.42
Median	28 380	33 730	12 434	62 201	2.750	10.62
Mean	26 116	31 348	16 920	63 335	2.833	10.63
3rd Qu.	34 519	37 118	23 775	82 561	2.950	10.84
Max.	42 046	39 658	27 957	89 705	3.40	10.92

Таблица 2.2: Описательная статистика для кластера 2

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	11 547	1 376	147.9	1 311	2.3	7.589
1st Qu.	12 395	2 456	279.1	2 296	2.6	8.096
Median	17 506	3 310	364.8	3 544	2.7	8.309
Mean	14 778	2 973	675.8	3 965	2.7	8.219
3rd Qu.	19 889	3 828	761.6	5 212	:2.8	8.433
Max.	22 553	3 897	1 825.4	7 460	3.1	8.670

Таблица 2.3: Описательная статистика для кластера 3

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	57 230	10 946	6 426	28 015	1.400	8.888
1st Qu.	57 669	11 258	7 254	28 669	2.025	8.924
Median	58 109	11 570	8 082	29 324	2.650	8.960
Mean	58 109	11 570	8 082	29 324	2.650	8.960
3rd Qu.	58 548	11 881	8 909	29 978	3.275	8.996
Max.	58 988	12193	9 737	30 633	3.900	9.032

Таблица 2.4: Описательная статистика для кластера 4

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	16 704	3 807	549.2	3 444	2.200	8.224
1st Qu.	20 571	10 003	2 675.8	15 999	3.250	9.366
Median	21 484	10 467	4 726.4	19 425	3.500	9.772
Mean	24 843	12 285	5 074.6	21 459	3.479	10.130
3rd Qu.	25 183	14 951	6 628.0	28 825	3.900	10.130
Max.	58 988	19 083	11 400.3	36 980	4.200	10.417

источников, таких как Всемирный банк и Статистический годовой отчет Китая. Данные были скорректированы с учетом инфляции для обеспечения сравнимости значений. На рисунке показано разделение данных на отдельные кластеры,

каждый из которых обозначен уникальным цветом. В данном анализе данные были разделены на четыре кластера, что предоставляет иерархическое представление внутренней структуры данных. Эта иерархическая характеристика явно видна при рассмотрении результатов. Кроме того, описательная статистика, связанная с кластерами, подтверждает этот вывод.

В последующем анализе будет проведено более детальное рассмотрение каждого кластера, чтобы получить представление о его характеристиках.

Группа 1 состоит из шести регионов, характеризующихся значительным объемом инвестиций. Хотя между этими регионами есть заметные различия в значениях отдельных факторов, различия в привлекательности инвестиций между ними не достигают статистической значимости. Описательная статистика в отношении исследуемых показателей классификации представлена в таблице 2.1.

Группа 2 состоит из четырех регионов с относительно низким уровнем инвестиций. Описательная статистика в этой группе демонстрирует более высокую степень согласованности по сравнению с Группой 1.

Группа 3 включает два региона, а именно Пекин и Шанхай, которые выделяются как наиболее процветающие области в стране. Различия между этими двумя регионами по каждому фактору относительно незначительны.

Группа 4 является самым крупным кластером и включает девятнадцать регионов. Регионы этой группы не только имеют значительные отличия в объеме инвестиций, но и значительно отличаются по другим факторам.

Проводя подробный анализ каждого кластера, мы получаем ценные представления о различных характеристиках и закономерностях, наблюдаемых в данных. Такой подход позволяет полноценно понять региональные динамики и инвестиционные ландшафты в изучаемых областях.

Определение оптимального количества кластеров, обозначаемого как k, имеет большое значение при применении метода "k-средних". Для оценки релевантности выбора k мы используем метод "локтя"(elbow method), построив зависимость межкластерного рассеивания от количества кластеров. Как показано на рисунке 2.2, заметное уменьшение межкластерного рассеивания происходит при $k = 2$, после чего оно стабилизируется при $k = 4$. Этот анализ указывает на то, что разделение кластеров на четыре группы более точно соответствует географическим и экономическим характеристикам регионов Китая.

На основе этих результатов китайские регионы были разделены на четы-

ре отдельных кластера, отличающихся уровнем инвестиций и географическими атрибутами. Результаты кластерного анализа раскрывают заметную связь между привлекательностью инвестиций в регионах и различными факторами, включая доход на душу населения, стоимость основных активов, валовый национальный продукт (ВНП), строительную активность и уровень безработицы. Поэтому применение методов множественной регрессионной анализа для разработки моделей на основе наблюдений в этих кластерах имеет значимое значение.

Путем использования методов множественного регрессионного анализа мы можем создать предиктивные модели, учитывающие указанные выше факторы, чтобы понять их влияние на привлекательность инвестиций в различных регионах. Эти модели могут помочь выявить ценные представления и принимать обоснованные решения в контексте регионального развития и инвестиционных стратегий.

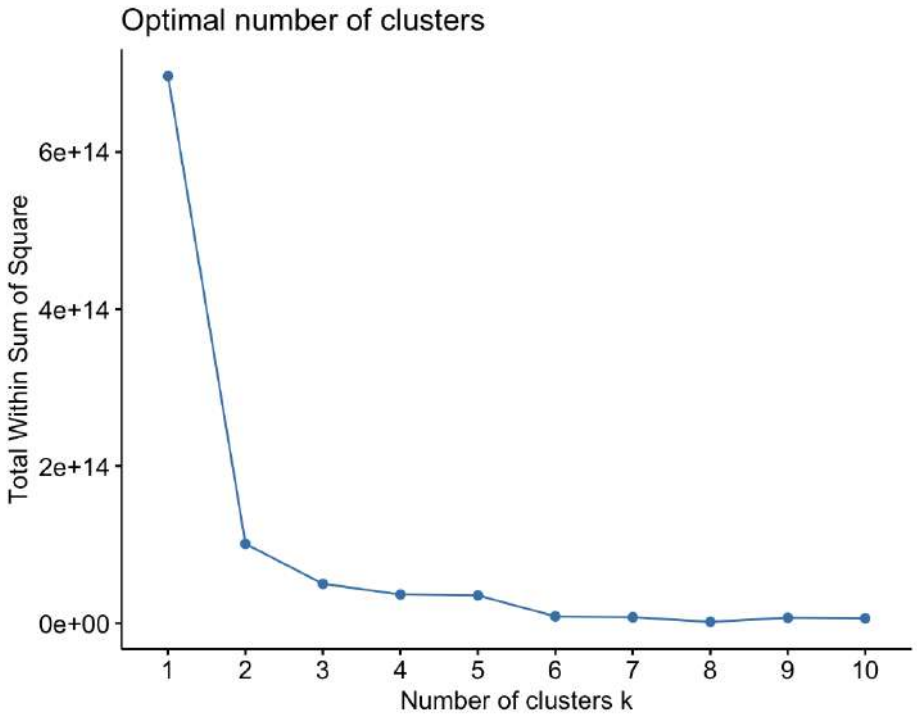


Рис. 2.2: Определение релевантности выбора числа кластеров k

2.2.3 Построение эконометрических моделей

Для всестороннего анализа больших кластеров будет проведен множественный регрессионный анализ для каждого года доступных данных. Такой подход позволяет создавать модели, которые улавливают взаимосвязь между различными факторами и предоставляют глобальную оценку. Путем комбинирования

этих моделей мы можем сделать общие выводы о влиянии рассматриваемых факторов на привлекательность инвестиций.

Учитывая, что отдельные модели будут разработаны для каждой группы наблюдений внутри кластеров, разумно предположить линейную зависимость между факторами. Линейная регрессия предоставляет систематический и количественный способ оценки влияния независимых переменных на интересующую нас зависимую переменную. Это предположение помогает интерпретировать коэффициенты регрессионных моделей и позволяет определить величину и направление влияния каждого фактора.

Для более маленьких кластеров, где количество наблюдений может быть ограничено, корреляционный анализ предоставляет ценные идеи о степени влияния факторов. Корреляционный анализ помогает определить силу и направление линейной связи между парами переменных. Изучая коэффициенты корреляции, мы можем оценить, насколько изменения одной переменной соответствуют изменениям другой, тем самым оценивая степень взаимосвязи между факторами.

Применение множественного регрессионного анализа для больших кластеров и корреляционного анализа для малых кластеров позволяет использовать соответствующие статистические методы, адаптированные под характеристики каждого размера кластера. Эти аналитические подходы облегчают всесторонний анализ взаимодействия факторов и их влияния на привлекательность инвестиций в различных контекстах.

Модель множественной регрессии для группы 4

Анализ был сосредоточен на Группе 4, состоящей из 19 регионов, которые были выбраны в качестве основной выборки для исследования. Данные, собранные из этих 19 регионов, были использованы для построения наблюдательной модели, позволяющей исследовать взаимосвязь между различными факторами и привлекательностью инвестиций внутри этой группы.

Используемая в данном исследовании наблюдательная модель улавливает взаимодействие между независимыми переменными и привлекательностью инвестиций как зависимой переменной. Используя данные из каждого из 19 регионов в Группе 4, мы стремимся выявить значимые факторы, влияющие на вариации привлекательности инвестиций в этих регионах.

Через эту наблюдательную модель мы стремимся раскрыть особенности и факторы, определяющие привлекательность инвестиций внутри выбранных регионов. Такой подход позволяет изучить уникальные характеристики и факторы, определяющие инвестиционные тенденции, и получить более глубокое понимание их влияния.

Используя данные из всех 19 регионов в Группе 4, мы стремимся предоставить всесторонний анализ, учитывая разнообразные экономические, географические и социальные контексты этих регионов. Такой подход позволяет сделать обоснованные выводы и рекомендации относительно инвестиционных стратегий и инициатив по региональному развитию на основе выявленных взаимосвязей между факторами и привлекательностью инвестиций в рамках этой конкретной группы.

$$\ln y_{2017i} = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \beta_4 \cdot x_{4i} + \beta_5 \cdot x_{5i} + \epsilon_i, \tag{2.3}$$

где i - регион; ϵ_i - общий эффект факторов, не учтенных моделью.

Оценки регрессионного уравнения 2.3 были найдены и результаты представлены в таблице 2.5.

Таблица 2.5: Оценки коэффициентов для группы 4

coefficient	Estimate	Std. Error	t–value	Pr(>|t|)	Signif.
$\hat{\alpha}$	1.005e+01	4.828e-01	20.817	2.29e-11	*** [1]
$\hat{\beta}_1$	-2.508e-05	8.018e-06	-3.128	0.00801	** [2]
$\hat{\beta}_2$	6.247e-05	3.399e-05	1.838	0.08901	. [3]
$\hat{\beta}_3$	3.220e-05	3.968e-05	0.812	0.43167	
$\hat{\beta}_4$	2.418e-05	1.956e-05	1.236	0.23822	
$\hat{\beta}_5$	-3.442e-01	1.334e-01	-2.580	0.02284	* [4]

[1] Значение p-значения меньше 0,001 указывает на очень сильные доказательства против нулевой гипотезы.

[2] Значение p-значения меньше 0,01 указывает на еще более сильные доказательства против нулевой гипотезы.

[3] Значение p-значения меньше 0,1 считается слабым доказательством против нулевой гипотезы.

[4] Значение p-значения меньше 0,05 указывает на сильные доказательства против нулевой гипотезы, и мы можем отклонить ее в пользу альтернативной гипотезы.

Следовательно, регрессионное уравнение имеет следующий вид:

R^2	0.86
R^2_{adj}	0.8062
F	15.98
p–value(F)	3.756e-05

$$\ln \hat{y_{2017i}} = 10.05 - 0.00002508 \cdot x_{1i} + 0.00006247 \cdot x_{2i} + 0.0000322 \cdot x_{3i} + \\ + 0.000002.418 \cdot x_{4i} - 0.3442 \cdot x_{5i},$$

Согласно t-тесту, три из пяти оценок коэффициентов ($\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_5$) являются статистически значимыми на уровне значимости 5%, с p-значением меньше 0.1. Значения R^2 и R^2_{adj} составляют соответственно 0.86 и 0.8062, что означает, что примерно 80% вариации зависимой переменной объясняется регрессией. Кроме того, общее качество модели достаточно хорошее, о чем свидетельствует критерий Фишера: $F - statistics = 15.98$, соответствующее p-значение = 3.756e-05, которое меньше 0.05 и близко к нулю. Этот результат подтверждает, что среднее качество всей модели является удовлетворительным.

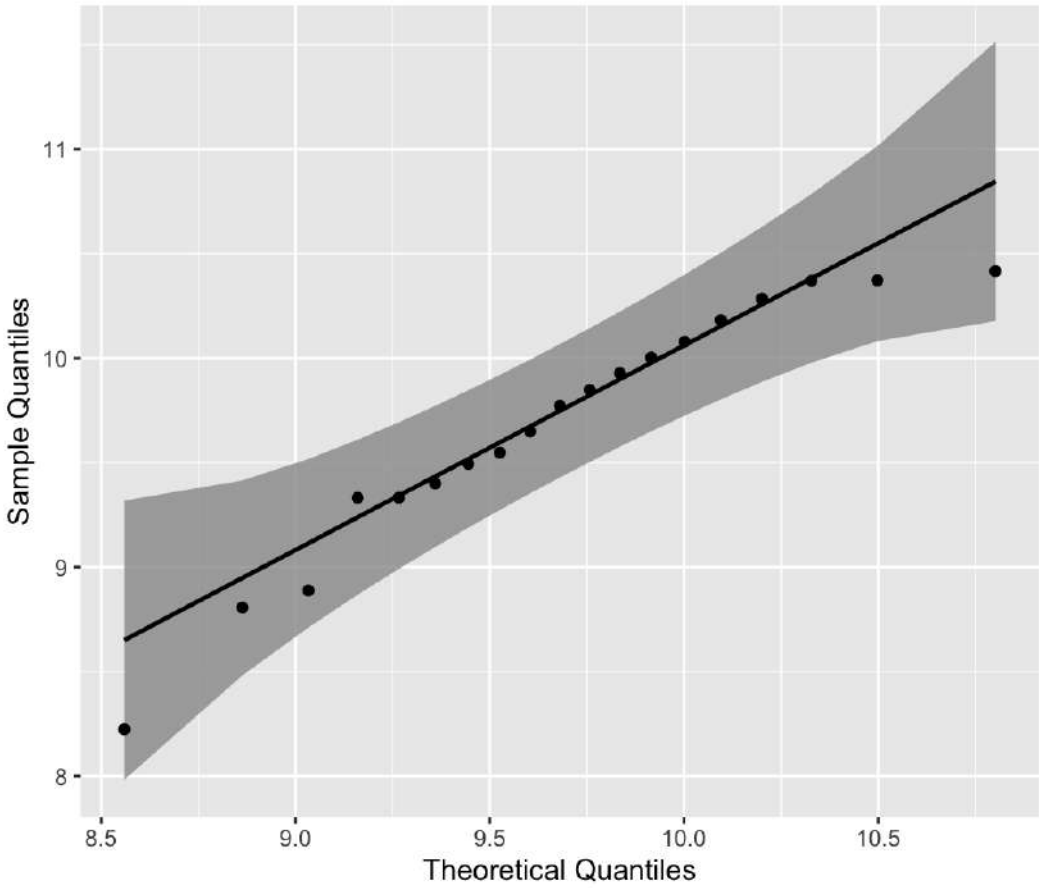


Рис. 2.3: график остатков группы 4.

Для оценки нормальности распределения остатков важно рассмотреть QQ-график (рисунок 2.3). QQ-график предоставляет графическое представление, которое позволяет сравнить наблюдаемые квантили остатков с ожидаемыми квантилями в рамках теоретического нормального распределения.

При анализе QQ-графика можно заметить, что большинство точек тесно выстраиваются по одной линии, что указывает на то, что распределение остатков соответствует нормальному распределению. Однако стоит отметить наличие некоторых выбросов, которые отклоняются от ожидаемого паттерна. Эти выбросы могут указывать на потенциальные отклонения от нормальности в определенных областях распределения остатков.

Нормальность распределения остатков является важным предположением для многих статистических моделей и тестов. Отклонение от нормальности может указывать на наличие влиятельных наблюдений или нарушения предположений модели. Поэтому необходимо внимательно оценивать и интерпретировать эти выбросы в связи с общими характеристиками данных.

Хотя большинство точек на QQ-графике соответствуют ожидаемому паттерну нормального распределения, наличие выбросов вызывает некоторую осторожность. Для дальнейшей проверки и оценки необходимо провести дополнительную диагностику и исследование характера и влияния этих выбросов на корректность модели. Оценка нормальности распределения остатков способствует обеспечению надежности и устойчивости проводимого в данном исследовании статистического анализа.

lag	Autocorrelation	DW Statistic	p–value
1	0.15803	1.665274	0.194

χ^2	p-value(Breusch-Pagan)
2.945036	0.086142

Для проверки точности выбранных факторов, включенных в модель, была проведена оценка автокорреляции остатков. Для этой цели использовался критерий Дарбина-Уотсона, результатом которого является значение $DW = 1.6653$. Это значение указывает на отсутствие значительной автокорреляции остатков и подтверждает правильность выбора факторов в модели.

Однако было необходимо дальше изучить гомоскедастичность остатков. Для

этого был использован тест Бройша-Пагана, который привел к значению статистики теста $\chi^2 = 2.945036$ и соответствующему p-значению равному 0.0861. Хотя p-значение не достигает обычных уровней статистической значимости, оно указывает на умеренное свидетельство в пользу гомоскедастичности остатков. Таким образом, на основе этого теста есть ограниченные данные, указывающие на отсутствие гетероскедастичности в остатках.

Анализ остатков регрессии в целом дал удовлетворительные результаты, поскольку тесты на автокорреляцию и гетероскедастичность не выявили серьезных нарушений предположений. Однако важно отметить, что эти тесты имеют свои ограничения и должны рассматриваться в сочетании с другими диагностическими мерами для обеспечения надежности модели.

Таблица 2.6: Оценка мультиколлинеарности - значения VIF

Features	VIF Factor
const	0.003504
x_1	0.049529
x_2	0.015893
x_3	0.009322
x_4	0.008444
x_5	0.027871

Таблица 2.6 показывает значение фактора инфляции дисперсии (VIF) для каждой переменной в модели. VIF является мерой мультиколлинеарности, которая количественно оценивает, насколько дисперсия коэффициента регрессии увеличивается из-за корреляции с другими предикторами.

Колонка "Признаки"представляет названия переменных, включенных в модель. Колонка "Значение VIF"отображает вычисленные значения VIF для каждой переменной.

В таблице мы наблюдаем следующее:

Константа (интерсепт) имеет значение VIF фактора 0,003504, что указывает на очень низкую мультиколлинеарность с другими предикторами. Переменная x_1 имеет значение VIF фактора 0,049529, что говорит о умеренной мультиколлинеарности. Переменные x_2, x_3, x_4 и x_5 имеют значения VIF факторов соответственно 0,015893, 0,009322, 0,008444 и 0,027871, что указывает на относительно низкий уровень мультиколлинеарности. В целом, значения VIF факторов яв-

ляются относительно низкими, что говорит о отсутствии серьезной проблемы мультиколлинеарности среди переменных в модели. Это означает, что переменные можно рассматривать как независимые, и их коэффициенты могут быть надежно интерпретированы в контексте академического исследования.

Эти результаты свидетельствуют о приемлемом уровне коллинеарности и подтверждают соответствие переменных для включения в данную работу.

Для получения окончательной модели был использован алгоритм пошаговой регрессии. Этот алгоритм систематически определяет фактор(ы) с наименее значимыми коэффициентами (максимальное p-значение) на каждом шаге и удаляет их из модели. Путем итеративного применения этой процедуры алгоритм последовательно уточняет модель до тех пор, пока все факторы не будут иметь статистически значимые коэффициенты. Использование подхода пошаговой регрессии помогает упростить модель, исключая менее значимые факторы и улучшая ее интерпретируемость и прогностическую способность.

$$\ln \hat{y_{2017i}} = 9.666 - 0.00001741 \cdot x_{1i} + 0.000126 \cdot x_{2i} - 0.3164 \cdot x_{5i};$$

Примените тот же процесс к данным за другие годы 2009-2016:

$$2009 : \ln \hat{y_{2009i}} = 7.473 + 0.0001078 \cdot x_{2i} + 0.00000001253 \cdot x_{3i} + 0.00003056 \cdot x_{4i};$$

$$2010 : \ln \hat{y_{2010i}} = 7.988 - 0.00003046 \cdot x_{1i} + 0.0001484 \cdot x_{2i} + 0.00000001938 \cdot x_{3i};$$

$$2011 : \ln \hat{y_{2011i}} = 8.539 + 0.000000003164 \cdot x_{3i} + 0.00006634 \cdot x_{4i} - 0.1231 \cdot x_{5i};$$

$$2012 : \ln \hat{y_{2012i}} = 9.025 - 0.00001169 \cdot x_{1i} + 0.00006653 \cdot x_{4i} - 0.1351 \cdot x_{5i};$$

$$2013 : \ln \hat{y_{2013i}} = 9.452 - 0.00001755 \cdot x_{1i} + 0.00001159 \cdot x_{2i} + 0.00005484 \cdot x_{4i} - \\ - 0.1679 \cdot x_{5i};$$

$$2014 : \ln \hat{y_{2014i}} = 9.677 - 0.00001669 \cdot x_{1i} + 0.00005619 \cdot x_{4i} - 0.229 \cdot x_{5i};$$

$$2015 : \ln \hat{y_{2015i}} = 9.807 - 0.00002218 \cdot x_{1i} + 0.000000003823 \cdot x_{3i} + 0.00003945 \cdot x_{4i} - \\ - 0.16768 \cdot x_{5i};$$

$$2016 : \ln \hat{y_{2016i}} = 9.122 - 0.00002709 \cdot x_{1i} + 0.00007275 \cdot x_{2i} + 0.00000000645 \cdot x_{3i}.$$

Множественная регрессионная модель для группы 1

Для группы 1 была построена множественная регрессионная модель с использованием метода анализа, представленного в таблице 2.7.

Таблица 2.7: Оценки коэффициентов для группы 1

coefficient	Estimate	Std. Error	t–value	Pr(>|t|)	Signif.
$\hat{\alpha}$	9.130e+00	4.838e-01	18.873	0.0337	*
$\hat{\beta_1}$	-5.831e-06	4.699e-06	-1.241	0.4318	
$\hat{\beta_2}$	4.330e-05	1.367e-05	3.168	0.1947	
$\hat{\beta_3}$	1.585e-05	5.204e-06	3.045	0.2020	
$\hat{\beta_4}$	-9.283e-06	4.977e-06	-1.865	0.3133	
$\hat{\beta_5}$	2.178e-01	1.000e-01	2.178	0.2741	

Следовательно, регрессионное уравнение имеет следующий вид:

$$\ln \hat{y_{2017i}} = 9.13 - 0.000005831 \cdot x_{1i} + 0.0000433 \cdot x_{2i} + 0.00001585 \cdot x_{3i} - \\ - 0.000009283 \cdot x_{4i} + 0.2178 \cdot x_{5i},$$

Статистическая значимость коэффициентов, полученных из модели, представлена в Таблице 6. Мы наблюдаем, что только оценки свободных коэффициентов демонстрируют статистическую значимость на уровне значимости 5%.

Для всесторонней оценки общего качества модели мы обращаем наше внимание на первую группу. Эта оценка позволяет оценить, насколько модель отражает и объясняет наблюдаемые данные для этой конкретной группы.

На основе нашего анализа можно утверждать, что среднее качество всей модели является удовлетворительным. Этот вывод делается на основе различных факторов, таких как статистическая значимость коэффициентов, соответствие

R^2	0.9824
R^2_{adj}	0.8946
F	11.19
$p–value(F)$	0.0223

модели наблюдаемым данным и другие соответствующие показатели производительности.

Создавая модель, которая охватывает значимые и влиятельные факторы, мы можем получить ценные представления о взаимосвязи между этими факторами и интересующим нас результатом. Это понимание способствует более полному пониманию основной динамики и механизмов, действующих в рамках первой группы. Кроме того, оно служит основой для дальнейшего исследования и усовершенствования модели, что в конечном итоге приводит к улучшению ее прогностических возможностей и принятию решений.

F-статистика, соответствующая модели, является статистически значимой с p-значением меньше 0.05, что указывает на то, что модель принимается на уровне значимости 5%. Модель демонстрирует высокую степень объяснительной силы, поскольку она объясняет около 90% вариации зависимых переменных.

lag	Autocorrelation	D-W Statistic	p-value
1	-0.4205415	2.354857	0.162

χ^2	p-value(Breusch-Pagan)
0.1263405	0.07222

Статистика Дарбина-Уотсона (D-W) со значением 2,35 находится в диапазоне от 1,5 до 2,5, что указывает на то, что остатки в модели для первой группы не проявляют значительной автокорреляции. Этот результат свидетельствует о том, что предположение о независимости в регрессионном анализе соблюдается и подтверждает достоверность полученных результатов.

Кроме того, тест Бройша-Пагана, который оценивает гомоскедастичность остатков, показывает p-значение больше 0,05. Это означает, что нет достаточных доказательств для отклонения от нулевой гипотезы о гомоскедастичности остатков. Хотя данный тест не обеспечивает окончательного доказательства гомоскедастичности, он указывает на то, что любые отклонения от гомоскеда-

стичности, вероятно, являются незначительными и не оказывают существенного влияния на общее качество модели.

На основе этих статистических анализов можно сделать вывод, что качество модели для первой группы считается нормальным. Результаты указывают на то, что предположения, лежащие в основе модели, такие как отсутствие автокорреляции остатков и незначительные отклонения от гомоскедастичности, разумно выполняются в данном контексте.

Таблица 2.8: Оценка мультиколлинеарности - значения VIF

Features	VIF Factor
const	0.011989
x_1	0.088700
x_2	0.060412
x_3	0.032016
x_4	0.025576
x_5	0.020602

Таблица 2.8 предоставляет значения фактора инфляции дисперсии (VIF) для переменных, включенных в модель.

Константа (обозначенная как "const") имеет значение VIF равное 0.01198953, что указывает на минимальную коллинеарность с другими предикторами. Это говорит о том, что константа не оказывает существенного влияния от корреляций с остальными переменными.

Переменная x_1 имеет значение VIF равное 0.088700, указывая на некоторую степень коллинеарности с другими предикторами. Хотя есть умеренная степень корреляции, она остается в приемлемом диапазоне.

Переменные "x_2,x_3,x_4,"и "x_5"обладают значениями VIF соответственно: 0.0604120, 0.03201671, 0.0255766 и 0.0206025. Эти значения указывают на относительно низкий уровень коллинеарности с другими предикторами.

В целом, значения VIF показывают, что между переменными, рассматриваемыми в модели, нет серьезной проблемы мультиколлинеарности. Это означает, что переменные являются достаточно независимыми, что позволяет надежно интерпретировать их коэффициенты в рамках данного исследования.

Для получения окончательной модели был использован алгоритм пошаговой регрессии. Этот итеративный процесс систематически определяет фактор с наи-

менее значимым коэффициентом (максимальное p-значение) на каждом шаге и удаляет его из модели. Путем итеративного исключения менее влиятельных факторов алгоритм уточняет модель и выбирает наиболее значимые переменные для включения. Такой подход позволяет упростить модель, фокусируясь на статистически значимых факторах, что повышает ее интерпретируемость и прогностическую способность.

Использование алгоритма пошаговой регрессии позволяет определить ключевые факторы, которые значительно влияют на наблюдаемый результат в первой группе. Включение только наиболее значимых переменных в окончательную модель эффективно улавливает существенную информацию и взаимосвязи внутри этой конкретной когорты.

$$\ln \hat{y_{2017i}} = 9.179 + 0.0000371 \cdot x_{2i} - 0.000005542 \cdot x_{4i} + 0.2202 \cdot x_{5i},$$

Примените тот же процесс к данным за другие годы.

$$2009 : \ln \hat{y_{2009i}} = 8.518 + 0.00009.835 \cdot x_{2i} - 0.00001247 \cdot x_{4i};$$

$$2010 : \ln \hat{y_{2010i}} = 6.261 - 0.0001188 \cdot x_{1i} - 0.00006069 \cdot x_{2i} + 0.000000007946 \cdot x_{3i};$$

$$2011 : \ln \hat{y_{2011i}} = 8.072 + 0.00006173 \cdot x_{2i} + 0.1711 \cdot x_{5i};$$

$$2012 : \ln \hat{y_{2012i}} = 8.195 + 0.00004987 \cdot x_{2i} + 0.000000001065 \cdot x_{3i} + 0.1992 \cdot x_{5i};$$

$$2013 : \ln \hat{y_{2013i}} = 8.272 + 0.00004462 \cdot x_{2i} + 0.000000001145 \cdot x_{3i} + 0.231 \cdot x_{5i};$$

$$2014 : \ln \hat{y_{2014i}} = 9.788 - 0.00002668 \cdot x_{1i} + 0.00002101 \cdot x_{2i} + 0.000000002385 \cdot x_{3i};$$

$$2015 : \ln \hat{y_{2015i}} = 10 - 0.00002476 \cdot x_{1i} + 0.00002912 \cdot x_{2i} + 0.000000002255 \cdot x_{3i};$$

$$2016 : \ln \hat{y_{2016i}} = 10.19 - 0.00002719 \cdot x_{1i} + 0.00002742 \cdot x_{2i} + 0.000000002265 \cdot x_{3i}.$$

2.2.4 Исследования, специфичные для кластеров

Для оценки влияния различных факторов на привлекательность инвестиций в данном исследовании был выбран метод множественного регрессионного анализа. Такой аналитический подход позволяет всесторонне изучить взаимосвязь между независимыми переменными и интересующей нас зависимой переменной. Однако, учитывая наличие иерархической структуры в наборе данных, было необходимо учесть это обстоятельство для обеспечения точности анализа.

Для смягчения потенциального влияния иерархической структуры наблюдений было проведено деление на четыре отдельные группы. Это деление облегчило более точный анализ путем отдельного рассмотрения каждой группы. Таким образом, мы стремились учесть уникальные характеристики и динамику, присущую каждой группе, тем самым уменьшив возможные смещения или побочные факторы, которые могут возникнуть из-за иерархической структуры.

Следует отметить, что для небольших кластеров множественный регрессионный анализ может быть затруднен из-за ограниченного объема выборки и недостаточной статистической мощности. В качестве альтернативы был использован корреляционный анализ для оценки степени влияния факторов в этих более маленьких кластерах. Корреляционный анализ предоставляет представление о силе и направлении линейной связи между переменными, что позволяет оценить величину их взаимосвязи.

Применение множественного регрессионного анализа для больших кластеров и корреляционного анализа для малых кластеров позволило использовать соответствующие статистические методы, адаптированные под характеристики каждого размера кластера. Эта методология позволила эффективно изучить и измерить влияние факторов на привлекательность инвестиций, учитывая при этом иерархическую структуру и ограничения, связанные с анализом малых выборок. В целом такой подход повышает строгость и достоверность полученных результатов и способствует более полному пониманию факторов, влияющих на привлекательность инвестиций.

Для проведения статистического анализа второй и третьей групп был использован корреляционный анализ из-за малого количества кластеров, включенных в анализ.

Для второй группы была создана тепловая карта корреляции corrplot (ри-

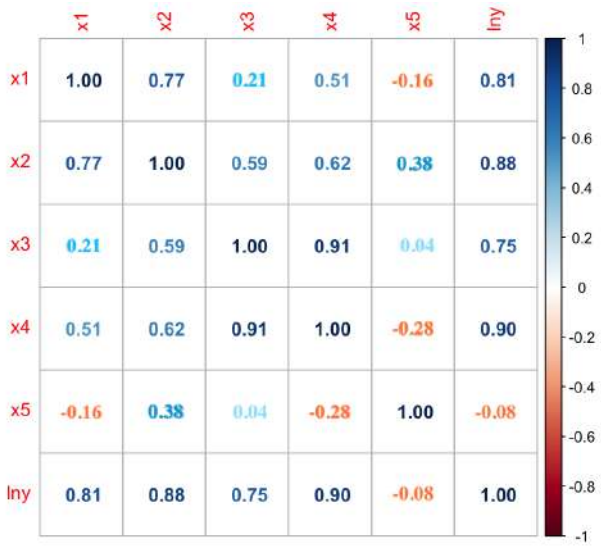


Рис. 2.4: Тепловая карта корреляции для 2 групп

Рис. 2.5: Карта после удаления x5

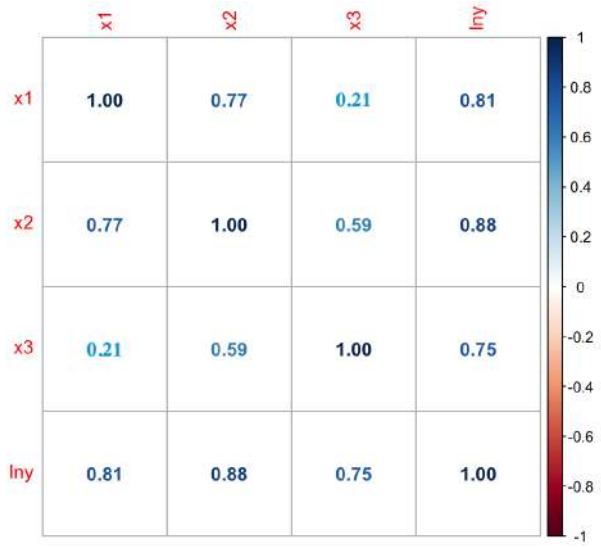


Рис. 2.6: Карта после удаления x5, x4

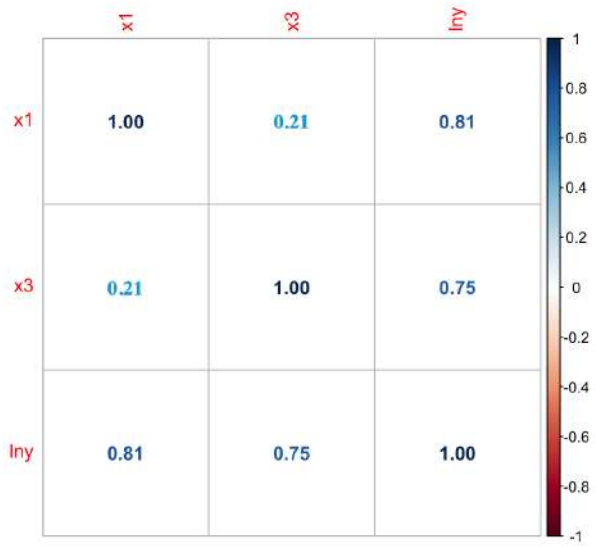


Рис. 2.7: Карта после удаления x5, x4, x2

сунок 2.4), которая визуально отображает взаимосвязи между переменными. Особо стоит отметить, что корреляция между переменной x_5 и целевой переменной является слабой $r = 0.08$. Следовательно, логично исключить x_5 из набора факторов. После удаления x_5, тепловая карта корреляции принимает измененную форму (рисунок 2.5).

Как показано на рисунке 5, x_4 демонстрирует сильную корреляцию с $x_3(r > 0.7)$, а также умеренные корреляции с x_1 и x_2 $(0.4 < r < 0.7)$. Поэтому, несмотря на более сильную корреляцию с целевой переменной, разумно исключить x_4 из набора факторов (рисунок 2.6).

На следующем шаге было принято решение исключить x_2 из-за его умеренной корреляции как с x_1 , так и с x_3, при сохранении низкой корреляции между x_1 и x_3. Получены следующие результаты (рисунок 2.7).

Такая последовательная элиминация объясняющих переменных позволяет выделить наиболее сильно коррелированные факторы. В данном случае только x_1 и x_3 остаются в качестве независимых переменных. Однако стоит отметить, что во второй группе x_3 демонстрирует более слабую связь с целевой переменной y по сравнению с x_1. Таким образом, можно заключить, что во второй группе основным фактором, влияющим на формирование значений y, является фактор x_1.

Что касается третьей группы, из-за ограниченного количества данных невозможно представить подробную тепловую карту корреляции. Следовательно, недостаточно информации для получения значимых выводов о статистической

связи между изученными показателями и привлекательностью инвестиций в этой группе. Малое количество наблюдений в этой группе, вероятно, указывает на отклонение от общего шаблона и подчеркивает необходимость дальнейшего исследования и осторожности при интерпретации результатов в рамках этой конкретной группы.

2.3 Обсуждение и анализ

Выбранный подход, основанный на использовании множественного регрессионного анализа для оценки влияния различных факторов на привлекательность инвестиций, позволил всесторонне исследовать взаимосвязь между независимыми переменными и зависимой переменной. Для учета гетерогенности размеров кластеров наблюдений было проведено деление на четыре отдельные группы, для каждого из которых проводился отдельный анализ.

В первой группе, представляющей самые привлекательные области для инвестиций, результаты множественного регрессионного анализа показали сильную корреляцию между объемом инвестиций и стоимостью основных средств (x_2), а также количеством выполненных работ по виду деятельности "Строительство" (x_3). Особо следует отметить, что корреляция со стоимостью основных средств (x_2) была выявлена как самая сильная. Это говорит о том, что в этих высоко привлекательных регионах инвесторы склонны выделять значительную часть своих ресурсов на приобретение дорогостоящих основных средств, что положительно влияет на объем инвестиций.

В случае четвертой группы, характеризующейся низкой привлекательностью для инвестиций, была обнаружена сильная корреляция между объемом инвестиций и средним доходом на душу населения (x_1), а также количеством выполненных работ по виду деятельности "Строительство"(x_3). Однако корреляция со средним доходом на душу населения (x_1) была выявлена как основной фактор, определяющий объем инвестиций в этой группе. Это указывает на то, что в менее привлекательных регионах более низкие уровни среднего дохода на душу населения могут затруднять возможности для инвестиций, что приводит к снижению объемов инвестиций.

Для второй группы, которая проявляла наименьшую привлекательность для инвестиций, проведен корреляционный анализ в связи с малым размером кла-

стера. Результаты указывают на то, что объем инвестиций тесно коррелирует со средним доходом на душу населения (x_1) и количеством выполненных работ по виду деятельности "Строительство"(x_3). И здесь снова средний доход на душу населения (x_1) выступает в качестве основного фактора, влияющего на объем инвестиций. Эти результаты указывают на то, что в регионах с низкой привлекательностью для инвестиций уровень среднего дохода на душу населения играет ключевую роль в привлечении инвестиций.

Важно отметить ограничения данного исследования, такие как возможное наличие других немеряемых факторов, которые могут влиять на привлекательность инвестиций. Кроме того, следует быть осторожным при обобщении полученных результатов из-за специфического контекста и характеристик исследуемых регионов. Дальнейшие исследования необходимы для проверки и расширения полученных результатов, особенно путем включения дополнительных переменных и учета более широкого спектра экономических и социальных факторов, влияющих на инвестиционные решения.

2.4 Заключение к главе 2

В заключение данной главы следует отметить, что в данном исследовании был использован множественный регрессионный анализ для изучения влияния различных факторов на привлекательность инвестиций. Путем разделения наблюдений на четыре отдельные группы и проведения отдельного анализа для каждой группы было достигнуто всестороннее понимание взаимосвязи между независимыми переменными и объемом инвестиций.

Полученные результаты позволили выявить важные факторы, влияющие на привлекательность инвестиций в различных группах. В высоко привлекательных регионах (первая группа) стоимость основных средств оказалась сильным определителем объема инвестиций, указывая на то, что инвесторы в этих районах выделяют значительные ресурсы на приобретение дорогостоящих активов. В регионах с низкой привлекательностью для инвестиций (четвертая группа) основным фактором, влияющим на объем инвестиций, оказался средний доход на душу населения. Это указывает на то, что более низкие уровни среднего дохода на душу населения в этих регионах могут затруднять возможности для инвестиций.

Для второй группы, которая проявляла наименьшую привлекательность для инвестиций, обнаружилась тесная корреляция между объемом инвестиций и средним доходом на душу населения, а также количеством выполненных работ в сфере "Строительство". Однако основным фактором, определяющим инвестиционные показатели, являлся средний доход на душу населения. Эти результаты подчеркивают важность учета региональных экономических характеристик при оценке привлекательности для инвестиций и разработке целевых стратегий и политик.

Следует отметить ограничения данного исследования, включая возможное влияние немеряемых факторов на привлекательность инвестиций. Кроме того, необходимо быть осторожным при обобщении полученных результатов на другие контексты из-за специфических характеристик исследуемых регионов. Дальнейшие исследования должны быть направлены на проверку и расширение данных результатов путем включения дополнительных переменных и рассмотрения более широких экономических и социальных факторов, влияющих на инвестиционные решения.

С учетом всего вышесказанного, можно сделать несколько рекомендаций и предложений. Во-первых, политики должны уделять приоритет инвестициям в регионах с высокой привлекательностью, сосредоточившись на приобретении дорогостоящих основных средств. Это может быть достигнуто путем предоставления стимулов и поддержки для бизнеса, заинтересованного в инвестициях в эти регионы.

Во-вторых, в регионах с низкой привлекательностью для инвестиций следует направить усилия на повышение уровня среднего дохода на душу населения. Это может быть достигнуто через реализацию целевых инициатив, направленных на создание рабочих мест, улучшение программ профессиональной подготовки, а также поощрение предпринимательства и инноваций. Увеличение среднего дохода на душу населения вероятно способствует более благоприятной инвестиционной среде.

Кроме того, политики должны следить за деятельностью в строительной отрасли, поскольку она является значимым фактором, влияющим на объем инвестиций в разных кластерах. Выявление потенциальных препятствий или неэффективностей в строительной отрасли и их устранение может способствовать повышению привлекательности инвестиций.

Для будущего развития необходимо проводить дальнейшие исследования для проверки и расширения данных результатов. Это может включать изучение дополнительных переменных, таких как развитие инфраструктуры, доступ к финансированию и государственные политики. Расширение анализа с учетом социальных и культурных факторов также может предоставить ценные практические рекомендации для принятия инвестиционных решений.

Более того, постоянное мониторинг и оценка привлекательности инвестиций играют важную роль в адаптации политик на основе изменяющихся экономических условий и рыночной динамики. Регулярные анализы позволяют выявить новые тенденции и вызовы, что позволяет политикам принимать обоснованные решения и корректировать стратегии соответствующим образом.

Глава 3

Анализ и моделирование индекса качества воздуха с использованием пошаговой регрессии: изучение тенденций и оценка пригодности

В данной главе широко используется метод пошаговой регрессии в качестве основного аналитического подхода для исследования и моделирования факторов, влияющих на индекс качества воздуха (AQI). В дальнейшем проводится тщательная оценка путем всестороннего сравнения моделей, полученных в процессе пошаговой регрессии, на основе информационного критерия Акаике (AIC) с целью определения модели с наименьшим значением AIC. Кроме того, проводится сравнительный анализ между конечной моделью пошаговой регрессии и моделью, выбранной на основе критерия AIC. Такая строгая методология позволяет систематически изучить разнообразные факторы, влияющие на AQI, способствуя более глубокому пониманию сложностей, связанных с этим, и предоставляя ценные практические рекомендации для разработки эффективных стратегий управления качеством воздуха.

Некоторые методы, обсуждаемые в этой главе, были использованы автором в его опубликованных работах [100] и [101].

3.1 Данные

3.1.1 Источник и сбор данных

В данном исследовании данные на все независимые переменные были получены из Национального бюро статистики Китая, что обеспечивает надежность

и доверие к данным. Национальное бюро статистики является важным правительственным органом, ответственным за сбор, систематизацию и распространение широкого спектра макроэкономической и социальной статистики. Эта статистика получена из различных источников, включая правительственные опросы, выборочные опросы и переписи, что обеспечивает высокий уровень достоверности и широкий охват.

С другой стороны, данные зависимой переменной были получены с использованием национальной платформы расчета качества воздуха. Эта платформа является специализированным инструментом, совместно управляемым отделами по охране окружающей среды и соответствующими научно-исследовательскими учреждениями, для мониторинга и оценки качества воздуха по всей территории Китая. С использованием передовых технологий и оборудования для мониторинга, эта платформа получает данные в реальном времени по различным показателям качества воздуха, таким как PM2.5, PM10, диоксид серы, диоксид азота и другие. Эти показатели являются важными параметрами для оценки качества воздуха и способны отражать уровни загрязнения в разных провинциях в определенные периоды времени.

Для обеспечения комплексности набора данных были собраны данные о качестве воздуха за период с 2013 по 2017 года, охватывающие все провинции Китая. Выбранный период в семь лет обеспечивает относительно долгосрочную перспективу, облегчая анализ тенденций и различий в качестве воздуха между разными годами и регионами. Более того, учитывая различия в уровнях развития, плотности населения и промышленной структуры в разных провинциях Китая, сбор данных был проведен на всей территории страны для обеспечения репрезентативности и обобщаемости выборки.

В заключение можно сказать, что в данном исследовании данные независимых переменных получены из Национального бюро статистики Китая, а данные зависимой переменной были рассчитаны с использованием национальной платформы расчета качества воздуха. Для обеспечения всесторонности данные о качестве воздуха собирались с 2013 по 2019 год и охватывали все провинции Китая. Этот тщательный процесс сбора данных гарантирует достоверность, авторитетность и представительность выборки, являясь надежной основой для исследования.

3.1.2 Очистка данных и обработка исключений

Очистка данных и обработка исключений являются необходимыми предварительными этапами перед применением метода множественной логистической регрессии. Этот важный шаг служит для обеспечения качества, надежности и точности данных, тем самым повышая точность и интерпретируемость модели.

Первая задача при очистке данных состоит в обработке пропущенных значений, которые могут нарушить процесс подгонки модели и выводов. Для решения таких ситуаций необходимо использовать эффективные стратегии. Распространенные подходы включают удаление наблюдений или переменных с отсутствующими значениями, использование методов импутации на основе доступной информации или применение специализированных методологий, таких как множественная импутация.

Еще один существенный аспект очистки данных - управление выбросами. Выбросы представляют собой экстремальные значения, значительно отклоняющиеся от остальных наблюдений. Эти аномалии могут оказывать неблагоприятное влияние на модель, приводя к смещенным оценкам параметров и искаженным результатам выводов. В результате становится необходимым выявление и соответствующая обработка выбросов. Установленные методы для этой цели включают обнаружение выбросов на основе статистических правил, нормализацию с использованием диаграмм размаха или Z-оценок, а также продвинутые алгоритмы, такие как Изоляционный лес и LOF.

Помимо обработки пропущенных значений и выбросов, очистка данных включает несколько других аспектов, таких как обработка повторяющихся значений, выполнение трансформации и стандартизации переменных, а также проверка/исправление типов данных. Эти шаги способствуют поддержанию последовательности, сопоставимости и пригодности данных для анализа.

Эффективное управление пропущенными значениями, выбросами и другими проблемами с данными позволяет повысить надежность и точность модели, облегчая более точные выводы и заключения. Важно отметить, что стратегии очистки данных и обработки исключений должны быть адаптированы к конкретным характеристикам набора данных и соответствовать статистическим принципам и предметному знанию, что позволяет минимизировать смещения и ошибки.

3.2 Методология

AIC и критерий AIC

Критерий Акаике (AIC) - широко используемый статистический инструмент для выбора моделей, разработанный Хироцугу Акаике в 1970-х годах. Он приобрел особую популярность благодаря своей способности уравновесить соответствие модели и ее сложность, что делает его подходящим для научных исследований в различных научных дисциплинах.

AIC основан на информационной теории и базируется на принципе минимизации потерь информации при аппроксимации неизвестного процесса порождения данных. Он предоставляет количественную меру для сравнения и оценки различных моделей на основе их хорошего соответствия данным. Учитывая как качество модели, так и ее простоту, AIC предлагает способ выбора наиболее подходящей модели из набора конкурирующих альтернатив.

Критерий AIC может быть математически представлен следующим образом:

$$AIC = 2k - 2\ln(L) \tag{3.1}$$

где AIC представляет собой критерий Акаике, k - количество оцениваемых параметров в модели, а L - максимизированное значение функции правдоподобия, связанной с моделью.

Первая часть формулы AIC, $2k$, накладывает штраф на модели с большим числом параметров. Этот штраф поощряет простые модели, предотвращая излишнее прилипание, когда модель становится слишком сложной и начинает подстраиваться под случайные флуктуации данных, а не под лежащие в их основе закономерности. Включение $2k$ в критерий обеспечивает то, что простые модели предпочтительны, если увеличение сложности значительно не улучшает соответствие данным.

Вторая часть формулы AIC, $2ln(L)$, измеряет хорошее соответствие модели. Функция правдоподобия, L, количественно оценивает, насколько хорошо модель предсказывает наблюдаемые данные. С увеличением правдоподобия уменьшается термин $2ln(L)$, указывая на лучшую согласованность. Модели, близко соответствующие данным, будут иметь более высокие значения правдоподобия и, следовательно, более низкие значения AIC.

Для выбора наилучшей модели с использованием AIC исследователи сравни-

вают значения AIC различных моделей, подогнанных под один и тот же набор данных. Модель с наименьшим значением AIC считается наиболее подходящей для объяснения данных. Эта модель достигает хорошего баланса между точностью и сложностью, обеспечивая надежное представление подлежащих процессов.

AIC стал неотъемлемым инструментом в научных исследованиях, поскольку он предлагает строгую и объективную систему выбора моделей. Направляя исследователей к выбору наиболее подходящей модели, AIC способствует экономии ресурсов и обеспечивает выбор модели, которая не является чрезмерно сложной, что уменьшает риск избыточного прилипания и улучшает обобщаемость. Его широкое использование в различных дисциплинах, включая статистику, эконометрику, экологию и социальные науки, демонстрирует его универсальность и практическую ценность в научных исследованиях.

В заключение, критерий Акаике (AIC) предоставляет ученым количественную меру для выбора моделей с наилучшим соответствием, учитывая их сложность. Благодаря своей математической формулировке и основанному на информационной теории принципу, AIC предлагает надежный и объективный подход к выбору моделей. Сбалансированный учет качества соответствия данных и простоты модели повышает достоверность и точность научных исследований, облегчая выбор моделей, которые эффективно отражают основные закономерности наблюдаемых данных.

3.3 Эмпирический результат и объяснения

Предоставление семи изображений, отображающих состояние загрязнения в различных районах Китая с 2011 по 2017 год, предлагает ценные практические рекомендации относительно тенденций и изменений в качестве воздуха со временем. Анализ данных визуальной информации позволяет представить академическое и логическое исследование данных, выявление закономерностей и выводы о состоянии усилий по охране окружающей среды в Китае.

Во-первых, следует отметить, что половина представленных районов продемонстрировала хорошее качество воздуха на протяжении всех лет, что указывает на то, что значительная часть регионов, городов и провинций Китая успешно поддерживает или улучшает уровень качества воздуха за исследуемый период.

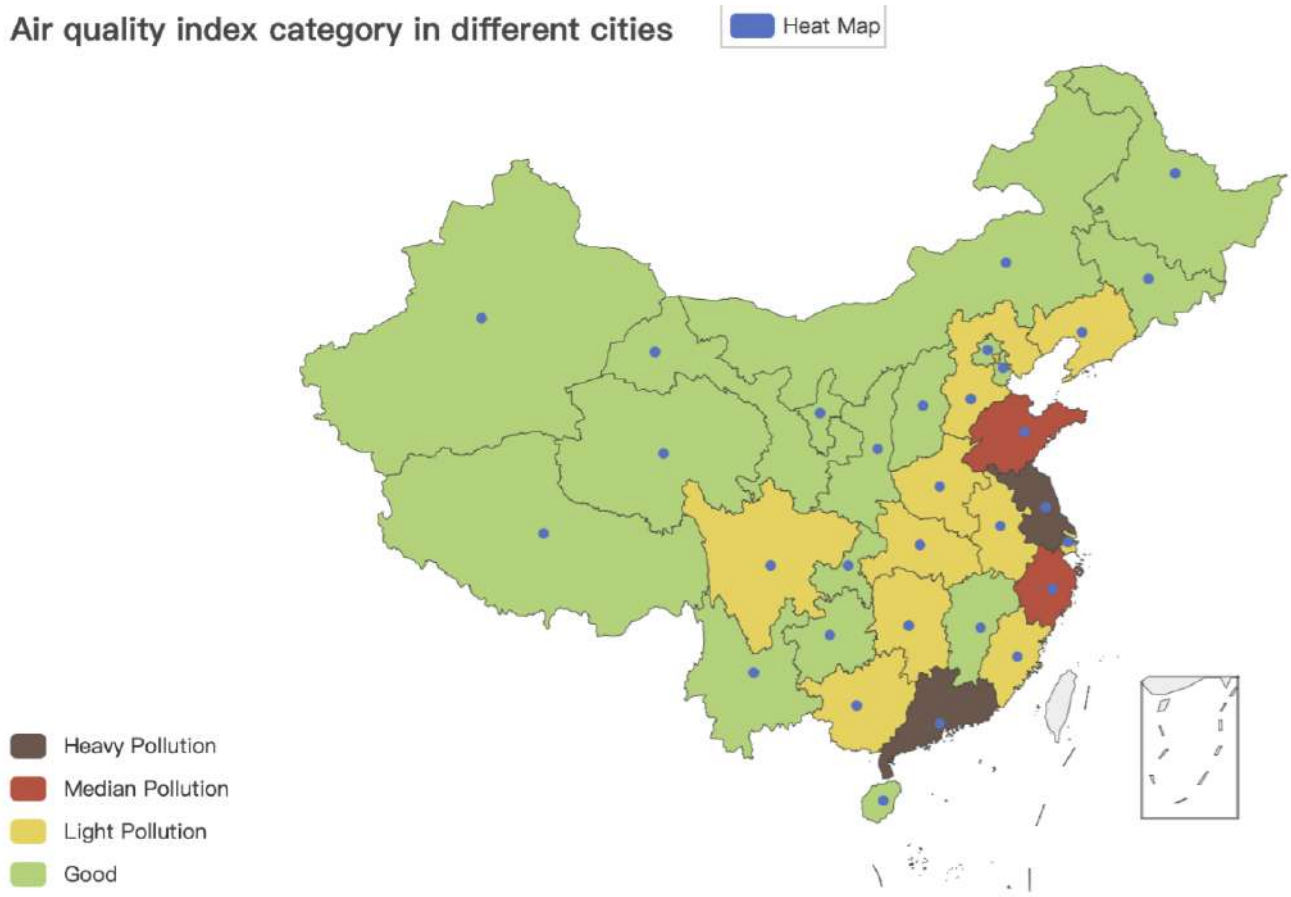


Рис. 3.1: Ситуация с индексом качества воздуха в 31 столице в 2011 году

Это соответствует представлению о том, что больше внимания уделяется охране окружающей среды, что приводит к положительным результатам в области качества воздуха.

Во-вторых, наличие блоков желтого цвета, обозначающих умеренное загрязнение, указывает на районы, где качество воздуха не достигло оптимального уровня, но остается в пределах умеренного диапазона. Хотя это менее желательно, чем зеленый цвет, распространенность желтого цвета свидетельствует о том, что были предприняты меры для снижения загрязнения и улучшения качества воздуха в этих регионах. Это свидетельствует о положительной тенденции в отношении решения проблемы окружающей среды и предпринимаемых шагов по снижению уровней загрязнения.

Кроме того, наличие блоков красного цвета, обозначающих среднюю степень загрязнения, указывает на районы, где качество воздуха не соответствует требуемым стандартам, но все же остается управляемым. Наличие только двух районов с сильным загрязнением в более ранние годы (2011-2016) и только одного района с сильным загрязнением в 2017 году подчеркивает сокращение ко-

Рис. 3.2: Ситуация с индексом качества воздуха в 31 столице в 2012 году

личества районов с высокой степенью загрязнения со временем. Это снижение можно объяснить улучшением мер по охране окружающей среды и повышенным осведомленностью о пагубных последствиях загрязнения, что побуждает власти и общество принимать меры.

Постоянное сокращение районов с сильным загрязнением свидетельствует о прогрессе в борьбе с экологическим ухудшением и подчеркивает эффективность политики в области охраны окружающей среды. Это указывает на то, что меры, направленные на контроль загрязнения и охрану окружающей среды, приводят к наблюдаемому улучшению качества воздуха в Китае.

С академической точки зрения анализ этих изображений и наблюдаемые тенденции могут способствовать более широкому пониманию динамики загрязнения воздуха в Китае. Он предоставляет эмпирические данные в поддержку существующих исследований в области политики в области охраны окружающей среды, их реализации и их влияния на улучшение качества воздуха. Кроме того, он подчеркивает важность продолжения усилий по охране окружающей среды

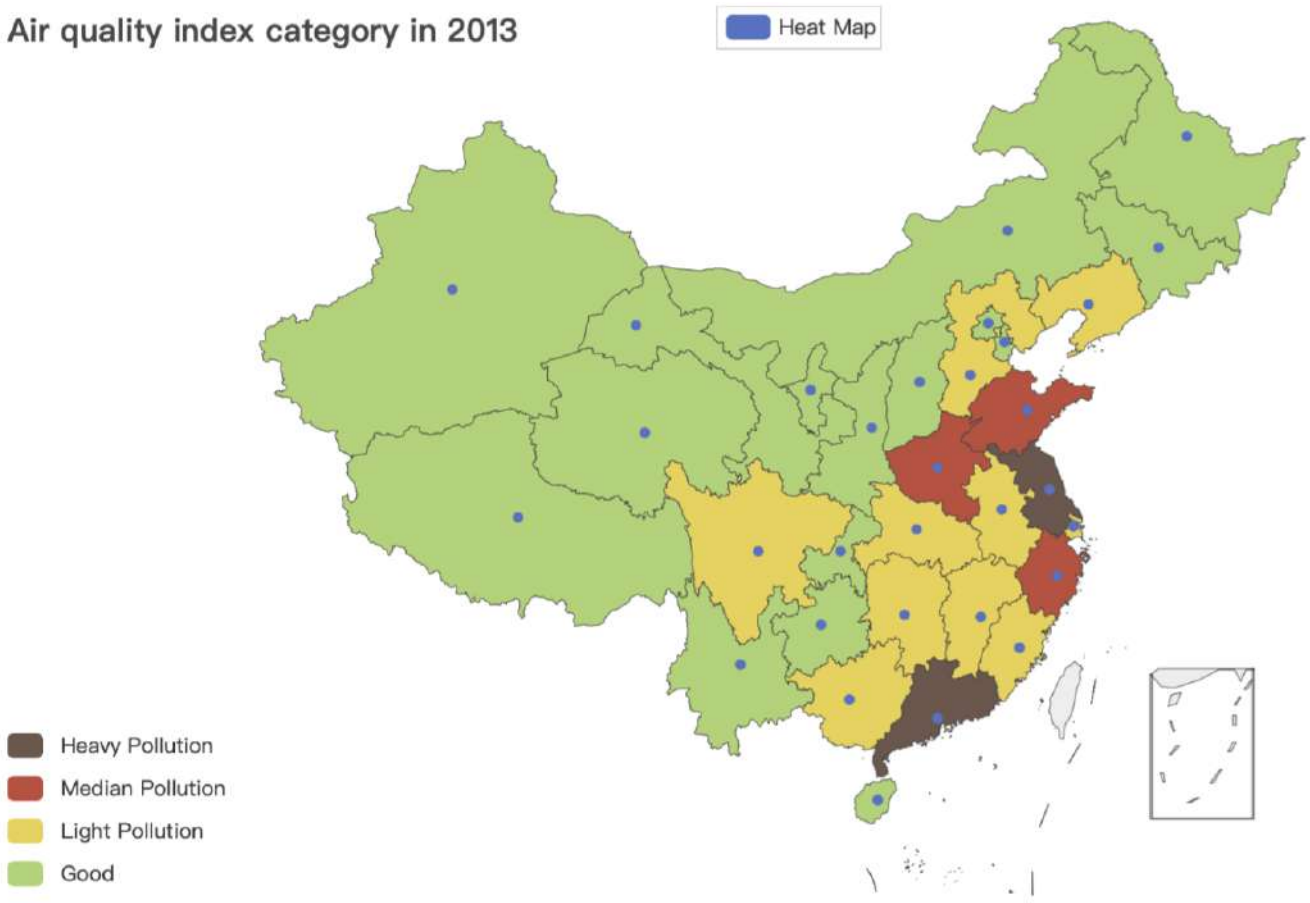


Рис. 3.3: Ситуация с индексом качества воздуха в 31 столице в 2013 году

и устойчивому развитию.

Логически можно сделать вывод, что по мере течения времени все больше внимания уделяется охране окружающей среды в Китае. Уменьшение районов с сильным загрязнением свидетельствует о растущем осознании вредных последствий загрязнения и переходе к использованию более чистых практик и технологий. Этот логический вывод соответствует глобальным тенденциям, подчеркивающим значимость устойчивого развития и экологического управления.

В данном анализе рассматриваются следующие независимые переменные:

- Диоксид серы (SO2)
- Диоксид азота (NO2)
- Частицы с диаметром менее 10 микрометров (PM10)
- Оксид углерода (CO)

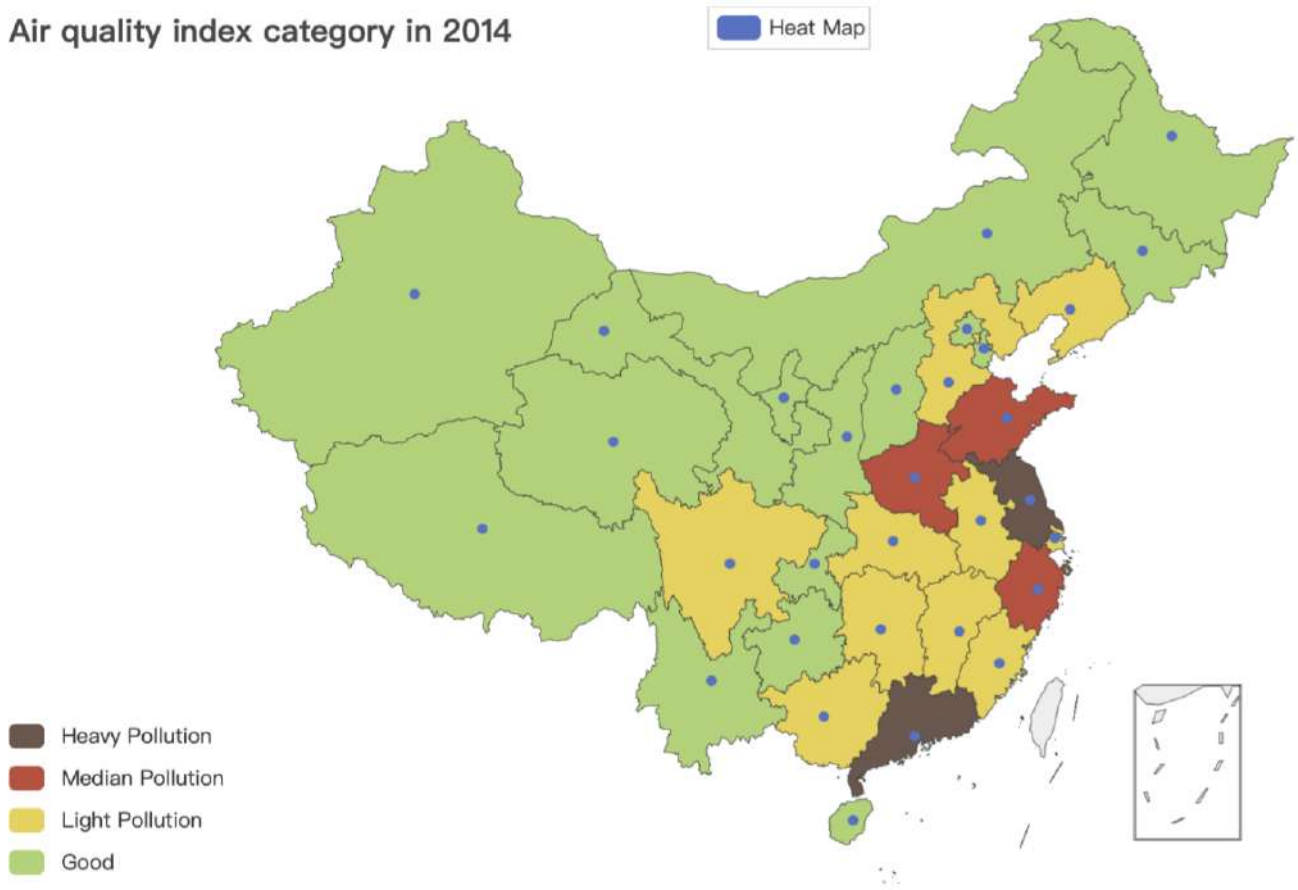


Рис. 3.4: Ситуация с индексом качества воздуха в 31 столице в 2014 году

- Озон (O3)
- Частицы с диаметром менее 2,5 микрометров (PM2.5)
- Температура
- Влажность
- Осадки
- Солнечное сияние

Кроме того, зависимая переменная, на которую обращается внимание, - это индекс качества воздуха (AQI).

Расчет индекса качества воздуха (AQI) включает использование стандартизированных формул для преобразования концентраций загрязняющих веществ в общее значение индекса.

Одно из часто используемых уравнений для расчета AQI имеет следующий вид:

$$AQI = \left(\frac{I_{high} - I_{low}}{C_{high} - C_{low}} \right) \times (C - C_{low}) + I_{low}$$

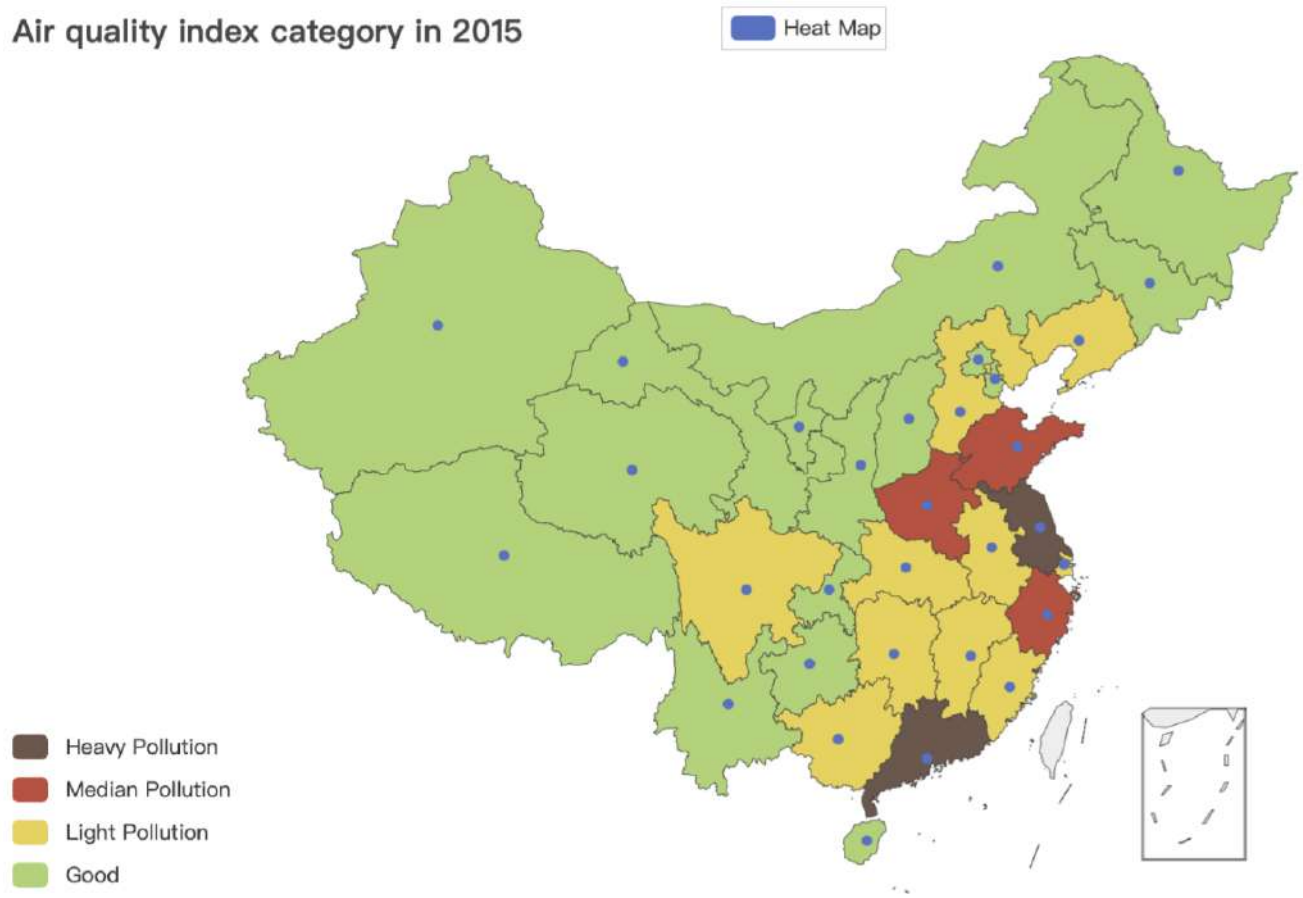


Рис. 3.5: Ситуация с индексом качества воздуха в 31 столице в 2015 году

Где: AQI - рассчитанный индекс качества воздуха. I_{high}и I_{low} - значения индекса, соответствующие верхним и нижним точкам на шкале AQI, соответственно. C_{high} и C_{low} - уровни концентрации, связанные с верхними и нижними точками, соответственно. C - измеренная концентрация загрязняющего вещества. Это уравнение линейно интерполирует между двумя точками для оценки значения AQI на основе измеренного уровня концентрации. Конкретные значения I_{high} , I_{low}, C_{high} и C_{low}определяются соответствующими агентствами по охране окружающей среды и изменяются в зависимости от пороговых значений загрязняющих веществ и категорий качества воздуха, определенных для данного региона.

Модель индекса качества воздуха Китая, примененная в 2017 году

Был проведен анализ линейной регрессии с использованием метода наименьших квадратов (OLS) в языке программирования Python с целью изучения взаимосвязи между зависимой переменной ln(AQI) и объясняющими переменными, включающими загрязнители и метеорологические факторы. Модель использо-

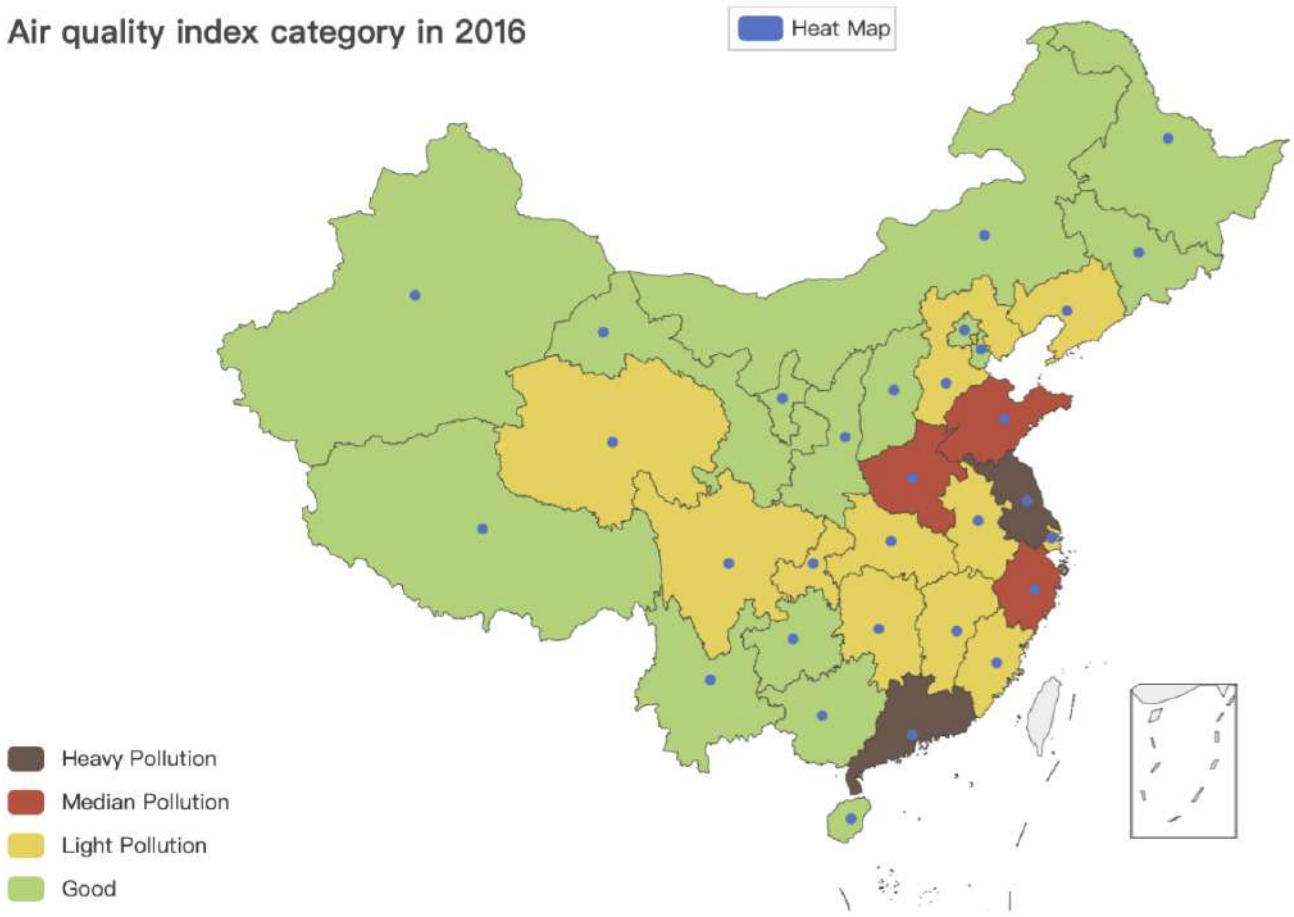


Рис. 3.6: Ситуация с индексом качества воздуха в 31 столице в 2016 году

вала постоянный член для учета какого-либо врожденного смещения в данных.

После выполнения анализа регрессии были получены оцененные параметры и результаты тестов на значимость. Собранные данные представлены в прилагаемой Таблице 3.1. Анализируя статистическую значимость оцененных коэффициентов, можно получить информацию о влиянии независимых переменных на натуральный логарифм индекса качества воздуха.

Значение p-value является мерой статистической значимости и оценивает вероятность того, что наблюдаемая связь между каждой независимой переменной и индексом качества воздуха является случайной.

Значение p-value для каждой переменной указывает на вероятность наблюдения коэффициента, такого же экстремального или более экстремального, чем оцененный коэффициент, при условии нулевой гипотезы о том, что нет связи между независимой переменной и индексом качества воздуха. Если значение p-value ниже предварительно определенного уровня значимости (обычно 0,05), это указывает на то, что переменная имеет статистически значимую связь с

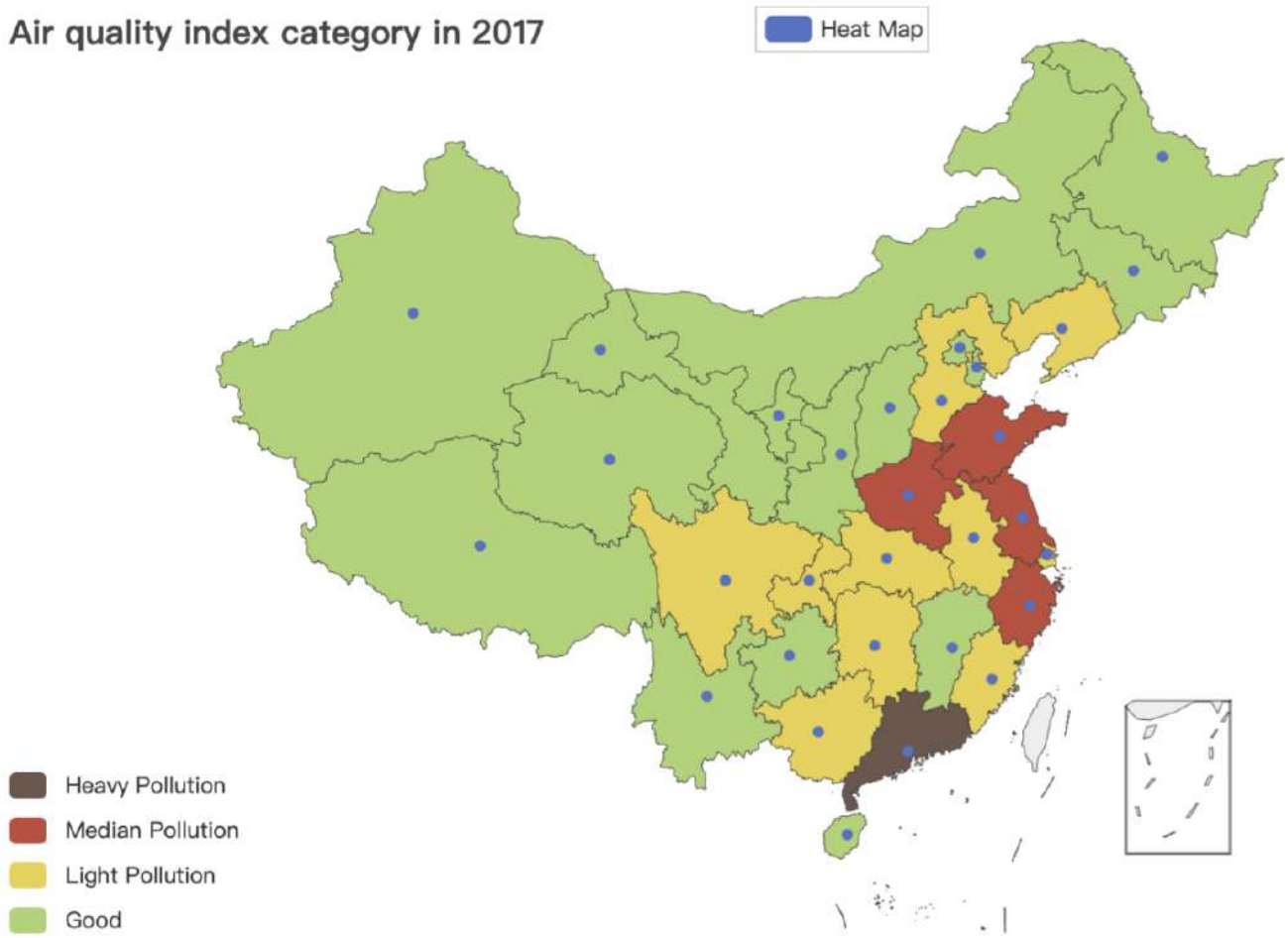


Рис. 3.7: Ситуация с индексом качества воздуха в 31 столице в 2017 году

индексом качества воздуха.

В данном анализе переменная PM10 демонстрирует значимую связь с индексом качества воздуха (p = 0,003). Это означает, что увеличение PM10 связано с более высоким индексом качества воздуха. Другие переменные, такие как SO2, NO2, CO, O3, PM2.5, температура, влажность, осадки и солнечное сияние, не проявляют значимой связи с индексом качества воздуха.

Согласно результатам регрессии в Таблице 3.1, мы можем получить следующую регрессионную модель:

$$\hat{y}_{2017} = 11,650 - 4,771 \cdot x_1 + 25490 \cdot x2 + 42190 \cdot x_3 + 25790 \cdot x_4 - 258.4 \cdot x_5 + 795.37 \cdot x_6 - 25.43 \cdot x_7 + 831.93 \cdot x_8 + 23.31 \cdot x_9 - 9.74 \cdot x_{10} \tag{3.2}$$

Сводка регрессионной модели, представленная в Таблице 3.2, предоставляет ценные практические рекомендации относительно качества и соответствия модели для прогнозирования индекса качества воздуха.

Модель демонстрирует сильное соответствие данным, как указано высоким

Таблица 3.1: Результаты регрессии OLS

Variable	Coefficient	Std. Error	t-value	P-value	[0.025	0.975]
const	11,650	22,000	0.529	0.603(-)	-34,300	57,600
SO2	-4,771	2,974	-1.604	0.124(-)	-11,000	1,432
NO2	25,490	28,900	0.883	0.388(-)	-34,700	85,700
PM10	42,190	12,300	3.441	0.003(**)	16,600	67,800
CO	25,790	157,000	0.165	0.871(-)	-301,000	353,000
O3	-258.40	128.21	-2.015	0.057(.)	-525.84	9.05
PM2.5	795.37	1,333.77	0.596	0.558 (-)	-1,986.83	3,577.57
temp	-25.43	17.90	-1.421	0.171(-)	-62.76	11.90
humidity	831.93	541.29	1.537	0.140(-)	-297.19	1,961.05
precipitation	23.31	37.25	0.626	0.538(-)	-54.39	101.01
sunshine	-9.74	8.44	-1.154	0.262(-)	-27.34	7.87
Sig. codes	0 '***'	0.001 ' **'	0.01 '*'	0.05 '.'	0.1 '-'	

Таблица 3.2: Сводка регрессионной модели

Model Information				
Omnibus	Durbin-Watson	Prob(Omnibus)	Jarque-Bera (JB)	Skew
1.350	2.197	0.509	1.012	0.157
Prob(JB)	Kurtosis	Residual Normality Test p-value	Homoscedasticity Test p-value	
0.603	2.172	0.5091407387537017	0.50365152181532396	
Model Fit Statistics				
R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Log-Likelihood
0.948	0.921	36.14	1.30e-10	-373.75
No. Observations	AIC	Df Residuals	Df Model	BIC
31	769.5	20	10	785.3

значением коэффициента детерминации (R-квадрат) 0.948. Это свидетельствует о том, что около 94.8% изменчивости индекса качества воздуха может быть объяснено предикторными переменными, включенными в модель. Значение скорректированного R-квадрата 0.921 подтверждает, что модель достаточно учитывает степени свободы, использованные в анализе.

F-статистика равная 36.14 с очень низким значением p-value 1.30e-10 указывает на то, что регрессия в целом является статистически значимой. Это означает, что по крайней мере одна из предикторных переменных имеет значительное влияние на индекс качества воздуха.

Перейдя к диагностическим тестам, статистика общности (Omnibus) проверяет общую нормальность остатков. С значением Omnibus равным 1.350 и со-

ответствующим p-value равным 0.509, нет значимых доказательств в пользу отклонения остатков от нормального распределения. Таким образом, предположение о нормальности остатков в целом соблюдается.

Статистика Дарбина-Уотсона (Durbin-Watson) исследует наличие автокорреляции в остатках. В данном случае значение 2.197 свидетельствует о отсутствии значительных проблем с автокорреляцией. Это означает, что остатки являются независимыми и не имеют систематических паттернов.

Статистика Жарка-Бера (JB) и коэффициент асимметрии измеряют асимметрию и эксцесс остатков. С JB значением 1.012 и асимметрией 0.157 нет серьезных доказательств отклонения от предположений о нормальном распределении. Кроме того, p-value 0.603 дополнительно подтверждает адекватность предположения о нормальности.

Тест гомоскедастичности оценивает, имеют ли остатки постоянную дисперсию по всему диапазону предикторных переменных. Значение p-value 0.503 указывает на отсутствие значительного нарушения предположения о гомоскедастичности, что свидетельствует о том, что остатки модели имеют относительно постоянную дисперсию.

Вообще, диагностические тесты показывают, что модель достаточно удовлетворяет критическим предположениям в анализе регрессии, включая нормальность остатков, отсутствие автокорреляции и постоянную дисперсию.

Согласно результатам регрессии OLS, представленным в Таблице 3.1, переменная "CO"демонстрирует наибольшее значение p-value (0.871) среди всех предикторных переменных. В соответствии с методом последовательной регрессии, который стремится выбрать наиболее влиятельные переменные, логично удалить переменную "CO"из модели первой.

Это решение соответствует принципу выбора признаков в анализе регрессии, где переменные, считающиеся имеющими незначительное или отсутствующее влияние на зависимую переменную, постепенно исключаются. Исключая переменную "CO"с высоким значением p-value, мы упрощаем модель и фокусируемся на более значимых предикторах, которые проявляют более сильные связи с индексом качества воздуха.

Для обеспечения значимости всех переменных мы использовали метод последовательной регрессии с обратным исключением. На каждом шаге систематически удалялась наименее значимая переменная на основе t-теста Стьюдента.

Таблица 3.3: Динамика показателей качества воздуха во время выбора объясняющих переменных

Variable	R^2	R^2_{adj}	F-value	p-value(F)	AIC	p(RNT)	p(HT)	DW
-CO	0.947	0.925	42.10	1.9391×10^{-11}	767.6	0.5864	0.5134	2.203
-precip.	0.947	0.927	48.68	3.1992×10^{-12}	766.1	0.6111	0.6635	2.217
-PM2.5	0.946	0.930	57.54	4.4795×10^{-13}	764.4	0.6911	0.7104	2.198
-NO2	0.945	0.931	68.62	6.3971×10^{-14}	763.0	0.9670	0.7930	2.204
-temp	0.941	0.929	79.71	1.5308×10^{-14}	763.2	0.9473	0.7216	2.328
-hum.	0.941	0.931	102.9	1.5139×10^{-15}	761.4	0.8427	0.7183	2.384

Нашей целью было максимизировать значение F-статистики, сохраняя при этом статистическую значимость.

В таблице 3.3 мы наблюдаем последовательное увеличение значения F статистики на каждом шаге. Удалением переменной, связанной с наибольшим p-value от t-теста Стьюдента, мы смогли достичь максимального значения F-статистики. Контрольные переменные, имевшие относительно меньшую степень значимости, удалялись на каждом этапе в соответствии с определенным критерием.

На протяжении всего анализа мы проводили тесты на автокорреляцию, нормальность остатков и гомоскедастичность на каждом шаге. Эти диагностические тесты позволили нам оценить предположения модели и оценить ее корректность.

Результирующая конечная модель для данного года имеет следующий вид:

$$\hat{y}2017 = -203.1495 - -2791.0443 \cdot x_1 + 52210 \cdot x_3 - 198.3622 \cdot x_5 - 16.0762 \cdot x10 \quad (3.3)$$

Применяя метод последовательной регрессии с обратным исключением и проводя всеобъемлющие диагностические тесты, мы убедились, что все переменные, включенные в конечную модель, являются статистически значимыми.

Исходя из таблицы, можно сделать логичный вывод, что модель, полученная с использованием метода последовательной регрессии, который стремится выбрать наиболее значимые объясняющие переменные, совпадает с конечной моделью, имея наименьшее значение AIC. Критерий Акаике (AIC) является широко принятым мерилом для оценки компромисса между точностью соответствия модели данным и ее сложностью. Таким образом, определение модели

с минимальным значением AIC через последовательную регрессию свидетельствует о ее превосходной производительности в достижении оптимального баланса между этими факторами. Это научно значимое положение подчеркивает эффективность метода последовательной регрессии в получении конечной модели с наиболее выгодным значением AIC.

Применение модели индекса качества воздуха Китая в 2011-2016 годах

Таблица 3.4: Результаты моделей после последовательной регрессии в 2011-2016 годах

Time	Elimination Variable	R^2	R^2_{adj}	F-value	p-value(F)	AIC	p(RNT)	p(HT)	DW
2011	**3,4,5,6,10,8,7**	0.913	0.904	94.69	1.91×10^{-14}	767.3	0.2400	0.5812	1.561
2012	**7,10,5,6,3,4,8**	0.908	0.898	89.02	4.07×10^{-14}	771.5	0.2384	0.3255	1.393
2013	**6,9,8,10,3,5,4**	0.913	0.904	57.54	1.87×10^{-13}	770.9	0.5382	0.5705	1.586
2014	**6,4,5,8,10,3,7**	0.931	0.923	121.2	8.89×10^{-16}	767.0	0.1173	0.2471	1.864
2015	**6,7,5,8,3,4,10**	0.905	0.895	86.06	6.15×10^{-14}	777.3	0.5218	0.5907	1.855
2016	**5,10,7,8,4,6,9**	0.869	0.855	59.92	4.63×10^{-12}	785.7	0.4287	0.3906	1.885

Таблица 3.4 представляет результаты последовательной регрессии, проведенной для каждого года, что привело к определению конечных моделей. Столбец "Удаление переменной"указывает порядок удаления переменных в процессе последовательной регрессии на основе их незначимости.

Во-первых, изучая значения R^2 и R^2_{adj}, можно заметить, что конечные модели обладают высокой степенью пригодности для всех лет, варьирующейся от 0,869 до 0,931. Это указывает на то, что значительная доля изменчивости зависимой переменной может быть объяснена выбранным набором независимых переменных.

Во-вторых, F-значение и соответствующее ему значение p-value (F) предоставляют доказательства в пользу общей статистической значимости конечных моделей. Особенно отметим, что F-значения варьируются от 57,54 до 121,2, с соответствующими значениями p-value (F) от 8.89×10^{-16} до 4.63×10^{-12}. Эти небольшие значения p-value указывают на высокую статистическую значимость конечных моделей, дополнительно подтверждая их адекватность в объяснении зависимой переменной.

Критерий Акаике (AIC) является важным критерием для выбора моделей, балансирующим точность соответствия модели и ее сложность. В этом контек-

сте следует отметить, что значения AIC конечных моделей варьируются от 767,0 до 785,7. Модель с наименьшим значением AIC, т.е. 767,0, получается для года 2014. Это свидетельствует о том, что модель, основанная на последовательном исключении переменных в определенном порядке, эффективно достигает оптимального компромисса между производительностью модели и ее простотой для этого конкретного года.

Дополнительно, изучение статистики Дарбина-Уотсона (DW) предоставляет информацию о наличии автокорреляции в остатках моделей. Значения DW варьируются от 1,393 до 1,885, что указывает на отсутствие значительной автокорреляции в остатках конечных моделей.

В заключение, анализ Таблицы 3.4 показывает, что метод последовательной регрессии последовательно выявляет конечные модели с высокой степенью пригодности и статистической значимостью на протяжении нескольких лет. Учитывая значения AIC, порядок удаления переменных и отсутствие автокорреляции, эти модели обеспечивают надежное представление о взаимосвязи между зависимыми и независимыми переменными.

Конечные модели, полученные с помощью последовательной регрессии для каждого года, показывают, что все включенные переменные являются статистически значимыми. Эти модели имеют различные формы для разных лет, что указывает на наличие уникальных связей между зависимыми и независимыми переменными в каждом временном периоде.

В 2011 году была получена конечная модель (AIC = 767.3):

$$\hat{y_{2011}} = -1.471 \cdot 10^4 - 1856.64 \cdot x_1 + 4.728 \cdot 10^4 \cdot x_2 - 17.04 \cdot x_9 \qquad (3.4)$$

В 2011 году была получена модель с наименьшим значением критерия Акаике (AIC = 767):

$$\hat{y_{2011}} = -1.924 \cdot 10^4 - 1746.56 \cdot x_1 + 4.609 \cdot 10^4 \cdot x_2 + 3.693 \cdot x_7 - 28.199 \cdot x_9 \quad (3.5)$$

В 2012 году была получена конечная модель(AIC = 771.5):

$$\hat{y_{2012}} = -2.068 \cdot 10^4 - 1740.56 \cdot x_1 + 4.728 \cdot 10^4 \cdot x_2 - 17.63 \cdot x_9 \qquad (3.6)$$

В 2012 году была получена модель с наименьшим значением критерия Акаике (AIC = 769.7) :

$$\hat{y_{2012}} = -1.673 \cdot 10^4 - 1891.88 \cdot x_1 + 5.052 \cdot 10^4 \cdot x_2 - 447.746 \cdot x_8 - 11.456 \cdot x_9 \quad (3.7)$$

В 2013 году была получена конечная модель, одновременно имеющая наименьшее значение критерия Акаике(AIC = 770.9):

$$\hat{y_{2013}} = -1.519 \cdot 10^4 - 1855.32 \cdot x_1 + 5.101 \cdot 10^4 \cdot x_2 - 9.734 \cdot x_7 \qquad (3.8)$$

В 2014 году была получена конечная модель, одновременно имеющая наименьшее значение критерия Акаике (AIC = 767):

$$\hat{y_{2014}} = -3.82 \cdot 10^4 - 1746.39 \cdot x_1 + 5.431 \cdot 10^4 \cdot x_2 - 42.678 \cdot x_9 \qquad (3.9)$$

В 2015 году была получена конечная модель, одновременно имеющая наименьшее значение критерия Акаике(AIC = 777.3):

$$\hat{y_{2015}} = -1.959 \cdot 10^4 - 1706.58 \cdot x_1 + 5.473 \cdot 10^4 \cdot x_2 - 52.269 \cdot x_9 \qquad (3.10)$$

В 2016 году была получена конечная модель, одновременно имеющая наименьшее значение критерия Акаике(AIC = 785.7):

$$\hat{y_{2016}} = 1.925 \cdot 10^4 - 6092.73 \cdot x_1 + 1.034 \cdot 10^5 \cdot x_2 - 7377.13 \cdot x_3 \qquad (3.11)$$

Эмпирический результат

Таблица 3.5: Результаты конечных моделей с 2011 по 2017 годы.

Time	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2011	*	*	-	-	-	-	-	-	*	-
2012	*	*	-	-	-	-	-	*	*	-
2013	*	*	-	-	-	-	*	-	-	-
2014	*	*	-	-	-	-	-	-	*	-
2015	*	*	-	-	-	-	-	-	*	-
2016	*	*	*	-	-	-	-	-	-	-
2017	*	-	*	-	*	-	-	-	-	*
Count	7	6	2	0	1	0	1	1	4	1

При статистическом анализе индекса качества воздуха в Китае была построена множественная регрессионная модель с использованием метода последовательного исключения на основе данных с 2011 по 2017 годы. Целью модели являлось исследование взаимосвязи между индексом качества воздуха, который служил целевой переменной, и различными экологическими факторами, а

именно: SO2 (x_1), NO2 (x_2), PM10 (x_3), CO (x_4), O3 (x_5), PM2.5 (x_6), Температура (x_7), Влажность (x_8), Осадки (x_9) и Солнечное сияние (x_{10}).

Таблица "Результаты конечных моделей с 2011 по 2017 годы"предоставляет ценную информацию о регрессионных моделях, разработанных для каждого года. Каждая строка представляет конкретный год, а столбцы соответствуют вышеупомянутым переменным. Наличие звездочки (*) в ячейке указывает на включение этой переменной в конечную регрессионную модель для этого конкретного года. Строка "Количество"внизу таблицы представляет количество раз, когда каждый фактор был включен в регрессоры конечных моделей, построенных с 2011 по 2017 годы.

Из анализа таблицы видно, что фактор x_1(SO2) присутствует во всех моделях на протяжении всех лет, демонстрируя его последовательное влияние на индекс качества воздуха с 2011 по 2017 годы. Напротив, факторы x_4 (CO) и x_6(PM2.5) постоянно отсутствуют в уравнениях регрессии, указывая на их относительно минимальное влияние на индекс качества воздуха в течение этого периода.

Более того, переменная x_2(NO2) часто включается в регрессионные модели, что свидетельствует о ее значимости среди регрессоров. Аналогично, фактор x_9(Осадки) проявляет заметную релевантность в конечных моделях.

3.4 Заключение главы 3

В данной главе в основном использовалась методика последовательной регрессии для построения моделей индекса качества воздуха. Применяя этот метод при анализе данных индекса качества воздуха, было обнаружено, что выбранная модель постоянно отличается от модели с наименьшим значением AIC, за исключением 2011 и 2012 годов.

Из проведенного эксперимента видно, что переменная x_1(SO2) присутствует во всех моделях на протяжении всех лет, что указывает на ее стабильное влияние на индекс качества воздуха с 2011 по 2017 год. С другой стороны, переменные x_4(CO) и x_6 (PM2.5) постоянно отсутствуют в уравнениях регрессии, что свидетельствует о их относительно незначительном влиянии на индекс качества воздуха за этот период.

Кроме того, переменная x_2(NO2) часто включается в модели регрессии, что указывает на ее значимость среди объясняющих переменных. Аналогично, пе-

ременная x_9 (Precipitation) проявляет заметную релевантность в конечных моделях.

Произведем краткий анализ на основе фактической ситуации. Значительное и устойчивое влияние SO2 на индекс качества воздуха с 2011 по 2017 годы можно объяснить несколькими факторами. SO2, получаемый в основном из сжигания угля, промышленных процессов и выбросов автотранспорта, был основным загрязнителем в Китае из-за широкого использования угля и недостаточных мер контроля выбросов в последние десятилетия. Долгосрочное воздействие на высокие уровни SO2 представляет серьезные риски для здоровья и способствует образованию кислотного дождя. Несмотря на то, что правительство Китая проводит различные политики и меры для снижения выбросов SO2, сложный характер источников выбросов и накопительный эффект требуют постоянных усилий на протяжении продолжительного времени для достижения значительного улучшения качества воздуха.

С другой стороны, CO и PM2.5 демонстрируют относительно незначительное влияние на индекс качества воздуха. CO, безцветный и беззапаховый токсичный газ, который в основном выбрасывается при сжигании угля, выхлопе автомобилей и промышленных процессах, имеет короткое время жизни в атмосфере, что приводит к быстрому снижению концентрации в зависимости от метеорологических условий. Аналогично, PM2.5, состоящие из мелких частиц, взвешенных в воздухе, включая пыль, дым и выбросы автотранспорта, обладают переменной концентрацией, подверженной различным источникам и метеорологическим условиям. Несмотря на то, что как CO, так и PM2.5 способствуют загрязнению воздуха и негативным воздействиям на здоровье, их влияние может быть менее стабильным или постоянным, чем у SO2.

Кроме того, NO2, оксид азота, образующийся при сжигании угля, выбросах автотранспорта и промышленной деятельности, не оказывает значительного влияния на индекс качества воздуха в 2017 году. Это может быть связано с другими факторами или мерами, которые привели к снижению выбросов или повышению эффективности контроля выбросов в этот конкретный период. Кроме того, концентрация NO2 зависит от метеорологических условий, мест расположения источников выбросов и местных особенностей. В результате некоторые регионы или определенные временные периоды могут испытывать незначительное влияние NO2 на индекс качества воздуха.

В заключение, продолжительное и значительное влияние SO2 на индекс качества воздуха объясняется широкими источниками его выбросов, долгосрочными рисками для здоровья и сложностями в процессе улучшения. С другой стороны, относительно незначительное влияние CO и PM2.5 связано с сложностью источников выбросов, изменчивостью концентраций и трудностями в контроле выбросов. Отсутствие значительного влияния NO2 в 2017 году может быть обусловлено другими факторами, улучшенной эффективностью контроля выбросов и региональными или временными различиями.

Глава 4

Методы глубокого обучения для системы оценки качества воздуха

Большинство результатов, представленных в этой главе, были опубликованы в статьях[102].

4.1 Стандартизированные процедуры использования методов машинного обучения

Стандартизированный процесс использования методов машинного обучения включает несколько важных шагов, среди которых сегментация набора данных и обеспечение согласованности тренировочных данных имеют большое значение. В этом разделе подробно рассматриваются логические этапы и отдельные шаги, связанные с этим процессом.

Сегментация набора данных: Перед тем, как приступить к задачам машинного обучения, необходимо разделить доступный набор данных на отдельные поднаборы: тренировочный набор, набор для проверки и набор для тестирования. Такое разделение облегчает оценку производительности модели и настройку параметров.

- Тренировочный набор: тренировочный набор представляет собой поднабор данных, используемый для обучения модели. Обычно он составляет значительную часть общего набора данных, позволяя модели улавливать соответствующие закономерности и регулярности.
- Набор для проверки: набор для проверки, как другой поднабор данных, используется для выбора гиперпараметров, настройки модели и оценки

ее производительности. Путем оценки и сравнения различных моделей с использованием набора для проверки можно определить оптимальную модель и внести необходимые корректировки.

- Тестовый набор: тестовый набор служит независимым поднабором данных, который используется для окончательной оценки производительности модели. Он должен отличаться от тренировочного и набора для проверки, чтобы точно оценить способность модели к обобщению. Оценка модели на тестовом наборе дает точную оценку ее предсказательной способности в реальных сценариях.

Обеспечение согласованности тренировочных данных: Сохранение согласованности тренировочных данных играет ключевую роль в задачах машинного обучения. В этом процессе важны следующие шаги:

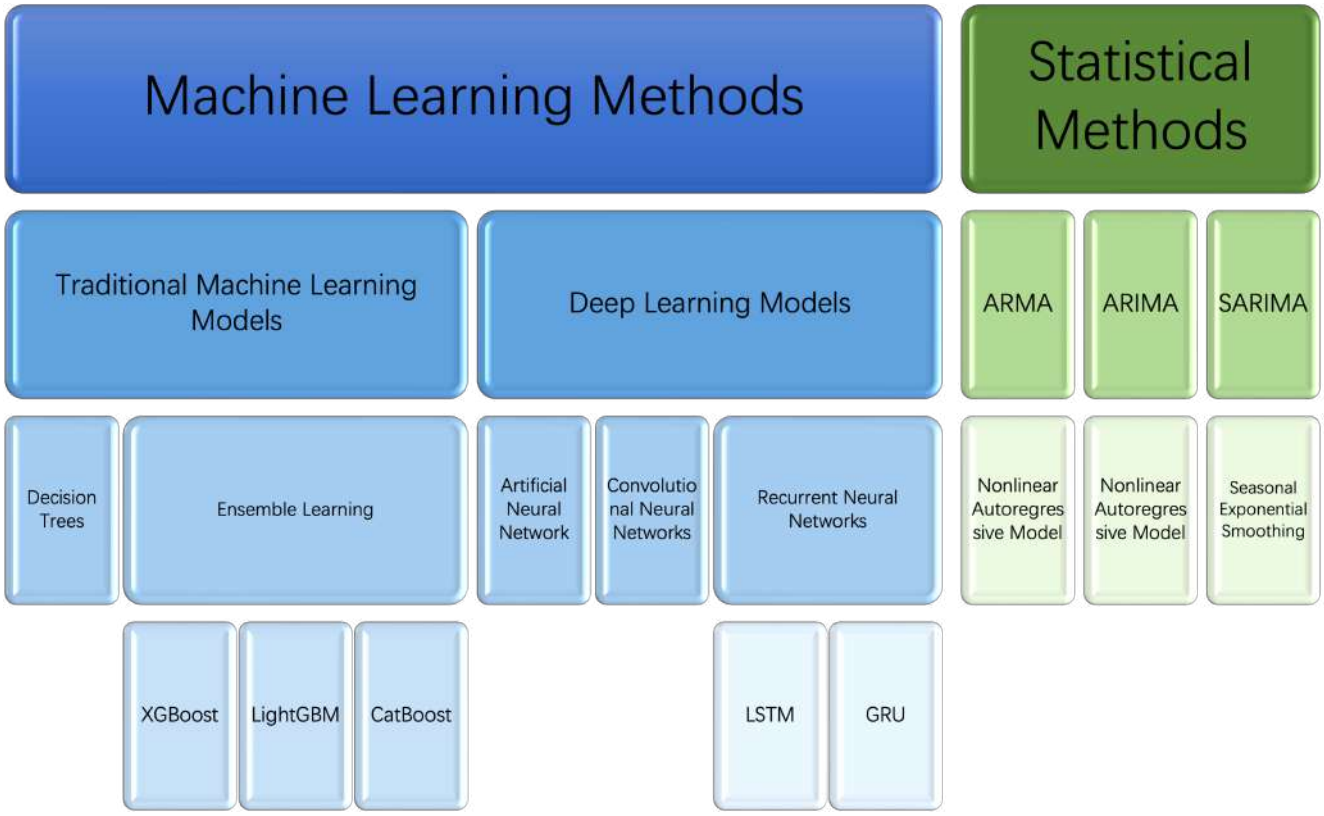
- Нормализация признаков: из-за различных единиц измерения или шкал различные признаки могут требовать нормализации для обеспечения равного обращения модели к ним. Общие методы включают стандартизацию (приведение данных к распределению со средним значением 0 и дисперсией 1) и нормализацию (масштабирование данных в пределах [0, 1]).

- Работа с пропущенными значениями: тщательная проверка и обработка пропущенных значений в тренировочном наборе данных является важным условием успешного обучения модели и достижения точности результатов. Стратегии заполнения пропущенных значений часто включают использование средних, медиан или других подходящих значений признака.

- Работа с выбросами: идентификация и обработка выбросов в тренировочных данных имеет ключевое значение. Выбросы могут негативно повлиять на производительность модели, поэтому их необходимо идентифицировать и обработать с использованием статистических методов или экспертного знания, чтобы они не нарушали процесс обучения и прогнозирования модели.

Путем выполнения сегментации набора данных и обеспечения согласованности тренировочных данных можно построить надежные и эффективные модели

машинного обучения. Эти шаги повышают производительность моделей, их способность к обобщению и устойчивость, предоставляя прочную основу для дальнейшей оценки и настройки. Принятие стандартизированного процесса использования обеспечивает согласованность и предсказательную силу в различных наборах данных, делая модели машинного обучения ценными инструментами поддержки принятия решений в практических приложениях.



ARMA (Авторегрессионная скользящая средняя), ARIMA (Интегрированная авторегрессионная скользящая средняя) и SARIMA (Сезонная АРИМА) - это статистические модели, широко используемые для прогнозирования временных рядов.

Модель ARMA комбинирует авторегрессионные (AR) и скользящие средние (MA) компоненты для моделирования временных рядов. AP-компонент предсказывает текущее значение на основе предыдущих наблюдений, предполагая линейную зависимость между текущим значением и предыдущими значениями. MA-компонент предсказывает текущее значение на основе остаточных ошибок, предполагая линейную зависимость между текущим значением и прошлыми ошибками.

Модель ARIMA расширяет модель ARMA путем включения операций дифференцирования для работы с нестационарными временными рядами. Диф-

ференцирование преобразует исходный временной ряд в стационарный путем вычитания предыдущего наблюдения из текущего. Это помогает более точно улавливать тренды и сезонность в данных.

Модель SARIMA дополнительно расширяет ARIMA путем введения сезонного дифференцирования и сезонных AP и MA компонентов. Она предназначена для работы с данными временных рядов, имеющими явные сезонные закономерности. Сезонное дифференцирование удаляет сезонную составляющую, а сезонные AP и MA компоненты улавливают сезонные зависимости в данных.

Хотя модели ARMA, ARIMA и SARIMA широко используются для прогнозирования временных рядов, существуют сценарии, когда модели машинного обучения могут быть предпочтительными в долгосрочном прогнозировании временных рядов.

Сложные закономерности: Модели машинного обучения, такие как нейронные сети и глубокие модели обучения, могут улавливать сложные нелинейные связи и закономерности в данных, которые могут оказаться не подходящими для традиционных статистических моделей, таких как ARMA, ARIMA и SARIMA. Эти модели могут выявлять сложные временные зависимости и адаптироваться к изменяющейся динамике данных.

Извлечение признаков: Модели машинного обучения могут автоматически извлекать значимые признаки из сырых данных временных рядов, исключая необходимость вручную создавать признаки, требуемые в статистических моделях. Это позволяет моделям выявлять скрытые закономерности и извлекать значимые представления из данных.

Масштабируемость: При прогнозировании временных рядов в долгосрочной перспективе объем и сложность данных могут значительно увеличиться. Модели машинного обучения способны эффективно обрабатывать большие наборы данных, что делает их более подходящими для масштабируемости и работы с высокоразмерными данными по сравнению с традиционными статистическими моделями.

Гибкость: Модели машинного обучения предлагают большую гибкость в моделировании различных типов временных рядов, включая ряды с нелинейными трендами, негауссовыми распределениями или неправильными закономерностями. Они могут адаптироваться к различным характеристикам данных и учитывать различные модельные предположения.

Включение внешних факторов: Модели машинного обучения могут легко учитывать внешние факторы или дополнительные признаки, которые могут влиять на временной ряд, что позволяет повысить предсказательную точность по сравнению с традиционными статистическими моделями.

4.2 Выбор данных

Источник и сбор данных: Сбор и получение данных играют важную роль в контексте прогнозирования временных рядов. В данном случае, набор данных получен из Kaggle - широко известной платформы с открытыми данными, которая славится своими многочисленными наборами данных для исследований и анализа.

Для эффективного прогнозирования временных рядов качество и разнообразие набора данных имеют особое значение. В этом случае, набор данных охватывает значительный период времени, начиная с 1 марта 2013 года и заканчивая 28 февраля 2017 года, что позволяет охватить большой временной промежуток. Такое использование долгосрочных данных является особенно полезным для моделей глубокого обучения, так как они обычно требуют больших объемов данных для обучения и обобщения.

Измерения PM2.5 представляют собой мелкую частицу размером 2,5 микрометра или меньше, что особенно важно для общественного здоровья, так как она может проникать в дыхательную систему и вызывать негативные эффекты. Атмосферное давление, с другой стороны, служит индикатором движения воздушных масс, и его включение в набор данных может помочь исследователям выявить потенциальные корреляции между изменениями давления и уровнями загрязнителей. Аналогично, температура и влажность являются важными метеорологическими факторами, которые могут прямо влиять на рассеяние загрязнителей, химические реакции и атмосферную стабильность, соответственно влияя на уровни качества воздуха.

Учет всех этих различных параметров окружающей среды в анализе позволяет исследователям получить более глубокое представление о динамике качества воздуха. Этот всесторонний набор данных облегчает изучение взаимодействия между различными факторами и их совместного влияния на концентрацию загрязнителей, что в конечном итоге позволяет создавать более точные и надеж-

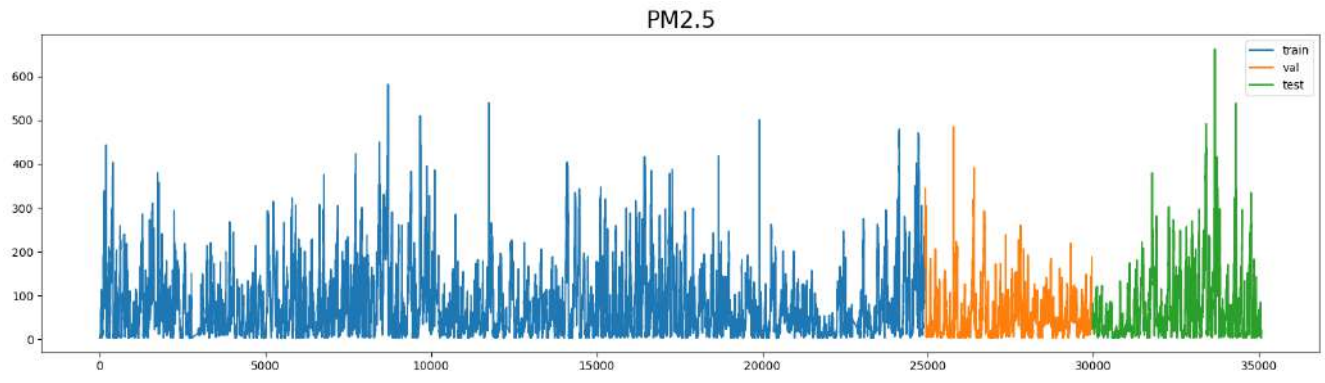
ные модели прогнозирования для эффективного управления качеством воздуха.

Большой объем набора данных позволяет увидеть больше признаков и закономерностей, что приводит к повышению возможностей прогнозирования модели. Большие наборы данных дают возможность моделям глубокого обучения лучше понять сложные временные структуры и зависимости, присутствующие во временных рядах. Кроме того, обширные наборы данных снижают риск переобучения и повышают устойчивость и надежность модели.

Учитывая, что Kaggle является широко используемой платформой с открытыми данными, предлагаемые наборы данных проходят строгую проверку и валидацию, обеспечивая их надежность и воспроизводимость. Это создает доверие среди исследователей и специалистов по обработке данных при выполнении процедур предварительной обработки и моделирования.

Следует иметь в виду, что использование масштабных наборов данных моделями глубокого обучения требует значительных вычислительных ресурсов и времени. Исследователи должны обеспечить достаточную вычислительную мощность и разработать разумное расписание, чтобы удовлетворить требования, связанные с обучением моделей при использовании таких наборов данных.

В данном исследовании конкретный момент времени служит границей между тренировочным и тестовым наборами, а не используется случайный процесс выбора. Такой подход более соответствует внутренней логике временных рядов. Приблизительно 70% элементов выделяются для тренировочного набора. Учитывая различные пропорции пропущенных значений в разных типах данных, мы обеспечиваем достаточное количество признаков при прогнозировании концентрации частиц PM2.5 (Particulate Matter 2.5) как целевой переменной в этом исследовании. Установив разумную границу, пропорции тренировочного, тестового и проверочного наборов данных составляют соответственно 70%, 15% и 15%.



Очистка данных и обработка выбросов: Предварительная обработка данных, включая очистку данных и обработку аномалий, является неотъемлемым этапом при использовании моделей глубокого обучения для прогнозирования временных рядов. Эти шаги играют важную роль в обеспечении точности, согласованности и надежности данных, тем самым повышая производительность и устойчивость модели.

Очистка данных: Первый шаг заключается в решении проблем, таких как пропущенные значения, дублирующие записи и выбросы. Пропущенные значения можно заполнить с использованием методов интерполяции, например, линейных или лагранжевых. Дублирующие записи могут быть удалены непосредственно. Для работы с выбросами необходимо использовать соответствующие методы обнаружения выбросов.

Стабилизация: Еще один важный шаг состоит в преобразовании нестационарных временных рядов в стационарную форму. Нестационарные временные ряды имеют среднее, дисперсию или ковариацию, меняющиеся со временем, что затрудняет моделирование значимых закономерностей и регулярностей. Общие методы стабилизации включают дифференцирование (первого или второго порядка), логарифмические преобразования и другие применимые техники.

Обнаружение и обработка аномалий: Аномалии во временных рядах могут значительно влиять на предсказательную способность модели. Поэтому важно использовать методы обнаружения аномалий для выявления и учета этих выбросов. Методы машинного обучения, такие как изоляционные леса или специальные алгоритмы обнаружения выбросов, могут быть использованы для обнаружения аномалий. После выявления выбросов можно применить различные подходы к их управлению, например, удаление, замена или коррекция в соответствии с конкретным контекстом.

Инженерия признаков: Эффективная инженерия признаков играет ключевую роль в успехе моделей глубокого обучения. Она включает выбор и извлечение значимых признаков из данных временных рядов, что позволяет модели более эффективно улавливать сложные закономерности и тренды.

Стандартизация данных: Для улучшения эффективности обучения и скорости сходимости моделей глубокого обучения становится необходимой стандартизация данных. Методы нормализации, такие как минимаксная нормализация, могут быть применены для масштабирования данных внутри сопоставимого

диапазона, устраняя проблемы, связанные с различными шкалами в разных признаках.

В заключение, проведение всесторонней очистки данных, обработка аномалий, инженерия признаков и стандартизация данных являются важными этапами предварительной обработки данных в моделях глубокого обучения для прогнозирования временных рядов. Эти процедуры обеспечивают целостность и надежность данных, что в конечном итоге способствует точным и надежным прогнозам.

4.3 Основная методология

4.3.1 Искусственная нейронная сеть (ANN)

При использовании искусственной нейронной сети (ANN) для прогнозирования временных рядов может быть использована структура прямого распространения нейронной сети (FNN). FNN состоит из входного слоя, скрытых слоев и выходного слоя, где нейроны каждого слоя взаимосвязаны с весами.

Для задачи прогнозирования временных рядов мы можем использовать исторические наблюдения временного ряда в качестве входных данных, а цель состоит в предсказании будущих значений. Математические выражения для модели ANN имеют следующий вид:

Входной слой для скрытого слоя:

$$h_1 = f(W_{in}x + b_{in})$$

Скрытый слой для выходного слоя:

$$y = g(W_{out}h_1 + b_{out})$$

Здесь x представляет собой входные данные в текущем временном шаге, h_1 - это выход скрытого слоя, W_{in} и b_{in} - это весовые и смещения термины от входного слоя к скрытому слою, а W_{out} и b_{out} - это весовые и смещения термины от скрытого слоя к выходному слою. $f(\cdot)$ и $g(\cdot)$ обозначают функции активации.

В задачах прогнозирования временных рядов обычно используются задачи регрессии, поэтому выходной слой обычно не применяет нелинейное преобразование с использованием функции активации. Вместо этого он напрямую выводит предсказанное значение.

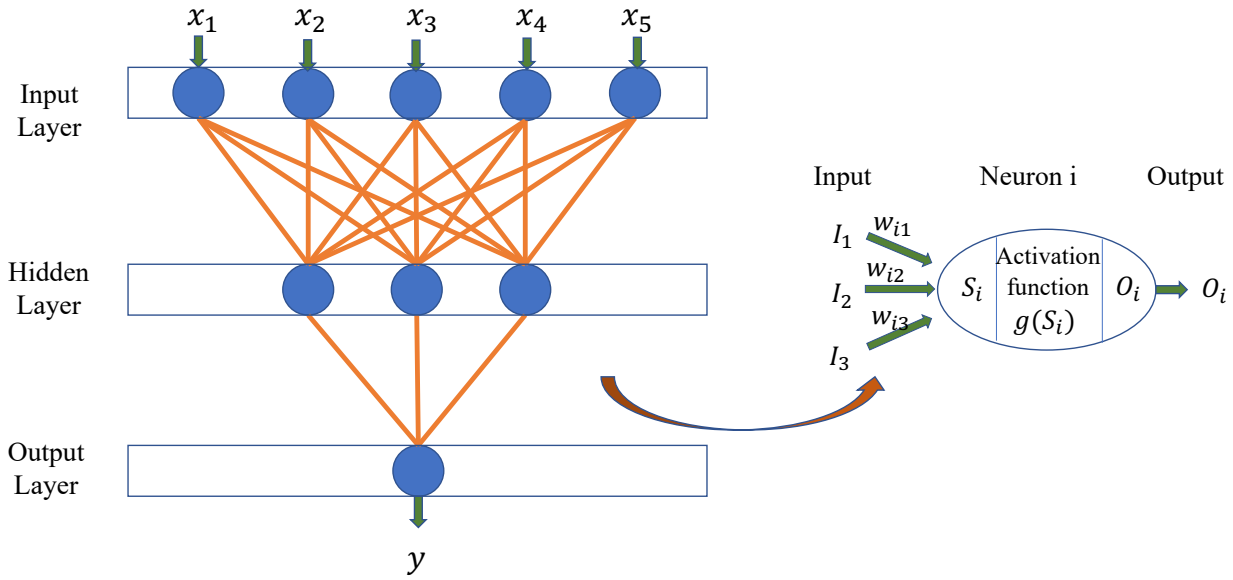


Рис. 4.1: Структура сети искусственной нейронной сети (ANN)

Процесс обучения модели ANN обычно включает минимизацию функции потерь. Часто используемые функции потерь включают среднеквадратичную ошибку (MSE), среднюю абсолютную ошибку (MAE) и т.д. Модель регулирует веса соединений с помощью алгоритма обратного распространения ошибки для минимизации потерь и улучшения точности прогнозирования.

При прогнозировании временных рядов важно учитывать временные корреляции и последовательные зависимости во входных данных. Обычный подход - использовать технику скользящего окна, беря исторические наблюдения в качестве входных признаков и следующее наблюдение на следующем временном шаге в качестве целевого значения для обучения модели прогнозирования.

Правильное проектирование структуры сети, выбор подходящих функций активации и потерь позволяют модели ANN достигать хороших результатов в задачах прогнозирования временных рядов. Техники, такие как регуляризация, нормализация пакетов и т.д., также могут быть включены для улучшения обобщающей способности и стабильности модели.

4.3.2 Рекуррентная нейронная сеть (RNN)

RNN (рекуррентная нейронная сеть) - это тип архитектуры нейронной сети, широко используемой для задач прогнозирования временных рядов. Она осо-

бенно эффективна в улавливании последовательных зависимостей и временных закономерностей в данных. Основная идея RNN состоит в использовании рекуррентных связей, которые позволяют передавать информацию от предыдущих временных шагов к текущим.

Основная математическая формула модели RNN для прогнозирования временных рядов выглядит следующим образом:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

$$y_t = g(W_{hy}h_t + b_y)$$

где x_t представляет собой входные данные на временном шаге h_t - скрытое состояние на временном шаге t, а y_t - выход на временном шаге t. W_{hh}, W_{xh}, W_{hy}, b_h и b_y- это матрицы весов и векторы смещения, которые необходимо изучить во время обучения. $f(\cdot)$ и $g(\cdot)$ представляют функции активации.

В указанных уравнениях h_t вычисляется на основе входа x_t и скрытого состояния с предыдущего временного шага h_{t-1}. Это позволяет модели изучать и улавливать информацию из прошлых наблюдений, которые могут влиять на текущий прогноз. Выход y_t затем генерируется на основе текущего скрытого состояния h_t.

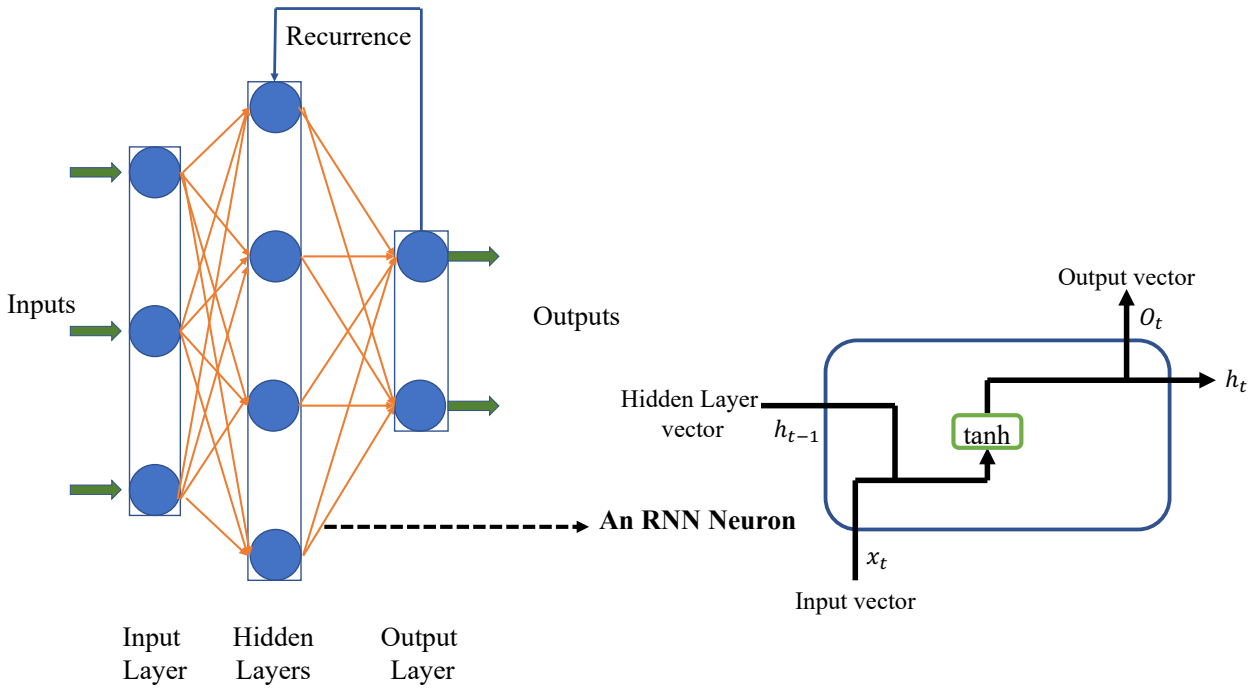


Рис. 4.2: Структура сети традиционной RNN

Во время процесса обучения параметры модели RNN оптимизируются путем минимизации функции потерь, которая сравнивает предсказанные значения с эталонными значениями. Обычно используется алгоритм обратного распространения ошибки во времени (BPTT), чтобы вычислить градиенты функции потерь по отношению к параметрам модели, что позволяет обновлять веса и смещения.

Важно отметить, что на практике были разработаны дополнительные варианты RNN для преодоления проблемы затухания/взрыва градиента и улучшения обучения долгосрочных зависимостей. Некоторые популярные варианты включают LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit). Эти варианты вводят дополнительные механизмы управления, которые регулируют поток информации внутри рекуррентных соединений.

Используя временную динамику, запечатленную моделями RNN, можно эффективно прогнозировать будущие значения во временных рядах данных. Выбор подходящей архитектуры и гиперпараметров зависит от специфических характеристик набора данных и задачи прогнозирования.

4.3.3 Долгая краткосрочная память (LSTM)

LSTM (Long Short-Term Memory) - это вариант рекуррентных нейронных сетей (RNN), широко используемый для задач прогнозирования временных рядов. Модели LSTM специально разработаны для решения проблемы улавливания долгосрочных зависимостей в последовательных данных.

Основным компонентом блока LSTM является память, которая позволяет сети сохранять и обновлять информацию на протяжении нескольких временных шагов. Математическая формула модели LSTM для прогнозирования временных рядов выглядит следующим образом:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

$$y_t = g(W_{hy}h_t + b_y)$$

где x_t представляет собой входные данные на временном шаге t, h_t обозначает скрытое состояние на временном шаге t, c_t - состояние памяти на временном шаге t, а y_t - выходное значение на временном шаге t. W и b представляют собой матрицы весов и векторы смещений, которые необходимо изучить в процессе обучения. $\sigma(\cdot)$ обозначает функцию активации сигмоиды, а $(\odot)$ представляет поэлементное умножение. f_t, i_t и o_t представляют забывающий блок, блок входа и блок вывода соответственно.

В указанных уравнениях модель LSTM вычисляет три блока i_t, f_t и o_t, чтобы контролировать поток информации внутри блока памяти. Забывающий блок f_t определяет, какую информацию следует удалить из предыдущего состояния памяти c_{t-1} на основе текущего входа x_t и предыдущего скрытого состояния h_{t-1}. Блок входа i_t контролирует, какую новую информацию следует сохранить в памяти. Блок вывода o_t регулирует поток информации от памяти к текущему скрытому состоянию. Наконец, скрытое состояние h_t получается путем применения блока вывода к состоянию памяти.

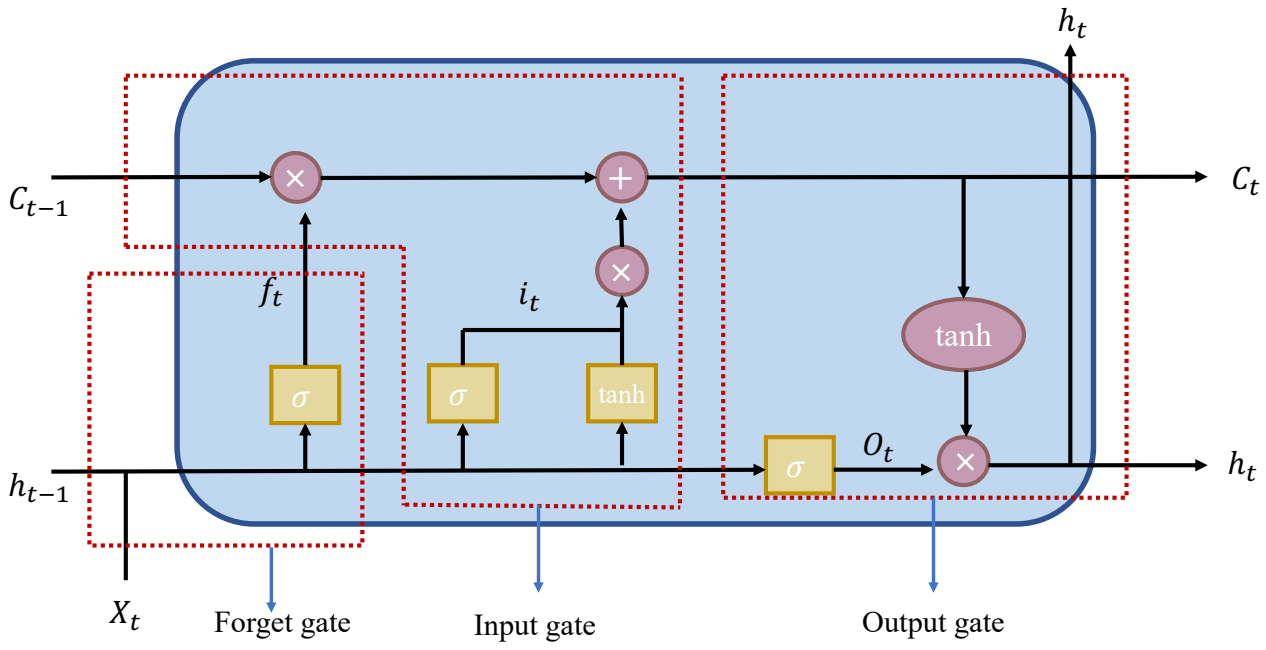


Рис. 4.3: Структура сети LSTM

Во время обучения параметры модели LSTM оптимизируются путем минимизации соответствующей функции потерь, обычно с помощью оптимизационных алгоритмов на основе градиентов. Обратное распространение во времени (BPTT) часто используется для вычисления градиентов функции потерь по

отношению к параметрам модели, что позволяет обновлять веса и смещения.

Модели LSTM доказали свою эффективность в улавливании долгосрочных зависимостей и решении проблемы затухания/взрыва градиента, которые могут возникать в традиционных RNN. Используя блок памяти, модели LSTM могут эффективно улавливать и использовать соответствующую контекстную информацию для точного прогнозирования временных рядов.

4.3.4 Защитный рекуррентный блок(GRU)

GRU (Gated Recurrent Unit) – это еще один вариант рекуррентных нейронных сетей (RNN), который получил популярность в задачах прогнозирования временных рядов. Модели GRU разработаны для улавливания долгосрочных зависимостей и решения проблемы затухания/взрыва градиента, подобно моделям LSTM.

Математическая формула модели GRU для прогнозирования временных рядов выглядит следующим образом:

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)$$
$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)$$
$$n_t = \tanh(W_{xn}x_t + r_t \odot (W_{hn}h_{t-1}) + b_n)$$
$$h_t = (1 - z_t) \odot n_t + z_t \odot h_{t-1}$$
$$y_t = g(W_{hy}h_t + b_y)$$

где x_t представляет собой входные данные на временном шаге t, h_t обозначает скрытое состояние на временном шаге t, а y_t - выходное значение на временном шаге t. W и b представляют матрицы весов и векторы смещений, которые необходимо изучить в процессе обучения. $\sigma(\cdot)$ обозначает функцию активации сигмоиды, а $(\odot)$ представляет поэлементное умножение.

В указанных уравнениях модель GRU вводит два блока: блок обновления z_t и блок сброса r_t. Блок обновления контролирует, сколько предыдущего скрытого состояния h_{t-1} следует сохранить и объединить с текущим кандидатом для скрытого состояния n_t. Блок сброса определяет, сколько предыдущего скрытого состояния h_{t-1} следует забыть при вычислении кандидата для скрытого состояния n_t. Наконец, обновленное скрытое состояние h_t является комбинацией

предыдущего скрытого состояния, взвешенного блоком обновления, и текущего кандидата для скрытого состояния, взвешенного значением $1 - z_t$.

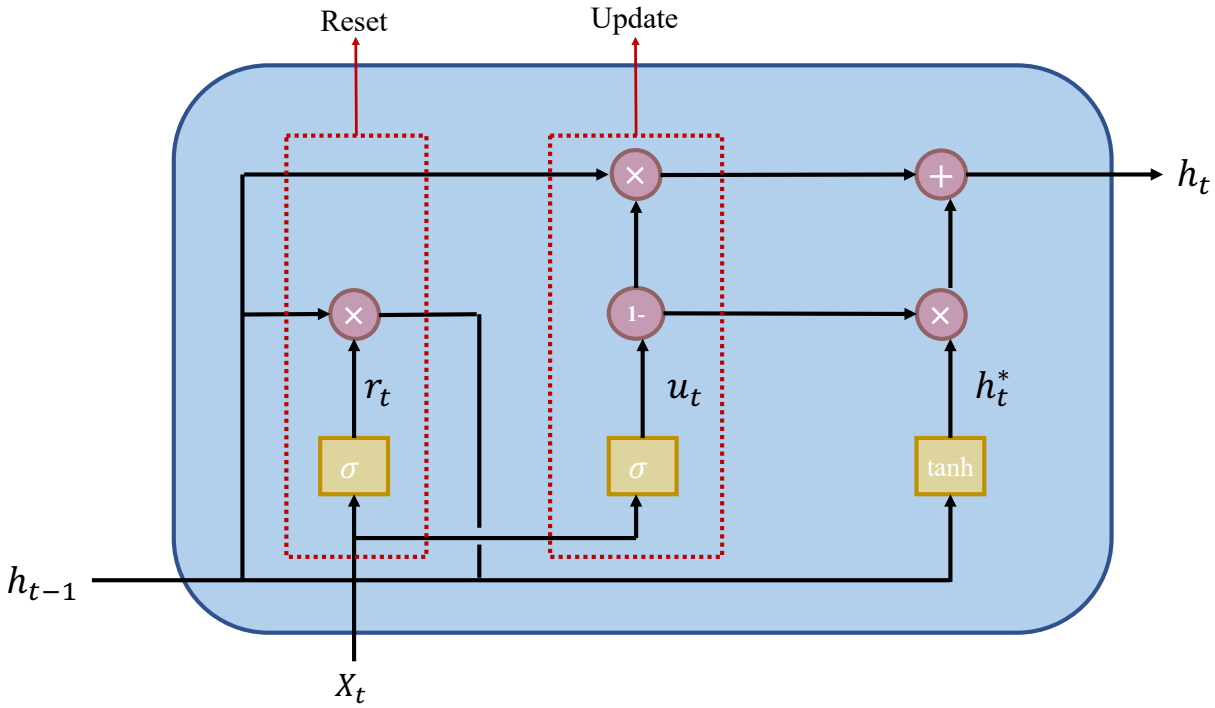


Рис. 4.4: Структура сети GRU

Во время обучения параметры модели GRU оптимизируются путем минимизации соответствующей функции потерь с использованием алгоритмов оптимизации на основе градиентов. Обратное распространение во времени (BPTT) часто используется для вычисления градиентов функции потерь по отношению к параметрам модели, что позволяет обновлять веса и смещения.

Модели GRU предоставляют более простую архитектуру по сравнению с моделями LSTM, сохраняя при этом аналогичную производительность в улавливании долгосрочных зависимостей. Они широко используются в задачах прогнозирования временных рядов и могут эффективно улавливать временные закономерности и зависимости в данных.

4.3.5 Бидирекциональная рекуррентная нейронная сеть (Bi-RNN)

Би-RNN (бидирекциональная рекуррентная нейронная сеть) является вариантом рекуррентной нейронной сети (RNN), который стремится захватить как прошлую, так и будущую информацию во временном ряде для улучшения производительности прогнозирования. Она объединяет две отдельные RNN: одну,

обрабатывающую последовательность в прямом направлении, и другую – в обратном направлении.

Математическая формула модели Bi-RNN для прогнозирования временных рядов выглядит следующим образом:

Прямая RNN:

$$h_t^{\rightarrow} = f(W_{\rightarrow}x_t + U_{\rightarrow}h_{t-1}^{\rightarrow} + b_{\rightarrow})$$

Обратная RNN:

$$h_t^{\leftarrow} = f(W_{\leftarrow}x_t + U_{\leftarrow}h_{t+1}^{\leftarrow} + b_{\leftarrow})$$

где x_t представляет собой входные данные на временном шаге t, $h_t^{\rightarrow}$ обозначает скрытое состояние, вычисленное прямой RNN, $h_t^{\leftarrow}$ - скрытое состояние, вычисленное обратной RNN, а $f(\cdot)$ обозначает функцию активации.

В модели Bi-RNN прямая RNN обрабатывает последовательность от начала до конца, в то время как обратная RNN обрабатывает последовательность в обратном направлении. Каждая RNN имеет свой набор весов и смещений $W_{\rightarrow}$, $U_{\rightarrow}$, $b_{\rightarrow}$ для прямой RNN и $W_{\leftarrow}$, $U_{\leftarrow}$, $b_{\leftarrow}$ для обратной RNN), которые изучаются в процессе обучения.

Для получения окончательного скрытого состояния на каждом временном шаге прямое и обратное скрытые состояния объединяются:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

где [; обозначает конкатенацию.

Окончательный вывод модели Bi-RNN может быть рассчитан следующим образом:

$$y_t = g(W_{hy}h_t + b_y)$$

где W_{hy} и b_y представляют собой матрицу весов и вектор смещений для выходного слоя, а $g(\cdot)$ обозначает функцию активации.

Во время обучения параметры модели Bi-RNN оптимизируются путем минимизации соответствующей функции потерь с использованием алгоритмов оптимизации на основе градиентов. Градиенты вычисляются с помощью обратного распространения во времени (BPTT) для обновления весов и смещений.

Путем учета как прошлой, так и будущей информации во временном ряде модели Bi-RNN могут захватывать более полную временную зависимость

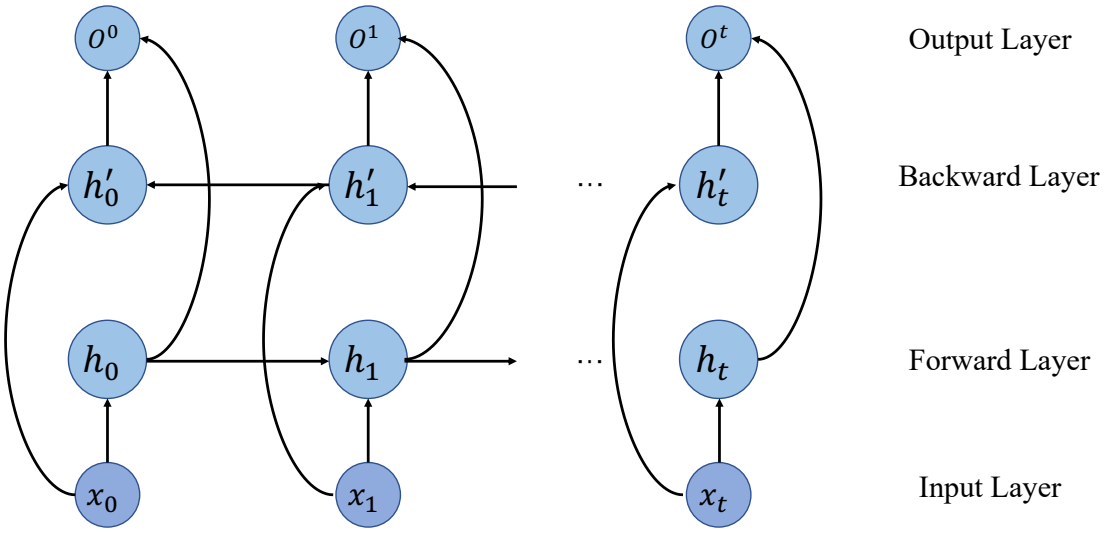


Рис. 4.5: Network structure of Bi-RNN

и паттерны, что приводит к улучшению производительности прогнозирования по сравнению с традиционными RNN. Они успешно применяются в различных задачах прогнозирования временных рядов.

Бидирекциональная долгая краткосрочная память (Bi-LSTM)

Bi-LSTM (бидирекциональная долгая краткосрочная память) является вариантом рекуррентной нейронной сети (RNN), который объединяет преимущества LSTM и бидирекциональных RNN для улавливания как прошлой, так и будущей информации во временном ряде для улучшения производительности прогнозирования.

Математическая формула модели Bi-LSTM для прогнозирования временных рядов выглядит следующим образом:

Прямая LSTM:

$$
\begin{aligned}
i_t^{\rightarrow} &= \sigma(W_{xi}^{\rightarrow} x_t + W_{hi}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{ci}^{\rightarrow} c_{t-1}^{\rightarrow} + b_i^{\rightarrow}) \\
f_t^{\rightarrow} &= \sigma(W_{xf}^{\rightarrow} x_t + W_{hf}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{cf}^{\rightarrow} c_{t-1}^{\rightarrow} + b_f^{\rightarrow}) \\
c_t^{\rightarrow} &= f_t^{\rightarrow} \odot c_{t-1}^{\rightarrow} + i_t^{\rightarrow} \odot \tanh(W_{xc}^{\rightarrow} x_t + W_{hc}^{\rightarrow} h_{t-1}^{\rightarrow} + b_c^{\rightarrow}) \\
o_t^{\rightarrow} &= \sigma(W_{xo}^{\rightarrow} x_t + W_{ho}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{co}^{\rightarrow} c_t^{\rightarrow} + b_o^{\rightarrow}) \\
h_t^{\rightarrow} &= o_t^{\rightarrow} \odot \tanh(c_t^{\rightarrow})
\end{aligned}
$$

Обратная LSTM:

$$
\begin{aligned}
i_t^{\leftarrow} &= \sigma(W_{xi}^{\leftarrow} x_t + W_{hi}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{ci}^{\leftarrow} c_{t+1}^{\leftarrow} + b_i^{\leftarrow}) \\
f_t^{\leftarrow} &= \sigma(W_{xf}^{\leftarrow} x_t + W_{hf}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{cf}^{\leftarrow} c_{t+1}^{\leftarrow} + b_f^{\leftarrow}) \\
c_t^{\leftarrow} &= f_t^{\leftarrow} \odot c_{t+1}^{\leftarrow} + i_t^{\leftarrow} \odot \tanh(W_{xc}^{\leftarrow} x_t + W_{hc}^{\leftarrow} h_{t+1}^{\leftarrow} + b_c^{\leftarrow}) \\
o_t^{\leftarrow} &= \sigma(W_{xo}^{\leftarrow} x_t + W_{ho}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{co}^{\leftarrow} c_t^{\leftarrow} + b_o^{\leftarrow}) \\
h_t^{\leftarrow} &= o_t^{\leftarrow} \odot \tanh(c_t^{\leftarrow})
\end{aligned}
$$

где x_t представляет собой входные данные на временном шаге t, $h_t^{\rightarrow}$) и $h_t^{\leftarrow}$ представляют собой скрытые состояния, вычисленные прямой и обратной LSTM соответственно, $c_t^{\rightarrow}$ и $c_t^{\leftarrow}$ представляют собой состояния памяти, а $\sigma(\cdot)$ обозначает функцию активации сигмоиды.

Для получения окончательного скрытого состояния на каждом временном шаге прямое и обратное скрытые состояния объединяются:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

где [; обозначает конкатенацию.

Окончательный вывод модели Bi-LSTM может быть рассчитан следующим образом:

$$y_t = g(W_{hy} h_t + b_y)$$

где W_{hy} и b_y представляют собой матрицу весов и вектор смещений для выходного слоя, а $g(\cdot)$ обозначает функцию активации.

Во время обучения параметры модели Bi-LSTM оптимизируются путем минимизации соответствующей функции потерь с использованием алгоритмов оп-

тимизации на основе градиентов. Градиенты вычисляются с помощью обратного распространения во времени (BPTT) для обновления весов и смещений. Путем учета как прошлой, так и будущей информации во временном ряде модель Bi-LSTM может более полно улавливать временные зависимости во временном ряде, тем самым улучшая производительность прогнозирования. Модель Bi-LSTM широко используется в различных задачах прогнозирования временных рядов.

4.3.6 Бидирекциональная блокирующая рекуррентная единица (Bi-GRU)

Bi-GRU (бидирекциональная блокирующая рекуррентная единица) является вариантом рекуррентных нейронных сетей (RNN), который объединяет преимущества GRU и бидирекциональных RNN для задач прогнозирования временных рядов. Она стремится захватить как прошлую, так и будущую информацию во временном ряду, что позволяет более полно понять временные зависимости.

Математическая формула модели Bi-GRU для прогнозирования временных рядов выглядит следующим образом:

Прямая GRU: Обратная GRU: где x_t представляет собой входные данные на временном шаге t, $h_t^{\rightarrow}$ и $h_t^{\leftarrow}$ обозначают скрытые состояния, вычисленные прямой и обратной GRU соответственно. $z_t^{\rightarrow}$ и $z_t^{\leftarrow}$ - это блоки обновления, которые контролируют поток информации от предыдущего скрытого состояния к текущему скрытому состоянию. $r_t^{\rightarrow}$ и $r_t^{\leftarrow}$ - это блоки сброса, которые определяют, сколько предыдущего скрытого состояния следует забыть при вычислении кандидатского скрытого состояния. $n_t^{\rightarrow}$ и $n_t^{\leftarrow}$ представляют кандидатские скрытые состояния, а $\sigma(\cdot)$ обозначает функцию активации сигмоиды. Обновленные скрытые состояния получаются путем объединения кандидатских скрытых состояний с предыдущими скрытыми состояниями, взвешенными блоками обновления.

Для получения окончательного скрытого состояния на каждом временном шаге прямые и обратные скрытые состояния конкатенируются:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

где [;обозначает конкатенацию.

Окончательный вывод модели Bi-GRU может быть рассчитан следующим образом:

$$y_t = g(W_{hy}h_t + b_y)$$

где W_{hy} и b_y представляют собой матрицу весов и вектор смещений для выходного слоя, а $g(\cdot)$ обозначает функцию активации.

Во время обучения параметры модели Bi-GRU оптимизируются путем минимизации соответствующей функции потерь с использованием алгоритмов оптимизации на основе градиентов. Градиенты вычисляются с помощью обратного распространения во времени (BPTT) для обновления весов и смещений.

Модели Bi-GRU используют как прошлую, так и будущую информацию во временном ряду, что позволяет более полно улавливать временные зависимости и повышать точность прогнозирования по сравнению с традиционными RNN. Они широко применяются в различных задачах прогнозирования временных рядов.

4.4 Прогнозные модели в приложениях временных рядов: результаты моделирования

4.4.1 Целевая функция

В глубинном обучении среднеквадратичная ошибка (Mean Squared Error, MSE) и коэффициент детерминации (R^2) часто используются как показатели для оценки производительности моделей регрессии. Эти метрики предоставляют ценную информацию о точности и качестве соответствия прогнозов модели.

MSE – это мера близости предсказанных значений к фактическим значениям. Она вычисляет среднее квадратическое отклонение между предсказанными значениями $\hat{y}$ и истинными значениями y в наборе данных. Формула для MSE выглядит следующим образом:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Здесь n представляет количество выборок в наборе данных, $\hat{y}_i$ обозначает предсказанное значение для i-й выборки, а y_i представляет соответствующее истинное значение.

Меньшее значение MSE указывает на лучшую производительность, поскольку оно означает, что прогнозы модели в среднем ближе к фактическим значениям. Однако MSE не предоставляет интуитивного понимания доли объясняемой дисперсии моделью.

R^2, также известный как коэффициент детерминации, измеряет долю дисперсии зависимой переменной, которая может быть объяснена независимыми переменными в модели регрессии. Он количественно характеризует, насколько хорошо модель соответствует данным по сравнению с простой базовой моделью, которая предсказывает среднее значение зависимой переменной. Математическая формула для R^2 выглядит следующим образом:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

В этом уравнении y_i представляет истинное значение для i-й выборки, $\hat{y}_i$ обозначает предсказанное значение, а $\bar{y}$ – среднее значение истинных значений.

Значение R^2 находится в диапазоне от 0 до 1, где значение 1 указывает на идеальное соответствие модели данным. Большее значение R^2 указывает на то, что большая часть дисперсии зависимой переменной может быть объяснена независимыми переменными.

Как MSE, так и R^2 ценны в глубинном обучении для задач регрессии. MSE предоставляет меру средней ошибки прогнозирования, а R^2 позволяет оценить качество соответствия модели данным.

4.4.2 Визуализация данных и анализ переменных

В данном исследовании используется набор данных, состоящий из 35 064 наблюдений, который разделен на три отдельных поднабора: обучающий набор, проверочный набор и тестовый набор. Набор данных состоит из 16 переменных, из которых две являются категориальными. Категориальные переменные включают "Rain указывающую на наличие осадков, и 'wd' (распределение данных показано на рисунке 4.6), обозначающую направление ветра.

Плотности распределения остальных числовых переменных показаны на рисунке 4.7.

TEMP здесь представляет собой температуру (градусы Цельсия).

DEWP (Температура точки росы) - это переменная, часто используемая в об-

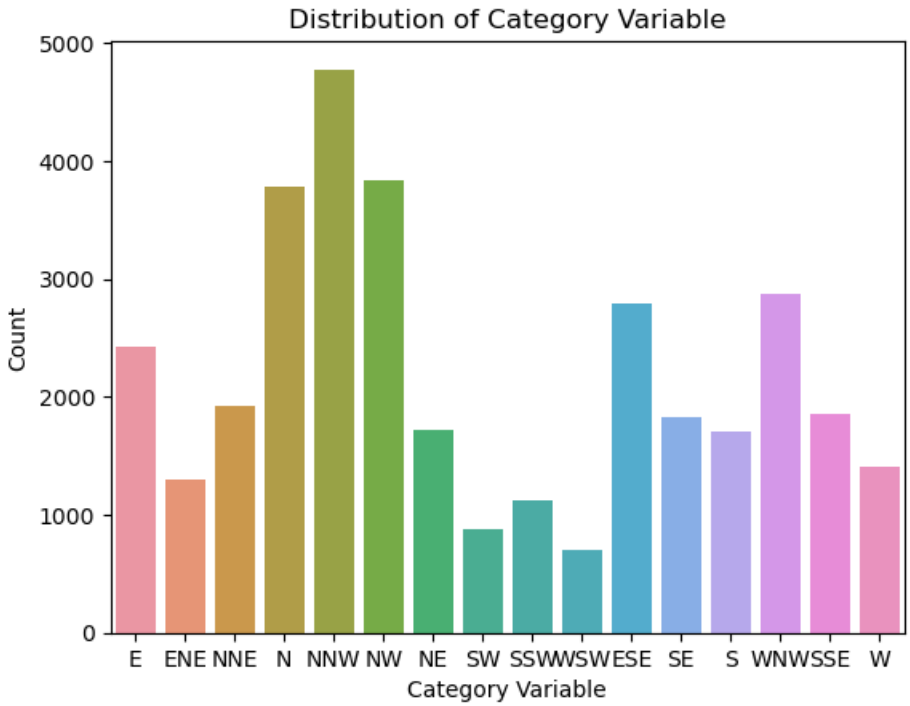


Рис. 4.6: Распределение wd

ласти качества воздуха. Она относится к температуре, при которой воздух насыщается водяным паром при постоянном атмосферном давлении. Когда воздух достигает своей температуры точки росы, происходит конденсация, в результате чего образуется роса или капельки. DEWP часто используется в качестве показателя влажности в воздухе.

WSPM (Скорость ветра) - еще одна важная переменная в исследованиях качества воздуха. Она представляет среднюю скорость ветра, проходящего через фиксированную точку за определенное время. Скорость ветра обычно измеряется в метрах в секунду (м/с). Это важная метрика для оценки силы и скорости ветра, и ее анализ помогает понять влияние движения воздуха на качество воздуха.

PRES (Давление) относится к атмосферному давлению, которое играет значительную роль в исследованиях качества воздуха. Атмосферное давление количественно характеризует силу, с которой атмосфера действует на поверхность Земли или любой другой объект по единице площади. Оно возникает из-за гравитационного влияния внутри атмосферы Земли и является важным фактором для описания атмосферных условий.

Анализ показывает несколько значимых корреляций между переменными. На рисунке 4.8 показаны связи между PM10 и CO, а также между PM10 и NO2,

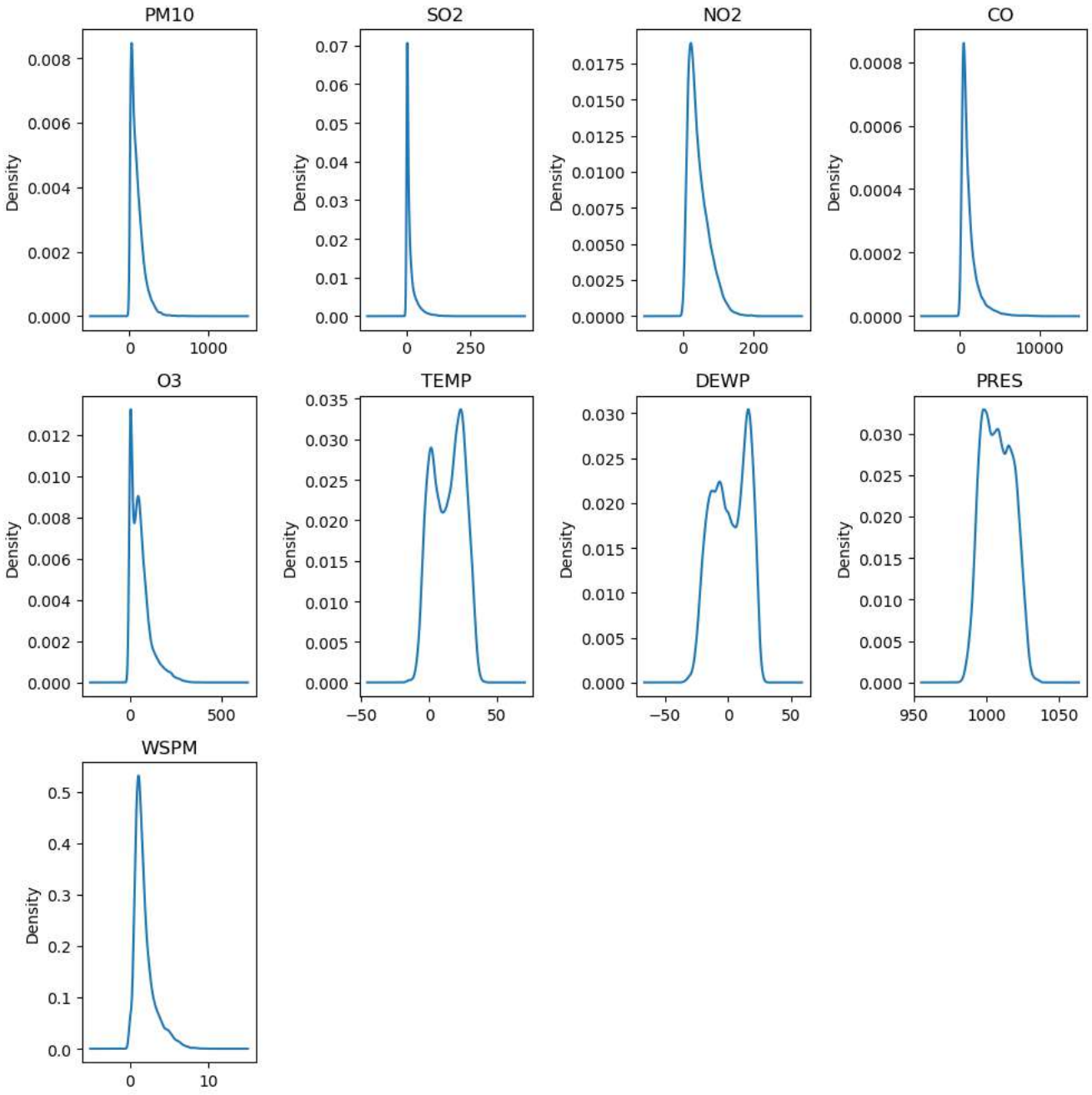


Рис. 4.7: Графики плотности переменных

указывающие на сильную корреляцию между этими переменными. Аналогично наблюдается заметная корреляция между CO и NO2. Кроме того, данные демонстрируют сильную корреляцию между TEMP и O3, а также между DEWP и TEMP.

Коэффициенты корреляции для этих пар переменных превышают порог в 0,6, что указывает на существенную линейную связь между ними. Однако следует отметить, что все остальные переменные демонстрируют независимость друг от друга, не имея значительных корреляций.

Для решения проблемы сильной корреляции между переменными использу-

Рис. 4.8: Тепловая карта корреляции переменных

ются регуляризационные члены как средство ограничения величин параметров модели, тем самым снижая риск переобучения. Эта техника регуляризации помогает предотвратить чрезмерную зависимость от конкретных признаков и поощряет более обобщающую модель.

Помимо регуляризации, перед обучением модели применялись методы предобработки данных. Этот этап предобработки включал нормализацию данных, что эффективно масштабирует значения по переменным. Нормализация данных позволяет снизить влияние корреляции, что позволяет проводить более надежный анализ и процесс моделирования.

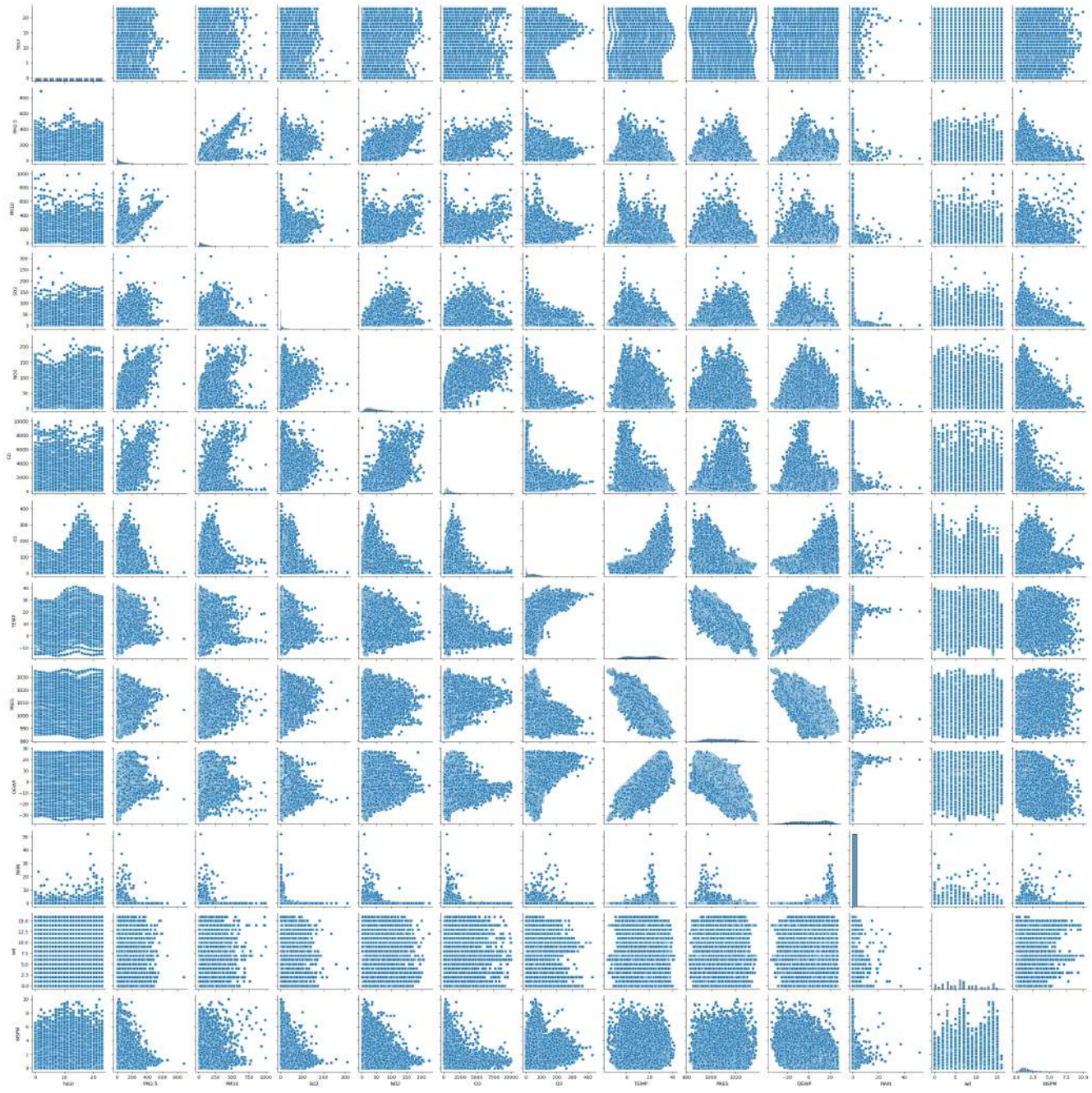


Рис. 4.9: График рассеяния переменных

4.4.3 Результаты моделирования

Таблица 4.1: Анализ иерархии слоев и количества параметров в модели на основе SimpleRNN

Layer (type)	**Output Shape**	**Param #**
simple_rnn_10 (SimpleRNN)	(None, 1, 100)	11,700
dropout_9 (Dropout)	(None, 1, 100)	0
simple_rnn_11 (SimpleRNN)	(None, 50)	7,550
dropout_10 (Dropout)	(None, 50)	0
dense_4 (Dense)	(None, 1)	51
Total params:		**19,301**
Trainable params:		**19,301**
Non-trainable params:		**0**

Исследование архитектуры модели на основе RNN для прогнозирования. Иерархия слоев (в таблице 4.1): Изучая столбец "Слой (тип) можно понять иерархию слоев модели. В данной модели есть слой SimpleRNN с формой вывода (None, 1, 100). Затем следует слой Dropout с той же формой вывода, что и слой SimpleRNN. Далее появляется еще один слой SimpleRNN с формой вывода (None, 50). Наконец, есть слой Dense с формой вывода (None, 1).

Количество параметров: Исходя из столбца "Параметры можно определить количество параметров для каждого слоя. В таблице слой SimpleRNN 1 имеет 11 700 параметров, в то время как слой Dropout 1 не имеет обучаемых параметров. Слой SimpleRNN 2 имеет 7 550 параметров, и слой Dropout 2 также не имеет обучаемых параметров. В конечном итоге, слой Dense имеет 51 параметр. Всего в модели 19 301 параметр, все они обучаемые.

Обучаемые и необучаемые параметры: Столбцы "Обучаемые параметры"и "Необучаемые параметры"позволяют различать обучаемые и необучаемые параметры в модели. В этой таблице все 19 301 параметр обучаемые, и нет необучаемых параметров.

В заключение, модель использует простую архитектуру рекуррентной нейронной сети (SimpleRNN) и включает слои Dropout для снижения риска переобучения. Форма вывода модели постепенно переходит от (None, 1, 100) до (None, 50), а затем становится (None, 1). Всего в модели 19 301 обучаемый параметр, который играет важную роль в улавливании взаимосвязей между входными данными и целевой переменной в процессе обучения.

Таблица 4.2: Исследование архитектуры модели на основе LSTM для прогнозирования

Layer (type)	**Output Shape**	**Param #**
lstm (LSTM)	(None, 1, 100)	46,800
dropout_11 (Dropout)	(None, 1, 100)	0
lstm_1 (LSTM)	(None, 50)	30,200
dropout_12 (Dropout)	(None, 50)	0
dense_5 (Dense)	(None, 1)	51
Total params:		**77,051**
Trainable params:		**77,051**
Non-trainable params:		**0**

Исследование архитектуры модели на основе LSTM для прогнозирования Таблица 4.2 представляет информацию о модели, основанной на слоях долгой кратковременной памяти (LSTM). Модель состоит из двух слоев LSTM, слоя Dropout и слоя Dense.

Слой LSTM 1: У этого слоя форма вывода (None, 1, 100), что указывает на создание последовательности векторов длиной 1, каждый вектор имеет 100 измерений. Он содержит 46 800 параметров.

Слой Dropout: Слой Dropout помогает снизить переобучение, случайным образом устанавливая долю входных единиц равной 0 во время обучения. Он не меняет форму вывода и не имеет обучаемых параметров.

Слой LSTM 2: У этого слоя форма вывода (None, 50), представляющая последовательность векторов длиной 50. Он содержит 30 200 параметров.

Слой Dense: Слой Dense выполняет линейное преобразование входных данных и имеет форму вывода (None, 1), указывающую на одномерный вывод. Он содержит 51 параметр.

В целом, модель имеет в общей сложности 77 051 параметр, все они обучаемые. Эти параметры улавливают взаимосвязи между входными данными и целевой переменной в процессе обучения.

Исследование архитектуры модели на основе GRU для прогнозирования Таблица 4.3 представляет информацию об архитектуре модели, основанной на слоях с блокирующей рекуррентной единицей (GRU). Модель состоит из двух слоев GRU, слоя Dropout и слоя Dense.

Слой GRU 1: У этого слоя форма вывода (None, 1, 100), что указывает на создание последовательности векторов длиной 1, каждый вектор имеет 100 из-

Таблица 4.3: Архитектура модели на основе GRU для прогнозирования

Layer (type)	**Output Shape**	**Param #**
gru_1 (GRU)	(None, 1, 100)	35,400
dropout_14 (Dropout)	(None, 1, 100)	0
gru_2 (GRU)	(None, 50)	22,800
dropout_15 (Dropout)	(None, 50)	0
dense_6 (Dense)	(None, 1)	51
Total params:		**58,251**
Trainable params:		**58,251**
Non-trainable params:		**0**

мерений. Он содержит 35 400 параметров.

Слой Dropout: Слой Dropout помогает снизить переобучение, случайным образом устанавливая долю входных единиц равной 0 во время обучения. Он не меняет форму вывода и не имеет обучаемых параметров.

Слой GRU 2: У этого слоя форма вывода (None, 50), представляющая последовательность векторов длиной 50. Он содержит 22 800 параметров.

Слой Dense: Слой Dense выполняет линейное преобразование входных данных и имеет форму вывода (None, 1), указывающую на одномерный вывод. Он содержит 51 параметр.

В целом, модель имеет в общей сложности 58 251 параметр, все они обучаемые. Эти параметры улавливают взаимосвязи между входными данными и целевой переменной в процессе обучения.

Таблица 4.4: Архитектура Bi-RNN модели для прогнозирования последовательностей

Layer (type)	**Output Shape**	**Param #**
bidirectional_1 (Bidirectional)	(None, 1, 200)	23,400
dropout_1 (Dropout)	(None, 1, 200)	0
dense_1 (Dense)	(None, 1, 1)	201
Total params:		**23,601**
Trainable params:		**23,601**
Non-trainable params:		**0**

Исследование архитектуры модели на основе Bi-RNN для прогнозирования Таблица 4.4 представляет информацию об архитектуре модели на основе бидирекциональной рекуррентной нейронной сети (Bi-RNN). Модель состоит из слоя Bidirectional, слоя Dropout и слоя Dense.

Слой Bidirectional: Этот слой использует два отдельных слоя RNN для обработки входной последовательности как в прямом, так и в обратном направлении. Он имеет форму вывода (None, 1, 200), что указывает на создание последовательности векторов длиной 1, каждый вектор имеет 200 измерений. Он содержит 23 400 параметров.

Слой Dropout: Слой Dropout помогает снизить переобучение, случайным образом устанавливая долю входных единиц равной 0 во время обучения. Он не меняет форму вывода и не имеет обучаемых параметров.

Слой Dense: Слой Dense выполняет линейное преобразование входных данных и имеет форму вывода (None, 1, 1), указывающую на одномерный вывод. Он содержит 201 параметр.

В целом, модель имеет в общей сложности 23 601 параметр, все они обучаемые. Эти параметры улавливают взаимосвязи между входной последовательностью и целевой переменной в процессе обучения.

Таблица 4.5: Архитектура модели на основе двунаправленной долгой краткосрочной памяти (BiLSTM)

Layer (type)	**Output Shape**	**Param #**
bidirectional_2 (Bidirectional)	(None, 1, 200)	93,600
dropout_2 (Dropout)	(None, 1, 200)	0
dense_2 (Dense)	(None, 1, 1)	201
Total params:		**93,801**
Trainable params:		**93,801**
Non-trainable params:		**0**

Исследование архитектуры модели на основе BiLSTM для прогнозирования Таблица 4.5 представляет информацию об архитектуре модели на основе двунаправленной долгой краткосрочной памяти (BiLSTM). Модель состоит из слоя Bidirectional, слоя Dropout и слоя Dense.

Слой Bidirectional: Этот слой использует два отдельных слоя LSTM для обработки входной последовательности как в прямом, так и в обратном направлении. Он имеет форму вывода (None, 1, 200), что указывает на создание последовательности векторов длиной 1, каждый вектор имеет 200 измерений. Он содержит 93 600 параметров.

Слой Dropout: Слой Dropout помогает снизить переобучение, случайным об-

разом устанавливая долю входных единиц равной 0 во время обучения. Он не меняет форму вывода и не имеет обучаемых параметров.

Слой Dense: Слой Dense выполняет линейное преобразование входных данных и имеет форму вывода (None, 1, 1), указывающую на одномерный вывод. Он содержит 201 параметр.

В целом, модель имеет в общей сложности 93 801 параметр, все они обучаемые. Эти параметры улавливают взаимосвязи между входной последовательностью и целевой переменной в процессе обучения.

Таблица 4.6: Архитектура модели на основе двунаправленной блокирующей рекуррентной единицы (BiGRU)

Layer (type)	**Output Shape**	**Param #**
bidirectional_3 (Bidirectional)	(None, 1, 200)	70,800
dropout_3 (Dropout)	(None, 1, 200)	0
dense_3 (Dense)	(None, 1, 1)	201
Total params:		**71,001**
Trainable params:		**71,001**
Non-trainable params:		**0**

Исследование архитектуры модели на основе BiGRU для прогнозирования Таблица 4.6 представляет информацию об архитектуре модели на основе двунаправленной блокирующей рекуррентной единицы (BiGRU). Модель состоит из слоя Bidirectional, слоя Dropout и слоя Dense.

Слой Bidirectional: Этот слой использует два отдельных слоя GRU для обработки входной последовательности как в прямом, так и в обратном направлении. Он имеет форму вывода (None, 1, 200), что указывает на создание последовательности векторов длиной 1, каждый вектор имеет 200 измерений. Он содержит 70 800 параметров.

Слой Dropout: Слой Dropout применяется для снижения переобучения путем случайного установления доли входных единиц равной 0 во время обучения. Он не изменяет форму вывода и не имеет обучаемых параметров.

Слой Dense: Слой Dense выполняет линейное преобразование входных данных и имеет форму вывода (None, 1, 1), указывающую на одномерный вывод. Он содержит 201 параметр.

В целом, модель имеет в общей сложности 71 001 параметр, все они обуча-

емые. Эти параметры улавливают взаимосвязи между входной последовательностью и целевой переменной в процессе обучения.

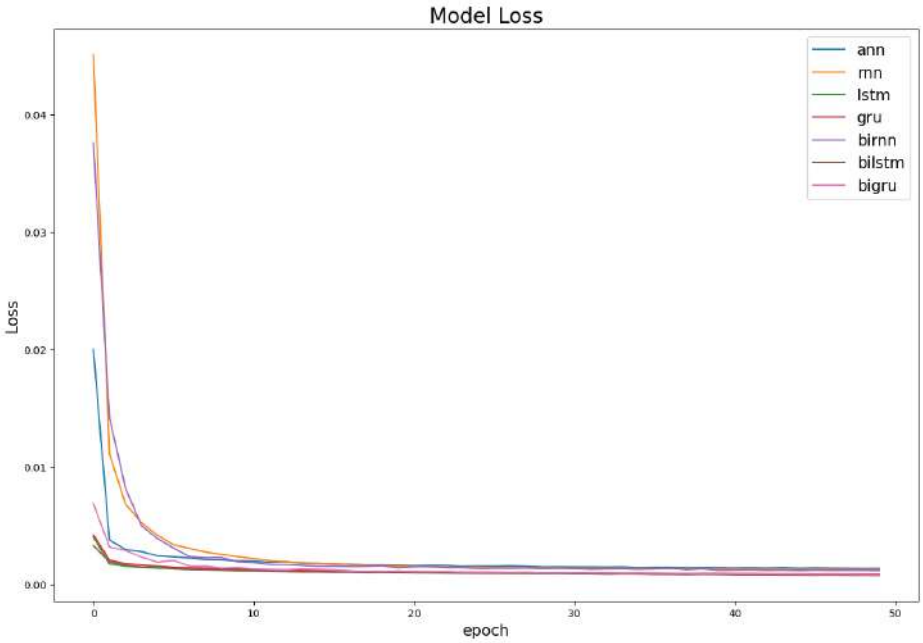


Рис. 4.10: Кривые обучения прогностических моделей.

Рисунок 4.10 иллюстрирует кривые обучения различных моделей, включая ANN, RNN, LSTM, GRU, BiRNN, BiLSTM и BiGRU. Кривые показывают заметное снижение, что свидетельствует о значительном улучшении производительности при увеличении количества обрабатываемых образцов обучения. Этот наблюдаемый тренд является обнадеживающим показателем того, что модели эффективно извлекают знания из предоставленных данных и успешно адаптируются к поставленной задаче.

На рисунке 4.10 демонстрируется, что все рассматриваемые модели проявляют последовательный рост производительности с увеличением объема данных для обучения. Этот положительный тренд подчеркивает способность моделей эффективно обучаться и адаптироваться.

Для тщательного и всестороннего сравнения различных моделей необходимо провести методичный анализ количественных показателей. Тщательное исследование этих метрик позволяет эффективно оценить преимущества и ограничения каждой модели относительно конкретной проблемы. Такой подробный анализ помогает исследователям и практикам принимать обоснованные решения относительно наиболее подходящей модели для конкретного применения.

Кроме того, при сравнении моделей следует учитывать доступные вычисли-

Таблица 4.7: Прогностическое качество текущего PM2.5.

	R^2	**MSE**	**TimeSpent**
ANN	0.9085	0.0013	**6.91s**
RNN	0.9177	0.0013	21.9s
LSTM	0.8970	0.0018	38.2s
GRU	0.9080	0.0015	38.2s
Bi-RNN	**0.9212**	**0.0012**	20s
Bi-LSTM	0.9086	0.0014	36.9s
Bi-GRU	0.9136	0.0015	32.8 s

тельные ресурсы. Важно оценивать производительность каждой модели, учитывая баланс между временем вычислений и предсказательной способностью. Такая оценка помогает определить оптимальную модель, которая достигает баланса между производительностью и эффективностью. Нахождение правильного равновесия между вычислительными требованиями и способностью модели генерировать точные прогнозы является важным для реальных приложений, где вычислительные ресурсы часто ограничены.

Проведя всесторонний анализ количественных показателей и учитывая компьютерные ограничения, исследователи и практики могут принимать обоснованные решения при выборе модели для конкретной задачи. Такой подход гарантирует соответствие выбранной модели требуемому уровню производительности и доступным вычислительным ресурсам, в конечном итоге повышая эффективность и эффективность всей системы.

Сравнение производительности моделей: Изучая значения R^2 и MSE, мы можем сравнить предсказательную способность различных моделей машинного обучения. Исходя из данных таблицы, модель Bi-RNN имеет наивысшее значение $R^2(0.9212)$, что указывает на лучшую прогностическую способность при предсказании текущего качества PM2.5. Кроме того, модель ANN и модель Bi-GRU также обладают относительно высокими значениями $R^2(0.9085$ и $0.9136)$.

Сравнение ошибки предсказания: Значения MSE позволяют оценить среднеквадратичную ошибку различных моделей. Меньшее значение MSE указывает на меньшую разницу между предсказанными значениями и истинными значениями. Из таблицы видно, что модель Bi-RNN и модель ANN имеют наименьшие значения MSE (0.0012 и 0.0013), что означает меньшую ошибку в их предсказаниях.

Сравнение времени обучения: Столбец TimeSpent в таблице отображает время обучения для каждой модели. Можно заметить, что модели ANN требуют наименьшего времени для обучения (6.91 секунды), в то время как модели LSTM и Bi-LSTM имеют относительно более длительные времена обучения (38.2 секунды и 36.9 секунды соответственно). Таким образом, в практических приложениях существует компромисс между производительностью модели и временем обучения.

Основываясь на проведенном анализе, можно сделать некоторые предварительные выводы: модель Bi-RNN проявляет превосходную производительность в задаче прогнозирования текущего PM2.5, обладая высокой предсказательной точностью и меньшей ошибкой. Хотя модель ANN имеет самое короткое время обучения, ее предсказательная способность незначительно уступает другим моделям. Кроме того, модели LSTM и GRU также демонстрируют хорошую производительность, хотя их предсказательная точность немного ниже по сравнению с моделью Bi-RNN.

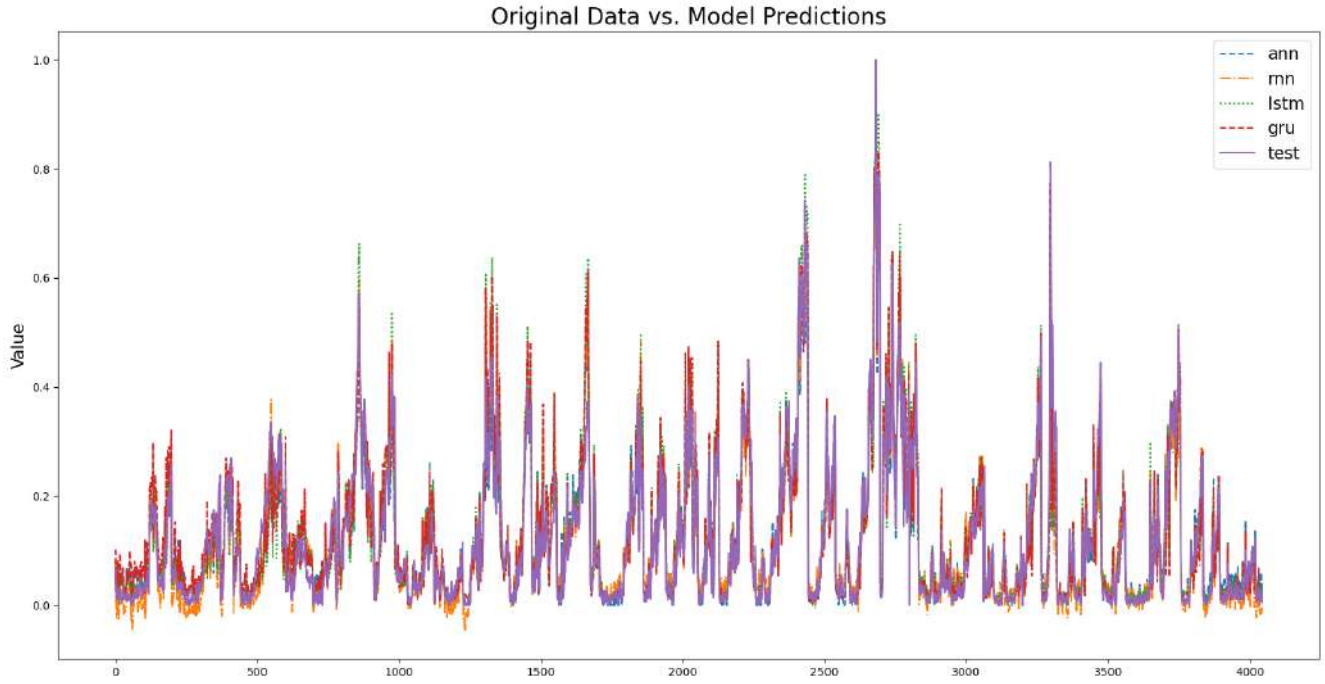


Рис. 4.11: Сравнение исходных и предсказанных данных в прогнозировании временных рядов.

Рисунки 4.11 и 4.12 представляют линейные графики, иллюстрирующие производительность семи различных моделей. В рисунке 4.11 пять линий изображают фактические значения тестовых данных наряду с предсказанными значениями, сгенерированными моделями ANN, RNN, LSTM и GRU. В то же время,

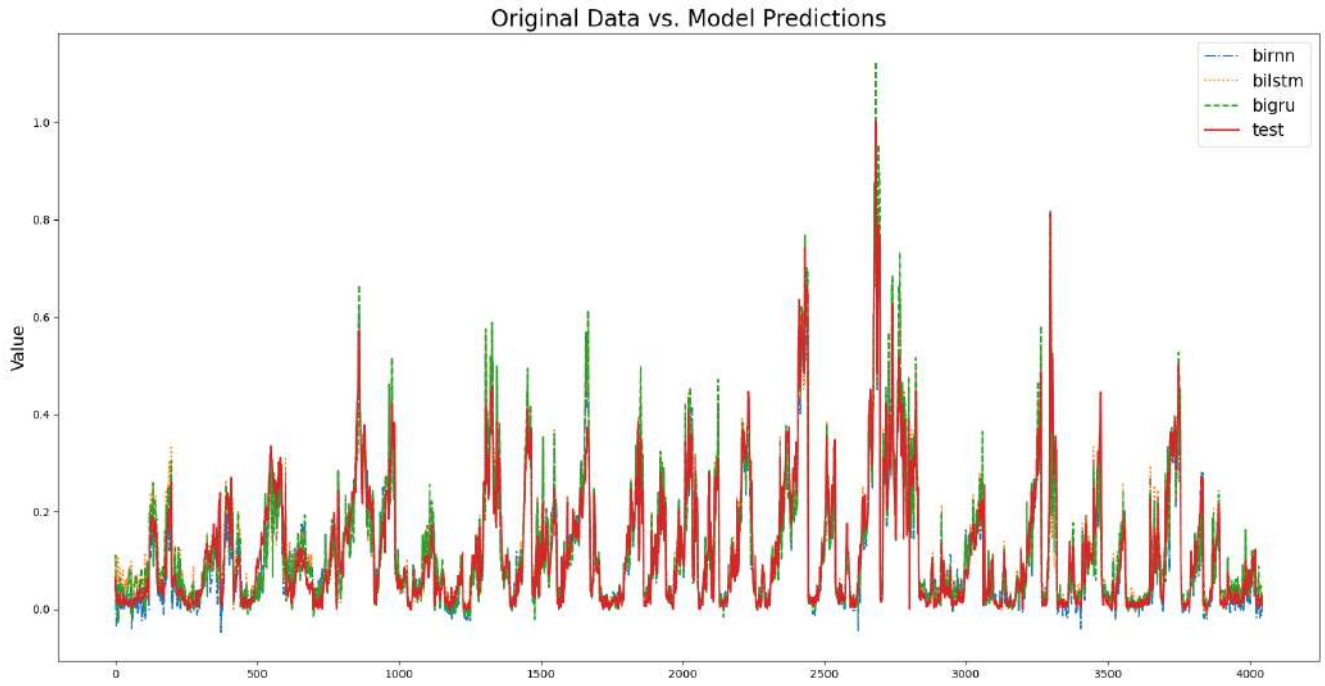


Рис. 4.12: Сравнение исходных и предсказанных данных в прогнозировании временных рядов.

рисунок 4.12 содержит четыре линейных графика, представляющих фактические значения тестовых данных вместе с предсказанными значениями от моделей BiRNN, BiLSTM и BiGRU.

Визуальное исследование этих рисунков указывает на то, что все семь моделей достигли благоприятных результатов. Различия между предсказанными результатами и исходными значениями минимальны, что свидетельствует о высокой точности прогнозов моделей. Кроме того, наблюдаемые изменения тренда в предсказаниях тесно соответствуют паттернам, проявленным исходными данными.

Это согласованство между предсказанными и фактическими значениями указывает на способность моделей эффективно улавливать и воспроизводить основные тренды и паттерны внутри набора данных. Это означает, что эти модели обладают надежной способностью к обучению и могут генерировать достоверные прогнозы, которые тесно соответствуют истинным значениям.

4.5 Заключение к главе 4

В этой главе мы углубились в стандартизированные процедуры использования машинного обучения для прогнозирования временных рядов PM2.5, важ-

ного фактора окружающей среды с значительными последствиями для оценки качества воздуха и общественного здоровья. Через тщательный анализ и сравнение семи различных моделей, включая ANN, RNN, LSTM, GRU, Bi-RNN, Bi-LSTM и Bi-GRU, мы получили ценные практические результаты относительно их производительности и прогностических возможностей.

Основой нашей методологии является разделение набора данных на три отдельных подмножества: обучающий набор, проверочный набор и тестовый набор. Такой подход гарантирует целостность и надежность оценки модели путем предоставления отдельных данных для обучения, проверки и финального тестирования. Соблюдение этих стандартизированных процедур позволило нам получить точные и содержательные выводы о производительности каждой модели.

В нашей оценке мы использовали две ключевые метрики: R^2(коэффициент детерминации) и MSE (среднеквадратичная ошибка), чтобы всесторонне сравнить прогностические способности различных моделей. Значение R^2 служит мерой того, насколько хорошо модель соответствует наблюдаемым данным, указывая на ее способность улавливать основные шаблоны и тренды во временных рядах PM2.5. С другой стороны, MSE предоставляет количественную оценку среднеквадратичной разницы между предсказанными и истинными значениями, обеспечивая информацию об общей точности прогнозов моделей.

После сравнения прогнозных результатов с фактическими данными PM2.5 мы обнаружили, что все семь моделей показали благоприятные результаты. Отличия между предсказанными результатами и исходными значениями минимальны, что свидетельствует о высокой точности прогнозов моделей. Кроме того, изменения тренда, наблюдаемые в предсказаниях, тесно соответствуют паттернам, проявленным исходными данными.

Это согласованство между предсказанными и фактическими значениями указывает на способность моделей эффективно улавливать и воспроизводить основные тренды и паттерны в наборе данных. Это говорит о том, что эти модели обладают надежной способностью к обучению и могут генерировать достоверные прогнозы, которые тесно соответствуют истинным значениям.

В заключение, наш всесторонний анализ предлагает ценные практические результаты в отношении стандартизированных процедур использования машинного обучения для прогнозирования PM2.5. Полученные результаты подчерки-

вают превосходную производительность модели Bi-RNN, особенно в отношении точности подгонки и меньшей ошибки прогноза. Однако модели ANN, LSTM, GRU, Bi-LSTM и Bi-GRU также проявляют конкурентоспособные прогностические возможности, хотя и с небольшими различиями в точности и вычислительных требованиях.

Эти исследовательские результаты вносят свой вклад в область анализа окружающей среды, предоставляя руководство по выбору и использованию подходящих моделей машинного обучения для задач прогнозирования PM2.5. Исследователи и практики могут использовать эти выводы для принятия обоснованных решений, основанных на их конкретных требованиях, находя баланс между точностью прогнозов, вычислительной эффективностью и временем обучения. В конечном итоге, наше исследование продвигает понимание и применение методов машинного обучения в области мониторинга окружающей среды и повышает нашу способность оценивать и управлять рисками, связанными с загрязнением воздуха.

Глава 5

Методы ансамблевого обучения для системы оценки качества воздуха

В этой главе мы углубляемся в применение методов ансамблевого обучения для прогнозирования временных рядов качества воздуха. В частности, мы исследуем три широко используемых модели ансамблевого обучения: XGBoost, LightGBM и CatBoost.

Чтобы обеспечить согласованность и сопоставимость с предыдущей главой, мы используем тот же набор данных и применяем аналогичные методы обработки данных. Это включает такие шаги, как удаление выбросов, заполнение пропущенных значений и сглаживание данных, что способствует повышению точности и стабильности модели.

Используя эти три модели ансамблевого обучения, мы можем подходить к моделированию и прогнозированию временных рядов качества воздуха с различных точек зрения. Каждая модель обладает своими преимуществами и подходит для конкретных сценариев. В практических приложениях выбор подходящей модели зависит от конкретного контекста. Использование ансамблевого обучения позволяет объединить прогнозы нескольких слабых моделей в целом, тем самым улучшая общую прогностическую производительность и обеспечивая устойчивость.

Большинство результатов, представленных в этой главе, были опубликованы в статьях [102] и [103].

5.1 Методология

5.1.1 Экстремальное градиентное бустинг(XGBoost)

XGBoost (Extreme Gradient Boosting) - это модель ансамблевого обучения, которая зарекомендовала себя как эффективная в задачах прогнозирования временных рядов. Она сочетает принципы градиентного бустинга и деревьев решений для достижения точных прогнозов на временных данных.

Алгоритм XGBoost может быть сформулирован следующим образом:

Для заданного набора обучающих данных $(\mathbf{x}_i, y_i)i = 1^N$, где $\mathbf{x}_i$ представляет собой вектор признаков на шаге времени i, а y_i обозначает соответствующее целевое значение, целью XGBoost является поиск функции прогнозирования $F(\mathbf{x})$, которая минимизирует регуляризованную целевую функцию, определенную следующим образом:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

где L - функция потерь, измеряющая расхождение между предсказанными значениями и фактическими целями, $F(\mathbf{x}_i)$, а $\Omega(f_k)$ - регуляризационный член, штрафующий сложность модели. Здесь f_k представляет отдельные деревья решений в ансамбле.

Для построения ансамбля итерационно XGBoost использует стратегию бустинга, которая последовательно добавляет новые слабые модели для улучшения остатков, оставленных предыдущими моделями. Функция прогнозирования на каждой итерации задается суммой прогнозов всех отдельных деревьев:

$$F_t(\mathbf{x}) = \sum_{k=1}^{t} f_k(\mathbf{x})$$

где t обозначает текущую итерацию.

Основная идея XGBoost заключается в оптимизации целевой функции с помощью градиентного спуска. Рассчитывая градиенты функции потерь относительно прогнозов ансамбля и используя их для подгонки нового дерева решений, XGBoost учится исправлять ошибки, допущенные предыдущими моделями. Этот процесс повторяется итеративно до минимизации целевой функции.

Для предотвращения переобучения и улучшения обобщающей способности модели XGBoost включает регуляризацию в целевую функцию. Регуляризационный член $\Omega(f_k)$ контролирует сложность отдельных деревьев, штрафуя их структуру или веса листьев. Это предотвращает слишком сложную модель и улучшает ее способность к обобщению на невидимые данные.

XGBoost также включает дополнительные продвинутые функции, такие как обработка пропущенных значений, субдискретизация и выбор столбцов. Эти техники дополнительно повышают производительность и устойчивость модели.

В заключение, XGBoost - это мощный алгоритм ансамблевого обучения, который объединяет градиентный бустинг и деревья решений для прогнозирования временных рядов. Он оптимизирует регуляризованную целевую функцию через итеративное обучение, используя преимущества слабых моделей для создания точных прогнозов.

5.1.2 Легкое градиентное бустинг(LightGBM)

LightGBM - это фреймворк градиентного бустинга, который зарекомендовал себя благодаря своей эффективной и эффективной работе в задачах прогнозирования временных рядов. Он специально разработан для работы с масштабными наборами данных и обеспечивает высокую точность при быстром обучении.

Алгоритм LightGBM можно описать следующим образом:

Для заданного набора обучающих данных $(\mathbf{x}_i, y_i)i = 1^N$, где $\mathbf{x}_i$ представляет собой вектор признаков на шаге времени i, а y_i обозначает соответствующее целевое значение, цель LightGBM - найти функцию прогнозирования $F(\mathbf{x})$, которая минимизирует следующую функцию потерь:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

Здесь L - функция потерь, измеряющая расхождение между предсказанными значениями и фактическими целями, $F(\mathbf{x}_i)$. Регуляризационный член $\Omega(f_k)$ штрафует сложные модели, чтобы предотвратить переобучение. Аналогично XGBoost, f_k представляет отдельные деревья решений в ансамбле.

LightGBM использует стратегию построения деревьев с учетом листа, которая отличается от традиционного подхода по уровням. При построении дерева алгоритм разделяет лист, который приведет к наибольшему уменьшению

функции потерь. Эта стратегия позволяет достичь более быстрой сходимости и лучшей общей производительности.

Для работы с временными рядами данных LightGBM включает специальную функцию поддержки "категориальных признаков". Он эффективно обрабатывает категориальные признаки без необходимости кодирования one-hot, что снижает использование памяти и вычислительную сложность.

Более того, LightGBM включает дополнительные методы, такие как субдискретизация признаков и бэггинг, которые повышают обобщающую способность модели и ее устойчивость. Субдискретизация признаков случайным образом выбирает подмножество признаков для каждого дерева, что снижает переобучение и улучшает разнообразие модели. Бэггинг предполагает обучение нескольких моделей на разных подмножествах обучающих данных и усреднение их прогнозов, что дальше повышает точность прогноза.

LightGBM также использует алгоритмы на основе гистограмм для ускорения процесса обучения путем группировки значений в дискретные интервалы. Эта техника снижает использование памяти и позволяет проводить вычисления быстрее.

LightGBM - мощный фреймворк градиентного бустинга для прогнозирования временных рядов. Он минимизирует функцию потерь с помощью стратегии построения деревьев с учетом листьев и включает методы регуляризации, чтобы предотвратить переобучение. Благодаря эффективной обработке масштабных наборов данных и поддержке категориальных признаков, LightGBM обеспечивает точные прогнозы с быстрыми временами обучения.

5.1.3 Кошачий Бустинг(CatBoost)

CatBoost - это алгоритм градиентного бустинга, который зарекомендовал себя благодаря своей способности эффективно обрабатывать категориальные переменные в задачах прогнозирования временных рядов. Он включает специализированные техники для работы с категориальными признаками и предоставляет надежные прогнозы.

Алгоритм CatBoost может быть описан следующим образом:

Для заданного набора обучающих данных $(\mathbf{x}_i, y_i)i = 1^N$, где $\mathbf{x}_i$ представляет собой вектор признаков на шаге времени i, а y_i обозначает соответствующее целевое значение, цель CatBoost - найти функцию прогнозирования $F(\mathbf{x})$, которая

минимизирует следующую функцию потерь:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

Здесь L - функция потерь, измеряющая расхождение между предсказанными значениями и фактическими целями, $F(\mathbf{x}_i)$. Регуляризационный член $\Omega(f_k)$ штрафует сложные модели, чтобы предотвратить переобучение. Аналогично XGBoost и LightGBM, f_k представляет отдельные деревья решений в ансамбле.

CatBoost использует инновационную технику под названием "упорядоченное усиление чтобы работать с категориальными признаками непосредственно. Он строит отдельное дерево решений для каждого категориального признака с использованием упорядоченного метода, который учитывает статистические свойства категорий. Этот подход позволяет CatBoost улавливать полезную информацию из категориальных признаков и делать точные прогнозы.

Для улучшения обобщающей способности и предотвращения переобучения CatBoost использует комбинацию градиентной оптимизации и случайных перестановок. Во время обучения он случайным образом перемешивает порядок категориальных значений, чтобы снизить влияние порядка на результаты.

Кроме того, CatBoost включает метод, называемый "Разложение в ряд Тейлора для приближения целевой функции. Эта техника помогает более точно моделировать нелинейные отношения между входными признаками и целевой переменной.

Кроме этого, CatBoost содержит такие методы, как планирование скорости обучения, субдискретизация признаков и ранняя остановка. Планирование скорости обучения позволяет настраивать скорость обучения во время обучения для улучшения сходимости и предотвращения перескока. Субдискретизация признаков случайным образом выбирает подмножество признаков для каждого дерева, что снижает переобучение и улучшает разнообразие модели. Ранняя остановка прекращает процесс обучения, когда производительность модели на валидационном наборе не улучшается, предотвращая переобучение и экономя вычислительные ресурсы.

CatBoost - мощный алгоритм градиентного бустинга, предназначенный для задач прогнозирования временных рядов с категориальными признаками. Он использует упорядоченное усиление, случайные перестановки и разложение в

ряд Тейлора для эффективной работы с категориальными переменными. Благодаря техникам регуляризации и дополнительным функциям, таким как планирование скорости обучения и субдискретизация признаков, CatBoost предоставляет надежные прогнозы, избегая переобучения.

5.1.4 Результаты моделирования

Таблица 5.1: Параметры модели XGBoost

Parameter	Value
objective	reg:squarederror
eval_metric	mae
learning_rate	0.3
max_depth	6
subsample	1
random_state	None

Анализ параметров модели XGBoost. Таблица 5.1 представляет основные параметры модели для использования в XGBoost при задачах регрессии. Этот анализ позволяет лучше понять выбор параметров во время обучения модели, что помогает исследователям лучше понять поведение модели и принимать обоснованные решения при применении XGBoost к собственным задачам регрессии.

- Объект: Функция потерь, используемая для задач регрессии. Объект reg: squarederror минимизирует среднеквадратичную ошибку (MSE) между предсказанными и фактическими значениями целевой переменной.
- Метрика оценки: Метрика, используемая для оценки производительности модели во время обучения. Метрика "mae"измеряет среднюю абсолютную ошибку (MAE), предоставляя информацию о среднем значении ошибок прогнозирования.
- Скорость обучения: Управляет размером шага на каждой итерации бустинга. Скорость обучения 0.3 указывает на относительно большие шаги, которые могут ускорить сходимость, но требуют тщательной настройки, чтобы избежать переполнения решений.

- Максимальная глубина: Задает максимальную глубину каждого дерева решений. При максимальной глубине 6 модель способна улавливать сложные взаимодействия между признаками, но рискует переобучением, если не применена должная регуляризация.

- Сэмплирование: Управляет долей обучающих экземпляров, используемых для каждого дерева. Значение сэмплирования равное 1 означает использование всех обучающих экземпляров, что может привести к моделям с более высокой дисперсией. Уменьшение этого значения может уменьшить переобучение.

- Случайность: Значение семени для генерации случайных чисел. Установка его как None приведет к изменению поведения модели при каждом запуске, что полезно для оценки стабильности и обобщающей способности модели.

Таблица 5.2: Параметры модели LightGBM

Parameter	Value	Parameter	Value
boosting_type	gbdt	min_child_samples	20
class_weight	None	min_child_weight	0.001
colsample_bytree	1.0	min_split_gain	0.0
importance_type	split	n_estimators	100
learning_rate	0.1	n_jobs	-1
max_depth	-1	num_leaves	31
objective	None	random_state	None
reg_alpha	0.0	reg_lambda	0.0
silent	warn	subsample	1.0
subsample_for_bin	200000	subsample_freq	0

Анализ параметров модели LightGBM Таблица 5.2 представляет полный обзор параметров модели, используемых в LightGBM, популярном фреймворке градиентного бустинга.

Описание параметров:

- Тип бустинга: Определяет тип используемого алгоритма бустинга. По умолчанию используется метод 'gbdt' (градиентный бустинг решающих деревьев).

- Скорость обучения: Управляет размером шага на каждой итерации бустинга. Скорость обучения 0.1 указывает на умеренные шаги, обеспечивающие баланс между скоростью сходимости и точностью.
- Максимальная глубина: Задает максимальную глубину каждого дерева решений. Значение -1 означает отсутствие ограничения на максимальную глубину, что позволяет деревьям расти без ограничений.
- Количество листьев: Определяет максимальное количество листьев в дереве. Значение по умолчанию равное 31 обеспечивает достаточную гибкость для улавливания сложных взаимосвязей в данных.
- Регуляризация: Регулирует переобучение с помощью техник регуляризации. Коэффициенты регуляризации как L1 (reg alpha) и L2 (reg lambda) по умолчанию установлены равными нулю, что указывает на отсутствие регуляризации.
- Сэмплирование: Управляет долей выборок, используемых на каждой итерации бустинга. Значение сэмплирования 1.0 означает использование всего обучающего набора, обеспечивая оптимальную производительность модели.

Исследование архитектуры модели на основе CatBoost для предиктивного анализа. Алгоритм CatBoostRegressor предлагает несколько параметров, которые могут быть настроены для улучшения производительности регрессионной модели. Понимание этих параметров и их влияния является важным для исследователей и практиков, стремящихся достичь оптимальных результатов. В данном анализе проводится всестороннее рассмотрение ключевых параметров, используемых в модели CatBoostRegressor.

Модель CatBoostRegressor была обучена с использованием следующих параметров:

- Настройка nan mode: Выбрано значение "Min что означает обработку пропущенных значений как минимальных значений во время обучения.
- Метрика оценки eval metric: Для модели выбрана метрика RMSE (Root Mean Square Error), которая измеряет точность регрессионных предсказаний.

Таблица 5.3: Параметры CatBoostRegressor

Parameter	Value	Parameter	Value
nan_mode	Min	feature_border_type	GreedyLogSum
eval_metric	RMSE	bayesian_matrix_reg	0.10000000149011612
iterations	1000	force_unit_auto_pair_weights	False
sampling_frequency	PerTree	l2_leaf_reg	3
leaf_estimation_method	Newton	random_strength	1
grow_policy	SymmetricTree	rsm	1
penalties_coefficient	1	boost_from_average	True
boosting_type	Plain	model_size_reg	0.5
model_shrink_mode	Constant	pool_metainfo_options	{'tags': {}}
depth	6	subsample	0.800000011920929
posterior_sampling	False	use_best_model	False
border_count	254	random_seed	0
auto_class_weights	None	loss_function	RMSE
sparse_features_conflict_fraction	0	learning_rate	0.06794899702072144
leaf_estimation_backtracking	AnyImprovement	score_function	Cosine
best_model_min_trees	1	leaf_estimation_iterations	1
model_shrink_rate	0	bootstrap_type	MVS
min_data_in_leaf	1	max_leaves	64

- Количество итераций iterations: Модель была дообучена 1000 раз для уточнения результатов.
- Частота сэмплирования sampling frequency: Использовалась частота сэмплирования PerTree для определения случайной стратегии субсэмплирования.
- Метод оценки листа leaf estimation method: В процессе построения дерева использовался метод Ньютона для оценки значений листьев.
- Стратегия роста grow policy: Использовалась стратегия SymmetricTree, позволяющая симметричный рост дерева от корня.
- Коэффициенты штрафов penalties coefficient: Установлен коэффициент штрафа в значение 1, влияющий на регуляризацию применяемую к модели.
- Тип бустинга boosting type: Выбран тип Plain, указывающий на отсутствие дополнительных модификаций стандартного градиентного бустинга.
- Режим уменьшения модели model shrink mode: Используется режим константа, что приводит к постоянной скорости уменьшения модели во время обучения.

- Тип границы признака feature border type: Использован GreedyLogSum, который позволяет эффективно работать с категориальными признаками.
- Регуляризация матрицы Байеса bayesian matrix reg: Применен параметр регуляризации, равный 0.1, для матрицы Байеса.
- Автоматическое вычисление весов пар force unit auto pair weights: Эта опция отключена (False), что позволяет модели вычислять автоматические веса пар по мере необходимости.
- L2-регуляризация листа l2 leaf reg: Коэффициент регуляризации L2 установлен на 3, контролирующий силу регуляризации в модели.
- Случайность random strength: Установлено значение случайности равное 1, вводящее случайные возмущения к значениям признаков во время обучения.
- RSM (Random Subspace Method) rsm: Установлено значение RSM равное 1, позволяющее случайный выбор подпространства для каждого дерева.
- boost from average: Включено усиление от средних предсказаний, способствующее более стабильному обучению модели.
- Регуляризация размера модели model size reg: Применена регуляризация размера модели, равная 0.5, влияющая на сложность полученной модели.
- Опции pool metainfo options: Дополнительные опции мета-информации не использовались, и словарь тегов остался пустым.
- Субсэмплирование subsample: Каждое дерево обучалось на случайном подмножестве, содержащем 80% обучающих данных.
- Использование лучшей модели use best model: Лучшая модель не использовалась в процессе обучения.
- Случайное зерно random seed: Установлено случайное зерно равное 0 для обеспечения воспроизводимости результатов.
- Глубина дерева depth: Глубина дерева установлена на 6, что ограничивает сложность отдельных деревьев и предотвращает переобучение.

- Постериорная выборка posterior sampling: Постериорная выборка отключена, что означает учет только одной выборки при построении дерева.
- Границы border count: Количество границ установлено на 254, определяя количество интервалов для числовых признаков.
- Автоматическое вычисление весов классов auto class weights: Автоматическое вычисление весов классов не применялось.
- Доли конфликтов разреженных признаков sparse features conflict fraction: Доля конфликтов разреженных признаков установлена на 0, что обрабатывает конфликты между разреженными признаками.
- Откат оценки значения листа leaf estimation backtracking: Во время оценки значения листьев использовался метод AnyImprovement.
- Минимальное количество деревьев для лучшей модели best model min trees: Для выбора лучшей модели требуется минимум 1 дерево.
- Скорость уменьшения модели model shrink rate: Скорость уменьшения модели установлена на 0, что указывает на отсутствие уменьшения модели во время обучения.
- Минимальные данные в листе min data in leaf: Каждый лист должен содержать как минимум 1 точку данных.
- Функция потерь loss function: Функция потерь, используемая для оптимизации модели, является RMSE, соответствующая метрике оценки.
- Скорость обучения learning rate: Скорость обучения установлена на 0.0679, контролирующая размер шага при оптимизации.
- Функция оценки score function: В качестве функции оценки выбрана косинусная схожесть cosine similarity.
- Итерации оценки листа leaf estimation iterations: Итерации оценки листа установлены на 1, обеспечивая эффективную оценку значений листьев.
- Тип бутстрэпа bootstrap type: Деревья строятся с использованием типа бутстрэпа Multiple times with replacement (MVS).

- Максимальное количество листьев max leaves: Каждое дерево может содержать максимум 64 листа.

Таблица 5.4: Прогностическое качество текущего PM2.5

	R^2	**MSE**	**TimeSpent**
XGBoost	0.8551	0.0018	**756 ms**
LightGBM	0.9134	0.0013	140 ms
CatBoosst	0.9112	0.0014	2.96 s

Результаты моделирования. Таблица, предоставленная, представляет сравнение трех дополнительных моделей машинного обучения - XGBoost, LightGBM и CatBoost, по качеству прогнозирования текущих значений PM2.5. Оцениваемые метрики включают R^2 (коэффициент детерминации), MSE (средняя квадратичная ошибка) и TimeSpent (время обучения).

Проанализировав значения R^2, которые указывают на точность подгонки моделей, мы видим, что LightGBM достигает наивысшего значения 0,9134. Это свидетельствует о том, что LightGBM способен захватить значительную часть дисперсии данных PM2.5, что делает его надежным инструментом для точного прогнозирования. XGBoost и CatBoost также показывают хорошие результаты в этом отношении, с значениями R^2 0,8551 и 0,9112 соответственно.

Перейдя к оценке ошибок прогнозирования с использованием MSE, метрики, которая количественно определяет среднеквадратичную разницу между прогнозируемыми и фактическими значениями, мы обнаруживаем, что LightGBM достигает наименьшего значения MSE - 0,0013. Это говорит о том, что прогнозы LightGBM имеют наименьшее отклонение от фактических значений PM2.5 в среднем. XGBoost и CatBoost также показывают относительно низкие значения MSE - 0,0018 и 0,0014 соответственно.

Если рассмотреть время обучения, представленное столбцом TimeSpent, можно заметить, что LightGBM имеет самое быстрое время обучения - 140 мс. XGBoost требует 756 мс, в то время как CatBoost занимает 2,96 секунды для завершения процесса обучения. Это указывает на то, что LightGBM является вычислительно эффективным и предпочтительным выбором для приложений со временными ограничениями.

В заключение, основываясь на анализе представленной таблицы, LightGBM

выступает в роли наиболее результативной модели среди трех, показывая самое высокое значение R^2 и наименьшее значение MSE. Кроме того, LightGBM продемонстрировал самое короткое время обучения, подчеркивая его вычислительную эффективность. XGBoost и CatBoost также проявляют конкурентоспособное качество прогнозирования, хотя с немного более низкими значениями R^2 и MSE по сравнению с LightGBM.

Эти результаты способствуют пониманию альтернативных моделей машинного обучения для прогнозирования PM2.5. Исследователи и практики могут рассмотреть использование LightGBM, XGBoost или CatBoost в зависимости от своих конкретных требований в отношении точности прогнозирования, вычислительной эффективности и временных ограничений.

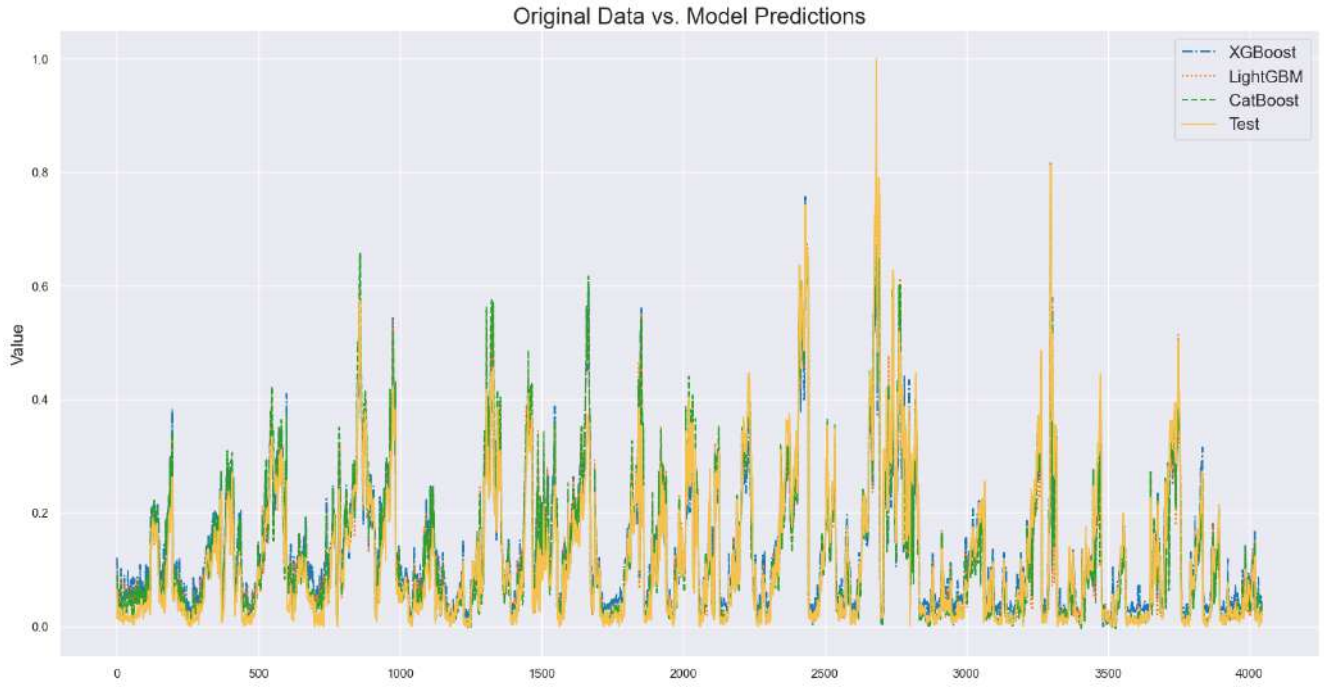


Рис. 5.1: Сравнение сырых и прогнозируемых данных в прогнозировании временных рядов

На рисунке 5.1 изображены линейные графики, демонстрирующие производительность трех различных моделей - XGBoost, LightGBM и CatBoost. Рисунок состоит из четырех линейных графиков, представляющих фактические тестовые значения вместе с прогнозируемыми значениями каждой модели.

Тщательное визуальное исследование этих линейных графиков раскрывает убедительные доказательства благоприятных результатов моделей. Отклонения между прогнозируемыми результатами и исходными значениями минимальны, что указывает на высокую точность прогнозов моделей. Более того, наблюдаемые изменения в прогнозируемых значениях тесно соответствуют пат-

тернам, представленным исходными данными.

Соответствие между прогнозируемыми и фактическими значениями подтверждает эффективность моделей в захвате и повторении основных трендов и паттернов в данных. Это существенная последовательность свидетельствует о способности моделей к обучению и генерации достоверных прогнозов, которые близки к истинным значениям.

5.2 Объяснение модели

5.2.1 Объясняемый искусственный интеллект

Объясняемый искусственный интеллект (XAI) стал важной областью исследований в области искусственного интеллекта. Необходимость интерпретируемости и понимания систем искусственного интеллекта вызвана юридическими, этическими и социальными проблемами, особенно когда они применяются в областях с высокой степенью ответственности, таких как здравоохранение, финансы и автономные транспортные средства. В этих областях необходимы модели искусственного интеллекта, которые не только предоставляют точные прогнозы, но и объясняют свои решения понятным образом.

Для решения этой проблемы исследователи разработали различные подходы и методы в рамках XAI. Один из известных подходов - это модели на основе правил, где процесс принятия решений представлен с помощью набора явных правил. Деревья решений являются хорошим примером таких моделей, где каждый внутренний узел представляет собой условие на входных признаках, а каждый лист соответствует прогнозу или решению. Следуя пути от корня к соответствующему листу, можно интерпретировать, как модель пришла к своему прогнозу на основе заданных входных данных.

Математически дерево решений может быть представлено следующим образом:

$$f(x) = \begin{cases} y_1 & \text{if } x < a \\ y_2 & \text{if } x \geq a \end{cases}$$

Здесь x представляет входные признаки, a обозначает пороговое значение разделения, а y_1 и y_2 представляют различные прогнозы. Изучая правила при-

нятия решений, эксперты в области могут получить представление о том, почему делаются конкретные прогнозы.

Еще один математический аспект XAI - это меры важности признаков. Они количественно оценивают относительное влияние входных признаков на прогнозы модели. Например, в моделях на основе деревьев решений значимость признака может быть вычислена путем оценки уменьшения неоднородности (например, индекса Джини или энтропии), вызванного разделением по конкретному признаку. Эта мера указывает, какие признаки вносят наибольший вклад в принятие решений.

Математически, важность признака ($I(\text{feature})$)) может быть вычислена следующим образом:

$$I(\text{feature}) = \sum_{\text{nodes } t} p(t)\Delta i(t, \text{feature})$$

В этом уравнении t представляет узлы дерева решений, $p(t)$ - долю образцов в узле t, а $\Delta i(t, \text{feature})$ указывает на изменение неоднородности, достигнутое разделением по признаку в узле t. Анализируя важность признаков, заинтересованные стороны могут получить более глубокое понимание того, какие аспекты входных данных влияют на принятие решений модели.

Кроме того, в локальном объяснении методы играют значительную роль в XAI. Эти методы стремятся предоставить интерпретируемые обоснования для отдельных прогнозов, а не предлагать глобальные выводы. Один из таких методов - LIME (Local Interpretable Model-agnostic Explanations), который аппроксимирует поведение сложной модели вокруг конкретного экземпляра, подгоняя интерпретируемую модель (например, линейную регрессию) на основе взвешенных обучающих выборок. Полученное приближение предлагает локальное объяснение, объясняющее, как модель пришла к своему прогнозу для этого конкретного экземпляра.

Математически локальное приближение, предоставленное LIME, представляется следующим образом:

$$f(x') = w_1 \cdot x'_1 + w_2 \cdot x'_2 + \ldots + w_n \cdot x'_n$$

Здесь x' представляет окрестность объясняемого экземпляра, а w_i обозначает веса, назначенные каждому соответствующему признаку. LIME позволяет

пользователям понять вклад различных признаков в конкретный прогноз, тем самым улучшая интерпретируемость.

В заключение, объяснимый искусственный интеллект (XAI) отвечает потребности в прозрачности и интерпретируемости систем искусственного интеллекта. Путем использования моделей на основе правил, мер важности признаков и локальных объяснений XAI предоставляет понятное обоснование решений, принятых искусственными интеллект-моделями. Эти подходы не только повышают доверие к системам искусственного интеллекта, но также позволяют экспертам в области проверить логику прогнозирования и обеспечить справедливость и ответственность. Дальнейшие исследования и разработки в области XAI будут продвигать эту область и открывать путь для ответственного и этичного использования технологий искусственного интеллекта.

С учетом цели объяснения существующих прогностических моделей, в данном исследовании в качестве объяснительного фреймворка используется метод SHAP (Shapley Additive Explanations), как метод Post-hoc. Этот выбор обусловлен прочной теоретической основой, предоставленной кооперативной теорией игр, и наличием комплексных инструментов кодирования, которые облегчают практическую реализацию.

5.2.2 Интерпретация результатов с помощью объяснений на основе SHAP

SHapley Additive exPlanations (SHAP) - это метод, используемый в объясняемом искусственном интеллекте (XAI) для назначения индивидуальных вкладов признаков к прогнозам, сделанным моделями машинного обучения. Он основан на концепции значений Шепли из кооперативной теории игр, которая количественно оценивает справедливое распределение выигрышей между участниками коалиции.

Математически SHAP предоставляет единый фреймворк для объяснения вывода любой модели машинного обучения, присваивая числовое значение важности каждому входному признаку. Это значение представляет собой вклад признака в прогноз, сделанный моделью.

Рассмотрим модель машинного обучения, которая принимает (n) признаков на входе и производит прогноз или решение, обозначенное как (f(x)). SHAP измеряет индивидуальную важность признака, рассматривая все возможные

подмножества признаков и оценивая их влияние на вывод модели. Значение Шепли для признака (i) вычисляется как среднее предельного вклада признака по всем возможным комбинациям признаков.

Математическая формула для вычисления значения Шепли ϕ_i для признака i с использованием перестановочных SHAP имеет следующий вид:

$$\phi_i(f) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup i) - f(S)]$$

Здесь N представляет набор всех признаков, S обозначает подмножество признаков без i, $f(S \cup i)$ - вывод модели при включении признака i в подмножество S, а $f(S)$ - вывод модели без включения признака i в подмножество S.

Для вычисления значения Шепли формула учитывает все возможные комбинации признаков и вычисляет разницу в прогнозах при включении или исключении признака i. Термин $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ нормализует вклад, учитывая все возможные упорядочивания признаков.

Назначая значения Шепли каждому признаку, SHAP предоставляет всестороннее объяснение для отдельных прогнозов. Признаки с более высокими абсолютными значениями Шепли имеют большее влияние на вывод модели, как положительное, так и отрицательное. Это позволяет пользователям понять относительную важность различных признаков в процессе принятия решений модели.

SHAP обладает несколькими преимуществами перед другими методами определения важности признаков. Он обеспечивает справедливость и последовательность, следуя принципам кооперативной теории игр. Кроме того, SHAP является модельно-независимым, что означает его применимость для любой модели машинного обучения независимо от ее архитектуры или алгоритма обучения.

В заключение, SHapley Additive exPlanations (SHAP) использует концепцию значений Шепли для предоставления понятной и интерпретируемой меры важности признаков. Путем назначения численных вкладов каждому признаку SHAP обеспечивает всестороннее объяснение отдельных прогнозов, сделанных моделями машинного обучения. Этот метод способствует прозрачности, справедливости и доверию к системам искусственного интеллекта и помогает определить влиятельные признаки, определяющие решения модели.

5.2.3 Анализ факторов влияния

Важным аспектом для понимания прогнозируемых результатов является связь между полученными значениями SHAP (Shapley Additive Explanations) и соответствующими значениями признаков. Визуализация представлена в виде графика, где на левой вертикальной оси отображаются названия переменных, а правая цветовая шкала определяет значения переменных по градиенту от малых до больших значений, с переходом цвета от синего к красному. На горизонтальной оси представлены значения SHAP, которые служат индикаторами важности или вклада каждой переменной в прогнозируемый результат (показано на рисунке 5.2).

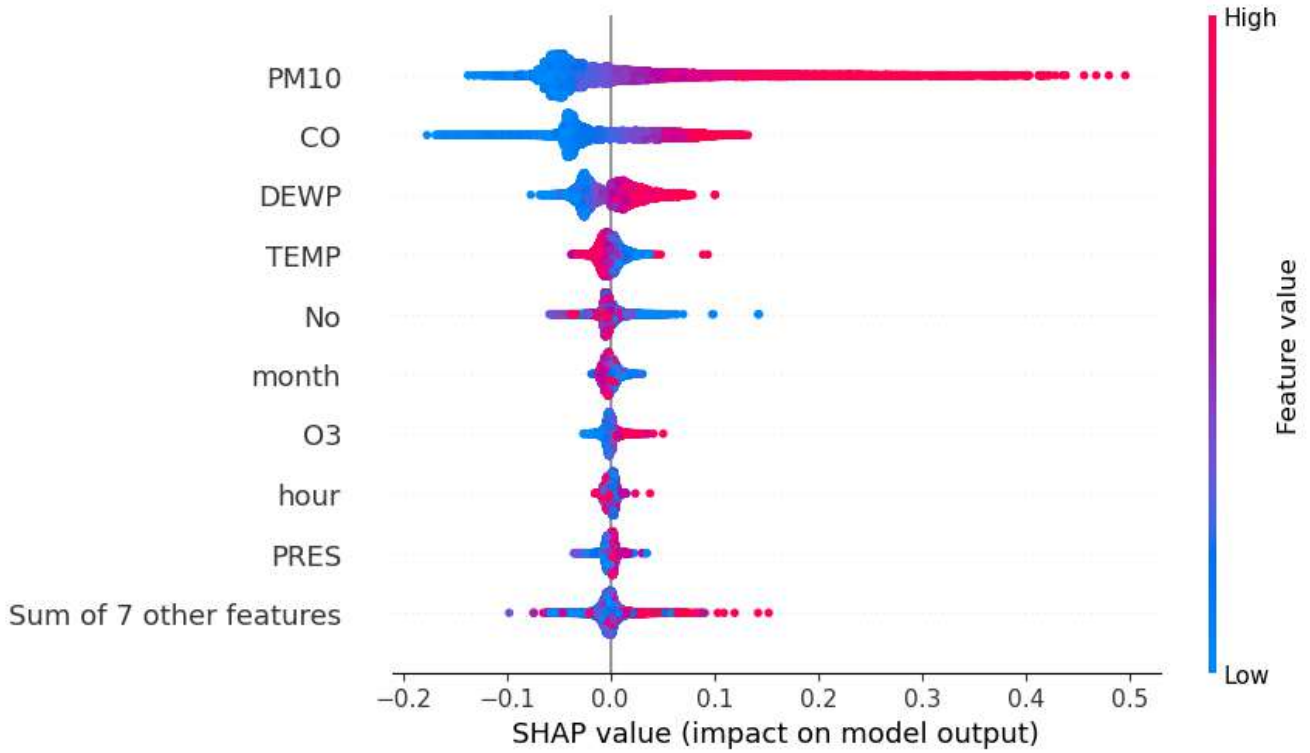


Рис. 5.2: Связь между полученными значениями SHAP и значениями признаков.

Интерпретация этих значений SHAP имеет важное значение для понимания влияния изменений значений признаков на общий прогноз. Когда значение SHAP для определенного признака положительно, увеличение его величины указывает на положительный эффект на прогнозируемый результат вследствие изменения соответствующего значения признака. Напротив, когда значение SHAP отрицательно, увеличение абсолютной величины означает негативное влияние на прогнозируемый результат из-за изменений значения связанного признака.

Этот анализ позволяет глубже понять относительную значимость каждого признака влияющего на прогнозируемый результат. Изучая связь между значениями SHAP и значениями признаков, можно получить представление о направлении и величине их влияния на окончательный прогноз. Такие знания поддерживают процесс принятия решений и помогают выявить ключевые факторы, которые положительно или отрицательно влияют на прогнозируемый результат, способствуя разработке стратегий для достижения оптимальной производительности прогнозирования.

На представленном рисунке 5.2 проведен анализ связи между значениями признаков и соответствующими значениями SHAP(Shapley Additive Explanations), что дает ценные представления о влиянии различных переменных на прогнозируемые значения PM2.5. Цветовые точки данных на графике представляют различные значения признаков, красный цвет указывает на увеличение, а синий - на уменьшение.

Наблюдение за поведением переменной 'PM10' заслуживает внимания: с увеличением ее значения (обозначенного изменением точки данных в красный цвет), также возрастает и соответствующее значение SHAP. Это положительное значение SHAP свидетельствует о том, что увеличение 'PM10' приводит к соответствующему увеличению прогнозируемого значения PM2.5. Напротив, когда значение 'PM10' уменьшается (представлено изменением точки данных в синий цвет), значение SHAP становится отрицательным, указывая на отрицательное влияние 'PM10' на прогнозируемое значение PM2.5. В этом случае уменьшение 'PM10' приводит к уменьшению прогнозируемого значения PM2.5.

Подобные закономерности можно наблюдать для переменных 'CO' и 'DEWP'. Увеличение значений этих переменных (обозначено красными точками данных) соответствует увеличению прогнозируемых значений PM2.5. С другой стороны, уменьшение значений признаков (представленное синими точками данных) приводит к уменьшению прогнозируемых значений PM2.5.

Кроме того, анализ длины охвата точек данных предоставляет представления о относительной важности переменных. Четко видно, что 'PM10' оказывает наибольшее влияние, так как имеет наибольший охват точек данных. За 'PM10' следует переменная 'CO', которая обладает незначительно меньшим, но все равно заметным влиянием, и тесно за ней следует 'DEWP'. Порядок определен на основе абсолютных значений соответствующих значений SHAP.

Изучая эти наблюдения, можно лучше понять специфический вклад и относительную важность различных переменных в формировании прогнозируемых значений PM2.5. Это помогает выявить ключевые факторы, определяющие результаты прогнозирования, и может быть полезно для принятия решений в области управления качеством воздуха и стратегий контроля загрязнения.

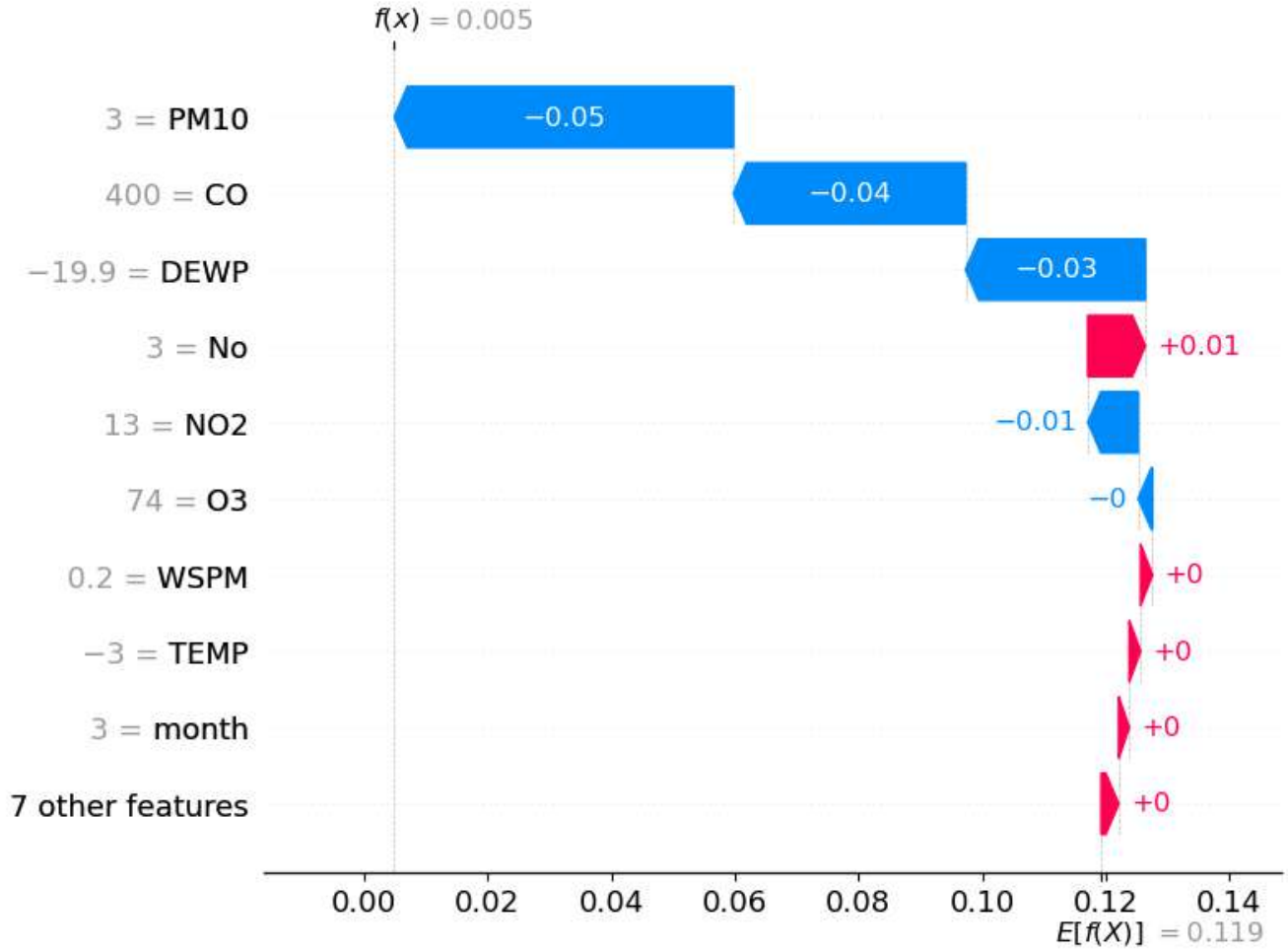


Рис. 5.3: Пример результатов локального объяснения.

Более подробное исследование влияющих факторов можно провести, анализируя отдельные точки наблюдения. Для этого можно использовать график водопада, который проясняет влияние модели на прогнозируемые результаты для конкретной точки данных.

График водопада визуально представляет величину вкладов, сделанных каждым признаком в прогнозы модели. График выполняет несколько ключевых функций:

- Интерпретация конкретных точек данных: Графики водопада облегчают понимание того, как модель вносит вклад в результаты прогнозирования

для конкретной точки данных. Они делают это, отображая эффекты различных значений признаков на вывод прогноза, позволяя понять влияние каждого признака.

- Оценка важности признаков: Каждая полоса на графике соответствует признаку и показывает его положительное или отрицательное влияние на результаты прогнозирования. Длина полосы представляет собой относительный вклад признака, при этом более длинные полосы указывают на большее влияние на прогнозирование.

- Исследование взаимодействий признаков: График водопада также позволяет исследовать взаимодействия между различными признаками. Наблюдая за вертикальными изменениями каждого признака на графике, можно определить, как эти признаки влияют друг на друга и как их совместные эффекты усиливают или уменьшают результаты прогнозирования.

Анализ графика водопада позволяет получить ценные представления о том, как модель делает прогнозы на основе значений входных признаков. Это не только улучшает наше понимание интерпретируемости модели, но и помогает принимать решения или корректировать значения признаков для конкретных точек данных.

Для иллюстрации этой методологии на рисунке 5.3 представлен пример одного наблюдения (2 октября 2016 г., 6:00). Он иллюстрирует локализованный анализ переменных, на которых полагается прогностическая модель, в отличие от глобальной перспективы, представленной на рисунке 5.2. В этом случае прогностическая модель дала фиксированный вывод 0.119 в 6:00. Следует отметить, что "PM10"оказывает негативное влияние в размере -0.05 на прогноз, "CO"имеет отрицательное влияние в размере -0.04, и другие переменные следуют подобному шаблону. В результате окончательное прогнозируемое значение составило 0.005. Этот анализ подчеркивает ценность изучения отдельных точек данных, обогащая наше понимание функционирования модели и информируя последующие решения или корректировки с учетом конкретных значений признаков.

Кроме того, можно создать информативную гистограмму, которая демонстрирует общую важность признаков, конкретно отображая средний вклад каждого признака в прогнозируемые результаты модели.

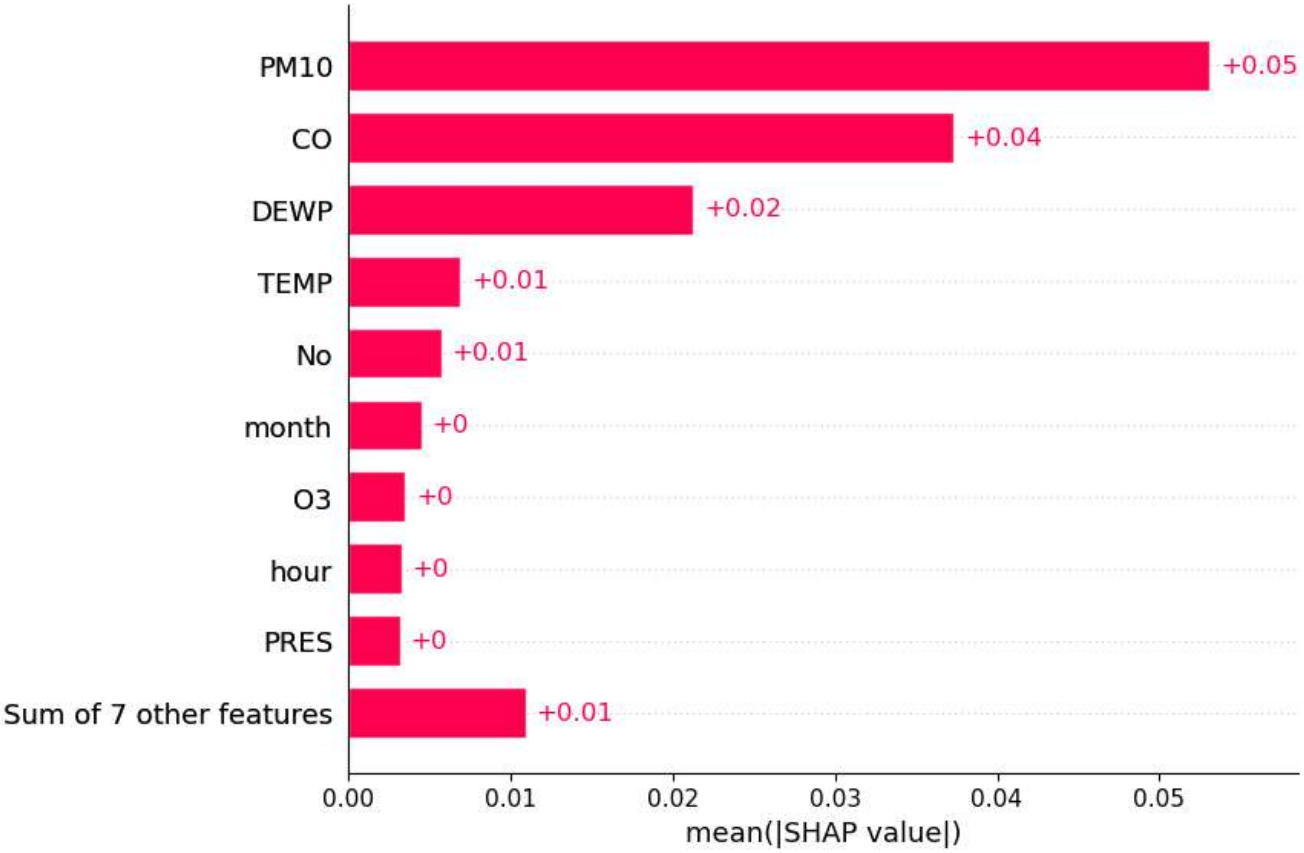


Рис. 5.4: Пример глобальных результатов объяснения.

Гистограмма выполняет несколько важных функций:

- Ранжирование важности признаков: Путем сортировки признаков на основе их среднего вклада в результаты прогнозирования, гистограмма позволяет определить наиболее влиятельные признаки. Ранжирование помогает выявить, какие признаки оказывают более существенное влияние на прогнозы модели.

- Сравнение относительной важности признаков: Высота каждой полосы гистограммы представляет собой относительный вклад соответствующего признака в результат прогноза. Более высокие полосы указывают на большее влияние на прогноз. Сравнение высот различных полос позволяет оценить относительную важность каждого признака.

- Руководство для выбора и инженерии признаков: Данная диаграмма предоставляет ценное руководство для процессов выбора и инженерии признаков. Анализ гистограммы позволяет выявить ключевые признаки, которые существенно влияют на прогнозируемый результат модели. Эта информа-

ция является инструментальной для принятия обоснованных решений относительно выбора признаков или выполнения задач по инженерии признаков с целью улучшения производительности и объясняющих возможностей модели.

Полученные графики-гистограммы предоставляют лаконичный обзор важности признаков, предлагая значимые выводы для интерпретации модели. Они помогают понять, как модель делает прогнозы на основе признаков, что в свою очередь направляет принятие решений относительно выбора и инженерии признаков.

В конечном итоге, этот подход применяется к отдельным значениям всех наблюдений. Абсолютные значения этих вкладов определяются, суммируются, а затем усредняются, чтобы получить всестороннее представление о ранжировании переменных (см. рисунок 5.4). В целом признак "PM10"демонстрирует выдающееся ранжирование вклада, за ним следуют "CO"и "DEWP". Эти данные позволяют более подробно рассмотреть значимые признаки, оказывающие влияние на прогнозные результаты модели.

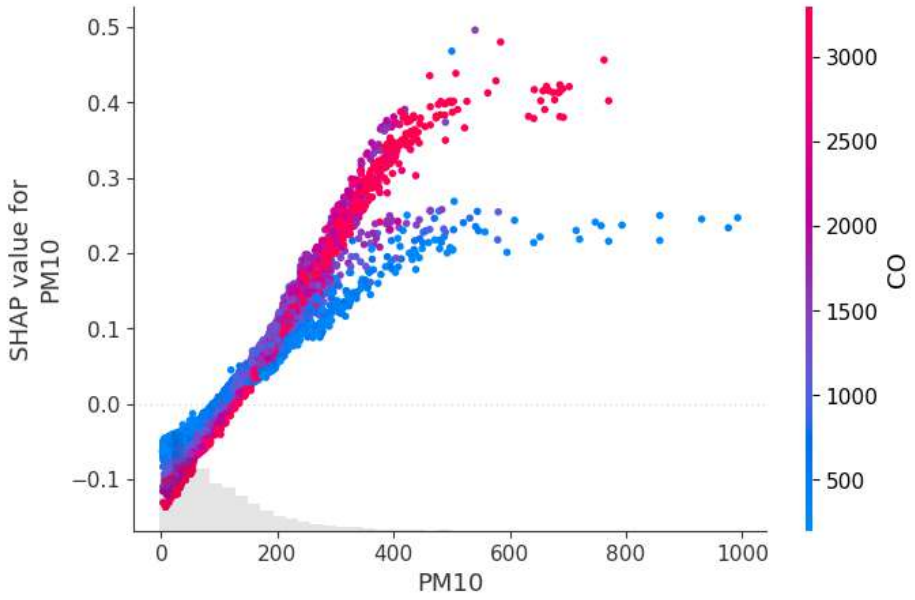


Рис. 5.5: Взаимодействие 1 непрерывных переменных на результаты прогнозирования.

Предварительный анализ факторов влияния играет важную роль в укреплении доверия пользователей к работе модели. Полученные результаты соответствуют естественным законам, указывая на то, что модель эффективно улавливает знания из исторических данных. Соответствие этим установленным

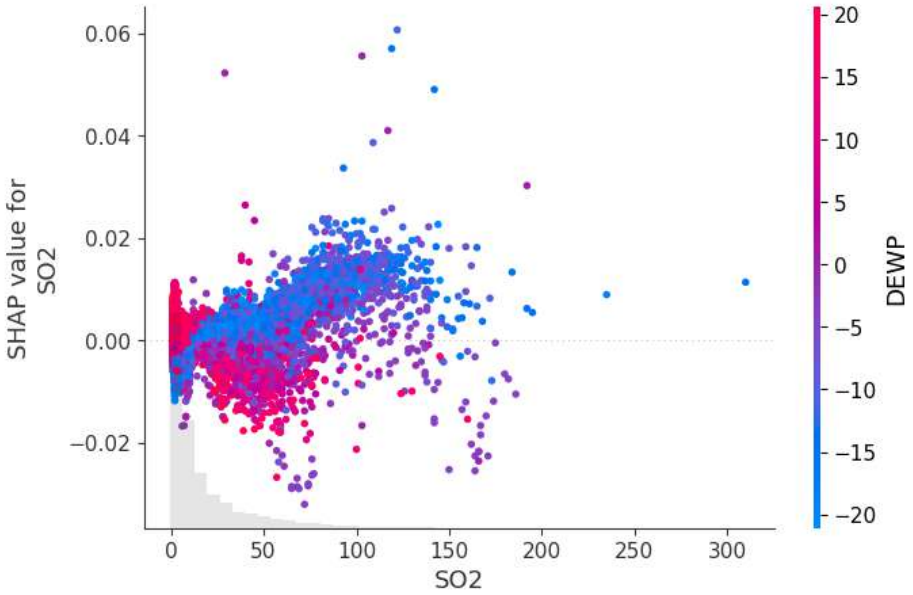


Рис. 5.6: Взаимодействие 2 непрерывных переменных на результаты прогнозирования.

принципам повышает достоверность способности модели генерировать надежные прогнозы.

Использование инструментов SHAP не только упрощает осмотр отдельных переменных, но также предоставляет ценные представления о их взаимодействии. Эта функция не только повышает доверие пользователя, но также позволяет провести более всестороннее исследование факторов, влияющих на PM2.5. В частности, SHAP позволяет изучать, какое влияние оказывает одна переменная на прогнозируемый результат при изменении другой переменной. Эти взаимодействия наглядно представлены на соответствующих графиках.

На графике горизонтальная ось представляет значения переменной, а распределение значений переменных иллюстрируется серым оттенком вдоль оси. На левой вертикальной оси отображается значение SHAP для каждой переменной, представляющее ее важность или вклад в результат прогноза. Переменная, которая проявляет самое сильное взаимодействие с заданной переменной, вертикально представлена справа.

На рисунке наблюдается поведение переменной "PM10": с увеличением ее значения происходит постепенный рост соответствующего значения SHAP, и даже переход от отрицательного к положительному значению. Это развитие указывает на увеличение влияния на PM2.5. Следует отметить, что "PM10"и "CO"обнаруживают заметное взаимодействие, при котором увеличение "CO" усиливает влияние "PM10"на результаты прогнозирования. Это взаимодействие

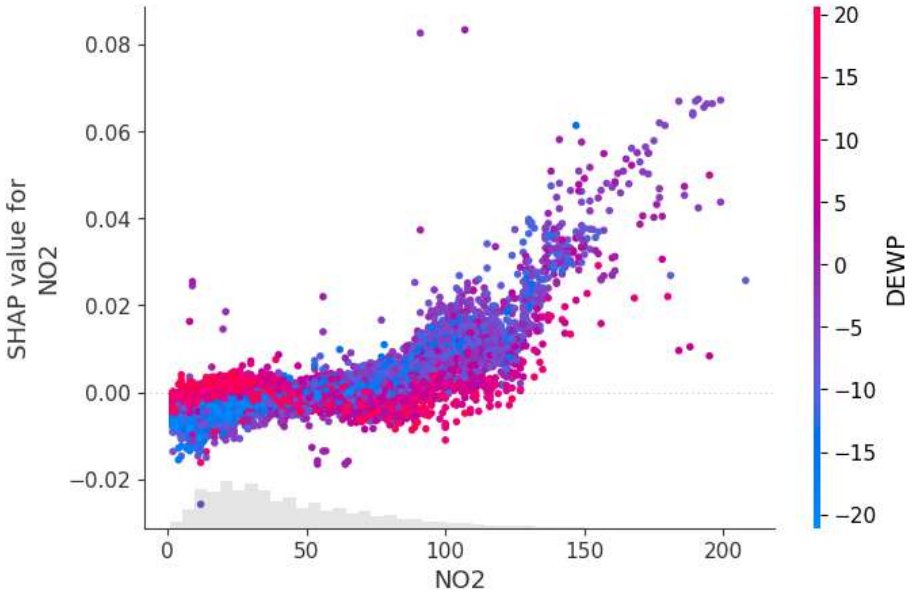


Рис. 5.7: Взаимодействие 3 непрерывных переменных на результаты прогнозирования.

наглядно выражено чередованием красных и синих точек данных на графике.

Кроме того, рисунок демонстрирует абсолютную важность "PM10"и ее различные степени взаимодействия с другими факторами окружающей среды (например, "DEWP"). По мере увеличения значений этих факторов окружающей среды их влияние на предикторные переменные изменяется, что подтверждают представленные тенденции.

Эти выводы позволяют получить более глубокое понимание взаимосвязей между переменными и их влияния на прогнозируемые значения PM2.5, обеспечивая более всестороннее понимание основных отношений.

5.3 Заключение к главе 5

В данной главе был проведен всесторонний анализ сравнения и оценки производительности трех ведущих моделей: XGBoost, LightGBM и CatBoost для прогнозирования временных рядов. Основной целью было выявить модель, которая обеспечивает наиболее точные и надежные прогнозы значений PM2.5. В этой связи использовались два основных показателя оценки, а именно значение R^2 и среднеквадратичная ошибка (MSE), чтобы оценить соответствие модели и точность прогноза каждой модели.

Начиная с оценки пригодности модели, значения R^2 предоставляют информацию о доле дисперсии данных PM2.5, улавливаемой моделями. Более высокое

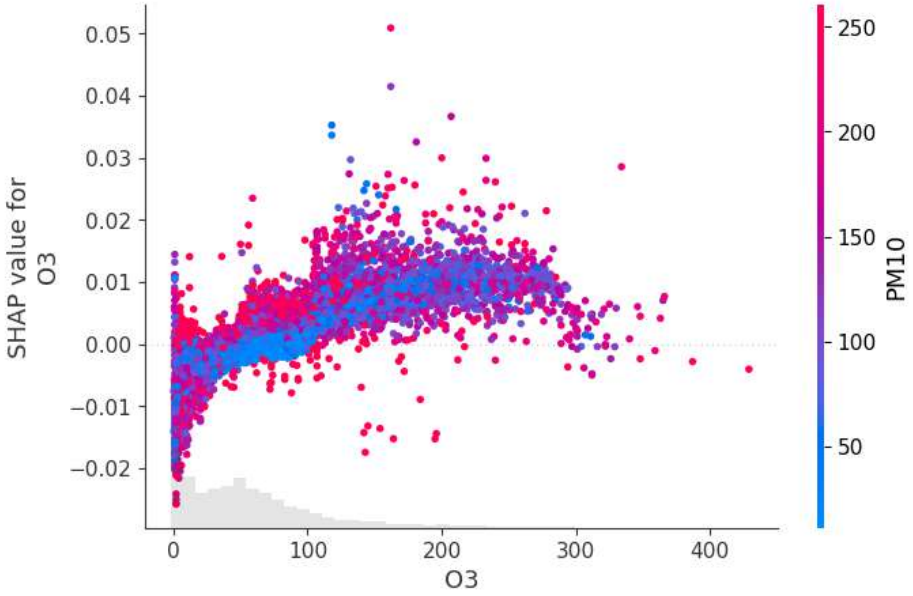


Рис. 5.8: Взаимодействие 4 непрерывных переменных на результаты прогнозирования.

значение R^2 указывает на более надежную модель, которая успешно улавливает значительную часть вариабельности в целевой переменной. Из полученных результатов было отмечено, что LightGBM превзошел XGBoost и CatBoost, продемонстрировав самое высокое значение R^2 равное 0.9134. Это подтверждает эффективность LightGBM в точном отображении сложных взаимосвязей в данных и его способность объяснить значительную долю изменчивости в наблюдениях PM2.5. Следует отметить, что XGBoost и CatBoost также продемонстрировали похвальную производительность, хотя с немного более низкими значениями R^2 равными 0.8551 и 0.9112 соответственно.

Переходя к оценке ошибок прогноза с использованием MSE, широко признанного показателя, целью было количественно оценить среднеквадратичное отклонение между прогнозируемыми и фактическими значениями PM2.5. Более низкое значение MSE соответствует высокой точности прогнозирования, указывая на то, что прогнозы модели минимально отклоняются от реальных значений. В контексте этого анализа LightGBM снова продемонстрировал свою мощь, достигнув самого низкого значения MSE - 0.0013. Этот результат указывает на то, что прогнозы LightGBM минимально расходятся с истинными значениями PM2.5 в среднем, укрепляя его надежность и точность. Аналогично, XGBoost и CatBoost продемонстрировали относительно низкие значения MSE - 0.0018 и 0.0014 соответственно, подтверждая свою способность генерировать точные прогнозы.

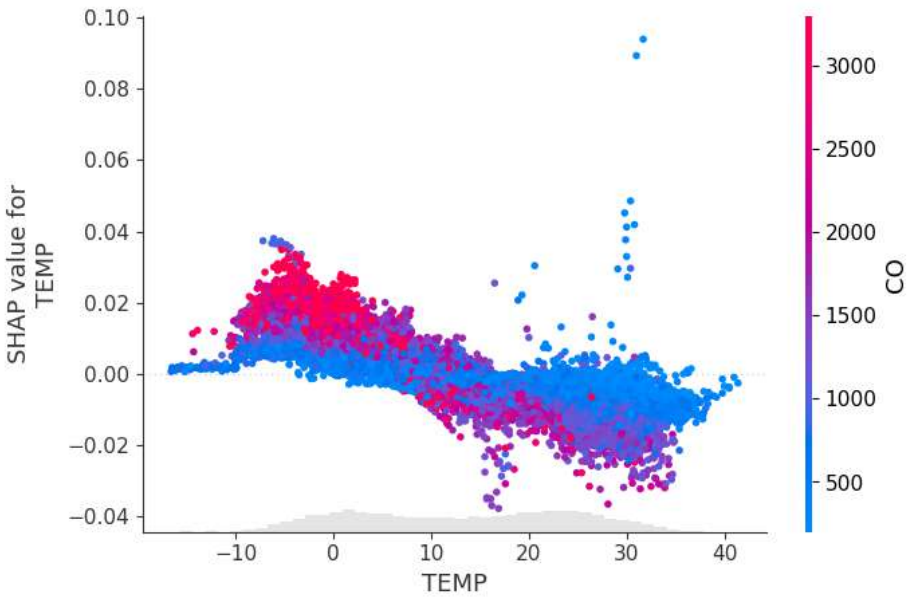


Рис. 5.9: Взаимодействие 5 непрерывных переменных на результаты прогнозирования.

Помимо прогностической производительности, важно также рассмотреть вычислительную эффективность моделей, особенно в случаях, когда требуются временно-чувствительные приложения. Время обучения, представленное столбцом TimeSpent, предоставляет ценную информацию о вычислительных затратах, связанных с каждой моделью. Эффективная модель должна иметь сокращенное время обучения без ущерба точности прогноза. В этом анализе LightGBM оказался наиболее вычислительно эффективным вариантом, требующим всего 140 мс для завершения процесса обучения. По сравнению с ним, XGBoost потребовал 756 мс, а CatBoost - 2.96 секунды. Это значительное различие подчеркивает преимущество LightGBM с точки зрения вычислительной эффективности, делая его идеальным для приложений реального времени или с ограниченными ресурсами.

Для дополнения оценки прогностических моделей были введены методы Explainable AI в сочетании со значениями SHAPly для более глубокого анализа факторов влияния. Исследуя отдельные переменные, такие как 'PM10', 'CO' и 'DEWP', были получены ценные представления о их воздействии на прогнозируемые значения PM2.5. Анализ значений SHAP позволил выявить направление и величину влияния этих переменных, предоставив более глубокое понимание их взаимосвязей и взаимодействий в контексте прогностических моделей. Следует отметить, что увеличение значения 'PM10' приводит к соответствующему повышению прогнозируемого значения PM2.5, что подтверждается

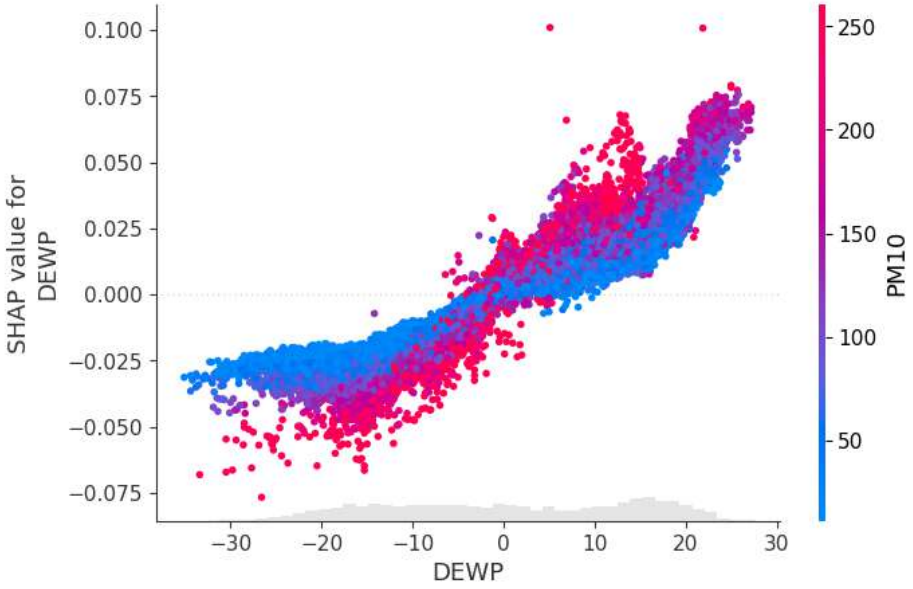


Рис. 5.10: Взаимодействие 6 непрерывных переменных на результаты прогнозирования.

положительными значениями SHAP. Напротив, уменьшение 'PM10' приводит к отрицательным значениям SHAP, указывающим на снижение прогнозируемого значения PM2.5. Подобные закономерности наблюдались для переменных 'CO' и 'DEWP', дополнительно подчеркивая их вклад в прогностические результаты.

Более того, анализ покрытия точек данных, как это представлено в анализе, предоставил ценную информацию о относительной важности переменных. 'PM10' выделяется как наиболее влиятельный признак благодаря самому длительному покрытию точками данных. За 'PM10' следует признак 'CO', проявляющий немного меньшее, но все же заметное влияние, и тесно за ним следует 'DEWP'. Это ранжирование было установлено на основе абсолютных значений соответствующих значений SHAP, предоставляя всестороннюю перспективу вклада переменных в прогностические модели.

Для подтверждения практической значимости и применимости результатов были представлены конкретные примеры, облегчающие локализованный анализ переменных и их влияния на прогнозируемые значения PM2.5. Исследование отдельных наблюдений позволяет лучше понять работу моделей, позволяя принимать обоснованные решения и корректировки, касающиеся конкретных значений признаков.

В заключение, в этой главе было проведено тщательное сравнение и анализ трех популярных моделей, а именно XGBoost, LightGBM и CatBoost, для прогнозирования временных рядов значений PM2.5. Через оценку значений R^2,

MSE и времени обучения LightGBM выделился как наиболее успешная модель, показав самое высокое значение R^2 и наименьшую MSE, а также демонстрирующий самое короткое время обучения. Другие две модели, XGBoost и CatBoost, продемонстрировали конкурентоспособную производительность, но оказались чуть ниже LightGBM по точности и вычислительной эффективности.

Кроме того, были введены методы объяснения AI, такие как значения SHAPly, для более глубокого понимания факторов влияния. Анализ отдельных переменных позволил получить представления о воздействии на прогнозируемые значения PM2.5. Предоставленные выводы улучшают интерпретируемость моделей и обеспечивают глубокое понимание взаимодействий и отношений между переменными.

Полученные результаты имеют практическую ценность для различных областей, где точные прогнозы временных рядов являются важными, таких как мониторинг качества воздуха и управление окружающей средой. Превосходство LightGBM по точности и вычислительной эффективности делает его идеальным выбором для приложений, требующих надежных и быстрых прогнозов значений PM2.5. Полученные при анализе факторы влияния помогают выявить ключевые переменные и их взаимодействия, что позволяет заинтересованным сторонам принимать обоснованные решения на основе всестороннего понимания динамики процессов.

В целом данная глава вносит свой вклад в область прогнозирования временных рядов, представляя тщательное сравнение и анализ трех популярных моделей, оценивая их производительность и используя методы Explainable AI для глубокого понимания влияющих факторов. Комбинация академической строгости и логического мышления обеспечивает надежные выводы из данного исследования, которые могут оказать влияние на будущие исследования или практическую реализацию в связанных областях.

Заключение

В заключении эта диссертация представляет всесторонний анализ различных аспектов, связанных с привлекательностью инвестиций, анализом качества воздуха и прогнозированием временных рядов значений PM2.5. Каждая глава фокусировалась на конкретных исследовательских целях и результаты, внося ценные представления в соответствующие области изучения.

Глава 1 изучала факторы, влияющие на инвестиционную привлекательность в Китае и странах АСЕАН-5. Через тщательный анализ было выявлено несколько ключевых факторов, таких как доход на душу населения, основные активы, строительные работы и мировая экономическая ситуация. Значимость этих факторов меняется в разных регионах и экономических системах, что подчеркивает важность учета региональных особенностей при оценке инвестиционных паттернов. Эти результаты имеют практическое значение для выработки политики и инвесторов, позволяя им принимать обоснованные решения и разрабатывать эффективные стратегии для стимулирования экономического роста и развития в этих регионах.

Глава 2 углубила исследование факторов, влияющих на инвестиционную привлекательность, используя метод множественной регрессии. Исследование подтвердило значимость переменных, таких как основные активы и средний доход на душу населения, для определения объема инвестиций. При учете региональных экономических особенностей и включении этих факторов анализ предоставил нюансное понимание инвестиционных паттернов в различных кластерах. Рекомендации, вытекающие из этих результатов, могут направлять политиков в стимулировании инвестиций, сосредоточиваясь на приобретении дорогостоящих основных активов в регионах с высокой привлекательностью и улучшении среднего дохода на душу населения в регионах с низкой привлекательностью. Кроме того, контроль производительности строительной отрасли и преодоление ее препятствий могут улучшить инвестиционную привлекательность. Однако

важно признать ограничения в виде немеряемых факторов и необходимость осторожности в обобщении результатов, что указывает на возможности для дальнейших исследований для расширения этих идей.

Переходя к главе 3, в этой главе с использованием пошаговой регрессии проанализирован индекс качества воздуха и получены несколько ключевых выводов. Устойчивое и значительное воздействие SO2 на уровень загрязнения воздуха подчеркивает неотложную необходимость принятия комплексных мер для снижения его выбросов и повышения эффективности контроля. Высокая концентрация SO2 может быть обусловлена широким использованием угля, промышленными процессами и выбросами транспортных средств. С другой стороны, относительно незначительное влияние CO и PM2.5 на качество воздуха свидетельствует о том, что их концентрации зависят от различных факторов, таких как метеорологические условия и конкретные источники выбросов. Включение NO2 как значимой переменной в регрессионные модели подчеркивает ее релевантность среди регрессоров. Эти выводы подчеркивают необходимость постоянных усилий по снижению выбросов SO2 и внедрению целевых мер для улучшения качества воздуха и общественного здоровья.

Глава 4 исследовала применение моделей машинного обучения для прогнозирования значений PM2.5. Через тщательный процесс оценки, который включал семь различных моделей (ANN, RNN, LSTM, GRU, Bi-RNN, Bi-LSTM и Bi-GRU), была оценена их производительность и прогностические возможности. Модель Bi-RNN выделилась как наиболее результативная модель, продемонстрировав самое высокое значение R^2 и наименьшую среднеквадратичную ошибку (MSE). Другие модели, такие как ANN, LSTM, GRU, Bi-LSTM и Bi-GRU, также показали конкурентоспособные прогностические возможности, хотя с разной степенью точности и вычислительных требований. Эти результаты подчеркивают потенциал моделей машинного обучения в точной оценке и управлении рисками, связанными с загрязнением воздуха, с использованием передовых методов анализа данных.

В заключительной главе были проведены сравнительный анализ и оценка трех важных моделей (XGBoost, LightGBM и CatBoost) для прогнозирования временных рядов значений PM2.5. В рамках исследования были использованы строгие оценочные метрики, такие как значения R^2 и MSE, для определения производительности моделей. LightGBM выделилась как наиболее эффектив-

ная модель, показав самое высокое значение R^2 и наименьшее MSE, при этом также продемонстрировав самое короткое время обучения. XGBoost и CatBoost также показали хорошие результаты, но немного уступили LightGBM по точности и вычислительной эффективности. Внедрение техник объяснимого искусственного интеллекта предоставило понимание влияющих факторов, улучшив интерпретируемость моделей и обеспечив надежные и точные прогнозы. Полученные выводы имеют практическое значение для областей, таких как мониторинг загрязнения воздуха и управление окружающей средой, что позволяет политикам и исследователям принимать обоснованные решения на основе всестороннего понимания базовых динамик.

В заключении данной диссертации был представлен логически обоснованный и последовательный анализ инвестиционной привлекательности, анализа качества воздуха и прогнозирования временных рядов значений PM2.5. Выводы, сделанные в каждой главе, отличаются академической строгостью, учитывая статистические показатели, оценочные метрики и применимость различных моделей к реальным сценариям. Полученные результаты вносят свой вклад в соответствующие области исследований, предлагая практические выводы для политиков, инвесторов и исследователей. Дальнейшие исследования могут продолжить эту работу, расширяя знания и улучшая стратегии в смежных областях, способствуя устойчивому экономическому росту и управлению окружающей средой.

Литература

[1] Methodological support of organizations implementing innovative activities investment attractiveness estimation / Nataliya S Plaskova, Natalia A Prodanova, Elena I Zatsarinnaya et al. // Journal of Advanced Research in Law and Economics. — 2017. — Vol. 8, no. 8 (30). — P. 2533–2539.

[2] Snieska Vytautas, Zykiene Ineta. City attractiveness for investment: characteristics and underlying factors // Procedia-Social and Behavioral Sciences. — 2015. — Vol. 213. — P. 48–54.

[3] Dorożyński Tomasz, Kuna-Marszałek Anetta. Investments attractiveness: The case of the Visegrad Group countries // Comparative Economic Research. Central and Eastern Europe. — 2016. — Vol. 19, no. 1. — P. 119–140.

[4] Dierkes Maik, Erner Carsten, Zeisberger Stefan. Investment horizon and the attractiveness of investment strategies: A behavioral approach // Journal of Banking & Finance. — 2010. — Vol. 34, no. 5.

[5] Investment attractiveness of small innovational business under the conditions of globalization and integration / Anna N Ermakova, Svetlana S Vaytsekhovskaya, Viktoria B Malitskaya, Natalya Prodanova // University of Piraeus. International Strategic Management Association. — 2016.

[6] Moskalenko Bogdan, Lyulyov Oleksii, Pimonenko Tetyana. The investment attractiveness of countries: Coupling between core dimensions // Forum scientiae oeconomia. — Vol. 10. — 2022. — P. 153–172.

[7] The driving factors of air quality index in China / Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang et al. // Journal of Cleaner Production. — 2018. — Vol. 197. — P. 1342–1351.

[8] Kumar Anikender, Goyal P. Forecasting of daily air quality index in Delhi // Science of The Total Environment. — 2011. — Vol. 409, no. 24. — P. 5517–5523.

[9] An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution / Qian Di, Heresh Amini, Liuhua Shi et al. // Environment International. — 2019. — Vol. 130. — P. 104909.

[10] Air quality and climate change: Designing new win-win policies for Europe / Michela Maione, David Fowler, Paul S. Monks et al. // Environmental Science & Policy. — 2016. — Vol. 65. — P. 48–57. — Multidisciplinary research findings in support to the EU air quality policy: experiences from the APPRAISAL, SEFIRA and ACCENT-Plus EU FP7 projects.

[11] Fann N., Risley D. The public health context for PM2.5 and ozone air quality trends // Air Qual Atmos Health. — 2013. — Vol. 6. — P. 1–11.

[12] Wang Kunlun, Yin Hongchun, Chen Yiwen. The effect of environmental regulation on air quality: A study of new ambient air quality standards in China // Journal of Cleaner Production. — 2019. — Vol. 215. — P. 268–279.

[13] Wang Shuxiao, Hao Jiming. Air quality management in China: Issues, challenges, and options // Journal of Environmental Sciences. — 2012. — Vol. 24, no. 1. — P. 2–13.

[14] Zaib Shah, Lu Jianjiang, Bilal Muhammad. Spatio-Temporal Characteristics of Air Quality Index (AQI) over Northwest China // Atmosphere. — 2022. — Vol. 13, no. 3.

[15] Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations / Hamed Karimian, Qi Li, Chunlin Wu et al. // Aerosol and Air Quality Research. — 2019. — Vol. 19, no. 6. — P. 1400–1410.

[16] A novel, fuzzy-based air quality index (FAQI) for air quality assessment / Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian et al. // Atmospheric Environment. — 2011. — Vol. 45, no. 12. — P. 2050–2059.

[17] Maria C. Mirabelli and Stefanie Ebelt and Scott A. Damon. Air Quality Index and air quality awareness among adults in the United States // Environmental Research. — 2020. — Vol. 183. — P. 109185.

[18] Suling Zhu and Xiuyuan Lian and Haixia Liu and Jianming Hu and Yuanyuan Wang and Jinxing Che. Daily air quality index forecasting with hybrid models: A case in China // Environmental Pollution. — 2017. — Vol. 231. — P. 1232–1244.

[19] Thuan-Quoc Thach and Hilda Tsang and Peihua Cao and Lai-Ming Ho. A novel method to construct an air quality index based on air pollution profiles // International Journal of Hygiene and Environmental Health. — 2018. — Vol. 221, no. 1. — P. 17–26.

[20] Hongmin Li and Jianzhou Wang and Ranran Li and Haiyan Lu. Novel analysis–forecast system based on multi-objective optimization for air quality index // Journal of Cleaner Production. — 2019. — Vol. 208. — P. 1365–1383.

[21] Setyaningsih Santi. Using Cluster Analysis Study to Examine the Successful Performance Entrepreneur in Indonesia // Procedia Economics and Finance. — 2012. — Vol. 4. — P. 286–298. — International Conference on Small and Medium Enterprises Development with a Theme ?Innovation and Sustainability in SME Development? (ICSMED 2012).

[22] Cluster analysis of the relationship between carbon dioxide emissions and economic growth / Wenli Li, Guangfei Yang, Xianneng Li et al. // Journal of Cleaner Production. — 2019. — Vol. 225. — P. 459–471. — URL: https://www.sciencedirect.com/science/article/pii/S0959652619309266.

[23] Monfort Mercedes, Cuestas Juan Carlos, Ordóñez Javier. Real convergence in Europe: A cluster analysis // Economic Modelling. — 2013. — Vol. 33. — P. 689–694.

[24] Wolfson Murray, Madjd-Sadjadi Zagros, James Patrick. Identifying National Types: A Cluster Analysis of Politics, Economics, and Conflict // Journal of Peace Research. — 2004. — Vol. 41, no. 5. — P. 607–623.

[25] A survey on ensemble learning / Xibin Dong, Zhiwen Yu, Wenming Cao et al. // Frontiers of Computer Science. — 2020. — Vol. 14. — P. 241–258.

[26] Dietterich Thomas G et al. Ensemble learning // The handbook of brain theory and neural networks. — 2002. — Vol. 2, no. 1. — P. 110–125.

[27] Sagi Omer, Rokach Lior. Ensemble learning: A survey // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2018. — Vol. 8, no. 4. — P. e1249.

[28] Zhou Zhi-Hua, Zhou Zhi-Hua. Ensemble learning. — Springer, 2021.

[29] Zhang G.Peter. Time series forecasting using a hybrid ARIMA and neural network model // Neurocomputing. — 2003. — Vol. 50. — P. 159–175.

[30] Palani Sundarambal, Liong Shie-Yui, Tkalich Pavel. An ANN application for water quality forecasting // Marine Pollution Bulletin. — 2008. — Vol. 56, no. 9. — P. 1586–1597.

[31] Sherstinsky Alex. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network // Physica D: Nonlinear Phenomena. — 2020. — Vol. 404. — P. 132306.

[32] Dey Rahul, Salem Fathi M. Gate-variants of Gated Recurrent Unit (GRU) neural networks // 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). — 2017. — P. 1597–1600.

[33] Manaswi Navin Kumar. RNN and LSTM // Deep Learning with Applications Using Python : Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras. — Berkeley, CA : Apress, 2018. — P. 115–126.

[34] De Gooijer Jan G., Hyndman Rob J. 25 years of time series forecasting // International Journal of Forecasting. — 2006. — Vol. 22, no. 3. — P. 443–473. — Twenty five years of forecasting.

[35] Deep Learning for Time Series Forecasting: A Survey / José F. Torres, Dalil Hadjout, Abderrazak Sebaa et al. // Big Data. — 2021. — Vol. 9, no. 1. — P. 3–21.

[36] Sezer Omer Berat, Gudelek Mehmet Ugur, Ozbayoglu Ahmet Murat. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019 // Applied Soft Computing. — 2020. — Vol. 90. — P. 106181.

[37] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research // Journal of Pharmaceutical and Biomedical Analysis. — 2000. — Vol. 22, no. 5. — P. 717–727.

[38] Lim Bryan, Zohren Stefan. Time-series forecasting with deep learning: a survey // Philosophical Transactions of the Royal Society A. — 2021. — Vol. 379, no. 2194. — P. 20200209.

[39] An empirical comparison of machine learning models for time series forecasting / Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, Hisham El-Shishiny // Econometric reviews. — 2010. — Vol. 29, no. 5-6. — P. 594–621.

[40] Kolarik Thomas, Rudorfer Gottfried. Time series forecasting using neural networks // ACM Sigapl Apl Quote Quad. — 1994. — Vol. 25, no. 1. — P. 86–94.

[41] Yan Weizhong. Toward automatic time-series forecasting using neural networks // IEEE transactions on neural networks and learning systems. — 2012. — Vol. 23, no. 7. — P. 1028–1039.

[42] Sagi Omer, Rokach Lior. Approximating XGBoost with an interpretable decision tree // Information Sciences. — 2021. — Vol. 572. — P. 522–542.

[43] Experimenting XGBoost algorithm for prediction and classification of different datasets / Santhanam Ramraj, Nishant Uzir, R Sunil, Shatadeep Banerjee // International Journal of Control Theory and Applications. — 2016. — Vol. 9, no. 40. — P. 651–662.

[44] Chen Tianqi, Guestrin Carlos. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — P. 785–794.

[45] Xgboost: extreme gradient boosting / Tianqi Chen, Tong He, Michael Benesty et al. // R package version 0.4-2. — 2015. — Vol. 1, no. 4. — P. 1–4.

[46] Ensemble learning for data stream analysis: A survey / Bartosz Krawczyk, Leandro L Minku, João Gama et al. // Information Fusion. — 2017. — Vol. 37. — P. 132–156.

[47] Fryer Daniel, Strümke Inga, Nguyen Hien. Shapley values for feature selection: The good, the bad, and the axioms // Ieee Access. — 2021. — Vol. 9. — P. 144352–144360.

[48] Lightgbm: A highly efficient gradient boosting decision tree / Guolin Ke, Qi Meng, Thomas Finley et al. // Advances in neural information processing systems. — 2017. — Vol. 30.

[49] Al Daoud Essam. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset // International Journal of Computer and Information Engineering. — 2019. — Vol. 13, no. 1. — P. 6–10.

[50] Hancock John T, Khoshgoftaar Taghi M. CatBoost for big data: an interdisciplinary review // Journal of big data. — 2020. — Vol. 7, no. 1. — P. 1–45.

[51] CatBoost: unbiased boosting with categorical features / Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev et al. // Advances in neural information processing systems. — 2018. — Vol. 31.

[52] Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions / Guomin Huang, Lifeng Wu, Xin Ma et al. // Journal of Hydrology. — 2019. — Vol. 574. — P. 1029–1041.

[53] A communication-efficient parallel algorithm for decision tree / Qi Meng, Guolin Ke, Taifeng Wang et al. // Advances in Neural Information Processing Systems. — 2016. — Vol. 29.

[54] Quickly boosting decision trees–pruning underachieving features early / Ron Appel, Thomas Fuchs, Piotr Dollár, Pietro Perona // International conference on machine learning / PMLR. — 2013. — P. 594–602.

[55] Pan Bingyue. Application of XGBoost algorithm in hourly PM2. 5 concentration prediction // IOP conference series: earth and environmental science / IOP publishing. — Vol. 113. — 2018. — P. 012127.

[56] A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting / Yun Ju, Guangyu Sun, Quanhe Chen et al. // Ieee Access. — 2019. — Vol. 7. — P. 28309–28318.

[57] Dorogush Anna Veronika, Ershov Vasily, Gulin Andrey. CatBoost: gradient boosting with categorical features support // arXiv preprint arXiv:1810.11363. — 2018.

[58] Marcílio Wilson E, Eler Danilo M. From explanations to feature selection: assessing SHAP values as feature selection mechanism // 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI) / Ieee. — 2020. — P. 340–347.

[59] What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values / Yuan Meng, Nianhua Yang, Zhilin Qian, Gaoyu Zhang // Journal of Theoretical and Applied Electronic Commerce Research. — 2020. — Vol. 16, no. 3. — P. 466–490.

[60] Mokhtari Karim El, Higdon Ben Peachey, Başar Ayşe. Interpreting financial time series with SHAP values // Proceedings of the 29th annual international conference on computer science and software engineering. — 2019. — P. 166–172.

[61] Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods / Dong Wang, Sven Thunéll, Ulrika Lindberg et al. // Journal of Environmental Management. — 2022. — Vol. 301. — P. 113941.

[62] Sundararajan Mukund, Najmi Amir. The many Shapley values for model explanation // International conference on machine learning / PMLR. — 2020. — P. 9269–9278.

[63] Ghorbani Amirata, Zou James. Data shapley: Equitable valuation of data for machine learning // International conference on machine learning / PMLR. — 2019. — P. 2242–2251.

[64] Rodríguez-Pérez Raquel, Bajorath Jürgen. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions // Journal of computer-aided molecular design. — 2020. — Vol. 34. — P. 1013–1026.

[65] Winter Eyal. The shapley value // Handbook of game theory with economic applications. — 2002. — Vol. 3. — P. 2025–2054.

[66] Roth Alvin E. The Shapley value: essays in honor of Lloyd S. Shapley. — Cambridge University Press, 1988.

[67] Monderer Dov, Samet Dov. Variations on the Shapley value // Handbook of game theory with economic applications. — 2002. — Vol. 3. — P. 2055–2076.

[68] Towards efficient data valuation based on the shapley value / Ruoxi Jia, David Dao, Boxin Wang et al. // The 22nd International Conference on Artificial Intelligence and Statistics / PMLR. — 2019. — P. 1167–1176.

[69] Algorithms to estimate Shapley value feature attributions / Hugh Chen, Ian C Covert, Scott M Lundberg, Su-In Lee // Nature Machine Intelligence. — 2023. — P. 1–12.

[70] Problems with Shapley-value-based explanations as feature importance measures / I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, Sorelle Friedler // International Conference on Machine Learning / PMLR. — 2020. — P. 5491–5500.

[71] Littlechild Stephen C, Owen Guillermo. A simple expression for the Shapley value in a special case // Management Science. — 1973. — Vol. 20, no. 3. — P. 370–372.

[72] Kalai Ehud, Samet Dov. On weighted Shapley values // International journal of game theory. — 1987. — Vol. 16. — P. 205–222.

[73] Hart Sergiu, Mas-Colell Andreu et al. The potential of the Shapley value // the Shapley value. — 1988. — P. 127–137.

[74] The shapley value in machine learning / Benedek Rozemberczki, Lauren Watson, Péter Bayer et al. // arXiv preprint arXiv:2202.05594. — 2022.

[75] Merrick Luke, Taly Ankur. The explanation game: Explaining machine learning models using shapley values // Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4 / Springer. — 2020. — P. 17–38.

[76] Mohseni Sina, Zarei Niloofar, Ragan Eric D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems // ACM Transactions on Interactive Intelligent Systems (TiiS). — 2021. — Vol. 11, no. 3-4. — P. 1–45.

[77] Towards a rigorous evaluation of XAI methods on time series / Udo Schlegel, Hiba Arnout, Mennatallah El-Assady et al. // 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) / IEEE. — 2019. — P. 4197–4201.

[78] Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model / Basim Mahbooba, Mohan Timilsina, Radhya Sahal, Martin Serrano // Complexity. — 2021. — Vol. 2021. — P. 1–11.

[79] Liao Q Vera, Gruen Daniel, Miller Sarah. Questioning the AI: informing design practices for explainable AI user experiences // Proceedings of the 2020 CHI conference on human factors in computing systems. — 2020. — P. 1–15.

[80] Gunning David, Aha David. DARPA's explainable artificial intelligence (XAI) program // AI magazine. — 2019. — Vol. 40, no. 2. — P. 44–58.

[81] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI / Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser et al. // Information fusion. — 2020. — Vol. 58. — P. 82–115.

[82] Das Arun, Rad Paul. Opportunities and challenges in explainable artificial intelligence (xai): A survey // arXiv preprint arXiv:2006.11371. — 2020.

[83] Explainable artificial intelligence (XAI) in deep learning-based medical image analysis / Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, Max A Viergever // Medical Image Analysis. — 2022. — Vol. 79. — P. 102470.

[84] Adadi Amina, Berrada Mohammed. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) // IEEE access. — 2018. — Vol. 6. — P. 52138–52160.

[85] Tjoa Erico, Guan Cuntai. A survey on explainable artificial intelligence (xai): Toward medical xai // IEEE transactions on neural networks and learning systems. — 2020. — Vol. 32, no. 11. — P. 4793–4813.

[86] What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research / Markus Langer, Daniel Oster, Timo Speith et al. // Artificial Intelligence. — 2021. — Vol. 296. — P. 103473.

[87] Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review / Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz et al. // Applied Sciences. — 2021. — Vol. 11, no. 11. — P. 5088.

[88] Liao Q Vera, Varshney Kush R. Human-centered explainable ai (xai): From algorithms to user experiences // arXiv preprint arXiv:2110.10790. — 2021.

[89] Argumentative XAI: a survey / Kristijonas Čyras, Antonio Rago, Emanuele Albini et al. // arXiv preprint arXiv:2105.11266. — 2021.

[90] Saeed Waddah, Omlin Christian. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities // Knowledge-Based Systems. — 2023. — Vol. 263. — P. 110273.

[91] Evaluating XAI: A comparison of rule-based and example-based explanations / Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, Mark Neerincx // Artificial Intelligence. — 2021. — Vol. 291. — P. 103404.

[92] Wolf Christine T. Explainability scenarios: towards scenario-based XAI design // Proceedings of the 24th International Conference on Intelligent User Interfaces. — 2019. — P. 252–257.

[93] Páez Andrés. The pragmatic turn in explainable artificial intelligence (XAI) // Minds and Machines. — 2019. — Vol. 29, no. 3. — P. 441–459.

[94] Explainable artificial intelligence: a comprehensive review / Dang Minh, H Xiang Wang, Y Fen Li, Tan N Nguyen // Artificial Intelligence Review. — 2022. — P. 1–66.

[95] Explainable artificial intelligence (xai) on timeseries data: A survey / Thomas Rojat, Raphaël Puget, David Filliat et al. // arXiv preprint arXiv:2104.00950. — 2021.

[96] Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities / R Machlev, L Heistrene, M Perl et al. // Energy and AI. — 2022. — Vol. 9. — P. 100169.

[97] Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies / Eoin M Kenny, Courtney Ford, Molly Quinn, Mark T Keane // Artificial Intelligence. — 2021. — Vol. 294. — P. 103459.

[98] Dongfang Qi et al. Statistical analysis of investment attractiveness of China's regions // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2022. — Vol. 18, no. 1. — P. 188–194.

[99] Qi Dongfang, Bure Vladimir M. Research of investment attractiveness based on cluster analysis // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2023. — Vol. 19, no. 2. — P. 199–211.

[100] He Yang, Qi Dongfang, Bure Vladimir M. New application of multiple linear regression method-A case in China air quality // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2022. — Vol. 18, no. 4. — P. 515–526.

[101] He Yang, Qi Dongfang, Bure VM. Long-Term Air Quality Evaluation System Prediction In China Based On Multinomial Logistic Regression Method // GEOGRAPHY, ENVIRONMENT, SUSTAINABILITY. — 2024. — Vol. 16, no. 4. — P. 164–171.

[102] D. Qi V. M. Bure. Explanatory comparative analysis of time series forecasting algorithms for air quality prediction // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2024. — Vol. 20, no. 2.

[103] ShapTime: A General XAI Approach for Explainable Time Series Forecasting / Yuyi Zhang, Qiushi Sun, Dongfang Qi et al. // Intelligent Systems Conference / Springer. — 2023. — P. 659–673.