САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Чжан Юйи

Объяснимый искусственный интеллект в прогнозировании временных рядов

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 физико-математических наук, профессор
Петросян Ованес Леонович

Санкт-Петербург

2024

Оглавление

Введение

Актуальность темы диссертационного исследования

В последние годы модели искусственного интеллекта (ИИ) [1–3], такие, как ансамблевое обучение и глубокое обучение, продемонстрировали заметный успех в задаче прогнозирования временных рядов, особенно в долгосрочном прогнозировании [4–6]. Во многих конкурсах эти методы машинного обучения превосходили традиционные статистические методы по точности прогнозирования. Важным этапом стала победа LightGBM [7] в конкурсе M5 [8–11], которая привлекла всеобщее внимание к возможностям искусственного интеллекта. Профессор С. Макридакис, основатель серии M competition[1] и выдающийся специалист в области прогнозирования, недавно сравнил эффективность статистических алгоритмов машинного обучения и глубокого обучения [12]. Результаты экспериментов подтверждают значительный потенциал ИИ для долгосрочного прогнозирования временных рядов, превосходящий традиционные статистические подходы.

Competitions	Year	Winning solution	Types
Schneider competition	2018	LightGBM	AI
M4 competition	2020	ES-RNN	Statistics + AI
M5 competition	2021	LightGBM	AI
M6 competition	2022	Neural network	AI

Рис. 1: Решения, выигравшие в крупных конкурсах по прогнозированию временных

На рисунке 1 показаны решения, выигравшие в крупных конкурсах по прогнозированию временных рядов за последние годы, что еще раз подчеркивает преобладающую тенденцию к использованию методов,

[1]Домашняя страница конкурса M

основанных на ИИ.

Методы искусственного интеллекта доказали свою высокую эффективность при моделировании сложных моделей и составлении точных прогнозов, что делает их незаменимыми в различных отраслях, таких как энергетика [13, 14], здравоохранение [15, 16] и финансы [17, 18]. Однако присущая этим моделям непрозрачность представляет собой серьезную проблему. Отсутствие прозрачности в моделях «черного ящика» подрывает доверие и ограничивает их широкое признание. Эта проблема особенно актуальна в приложениях, где понимание процесса принятия решений так же важно, как и результаты прогнозирования. Чтобы решить эту проблему, растет интерес к объяснимому искусственному интеллекту (XAI) [19–21, 23, 26, 28–31]. XAI стремится сделать модели чёрного ящика более прозрачными и заслуживающими доверия, разъясняя их процессы принятия решений. Эта технология имеет существенные результаты для развития искусственного интеллекта и обеспечения его более безопасного применения в обществе [32–34]. Рисунок 2 иллюстрирует два различных типа объяснений в задаче прогнозирования временных рядов.

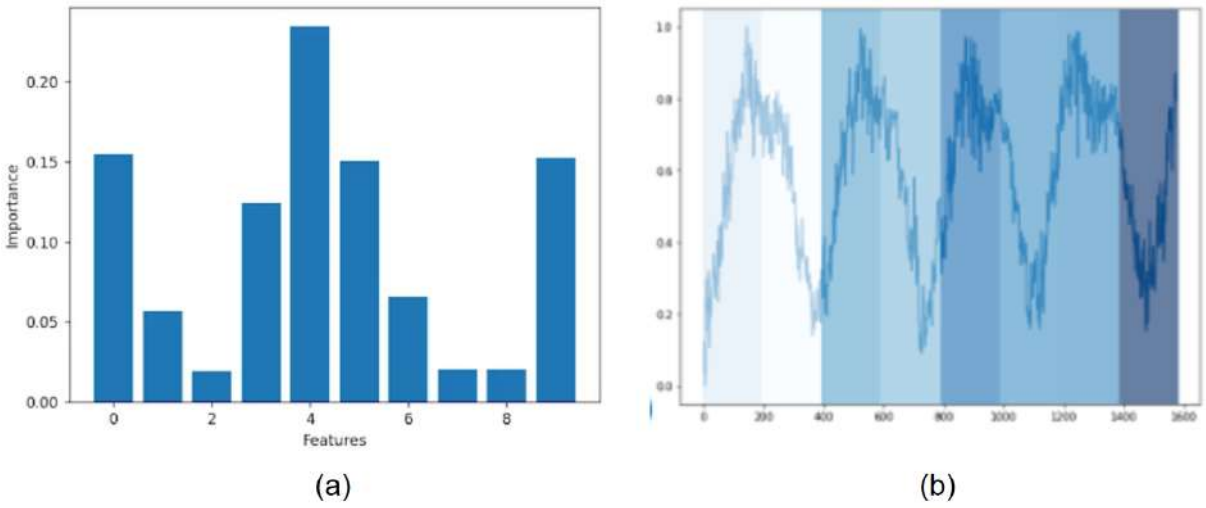


Рис. 2: Примеры объяснения результатов при прогнозировании временных рядов. (a) показывает, как каждая переменная влияет на целевую переменную; (b) показывает, как исторические данные влияют на целевую переменную.

Рисунок 2(a) представляет объяснение с точки зрения признаков, показывая, как каждая переменная влияет на целевую переменную [19–23, 28–31]. Этот тип объяснения особенно полезен для задач прогнозирования многомерных временных рядов, в которых задействовано

множество переменных. Напротив, Рисунок 2(b) дает объяснение с временной точки зрения. Он подчеркивает, как исторические данные влияют на результаты прогнозирования [26]. Этот подход применим как для одномерного, так и для многомерного прогнозирования временных рядов, что делает его универсальным для различных типов наборов данных. Эти визуализации подчеркивают важность объяснимости в моделях ИИ, используемых для прогнозирования временных рядов. Понимая, как признаки и исторические данные способствуют прогнозам, заинтересованные стороны могут получить ценную информацию, что приведет к более обоснованному принятию решений.

В последние годы быстрое развитие моделей ИИ в прогнозировании временных рядов подчеркнуло необходимость XAI в этой области. Понимание и объяснение этих сложных моделей имеют решающее значение для их принятия и надежности. Следовательно, исследование XAI в задаче прогнозирования временных рядов имеет большое значение [35–44]. Современную концепцию XAI можно проследить до 2007 года с исследованиями В. Щетинина [45] по объяснимости байесовских деревьев решений. С тех пор был достигнут значительный академический прогресс, включая многочисленные теоретические основы, алгоритмы [46–48] и библиотеки поддерживающего кода[2].

Алгоритмы XAI можно в целом разделить на **модель-специфические**, **модель-агностические** и **гибридные подходы**. Модель-специфические алгоритмы, такие как важность признаков (Feature Importance - FI) [49, 50], встроенная в алгоритмы повышения, адаптированы к определенным типам моделей черного ящика. С другой стороны, модельно-агностические алгоритмы, такие как SHAP [51] и LIME [52], предлагают универсальность, будучи применимыми ко всем типам моделей черного ящика, обеспечивая сильную обобщаемость. Гибридные подходы объединяют интерпретируемый модуль [53, 54] в модели прогнозирования, где выходные данные модуля объясняют общее поведение модели. Хотя эти алгоритмы дали значительные результаты исследований для задач классификации и регрессии, их применение к данным временных рядов остается ограниченным. Это ограничение подчеркивает несколько ключевых

[2]Официальная домашняя страница SHAP

проблем, которые необходимо решить:

Проблема 1: Как оценить эффективность результатов объяснений в прогнозировании временных рядов?

В отличие от таких задач, как регрессия, где для оценки доступны реальные целевые значения, область XAI сталкивается с трудностями в получении маркированных значений в практических приложениях. Этот недостаток доступных маркированных данных усложняет объективную оценку качества XAI [23, 55, 56]. Следовательно, в отличие от устоявшихся метрик в других областях, таких как средняя квадратическая ошибка (MSE) и площадь под кривой (AUC), в области XAI наблюдается дефицит общепринятых количественных метрик. Это отсутствие затрудняет стандартизацию оценки алгоритмов XAI.

Проблема 2: Как построить алгоритмы XAI, которые объясняют влияние исторических данных на все типы моделей искусственного интеллекта?

Недавние исследования подчеркнули растущую важность XAI в области прогнозирования временных рядов. Большая часть этих исследований была сосредоточена на алгоритмах, специфичных для конкретных моделей [35, 38, 39]. Как правило, эти исследования включают разработку новой модели прогнозирования временных рядов, а затем добавление интерпретируемых модулей, чтобы представить ее в качестве объяснимой модели. Хотя у этой методологии есть свои достоинства, мы утверждаем, что использование алгоритмов, не зависящих от модели, имеет решающее значение, особенно в инженерных приложениях. Возможность обобщения этих методов, независимо от используемой модели прогнозирования или метода объяснения, имеет большое значение.

Проблема 3: Как построить алгоритмы XAI, основанные на новых гибридных механизмах прогнозирования временных рядов?

В настоящее время гибридный метод предполагает включение механизма учета внимания в модели прогнозирования для улучшения их общей объяснимости [42–44]. Несмотря на то, что этот подход был применен при прогнозировании временных рядов, объяснимость самого механизма внимания по- прежнему остается предметом постоянных дискуссий.

Проблема 4: Как преобразовать результаты объяснения в реальную экономическую ценность?

Многие предыдущие исследования представляли концепцию объяснимости как новую идею, но часто упускали из виду ее практическое применение для решения текущих задач [35–44].

Эти алгоритмы, разработанные в данной работе, основаны на значении Шепли, концепции, введенной Л. Шепли в 1951 году. Значение Шепли [57] является основополагающим методом в теории кооперативных игр, предназначенным для справедливого распределения выплат среди участников. Этот принцип был эффективно адаптирован для объяснимого искусственного интеллекта (XAI), особенно в форме алгоритма SHAP, разработанного С. Лундбергом и Су-ин Ли. Хотя традиционный метод SHAP рассчитывает вклад переменных в выходные данные модели, он не способен предоставлять объяснения с временной точки зрения. Решение этой проблемы необходимо для повышения интерпретируемости прогнозирования временных рядов. Значительный вклад в алгоритмы атрибуции также внесли исследователи, такие как С. Бах, А. Чарнес, А. Датта, С. Липовецкий, М. Т. Рибейро, А. Шрикумар, Е. Струмбель и Х. П. Янг [52, 58–65]. Эти достижения в совокупности улучшают наше понимание и применение справедливых и разумных методов распределения в контексте XAI.

Цели и задачи диссертации

Эта работа направлена на развитие области XAI в прогнозировании временных рядов. Поскольку модели искусственного интеллекта с использованием черного ящика; по-прежнему широко распространены, наша цель - создать алгоритмы XAI, специально разработанные для этой области, и исследовать их применение в экономике. Достижение этой цели требует решения четырех основных задач в этой области. Во-первых, получение помеченных данных для XAI в практических приложениях остается сложной задачей, что приводит к отсутствию общепринятых количественных показателей. Цель данной работы - предложить новые оценочные показатели, которые объективно оценивают качество алгоритма XAI. Во- вторых, важно использовать подходы, не зависящие от модели, особенно в инженерных контекстах, где требуется общая применимость

различных моделей прогнозирования. Мы намерены разработать алгоритмы XAI, не зависящие от модели, для повышения гибкости и удобства использования. В-третьих, хотя в моделях обычно используются модули внимания, их интерпретируемость часто ставится под сомнение. В этом исследовании будут рассмотрены альтернативные механизмы для разработки гибридных подходов, что улучшит объяснимость моделей прогнозирования временных рядов. Наконец, многие исследования подчеркивают объяснимость, не затрагивая в полной мере ее конкретных целей или вклада в решение текущих задач в этой области. Цель данной работы - восполнить этот пробел путем проведения всесторонних поисковых исследований, применения алгоритмов XAI к реальным сценариям и разъяснения их преимуществ. Для достижения этих целей будут решены следующие задачи:

1. Определение количественных показателей оценки: Разработка общих показателей оценки на основе признанных теорем для выбора применимых алгоритмов XAI для различных моделей искусственного интеллекта. Эта основа позволит разрабатывать новые алгоритмы, основанные на наиболее подходящих методологиях XAI.

2. Разработка алгоритмов, не зависящих от модели: Создайте алгоритм, не зависящий от модели, для прогнозирования временных рядов, который может быть применен к любой модели прогнозирования. Алгоритм объяснит значимость исторических данных за разные периоды времени, что впоследствии повысит точность прогнозирования.

3. Механизм гибридного подхода: Внедрите комбинацию алгоритмов, не зависящих от модели, и алгоритмов, зависящих от конкретной модели, в качестве нового механизма для гибридных подходов. Это расширит технические возможности алгоритмов XAI, предоставив объяснения переменной важности и повысив точность прогнозирования.

4. Изучение экономической ценности: Изучите экономическое воздействие XAI, применив его к практическим вопросам, таким как онлайн-адаптация, повышение точности прогнозирования и

возможность анализа импакт-факторов. Посредством этих задач диссертация направлена на то, чтобы связать теоретические идеи с ощутимыми результатами в реальных сценариях, тем самым внося свой вклад как в академическом, так и в практическом плане в эту область.

Научная новизна

Проблема необъяснимости алгоритмов искусственного интеллекта существенно затрудняет их более широкое применение в прогнозировании временных рядов. В данной статье рассматривается эта проблема путем изучения методов XAI в задаче прогнозирования временных рядов - области, где алгоритмы искусственного интеллекта имеют разнообразные применения. Новшества этой работы заключаются в следующем:

1. В отличие от задач прогнозирования временных рядов, где эффективность алгоритма оценивается с помощью таких показателей, как среднеквадратичная ошибка, при оценке объяснений с помощью моделей искусственного интеллекта отсутствуют истинные базовые значения. Это усложняет оценку высококачественных методов объяснения. В нашем исследовании представлена количественная структура, разработанная для оценки этих объясняющих результатов, позволяющая выявлять и выделять превосходные методы XAI.

2. В отличие от исходного SHAP, наш алгоритм ShapTime обеспечивает объяснение в измерении временных рядов, т.е. он может выводить важность исторических данных для результатов прогнозирования, что невозможно с другими алгоритмами, не зависящими от модели.

3. Сочетание методов, не зависящих от модели, и методов, зависящих от конкретной модели, позволяет неявно сделать объяснение результатов более информативным, что помогает оптимизировать точность прогнозирования.

4. Результаты, полученные с помощью методов XAI, просто представлены без практического применения в предыдущих работах. Наша статья устраняет этот пробел, демонстрируя, как эти идеи могут быть преобразованы в ощутимую экономическую ценность. Мы

подчеркиваем полезность результатов, применяя их к реальным проблемам. Это включает в себя анализ влияющих факторов, прогнозирование повышения производительности и проблемы онлайн-адаптации.

Благодаря этим инновациям данная работа направлена на повышение надежности и применимости ИИ в прогнозировании временных рядов, делая модели ИИ более интерпретируемыми и практически полезными.

Теоретическая и практическая значимость

Это исследование имеет важное теоретическое и практическое значение. Теоретически, сравнивая эффективность ускоряющих моделей и моделей нейронных сетей в задачах прогнозирования временных рядов, мы создаем основу для выбора наиболее подходящей модели для объяснения. Это косвенно указывает на необходимость разработки общих алгоритмов, не зависящих от модели. Кроме того, благодаря разработке не зависящих от модели алгоритмов прогнозирования временных рядов (ShapTime) и гибридного подхода, сочетающего алгоритмы определения важности признаков и SHAP- алгоритмов (FI-SHAP), это исследование расширяет область применения объяснимой технологии искусственного интеллекта в прогнозировании временных рядов и открывает новые перспективы для объяснения.

На практике это исследование предлагает ценные рекомендации по выбору и внедрению модели, объяснению модели черного ящика, поддержке принятия решений и управлению рисками. Применяя алгоритмы XAI, мы можем анализировать влияющие факторы и обеспечивать надежную поддержку принятия решений, тем самым стимулируя разработку и применение технологии искусственного интеллекта. В целом, это исследование положительно влияет на понятность и надежность искусственного интеллекта, способствуя его широкому использованию в реальных сценариях.

В заключение, это исследование повышает объяснимость и надежность искусственного интеллекта, способствуя его широкому применению на практике. Внедряя новые подходы и перспективы в области поддающегося

объяснению ИИ для прогнозирования временных рядов, мы лучше подготовлены к решению задач и удовлетворению требований в реальных условиях. Это исследование не только повышает эффективность систем поддержки принятия решений, но и закладывает основу для широкого применения технологий искусственного интеллекта в различных областях. Благодаря глубокому пониманию и применению алгоритмов XAI мы обеспечиваем прозрачность, объяснимость и надежность в продвижении искусственного интеллекта, способствуя тем самым его широкому внедрению и общественным выгодам.

Структура диссертации

В первой главе представлен сравнительный анализ моделей искусственного интеллекта, в котором основное внимание уделяется как теоретическим аспектам, так и прогнозированию производительности. В разделе 1.1 разъясняются основные характеристики различных моделей ИИ. В разделе 1.2 мы сравниваем различные методы ИИ, чтобы определить наиболее эффективную модель для различных задач.

Во второй главе мы представляем метод искусственного интеллекта (XAI), разработанный специально для прогнозирования временных рядов и не зависящий от модели. В разделе 2.1 рассматриваются проблемы, с которыми приходится сталкиваться при разработке этой временной перспективы. В разделе 2.2 подробно описывается структура предлагаемого алгоритма ShapTime и объясняется, как он устраняет эти проблемы.

Третья глава посвящена совершенствованию существующих методов XAI путем объединения алгоритмов, не зависящих от модели, и алгоритмов, зависящих от конкретной модели. Результатом такой интеграции является новый подход XAI под названием FI-SHAP, разработанный специально для повышения эффективности моделей.

В четвертой главе демонстрируется практическое применение методов XAI с помощью имитационного моделирования, включая факторный анализ в прогнозном моделировании и решение проблем онлайн-адаптации. В разделах 4.1 и 4.2 используются доступные методы XAI для объяснения оптимальной модели и проведения факторного анализа переменных,

связанных с выработкой солнечной энергии и концентрациями PM2.5, с предоставлением рекомендаций по оптимальному размещению солнечных электростанций. В разделе 4.3 метод XAI применяется для управления изменением концепции в сценариях онлайн-адаптации.

Методы исследования

В этом исследовании используется анализ литературы в качестве теоретического подхода для изучения применения методов искусственного интеллекта в задачах прогнозирования временных рядов, а также исторического развития и текущих проблем, с которыми сталкиваются методы искусственного интеллекта в этой области. Эмпирически мы провели сравнительное исследование, чтобы сравнить различные методы прогнозирования и различные технологии искусственного интеллекта. Кроме того, были проведены экспериментальные исследования, чтобы продемонстрировать работоспособность алгоритмов. Кроме того, были проведены имитационные исследования для моделирования сценариев применения объяснимых методов искусственного интеллекта при прогнозировании временных рядов с использованием реальных данных.

Утверждение полученных результатов

Результаты, представленные в диссертации, были доложены и **одобрены** на следующих международных конференциях и семинарах (с достаточным числом иностранных участников):

- NeuroNT 2021-2022: 2-я и 3-я Международная конференция по нейронным сетям и нейротехнологиям, Санкт-Петербург (Россия).
- IntelliSys 2022-2023: Конференция по интеллектуальным системам, Амстердам (Нидерланды).
- MLIS 2022: Машинное обучение и интеллектуальные системы, Тэгу (Южная Корея).

Во время обучения в аспирантуре автор участвовала в совместном проекте Санкт-Петербургского государственного университета и китайской

коммерческой компании. Этот проект был посвящен прогнозированию и управлению энергией, и результаты этой дипломной работы были частично реализованы в проекте, что привело к успешному достижению ожидаемых результатов. О результатах исследования неоднократно сообщали в соавторстве профессор Петросян Ованес Леонович, аспиранты аспиранты Jinying Zou, Feiran Xu, Ruimin Ma, Jing Liu, Dongfang Qi and Qiushi Sun.

Публикации. Автором выполнено 13 научных работ [19–31] (5 из них Web of Science/Scopus [19, 21–23, 26]), из них 9 работы по теме исследования [19–23, 26, 27, 30, 31], из которых 4 работ опубликованных в научных изданиях, включенных в перечень рецензируемых научных изданий, рекомендованных ВАК РФ [19, 21, 22, 27], 4 работ - в изданиях, индексируемых в международных наукометрических базах SCOPUS/Web of Science [19, 21, 22, 27]. Основные научные результаты, выносимые на защиту, опубликованы в рецензируемых научных изданиях и отражены в следующих работах: [23, 27] - пункт 1, [26] - пункт 2, [20] - пункт 3, [19, 21, 22] - пункт 4.

Все научные работы автора получили в общей сложности 302 цитирований в GoogleScholar[3] за время обучения в аспирантуре (с 2020 по 2024 год). Кроме того, весь код опубликован на GitHub[4].

Личный вклад автора

Работа выполнена в Санкт-Петербургском государственном университете. Часть исследований выполнена совместно с Петросяном Ованесом Леоновичем. Большинство результатов исследований, представленных в диссертации, опубликованы в соавторстве; во избежание двусмысленности в диссертации соответствующие ссылки помечены полным списком имен. Между тем, результаты защиты диссертации принадлежат только авторам.

Основные научные результаты

- Метрика оценки XAI - MDMC создана для измерения точности результатов объяснения, чтобы можно было отфильтровать

[3]The Google Scholar homepage of Zhang Yuyi
[4]The GitHub homepage of Zhang Yuyi

соответствующие наиболее подходящие методы объяснения для различных алгоритмов ИИ. Это обобщенная метрика оценки, т. е. ее можно использовать для любой задачи, включая задачи прогнозирования временных рядов, см. работу [23, 27] и главу 1 (стр. 22) в этой работе (с индивидуальным вкладом не менее 80%).

- Новый алгоритм XAI - ShapTime создан для визуализации важности исторических временных периодов для результатов прогнозирования. Этот алгоритм делит временные шаги на временные периоды во временном измерении и вычисляет значения Шепли для каждого времени как его важность для результатов прогнозирования, см. работу [26] и главу 2 (стр. 39) в этой работе (с индивидуальным вкладом не менее 80%).

- Создан новый гибридный механизм XAI - FI-SHAP, чтобы улучшить точность объяснения текущего алгоритма SHAP. Этот механизм позволяет объяснительным результатам содержать более богатую информацию по сравнению с предыдущими методами SHAP за счет объединения модельно-агностических алгоритмов и модельно-специфических алгоритмов, см. работу [20] и главу 3 (стр. 56) в этой работе (с индивидуальным вкладом не менее 80%).

- Разработанные методы и подходы XAI применяются для анализа влияющих факторов, решения дрейфа концепций в задачах адаптации и повышения точности прогнозирования временных рядов, см. работу [19, 21] и главу 4 (стр. 71) в этой работе. (с индивидуальным вкладом не менее 80%).

Положения, выносимые на защиту

- Количественная оценка методов XAI - MDMC. Построена метрика оценки точности объяснительных результатов, которая способна показать наиболее подходящий метод XAI для заданной модели искусственного интеллекта черного ящика. Таким образом, можно определить оптимальный метод XAI, и в этой работе подтверждено, что метод SHAP, основанный на значении Шепли, является

оптимальным. Поэтому последующая разработка новых алгоритмов XAI основана на значении Шепли.

- Обобщенный метод XAI для измерения временных рядов - ShapTime, который вычисляет значения Шепли для разных периодов времени в измерении времени, что представляет вклад разных периодов времени в результаты прогнозирования. ShapTime больше подходит для задач прогнозирования временных рядов, чем предыдущие методы XAI, которые выводят вклад переменных. Результаты ShapTime также согласуются с метриками MDMC, что в значительной степени доказывает его эффективность.

- Метод XAI, основанный на новом гибридном механизме - FI-SHAP. В алгоритмах бустинга важность признаков и SHAP являются двумя распространенными объяснительными методами, первый содержит информацию из самой модели, а второй - информацию из набора данных. FI-SHAP объединяет оба, так что он содержит больше информации, и, соответственно, основанная на нем инженерия признаков достигает лучших результатов повышения производительности.

- Изучить сценарии, в которых объяснимый ИИ может быть использован в реальных приложениях. Это имеет решающее значение для ИИ, чтобы лучше служить обществу, и является ключом к повышению прозрачности ИИ, а также доверия людей. Это включает в себя анализ факторов влияния, а также решение проблем адаптации в режиме онлайн.

Обзор литературы

Современные подходы к прогнозированию объяснимых временных рядов можно разделить на три категории: Модель-агностические алгоритмы, Модельно-специфические алгоритмы и Гибридные методы.

Модель-агностические алгоритмы. Это достигается путем изменения набора входных данных и изменения выходных данных, и, наконец, приписывания этого изменения входным характеристикам, чтобы реализовать объяснение модели. Это один из подходов, который широко используется по сравнению с другими из-за его универсальности (например, [19, 20, 35–37]). Однако этот тип метода XAI изначально был разработан на основе задач классификации и регрессии, поэтому они часто выводят значение признака вместо значения самого времени, то есть они не могут вывести $\Phi(X_{T_i})$.

Модельно-специфические алгоритмы. Это особый метод разработки моделей прогнозирования временных рядов, то есть встраивание функции интерпретации в модель (например, [38, 40]) для достижения лучшего эффекта интерпретации для прогнозирования временных рядов. Хотя некоторые работы уделяют внимание объяснению временного измерения (например, [39, 41]), они все еще основаны на признаках, то есть подход XAI выводит важность признака в каждый момент времени и сшивает их вместе для достижения объяснения временного измерения. Строго говоря, такое объяснение также не выводит $\Phi(X_{T_i})$, и ему значительно не хватает универсальности.

гибридные методы. Это подход к достижению эффекта объяснения путем гибридизации модулей с определенной степенью функции интерпретации в модели прогнозирования временных рядов. Наиболее характерный подход заключается в гибридизации механизма внимания с моделью прогнозирования временных рядов (например, [42–44]) и получении

объяснения за счет объяснимости механизма внимания. В некоторых работах обсуждалась объяснимость механизма внимания. Даже, несмотря на некоторые разногласия (например, [66]), исследователи по-прежнему положительно относятся к его объяснимости (например, [67, 68]). Гибридные методы страдают от той же проблемы, что и вышеупомянутые подходы, то есть, даже если есть объяснение во временном измерении, это объяснение зависит от особенностей. С другой стороны, оно также требует разработки новых моделей и, следовательно, не имеет общего характера.

Методы машинного обучения и глубокого обучения продемонстрировали высокую эффективность в области прогнозирования временных рядов, особенно в задачах долгосрочного прогнозирования. Методы ансамблевого обучения [69] и глубокого обучения продемонстрировали высокую эффективность при работе с нелинейными и нестационарными данными. Метод ансамблевого обучения включает в себя объединение прогнозов нескольких моделей для повышения точности, а метод глубокого обучения использует сложные структуры сетей для выявления взаимосвязей между входными и выходными переменными. Эти подходы имеют особенно важное значение в таких областях, как прогнозирование солнечной и ветровой энергии [70], где часто наблюдаются нелинейные взаимосвязи и закономерности. Ансамблевое обучение включает в себя алгоритмы бустинга и бэггинга, при этом широко используются алгоритмы бустинга, такие как XGBoost [71], LightGBM [72] и CatBoost [73]. С другой стороны, глубокое обучение можно разделить на три разных типа по структуре сети: Искусственные нейронные сети (ANN) [27, 74], конволюционные нейронные сети (CNN) [75] и рекуррентные нейронные сети (RNN) [76].

XGBoost, LightGBM и CatBoost продемонстрировали высокую эффективность в конкурсах по прогнозированию временных рядов. В частности, XGBoost широко использовался в конкурсе M4 [77]. LightGBM продемонстрировал свою эффективность в обработке обширных наборов данных и в быстром обучении, что сделало его популярным выбором в различных конкурсах. На самом деле LightGBM опередил другие фреймворки и победил в конкурсе прогнозирования M5 [78], продемонстрировав тем самым свою высокую эффективность в реальных сценариях. CatBoost также широко используется в конкурсах Kaggle по

прогнозированию временных рядов. Все три системы бустинга продемонстрировали свою эффективность для долгосрочного прогнозирования временных рядов, и каждая из них обладает уникальными характеристиками, которые делают их подходящими для разных случаев использования [79] [80]. Bae DJ и др. [81] предложили алгоритм прогнозирования на основе XGBoost, который повысил точность на 21% и 29% в 2019 и 2020 годах, по сравнению с предыдущими моделями. Zhang Y и др. [19] оценили основные методы прогнозирования на различных наборах данных, включая производство солнечной энергии. Результаты эксперимента показали, что LightGBM является лучшим алгоритмом в целом, превосходящим другие в трех наборах данных. CatBoost, еще один заметный алгоритм бустинга, также демонстрирует очевидный потенциал в прогнозировании выработки солнечной энергии [82]. Методы глубокого обучения часто используются для прогнозирования временных рядов, в частности искусственные нейронные сети (ANN) [83], модели на основе рекуррентных нейронных сетей (RNN [84], LSTM [85], GRU [86]) и модели на основе двунаправленных RNN (Bi-RNN, Bi-LSTM, Bi-GRU) [87]. Эти модели могут автоматически извлекать нужные признаки из входных данных, что дает им преимущество в сравнении с ансамблевым обучением. ANN демонстрирует превосходство в улавливании сложных нелинейных закономерностей в сценариях, где связи между входными признаками и целевыми переменными неясны или нелинейны. Хотя модели на основе RNN и Bi-RNN изначально были разработаны для обработки естественного языка, они получили широкое распространение в прогнозировании выработки солнечной энергии из-за их способности улавливать временные зависимости. Учитывая, что различные модели могут обладать разной способностью к обучению на основе данных, необходимо проводить всесторонние сравнения и анализы с учетом конкретных случаев [19]. Ансамблевое обучение и глубокое обучение представляют собой важные методологии прогнозирования временных рядов, особенно в контексте сложных и долгосрочных прогнозов. Тем не менее, их неинтерпретируемые свойства создают проблему для выявления факторов, влияющих на точность этих прогнозов [89]. Напротив, выявление таких факторов приобретает большое значение при оптимизации солнечных энергетических

систем. Такие факторы, как погодные условия [90], затенение [91] и характеристики оборудования [92], играют ключевую роль в производительности системы, снижении затрат и выборе места. Изучение этих факторов позволяет уменьшить риски, оптимизировать производство энергии и помочь в выборе наиболее подходящего места для солнечной электростанции. Следовательно, для успешного производства солнечной энергии требуется критический анализ влияющих факторов и эффективные методы прогнозирования. Однако отсутствие объяснимости представляет собой серьезное затруднение для их дальнейшего развития. Необъяснимость означает то, что модель не может быть понятна человеку. Объяснимость [93–95] включает в себя следующие аспекты:

- Интерпретируемость моделей
- Интерпретируемость параметров

Если модели и параметры могут быть легко поняты человеком, их называют «моделями белого ящика». Напротив, модели, внутренние механизмы которых неизвестны человеку, называются «моделями черного ящика». Следует отметить, что модели «черного ящика» все чаще демонстрируют превосходную эффективность в задачах прогнозирования временных рядов, особенно в задачах долгосрочного прогнозирования. Несмотря на высокую эффективность моделей «черного ящика», присущая им неинтерпретируемость вызывает две серьезные проблемы: кризис доверия и недостаток знаний [96–98].

- Кризис доверия связан с тем, что люди, будь то пользователи или разработчики, не могут понять закономерности и процессы принятия решений в этих моделях «черного ящика». Отсутствие транспарентности приводит к возникновению опасений по поводу возможной дискриминации, предвзятости или других нежелательных факторов, заложенных в архитектуру модели. Кроме того, отсутствие ясного понимания внутренней работы модели сдерживает ее последующее совершенствование и оптимизацию. В целом эти факторы усугубляют недоверие людей к моделям «черного ящика».

- Недостаток знаний - еще одна критическая проблема. Хотя модели «черного ящика» могут приносить относительно точные результаты в задачах прогнозирования временных рядов, они просто генерируют выходные данные, но не предоставляют дополнительной информации или знаний. Напротив, традиционные математические и статистические методы со своими интерпретируемыми моделями предлагают дополнительные знания и информацию, помогающие человеку принимать решения. Эти методы могут определить переменные и временные периоды, которые наиболее сильно влияют на прогнозируемые результаты, и тем самым помочь человеку принимать более обоснованные и разумные решения.

В качестве возможного решения проблем, связанных с непрозрачными моделями искусственного интеллекта (ИИ), была предложена концепция объяснимого искусственного интеллекта (Explainable Artificial Intelligence, XAI) [32–34, 46–48]. Целью XAI является повышение интерпретируемости и прозрачности сложных моделей «черного ящика», чтобы сделать их понятными для пользователей, даже если только частично или в определенной степени. При этом XAI стремится предоставить человеку как можно больше информации и знаний, что позволяет лучше понять процессы принятия решений и обоснования, заложенные в эти модели.

В связи с поиском возможности интерпретации моделей машинного обучения были разработаны различные алгоритмы, включая LIME (Local Interpretable Model-agnostic Explanations) [52], SHAP (SHapley Additive exPlanations) [51], Deep-LIFT [99, 100], Integrated Gradients (IG) [101] и LRP (Layer-wise Relevance Propagation) [102]. Следует отметить, что значение Шэпли, на котором основан SHAP, является единственным алгоритмом, удовлетворяющим определенным детерминированным свойствам, что способствует его широкому применению в исследованиях, связанных с интерпретируемостью. Рисунок 3 иллюстрирует индексацию оригинальных статей по этим подходам, демонстрируя явное преимущество SHAP. Важно подчеркнуть, что эффективная реализация SHAP в инженерных технологиях использует идеи LIME, что прямо приводит к высокому уровню интереса к LIME. Кроме того, концепции, на которых базируется LIME, являются источником вдохновения для ключевых алгоритмов,

используемых в данном исследовании.

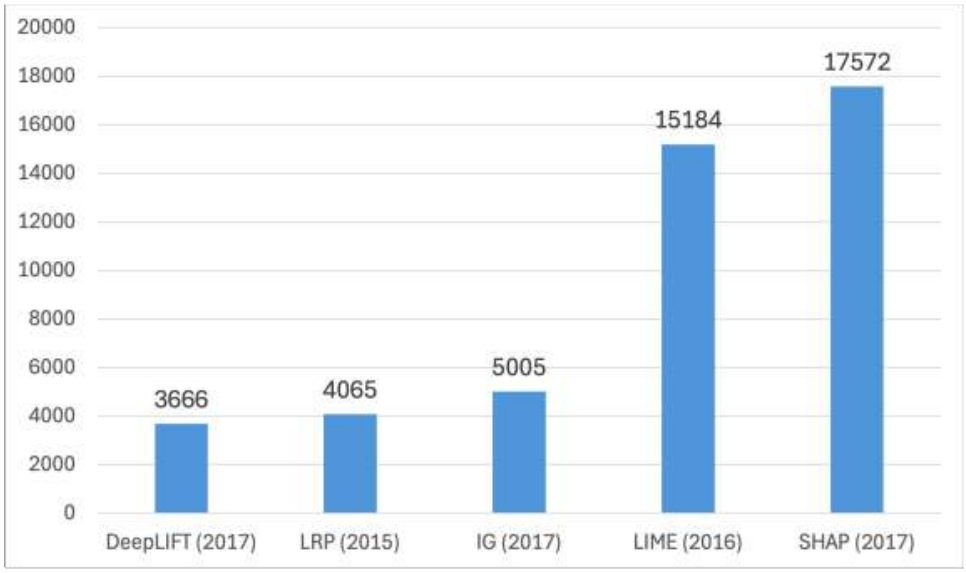


Рис. 3: Ссылки на статьи для алгоритмов с датой сбора данных 07.12.2023.

Как известно, SHAP является широко распространенным методом [51, 103, 104]. Однако его прямое применение к задачам прогнозирования временных рядов нецелесообразно. В рамках классического SHAP делается предположение о том, что выборки независимы, а значит, временные зависимости между выборками не учитываются. Следовательно, при работе с данными, имеющими временные зависимости, объяснительные результаты, получаемые с помощью SHAP, ограничены важностью признаков и не отражают внутреннюю динамику, управляющую временным рядом. Например, при прогнозировании температуры [105] традиционный SHAP может объяснить относительную важность факторов температуры, таких как солнечное облучение и влажность. Однако он не способен определить влияние температурных значений предыдущих временных шагов на последующие, а ведь это важный аспект при анализе моделей прогнозирования временных рядов.

Глава 1

Сравнение алгоритмов прогнозирования и объяснимого ИИ для прогнозирования временных рядов

Объяснимый искусственный интеллект (XAI) фокусируется на том, чтобы сделать модели ИИ более понятными, поэтому важно сравнить производительность этих моделей. В этой главе мы сравниваем основные модели ИИ, чтобы определить наиболее подходящие для различных задач, и **наши результаты были опубликованы в научных журналах [19, 21, 27]**. Важно отметить, что различные алгоритмы XAI могут давать разные результаты даже при применении к одной и той же модели искусственного интеллекта. Поэтому оценка качества этих результатов объяснения имеет решающее значение. Чтобы решить эту проблему, мы разработали систему оценки XAI под названием MDMC, которая помогает отфильтровывать наиболее подходящие алгоритмы XAI для различных моделей искусственного интеллекта, и **наши результаты были опубликованы на конференции [23]**.

1.1 Сравнение алгоритмов объяснимого искусственного интеллекта

1.1.1 Алгоритмы прогнозирования

Нейронные сети и ансамблевые модели считаются образцовыми представителями моделей "черного ящика"в этой области. В частности, искусственные нейронные сети (ANN) [27, 74], а также ансамблевые модели,

такие как Random Forest (RF) [106] и LightGBM [78–80] демонстрируют выдающиеся прогностические возможности.

Сфера прогностического моделирования включает в себя различные модели "черного ящика среди которых выделяются нейросетевые модели и ансамблевые модели. Среди них модель искусственной нейронной сети (ANN) является представителем нейросетевой парадигмы. В то же время LightGBM и Random Forest считаются представителями ансамблевых моделей. LightGBM, основанная на алгоритме boosting, является олицетворением ансамблевой модели, в то время как Random Forest основана на алгоритме bagging. И бустинг, и бэггинг по сути состоят из нескольких упрощенных моделей деревьев. При отсутствии взаимосвязи между этими моделями деревьев алгоритм суммирования использует механизм "голосования"для получения конечного результата. С другой стороны, когда между этими древовидными моделями существует сильная взаимозависимость, они образуют алгоритм boosting, в котором результаты предыдущих древовидных моделей влияют на последующие. Кроме того, в задачах прогнозирования для сравнительного анализа используются линейная регрессия и деревья решений на основе моделей белого ящика.

Нейронную сеть можно описать как адаптивную нелинейную динамическую систему, состоящую из множества базовых единиц, обычно называемых нейронами. Эти нейроны связаны между собой посредством функций активации, что приводит к возникновению сложной сети взаимодействий. Хотя структура и функции каждого отдельного нейрона относительно просты, коллективное поведение всей сети становится чрезвычайно сложным и труднообъяснимым. Архитектура ANN обычно состоит из трех отдельных слоев: входного, скрытого и выходного. Такое расположение слоев позволяет сети обрабатывать информацию в последовательном порядке. На входной слой поступают внешние стимулы или данные, которые затем передаются на скрытый слой. В скрытом слое происходят промежуточные вычисления, позволяющие осуществлять сложные преобразования и выделение признаков. Наконец, обработанная информация поступает на выходной слой, который генерирует ответ или предсказание сети.

GBDT [78–80] занимает чрезвычайно важное место в области машинного

обучения. По своей сути эта модель основана на использовании слабых классификаторов, в частности деревьев решений, для постепенного обучения и получения оптимальной модели. Использование такого подхода делает GBDT очень привлекательной благодаря исключительной эффективности обучения и способности избегать перебора. В соответствии с GBDT, LightGBM, также известный как Light Gradient Boosting Machine, является замечательной системой, реализующей алгоритм GBDT. Этот фреймворк обладает рядом примечательных преимуществ, включая ускоренное обучение, уменьшенное потребление памяти и повышенную точность. Способность LightGBM обеспечить эти преимущества еще больше повышает его ценность как важного инструмента в области машинного обучения.

Случайный лес - это мощный метод ансамблевого обучения, который объединяет несколько деревьев решений для повышения точности и стабильности прогнозов. Объединяя эти деревья решений, классификатор случайного леса использует совокупные знания всех отдельных классификаторов, в то же время используя гиперпараметры, связанные с классификатором мешков, для регулирования его общей структуры. Важно отметить, что помимо классификации, концепция случайного леса может быть применена и для решения задач регрессии, когда используется регрессор случайного леса.

1.1.2 Тесты эффективности прогнозирования

Чтобы наглядно убедиться в преимуществах модели "черного ящика"в прогнозировании, мы используем две традиционные модели, линейную регрессию (LR) и классическую модель прогнозирования "белого ящика дерево решений (DT) - в качестве контрольной группы. Эти модели используются наряду с искусственными нейронными сетями (ANN), LightGBM и случайным лесом (RF) для прогнозирования набора данных по жилью в Бостоне. Результаты прогнозирования с точки зрения качества представлены в следующей таблице 1.1, учитывая стандартный набор данных [1].

Анализ метрик показывает, что искусственные нейронные сети (ANN),

[1]Набор данных по жилью в Бостоне

Таблица 1.1: Статистика качества

$Model$	R^2	MSE	MAE
LR	0.64856	28.40585	3.69136
DT	0.74361	20.72322	3.05065
ANN	0.79477	16.58769	2.57977
LGBM	0.80417	15.82852	2.53292
RF	0.81888	14.63918	2.35044

LightGBM и случайные леса (RF) демонстрируют существенные преимущества при использовании в задачах прогнозирования. Тем не менее, в отличие от линейной регрессии и деревьев решений, ANN, LightGBM и RF сложны для понимания человеком из-за их запутанной внутренней структуры. Следовательно, несмотря на то, что модели "черного ящика"с повышенной точностью превосходят традиционные модели и модели "белого ящика"с пониженной точностью в различных областях, они не могут полностью их заменить. Следовательно, возникает необходимость в разработке эффективных подходов, направленных на понимание и прояснение работы этих непрозрачных моделей.

1.1.3 Алгоритмы объяснимого искусственного интеллекта

Согласно рисунку 3, в SHAP и LIME наблюдается явное превосходство внимания по сравнению с другими методами. Поэтому в данном разделе мы сосредоточимся на этих двух подходах, используя количественные метрики для оценки их эффективности.

SHAP (SHapley Additive exPlanations) и LIME (Local Interpretable Model-agnostic Explanations) представляют собой две известные методологии, используемые в области интерпретации моделей машинного обучения и понимания их предсказаний. В случае с SHAP ее основная задача заключается в присвоении значений значимости входным признакам данной модели, чтобы пролить свет на ее выходные данные. Используя принципы теории игр и значения Шэпли, SHAP эффективно вычисляет атрибуции признаков. Исчерпывающе рассматривая все возможные комбинации признаков и соответствующих результатов, этот подход точно оценивает вклад каждого признака в предсказание случая. Примечательно, что данная методология использует всесторонний взгляд на взаимодействие

признаков, что позволяет получить объяснения, которые являются одновременно локально точными и глобально согласованными. Значения SHAP представляют собой комплексную основу для интерпретации различных алгоритмов машинного обучения, включающих глубокие нейронные сети, древовидные и линейные модели. Следовательно, это облегчает понимание того, как отдельные характеристики влияют на предсказания модели, давая нам возможность понять, что лежит в основе конкретных решений.

В рамках SHAP переменные $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ рассматриваются как набор игроков, а модель черного ящика f заменяется объяснителем g, таким образом, что $f(\mathbf{x}) = g(\mathbf{x}) = \phi_{x_0} + \sum_{i=1}^{n} \phi_{x_i}$. Это означает, что SHAP предполагает, что любая модель "черного ящика"будет представлена этим линейным выражением, где ϕ_i представляет собой вклад соответствующей переменной в предсказание, а Φ_{x_0} представляет собой выход модели, когда все переменные неэффективны. На самом деле, при моделировании реальных проблем практически невозможно перечислить все влияющие факторы. В физическом смысле Φ_{x_0} представляет собой вклад этих переменных, не включенных в область анализа. В этих рамках математические выражения всех моделей "черного ящика"принимаются за вышеупомянутое линейное выражение, а вклад каждой переменной, представленный Φ_{x_i}, служит результатом объяснения, тем самым достигается общность XAI. Вклад каждой переменной вычисляется по формуле ценности Шэпли,
$\phi(x_i) = \sum_{\mathbf{x_s} \subseteq \{x_1 \cdots x_n\} \setminus \{x_i\}} \frac{|s|!(n-|s|-1)!}{n!}(v(\mathbf{x_s} \cup \{x_i\}) - v(\mathbf{x_s}))$, которая рассматривает результаты прогнозирования модели "черного ящика" $f(\mathbf{x}_s)$ (где s - подмножество n) как функцию прибыли $v(\mathbf{x}_s)$. Это позволяет нам вычислить значение вклада $\phi(x_i)$ для каждой переменной x_i, служащее объяснением модели "черного ящика".

По сравнению с этим LIME делает акцент на предоставлении объяснений на локальном уровне для отдельных предсказаний, а не на стремлении к глобальной интерпретируемости. Это достигается путем аппроксимации сложных моделей машинного обучения интерпретируемыми моделями, которые сравнительно проще для понимания. Подход LIME заключается в возмущении входных данных вокруг конкретного интересующего нас

случая и наблюдении за изменениями в прогнозах. Создавая новый набор данных путем выборки экземпляров из исходных данных, LIME подгоняет локальную модель под эти выборочные данные, тем самым генерируя объяснения.

Аппроксимированная модель, созданная LIME, проливает свет на значимые особенности и их влияние на предсказание для данного конкретного случая. Примечательным аспектом LIME является его модельно-агностическая природа, позволяющая применять его к любой модели "черного ящика"без необходимости понимания внутренней работы этой модели. Одной из сильных сторон LIME является предоставление понятных человеку объяснений, таких как важность функций или выделение значимых частей входных данных. Эта возможность способствует развитию доверия к моделям машинного обучения.

Как SHAP, так и LIME служат мощными инструментами для интерпретации сложных моделей машинного обучения. В то время как SHAP дает глобальные объяснения, учитывающие взаимодействие признаков, LIME в основном фокусируется на предоставлении локальных объяснений, используя упрощенную приближенную модель. Выбор между этими методологиями зависит от конкретной задачи и специфических требований к анализу интерпретируемости.

1.1.4 Рамки оценки алгоритмов объяснимого ИИ

Создание системы оценки объяснимого ИИ основано на предпосылке: *Удаление или изменение признаков с большим вкладом (рассчитанным методом XAI) в наборе данных приведет к значительному снижению точности предсказания модели.* Поэтому, исходя из этой предпосылки, степень изменения метрик (R2, MSE, MAE) может быть использована в качестве ядра системы оценки.

Средняя степень изменения показателей (MDMC)

При возмущении исходных данных в соответствии с результатами метода XAI и вводе измененных данных в модель прогнозирования для получения новых метрик (R2*, MSE*, MAE*), тогда степень изменения метрик (D) может быть определена как:

$$D = f(M - M^*) \tag{1.1}$$

M - исходная метрика, а M* - измененная метрика.

Объединив степень изменения всех показателей, можно определить итоговую систему оценки (MDMC):

$$\begin{aligned} MDMC &= \frac{1}{n}\sum_{i=1}^{n} D = \frac{1}{n}\sum_{i=1}^{n} f(M - M^*) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[(R_0^2 - R_i^2) + (MSE_i - MSE_0) + (MAE_i - MAE_0)\right] \end{aligned} \tag{1.2}$$

Следует отметить, что значения MSE и MAE в уравнении 1.2 необходимо использовать после нормализации. Теоретически, чем больше значение MDMC, это означает, что модель предсказания "черного ящика"внесла значительные изменения в набор данных, тем самым доказав эффективность XAI.

Визуализация качественной оценки

Чтобы интуитивно понять фреймворк MDMC, возьмем в качестве примера модель ANN для оценки методов XAI. Изменения в метриках показаны на рисунке 1.1. Стоит отметить, что если стабильность самой ANN-модели недостаточна, то случайное удаление признаков также приведет к снижению точности предсказания модели. Чтобы исключить эту возможность, в процессе оценки набор данных "случайно удаленных признаков"использовался в качестве сравнения методов XAI для подтверждения устойчивости построенной нами ANN-модели.

Как показано на рисунке 1.1, по сравнению со случайным удалением признаков, возмущение, основанное на результатах методов XAI, значительно снижает точность ANN-модели предсказания.

Метрики для количественной оценки

Визуализация позволяет интуитивно оценить метод XAI, но для задач с высокими требованиями к точности необходима матрица количественной оценки. Процесс применения методов XAI на LightGBM и Random Forest

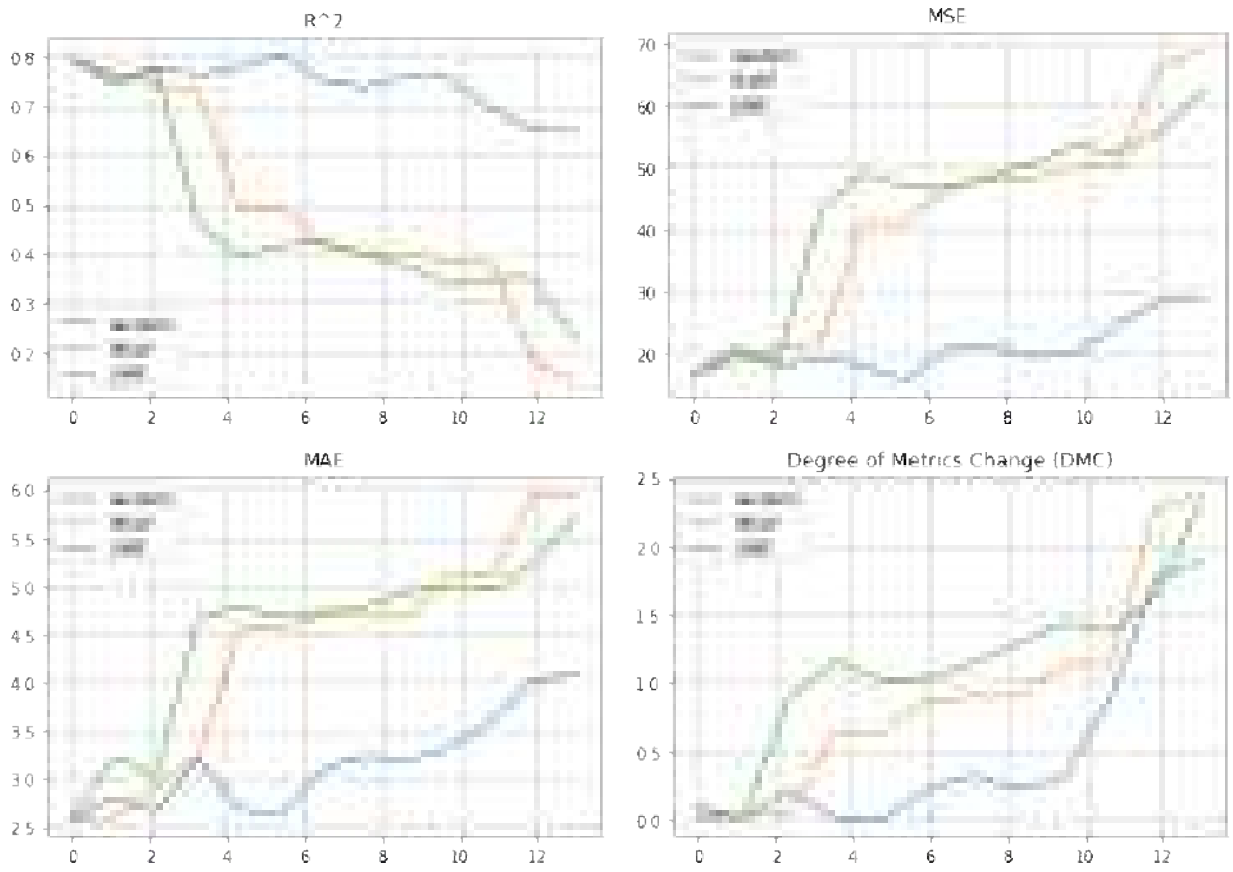


Рис. 1.1: Визуализация изменений в метриках для ANN

аналогичен процессу применения ANN и подчиняется той же формуле расчета. Значения в таблице 1.2 рассчитываются на основе уравнения 1.2. Чем больше значение, тем эффективнее метод XAI.

Таблица 1.2: Матрица оценки

MDCD	ANN	$LightGBM$	$Randomforest$
Random	0.8333	0.8222	0.6242
SHAP	1.4154	**1.5004**	1.3286
LIME	**1.6506**	1.3995	**1.6069**

Результаты экспериментов показывают, что в модели черного ящика возмущение, основанное на методах XAI, более эффективно, чем случайное возмущение. С другой стороны, в рамках метода XAI LIME лучше работает в модели ANN и модели случайного леса, а SHAP более эффективен в модели LightGBM.

1.2 Сравнение алгоритмов прогнозирования на основе искусственного интеллекта

1.2.1 Сравнение прогнозирования временных рядов энергии

Описание данных

Конкурс Schneider делает значительный акцент на фундаментальной роли планирования и прогнозирования для достижения эффективной работы в энергетическом секторе. Поэтому точное прогнозирование временных рядов становится обязательным в этой области. Одновременно Американское общество инженеров по отоплению, охлаждению и кондиционированию воздуха (ASHRAE) в своем конкурсе по прогнозированию энергопотребления зданий на Kaggle подчеркнуло, что существующие методы оценки не являются согласованными и не обеспечивают масштабируемости. Следовательно, отсутствие стандартизированного подхода мешает нам определить оптимальную модель для прогнозирования временных рядов. Ниже приводится описание набора данных по энергетике, использованного в конкурсе прогнозов:

- ASHRA: Большой энергетический прогнозист III[2]
  . Набор данных ASHRAE включает в себя данные об использовании энергии из различных источников в зданиях, включая счетчики охлажденной воды, электричества, горячей воды и пара. Этот обширный набор данных охватывает более трех лет и включает в себя более 1000 зданий. Для обеспечения точного прогнозирования набор данных включает 15 различных характеристик. Эти характеристики включают как внутренние характеристики зданий, такие как идентификатор здания, тип использования, площадь, год постройки и количество этажей, так и внешние факторы, такие как скорость ветра, направление ветра, температура и облачность. Точная оценка инвестиций в энергосбережение имеет первостепенное значение, поскольку привлекает повышенное внимание заинтересованных сторон, в частности финансовых учреждений. Такое повышенное внимание к

[2]Набор данных ASHRAE

этой области в конечном итоге стимулирует прогресс в области эффективности зданий.

- Законы силы: Прогнозирование энергопотребления[3]
  Набор данных, предоставленный компанией Schneider, включает в себя полный набор из 14 характеристик, включая идентификатор здания, температуру, информацию о праздниках и выходных, и охватывает период примерно в три года. Основная цель этого конкурса - повысить точность оценки глобального потребления энергии в зданиях. Используя этот набор данных, участники стремятся уточнить и улучшить существующие методики, используемые для оценки этих параметров потребления.

- Данные о генерации солнечной энергии[4]
  В нашем исследовании мы включили данные о солнечных электростанциях в качестве дополнительного ресурса. Основная цель этого проекта - улучшить управление сетью за счет точного прогнозирования ближайших мощностей по производству солнечной энергии. Используемый набор данных включает в себя пять различных характеристик, а именно: номер устройства, постоянный ток, переменный ток, температура и радиация. Эти переменные были тщательно отобраны, чтобы отразить основные аспекты, влияющие на выработку солнечной энергии.

В данном исследовании в качестве демаркации между обучающим и тестовым набором используется определенная временная точка, что соответствует логике, присущей временным рядам. Приблизительно 70% элементов должны быть включены в обучающий набор.

Целевая функция

Для оценки эффективности моделей машинного обучения были разработаны различные критерии. Две широко используемые метрики включают коэффициент детерминации (R^2) и среднюю квадратичную

[3]Набор данных Schneider
[4]Набор данных о солнечных электростанциях

ошибку (MSE), которые могут быть рассчитаны следующим образом:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.3}$$

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 \tag{1.4}$$

где y_i - истинное значение; $\hat{y}_i$ - прогнозное значение, $\bar{y}$ - среднее значение всех истинных значений, а n - количество наблюдений.

Сравнительные результаты

В этом исследовании мы приводим в качестве примера результаты конкурса ASHRAE по величине потерь, обращая особое внимание на их простоту. Для обеспечения эффективности вычислений подмножество данных было выбрано в зависимости от вычислительной мощности компьютера. В частности, обучающий набор состоял из 242,214 образцов, а тестовый набор - из 79,514 образцов. Чтобы облегчить анализ, на основе этих данных было построено 98 признаков.

Кривые обучения всех моделей демонстрировали последовательное снижение, как показано на рисунке 1.2. Примечательно, что модель Bi-RNN продемонстрировала более высокую производительность на валидационном наборе по сравнению с обучающим набором. Для того чтобы представить всестороннее сравнение, необходимо включить подробный количественный анализ показателей.

Очевидно, что все модели прогнозирования продемонстрировали исключительную производительность на наборе данных ASHRAE, как показано на рисунке 1.2 и в таблице 1.3. Среди этих моделей Bi-RNN выделяется наиболее впечатляющей производительностью. Однако стоит отметить, что M5 LightGBM демонстрирует значительно меньшее время работы по сравнению со случайным лесом и нейросетевыми моделями. Это означает, что M5 LightGBM может достичь уровня точности, аналогичного нейросетевым моделям в контексте ASHRAE, при этом требуя меньше вычислительного времени.

После первоначального применения этих моделей прогнозирования мы продолжили использовать их на различных наборах данных, включая конкурс Schneider, а также набор данных по солнечным электростанциям,

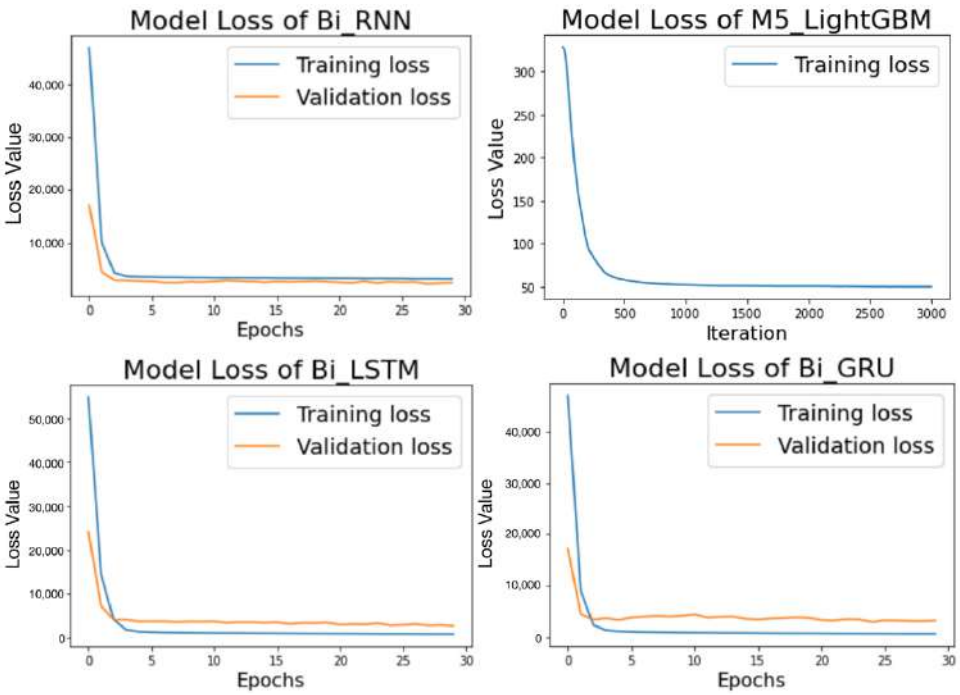


Рис. 1.2: ASHRAE: Кривые обучения моделей прогнозирования. Примечание: Кривая обучения ансамблевого алгоритма в модели LightGBM - RMSE, а нейронной сети - MSE.

Таблица 1.3: Прогноз качества электроэнергии ASHRAE

ASHRAE	R^2	MSE	TimeSpent
M5LightGBM	0.9676	2516.83	39.5 s
Random Forest	0.9673	2538.70	3 min 8 s
Bi-RNN	0.9688	2426.72	6 min 33 s
Bi-LSTM	0.9655	2678.17	13 min 47 s
Bi-GRU	0.9568	3358.26	12 min 11 s

доступный на Kaggle. В наборе данных Schneider мы тщательно отобрали часть данных для анализа, состоящую из обучающего набора, включающего 271,803 экземпляра, и тестового набора, содержащего 78,380 экземпляров. Затем на основе этого набора данных мы составили полный набор из 28 признаков.

В наборе данных по солнечной энергетике мы использовали все имеющиеся данные (которые были сравнительно небольшими по объему). Для завода 1 обучающий набор содержал 2,393 экземпляра, а тестовый набор состоял из 864 экземпляров. Аналогично, для Plant 2 обучающий набор включал 2,293 экземпляра, а тестовый набор состоял из 862 экземпляров. На основе этого набора данных мы создали в общей сложности 19 признаков для облегчения анализа.

Чтобы оценить эффективность моделей прогнозирования, важно измерить их производительность (см. таблицу 1.4). Точность оценивается в рамках каждого набора данных, а стабильность - по всем наборам. Примечательно, что значения целевой переменной в наборах данных Schneider и Kaggle значительно велики, что делает оптимизацию потерь всех моделей прогнозирования практически неэффективной. Поэтому в процессе прогнозирования применяется осторожный подход, при котором целевая переменная и данные о признаках обрабатываются отдельно с помощью алгоритма MinMax.

Таблица 1.4: Прогноз качества электроэнергии Schneider

Schneider	R^2	MSE	TimeSpent
M5LightGBM	0.9381	0.0001	42.1s
Random Forest	0.9297	0.0001	1min 34s
Bi-RNN	0.8595	0.0003	55.2 s
Bi-LSTM	0.9146	0.0001	1min 19s
Bi-GRU	0.9165	0.0001	1min 48s

Что касается наборов данных по солнечной энергетике, то производительность моделей прогнозирования для электростанции 1 неизменно высока, о чем свидетельствует таблица 1.5. В частности, модель M5 LightGBM выделяется своей точностью и эффективностью, превосходя все остальные модели по обоим показателям.

Таблица 1.5: Прогноз качества производства солнечной энергии - Завод 1

Solar-1	R^2	MSE	TimeSpent
M5LightGBM	0.9928	0.0008	0.74 s
Random Forest	0.9723	0.0034	0.99 s
Bi-RNN	0.9665	0.0041	6.17 s
Bi-LSTM	0.9887	0.0013	12.9 s
Bi-GRU	0.9865	0.0016	8.36 s

Когда речь идет об электростанции 2, основной задачей является поддержание высокой точности, достигающей достойного значения 0,9005. В этой связи модель M5 LightGBM оказывается наиболее оптимальным выбором, обеспечивающим эффективную работу при минимизации временных затрат. Эти результаты представлены в таблице 1.6 для справки.

Таблица 1.6: Прогноз качества производства солнечной энергии - Завод 2

Solar-2	R^2	MSE	TimeSpent
M5LightGBM	0.9005	0.0081	0.16 s
Random Forest	0.8689	0.0107	1.19 s
Bi-RNN	0.9329	0.0055	6.39 s
Bi-LSTM	0.8917	0.0088	13.7 s
Bi-GRU	0.9185	0.0066	8.76 s

Результаты показывают, что M5 LightGBM демонстрирует явные преимущества. По точности M5 LightGBM превосходит другие алгоритмы в обоих наборах данных (см. таблицу 1.4 и таблицу 1.5). Кроме того, он демонстрирует значительно меньшее время обработки, особенно по сравнению с нейросетевыми алгоритмами. Хотя M5 LightGBM и не достигает лучшей производительности в некоторых наборах данных (см. Таблицу 1.3 и Таблицу 1.6), он все же достигает сопоставимой точности прогнозирования ($R^2 > 0.9$) при меньших вычислительных затратах. Таким образом, проведя всестороннее сравнение, мы пришли к выводу, что M5 LightGBM является более совершенной моделью прогнозирования.

Алгоритмы ансамблевого обучения, несмотря на их кажущуюся простоту, часто демонстрируют исключительную эффективность в практических приложениях. Это объясняется их способностью объединять множество фундаментальных алгоритмических идей в единую структуру, усиливая сильные и минимизируя слабые стороны. Следовательно, такая оптимизация повышает устойчивость и обобщающую способность исходных базовых алгоритмов, что способствует стабильной работе модели. Для наглядного представления результатов прогнозирования, пожалуйста, обратитесь к рисунку 1.3.

1.2.2 Сравнение прогнозирования временных рядов PM2.5

Описание данных

Набор данных, используемый в данном исследовании, включает в себя данные о концентрации PM2.5, собранные в спортзале Олимпийского спортивного центра Пекина[5]. Данные охватывают период с 1 марта 2013

[5]Beijing PM2.5 data set

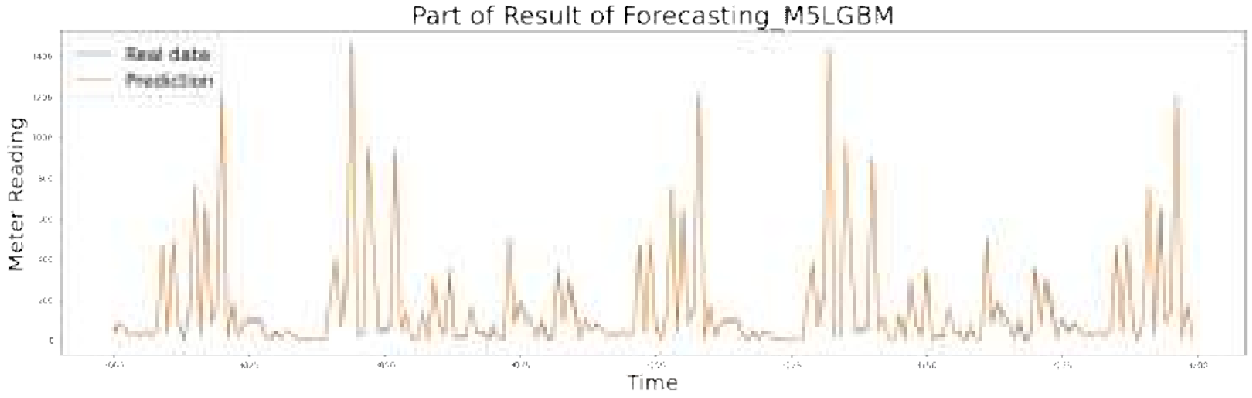


Рис. 1.3: Визуализация прогнозов M5LightGBM на наборе данных ASHRAE

года по 28 февраля 2017 года, измерения регистрировались с часовым интервалом.

Набор данных, используемый в исследовании, состоит из 31,815 наблюдений. Он включает 11 переменных, в том числе метеорологические условия и коэффициенты выбросов. Коэффициенты выбросов включают в себя твердые частицы ($PM10$, ug/m^3), диоксид серы (SO_2, ug/m^3), диоксид азота (NO_2, ug/m^3), монооксид углерода (CO, ug/m^3) и озон (O_3, ug/m^3). Среди метеорологических условий есть непрерывные переменные, такие как температура (TEMP, градусы Цельсия), давление (PRES, гПа), точка росы (DEWP, градусы Цельсия), количество осадков (RAIN, мм), скорость ветра (WSPM, м/с), и одна дискретная переменная - направление ветра (WD). Направление ветра кодируется по системе естественного порядка с 16 направлениями: NNW, N, NW, NNE, ENE, E, NE, W, SSW, WSW, SE, WNW, SSE, ESE, S, SW.

Целевая функция

Учитывая, что в данном разделе мы рассматриваем прогнозирование длинных временных рядов как более сложную задачу, становится необходимым использовать комплексные метрики оценки для эффективного определения различий в производительности различных моделей ИИ. Поэтому мы специально выбрали такие объективные функции, как R^2 (коэффициент детерминации), RMSE (среднеквадратичная ошибка) и MAE (средняя абсолютная ошибка), чтобы они послужили основой для наших оценок. Эти метрики позволят нам тщательно оценить эффективность различных моделей ИИ в решении этой сложной задачи прогнозирования.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.5}$$

$$RMSE\% = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{n} \times 100 \tag{1.6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.7}$$

где y_i - истинное значение; $\hat{y}_i$ - прогнозное значение, $\bar{y}$ - среднее значение всех истинных значений, а n - количество наблюдений.

Сравнительные результаты

Долгосрочное прогнозирование, основанное на статистических методах, является сложной задачей из-за нестационарности и шума. Глубокое обучение [107–111] и ансамблевое обучение [112–114] являются мощными методами для работы с нелинейными и нестационарными данными. Они предлагают эффективные способы выявления взаимосвязей в сложных наборах данных. Ансамблевое обучение достигается путем объединения прогнозов нескольких моделей, используя их коллективную мудрость для повышения точности. Алгоритмы бустинга [113, 114], такие как XGBoost, LightGBM и CatBoost, широко используются в ансамблевом обучении, наряду с алгоритмами бэггинга. С другой стороны, глубокое обучение использует сложную сетевую структуру для выявления сложных закономерностей и зависимостей между входными и выходными переменными. Эта техника подразделяется на три типа в зависимости от структуры сети: Искусственные нейронные сети (ANN) и рекуррентные нейронные сети (RNN). Эти методы находят широкое применение в таких областях, как долгосрочное прогнозирование PM2.5 [4–6], где преобладают нелинейные связи и закономерности. Используя ансамблевое обучение и глубокое обучение, исследователи и практики смогут более эффективно решать проблему нелинейных и нестационарных данных в этих областях.

Мы использовали коэффициенты выбросов и метеорологические условия в качестве входных данных для моделей "черного ящика"для прогнозирования PM2.5 и оценили эффективность этих моделей на различных горизонтах прогнозирования (30 дней, 90 дней и 180 дней).

Результаты обобщены в таблице 1.7.

Horizon	30			90			180		
Metrics	R^2	$RMSE$	MAE	R^2	$RMSE$	MAE	R^2	$RMSE$	MAE
XGBoost	0.9623	24.180	0.0144	0.9376	28.167	0.0239	0.9349	28.471	0.0210
LightGBM	0.9639	23.653	0.0148	0.9433	26.863	0.0227	0.9414	27.001	0.0203
Catboost	**0.9734**	**20.310**	**0.0128**	0.9297	29.893	0.0252	0.9349	28.471	0.0214
ANN	0.9278	33.298	0.0218	0.9498	26.305	0.0256	0.9304	28.036	0.0264
RNN	0.9614	24.462	0.0149	0.9639	21.418	0.0206	0.9564	23.274	0.0185
LSTM	0.9659	22.973	0.0148	0.9521	24.677	0.0216	0.9531	24.149	0.0182
GRU	0.9327	32.307	0.0227	0.9513	24.878	0.0237	0.9520	24.433	0.0189
Bi-RNN	0.9603	24.807	0.0165	0.9570	23.375	0.0262	0.9576	22.970	0.0189
Bi-LSTM	0.9419	30.005	0.0212	**0.9661**	**20.740**	**0.0196**	0.9576	22.949	0.0185
Bi-GRU	0.9648	23.345	0.0153	0.9656	20.912	0.0214	**0.9589**	**22.601**	**0.0182**

Таблица 1.7: Сравнение эффективности прогнозирования. Оптимальная модель фильтруется через валидационный набор и сравнивается с тестовым набором. Значения метрик эффективности, представленные здесь, взяты из тестового набора.

Анализ показывает, что Catboost лучше всего работает при горизонте прогнозирования 30 дней, Bi-LSTM обеспечивает наивысшую точность при 90 днях, а Bi-GRU - при 180 днях. Если рассматривать типы моделей, то модели ансамблевого обучения демонстрируют лучшую производительность для более коротких горизонтов прогнозирования, в то время как модели рекуррентных нейронных сетей, включая их производные двунаправленные аналоги, показывают лучшую точность для более долгосрочного прогнозирования. Напротив, модель многослойного перцептрона не показала достойных результатов в нашем исследовании.

Анализ показывает, что Catboost лучше всего работает при горизонте прогнозирования 30 дней, Bi-LSTM обеспечивает наивысшую точность при 90 днях, а Bi-GRU - при 180 днях. Если рассматривать типы моделей, то модели ансамблевого обучения демонстрируют лучшую производительность для более коротких горизонтов прогнозирования, в то время как модели рекуррентных нейронных сетей, включая их производные двунаправленные аналоги, показывают лучшую точность для более долгосрочного прогнозирования. Напротив, модель многослойного перцептрона не показала достойных результатов в нашем исследовании.

1.3 Вывод из главы 1

По сравнению с точностью традиционной линейной регрессии и моделей "белого ящика модели "черного ящика такие как нейронные сети и ансамблевые модели, включая алгоритмы повышения эффективности и алгоритмы упаковки, обладают абсолютными преимуществами в прогнозировании, однако необъяснимый характер делает модель "черного ящика"препятствием в процессе практического применения. Алгоритмы XAI, очевидно, могут решить эту проблему. Выявляя функции с относительно высоким вкладом, пользователи могут более четко понимать модель черного ящика при ее использовании для прогнозирования, тем самым повышая доверие. Однако, поскольку многие алгоритмы XAI имеют разные принципы и характеристики, разные алгоритмы XAI выдают разные результаты для одной и той же модели черного ящика. Поэтому необходимо разработать систему оценки алгоритмов XAI. Алгоритмы XAI оцениваются с помощью установленной системы оценки XAI - MDMC. Результаты показывают, что LIME больше подходит для моделей ANN и random forest, основанных на алгоритме формирования пакетов, а SHAP больше подходит для LightGBM, основанного на алгоритме повышения производительности.

Глава 2

Алгоритмы объяснимого искусственного интеллекта для вычисления важности периодов времени

Применение объяснимого искусственного интеллекта (XAI) в прогнозировании временных рядов постепенно привлекало внимание, учитывая широкое внедрение машинного обучения и глубокого изучения. ShapTime - общий подход XAI, основанный на значении Шепли, специально разработанный для прогнозирования объяснимых временных рядов, который позволяет исследовать более обширную информацию во временном измерении, вместо того, чтобы только грубо применять традиционные подходы XAI к прогнозированию временных рядов, как в предыдущих работах. **Результаты исследования были опубликованы в докладе конференции [23, 26]**. **Новизна** нашего метода заключается в том, что он позволяет получить объяснение в измерении временных рядов, т.е. позволяет выявить важность исторических данных для результатов прогнозирования, что невозможно при использовании других методов, не зависящих от модели.

2.1 Описание существующих проблем при вычислении важности периода времени

2.1.1 Отсутствие универсальности

Многочисленные соревнования по прогнозированию временных рядов, включая M4 [77] и M5 [78], показали, что ML и DL работают значительно

лучше, чем традиционные статистические методы, особенно для более сложных задач. Это привело к исследованию применения объяснимого ИИ (XAI) в прогнозировании временных рядов. Прогнозирование объяснимых временных рядов направлено на повышение надежности ML и DL в таких областях, как финансы, энергетика и метеорология. Существует два основных подхода к применению XAI в моделях прогнозирования временных рядов: (1) прямое использование существующего метода, не зависящего от модели, с высокой степенью общности; (2) разработка метода, специфичного для конкретной модели. Эти два подхода напрямую порождают две ключевые проблемы.

***Узкое место 1**: При прогнозировании временных рядов существующий метод диагностики моделей применяется грубо, что приводит к недостаточному объяснению.* Основная причина недостаточного объяснения заключается в том, что большинство существующих методов диагностики моделей - это методы атрибуции признаков, учитывая, что XAI изначально разрабатывался на основе задач регрессии и классификации, таких как SHAP [51], LIME [52] и т.д. В отличие от данных временных рядов, в задачах регрессии и классификации отсутствует временная связь между экземплярами данных, поэтому соответствующий метод диагностики модели уделяет больше внимания важности (или вкладу) признака. Однако для прогнозирования временных рядов этого недостаточно. В моделях прогнозирования временных рядов $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \omega_t$; $\omega_t = b + k_1 x_1 + k_2 x_2 + \cdots + k_n x_n$. y_{t-i} - это исторические данные целевой переменной, а ω_t состоит из признаков x_i и членов перехвата b. Эти традиционные модельно-диагностические методы могут вывести только важность признаков k_i, но не важность самого времени ϕ_i. На самом деле, в силу когнитивной инерции человека, даже в модельно-специфических методах принято использовать важность признака в качестве результата объяснения.

***Узкое место 2:** Специально разработанный метод, специфичный для модели, должен быть встроен в модель, что приводит к низкой общности и высокой стоимости применения.* В ряде работ была замечена проблема 1 и разработаны некоторые объяснимые модели прогнозирования временных рядов, которые могут показывать периодичность, а также тренд, который

действует на временном измерении. Однако эти методы все еще не способны вывести ϕ_i, а степень объяснения ограничена. Помимо вышеперечисленных проблем, существует также общая проблема в области XAI, которая в данной работе обозначена как проблема 3.

Узкое место 3: *Отсутствие сценариев применения.* В работах, посвященных XAI, обычно результаты объяснения представляются только как инновация. Во многих предыдущих работах сценарии применения XAI упоминаются только в вводной части, включая помощь пользователям в доверии к модели и отладку модели разработчиками, но эти сценарии не реализуются.

В целом, с одной стороны, текущий метод, не зависящий от модели, с высокой степенью общности не может полностью объяснить задачу прогнозирования временных рядов, то есть он не может реализовать объяснение временного измерения. С другой стороны, модельно-специфический метод, который в определенной степени может объяснить временное измерение, также имеет ограничения, а его низкая обобщенность также увеличивает стоимость использования. Поэтому необходим общий XAI-подход к объясняемому прогнозированию временных рядов, учитывая растущую важность ML и DL в прогнозировании временных рядов.

ShapTime[1] реализует атрибуцию времени путем вычисления значения Шэпли [57] на временном измерении, и в итоге выводит важность времени ϕ_i. Таким образом, ShapTime может реализовать объяснение во временном измерении и относится к модельно-агностическому методу, что означает, что он может быть развернут на любой модели прогнозирования с меньшими затратами. Основой ShapTime является значение Шэпли, которое лежит в основе многочисленных методов атрибуции, включая SHAP. Значение Шэпли пришло из теории кооперативных игр, которая изучает, как разумно распределить выгоды между игроками в альянсе, и доказало, что оно обладает некоторыми хорошими свойствами. Поэтому в последние годы развитие методов XAI на основе Shapley Value пытается стать стабильным путем [103], и наш ShapTime исследует объяснимое прогнозирование временных рядов как ответвление на этом пути. Его вклад включает:

[1]Библиотека github ShapTime

- Он реализует атрибуцию времени во временном измерении, то есть можно получить значение самого времени ϕ_i
- Как очень общий метод, не зависящий от модели, он может быть использован в любой модели прогнозирования
- Его результаты объяснения могут быть использованы в качестве руководства для улучшения эффективности прогнозирования модели

2.1.2 Высокая вычислительная сложность

Введение методов атрибуции признаков, таких как SHAP, направлено на решение проблемы интерпретируемости моделей "черного ящика". Среди различных методов атрибуции признаков SHAP выделяется как единственный метод, который удовлетворяет нескольким желательным свойствам. Кроме того, соответствующая программная библиотека хорошо совместима с алгоритмом Gradient boosting. В результате сочетание SHAP и Gradient boosting получило широкое распространение в отрасли благодаря своему удобству. Ежегодно публикуется значительное количество прикладных исследовательских работ, основанных на этой комбинации и посвященных задачам, связанным с прогнозированием временных рядов.

Однако первоначальная разработка SHAP была ориентирована в основном на задачи регрессии и классификации. Она предполагает, что выборки являются независимыми и одинаково распределенными (i.i.d.), поэтому может объяснить только влияние признаков X_i на y_i (рис. 1), т.е, прогнозное значение $\hat{y}_i = \Phi_i X_i = \phi_1 x_1 + \cdots + \phi_S x_S$, но не влияние исторических данных $Y_L = \{y_{i-L}, \cdots, y_{i-L}\}$ на y_i.

Эту проблему можно решить, перенеся Y_L из временной размерности в размерность признаков. Другими словами, исходные данные обрабатываются как $t_i = \{X_i, Y_L, y_i\}$. Таким образом, исторические данные Y_L из y_i интегрируются в размерность признаков. Следовательно, используя SHAP, прогнозное значение $\hat{y}_i$ может быть отнесено как $\hat{y}_i = \Phi_i X_i + \Psi_L Y_L$, $\Psi_L Y_L = \psi_{i-L} y_{i-L} + \cdots + \psi_{i-1} y_{i-1}$, что позволяет объяснять во временном измерении. Однако, когда нам нужно объяснить более длинный исторический диапазон данных во временном измерении, это приводит к высокоразмерным данным. Например, при временной частоте 1 час и

желании исследовать влияние исторических данных за прошлый год на текущее значение y_i, то есть $L = 24 \times 365 = 8760$, результирующая размерность обработанных временных данных становится $8760 + S$. Аналогично, при временной частоте 15 минут, L достигнет $96 \times 365 = 35040$. Такая высокая размерность данных создает две проблемы для достижения объяснимости:

Проблема 1: Она требует значительных вычислительных ресурсов, вплоть до того, что вычисления могут стать невыполнимыми. Основным алгоритмом SHAP является значение Шэпли, а его вычислительная сложность составляет 2^F (где F - общее количество признаков). Несмотря на существование в текущей библиотеке SHAP нескольких вариантов алгоритмов, позволяющих снизить сложность, высокоразмерные данные по-прежнему представляют собой значительную нагрузку для их вычислений. Введение в SHAP и его варианты можно найти в разделе 3.1.

Проблема 2: Извлечение значимой информации из высокоразмерных объяснительных результатов становится затруднительным. Предполагая незначительные затраты вычислительных ресурсов, мы получили временное объяснение этих высокоразмерных данных. Если взять в качестве примера временную частоту 1 час и $L = 8760$, то для исторических данных y_i будут выведены окончательные объяснительные результаты $\Psi_L = \{\psi_{i-8760}, \cdots, \psi_{i-1}\}$. На самом деле, извлечь полезную информацию из таких плотных данных довольно сложно, а наличие аномальных колебаний в данных временного ряда может помешать пониманию этих объяснительных результатов.

Короче говоря, чтобы преодолеть узкое место SHAP в прогнозировании временных рядов, необходимо решить две вышеупомянутые проблемы. Поэтому необходимо иметь разумную схему сокращения размерности Y_L, которая могла бы сохранить временную информацию. Рисунок 2.1 интуитивно демонстрирует подход к решению этих проблем путем сокращения Y_L до Y_K.

Выбор K зависит от практических требований, например, если нас интересует определение дня в течение прошлой недели, оказавшего наибольшее влияние на текущий y_i, с временной частотой в 1 час, то

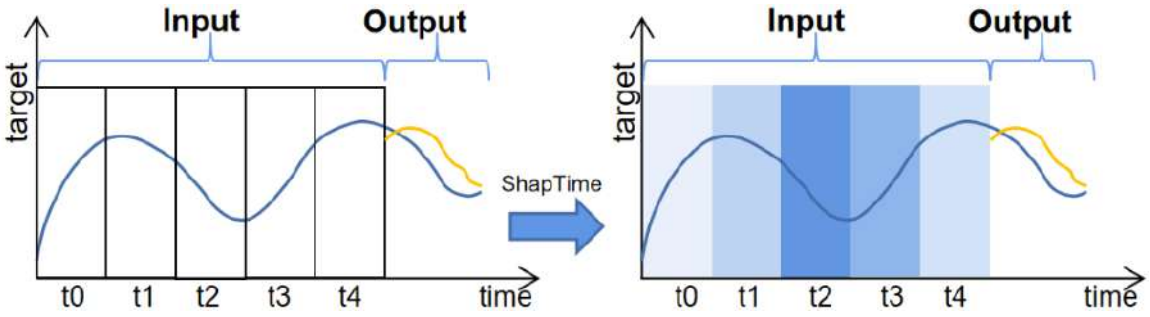


Рис. 2.1: Снижение размерности. В данном примере K = 5

$L = 7 \times 24 = 168$ и $K = 7$. В результате мы получаем новый набор данных $t_i = \{X_i, Y_K, y_i\}$, $Y_K = \{y_{T_0}, y_{T_1}, \cdots, y_{T_6}\}$. Вычисляя SHAP, $\hat{y}_i$ можно разложить в $\hat{y}_i = \Phi_i X_i + \Psi_K Y_K$, $\Psi_K Y_K = \psi_{T_0} y_{T_0} + \cdots + \psi_{T_6} y_{T_6}$, тем самым достигая нашей объяснительной цели.

2.2 ShapTime: алгоритм объяснимого ИИ с универсальностью и низкой вычислительной сложностью для вычисления важности периодов времени

Ценность Шэпли - одна из классических теорий кооперативных игр, цель которой - справедливое распределение выгоды между игроками в альянсе. Существует соответствие между Shapley Value и объяснением модели: признаки, используемые для обучения, соответствуют "игрокам а предсказания модели - "доходам". Таким образом, распределение, достигаемое с помощью Shapley Value, может приписать результат прогнозирования признакам, то есть вклад (важность) каждого признака в результат прогнозирования. Присваивание осуществляется по следующей формуле:

$$k_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \times (v(S \cup \{i\}) - v(S)). \tag{2.1}$$

где N - множество всех игроков (функций) (1,2,3...i... n), то есть полное множество; S - подмножество N, в котором удаляется объясняемый признак i, всего 2^N; v - функция выигрыша ($v(S) = E_{\hat{D}}[f(x) \mid x_S]$, где $\hat{D}$ - эмпирическое распределение обучающих данных, а f - черная модель).

Причина, по которой значение Шэпли стало основой многих подходов XAI, заключается в том, что оно обладает рядом желательных свойств, включая: Эффективность, Симметрия, Линейность и Нулевой игрок, и среди многих методов атрибуции это единственное отображение ($v : 2^N \to \mathbb{R}$), которое может удовлетворить вышеуказанным свойствам.

Значение Шэпли состоит из двух ключевых элементов: "игрок"и "функция выигрыша". Когда оба элемента определены, Shapley Value можно рассчитать теоретически. Соответственно, Shapley Value - это атрибуция "игрока".

Для достижения атрибуции по временному измерению в качестве "игроков"следует рассматривать временные точки. В соответствии с моделью прогнозирования временных рядов, y_{t-k} рассматривается как "игрок и после определения функции усиления, атрибуция y_{t-k}, а именно ψ_k, может быть реализована через Shapley Value. Модифицированная формула (ShapTime) имеет вид:

$$\psi_{t_k} = \sum_{S' \subseteq T \setminus t_k} \frac{|S'|!(|T| - |S'| - 1)}{|T|!} \times (v'(S' \cup t_k) - v'(S')). \tag{2.2}$$

где t_k - супервремя (игроков), T - множество всех супервремен t_k. S' - подмножество T, в котором удаляется объясненное супервремя t_k, всего 2^T. v' - функция выигрыша ShapTime.

2.2.1 Super-time: метод снижения вычислительной сложности

Так называемая временная атрибуция - это отнесение результатов прогнозирования ко времени, что означает рассмотрение временных точек как "игроков что неизбежно приведет к взрыву размерности и краху системы. Аналогичная проблема возникает и в подходе XAI для распознавания образов - при большом количестве пикселей вычислительные затраты растут экспоненциально. Чтобы решить эту проблему, [52] предложил концепцию суперпикселя, которая заключается в том, что пиксели с высокой степенью сходства объединяются в единое целое, а затем участвуют в вычислении вклада пикселя, тем самым значительно снижая стоимость вычислений.

Algorithm 1 Super-time

Вход: Исходный набор данных: X

Параметр: Количество Super-time: n

Вывод: Все Super-time t_i

1: Пусть $L = \text{int}(\text{length}(X)/n)$.
2: Пусть start = length(X) - $L \times n$.
3: Пусть $X = X\,[start :, :]$.
4: **for** i в диапазоне (n) **do**
5: $\quad t_i = X\,[i \times L : (i+1) \times L]$.
6: **end for**
7: возврат всех t_i

Ввиду схожести задачи, мы обращаемся к его методу построения супервремени во временном измерении, то есть набор данных разбивается на n супервремен t_i в соответствии с временным измерением.

$$t_i = \{y_{t-i}, y_{t-i-1}, y_{t-i-2}, \cdots\} \tag{2.3}$$

В ShapTime супервремя эквивалентно "игроку". Важно отметить, что хотя супервремя контролирует вычислительные затраты в приемлемом диапазоне, учитывая сложность $O(2^n)$ в ShapTime, мы рекомендуем, чтобы n не превышало 10 или 11. Алгоритм 1 показывает процесс построения Super-time.

2.2.2 Переопределение функции для универсальности

В ShapTime приписываемым объектом является супервремя, которое представляет собой совокупность временных точек в пределах периода времени, то есть "игрок это уже не временная точка в данный момент, а временной период.

Соответственно, наша цель прогнозирования y_t также является набором временных точек. Однако вычисление ценности Шэпли требует, чтобы для каждой комбинации "игроков"соответствующая функция усиления выдавала значение. Поэтому здесь в качестве функции усиления v' в ShapTime (строки 6 и 10 алгоритма 2) берется усредненный результат прогнозирования модели f.

Algorithm 2 Функция усиления ShapTime

Вход: Объясненная модель: f; Сверхвремя: t_i

Выход: Все значения выигрыша: v'

Пусть $S' \subseteq T \setminus t_i$.
Пусть $T = \{t_0, t_1, \cdots, t_n\}$.
for S' в T **do**
  if length(S') == 1 **then**
    $S' = \{t_i\}$.
    $v' = \text{sum}(f(t_i))/\text{length}(t_i)$.
  else
    $S' = \{t_i, t_j, \cdots\}$.
    $S_c = concat(t_i, t_j, \cdots)$.
    $v'(S_c) = \text{sum}(f(S_c))/\text{length}(S_c)$.
  end if
end for
возврат 2^T получаем значения: v'

Формула такова:

$$v'(S') = avg(f_{S'}(x_{S'})) \tag{2.4}$$

В конце концов, будет получено значение выигрыша v' для всех комбинаций супервремени, всего $2^{|T|}$. Таким образом, согласно уравнению (7), результат прогнозирования может быть отнесен к каждому супервремени t_i, то есть ϕ_{t_i}, чтобы реализовать объяснение во временном измерении.

2.2.3 Визуализация важности временных периодов

В настоящее время в области исследований XAI не хватает признанных правильных обозначений, что создает проблемы для оценки методов XAI. С другой стороны, наш ShapTime - это модельно-агностический метод объяснения временного измерения. В данной области такая форма объяснения встречается редко, что приводит к отсутствию объектов для сравнения, поэтому оценка по контрасту затруднена. Поэтому мы предлагаем несколько разумных критериев оценки, позволяющих в определенной степени эффективно оценить ShapTime. Это также дает базовый ориентир для последующих исследований.

***Свойство 1**: При условии использования одного и того же набора*

данных результаты объяснения с помощью подхода XAI для схожих типов моделей должны быть в определенной степени согласованы.

Свойство 2: *XAI-подход должен быть подвержен анализу чувствительности, то есть при возмущении жизненно важных "игроков"происходит значительное падение эффективности прогнозирования.*

Критерии оценки XAI очень важны для того, чтобы пользователи могли доверять машинам. Самый важный из них заключается в том, что результаты объяснения XAI должны сохранять определенную степень стабильности (свойство 1), что является основой для формирования доверия. Исходя из этого, метод XAI также гарантирует валидность (свойство 2), то есть при смене важных "игроков"производительность модели значительно снизится. Это в определенной степени доказывает, что результаты объяснения, полученные с помощью данного XAI-подхода, достоверны.

Для того чтобы проверить возможности практического применения ShapTime, мы выбрали 5 реальных наборов данных, включая: Климатические данные, Потребление энергии, Солнечная энергия, Цена золота, Акции Tesla. Среди них первые три являются периодическими данными, а последние два - трендовыми, чтобы проверить работу ShapTime при различных типах данных.

С другой стороны, существует два типа моделей "черного ящика участвующих в обучении: Бустинговая модель и модель на основе RNN. Первая состоит из XGBoost и LightGBM, вторая включает RNN, LSTM, GRU (на основе RNN) и Bi-RNN, Bi-LSTM, Bi-GRU (на основе Bi-RNN). Эти модели прогнозирования в основном включают в себя основные методы, используемые в настоящее время в конкурентной борьбе и на практике. Показатели эффективности их прогнозирования приведены в таблице 2.1

Мы применили ShapTime к 8 моделям и получили результаты объяснения (рисунок 2.2): Для краткости мы показываем только результаты объяснения ShapTime для XGBoost и LSTM на 5 наборах данных, а полные результаты объяснения находятся на Github).

В качестве примера для объяснения результатов взят набор данных Daily

Data	Climate	Energy	Gold	Solar	Tesla
XGB	0.7756	0.7570	**0.7988**	0.9307	0.7632
	(0.0081)	0.0095	0.0006	0.0055	0.0029
LGB	**(0.7814)**	**(0.8290)**	0.7779	**0.9322**	**0.7777**
	0.0082	**(0.0071)**	**0.0006**	**0.0053**	**0.0023**
RNN	0.7170	0.6574	0.7018	0.9587	0.8133
	0.0104	0.0128	0.0011	0.0051	0.0022
LSTM	**0.7507**	0.6249	**0.8182**	0.9544	**0.8586**
	0.0081	0.0153	**0.0005**	0.0057	**0.0017**
GRU	0.6719	**0.7130**	0.7096	**(0.9661)**	0.8176
	0.0117	**0.0119**	0.0009	**(0.0035)**	0.0025
Bi-R	**0.7257**	0.6804	0.8048	0.9312	0.7990
	0.0102	0.0120	0.0008	0.0099	0.0029
Bi-L	0.6903	**0.7122**	**(0.8791)**	0.9326	**(0.8689)**
	0.0103	**0.0119**	**(0.0005)**	0.0082	**(0.0016)**
Bi-G	0.7136	0.5927	0.8664	**0.9481**	0.7055
	0.0109	0.0154	0.0006	**0.0056**	0.0045

Таблица 2.1: Метрики эффективности прогнозирования. Модели прогнозирования делятся на три категории: Boosting; RNN-based; Bi-RNN-based. В качестве метрик эффективности прогнозирования используются r^2 и MSE, для каждой модели первая строка - r^2, вторая - MSE. Внутри каждого класса моделей лучшая модель выделена (жирным шрифтом), среди всех моделей лучшая модель выделена дополнительно (в скобках). Как видно из таблицы, модель на основе Bi-RNN лучше всего работает с трендовыми данными (цена на золото и акции Tesla), а Boosting - с периодическими данными (ежедневный климат, потребление энергии и солнечная генерация).

Climate (Figure.2.2(a)(f)). В этом примере число n супервремен задано равным 8, тогда соответствующая теоретическая модифицированная модель временного ряда имеет вид:

$$y_t = \phi_{t_0} t_0 + \phi_{t_1} t_1 + \cdots + \phi_{t_7} t_7 + \omega_t \tag{2.5}$$

После завершения обучения мы используем ShapTime для объяснения XGBoost и LSTM, чтобы приписать результаты прогнозирования модели к каждому супервремени, а затем получить каждое значение ϕ_{t_i} и визуально отобразить их с помощью тепловой карты. Мы можем ясно видеть, что обе модели фиксируют самое последнее супервремя t_7 как наиболее важный вход. В наборе данных Energy Consumption обе модели считают t_7 наиболее важным супервременем, в Gold Price - t_9, в Tesla Stock - t_{10}. Иные ситуации

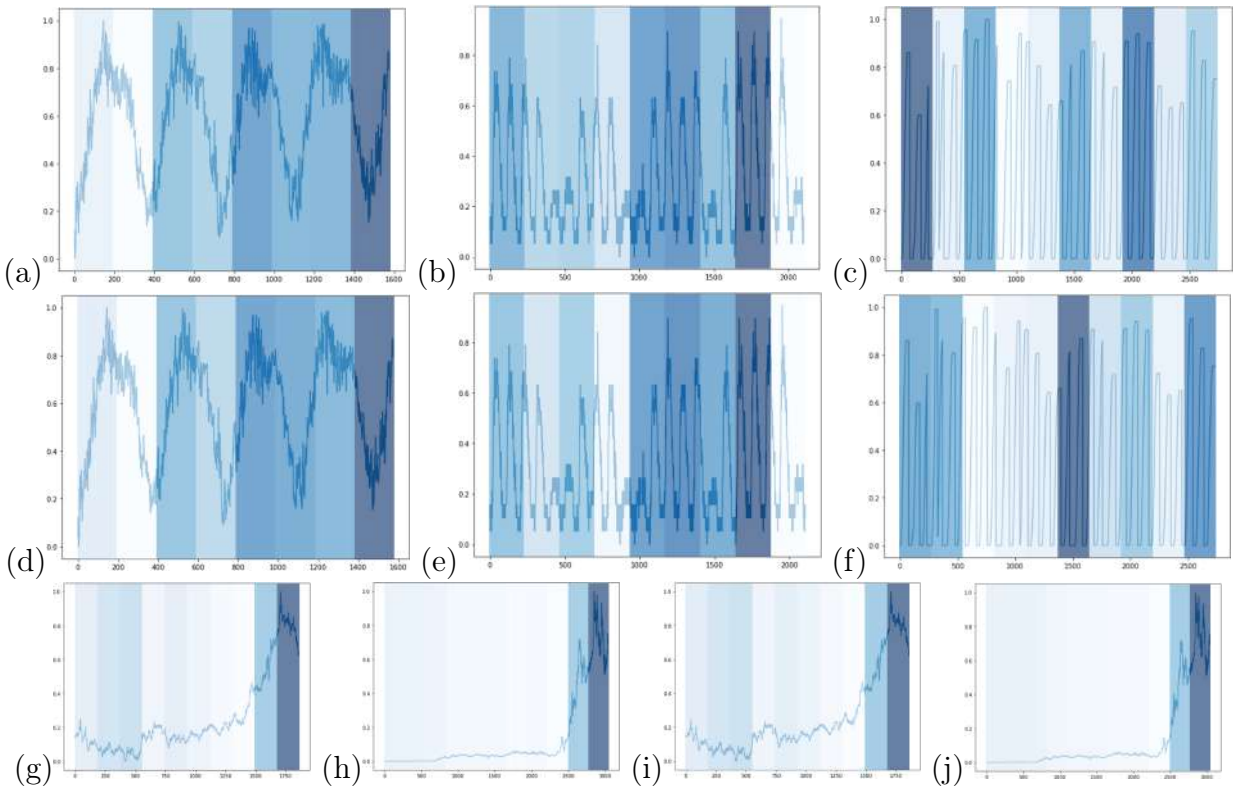


Рис. 2.2: Пример результатов объяснения от ShapTime. Это объяснение XGBoost и LSTM от ShapTime. a: XGBoost-Climate; b: XGBoost-Energy; c: XGBoost-Solard; d: LSTM-Climate; e: LSTM-Energye; f: LSTM-Solar; g: XGB-Gold; h: XGB-Tesla; i: LSTM-Gold; j: LSTM-Tesla. Тепловая карта используется для визуализации результатов объяснения входных данных. Чем темнее цвет, тем важнее супервремя. Это означает, что модель уделяет больше внимания этому супервремени в процессе обучения. Из рисунков видно, что объяснение трендовых данных с помощью LSTM и XGBoost в основном совпадает, а в объяснении периодических данных есть определенная разница, особенно в наборе данных "Солнечная генерация".

возникают в Solar Generation. Супервремя, захватываемое двумя моделями, отличается. XGBoost фиксирует t_0 как наиболее важное, а LSTM - t_5.

Если посмотреть на все результаты объяснения 5×8, то наиболее важным супервременем, которое улавливают все модели, является последнее время в трендовых данных, но в периодических данных нет очевидного правила. Однако, если мы проанализируем эффективность прогнозирования модели (табл.2.1), мы все же сможем найти некоторые потенциальные правила, то есть модель Boosting больше подходит для периодических данных, а модель на основе RNN больше подходит для периодических данных, когда результаты объяснения в целом совпадают.

В соответствии с вышеуказанными критериями оценки, мы оцениваем результаты объяснения ShapTime, и примеры оценки показаны на рисунке

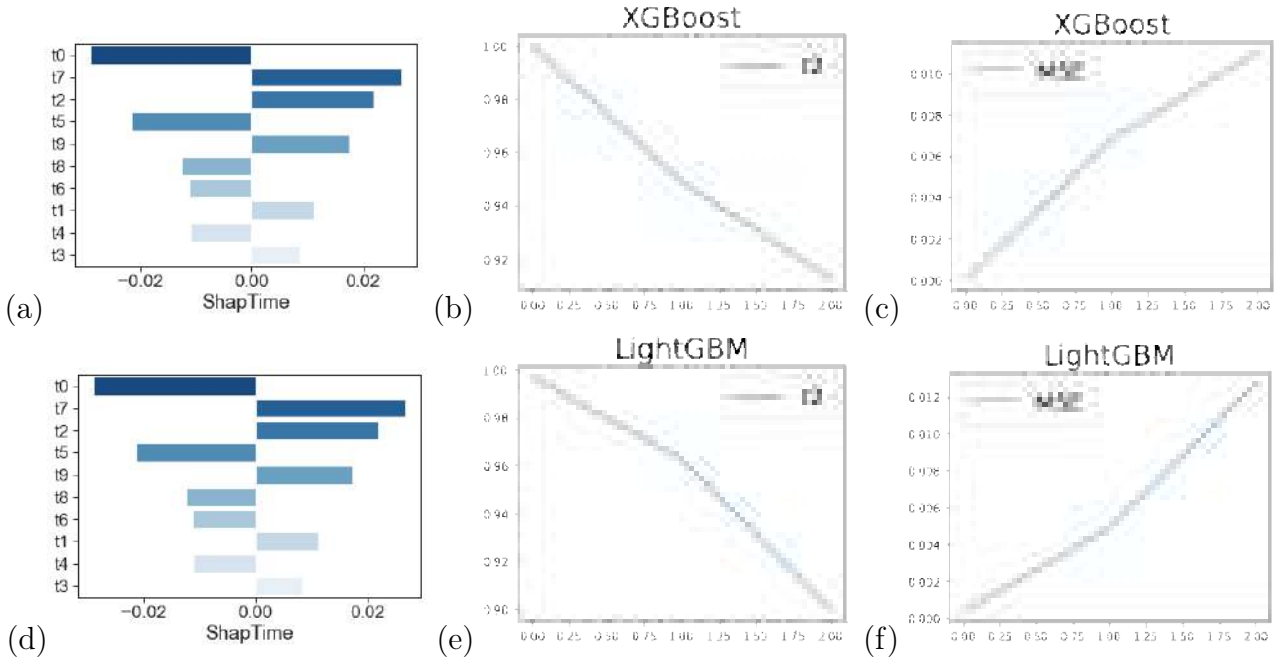


Рис. 2.3: Пример оценки результатов объяснения ShapTime. a: XGB-Solar; b: XGB-r^2; c: XGB-MSE; d: LGB-Solar; e: LGB-r^2; f: LGB-MSE. (a) и (b) удовлетворяют свойству 1, (b)(c) и (e)(f) удовлетворяют свойству 2.

2.3. Аналогично, для простоты, мы отображаем только результаты оценки модели Boosting. На рисунке 2.3(a)(d) представлены объяснения ShapTime для XGBoost и LightGBM, соответственно. Интуитивно видно, что объяснение ShapTime может сохраняться все время для одного и того же типа моделей, что является общей закономерностью во всех результатах объяснения. Это является ключевым фактором доверия пользователей из всех отраслей. Представьте себе, если результаты объяснения будут часто меняться в процессе использования, что приведет к недоверию пользователей или даже отказу от использования.

Именно для этой цели и было определено свойство 1. Однако в случае разных типов моделей в принципе необходимо допускать различия в результатах объяснения, поскольку модели разной архитектуры не одинаково чувствительны к характеру распределения данных. Как показано на рисунке 2.2(c)(h), результаты объяснения для различных моделей отличаются.

С другой стороны, теоретически, если мы возмутим обучающие данные в соответствии с результатами объяснения, то есть заменим важное супервремя на наименее важное, то эффективность прогнозирования модели значительно снизится. В данном примере наиболее важное t_0

заменяется на наименее способствующее t_3, а также t_7 заменяется на t_4. Соответственно, r^2 и MSE XGBoost и LightGBM демонстрируют значительное и постепенное снижение производительности (рис. 2.3(b)(c)(e)(f)). Эта схема оценки является относительно классическим методом оценки в области XAI, и она также известна как анализ чувствительности. В данной работе она обобщена в свойстве 2.

2.2.4 Повышение точности прогнозирования с помощью ShapTime

Несмотря на то, что в последние годы исследованиям в области объяснимости уделяется все больше внимания, в целом результаты объяснения представляются просто как образцы. Это явление наблюдается не только в области объяснимого прогнозирования временных рядов, но и является общей проблемой для всей области объяснимого ИИ (узкое место 3). Чтобы изучить сценарии применения подхода XAI, мы пытаемся использовать результаты объяснения ShapTime в качестве руководства, направленного на улучшение эффективности прогнозирования временных рядов.

Сначала входной набор данных делится поровну на несколько супервремен, затем с помощью ShapTime вычисляется важность каждого супервремени и визуализируется на тепловой карте, где более темный цвет означает большую важность, а более светлый - меньшую. Способ улучшения производительности заключается в замене малозначимых t_k на высокозначимые t_k и использовании этого нового набора данных для переобучения модели. Причина построения такой схемы улучшения (рисунок 2.4) заключается в том, что ShapTime на основе значения Шэпли - это метод, не зависящий от модели, поэтому логика его работы основана на наборе данных. В деталях, возмущение входных данных изменит результат прогнозирования, и этот тип метода реализует атрибуцию через это изменение. Фактически, $(v'(S' \cup t_k) - v'(S'))$ представляет собой это изменение. Поэтому при использовании результатов объяснения, полученных с помощью этого метода, естественным образом принимаются уточнения в зависимости от набора данных.

Основываясь на разработанных нами подходах, FI-SHAP и ShapTime, мы даем объяснения модели с двух разных точек зрения. Первый фокусируется

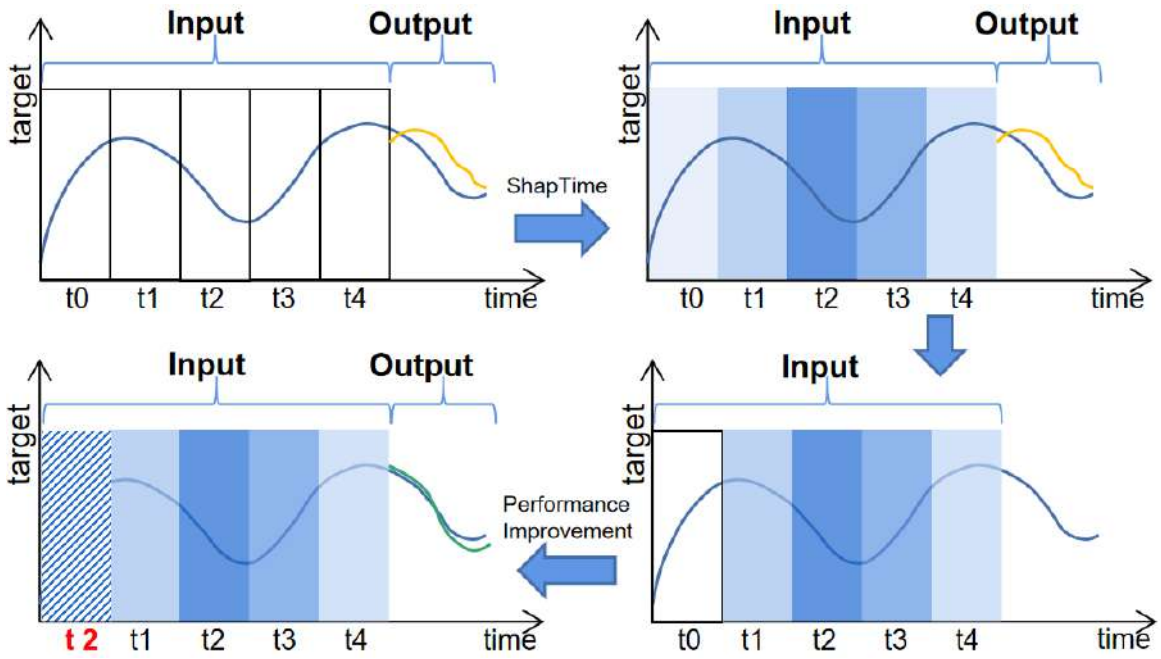


Рис. 2.4: Процесс повышения производительности на основе ShapTime. Синяя линия представляет исходный набор данных, а оранжевая - результаты прогнозирования модели. Синий фон представляет собой результат объяснения ShapTime. Чем темнее цвет, тем важнее t_k, и наоборот. После получения важности во временном измерении t_k с наименьшим значением заменяется на наибольшее. Зеленая линия представляет собой улучшенные результаты прогнозирования.

на уровне признаков (или переменных) и повышает производительность за счет усовершенствованных методов проектирования признаков. С другой стороны, второй работает на временном уровне и повышает производительность за счет методов дополнения данных.

Результаты улучшения

Чтобы создать ценные сценарии применения XAI, мы используем результаты объяснения ShapTime в качестве руководства для достижения улучшенной производительности прогнозирования, и процесс улучшения показан на рисунке 2.4. В таблице 2.2 показаны метрики улучшенной производительности (по сравнению с таблицей 2.1), и результаты показывают, что Bi-RNN-based и Boosting по-прежнему сохраняют свои первоначальные преимущества для трендовых и периодических данных, соответственно.

Мера степени улучшения представлена на рисунке 2.5 (a), где показана средняя степень улучшения для каждого типа модели в каждом наборе

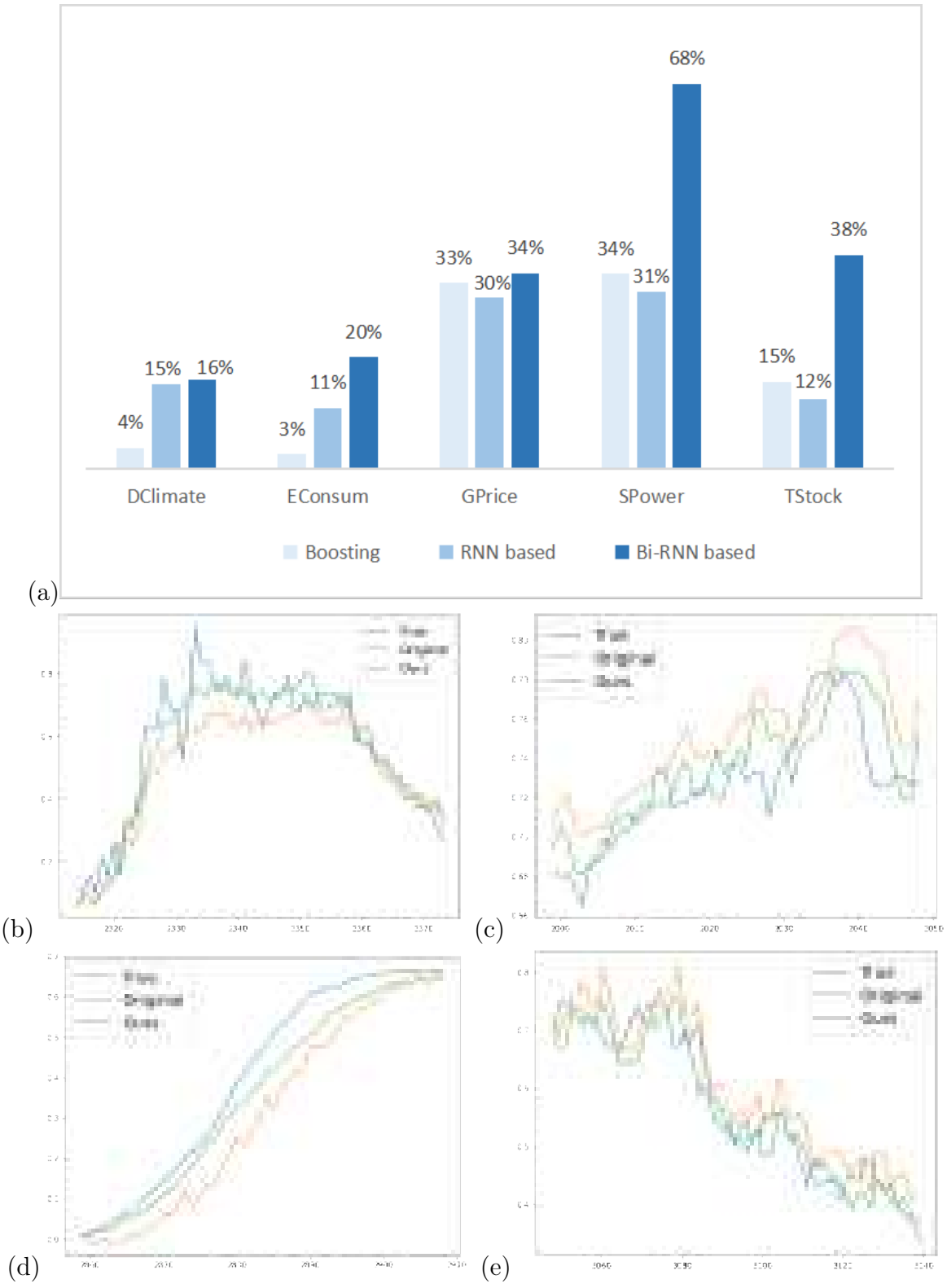


Рис. 2.5: Визуализация улучшения производительности. a: Результаты улучшения производительности прогнозирования; b: Улучшить Energy; c: Улучшить Gold; d: Улучшить Solar; e: Улучшить Tesla. Original - результат оригинальной модели прогнозирования, а Ours - результат модели прогнозирования, управляемой ShapTime.

Data	Climate	Energy	Gold	Solar	Tesla
XGB	**(0.7898)**	0.7975	**0.8880**	0.9521	**0.8198**
	0.0075	0.0093	**(0.0004)**	0.0038	**0.0021**
LGB	0.7847	**(0.8445)**	0.8715	**0.9573**	0.8038
	0.0082	**(0.0069)**	0.0004	**0.0033**	0.0022
RNN	0.7255	**0.7572**	**0.8535**	0.9701	0.8385
	0.0096	0.0116	0.0006	**0.0024**	0.0020
LSTM	**0.7751**	0.6412	0.8287	0.9597	**0.8812**
	(0.0070)	0.0142	**0.0005**	0.0047	**0.0016**
GRU	0.7470	0.7376	0.7825	**0.9740**	0.8498
	0.0092	**0.0101**	0.0006	0.0026	0.0020
Bi-R	0.7428	0.7425	0.8980	0.9693	0.8932
	0.0082	0.0101	0.0004	0.0036	0.0016
Bi-L	**0.7532**	**0.7604**	0.9002	0.9742	**(0.9080)**
	0.0084	**0.0100**	0.0004	0.0028	**(0.0013)**
Bi-G	0.7412	0.7092	**(0.9043)**	**(0.9840)**	0.8505
	0.0100	0.0112	**0.0004**	**(0.0014)**	0.0021

Таблица 2.2: Улучшение эффективности прогнозирования с помощью ShapTime. В качестве показателей эффективности прогнозирования используются r^2 и MSE, для каждой модели первая строка - r^2, вторая - MSE.

данных. Проценты рассчитаны на основе степени улучшения в таблице 2.2 по сравнению с таблицей 2.1. Интуитивно видно, что ShapTime показывает максимальное улучшение для модели на основе Bi-RNN и наиболее значительное улучшение для Solar Generation. В целом, влияние ShapTime на Boosting менее значительно, чем на другие типы моделей. В частности, ShapTime имеет самое значительное улучшение для Bi-GRU. Если в исходном варианте прогнозирования (таблица 2.1) Bi-GRU не обладает наилучшими показателями во всех наборах данных, то после улучшения (таблица 2.5) наилучшие показатели прогнозирования цены золота и солнечной генерации демонстрирует Bi-GRU.

Частичная визуализация эффекта улучшения показана на рисунке 2.5. Судя по эффекту прогнозирования, результаты прогнозирования после улучшения ShapTime все еще сохраняют примерно ту же картину, что и исходные результаты прогнозирования, однако они могут быть гораздо ближе к исходным данным, что позволяет добиться улучшения

производительности.

2.3 Вывод из главы 2

В этой главе разработан подход XAI, специально ориентированный на прогнозирование временных рядов, и мы называем его ShapTime, поскольку его расчет основан на значении Шепли. Он позволяет проводить атрибуцию во временном измерении, тем самым объясняя важность самого времени, что отличается от предыдущих работ. С другой стороны, с помощью точного определения времени мы смогли добиться повышения производительности при прогнозировании временных рядов. Заменяя данные в периоды с низким вкладом на данные с высоким вкладом, можно в некоторой степени повысить производительность.

Глава 3

Алгоритмы объяснимого искусственного интеллекта для вычисления важности признаков

Вышеупомянутые результаты исследований указывают на то, что как коллективное обучение, представленное бустингом, так и глубокое обучение, представленное нейронными сетями, важны в области прогнозирования временных рядов, особенно для долгосрочного прогнозирования последовательности событий. Однако обе они являются моделями черного ящика, которые не могут быть поняты людьми. В этой главе мы сосредоточимся на бустинговой модели как на представителе коллективного обучения. В области прогнозирования временных рядов применение бустинговых моделей часто требует разработки функциональных возможностей. Таким образом, пояснения к моделям повышения производительности могут стать эффективным руководством для разработки функциональных возможностей, что приведет к повышению производительности.

Алгоритм бустинга (BA) является передовым в основных состязаниях, особенно в состязаниях по прогнозированию временных рядов M4 и M5. Однако использование BA требует кропотливой работы по разработке функций в условиях слепоты и случайности, что приводит к серьезной потере времени. В этой работе мы пытаемся руководить начальными операциями по проектированию объектов на основе результатов, полученных с помощью метода SHAP, при этом также учитывается традиционный метод важности признаков (FI). **Результаты**

исследования были успешно опубликованы [20]. **Новизна** нашего метода заключается в том, что сочетание этих двух методов позволяет сделать объяснение результатов более информативным в неявном виде, что помогает оптимизировать точность прогнозирования.

3.1 Генерация признаков на основе важности признаков

Задача генерации признаков заключается в обогащении информации набора данных, чтобы модель прогнозирования могла получить больше "знаний"и показать лучшую производительность [115]. Существующие методы инженерии признаков в задачах прогнозирования временных рядов делятся на две категории: экзогенные и эндогенные. Экзогенная схема заключается в добавлении признаков, которые могут повлиять на временной ряд; эндогенная схема заключается в обогащении информации набора данных путем извлечения скрытых признаков из исходных признаков. В данной работе рассматривается только эндогенная схема.

Проектирование признаков не имеет фиксированного маршрута выполнения. Она часто требует, чтобы специалисты разрабатывали признаки, основываясь на своих профессиональных знаниях, поэтому она чрезвычайно зависит от человеческого опыта. Это ненадежно и отнимает много времени. Для решения этой проблемы было разработано большое количество автоматических методов разработки признаков, таких как [116–120]. Эти основные методы используют несколько групп признаков, содержащихся в исходном наборе данных, для поиска новых релевантных признаков, и все они в большей степени ориентированы на задачи классификации. Несмотря на то, что эти методы могут частично использоваться для прогнозирования временных рядов, они не отражают временные особенности. В отличие от этого, наш метод инженерии признаков, основанный на XAI, специально разработан для данных временных рядов и может отражать временные признаки.

Методы инженерии признаков для прогнозирования временных рядов включают [121–123]. Они, по сути, добавляют признаки запаздывания на основе вышеупомянутых методов, то есть вводят признаки авторегрессии, чтобы обеспечить достаточное количество временных признаков в обучении.

Кроме того, отбор признаков осуществляется путем построения большого количества признаков (включая признаки авторегрессии) заранее и вычисления важности признаков для модели после прогнозирования, как, например, [124–126]. Однако этот процесс имеет определенную степень слепоты и случайности. Однако наш метод проектирования признаков на основе XAI способен количественно рассчитать порядок необходимых запаздываний, тем самым устраняя эту слепоту и случайность.

3.1.1 Построение признаков запаздывания временного ряда

Для разработки признаков в прогнозировании временных рядов построение запаздывающих признаков является чрезвычайно важным. Во-первых, теоретически прогнозирование временных рядов является задачей авторегрессии, то есть использования собственных исторических данных для предсказания будущих данных, поэтому построение лаговых функций является обязательным для задач прогнозирования временных рядов. Во-вторых, чем больше лаговых признаков, тем не лучше, а слишком большое количество лаговых признаков приведет к снижению эффективности прогнозирования. Это связано с тем, что авторегрессионный процесс приводит к накоплению ошибок, а значит, чем больше лаговых признаков, тем больше ошибок будет накапливаться. Поэтому построение подходящего числа запаздывающих признаков особенно важно при прогнозировании временных рядов. Разработанная нами автоматическая инженерия признаков будет направлена на построение запаздывающих признаков при прогнозировании временных рядов.

3.1.2 Недостатки существующих алгоритмов вычисления важности признаков для генерации признаков

Популярность машинного обучения и глубокого обучения привела к тому, что все большее внимание уделяется объяснимому искусственному интеллекту (XAI). Модели машинного и глубокого обучения обычно рассматриваются как "черные ящики"с внутренними неизвестными характеристиками. Поэтому при применении таких моделей очень важно завоевать доверие человека, прояснить конкретный смысл их ошибок и надежность их предсказаний.

Важность признака в модели boosting - важная часть инженерии признаков. С одной стороны, она обладает совершенными математическими теоретическими знаниями [127]; с другой стороны, она применима практически ко всем моделям деревьев и чрезвычайно удобна. Важность признака - это, по сути, информационный выигрыш, который используется для отбора признаков при разбиении дерева решений, а его расчет основан на энтропии Шеннона. Признаки с большим информационным коэффициентом считаются важными. Процесс вычисления информационного выигрыша определяется как:

Ожидаемая информация (энтропия Шеннона):

$$Info(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \tag{3.1}$$

Получение информации:

$$Gain = Info(X) - \sum_{i=1}^{n} \frac{|X^i|}{X} Info(X^i) \tag{3.2}$$

X - случайная величина, P - вероятность всех случаев.

Как уже говорилось выше, чтобы построить нужное количество функций запаздывания, мы должны заранее знать их важность. В дополнение к функции FI, которая поставляется с моделью форсирования, XAI также является способом, который стоит попробовать.

Метод объяснения достигает цели объяснения путем построения более простой и понятной модели, позволяя ей постоянно приближаться к модели, которую нужно объяснить. Такой метод называется Post-Hoc, он отличается от Intrinsic, который интегрирует интерпретируемые функции в модель "черного ящика". Первый редко опирается на архитектуру модели "черного ящика"и может широко использоваться в моделях, прошедших обучение. Среди них обобщенные аддитивные модели [129], список правил Байеса [128] и нейроаддитивная модель [130] относятся к Intrinsic, и их работа тесно связана с моделью черного ящика. SHAP [51] и LIME [52] - два чрезвычайно популярных объяснения, диагностирующих модель (Post-Hoc). В данной работе мы рассматриваем только SHAP благодаря его полному хранилищу кода и строгой математической теории.

В основе создания SHAP лежит значение Шэпли, которое рассматривает каждый признак как "игрока чтобы построить систему, в которой

"одиночный игрок (отдельный признак)"и "альянс (комбинация признаков)"участвуют в "игре (модель черного ящика)". SHAP фактически приписывает выходное значение к значению формы каждой функции. Другими словами, он вычисляет значение Шэпли для каждого признака и на основе этого измеряет влияние признака на конечное выходное значение. Для линейных моделей с независимыми признаками сумма вкладов всех признаков выборки равна предсказанному значению минус среднее предсказанное значение, но для алгоритмов бустинга это явно не так. Поэтому для признаков буст-модели необходимо вычислить значение Шэпли для всех возможных комбинаций признаков (включая различные порядки), а затем взвесить и просуммировать, что определяется как:

$$\phi(val) = \sum_{S \subseteq \{x_1 \cdots x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} (val(S \cup \{x_j\}) - val(S)) \tag{3.3}$$

где S - подмножество признаков, используемых в модели, x - вектор значений признаков объясняемой выборки, p - количество признаков, а $val(S)$ - выходное значение модели при комбинации признаков S.

Мы можем количественно построить переменные запаздывания, основываясь на результатах объяснения XAI, что позволяет автоматизировать разработку признаков. Общий процесс показан на рисунке 3.1.

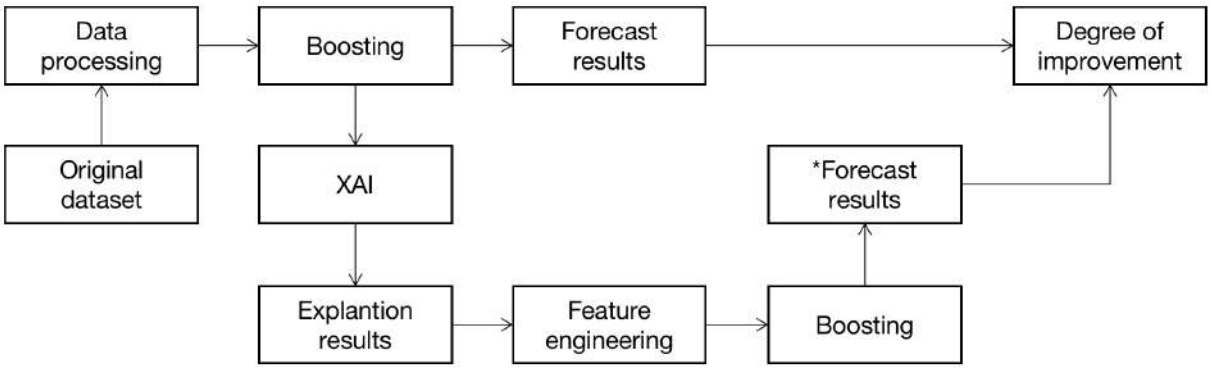


Рис. 3.1: Процесс разработки признаков, который XAI использует для прогнозирования временных рядов. Теоретически часть XAI может быть заменена другими методами, способными выдавать значение (вклад) признака. Результатом интерпретации является важность (вклад) признака, которая также является основой для обработки признаков в рамках нашего метода.

Поскольку в процессе выполнения задач машинного обучения на разработку признаков тратится много ресурсов, объяснение алгоритма

Boosting было рассмотрено еще на этапе его разработки. Важность признака (FI), как встроенный атрибут самого алгоритма бустинга, уже давно применяется для отбора признаков. Однако, как объяснение алгоритма бустинга, FI имеет много недостатков, которые нельзя игнорировать, в том числе:

- FI не может отразить, является ли влияние признаков на результаты прогноза положительным или отрицательным.
- FI не может отразить взаимосвязь между признаками и целевыми переменными. По сути, это означает, что интерпретация не является идеальной

В отличие от популярных в области объясняемого ИИ модельно-агностических приближений, ФИ является атрибутом самого алгоритма бустинга. Поэтому FI следует уделять внимание в рамках построения объяснения с помощью алгоритма Boosting, даже если он имеет крайне низкие показатели для объяснения, ориентированного на пользователя.

3.2 FI-SHAP: алгоритм объяснимого ИИ с гибридным механизмом для вычисления важности признаков

3.2.1 Описание гибридного механизма

Был разработан гибридный метод объяснения, сочетающий FI и XAI, с целью попытаться объединить инженерию признаков и объясняемый ИИ для улучшения прогнозной эффективности алгоритма Boosting Algorithm при прогнозировании временных рядов. Особенно для энергетических и других временных рядов, подверженных внешним помехам. Возьмем два метода объясняемого ИИ, упомянутых выше: SHAP и FI в качестве примеров, новый гибридный метод объяснения определяется как:

$$\varphi_{x_j} = \phi_{x_j} \times \frac{FI(x_j)}{\sum_{i=1}^{p} FI(x_i)} \tag{3.4}$$

φ_{x_j} представляет собой вклад j-го признака в результат прогноза в рамках объяснения SHAP.

FI-SHAP сочетает в себе традиционное проектирование признаков и развивающийся объясняемый ИИ. С одной стороны, он повышает специфичность XAI для алгоритма Boosting, а с другой - предоставляет пользователям более полные результаты объяснений.

3.2.2 Визуализация важности признаков

Для того чтобы эффективно измерить эффект улучшения этих методов объяснения при проектировании признаков, мы используем "объяснение со средним запаздыванием"в качестве базовой линии эксперимента. В "объяснении среднего запаздывания"объясненное значение всех признаков искусственно устанавливается равным 1 (или любому другому значению, если гарантируется, что объясненное значение всех признаков одинаково), то есть вес каждого признака одинаков. Таким образом, эффект заключается в том, что каждый признак должен построить одинаковое количество признаков запаздывания. Например, если N = 30, а исходное число признаков равно 5, то для "среднего объяснения запаздывания"необходимо построить 6 признаков запаздывания для каждого признака. Шэп может назначить 20 запаздывающих признаков для первого признака, 15 признаков для второго, и не назначать запаздывающие признаки для признаков с низкой объясняющей способностью. Наконец, сравним влияние этих методов на улучшение эффективности прогнозирования, которое также представляет собой эффект улучшения этих методов при конструировании признаков.

Данные временных рядов энергии, используемые в этой работе, были получены из Kaggle, которые были собраны с двух солнечных электростанций, расположенных в Индии, с периодом 34 дня (каждые 15 минут). Набор данных состоит из двух частей, первая часть - это набор данных о выработке электроэнергии, которая генерируется инвертором, включая постоянный ток, переменный ток, ежедневный выход и общий выход. Вторая часть - это данные, собранные датчиками, включая температуру и солнечную радиацию. Инверторы называются *sourcekey*, и каждая электростанция имеет 22 инвертора, в общей сложности 44 инвертора. Все эти инверторы генерируют данные о выработке электроэнергии в одно и то же время, в результате чего получается набор

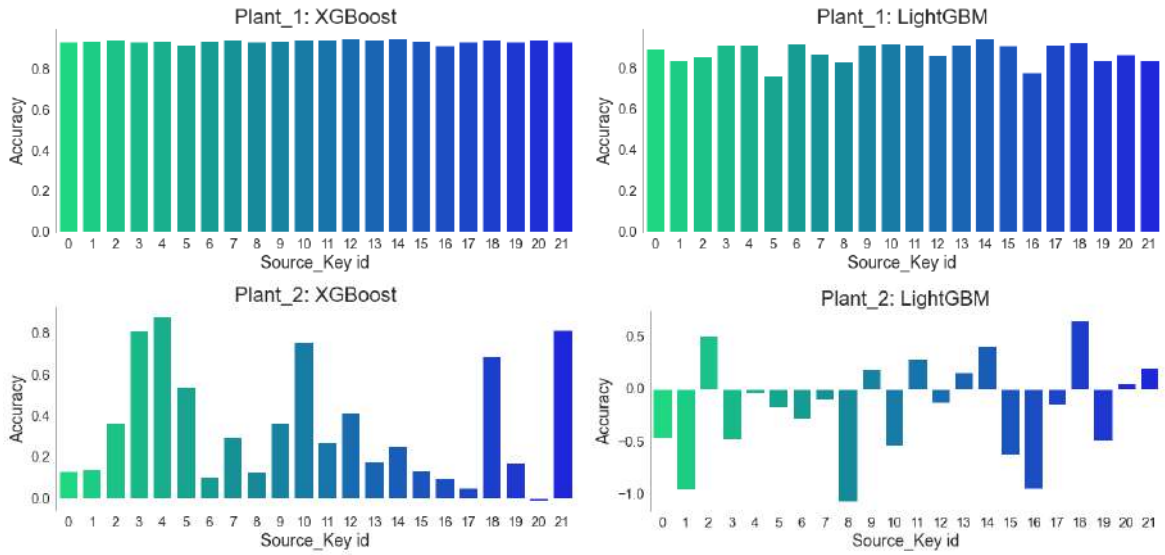


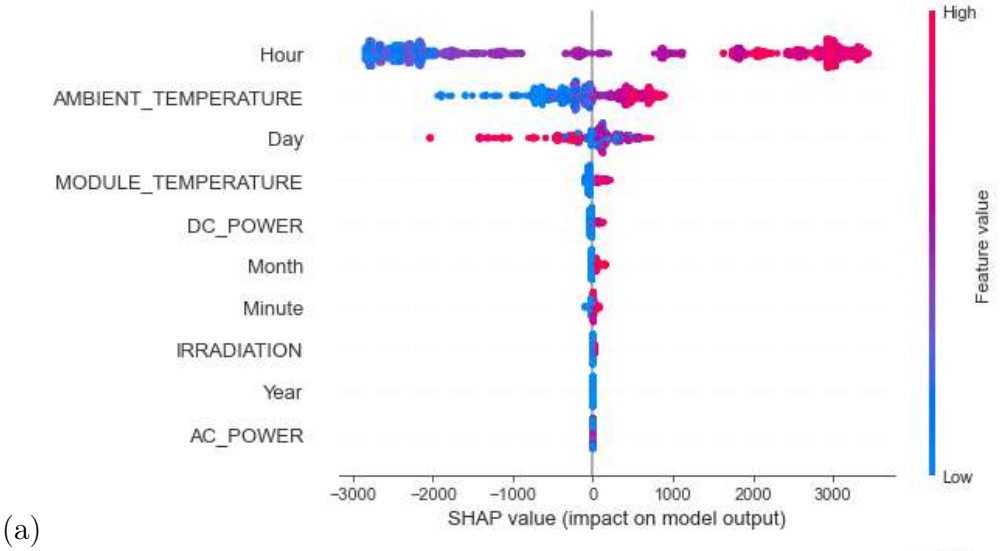
Рис. 3.2: Результаты прогнозирования в сыром виде (без использования функций). Результаты показывают, что набор данных электростанции 1 изначально обладает высоким качеством, в то время как набор данных электростанции 2 содержит большое количество аномалий, и даже LightGBM практически не справляется с ним.

данных, состоящий из 136 476 строк и 7 столбцов (за 34 дня).

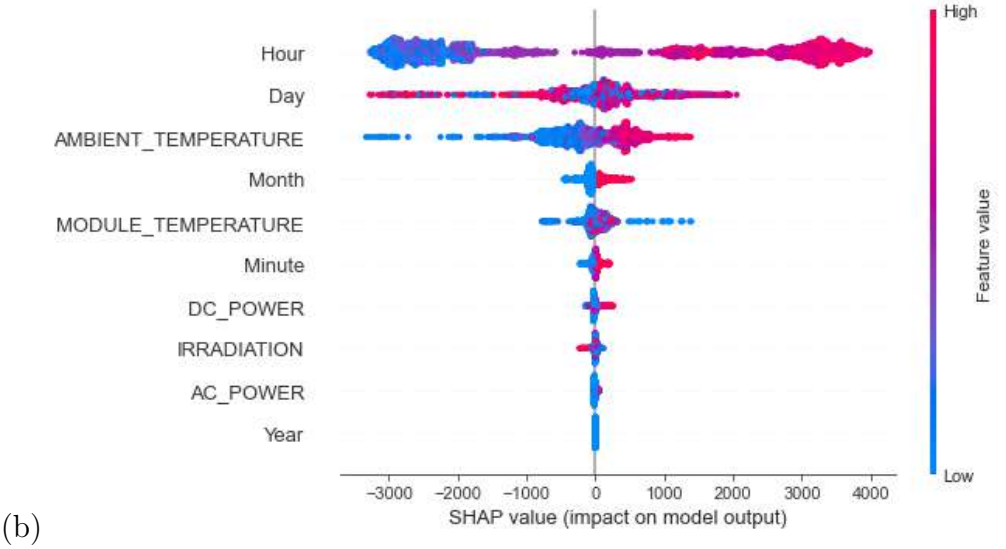
Мы использовали многопоточную схему обработки, то есть в соответствии с различными идентификаторами преобразователей (*sourcekey*) таблица данных была разделена на 44 набора данных, содержащих около 3000 строк и 7 столбцов (в течение 34 дней). Мы используем как необработанные модели бустинга, так и модели бустинга с функционалом на этих наборах данных, чтобы выделить наиболее эффективные методы объяснения.

Мы используем XGBoost с LightGBM для раздельного прогнозирования на 22 небольших наборах данных с солнечной электростанции. Результаты прогнозирования показаны на рисунке 3.2. Следует отметить, что LightGBM значительно превзошел XGBoost в соревнованиях по временным рядам в последние годы, поскольку масштаб используемых в соревнованиях наборов данных велик. Основой LightGBM по-прежнему является XGBoost, а добавление таких алгоритмов, как гистограмма, позволяет LightGBM обучаться быстрее, чем XGBoost, и обеспечивать точность. Однако на небольших наборах данных преимущества LightGBM соответственно исчезают. Это также является причиной низкой производительности LightGBM по сравнению с XGBoost в данной работе.

На основе этих наборов данных мы проверяем эффективность этих методов



(a)



(b)

Рис. 3.3: Объяснение SHAP (растение 1, идентификатор исходного ключа: 0): a: XGBoost; b: LightGBM

объяснения с помощью набора данных "Электростанция 1"и эффективность этих методов объяснения с помощью набора данных "Электростанция 2".

По сравнению со значением признака, результаты объяснения (рис. 4), выводимые SHAP, могут показать как положительное, так и отрицательное влияние признака и предоставить пользователям более полную информацию. Здесь мы приводим в качестве примера только результаты объяснения набора данных, полученных от инвертора.

Из результатов видно, что влияние "HOUR"является как положительным, так и отрицательным, и общий эффект относительно сбалансирован. Однако влияние "AMBIENT TEMPERATURE"на результаты прогнозирования смещено в отрицательную сторону, то есть в определенном диапазоне повышение температуры будет иметь определенное

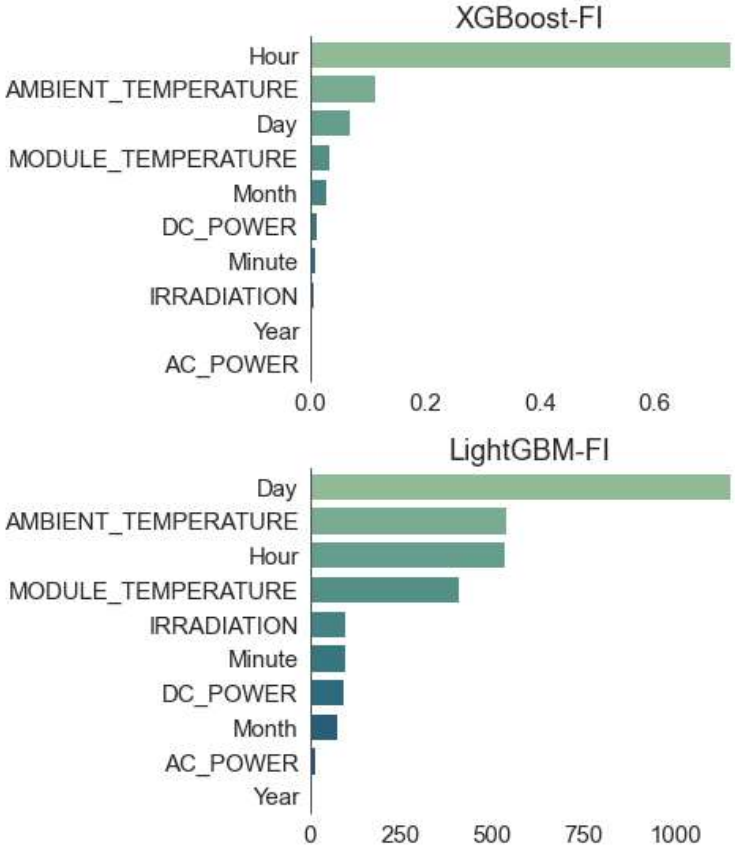


Рис. 3.4: Важность признака (растение 1, идентификатор исходного ключа: 0)

положительное влияние на выработку электроэнергии, и другие характеристики также объясняются в соответствии с той же логикой. Для "Важности признака"такой эффект интерпретации невозможен, и FI может только показать ранжирование важности признака, как показано на рисунке 3.4.

3.2.3 Повышение точности прогнозирования с помощью FI-SHAP

Как уже упоминалось выше, суть нашей автоматической инженерии признаков заключается в построении подходящего количества признаков запаздывания для достижения цели улучшения инженерии признаков. Изначально мы использовали признаки исходных данных для прогнозирования непосредственно с помощью модели boosting, то есть без применения инженерии признаков. После обучения модели необходимо определить важность признаков и рассчитать вес каждого признака, который определяется по следующей формуле:

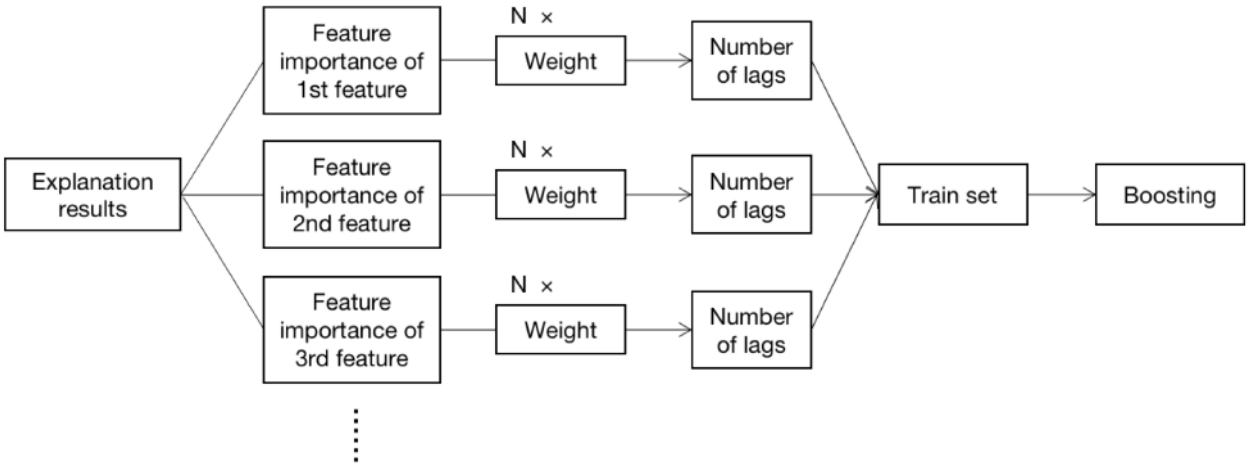


Рис. 3.5: Процесс проектирования признаков работает по результатам объяснения. N - общее количество признаков, которое может быть задано пользователем в соответствии с реальными потребностями.

$$Weight = \frac{F_j}{\sum_{j=1}^{n} F_j} \tag{3.5}$$

F_j представляет собой объяснительное значение каждого признака.

Для разных методов объяснения представление F_j также отличается. Для SHAP F - это $\phi(val)$ (Equ.3.3); для FI F - это Gain (Equ.3.2), а для нашего FI-SHAP F - это $\varphi(val)$ (Equ.3.4).

Рассчитайте вес каждого непрерывного признака по результатам объяснения SHAP и вывода Feature Importance (FI) и установите общее количество признаков для построения, чтобы в соответствии с весом можно было построить запаздывающие признаки (рис. 3.5). После получения весов нам также необходимо задать общее количество N признаков для построения и построить запаздывающие признаки для непрерывных признаков в исходном наборе данных в соответствии с результатом N*вес. Однако на самом деле мы не знаем точного значения N, поэтому автоматический инжиниринг признаков, который мы предоставляем, использует итерационный подход для построения запаздывающих признаков. Например, если пользователь установит N равным 50, наша система начнет с 1-50 и выберет лучший результат. Поэтому теоретически, если вычислительная мощность компьютера позволяет, значение N может быть установлено больше. Однако, исходя из реальной производительности, мы рекомендуем устанавливать значение N не более 200. Подробности см. в

этой библиотеке кода[1].

Таблица 3.1: Прогноз качества (R^2) электростанций 1

Source key	0	1	2	3	4	5	6	7	8	9	10
XGBoost	0.934	0.940	0.942	0.934	0.938	0.920	0.938	0.941	0.935	0.940	0.941
avg.	0.934	0.941	0.947	0.941	0.939	0.923	0.942	0.946	0.940	0.944	0.944
FI	0.938	0.953	**0.953**	0.942	**0.944**	0.927	0.938	0.951	0.951	0.942	0.944
SHAP	**0.940**	0.952	0.952	**0.944**	**0.944**	0.932	**0.943**	**0.952**	0.950	**0.943**	**0.945**
FI-SHAP	**0.940**	**0.954**	0.951	**0.944**	0.942	**0.933**	0.941	**0.952**	**0.953**	**0.943**	**0.945**
Source key	11	12	13	14	15	16	17	18	19	20	21
XGBoost	0.941	0.945	0.940	0.945	0.936	0.916	0.934	0.942	0.934	0.944	0.932
avg.	0.944	0.948	0.943	0.946	0.941	0.920	0.942	0.944	0.937	0.948	0.941
FI	**0.946**	0.953	0.943	0.948	0.942	0.926	0.938	0.946	0.947	0.953	0.949
SHAP	0.945	0.953	0.943	0.948	0.942	0.934	**0.939**	0.946	0.950	0.953	0.950
FI-SHAP	0.944	**0.954**	0.943	0.948	0.942	**0.935**	**0.939**	0.946	**0.951**	0.953	**0.951**
Source key	0	1	2	3	4	5	6	7	8	9	10
LightGBM	0.892	0.836	0.856	0.914	0.914	0.760	0.919	0.867	0.831	0.913	0.920
avg.	0.911	0.885	0.889	0.930	0.934	0.764	0.933	0.888	0.877	0.932	0.934
FI	0.928	0.889	0.893	0.940	0.936	0.774	**0.936**	0.895	0.879	0.935	0.940
SHAP	0.925	0.889	0.899	**0.944**	**0.939**	0.774	0.935	0.893	0.878	0.941	0.941
FI-SHAP	**0.931**	**0.892**	**0.900**	**0.944**	0.938	**0.859**	0.934	**0.904**	**0.888**	**0.942**	**0.945**
Source key	11	12	13	14	15	16	17	18	19	20	21
LightGBM	0.912	0.857	0.914	0.939	0.909	0.779	0.914	0.923	0.838	0.864	0.838
avg.	0.924	0.890	0.922	0.948	0.929	0.793	0.929	0.935	0.889	0.893	0.882
FI	0.935	0.891	0.936	0.947	0.926	0.827	0.937	0.940	0.881	0.887	0.880
SHAP	0.937	0.894	**0.942**	**0.953**	**0.932**	0.903	0.939	**0.944**	0.883	0.890	0.888
FI-SHAP	**0.939**	**0.904**	0.938	**0.953**	**0.932**	**0.915**	**0.942**	0.943	**0.890**	**0.894**	**0.893**

Результаты улучшения

Мы создаем различные виды признаков запаздывания на основе различных результатов объяснения, тем самым обогащая инженерию признаков для улучшения эффективности моделей прогнозирования. Эффект улучшения метода объяснения проверяется на наборе данных электростанции 1, а эффект восстановления метода объяснения проверяется на наборе данных электростанции 2.

[1]Библиотека FI-SHAP на github

Таблица 3.2: Прогноз качества (R^2) электростанций 2

Source key	0	1	2	3	4	5	6	7	8	9	10
XGBoost	0.132	0.141	0.365	0.810	0.878	0.535	0.102	0.297	0.124	0.367	0.755
avg.	0.494	0.159	0.372	0.869	0.904	**0.556**	0.153	0.327	0.174	0.449	0.821
FI	**0.569**	0.209	**0.465**	0.934	0.924	0.544	0.192	0.375	0.227	**0.554**	0.889
SHAP	0.568	**0.309**	0.384	0.946	**0.943**	0.538	**0.216**	0.394	0.362	0.398	0.890
FI-SHAP	0.558	0.252	**0.497**	**0.951**	0.937	0.536	0.212	**0.433**	**0.375**	0.401	**0.896**

Source key	11	12	13	14	15	16	17	18	19	20	21
XGBoost	0.271	0.415	0.176	0.251	0.134	0.093	0.051	0.689	0.173	-0.011	0.813
avg.	**0.318**	0.598	0.321	0.311	0.347	0.281	0.324	0.788	0.216	0.061	0.822
FI	0.301	0.608	0.456	**0.357**	0.449	0.343	0.360	0.938	0.247	0.337	0.826
SHAP	0.303	0.656	0.462	0.316	0.445	0.305	0.352	0.947	**0.258**	0.073	0.827
FI-SHAP	0.302	**0.682**	**0.541**	0.355	**0.465**	**0.471**	**0.368**	**0.948**	**0.249**	**0.296**	**0.828**

Source key	0	1	2	3	4	5	6	7	8	9	10
LightGBM	-0.472	-0.962	0.502	-0.486	-0.045	-0.176	-0.292	-0.107	-1.073	0.188	-0.546
avg.	-0.042	-0.314	0.576	0.654	0.405	0.080	0.004	0.318	-0.586	**0.398**	0.191
FI	**0.278**	-0.061	0.570	0.714	0.884	**0.205**	**0.169**	**0.396**	-0.179	0.367	0.500
SHAP	0.248	**0.206**	**0.597**	**0.770**	**0.903**	0.047	**0.022**	0.324	-0.008	0.309	0.615
FI-SHAP	0.253	0.074	0.585	0.688	0.889	0.135	0.117	0.356	0.022	0.324	**0.670**

Source key	11	12	13	14	15	16	17	18	19	20	21
LightGBM	0.288	-0.134	0.155	0.407	-0.628	-0.955	-0.148	0.653	-0.487	0.049	0.202
avg.	**0.327**	0.110	0.213	0.442	-0.051	-0.210	0.013	0.708	-0.288	0.193	0.258
FI	0.317	0.168	0.177	0.429	0.127	0.012	0.162	0.729	-0.224	0.177	0.325
SHAP	0.307	0.107	**0.503**	**0.477**	0.000	**0.026**	**0.190**	**0.901**	-0.290	**0.224**	0.313
FI-SHAP	0.290	**0.188**	0.403	0.475	**0.142**	-0.025	0.161	0.884	**-0.222**	0.188	**0.345**

Окончательные результаты улучшения показаны в таблице 3.1 и таблице 3.2, таблица 3.1 - это улучшение электростанции 1, а таблица 3.2 - это ремонт электростанции 2.

Согласно полученным результатам, в наборе данных более высокого качества Power Plant 1 все методы объяснения демонстрируют улучшенный эффект. В целом, эффект улучшения FI-SHAP является наилучшим, особенно в LightGBM. Вторым является SHAP, а FI оказывает общее влияние на улучшение высококачественных данных. На синтетическом наборе данных Power Plant 2 с низким качеством почти все методы объяснения имеют незначительный эффект восстановления. С одной стороны, это означает, что только построения авторегрессионных признаков

недостаточно для достижения хорошей производительности. С другой стороны, результаты также показывают, что для небольших наборов данных адаптивность LightGBM не так хороша, как у XGBoost.

Улучшение производительности за счет большего числа запаздывающих признаков все равно не является незначительным, поскольку при реальном построении признаков существует еще много способов участия, как было описано ранее. Однако в данной работе мы сосредоточимся только на построении запаздывающих признаков, чтобы более профессионально изучить задачи прогнозирования временных рядов. При восстановлении некачественных данных хорошо видно, что эффект восстановления FI-SHAP более очевиден для XGBoost, а эффект восстановления SHAP более очевиден для LightGBM.

3.3 Вывод по главе 3

В этой главе мы предлагаем новый гибридный метод объяснения, называемый FI-SHAP, для алгоритма бустинга, который объединяет как модельно-специфические, так модель-агностические. Используя оценки важности признаков, полученные на основе результатов FI-SHAP, мы разрабатываем наиболее важные признаки для расширения набора признаков. Наша цель - повысить эффективность задач прогнозирования временных рядов с помощью этого процесса обогащения признаков.

Глава 4

Применение объяснимого искусственного интеллекта

В этой главе мы применяем методы объяснимого искусственного интеллекта к реальным задачам, включая анализ влияющих факторов, повышение эффективности прогнозирования временных рядов и решение проблем онлайн-адаптации. **Результаты были опубликованы в статье [21, 22]** Результаты, изложенные в разделах 4.2 [30] и 4.3 [31], находятся на рассмотрении в журнале. Результаты, изложенные в разделе 4.1 [21, 22], были опубликованы в журналах. Новизна заключается в том, что в предыдущих работах приведены только пояснительные результаты, без рассмотрения сценариев применения, в нашем исследовании предлагаются три экономически выгодных сценария применения на практике.

4.1 Анализ факторов, влияющих на солнечную генерацию и качество воздуха

Новизна нашего метода заключается в том, что информация о важности признаков получается на основе метода XAI и используется для определения места установки солнечной батареи

4.1.1 Анализ факторов, влияющих на солнечную генерацию

Хотя в ходе сравнительных экспериментов были найдены более подходящие модели прогнозирования для набора данных по выработке солнечной энергии, эти модели отличаются от традиционных статистических моделей тем, что логика их вывода не является для нас прозрачной. Скорее, мы

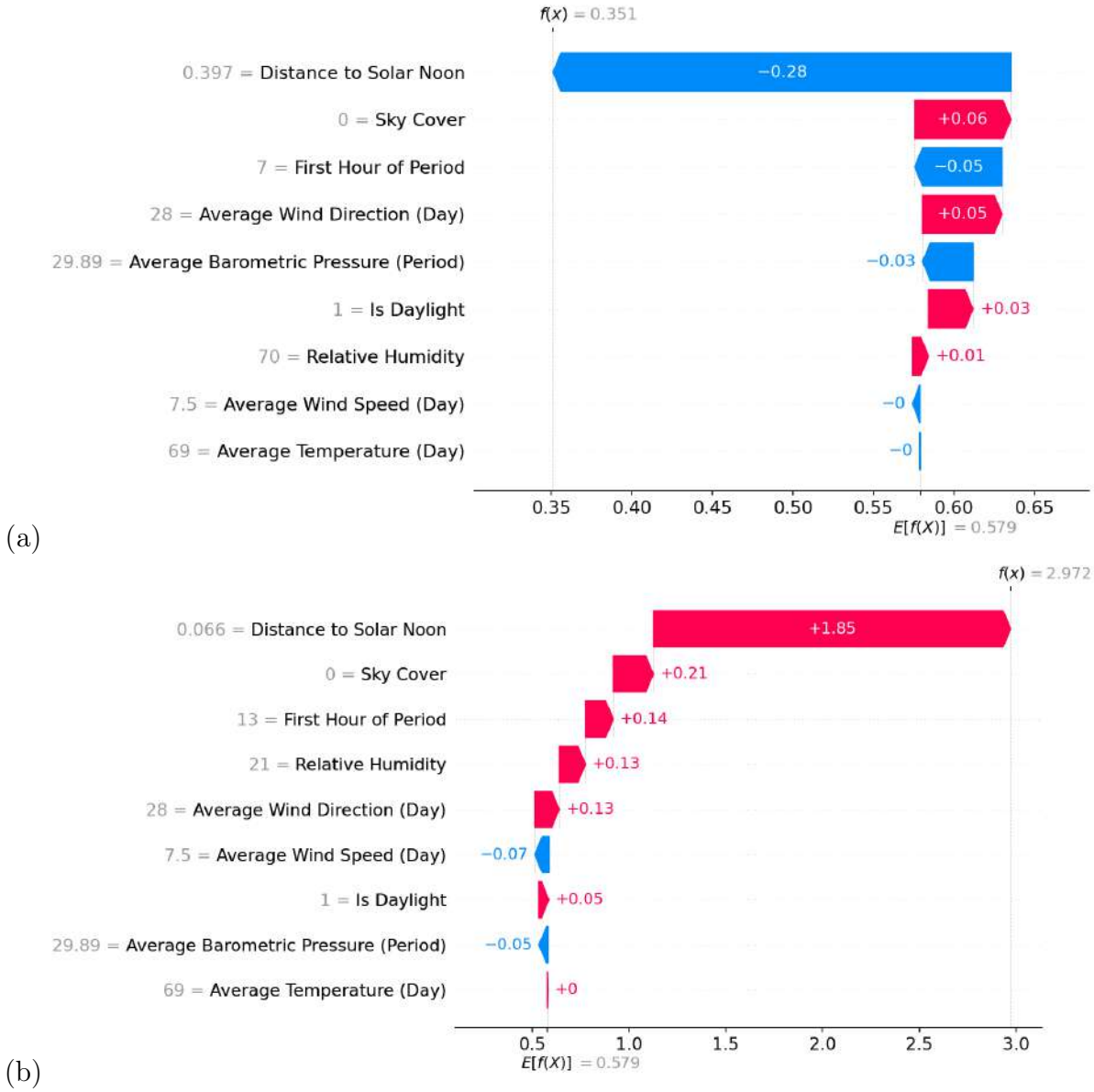


Рис. 4.1: Пример результатов локального объяснения.

знаем только, что в них существуют нелинейные зависимости. Отсутствие прозрачности не позволяет нам полностью полагаться на результаты их прогнозирования и ограничивает нашу способность анализировать факторы, влияющие на эти прогнозы. Анализ влияющих факторов - эффективный инструмент для прогнозирования и управления мощностью солнечных энергосистем. Выявление ключевых факторов, влияющих на производительность системы, таких как погодные условия, географическое положение и затенение, позволяет более точно прогнозировать выработку энергии и оптимизировать производительность системы с помощью усовершенствованных стратегий управления. Кроме того, исторические данные и тенденции, полученные в результате анализа влияющих факторов,

могут способствовать принятию обоснованных решений об инвестициях в солнечные технологии и инфраструктуру.

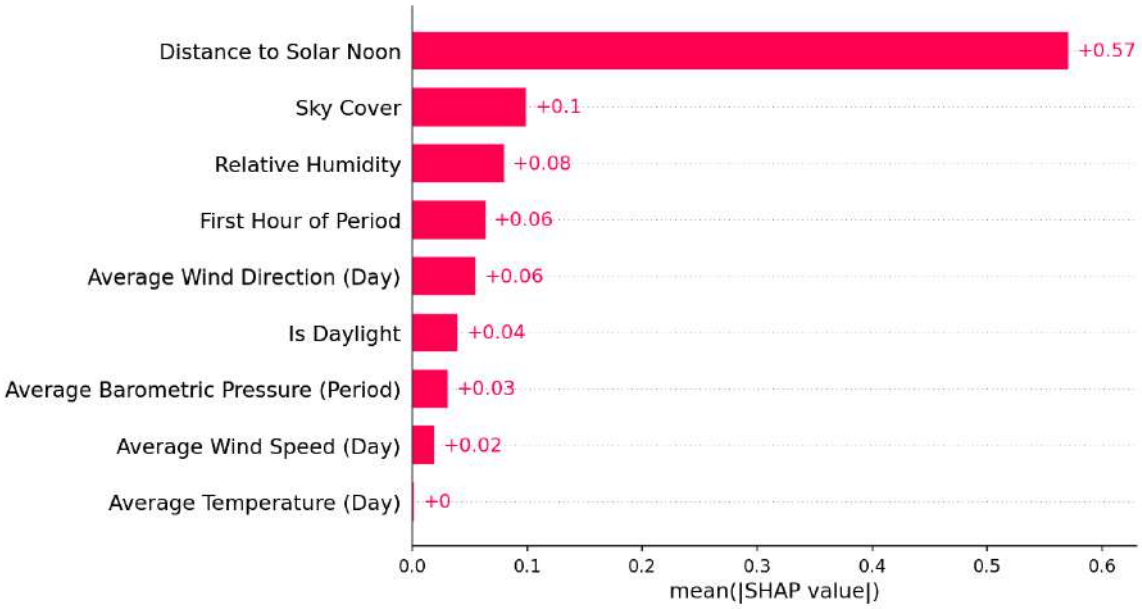


Рис. 4.2: Пример результатов глобального объяснения.

Анализ факторов влияния

В исследовании используется алгоритм SHAP как общий метод XAI для проведения анализа факторов влияния. Алгоритм SHAP, который вычисляет значения Шэпли, подробно описан в разделе 2.2. Мы используем LightGBM в качестве примера для демонстрации реализации SHAP. Этот выбор обусловлен поддержкой библиотеки исходного кода SHAP для ансамблевого обучения и превосходными возможностями визуализации, которые превосходят возможности глубокого обучения. Важно признать, что такое различие в технических возможностях ограничивается только кодовыми областями, и SHAP сохраняет свою роль в качестве схемы объяснения для всех моделей "черного ящика"в рамках теоретической структуры.

Важно подчеркнуть, что значение SHAP переменной отражает степень ее вклада в результаты прогнозирования, и чем больше значение, тем важнее соответствующая переменная. На примере рисунка 4.1 мы подробно рассмотрим анализ влияющих факторов с использованием SHAP в качестве основы методики.

На рисунке 4.1 показаны объяснения SHAP для двух точек выборки с временными метками "1 сентября 2008, 7:00"и "1 сентября 2008, 13:00".

Базовая линия, обозначаемая как $E[f(x)]$, представляет собой неучтенные переменные и заменяется средним значением всех прогнозируемых величин. В данном исследовании $E[f(x)] = 0,579$. На рисунке 11(a), начиная с нижней части, базовая линия модели равна 0,579. Включение средней температуры и средней скорости ветра не приводит к изменению прогнозного значения $f(x)$. Однако после включения относительной влажности $f(x)$ начинает меняться, и каждая последующая переменная вносит свой вклад в прогнозирование. Процесс расчета выглядит следующим образом: $f(X) = 0,579 - 0 - 0 + 0,01 + 0,03 - 0,03 + 0,05 - 0,05 + 0,06 - 0,28 = 0,351$.

Таким образом, SHAP присваивает каждой переменной прогнозное значение $f(x) = 0,351$. Аналогичный процесс происходит и на рисунке 11(b). Сравнивая рисунки 11(a) и 11(b), можно заметить, что при изменении времени с 7:00 до 13:00 вклады большинства переменных в прогнозируемое значение становятся положительными. В частности, при уменьшении расстояния до солнечного полудня ее вклад значительно возрастает до +1,85, оказывая решающее влияние. Однако это лишь объяснения для двух точек выборки. На практике невозможно проанализировать каждую точку данных в отдельности. Поэтому SHAP также предоставляет глобальные объяснения переменных, беря абсолютные значения результатов локальных объяснений и усредняя их. Это среднее значение служит глобальным объяснением, представленным на рисунке 4.2.

Глобальное объяснение может всесторонне оценить важность этих переменных. На рисунке 4.2 показано ранжирование важности переменных для прогнозирования исходов, причем переменные расположены сверху вниз в порядке убывания их важности. Для всего набора данных "расстояние до солнечного полудня"является наиболее важным, его значимость значительно превосходит значимость остальных переменных. Затем на первое место выходят покрытие неба и относительная влажность, в то время как другие переменные, такие как направление ветра, скорость ветра и средняя температура, не выделяются в этой комплексной оценке. До сих пор вышеупомянутый анализ основывался исключительно на статическом объяснении значений SHAP, без учета значений переменных. Далее мы изучим значения SHAP при изменении переменных величин,

чтобы провести динамический анализ. Глобальное динамическое объяснение представлено на рисунке 4.3.

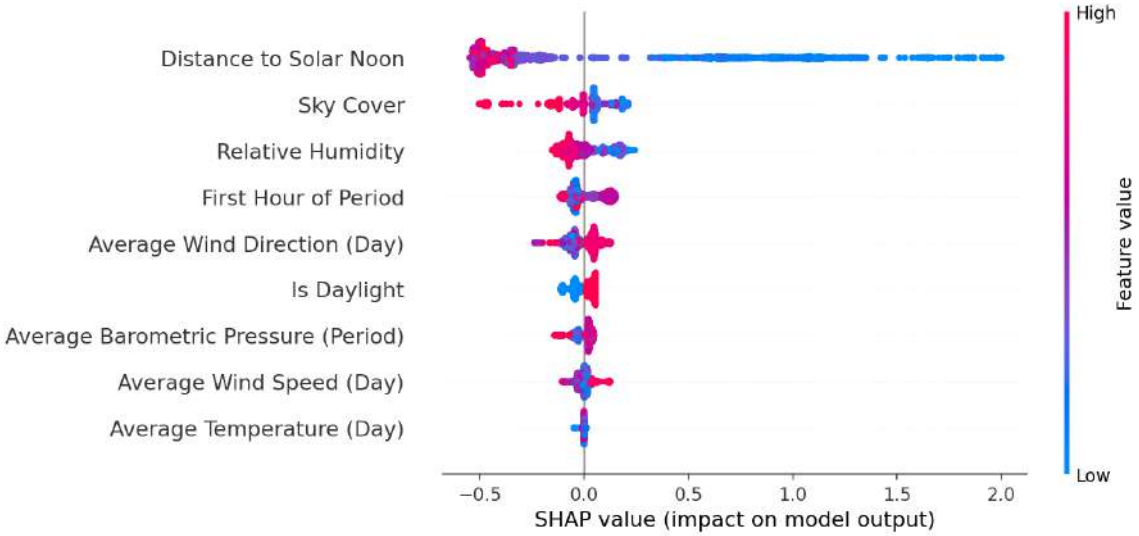


Рис. 4.3: Глобальное динамическое объяснение. Левая вертикальная ось показывает название переменной, а правая цветовая полоса от синего до красного представляет значение переменной от малого до большого. Горизонтальная ось представляет собой значение SHAP, указывающее на важность или вклад переменной в результаты прогнозирования.

Анализ рисунка 4.3 помогает нам выявить наличие сильной монотонной зависимости между значениями SHAP и значениями переменных. Ключевым аспектом анализа рисунка 4.3 является четкое разграничение между синей и красной областями. Например, если рассматривать наиболее важную переменную - расстояние до солнечного полудня, то когда эта переменная имеет более низкие значения, она представлена синими точками, расположенными в правой части. Это говорит о том, что более низкие значения этой переменной положительно влияют на прогнозирование выработки солнечной энергии. И наоборот, красные точки, представляющие более высокие значения переменной, сосредоточены в левой части, что говорит о том, что большие значения этой переменной оказывают негативное влияние на прогнозирование. Чем отчетливее разделение между красными и синими областями, тем сильнее монотонная связь между значениями переменных и соответствующими им значениями SHAP. Кроме того, SHAP предоставляет интерактивный пояснительный график (рис. 4.4, 4.5), который не только детально отображает монотонные зависимости, но и иллюстрирует эффекты взаимодействия между

переменными.

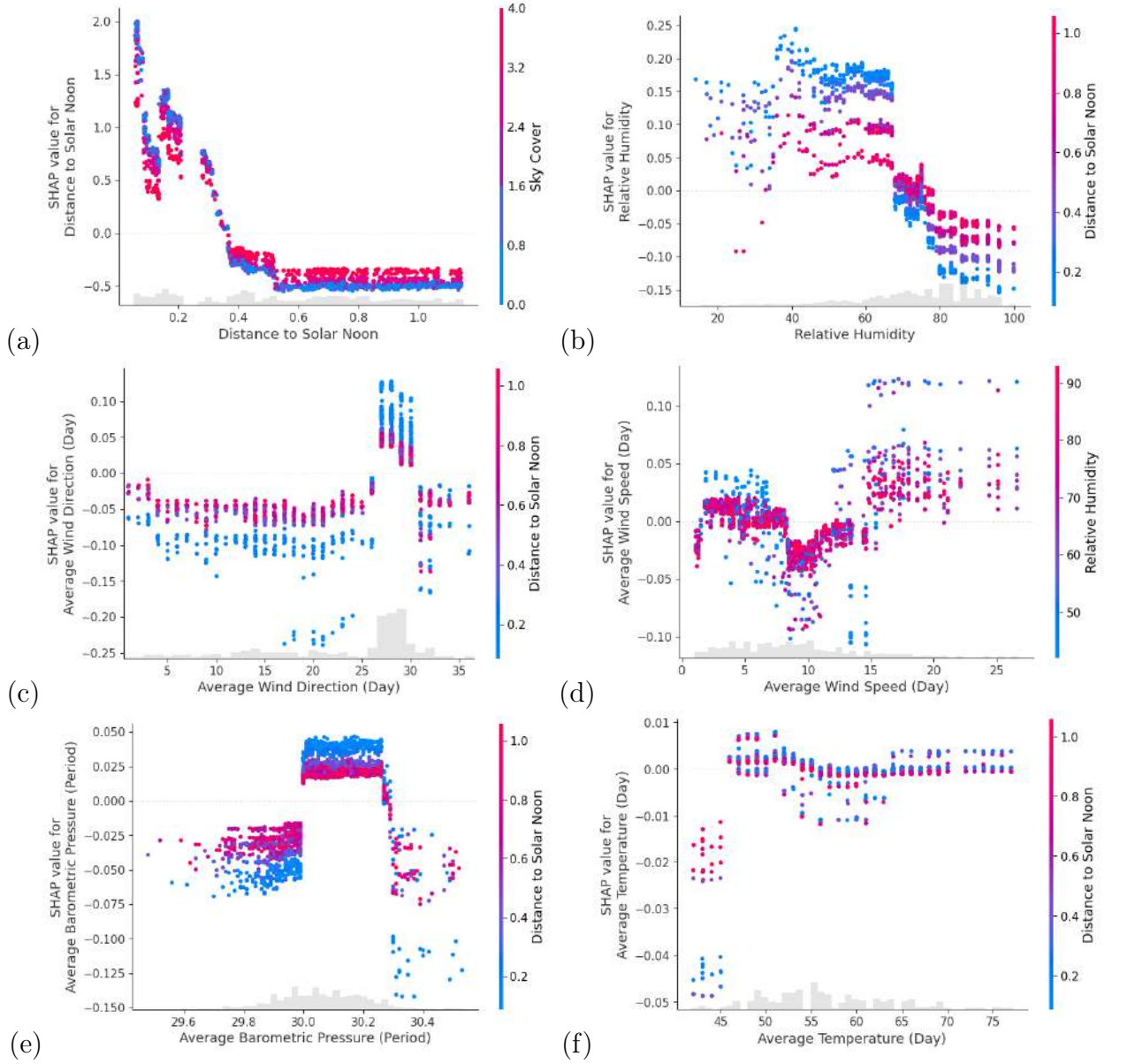


Рис. 4.4: Влияние взаимодействия непрерывных переменных на результаты прогнозирования. (a): Расстояние до солнечного полудня; (b): Относительная влажность; (c): Среднее направление ветра; (d): Средняя скорость ветра; (e): Среднее барометрическое давление; (f): Средняя температура. Серая штриховка показывает распределение соответствующих переменных

Начнем с анализа непрерывных переменных. Как показано на рисунке 4.4, тренд, демонстрируемый точками данных, представляет собой связь между значениями переменных и SHAP, а цвет указывает на значение переменной, которое имеет наиболее значимый эффект взаимодействия с данной переменной. Влиятельные факторы на основе интерактивных графиков анализируются следующим образом:

Дальность до солнечного полудня (рисунок(4.4a)). Тенденция, отображаемая точками данных, указывает на сильную монотонную зависимость, при которой значение SHAP значительно уменьшается при увеличении значения переменной. Это означает значительное уменьшение его вклада в прогнозируемое значение. Кроме того, когда значение переменной увеличивается примерно до 0,3, значение SHAP остается на относительно низком уровне (около 0,5); однако, когда значение достигает примерно 0,4, его вклад становится отрицательным. С другой стороны, правая часть графика демонстрирует переменную, которая проявляет заметный эффект взаимодействия, в данном случае это покрытие неба. Можно заметить, что даже когда расстояние находится в пределах 0,3 км, красные точки данных, представляющие высокий уровень покрытия неба, все равно снижают значение SHAP до более низкого уровня. Однако, как только расстояние превышает 0,3 км, из-за значительного снижения выработки электроэнергии, такие эффекты взаимодействия теряют свою аналитическую ценность. В целом, можно сделать вывод, что когда расстояние не превышает 0,3 км, а покрытие неба остается ниже 2, модель может генерировать более высокие прогнозные значения. Наконец, результаты SHAP-анализа показывают, что расстояние до солнечного полудня в пределах 0,3 км в сочетании с уровнем покрытия неба, не превышающим 60%, являются наиболее важными факторами окружающей среды для качественной выработки солнечной энергии.

Относительная влажность (рисунок(4.4b)). По мере увеличения относительной влажности соответствующее ей значение SHAP постепенно уменьшается, что свидетельствует о снижении вклада в прогнозные значения. Как видно из рисунка, когда расстояние до солнечного полудня находится в пределах 0,3 км, относительная влажность ниже 60% оказывает относительно положительное влияние на выработку солнечной энергии.

Среднее направление ветра (рисунок(4.4c)). В отношении набора данных можно заметить, что когда расстояние до солнечного полудня находится в пределах 0,3 км, среднее направление ветра находится в диапазоне от 25° до 30°, что благоприятствует выработке солнечной энергии. Однако следует подчеркнуть, что этот вывод применим только к данному набору данных.

Средняя скорость ветра (рисунок(4.4d)). Что касается данного набора данных, то при относительной влажности ниже 60% и средней скорости ветра от 15м/с до 25м/с наблюдается значительное увеличение соответствующих значений SHAP. Следовательно, это оказывает положительное влияние на выработку солнечной энергии.

Среднее барометрическое давление (рисунок(4.4e)). В этом наборе данных, когда расстояние до солнечного полудня находится в пределах 0,3 км, среднее барометрическое давление в диапазоне от 30 до 30,2 дюймов рт. ст. может генерировать положительные значения SHAP, что указывает на положительное влияние на значение прогноза. Однако величина этого влияния значительно меньше, чем у вышеупомянутых четырех факторов окружающей среды.

Средняя температура (рисунок(4.4f)). В целом, по сравнению с другими экологическими факторами, средняя температура оказывает относительно небольшое влияние на выработку солнечной энергии. Судя по результатам, большинство точек данных колеблются около SHAP=0, что указывает на нулевой вклад в выработку солнечной энергии. Единственное примечательное наблюдение - когда средняя температура опускается ниже 45°F, она оказывает негативное влияние на выработку солнечной энергии. Однако выше этого температурного порога его влияние остается ограниченным.

Рисунок 4.5 иллюстрирует анализ дискретных переменных, с особым вниманием к представлению "Is Daylight"для аналитической полноты. На практике, когда это значение равно 0, что означает ночное время, выработка солнечной энергии снижается до нуля. И наоборот, когда значение равно 1, что означает дневное время, выработка солнечной энергии начинает расти.

Анализ покрытия неба показывает, что когда расстояние до солнечного полудня находится в пределах 0,3 км, покрытие неба не превышает 2, что соответствует 60%. Такая ситуация положительно сказывается на мощности солнечной генерации, повышая ее эффективность. Однако, наоборот, это приводит к значительному негативному влиянию на производство солнечной энергии.

Пример приложения

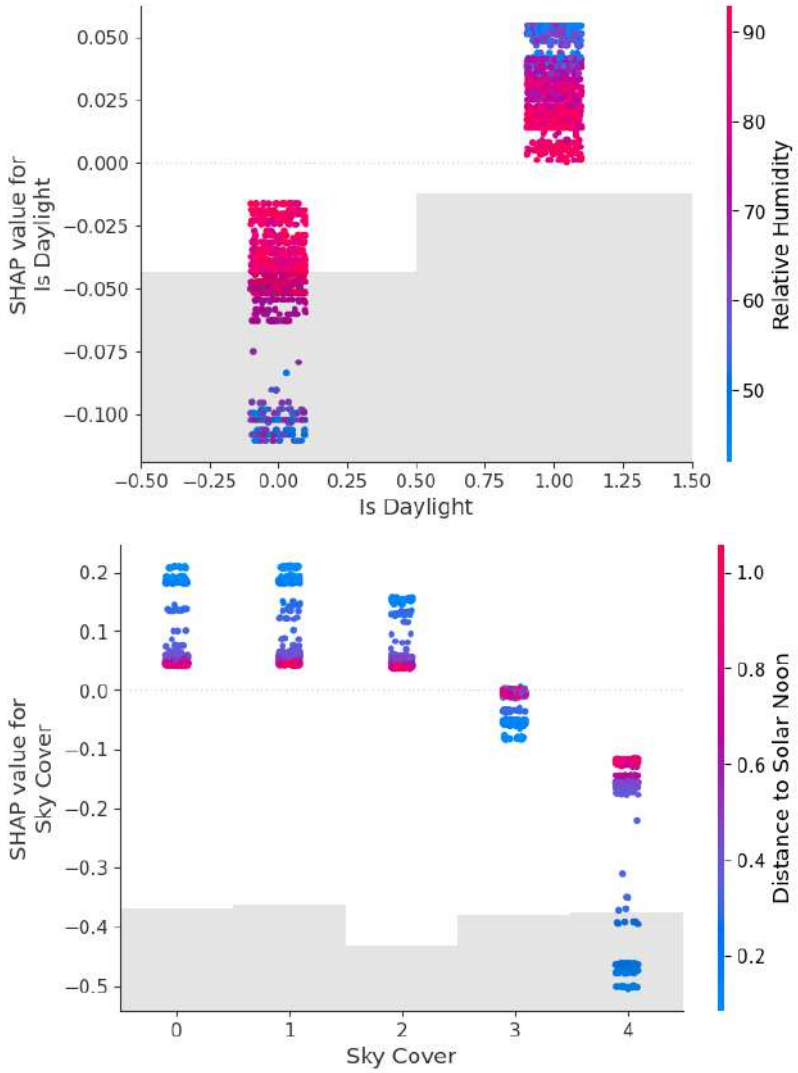


Рис. 4.5: Влияние взаимодействия дискретных переменных на результаты прогнозирования

Принимая во внимание эти факторы, мы предлагаем два потенциальных места для электростанции, основываясь на топографической карте Беркли (рисунок 4.6). Мы определили местоположение 1 со средней высотой 340 м и местоположение 2 со средней высотой 370 м. Оба места рекомендованы в качестве подходящих вариантов для выбора площадки под солнечную электростанцию.

На основе результатов важности признаков, полученных в результате моделирования с помощью LightGBM и методов объясняющего искусственного интеллекта (XAI), представленных SHAP, можно сделать следующие выводы. Очевидно, что расстояние от солнца в полдень является наиболее значимым экологическим фактором, формирующим основу для положительного влияния других экологических факторов. В частности, для

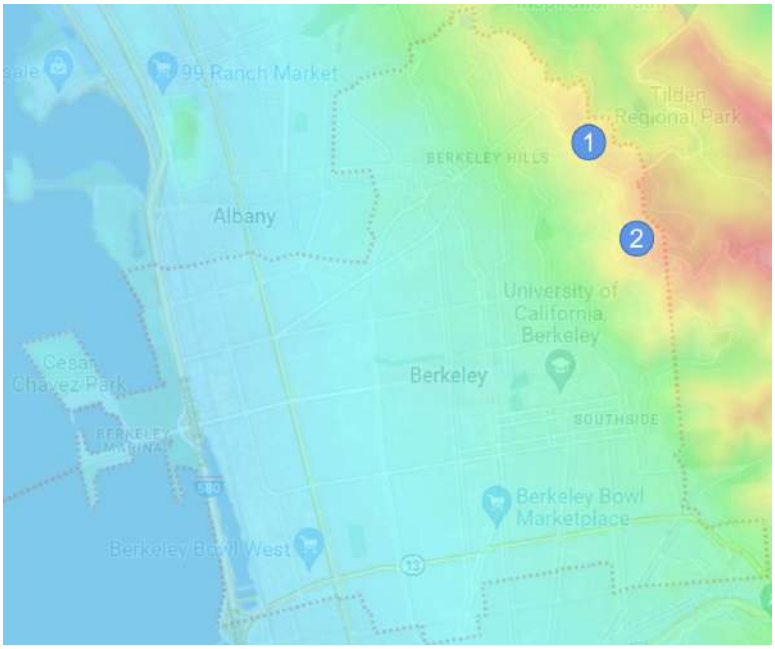


Рис. 4.6: Предлагаемые места для размещения электростанций

максимальной выработки солнечной энергии крайне важно поддерживать расстояние до солнца в пределах 0,3 км в полдень. Кроме того, увеличению выработки солнечной энергии способствуют следующие условия: относительная влажность воздуха ниже 60%, среднее направление ветра от 25° до 30°, средняя скорость ветра от 15 м/с до 25 м/с, среднее барометрическое давление от 30 дюймов рт. ст. до 30,2 дюймов рт. ст., средняя температура ниже 45°F и облачность не более 60%.

4.1.2 Анализ факторов, влияющих на качество воздуха

В данном исследовании для анализа факторов, влияющих на PM2.5, мы выбрали модель ансамблевого обучения, которая наилучшим образом поддерживалась SHAP. Мы сосредоточили наше сравнение на типе ансамблевой модели. В частности, мы использовали Catboost для изучения факторов, влияющих на PM2.5, с горизонтом прогноза 30 дней, а LightGBM - для анализа на 90 и 180 дней. Результаты объяснения SHAP позволили получить ценные сведения о том, как переменные влияют на результаты прогнозирования, отражая как влияние отдельных переменных, так и взаимодействие между ними. Чтобы наглядно представить вклад каждой переменной, мы использовали графики средних значений, как показано на рисунке 4.7.

В этом анализе объяснение SHAP трех горизонтов прогнозирования сохраняет последовательность. Выяснилось, что PM10 вносит наибольший вклад в результаты по PM2.5 во всех горизонтах, что свидетельствует о его значительном влиянии. Вслед за PM10, CO также демонстрирует более высокий уровень значимости в прогнозируемых результатах для всех горизонтов. Что касается таких факторов выбросов, как O_3 и SO_2, то их влияние считается приемлемым. Однако среди всех факторов выбросов NO_2 имеет наименее значимое влияние. Что касается метеорологических условий, то наиболее выраженное влияние на PM2.5 оказывает точка росы, а следующим по значимости фактором является температура. Фактически, эти два фактора перевешивают все факторы выбросов, кроме PM10 и CO. Давление и скорость ветра оказывают незначительное влияние на концентрацию PM2.5. Поскольку анализ сосредоточен исключительно на концентрации PM2.5 в конкретном районе, влияние направления ветра ограничено. Кроме того, в этом наборе данных переменная rain, представляющая количество осадков, в большинстве случаев записывается как 0. Следовательно, ее влияние на PM2.5 минимально, что подчеркивается климатическими особенностями Пекина.

Анализ, представленный выше, сосредоточен на отдельных переменных, но диаграмма рассеяния дает дополнительное представление о взаимосвязи между переменными и значениями SHAP. На диаграмме рассеяния (рисунок 4.7 вверху) высокие значения переменных показаны красными точками, а низкие значения - синими точками. Например, что касается PM10, то высокое значение PM10 соответствует высокому значению SHAP на горизонтальной оси, что указывает на положительный эффект. Другими словами, увеличение PM10 (красные точки) способствует увеличению PM2.5, в то время как уменьшение PM10 (синие точки) препятствует увеличению PM2.5. Применяя тот же аналитический процесс, мы наблюдали аналогичные закономерности для CO, точки росы и озона. Однако влияние температуры демонстрирует иной механизм корреляции. Повышение температуры препятствует увеличению PM2.5, в то время как понижение способствует его увеличению. Вероятно, на это явление влияет наличие централизованного зимнего отопления в районе Пекина. Поскольку основным способом отопления в зимний период является использование

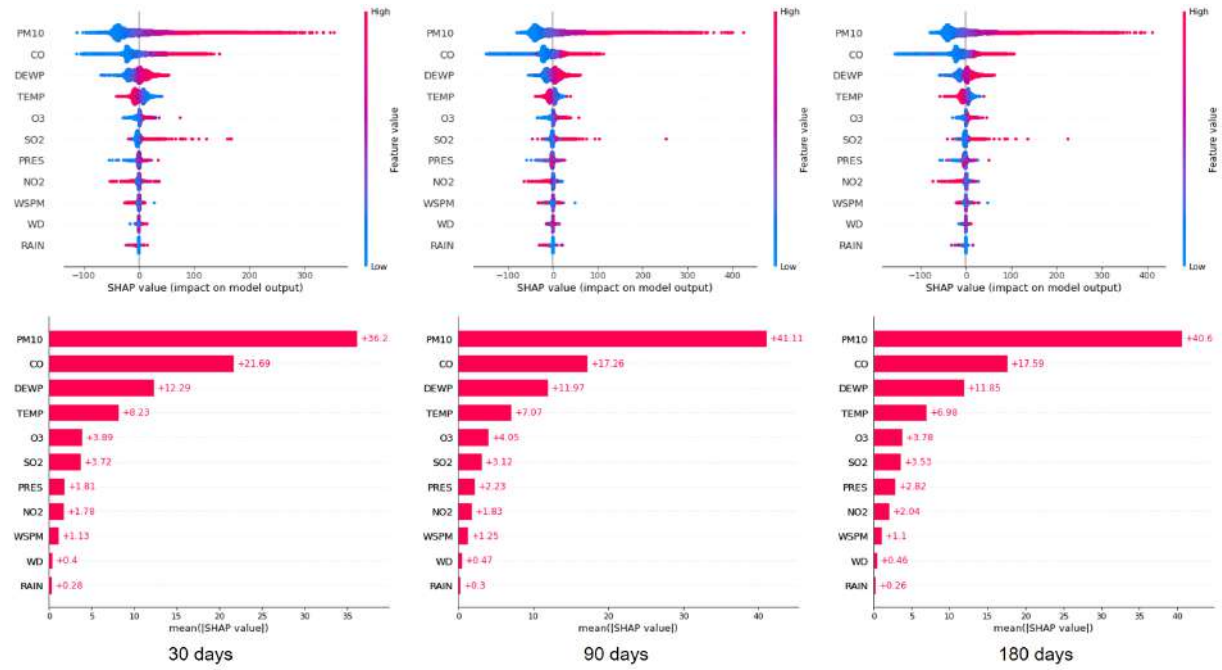


Рис. 4.7: Результаты одномерного объяснения SHAP для Catboost (слева: 30 дней) и для LightGBM (в середине: 90 дней, справа: 180 дней). Разброс результатов объяснения показан выше, а график средних значений - ниже. Для каждой переменной в каждом наблюдении SHAP рассчитывает ее вклад в прогнозируемый результат, все вклады обрабатываются в абсолютных значениях и усредняются для каждой переменной, что в конечном итоге приводит к глобальному значению вклада переменной. Значение SHAP для каждой точки данных отображается в виде точки на диаграмме рассеяния, а итоговый глобальный вклад переменной - на средней диаграмме. На диаграмме рассеяния горизонтальная ось представляет собой значение SHAP, где большее значение представляет собой большее значение целевой переменной, а вертикальная ось представляет собой рейтинг вклада всех переменных в прогноз, где переменная с более высоким рейтингом представляет собой большее влияние на конечный результат прогноза. Синий и красный переходы представляют значения переменных от малого к большому.

тепловой энергии, значительное количество ископаемого топлива сжигается при более низких температурах, что приводит к повышению концентрации PM2.5.

Анализ взаимодействия факторов

Учитывая значимость объяснительных графиков взаимодействия, мы решили сосредоточить наш анализ на четырех переменных, которые оказывают существенное влияние на результаты. Эти переменные включают в себя два фактора выбросов, а именно PM10 и CO, а также два метеорологических условия, а именно точку росы и температуру. График объяснения взаимодействия SHAP дает ценные сведения, в частности, о

взаимодействии между переменными. Он раскрывает механизмы влияния переменных на результаты прогноза, даже в тех случаях, когда между ними существует ковариация. Рисунок 4.8 наглядно иллюстрирует, как два важнейших фактора, PM10 и CO, влияют на концентрацию PM2.5.

Взаимосвязь между концентрацией PM10 и ее вкладом в результат прогнозирования, о чем свидетельствует значение SHAP, показывает, что увеличение PM10 приводит к соответствующему увеличению PM2.5. Особенно значителен эффект взаимодействия CO с PM10. Чтобы проанализировать этот эффект взаимодействия, мы изобразили концентрацию CO с помощью цветов: красный обозначает высокую концентрацию CO, а синий - низкую. Изучение данных показывает, что при высоких концентрациях CO наблюдается линейное увеличение PM2.5 с увеличением уровня PM10 (красные точки). И наоборот, снижение концентрации CO значительно уменьшает вклад PM10 в концентрацию PM2.5 (синие точки). Аналогичным образом, анализ графиков взаимодействия CO показывает, что увеличение CO также способствует увеличению PM2.5, хотя и не в такой степени, как PM10. Более того, при высоких концентрациях PM10 и PM10, и CO способствуют повышению уровня PM2.5, но подавление этой тенденции низкими уровнями PM10 не является значительным (наблюдается лишь несколько синих точек).

График взаимодействия SHAP дает представление, которое невозможно получить только с помощью графика разброса и среднего значения. Хотя PM10 занимает первое место по значимости, это наблюдается именно при высоких концентрациях CO. При низких концентрациях CO увеличение концентрации PM10, как правило, стабилизирует его влияние на PM2.5. Таким образом, можно сделать вывод, что CO играет решающую роль в возникновении этих эффектов.

Рисунок 4.9 иллюстрирует влияние одной переменной на прогнозируемые результаты. Переменная, проявляющая наиболее выраженное взаимодействие с этой переменной, показана на правой вертикальной оси. Графики взаимодействия дают более четкое представление о линейных ассоциациях между изменениями этих переменных и изменениями уровней PM2.5. В частности, точка росы положительно коррелирует с PM2.5, в то время как температура в целом показывает отрицательную корреляцию. С

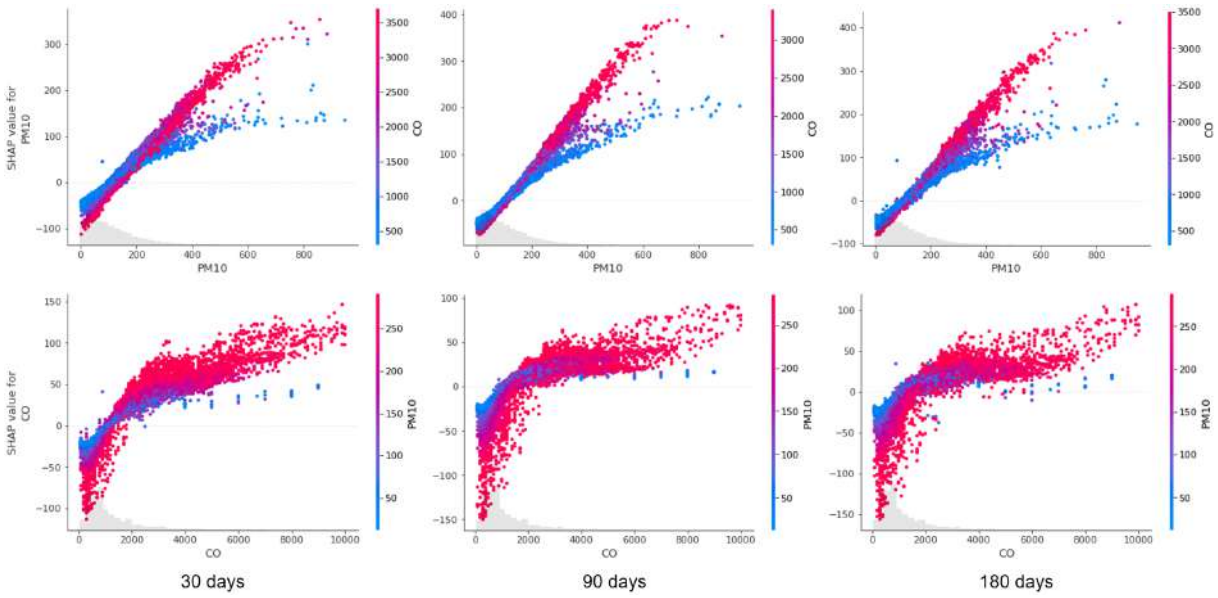


Рис. 4.8: График SHAP-эффектов взаимодействия PM10 и CO для Catboost (слева: 30 дней), для LightGBM (в середине: 90 дней, справа: 180 дней). Горизонтальная ось представляет значение переменной, заштрихованная область - распределение данных для этой переменной. Левая вертикальная ось представляет значение SHAP, а правая вертикальная ось показывает переменную, с которой переменная имеет наиболее очевидное взаимодействие, представляя от наименьшего до наибольшего ее значения переходом от синего к красному.

другой стороны, для прогнозирования PM2.5 с горизонтом в 30 дней "месяц"имеет наиболее очевидное взаимодействие с "росой"и "прессой соответственно. Однако, поскольку "месяц"не имеет линейной связи с результатами прогноза, он также не показывает четкой красно-синей границы.

Для остальных горизонтов существует значительное взаимодействие между температурой и точкой росы. Очевидно, что более высокие точки росы соответствуют более высоким температурам, что указывает на положительную корреляцию между точкой росы и температурой. Эту связь можно подтвердить, рассмотрев переменную "температура где высокие точки росы (представленные красными точками) наблюдаются в диапазоне более высоких температур. Кроме того, увеличение точки росы совпадает с увеличением значения SHAP, что указывает на рост концентрации PM2.5. И наоборот, высокие температуры, сопровождающиеся повышением точки росы, приводят к снижению концентрации PM2.5. Однако численно, когда точка росы превышает 25 градусов Цельсия, это может даже способствовать увеличению концентрации PM2.5 на 100 $\mu g/m^3$, в то время как

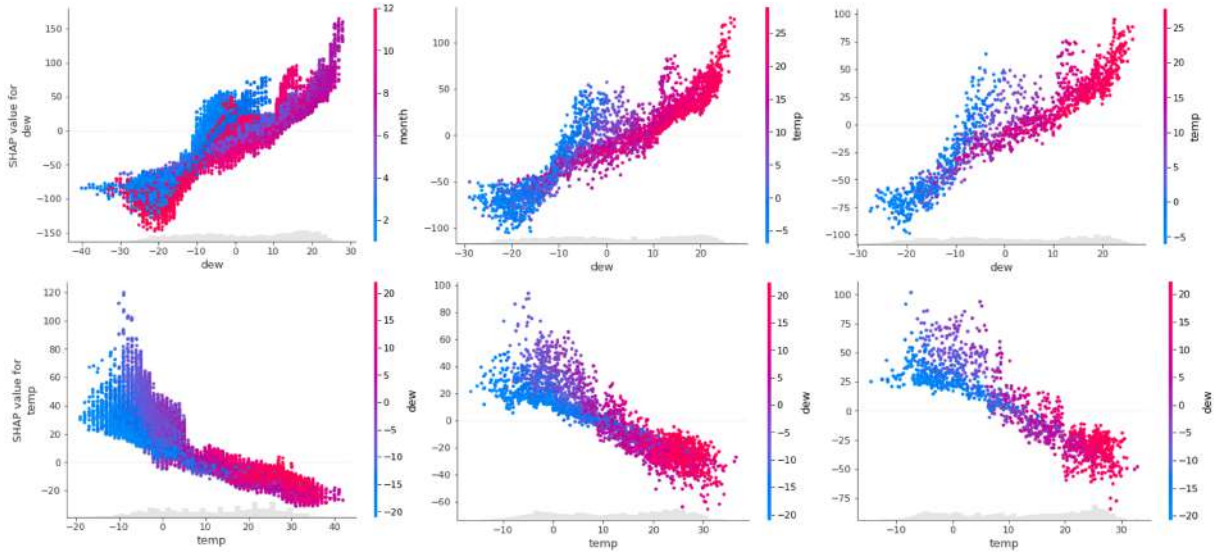


Рис. 4.9: График эффектов взаимодействия SHAP точки росы и температуры для Catboost (слева: 30 дней), для LightGBM (в середине: 90 дней, справа: 180 дней).

соответствующее повышение температуры приводит к снижению концентрации только на 60 $\mu g/m^3$. Таким образом, влияние высокой точки росы на концентрацию PM2.5 можно считать достаточно надежным. Следовательно, очень важно подчеркнуть положительное влияние высокой точки росы на концентрацию PM2.5.

Выводы из анализа

Результаты анализа показывают, что "PM10"оказывает самое сильное влияние на прогнозирование PM2.5, за ним следуют "CO "точка росы"и "температура". Эти влиятельные факторы демонстрируют различную степень линейной связи с PM2.5. В частности, температура отрицательно коррелирует с PM2.5, в то время как PM10, CO и точка росы в целом положительно коррелируют с PM2.5. Кроме того, влияние PM10 на PM2.5 уменьшается при более низких концентрациях CO, в то время как влияние CO на PM2.5 практически не зависит от PM10. Теоретически, эти коррелирующие переменные могут взаимно влиять на концентрацию PM2.5. Однако численно превышение точки росы на 25 градусов Цельсия приводит к значительному увеличению концентрации PM2.5 до 100 $\mu g/m^3$, по сравнению с максимальным снижением на 60 $\mu g/m^3$ при высоких температурах и 20-40 $\mu g/m^3$.

4.2 Разработка алгоритмов автоматической генерации признаков для задач прогнозирования временных рядов

Это исследование представляет собой автоматизированную структуру для инженерии признаков в прогнозировании временных рядов, которая использует объяснимый искусственный интеллект (XAI). Включив модуль XAI, данная структура повышает объяснимость и направляет усилия по инженерии признаков на улучшение точности прогнозирования. Мы применяем LightGBM (Light Gradient-Boosting Machine) и гибридную модель, которая включает алгоритм экспоненциального сглаживания (ES) для решения проблем экстраполяции тренда, присущих моделям на основе деревьев. Эффективность различных методов XAI в рамках данной структуры оценивается путем анализа производительности. Экспериментальные результаты показывают, что гибридный метод XAI, сочетающий модель-агностические и модель-специфические подходы, обеспечивает наибольшее улучшение производительности. Однако, консистентность этих улучшений варьируется между гибридной моделью и оригинальной LightGBM, указывая на ограниченности структуры. С инженерной точки зрения, эта работа демонстрирует, что XAI является крайне экономичным решением для повышения точности прогнозирования данных временных рядов. С точки зрения искусственного интеллекта, наши выводы подчеркивают потенциал гибридного подхода XAI как эффективной технической стратегии, что подтверждается превосходными результатами, достигнутыми с нашей разработанной гибридной методикой.

4.2.1 Описание структуры автоматической генерации признаков

Для одномерных временных рядов мы сосредоточены на разработке двух столбцов признаков. Первый столбец представляет собой информацию о временной метке, тогда как второй столбец соответствует целевой переменной. Предложенная в этом исследовании автоматизированная система генерации признаков высоко применима к моделям регрессии, особенно при применении к задачам прогнозирования временных рядов, в

частности к задачам прогнозирования многовременного шага для одномерных временных рядов. Эта система, как показано на Рисунке 4.10, включает три основных компонента: генерация отставленных признаков, генерация временного признака и выбор оптимального лага.

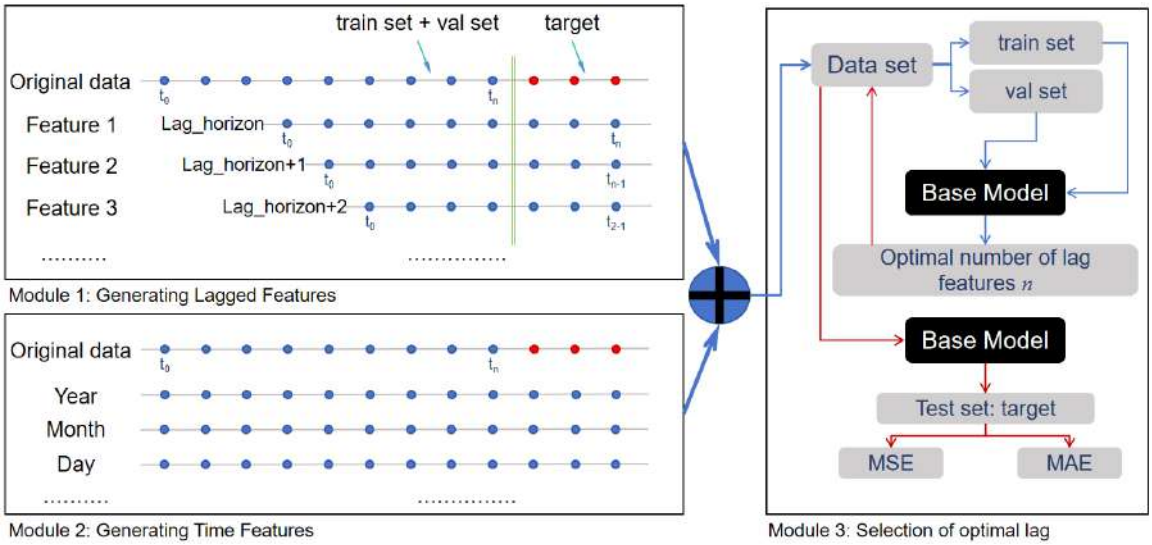


Рис. 4.10: Автоматизированная система разработки признаков, адаптированная для прогнозирования временных рядов

Модуль 1: Генерация отстающих признаков

В рамках этого модуля используется только один столбец, а именно целевая переменная, для генерации отстающих признаков в соответствии с желаемым горизонтом прогнозирования. Как показано на Рисунке 4.10, когда горизонт прогнозирования установлен на 3, первый созданный признак - это отстающий признак третьего порядка целевой переменной. Затем создается отстающий признак четвертого порядка, затем отстающий признак пятого порядка и так далее. Поэтому становится необходимым определить максимальное количество отстающих признаков, обозначаемых как N. В данной работе мы установили N равным 100.

Модуль 2: Генерация временных признаков

Этот модуль включает в себя генерацию соответствующих временных признаков на основе информации о метке времени – широко применяемая техника инженерии признаков в области прогнозирования временных рядов. Например, при заданной метке времени, такой как '2020-01-01 15:30', соответствующий временной признак будет состоять из различных компонентов: год=2020; месяц=01; день=01; час=15; мин=30.

Модуль 3: Выбор оптимального лага

Данные, сгенерированные в Модуле 1, включают всего 100 лаговых признаков. В этом модуле 100 лаговых признаков, извлеченных из набора данных, поочередно добавляются в модель во время этапа обучения. Посредством систематической оценки на валидационном наборе данных определяется оптимальное количество лаговых признаков, обозначаемое как n. Затем обученная модель, включающая эту оптимальную конфигурацию, используется для задачи прогноза на тестовом наборе данных.

4.2.2 Автоматическая генерация признаков с помощью объяснимого ИИ

Предлагаемая структура, как показано на Рисунке 4.10, позволяет разумно применять регрессионную модель к задаче прогнозирования временных рядов. Последующие экспериментальные результаты демонстрируют эффективность структуры с точки зрения производительности прогнозирования. Однако эта эффективность не освобождает структуру от присущих ей ограничений, а именно генерации избыточных признаков из-за большого количества запаздывающих переменных. Удаление этих избыточных признаков является ключевым для улучшения производительности прогнозирования. Поддерживаемая XAI автоматическая структура создания признаков, изображенная на Рисунке 4.11, решает эту проблему и включает три отдельных модуля:

Модуль 1: Объяснение

В этом модуле используется метод XAI для объяснения черного ящика модели и набора данных, генерируя вклад или важность каждой характеристики. Характеристики с меньшим вкладом считаются избыточными.

Модуль 2: Выбор данных

После получения ранжирования важности характеристик необходимо определить количество характеристик, которые нужно удалить. В этом модуле мы начинаем с итеративного удаления характеристик с наименьшей важностью, непрерывно формируя новый набор данных. В конечном итоге мы выбираем оптимальное количество удаляемых характеристик, что приводит к лучшему набору данных.

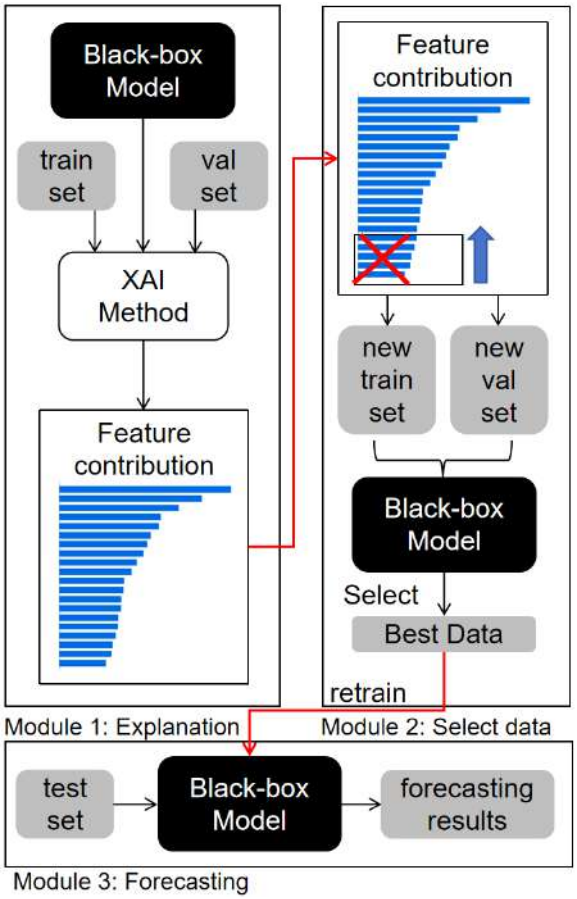


Рис. 4.11: Структура для автоматизированной разработки признаков под руководством XAI

Модуль 3: Прогнозирование

В этом модуле список характеристик из лучшего набора данных сохраняется и развертывается на тестовом наборе для получения результатов прогнозирования.

Предлагаемая структура, управляемая XAI, направлена на использование мощных техник XAI для идентификации и устранения избыточных характеристик, тем самым улучшая производительность регрессионной модели в задачах прогнозирования временных рядов.

4.2.3 Повышение точности прогнозирования

Описание данных

Это исследование использует четыре разнообразных набора данных, полученных из различных источников, для изучения отдельных явлений. Наборы данных охватывают генерацию солнечной энергии, объем трафика, температурные записи и данные по продажам яиц. Ниже приведено

подробное описание каждого набора данных, подчеркивающее их происхождение, временной охват, частоту и размер выборки.

Dataset 1 (D1): Генерация Солнечной Энергии. Набор данных по генерации солнечной энергии поступает из солнечной энергетической системы в Беркли, Калифорния, США. Он фиксирует выходную мощность этого возобновляемого источника энергии за период в один год, с 1 сентября 2008 года по 31 августа 2009 года. Данные собираются с интервалом в три часа, что предоставляет всего 2,920 образцов.

Dataset 2 (D2): Объем Дорожного Движения. Набор данных по объему дорожного движения происходит от западного потока транспорта на межгосударственных шоссе в Миннесоте, США. Он фиксирует поток транспортных средств на этих основных транспортных артериях за длительный период, с 2 октября 2012 года по 30 сентября 2018 года. Данные записываются с почасовым интервалом, что приводит к значительному объему выборки в 48,204 наблюдений.

Dataset 3 (D3): Температура. Набор данных по температуре был получен из Центра Предупреждения Погодных Условий в Дели, Индия. Он состоит из ежедневных записей температуры за период примерно в четыре года, с 1 января 2013 года по 24 апреля 2017 года. Набор данных включает в себя всего 1,575 образцов, предоставляя информацию о колебаниях температуры в регионе.

Dataset 4 (D4): Продажа Яиц. Набор данных по продажам яиц поступил из местного магазина в Шри-Ланке и охватывает период в 30 лет. Для обеспечения актуальности и точности данных первые 5,000 образцов, датирующиеся до 1993 года, были исключены из анализа. Измененный набор данных охватывает период с 9 сентября 2006 года по 31 декабря 2021 года с ежедневной частотой, что составляет всего 5,592 образцов.

Сравнение эффективности прогнозирования с поддержкой XAI

Ранжирование важности признаков, полученное различными методами объяснения, является непоследовательным. Эта непоследовательность предполагает, что результаты прогнозирования, созданные в рамках XAI, будут варьироваться в зависимости от используемого метода объяснения, что приведет к различным эффектам повышения производительности. В

результате списки признаков, сгенерированные каждым методом объяснения, первоначально проверяются с использованием валидационного набора данных. После определения оптимальных списков признаков оценивается эффект улучшения производительности инженерии признаков на прогнозирование временных рядов, руководствуясь различными методами объяснения, на тестовом наборе данных. Результаты улучшения для LightGBM на рисунке 4.12.

Models		LightGBM		FI		FI-SHAP		TreeSHAP		KernelSHAP		Partition		Additive		Permutation	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
D1	28	20.72	0.265	20.88	0.252	20.72	0.265	21.19	0.258	20.72	0.265	20.88	0.252	23.87	0.271	20.72	0.265
	56	11.02	0.199	14.24	0.228	12.49	0.209	12.49	0.209	12.49	0.209	12.50	0.210	12.62	0.217	12.49	0.209
	96	15.96	0.236	17.89	0.241	16.22	0.237	15.72	0.231	15.72	0.231	15.66	0.232	18.97	0.248	15.72	0.231
	avg	15.90	0.233	17.67	0.240	16.47	0.237	16.46	0.234	16.31	0.235	16.34	0.236	18.48	0.245	16.31	0.235
D2	48	10.04	0.201	11.34	0.217	9.831	0.198	10.24	0.201	12.15	0.245	9.831	0.198	10.20	0.199	10.03	0.198
	168	8.595	0.212	8.393	0.210	8.088	0.203	8.100	0.206	8.457	0.206	8.984	0.221	8.088	0.203	8.577	0.206
	240	8.908	0.220	8.595	0.223	8.595	0.223	8.595	0.223	8.848	0.221	8.699	0.217	8.916	0.226	8.595	0.223
	avg	9.181	0.211	9.442	0.216	8.838	0.208	8.978	0.210	9.818	0.224	9.171	0.212	9.068	0.209	9.067	0.209
D3	7	1.213	0.882	1.213	0.882	1.213	0.882	1.213	0.882	1.213	0.882	1.213	0.882	1.213	0.882	1.213	0.882
	15	5.583	1.966	5.583	1.966	5.583	1.966	5.583	1.966	5.583	1.966	5.583	1.966	5.583	1.966	5.583	1.966
	30	7.586	2.411	7.586	2.411	7.224	2.282	8.210	2.432	7.224	2.282	8.080	2.413	7.483	2.284	7.224	2.282
	avg	4.794	1.753	4.794	1.753	4.673	1.710	5.002	1.716	4.674	1.760	4.959	1.753	4.760	1.710	4.674	1.710
D4	24	13.29	0.299	10.17	0.259	8.426	0.232	9.119	0.262	8.426	0.232	23.22	0.420	23.33	0.423	8.426	0.232
	48	10.50	0.249	9.402	0.246	9.402	0.246	9.402	0.246	9.402	0.246	9.402	0.246	9.402	0.246	11.12	0.275
	120	9.233	0.258	9.233	0.258	7.431	0.222	9.233	0.258	9.666	0.250	7.431	0.222	9.233	0.258	9.666	0.250
	avg	11.00	0.268	9.601	0.254	8.419	0.233	9.251	0.255	9.164	0.242	13.35	0.296	13.98	0.309	9.737	0.252

Рис. 4.12: Результаты сравнения прогнозирования временных рядов. Заштрихованная рамка показывает оптимальную среднюю производительность.

Вкратце, использование объясняемого искусственного интеллекта (XAI) значительно повышает точность прогнозов во всех наборах данных, кроме Набора данных 1. Однако модель-специфический метод объяснения, FI, который интегрирован в LightGBM, не улучшает точность прогнозов удовлетворительно. Этот недостаток возникает из-за того, что метод FI не учитывает корреляции между признаками, что приводит к неточностям в его объяснительных результатах.

Хотя улучшение FI не является очевидным, улучшение FI-SHAP, которое интегрирует его с TreeSHAP в LightGBM, особенно значимо. Это связано с тем, что метод TreeSHAP не полагается на предположение независимости. Кроме того, комбинация модель-специфических и модель-агностических методов предлагает более эффективную пояснительную информацию, которая может предоставить ценное руководство для улучшения производительности. Это руководство, основанное на методах XAI, может

эффективно устранить избыточные признаки, что принципиально отличается от смягчения проблемы экстраполяции тренда. Смягчение проблемы экстраполяции тренда демонстрируется повышением прогнозов LightGBM, то есть предсказанные значения LightGBM увеличиваются до некоторой степени. Напротив, удаление избыточных признаков не связано со смещением результатов прогнозирования. Вместо этого оно направлено на ослабление проблемы переобучения, приближая результаты прогнозирования к фактическим значениям.

4.3 Решение проблемы дрейфа концепций при онлайн-адаптации

В задачах прогнозирования временных рядов решение проблемы дрейфа понятий практически сводится к разработке онлайн-модели, которая может обновляться в режиме реального времени. Реализация таких моделей в основном зависит от обновления обучающего набора и параметров модели. Мы предлагаем новую простую систему адаптации на основе онлайн-модели для дальнейшего решения проблемы дрейфа концепций в прогнозировании временных рядов - Tracker. В нашей системе адаптации признаки с низким вкладом будут немедленно удаляться на основе рейтинга вклада признаков, полученного из XAI, чтобы достичь динамического улучшения. По сравнению с предыдущими онлайн-моделями, в которых основное внимание уделяется обновлению параметров и обучающего набора, в нашей системе адаптации размерность обучающего набора также обновляется в режиме реального времени. Экспериментальные результаты доказывают, что наша система имеет очевидные улучшения. Новизна нашего метода заключается в том, что на основе методов XAI можно динамически обновлять характеристики в онлайн-модели прогнозирования. Ранее онлайн-адаптация могла только динамически обновлять параметры и учитывать новые данные. В результате наш метод позволяет лучше прогнозировать производительность. Код для эксперимента можно посмотреть по этой ссылке[1].

[1]Библиотека github для фреймворка XAI-адаптации

4.3.1 Дрейф концепции

Дрейф концепции [131] - это острая проблема, возникающая в задачах прогнозирования временных рядов, которая заключается в том, что точность обученной модели постепенно снижается со временем, пока не приводит к ее отказу. Причина [132] заключается в том, что распределение потока данных меняется со временем, что приводит к сбою статической модели, которая была обучена. Поэтому популярным решением является использование обновленного набора данных и параметров для обучения модели в реальном времени, чтобы модель прогнозирования могла улавливать информацию об изменении распределения данных во времени (рисунок 4.13). Все эти решения рассматриваются с точки зрения количества экземпляров набора данных и самой модели, и мы больше склоняемся к тому, чтобы попробовать рассмотреть их с точки зрения размерности. Исходя из этих предыдущих решений, добавление "изменения размерности"позволяет удалить неважные признаки в исходных данных.

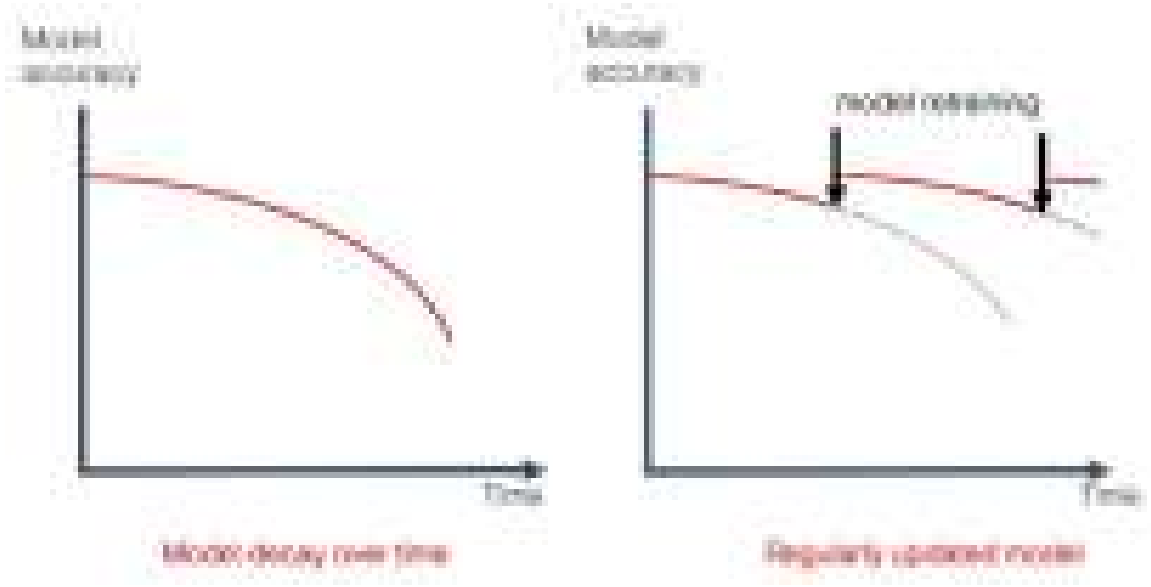


Рис. 4.13: Визуализация дрейфа концепций и их решений

Объясняемый ИИ [32–34], также называемый XAI, в широком смысле является разновидностью искусственного интеллекта, разработанного для того, чтобы позволить человеку понять принцип работы модели "черного ящика"через результаты ее объяснения. Учитывая различные личности пользователей, опыт взаимодействия человека и компьютера, реализуемый XAI, также отличается: Для пользователей в различных профессиональных

областях его метод склонен к пост-специальному объяснению, то есть измерению изменения результата предсказания через "функцию возмущения"для расчета вклада функции; Для разработчиков, ориентированных на модели, метод склонен к внутреннему, то есть предоставлению интерпретации для разработчиков в процессе разработки модели, что приводит к оптимизации модели. В последние годы последний метод постепенно превратился в интерпретируемый ИИ, в котором больше внимания уделяется разработке моделей, которые по своей сути являются интерпретируемыми.

Учитывая, что целью данной работы является построение универсального и обобщенного объясняемого фреймворка адаптации, нами принято объяснение post-hoc, что может обеспечить применение фреймворка к другим моделям "черного ящика". Построенная нами простая схема адаптации к дрейфу понятий в прогнозировании временных рядов, Трекер, не только охватывает предыдущие решения по обработке дрейфа понятий, но и вводит новые перспективы, изменения размерности, для решения проблемы дрейфа понятий, тем самым увеличивая верхний предел производительности модели прогнозирования.

4.3.2 Система онлайн-адаптации

Разработанный нами фрейм показан на рисунке 4.14. После получения объяснения каждой обновленной модели прогнозирования, признаки с низким вкладом из обновленного набора данных будут удалены. В соответствии с результатом объяснения модель прогнозирования переобучается на основе набора данных, размерность которого изменилась. В итоге мы обнаружили, что верхняя граница эффективности прогнозирования обновленной модели повышается при соответствующей корректировке параметров.

Обновление обучающего набора

В нашей задаче прогнозирования многомерных временных рядов, матрица признаков:

$$X = \{x_0, x_1, \cdots, x_m\}$$

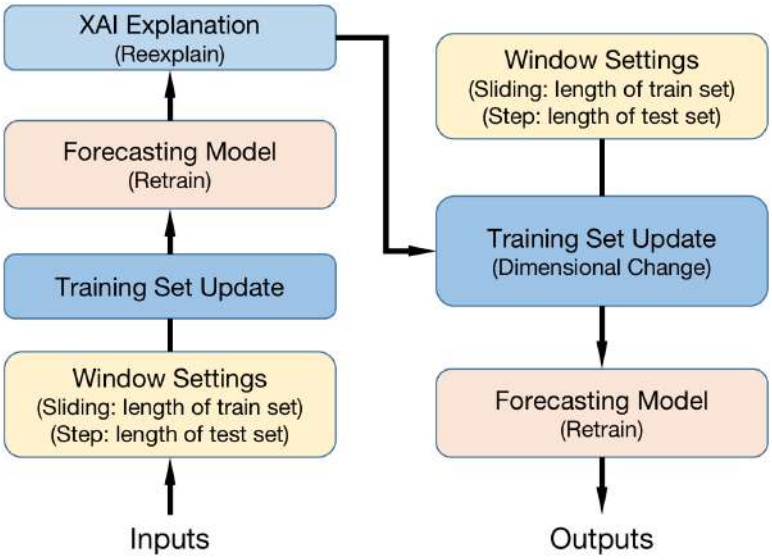


Рис. 4.14: The Tracker: адаптационная система на основе XAI

где признак x содержит экземпляры временного ряда, сопровождаемые временем $T = \{t_0, t_1, \cdots, t_n\}$, а именно $x = \{a_{t_0}, a_{t_1}, \cdots, a_{t_n}\}$. $Y = \{y_0, y_1, \cdots, y_n\}$ - целевая переменная в задаче, как показано на рисунке 4.15. Чтобы сделать наше изложение более кратким, мы оговариваем $t = \{a_{0,t}, a_{1,t}, \cdots, y\}$. В реальных задачах распределение данных X и Y будет меняться с течением времени. В результате информация об изменении распределения данных не может быть эффективно учтена в статической модели прогнозирования, что приводит к постоянному снижению точности прогноза.

Поскольку дрейф концепций вызван изменениями в распределении данных, непрерывное обновление обучающего набора имеет значительную эффективность в решении проблемы дрейфа концепций. Обучающий набор можно обновлять, постоянно добавляя в него новые данные (увеличивая t_n) и соответствующим образом регулируя вес, то есть уменьшая вес прошлого времени (w_0, w_1, w_1 и т. д.) и увеличивая вес недавнего времени (w_n, w_{n-1} и т. д.). t - это информация о времени, которая представляет собой строку обучающего множества в определенный момент времени (уравнение 4.1).

$$X_{expansion} = \{w_0 t_0, w_1 t_1, w_2 t_2, \cdots, w_n t_n\} \tag{4.1}$$

Он также может обновляться через скользящее окно. После установки длины окна L обучающий набор обновляется с шагом S. i - это количество раз скольжения окна, следовательно, значение i - это количество

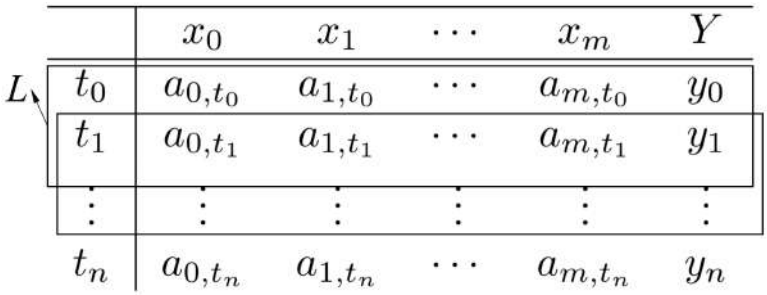


Рис. 4.15: The process of updating forecasting. Every time the window of length L moves backward by S steps, the forecasting model will be retrained and the corresponding prediction value matrix will be output.

обновлений модели прогнозирования (рис. 4.15).

Этот процесс можно обобщить в виде уравнений

$$\begin{aligned} X_{window} &= [i \times S, L + (i \times S)], i = 0, 1, 2, \cdots, k; L > S \\ X_{w0} &= \{t_0, t_1, \cdots, t_l\}, i = 0; l < n \\ X_{w1} &= \{t_s, t_{1+s}, \cdots, t_{l+s}\}, i = 1; l + s < n \\ &\cdots \\ X_{wl} &= \left\{t_{i\times s}, t_{1+(i\times s)}, \cdots, t_{l+(i\times s)}\right\}, l + (i \times s) < n \end{aligned} \tag{4.2}$$

В данной работе используется подход скользящего окна, учитывая его более высокую эффективность, то есть он позволяет максимально сохранить информацию о последних данных в режиме реального времени при небольших вычислительных затратах.

Модель прогнозирования

Поскольку обновление модели прогнозирования требует большого количества повторяющихся вычислений, мы больше склоняемся к более быстрой модели прогнозирования, то есть к той модели прогнозирования, которая может достичь большей точности при более высокой эффективности вычислений.

Результаты наших предыдущих экспериментальных исследований показывают, что LightGBM [19] является лучшей моделью прогнозирования, которая может учитывать оба фактора одновременно, кроме того, титул чемпиона конкурса прогнозирования M5 [8–10] также доказал ее отличную производительность.

Поэтому в этих рамках в качестве модели прогнозирования мы выбрали

LightGBM. На самом деле, модель прогнозирования не является фиксированной, теоретически она может быть заменена на любой тип модели, если это выгодно для экспериментальной цели.

Обновление модели прогнозирования зависит от обновления обучающего набора, поэтому количество обновлений модели прогнозирования также равно i:

$$f_i(X_{window}) = \left\{\hat{Y_0}, \hat{Y_1}, \hat{Y_2}, \cdots, \hat{Y_k}\right\}, k < n \tag{4.3}$$

Среди них f - модель прогнозирования, а $\hat{\text{Y}}$ - матрица прогнозных значений.

Как показано на рисунке 4.14, при непрерывном обновлении обучающего множества, даже если распределение данных изменится, модель прогнозирования будет отражать это изменение во времени с помощью постоянно обновляемого обучающего множества.

XAI объяснение

В объяснении данного вопроса рассматривается использование технологии XAI. Мы попытаемся использовать метод SHAP [51] для объяснения процесса и осуществить дальнейшее улучшение эффективности обновленной модели прогнозирования на основе результатов объяснения.

Вычисление значения SHAP (V) основано на значении Шэпли [57] в кооперативной игре (уравнение 4.4).

$$V_{x_i} = \sum_{U \subseteq M \setminus \{i\}} \frac{|U|!(M - |U| - 1)!}{M!} \left[F_x(U \cup \{i\}) - F_x(U)\right] \tag{4.4}$$

M - размерность всех признаков, U - размерность подмножества признаков, а F - функция признака. Вычисление V, которое представляет собой значение SHAP, является разницей между значениями признаков, присвоенными по принципу Шэпли.

Мы не довольствуемся получением чисел i матрицы прогнозируемых значений с большей точностью, но хотим знать больше информации о модели прогнозирования, включая информацию о признаках в процессе прогнозирования, например, вклад каждого признака в результат прогнозирования. Таким образом, мы получаем определенную степень понимания модели прогнозирования, относящейся к модели "черного

ящика то есть получаем объяснение модели, а результат объяснения дает нам возможность улучшить устойчивость, безопасность и производительность модели прогнозирования.

Объяснимый ИИ делится на локальный и глобальный. И то, и другое - это вклад V признаков в результаты прогнозирования. Разница в том, что область применения локального объяснения g ограничена одним случаем (одна строка), а область применения глобального объяснения G - периодом времени (несколько строк), что по сути является средним весом локальных объяснений за период времени.

$$\begin{aligned} G_i &= SHAP(f_i(X_{window})) \\ &= \{V_{x_0}, V_{x_1}, \cdots, V_{x_m}\} \\ g &= \{V_{a_0,t}, V_{a_1,t}, \cdots, V_{a_m,t}\} \end{aligned} \tag{4.5}$$

Таким образом, теоретически, для аддитивной модели после получения локального объяснения g будет вычислено глобальное объяснение G.

$$V_{x_z} = \sum_{i=0}^{n} V_{a_z,t_i}, z = [0, m], z \subset N \tag{4.6}$$

В задачах временных рядов, чтобы сделать терминологию более разумной, "объяснение"ниже представляет собой глобальное объяснение G, а локальное объяснение g мы используем вместо "объяснения в реальном времени". "Объяснение это вклад подмножества T в результат прогнозирования, а "объяснение в реальном времени это вклад t в результат прогнозирования.

Среди конкретных методов XAI приоритет отдается SHAP, не только потому, что он имеет полную библиотеку ресурсов, к которым можно обращаться напрямую, но, что более важно, этот метод объяснения, основанный на значении Шэпли, имеет математическое доказательство на основе теории игр.

Изменение размеров

Сдвиг окна приводит к обновлению модели прогнозирования, и результаты объяснения обновляются соответствующим образом. На основе каждого результата объяснения исходная матрица признаков улучшается, то есть

признаки с низким значением вклада удаляются, и модель прогнозирования переобучается. По сути, такое изменение размерности в реальном времени представляет собой операцию по снижению уровня шума, что повышает эффективность набора данных. Процесс изменения размерности показан в алгоритме 3.

Algorithm 3 Алгоритм трекера

Вход: Общее количество признаков: M; Объяснимые результаты: G_i; Длина окна L: Шаг перемещения: S

Выход: Результаты оптимизированного прогноза: $f_i\left(X_{new}\right)$

for $i = 0$ to k **do**
  $SORT(G_i) : ascending$
  $X_{window} = [i * S, L + (i * S)]\,,\ i = 0, 1, 2, ..., k$
  Set $h = number(h < M)$
  $Interval = [0, h]$
  $Featuredeleted = SORT(G_i).Interval$
  $X_{new} = X_{window}.DROP(Featuredeleted)$
  $f_i\left(X_{new}\right)$
end for
$f_i\left(X_{new}\right)$

Поскольку удаление размерностей основано на вкладе признака, который является результатом объяснений, на первом этапе нам необходимо отсортировать результат вклада признака G_i. Здесь рассматривается порядок возрастания. Зададим значение h $(h < k)$, построим интервал от 0 до h, а затем извлечем из G_i отсортированные признаки от 0 до h. Наконец, признаки с низким значением вклада удаляются из исходного обучающего множества X_{window} для формирования нового обучающего множества X_{new}, и модель прогнозирования переобучается.

4.3.3 Повышение точности прогнозирования

В данной работе основной процесс оптимизации заключается в использовании результатов объяснения, выдаваемых XAI, для улучшения характеристик исходных данных. Поэтому теоретически любая технология, которая может выводить важность признака (вклад) [113, 114], может достичь цели оптимизации, например, для LightGBM (то есть модели прогнозирования, используемой в данной работе), ее собственная функция

Feature Importance (FI) может выводить важность признака.

Однако у FI есть более очевидные ограничения, чем у технологии XAI, о которой мы говорим [103]. Например, она не может выдавать локальные объяснения, что означает, что она не может достичь объяснения в реальном времени в задачах, связанных с временными рядами. С другой стороны, результат объяснения - это глобальная средняя важность каждого признака, поэтому он не является положительным или отрицательным, а значит, пользователи не могут использовать его, чтобы понять, является ли влияние признаков на модель прогнозирования положительным или отрицательным.

Кроме того, в данной работе наша цель - построить объясняемый и адаптивный фреймворк, поэтому обобщение очень важно для всех частей фреймворка. Это побуждает нас не рассматривать уникальные методы объяснения определенной модели или определенного типа моделей, таких как FI. В целом мы больше склоняемся к технологии XAI, которая позволяет достичь локальных и глобальных объяснений, не зависящих от модели.

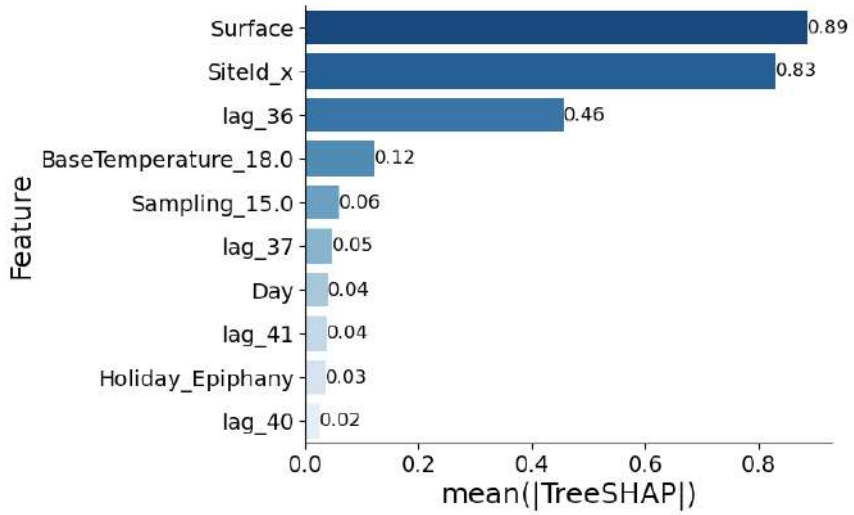


Рис. 4.16: Результаты объяснения.

Мы обучаем статическую модель прогнозирования, используя часть данных, извлеченных из исходных данных. Нас не устраивает простое получение результатов прогнозирования, мы надеемся получить больше информации о внутренней работе модели прогнозирования. Поэтому мы используем $SHAP$ в качестве инструмента объяснения модели

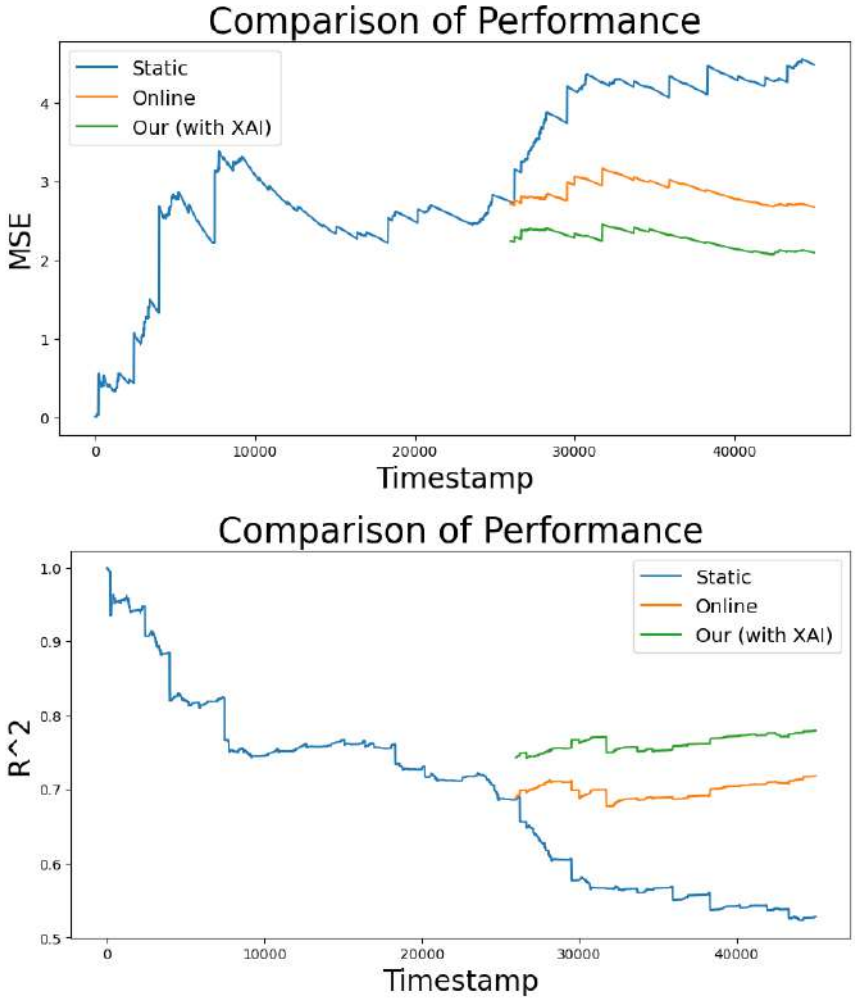


Рис. 4.17: Результаты моделирования

прогнозирования. В данной работе результаты объяснения показаны на рисунке 4.16.

На рисунке 4.16 хорошо видны некоторые очевидные особенности, такие как поверхность, идентификатор участка и особенности запаздывания, которые могут значительно повлиять на спрос на энергию. Увеличение Surface естественным образом приведет к увеличению спроса на энергию.

Судя по конечному результату (см. Рисунок 4.17), наш алгоритм адаптации в большей или меньшей степени оказывает оптимизирующее воздействие на обновленную модель.

С точки зрения оптимизированного тренда, тренд согласуется с исходной обновленной моделью. Таким образом, это также может свидетельствовать о том, что наша система адаптации является оптимизацией обновленной модели прогноза. Если обновленная модель прогноза работает плохо (производительность обновленной модели не так хороша, как статической

модели), то рамки адаптации также оптимизируются на этой основе.

4.4 Вывод из главы 4

В этом разделе рассматривается практическое применение методов искусственного интеллекта (XAI) и подчеркивается их экономическая ценность. Чтобы обеспечить всестороннее понимание, мы провели углубленный анализ многочисленных влиятельных факторов и рассмотрели их влияние. В результате мы смогли значительно повысить точность наших прогнозов за счет интеграции методов XAI. Кроме того, наше исследование внесло значительный вклад в решение проблемы изменения концепции в сценариях онлайн-адаптации. Это достижение особенно важно, поскольку оно открывает путь к решению, которое эффективно справится с этой постоянной проблемой. Объединяя эти ключевые элементы, наше исследование подчеркивает огромный потенциал методов XAI для прибыльного и эффективного применения в реальных приложениях.

Проливая свет на практическое значение и экономические выгоды XAI, наши результаты стимулируют внедрение этих методов для открытия новых возможностей для бизнеса и отраслей промышленности. XAI не только повышает точность и надежность прогнозов, но и предоставляет ценную информацию о процессе принятия решений, обеспечивая прозрачность и интерпретируемость. Это не только укрепляет доверие между заинтересованными сторонами, но и позволяет организациям делать осознанный выбор на основе понятных и обоснованных прогнозов, основанных на ИИ.

Заключение

В этой работе разработан подход XAI, специально ориентированный на прогнозирование временных рядов, и мы называем его ShapTime, поскольку его вычисление основано на значении Шепли. Он позволяет проводить атрибуцию во временном измерении, тем самым объясняя важность самого времени, что отличается от предыдущих работ.

Благодаря объяснению ShapTime мы можем в некоторой степени понять модель прогнозирования. В данных трендовых временных рядов все модели фокусируются на самых последних данных как на наиболее важном объекте обучения, в то время как в данных периодических временных рядов такой закономерности явно не существует, и разные модели не обязательно фокусируются на одном и том же периоде времени для обучения.

С другой стороны, с помощью точного определения времени мы смогли добиться повышения производительности при прогнозировании временных рядов. Заменяя данные в периоды с низким вкладом на данные с высоким вкладом, можно в некоторой степени повысить производительность. Улучшенные показатели производительности показывают, что модель Бустинга и модель, основанная на Bi-RNN, по-прежнему сохраняют свои первоначальные преимущества в отношении периодических данных и данных о тенденциях, соответственно. Таким образом, время формирования показало наиболее значительное улучшение для модели, основанной на Bi-RNN, со средним улучшением на 35%. В частности, время формирования показало наиболее значительное улучшение для двухгрупповой модели, при этом набор данных о солнечной генерации улучшился на 73,87%.

С другой стороны, мы подтверждаем, что построение функций с запаздыванием может повысить производительность прогнозных моделей в задачах прогнозирования временных рядов, но для данных более низкого качества функций с запаздыванием недостаточно. Результаты

моделирования также показывают, что разработанный нами метод автоматического определения запаздывания: эффект улучшения FI-SHAP наиболее стабилен для данных более высокого качества. Для данных более низкого качества FI-SHAP по-прежнему оказывает значительное влияние на прогнозирование XGBoost. В целом, XGBboost превосходит LightGBM по производительности для небольших наборов данных и лучше адаптируется к данным более низкого качества. Большинство существующих методов разработки признаков сосредоточены на задачах классификации, в то время как методы разработки признаков для прогнозирования временных рядов также уделяют мало внимания построению признаков с запаздыванием. В этой работе доказано, что разумное построение признаков с запаздыванием имеет решающее значение.

Благодарности

Я хотел бы выразить глубочайшую благодарность Ованесу Петросяну, моему научному руководителю, за его неизменное руководство, поддержку и бесценные идеи на протяжении всего исследования. Его опыт и постоянная поддержка сыграли решающую роль в формировании этой диссертации. Я также благодарен членам моего диссертационного комитета за их время и ценные отзывы. Их опыт и различные точки зрения значительно повысили качество этой работы.

Особая благодарность выражается Шисян Чжао, Дунфан Ци, Цзин Лю, Жуйминь Ма, Ци Чжао, Цюши Сунь, Цзиньин Цзоу, Фэйран Сюй за помощь на разных этапах работы над проектом. Их вклад и готовность поделиться своими знаниями сыграли важную роль в преодолении трудностей и достижении целей данного исследования. Я хотел бы выразить благодарность Китайскому совету по стипендиям за предоставление необходимых ресурсов, средств и финансирования, которые сделали это исследование возможным.

Я глубоко признателен своей маме и дорогой мисс Мэй за их безусловную любовь, понимание и поддержку на протяжении всей моей учебы. Их непоколебимая вера в меня была постоянным источником силы. И последнее, но не менее важное: я хочу выразить признательность бесчисленному множеству неназванных людей, которые косвенно способствовали проведению этого исследования своей работой,

публикациями или общими знаниями. Ваш вклад был неоценим в формировании моего понимания и методологии.

В заключение хочу сказать, что эта диссертация была бы невозможна без поддержки и вклада вышеупомянутых людей и многих других, сыгравших решающую роль в моем академическом пути. Благодарю вас всех от всего сердца.

Литература

[1] Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning //Electronic Markets. - 2021. - Vol. 31, No. 3. - P. 685-695.

[2] Moein M. M. et al. Predictive models for concrete properties using machine learning and deep learning approaches: A review //Journal of Building Engineering. - 2023. - Vol. 63, - P. 105444.

[3] Choi R. Y. et al. Introduction to machine learning, neural networks, and deep learning //Translational vision science and technology. - 2020. - Vol. 9, No. 2. - P. 14.

[4] Kaytez F. A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption //Energy. - 2020. - Vol. 197, - P. 117200.

[5] Wu H. et al. Autoformer:Decomposition transformers with autocorrelation for long-term series forecasting //Advances in neural information processing systems. - 2021. - Vol. 34, - P. 22419-22430.

[6] Zhou T. et al. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting //International conference on machine learning. - PMLR, - 2022. - P. 27268-27286.

[7] Wellens A. P., Udenio M., Boute R. N. Transfer learning for hierarchical forecasting: Reducing computational efforts of M5 winning methods //International Journal of Forecasting. - 2022. - Vol. 38, No. 4. - P. 1482-1491.

[8] Makridakis S., Petropoulos F., Spiliotis E. Introduction to the M5 forecasting competition Special Issue //International Journal of Forecasting. - 2022. - Vol. 38, No. 4. - P. 1279.

[9] Makridakis S., Spiliotis E., Assimakopoulos V. The M5 competition: Background, organization, and implementation //International Journal of Forecasting. - 2022. - Vol. 38, No. 4. - P. 1325-1336.

[10] Makridakis S., Petropoulos F., Spiliotis E. The M5 competition: Conclusions //International Journal of Forecasting. - 2022. - Vol. 38, No. 4. - P. 1576-1582.

[11] Makridakis S. et al. The M6 forecasting competition: Bridging the gap between forecasting and investment decisions // arXiv preprint arXiv:2310.13357. - 2023.

[12] Makridakis S. et al. Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward //Journal of the Operational Research Society. - 2023. - Vol. 74, No. 3. - P. 840-859.

[13] Schraagen J. M. Responsible use of AI in military systems: Prospects and challenges //Ergonomics. - 2023. - Vol. 66, No. 11. - P. 1729.

[14] Zhang Y. L. et al. Application of artificial intelligence in military: From projects view //2020 6th International Conference on Big Data and Information Analytics (BigDIA). - IEEE, 2020. - P. 113-116.

[15] Reddy S. et al. A governance model for the application of AI in health care //Journal of the American Medical Informatics Association. - 2020. - Vol. 27, No. 3. - P. 491-497.

[16] Morley J. et al. The ethics of AI in health care: a mapping review //Social Science and Medicine. - 2020. - Vol. 260, - P. 113172.

[17] Cao L. Ai in finance: challenges, techniques, and opportunities //ACM Computing Surveys (CSUR). - 2022. - Vol. 55, No. 3. - P. 1-38.

[18] Goodell J. W. et al. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis //Journal of Behavioral and Experimental Finance. - 2021. - Vol. 32, - P. 100577.

[19] Zhang Y. et al. Comparison and explanation of forecasting algorithms for energy time series //Mathematics. - 2021. - Vol. 9. - No. 21. - P. 2794.

[20] Zhang Y. et al. FI-SHAP: explanation of time series forecasting and improvement of feature engineering based on boosting algorithm //Intelligent Systems Conference. - Cham : Springer International Publishing, - 2022. - P. 745-758.

[21] Zhang, Y., Sun, Q., Liu, J. et al. Long-Term Forecasting of Air Pollution Particulate Matter (PM2. 5) and Analysis of Influencing Factors //Sustainability. - 2023. - Vol. 16, No. 1. - P. 19.

[22] Petrosian.O, and Yuyi Zhang. Solar Power Generation Forecasting in Smart Cities and Explanation Based on Explainable AI. //Smart Cities. - 2024. - Vol. 7, No. 6. -P. 3388-3411.

[23] Zhang Y. et al. XAI evaluation: evaluating black-box model explanations for prediction //2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT). - IEEE, 2021. - P. 13-16.

[24] Zou J. et al. High-dimensional explainable AI for cancer detection //International Journal of Artificial Intelligence. - 2021. - Vol. 19. - No. 2. - P. 195.

[25] Sun Q. et al. Resource Allocation in Heterogeneous Network with Supervised GNNs //International Conference on Swarm Intelligence. - Cham : Springer Nature Switzerland, 2023. - P. 350-361.

[26] Zhang Y. et al. ShapTime: A General XAI Approach for Explainable Time Series Forecasting //Intelligent Systems Conference. - Cham : Springer Nature Switzerland, 2023. - P. 659-673.

[27] Zhao, S., Petrov, Y. V., Zhang, Y. et al. Modeling of the thermal softening of metals under impact loads and their temperature-time correspondence //International Journal of Engineering Science. - 2024. - Vol. 194, - P. 103969.

[28] Ma, R., Zhang, Y., Liu, J. et al. Prediction of Next App in OS //2022 III International Conference on Neural Networks and Neurotechnologies (NeuroNT). - IEEE, 2022. - P. 28-31.

[29] Ma R, Zhang Y, Liu J, et al. Forecasting and XAI for Applications Usage in OS //Machine Learning and Artificial Intelligence. - IOS Press, 2022. - P. 17-27.

[30] Zhang Y. et al. Automated feature engineering based on explainable artificial intelligence for time series forecasting // Engineering Applications of Artificial Intelligence. - Under review.

[31] Zhang Y. et al. XAI-Based Explainable Adaptation Framework for Handling Concept Drift in Time Series Forecasting //Knowledge-based systems. - Under review.

[32] Dwivedi R. et al. Explainable AI (XAI): Core ideas, techniques, and solutions //ACM Computing Surveys. - 2023. - Vol. 55, No. 9. - P. 1-33.

[33] Saeed W., Omlin C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities //Knowledge-Based Systems. - 2023. - Vol. 263, - P. 110273.

[34] Lundberg S. M. et al. From local explanations to global understanding with explainable AI for trees //Nature machine intelligence. - 2020. - Vol. 2, No. 1. - P. 56-67.

[35] Pan Q., Hu W., Chen N. Two Birds with One Stone: Series Saliency for Accurate and Interpretable Multivariate Time Series Forecasting //IJCAI. - 2021. - P. 2884-2891.

[36] Ozyegen O., Ilic I., Cevik M. Evaluation of interpretability methods for multivariate time series forecasting //Applied Intelligence. - 2022. - P. 1-17.

[37] Jabeur S. B., Mefteh-Wali S., Viviani J. L. Forecasting gold price with the XGBoost algorithm and SHAP interaction values //Annals of Operations Research. - 2024. - Vol. 334. - No. 1. - P. 679-699.

[38] Oreshkin B. N. et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting //arXiv preprint arXiv:1905.10437. - 2019.

[39] Wang J. et al. Multilevel wavelet decomposition network for interpretable time series analysis //Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. - 2018. - P. 2437-2446.

[40] Shen Q. et al. Visual interpretation of recurrent neural network on multi-dimensional time-series forecast //2020 IEEE Pacific visualization symposium (PacificVis). - IEEE, 2020. - P. 61-70.

[41] Guo T., Lin T., Antulov-Fantulin N. Exploring interpretable LSTM neural networks over multi-variable data //International conference on machine learning. - PMLR, 2019. - P. 2494-2504.

[42] Lim B. et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting //International Journal of Forecasting. - 2021. - Vol. 37. - No. 4. - P. 1748-1764.

[43] Ding Y. et al. Interpretable spatio-temporal attention LSTM model for flood forecasting //Neurocomputing. - 2020. - Vol. 403. - P. 348-359.

[44] Zhou B. et al. Interpretable temporal attention network for COVID-19 forecasting //Applied soft computing. - 2022. - Vol. 120. - P. 108691.

[45] Schetinin V. et al. Confident interpretation of Bayesian decision tree ensembles for clinical applications //IEEE Transactions on Information Technology in Biomedicine. - 2007. - Vol. 11, No. 3. - P. 312-319.

[46] Speith T. A review of taxonomies of explainable artificial intelligence (XAI) methods //Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. - 2022. - P. 2239-2250.

[47] Dwivedi R. et al. Explainable AI (XAI): Core ideas, techniques, and solutions //ACM Computing Surveys. - 2023. - Vol. 55, No. 9. - P. 1-33.

[48] Arrieta A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI //Information fusion. - 2020. - Vol. 58. - P. 82-115.

[49] Alsahaf A. et al. A framework for feature selection through boosting //Expert Systems with Applications. - 2022. - Vol. 187. - P. 115895.

[50] Upadhyay D. et al. Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids //IEEE Transactions on Network and Service Management. - 2020. - Vol. 18, No. 9. - P. 1104-1116.

[51] Lundberg S. M., Lee S. I. A unified approach to interpreting model predictions //Advances in neural information processing systems. - 2017. - Vol. 30.

[52] Ribeiro M. T., Singh S., Guestrin C. "Why should i trust you?"Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. - 2016. - P. 1135-1144.

[53] Lim B. et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting //International Journal of Forecasting. - 2021. - Vol. 37, No. 4. - P. 1748-1764.

[54] Lin Y., Koprinska I., Rana M. Temporal convolutional attention neural networks for time series forecasting //2021 International joint conference on neural networks (IJCNN). - IEEE, 2021. - P. 1-8.

[55] Lopes P. et al. XAI systems evaluation: A review of human and computer-centred methods //Applied Sciences. - 2022. - Vol. 12, No. 19, - P. 9423.

[56] Schlegel U. et al. Towards a rigorous evaluation of XAI methods on time series //2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). - IEEE, 2019. - P. 4197-4201.

[57] Lloyd S Shapley. A value for n-person games. In: Contributions to the Theory of Games. - 1953. - Vol. 2, No. 28, - P. 307-317.

[58] Bach S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation //PloS one. - 2015. - Vol. 10, No. 7, - P. 0130140.

[59] Charnes A. et al. Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations //Econometrics of planning and efficiency. - 1988. - P. 123-133.

[60] Datta A., Sen S., Zick Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems //2016 IEEE symposium on security and privacy (SP). - IEEE, 2016. - C. 598-617.

[61] Lipovetsky S., Conklin M. Analysis of regression in game theory approach //Applied stochastic models in business and industry. - 2001. - Vol. 17, No. 4, - P. 319-330.

[62] Shrikumar A., Greenside P., Kundaje A. Learning important features through propagating activation differences //International conference on machine learning. - PMLR, 2017. - P. 3145-3153.

[63] Shrikumar A. et al. Not just a black box: Learning important features through propagating activation differences //arXiv preprint arXiv:1605.01713. - 2016.

[64] Strumbelj E., Kononenko I. Explaining prediction models and individual predictions with feature contributions //Knowledge and information systems. - 2014. - Vol. 41, - P. 647-665.

[65] Young H. P. Monotonic solutions of cooperative games //International Journal of Game Theory. - 1985. - Vol. 14, No. 4, - P. 65-72.

[66] Jain S., Wallace B. C. Attention is not explanation //arXiv preprint arXiv:1902.10186. - 2019.

[67] Serrano S., Smith N. A. Is attention interpretable? //arXiv preprint arXiv:1906.03731. - 2019.

[68] Wiegreffe S., Pinter Y. Attention is not not explanation //arXiv preprint arXiv:1908.04626. - 2019.

[69] Zhou Z. H., Zhou Z. H. Ensemble learning. - Springer Singapore, 2021. - P. 181-210.

[70] da Silva R. G. et al. A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting //Energy. - 2021. - Vol. 216. - P. 119174.

[71] Qiu R. et al. Generalized Extreme Gradient Boosting model for predicting daily global solar radiation for locations without historical data //Energy Conversion and Management. - 2022. - Vol. 258. - P. 115488.

[72] Ribeiro F., Gradvohl A. L. S. Machine learning techniques applied to solar flares forecasting //Astronomy and Computing. - 2021. - Vol. 35. - P. 100468.

[73] Fan J. et al. Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data //Journal of Cleaner Production. - 2020. - Vol. 248. - P. 119264.

[74] Cabaneros S M, Calautit J K, Hughes B R. A review of artificial neural network models for ambient air pollution prediction //Environmental Modelling and Software. - 2019. - Vol. 119. - P. 285-304.

[75] Kiranyaz S, Avci O, Abdeljaber O, et al. 1D convolutional neural networks and applications: A survey //Mechanical systems and signal processing. - 2021. - Vol. 151. - P. 107398.

[76] Cossu A, Carta A, Lomonaco V, et al. Continual learning for recurrent neural networks: an empirical evaluation //Neural Networks. - 2021. - Vol. 143. - P. 607-627.

[77] Makridakis S., Spiliotis E., Assimakopoulos V. The M4 Competition: 100,000 time series and 61 forecasting methods //International Journal of Forecasting. - 2020. - Vol. 36. - No. 1. - P. 54-74.

[78] Makridakis S., Spiliotis E., Assimakopoulos V. M5 accuracy competition: Results, findings, and conclusions //International Journal of Forecasting. - 2022. - Vol. 38. - No. 4. - P. 1346-1364.

[79] Al Daoud E. 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.

[80] Hong J. et al. An application of XGBoost, LightGBM, CatBoost algorithms on house price appraisal system //Housing Finance Research. - 2020. - Vol. 4. - P. 33-64.

[81] Bae D. J., Kwon B. S., Song K. B. XGBoost-based day-ahead load forecasting algorithm considering behind-the-meter solar PV generation //Energies. - 2021. - Vol. 15. - No. 1. - P. 128.

[82] Aksoy N., Genc I. Predictive models development using gradient boosting based methods for solar power plants //Journal of Computational Science. - 2023. - Vol. 67. - P. 101958.

[83] Pazikadin A. R. et al. Solar irradiance measurement instrumentation and power solar generation forecasting based on Artificial Neural Networks (ANN): A review of five years research trend //Science of The Total Environment. - 2020. - Vol. 715. - P. 136848.

[84] Vu B. H., Chung I. Y. Optimal generation scheduling and operating reserve management for PV generation using RNN-based forecasting models for stand-alone microgrids //Renewable Energy. - 2022. - Vol. 195. - P. 1137-1154.

[85] Neshat M. et al. Short-term solar radiation forecasting using hybrid deep residual learning and gated LSTM recurrent network with differential covariance matrix adaptation evolution strategy //Energy. - 2023. - Vol. 278. - P. 127701.

[86] Liu Y. et al. An attention-based category-aware GRU model for the next POI recommendation //International Journal of Intelligent Systems. - 2021. - Vol. 36. - No. 7. - P. 3174-3189.

[87] Peng T. et al. An integrated framework of Bi-directional long-short term memory (BiLSTM) based on sine cosine algorithm for hourly solar radiation forecasting //Energy. - 2021. - Vol. 221. - P. 119887.

[88] Alshemali B., Kalita J. Improving the reliability of deep neural networks in NLP: A review //Knowledge-Based Systems. - 2020. - Vol. 191. - P. 105210.

[89] Liang Y. et al. Explaining the black-box model: A survey of local interpretation methods for deep neural networks //Neurocomputing. - 2021. - Vol. 419. - P. 168-182.

[90] Alshawaf M., Poudineh R., Alhajeri N. S. Solar PV in Kuwait: The effect of ambient temperature and sandstorms on output variability and uncertainty //Renewable and sustainable energy reviews. - 2020. - Vol. 134. - P. 110346.

[91] Belhaouas N. et al. A new approach of PV system structure to enhance performance of PV generator under partial shading effect //Journal of Cleaner Production. - 2021. - Vol. 317. - P. 128349.

[92] Tu J. et al. Experimental study on the influence of bionic channel structure and nanofluids on power generation characteristics of waste heat utilisation equipment //Applied Thermal Engineering. - 2022. - Vol. 202. - P. 117893.

[93] Vilone G., Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence //Information Fusion. - 2021. - Vol. 76. - P. 89-106.

[94] Burkart N., Huber M. F. A survey on the explainability of supervised machine learning //Journal of Artificial Intelligence Research. - 2021. - Vol. 70. - P. 245-317.

[95] Heuillet A., Couthouis F., Díaz-Rodríguez N. Explainability in deep reinforcement learning //Knowledge-Based Systems. - 2021. - Vol. 214. - P. 106685.

[96] Vale D., El-Sharif A., Ali M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law //AI and Ethics. - 2022. - Vol. 2. - No. 4. - P. 815-826.

[97] Colin J. et al. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods //Advances in neural information processing systems. - 2022. - Vol. 35. - P. 2832-2845.

[98] Ferrario A., Loi M. How explainability contributes to trust in AI //Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. - 2022. - P. 1457-1466.

[99] Shrikumar A. et al. Not just a black box: Learning important features through propagating activation differences //arXiv preprint arXiv:1605.01713. - 2016.

[100] Shrikumar A., Greenside P., Kundaje A. Learning important features through propagating activation differences //International conference on machine learning. - PMLR, 2017. - P. 3145-3153.

[101] Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks //International conference on machine learning. - PMLR, 2017. - P. 3319-3328.

[102] Bach S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation //PloS one. - 2015. - Vol. 10. - No. 7. - P. e0130140.

[103] Sundararajan M., Najmi A. The many Shapley values for model explanation //International conference on machine learning. - PMLR, 2020. - P. 9269-9278.

[104] Owen A. B., Prieur C. On Shapley value for measuring importance of dependent inputs //SIAM/ASA Journal on Uncertainty Quantification. - 2017. - Vol. 5. - No. 1. - P. 986-1002.

[105] Ghafarian F. et al. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data //Environmental Modelling and Software. - 2022. - Vol. 156. - P. 105466.

[106] Altman N, Krzywinski M. Ensemble methods: bagging and random forests //Nature Methods. - 2017. Vol. 14. - No. 10. - P. 933-935.

[107] Benidis K. et al. Deep learning for time series forecasting: Tutorial and literature survey //ACM Computing Surveys. - 2022. - Vol. 55, No. 6. - P. 1-36.

[108] Mahmoud A., Mohammed A. A survey on deep learning for time-series forecasting //Machine learning and big data analytics paradigms: analysis, applications and challenges. - 2021. - P. 365-392.

[109] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network //Physica D: Nonlinear Phenomena. - 2020. - Vol. 404. - P. 132306.

[110] Hewamalage H., Bergmeir C., Bandara K. Recurrent neural networks for time series forecasting: Current status and future directions //International Journal of Forecasting. - 2021. - Vol. 37. - No. 1. - P. 388-427.

[111] Orvieto A. et al. Resurrecting recurrent neural networks for long sequences //International Conference on Machine Learning. - PMLR, 2023. - P. 26670-26698.

[112] Lewis R. J. An introduction to classification and regression tree (CART) analysis //Annual meeting of the society for academic emergency medicine in San Francisco, California. - San Francisco, CA, USA : Department of Emergency Medicine Harbor-UCLA Medical Center Torrance, 2000. - Vol. 14.

[113] Schapire R. E. et al. A brief introduction to boosting //IJCAI. - 1999. - Vol. 99, No. 999. - P. 1401-1406.

[114] Mayr A. et al. The evolution of boosting algorithms //Methods of information in medicine. - 2014. - Vol. 53, No. 06. - P. 419-427.

[115] Guyon I., Elisseeff A. An introduction to feature extraction //Feature extraction: foundations and applications. - Berlin, Heidelberg : Springer Berlin Heidelberg, 2006. - P. 1-25.

[116] Kanter J. M., Veeramachaneni K. Deep feature synthesis: Towards automating data science endeavors //2015 IEEE international conference on data science and advanced analytics (DSAA). - IEEE, 2015. - P. 1-10.

[117] Katz G., Shin E. C. R., Song D. Explorekit: Automatic feature generation and selection //2016 IEEE 16th International Conference on Data Mining (ICDM). - IEEE, 2016. - P. 979-984.

[118] Kaul A., Maheshwary S., Pudi V. Autolearn—automated feature generation and selection //2017 IEEE International Conference on data mining (ICDM). - IEEE, 2017. - P. 217-226.

[119] Khurana U. et al. Cognito: Automated feature engineering for supervised learning //2016 IEEE 16th international conference on data mining workshops (ICDMW). - IEEE, 2016. - P. 1304-1307.

[120] Lam H. T. et al. One button machine for automating feature engineering in relational databases //arXiv preprint: 1706.00327. - 2017.

[121] Cerqueira V., Moniz N., Soares C. Vest: Automatic feature engineering for forecasting //Machine Learning. - 2021. - P. 1-23.

[122] Li L. et al. Research on feature engineering for time series data mining //2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC). - IEEE, 2018. - P. 431-435.

[123] Zdravevski E. et al. Robust histogram-based feature engineering of time series data //2015 Federated Conference on Computer Science and Information Systems (FedCSIS). - IEEE, 2015. - P. 381-388.

[124] Selvam S. K., Rajendran C. tofee-tree: au to matic fe ature e ngineering framework for modeling tre nd-cycl e in time series forecasting //Neural Computing and Applications. - 2023. - Vol. 35, No. 16. - P. 11563-11582.

[125] Punmiya R., Choe S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing //IEEE Transactions on Smart Grid. - 2019. - Vol. 10, No. 2. - P. 2326-2329.

[126] Hu Y. et al. Faster clinical time series classification with filter based feature engineering tree boosting methods //Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability. - 2021. - P. 247-260.

[127] Shannon C. E. A mathematical theory of communication //The Bell system technical journal. - 1948. - Vol. 27, No. 3. - P. 379-423.

[128] Letham B. et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. - 2015. - P. 1350-1371.

[129] Caruana R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission //Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. - 2015. - P. 1721-1730.

[130] Agarwal R. et al. Neural additive models: Interpretable machine learning with neural nets //Advances in neural information processing systems. - 2021. - Vol. 34. - P. 4699-4711.

[131] Bayram F., Ahmed B. S., Kassler A. From concept drift to model degradation: An overview on performance-aware drift detectors //Knowledge-Based Systems. - 2022. - Vol. 245. - P. 108632.

[132] Agrahari S., Singh A. K. Concept drift detection in data stream mining: A literature review //Journal of King Saud University-Computer and Information Sciences. - 2022. - Vol. 34. - No. 10. - P. 9523-9540.