

Исследование методов прогнозирования временных рядов для предсказания качества воздуха: объяснительный сравнительный анализ

Д. Ци, В. М. Буре

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: Ци Д., Буре В. М. Исследование методов прогнозирования временных рядов для предсказания качества воздуха: объяснительный сравнительный анализ // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2024. Т. 20. Вып. 2. С. 206–219. <https://doi.org/10.21638/spbu10.2024.206>

Анализируется эффективность моделей прогнозирования временных рядов для предсказания качества воздуха. Для этого используются наборы данных от сенсора Purple Air Dual Laser Air Quality Sensor и платформы Kaggle Online. Полученные данные содержат достоверную информацию, необходимую для охраны окружающей среды. В ходе исследования основное внимание уделяется определению подходящих моделей прогнозирования для анализа окружающей среды, включая популярные структуры алгоритмов, такие как нейронные сети и ансамблевые модели. Также применяется метод объяснительного искусственного интеллекта, который обеспечивает объяснения для моделей с высокой производительностью и повышает их доверие и прозрачность. Производительность моделей оценивалась с помощью метрик, средней абсолютной ошибки, квадратного корня из средней квадратичной ошибки и R -квадрата. Результаты показывают, что нейронные сети и ансамблевые модели эффективны для прогнозирования временных рядов качества воздуха. Это исследование вносит вклад в развитие моделей прогнозирования временных рядов и предоставляет полезные рекомендации для будущих исследований в области прогнозирования качества воздуха.

Ключевые слова: качество воздуха, прогнозирование временных рядов, нейронные сети, ансамблевые модели, объяснительный искусственный интеллект.

1. Введение. Важность качества воздуха, особенно PM_{2.5}, и точное его прогнозирование — это ключевые темы в экологических исследованиях. PM_{2.5} относится к мелким аэрозольным частицам диаметром 2.5 мкм или меньше, которые могут оказывать негативное влияние на здоровье людей и окружающую среду. Точное прогнозирование качества воздуха, включая концентрацию PM_{2.5}, играет важную роль в оценке потенциальных рисков для здоровья, использовании эффективных мер по борьбе с загрязнением и разработке соответствующих стратегий управления окружающей средой. Своевременные и надежные прогнозы позволяют властям и отдельным лицам принимать проактивные меры, такие как предупреждения о здоровье, корректировка активностей на открытом воздухе или внедрение инициатив по сокращению загрязнений.

Линейная регрессия и ARIMA — это широко используемые методы прогнозирования временных рядов. Линейная регрессия имеет некоторые ограничения в моделировании нелинейных отношений и сложных закономерностей, а ARIMA может испытывать сложности при работе с нестационарными данными и сложными нелиней-

ными отношениями. Для преодоления этих ограничений все большую популярность набирают алгоритмы машинного и глубокого обучения в прогнозировании временных рядов. Методы машинного обучения предлагают гибкость в моделировании сложных отношений и работы с нелинейностями. Применение методов машинного и глубокого обучения в прогнозировании временных рядов положило начало новой эпохе в этой области и обеспечивает повышенную точность, гибкость и надежность по сравнению с традиционными методами. Работа [1] посвящена прогнозированию почасовой концентрации PM_{2.5} в Китае с использованием алгоритма XGBoost. В ней оценивается производительность XGBoost путем сравнения наблюдаемых и предсказанных концентраций PM_{2.5}, что подтверждает его превосходство по сравнению с другими методами машинного обучения. В статье [2] использовали модель нейронной сети для прогнозирования качества воздуха. LSTM модель рассмотрена в [3], а RNN модель — в [4].

Цель настоящего исследования — критическая оценка и выбор эффективной и интерпретируемой модели прогнозирования путем проведения всестороннего сравнительного анализа широко признанных моделей, применяемых в конкурсах по прогнозированию временных рядов. Для обеспечения практической значимости прогностических моделей будем использовать достоверные записи сенсоров в качестве наборов данных. Кроме того, сосредоточение на управлении окружающей средой гарантирует актуальность и применимость исследовательских результатов к реальным сценариям.

Идентификация факторов, таких как температура, давление, направление ветра и осадки, является ключевой для оптимизации систем прогнозирования качества воздуха. Эти факторы непосредственно воздействуют на формирование, рассеивание и концентрацию атмосферных загрязнителей. Понимание их влияния позволяет проводить целенаправленные мероприятия, такие как контроль выбросов и управление транспортом, для эффективного смягчения загрязнения. Оптимизация систем прогнозирования с применением этих факторов способствует принятию обоснованных решений и улучшает управление загрязнением. Действительно, при использовании нейронных сетей в большинстве случаев сложно объяснить, какой именно фактор влияет на зависимую переменную, и представить явную функциональную зависимость.

Внедрение технологии объяснительного искусственного интеллекта (XAI) предлагает новое видение оценки прогностических моделей. Будем использовать метод SHapley Additive exPlanations (SHAP) — широко применяемую технику XAI, которая дает возможность оценить вклад каждого признака в конечный прогноз. Определяя наиболее влиятельные признаки в процессе принятия решений модели, получаем более глубокое понимание базовых данных, что может раскрыть потенциальные области для улучшения модели. Такой метод помогает понять, как каждая переменная вносит свой вклад в предсказание, несмотря на то, что он не дает явную функциональную зависимость.

Уже опубликован ряд исследований, в которых применялись различные методы для прогнозирования индекса качества воздуха. Однако они требуют учета нескольких факторов, которые могут повлиять на их согласованность и надежность. Во-первых, отсутствует стандартизированный процесс применения этих методов, что приводит к несогласованным результатам. Во-вторых, не проводится достаточный анализ факторов, оказывающих влияние на качество воздуха. Направленное развитие будущих работ должно устранить эти проблемы и повысить качество и надежность подобных работ.

Основное внимание будем уделять следующим основным вкладам: предложение стандартного процесса использования методов машинного обучения для обеспечения справедливого сравнения производительности, введение интерпретируемых методов ХАИ для прогнозирования качества воздуха и комплексный анализ факторов, влияющих на качество воздуха.

2. Выбор данных. В сокращении набора данных Kaggle применяется комплексная коллекция экологических измерений, включая PM2.5, которая представляет собой аэрозольные частицы размером 2.5 мкм или меньше и оказывает значительное воздействие на качество воздуха [5]. В данном случае PM2.5 является зависимой переменной, а также присутствуют 16 независимых переменных, включая No (номер), year (год), month (месяц), day (день), hour (час), PM10, SO2, NO2, CO, O3, TEMP (температура), PRES (давление), DEWP (температуру точки росы), gain (указывает на выпадение осадков), wd (направление ветра) и WSPM (скорость ветра).

Измерения PM2.5 имеют большое значение для здоровья человека, поскольку эти частицы могут проникать в легкие и вызывать отрицательные последствия для организма [6]. Существует корреляция между уровнем PM2.5, изменениями давления и содержанием загрязнителей, связанных с метеорологическими факторами. Эти факторы могут оказывать прямое влияние на распространение загрязнителей, химический состав и климатическую стабильность, что, в свою очередь, влияет на уровни качества воздуха [7].

Анализируя все различные параметры окружающей среды, можно получить более подробное представление о динамике качества воздуха, особенно в контексте температуры воздуха. Рассмотрение таких факторов дает возможность лучше понять взаимосвязи и влияние различных параметров на качество воздуха.

Анализ временных рядов имеет несколько преимуществ при оценке качества воздуха, делая его важным инструментом для исследователей и экологических менеджеров. Использование моделей временных рядов дает возможность обнаружить скрытые тренды, сезонные закономерности и циклическое поведение данных о качестве воздуха. С помощью этих моделей аналитики могут выявлять и интерпретировать повторяющиеся закономерности, возникающие со временем, а также понимать сложные взаимосвязи между различными загрязнителями воздуха. Таким образом, комплексное понимание такой временной динамики — ключевое для эффективного управления окружающей средой, разработки политики и применения целевых стратегий вмешательства для смягчения загрязнения воздуха.

Применение моделей позволяет интегрировать передовые техники машинного обучения и технологии ХАИ, улучшая интерпретируемость и прозрачность анализа качества воздуха. Включение методов, таких как SHAP, дает возможность исследователям определить наиболее влиятельные признаки в процессе принятия решений моделями, что приводит к лучшему пониманию базовых данных и взаимосвязей, которое способствует доверию к моделям со стороны заинтересованных сторон и позволяет принять данные ориентированных подходов в управлении окружающей средой. Более того, непрерывное развитие методов моделирования временных рядов способствует повышению точности и надежности прогнозов качества воздуха, дополнительно усиливая их применимость и ценность в данной области.

В рамках настоящего исследования конкретный момент времени служит границей между обучающим и тестовым наборами данных вместо произвольного случайного выбора. Такой подход более отвечает внутренней логике временных рядов. Примерно 70 % элементов выделены для обучающего набора данных. Учитывая раз-

личные пропорции отсутствующих значений для разных видов данных, можно обеспечить достаточное количество признаков при прогнозировании концентрации частиц PM_{2.5} в качестве целевой переменной для данного исследования (рис. 1). Установив разумную границу, пропорции обучающего набора данных, тестового и валидационного наборов данных определены как 70, 15 и 15 % соответственно.

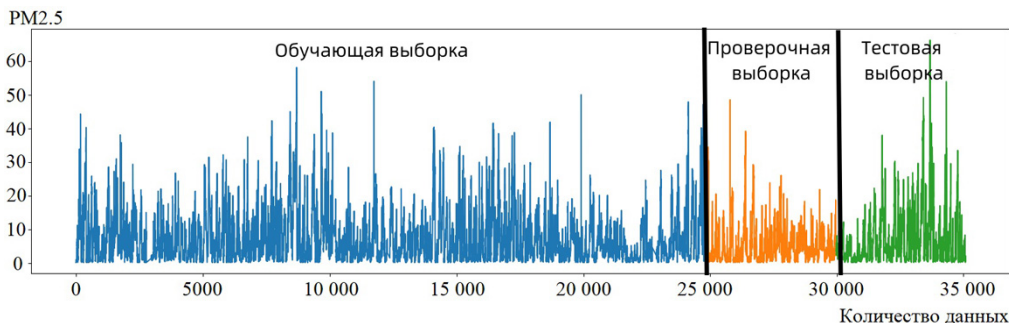


Рис. 1. Распределение данных

3. Модели прогнозирования в приложениях временных рядов.

3.1. Основные методы. В настоящее время структуры нейронных сетей и ансамблевых моделей широко применяются в различных приложениях. Поэтому важно выбрать представительные модели в этих структурах для всестороннего сравнительного анализа прогностических моделей. В качестве модели нейронной сети выбираем искусственные нейронные сети (ANN), рекуррентные нейронные сети (RNN), двунаправленные рекуррентные нейронные сети (Bi-RNN), двунаправленные долгосрочные нейронные сети с короткой памятью (Bi-LSTM) и двунаправленные блокирующие рекуррентные единицы (Bi-GRU). Ансамблевые модели включают алгоритмы бустинга и бэггинга; выбираем LightGBM, XGBoost и CatBoost.

Принципиально оба типа моделей — и ансамблевые, и нейронных сетей — решают проблемы путем последовательной декомпозиции, а не разбиения всего набора данных с помощью сложных границ, как в методе опорных векторов или логистической регрессии. Так как ясно, что алгоритмы на основе деревьев итеративно разделяют пространство признаков по различным признакам для оптимизации информационного выигрыша, нейронные сети работают аналогично, но менее очевидным образом. Каждый нейрон отслеживает конкретную область пространства признаков с некоторыми перекрытиями, активируя определенные нейроны при входе в эти области. Нейронные сети подходят к такому моделированию путем построения моделей с использованием метода вероятностной разбивки, тогда как в методах на основе деревьев применяется детерминированный подход. В любом случае производительность обоих типов моделей зависит от их глубины, поскольку их компоненты соответствуют различным сегментам пространства признаков. В настоящей работе в качестве объектов исследования выбраны LightGBM, ANN, RNN, Bi-RNN, Bi-LSTM и Bi-GRU. Напрямую используем версию LightGBM, победившую в соревнованиях M5, которая демонстрирует улучшенную адаптивность к сложным структурам признаков и значительно улучшенную предсказательную производительность по сравнению со стандартной версией.

3.2. Light GBM. Light Gradient Boosting Machine (LightGBM) — это усовершенствованная версия алгоритма градиентного бустинга на деревьях решений (GBDT),

характеризующаяся превосходным качеством работы в различных задачах машинного обучения, особенно с большими наборами данных и сложными пространствами признаков. В ней применяется метод на основе гистограмм для построения и разделения деревьев решений, что уменьшает использование памяти и вычислительную нагрузку путем дискретизации непрерывных значений признаков в дискретные блоки. В результате LightGBM эффективнее обрабатывает большие наборы данных по сравнению с традиционными реализациями GBDT.

Еще одна отличительная характеристика LightGBM — это стратегия Exclusive Feature Bundling (EFB), которая использует разреженные структуры данных для идентификации взаимоисключающих признаков и их объединения, что эффективно уменьшает размерность набора данных. Это позволяет избежать «проклятия размерности» и ускоряет процесс обучения без потери предсказательной способности модели.

LightGBM предлагает две инновационные стратегии роста деревьев: Leaf-wise и Depth-wise для улучшения производительности модели. Подход Leaf-wise разделяет узлы дерева с наибольшей дельтой функции потерь, в то время как метод Depth-wise способствует сбалансированному развитию дерева. Предоставление этих альтернативных подходов позволяет пользователям выбрать соответствующую стратегию роста в зависимости от конкретной задачи и вычислительных ограничений. Как исследователи, так и практики могут получить выгоду от преимуществ LightGBM, особенно при работе с обширными и высокоразмерными наборами данных. Версия M5 LightGBM относится к хорошо оптимизированной реализации, использующей тщательный подбор признаков для извлечения высококачественной информации из данных.

3.3. ANN. Artificial Neural Network (ANN) — это основная архитектура машинного обучения, которая моделирует сложные закономерности в данных, имитируя структуру и функцию биологических нейронов. ANN применяются в задачах распознавания образов, обработки естественного языка и кластеризации [8].

3.4. RNN, LSTM, GRU. Recurrent Neural Network (RNN) — это специализированная нейронная сеть, разработанная для работы с последовательными данными путем включения рекуррентных связей, что позволяет сохранять информацию на протяжении временных шагов [9]. RNN достигает значительной производительности в таких задачах, как обработка естественного языка, анализ временных рядов и распознавание речи.

Long Short-Term Memory (LSTM) — это продвинутая модификация RNN, которая использует механизмы гейтов для регулирования потока информации и устранения проблемы затухающего градиента, характерной для глубоких RNN. Внутренние вентили — входной, забывания и выходной — в LSTMs позволяют лучше понять долгосрочные зависимости и сложные шаблоны в последовательностях, что улучшает производительность в задачах машинного перевода и генерации текста.

Gated Recurrent Unit (GRU) — еще одно расширение RNN, упрощающее архитектуру LSTM с использованием вентиля обновления и сброса для регулирования потока информации [10]. Оно предлагает сбалансированный компромисс между выразительностью LSTM и простотой стандартных RNN, достигая конкурентоспособной производительности при сниженной вычислительной сложности в таких задачах как анализ тональности и моделирование последовательностей.

3.5. Bi-RNN, Bi-LSTM, Bi-GRU. Двухнаправленные рекуррентные нейронные сети (Bi-RNN), двухнаправленные долгосрочные нейронные сети с короткой па-

мятью (Bi-LSTM) и двунаправленные блокирующие рекуррентные единицы (Bi-GRU) являются передовыми архитектурами машинного обучения, которые расширяют возможности рекуррентных нейронных сетей (RNN) путем обработки последовательных данных в направлениях вперед и назад одновременно. Захватывая зависимости от предыдущих и будущих входов одновременно, данные модели достигают превосходной производительности в задачах обработки естественного языка, распознавания речи и анализа временных рядов.

Bi-RNN расширяют стандартные RNN с отдельным обратным слоем, который обрабатывает последовательности ввода в обратном порядке. Скрытые состояния прямого и обратного слоев объединяются на каждом временном шаге, обеспечивая комплексное представление последовательности ввода. Bi-LSTM и Bi-GRU — это варианты Bi-RNN, включающие специализированные ячейки памяти для решения проблемы затухающего градиента, часто возникающей при использовании глубоких RNN. LSTM использует механизмы управления, включая входные, забывающие и выходные вентили, для регулирования потока информации между временными шагами, тогда как GRU упрощает этот процесс с помощью вентиля обновления и сброса. В результате Bi-LSTM и Bi-GRU предлагают улучшенную способность к обучению долгосрочных зависимостей и сложных шаблонов в последовательностях, что в конечном итоге приводит к более точным и надежным моделям для различных задач машинного обучения.

3.6. XGBoost, CatBoost. XGBoost — оптимизированный алгоритм градиентного бустинга, который использует несколько классификаторов на основе деревьев решений последовательно для минимизации функций потерь. Он обладает отличной предсказательной способностью, масштабируемостью и возможностями параллелизации, что делает его подходящим для разных задач, включая классификацию и регрессию. CatBoost — еще один алгоритм градиентного бустинга, который обрабатывает категориальные переменные с помощью новаторского метода кодирования, основанного на статистике целевой переменной, это позволяет избежать переобучения и повысить производительность модели [11]. CatBoost имеет надежные значения гиперпараметров по умолчанию и демонстрирует исключительную эффективность в задачах с высокоразмерными категориальными признаками [12].

4. Результаты моделирования. При оценке моделей машинного обучения крайне важно учитывать вычислительную мощность используемого компьютера. Был выбран набор данных, содержащий 11 300 образцов, которые были разделены на три отдельных набора: обучающий, состоящий из 8000 образцов, тестовый с 1650 образцами и набор для проверки, также содержащий 1650 образцов. Такое разделение позволяет осуществить эффективное обучение, тестирование и оценку модели для обеспечения ее хорошей обобщающей способности. Для представления данных точно и при этом сохраняя вычислительную эффективность, 12 векторов признаков были тщательно выбраны на основе предметных знаний и предварительного анализа.

Кривая обучения всех рассматриваемых моделей показывает уменьшающуюся тенденцию, что указывает на улучшение производительности по мере обработки большего количества обучающих образцов. Это положительный признак того, что модели обучаются на данных и хорошо адаптируются к задаче.

Для обеспечения всестороннего сравнения различных моделей необходимо провести детальный анализ количественных метрик, который позволит исследователям выявить сильные и слабые стороны каждой модели в отношении конкретной проблемы. Более того, полезно оценить производительность моделей с учетом компромисса

между временем вычислений и предсказательной способностью с учетом доступных вычислительных ресурсов. Это поможет определить наиболее подходящую модель для развертывания, сбалансировав производительность и эффективность.

Тщательное изучение производительности моделей, измеряемой преимущественно по дисперсии, показывает, что они все обладают замечательной эффективностью (рис. 2 и таблица). Среди них сети ANN демонстрируют наиболее выдающуюся производительность. Тем не менее важно учитывать и другие факторы, например вычислительную эффективность, при выборе подходящей модели прогнозирования для решаемой задачи. В этом контексте XGBoost представляется весьма привлекательным выбором, благодаря значительно более низкому времени выполнения по сравнению с другими моделями. Такое снижение вычислительной нагрузки XGBoost приводит к достижению сравнимых уровней точности с другими моделями, но за гораздо меньшее время. В результате комплексного сравнения, учитывающего как точность предсказаний, так и скорость вычислений, XGBoost можно считать общей наилучшей моделью прогнозирования.

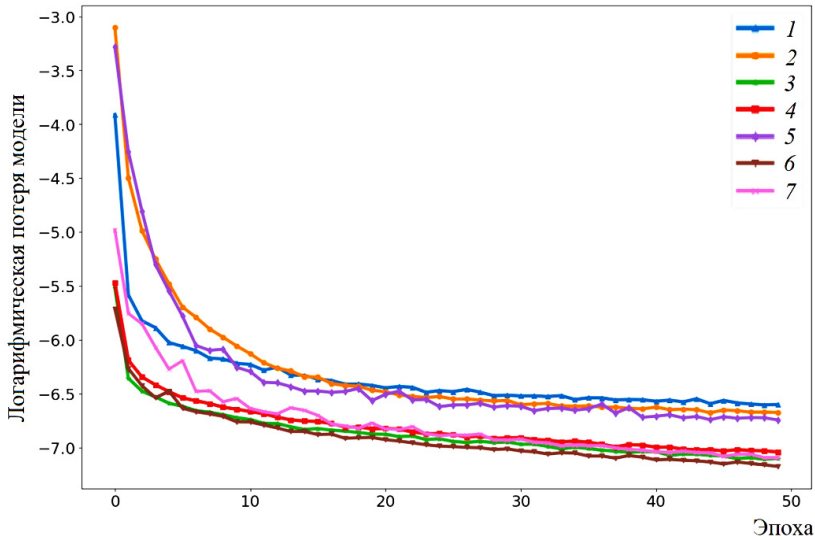


Рис. 2. Кривые обучения прогностических моделей
 1 — ANN; 2 — RNN; 3 — LSTM; 4 — GRU; 5 — Bi-RNN; 6 — Bi-LSTM; 7 — Bi-GRU.

Таблица. Качество прогноза текущего PM2.5

Модель	R^2	MSE	Затраченное время, с
ANN	0.8535	6.7295	15
RNN	0.8226	6.6638	23.5
LSTM	0.8094	6.1287	41
GRU	0.7833	7.2667	38
Bi-RNN	0.7668	8.2554	20.8
Bi-LSTM	0.8227	6.9529	42.4
Bi-GRU	0.8481	5.9622	39.6
LightGBM	0.8239	6.9952	4.05
XgBoost	0.8268	6.9714	3.64
CatBoost	0.8351	6.0393	9.52

Выдающаяся производительность XGBoost объясняется его основным принципом — интегрированными алгоритмами обучения. Несмотря на кажущуюся простоту, эти алгоритмы часто дают впечатляющие результаты на практике. Комбинируя несколько слабых учеников в одну сильную модель, интегрированные алгоритмы обучения, такие как XGBoost, создают устойчивую и эффективную систему прогнозирования. Это позволяет модели адаптивно обучаться на данных, улучшая ее предсказательную способность и способность к обобщению (рис. 3).

В заключение, учитывая как точность предсказаний, так и вычислительную эффективность, XGBoost выделяется среди других альтернативных моделей в качестве лучшей модели прогнозирования. Его основание на интегрированных алгоритмах обучения не только обеспечивает высокую производительность на практике, но и подчеркивает значимость таких, казалось бы, простых концепций для достижения впечатляющих результатов.

5. Объяснение модели.

5.1. Объяснение искусственного интеллекта. Искусственный интеллект с объяснениями направлен на повышение прозрачности традиционно непрозрачных «черных ящиков» моделей путем использования различных методологий. Среди распространенных подходов можно выделить вычисление вклада признаков, где количественно оценивается значимость каждого признака в модели. Более высокий вклад указывает на большее влияние на конечный результат прогнозирования. Этот метод также позволяет представить взаимосвязи между признаками. Предоставление интуитивно понятных и логически последовательных вкладов признаков вместе с прогнозами может значительно улучшить доверие пользователей к «черным ящикам» моделей. Важно подчеркнуть, что основная цель искусственного интеллекта с объяснениями не заключается в раскрытии всех деталей и процессов внутри «черного ящика» модели. Он сконцентрирован на представлении ключевых аспектов процесса принятия решений модели таким образом, чтобы это было понятно и значимо для пользователя. Такой подход способствует созданию среды, где системы искусственного интеллекта могут быть более легко приняты, поняты и эффективно использованы в различных приложениях.

Искусственный интеллект с объяснениями может быть классифицирован на основе объема объяснения и основополагающих принципов. С точки зрения объема объяснения могут быть локальными или глобальными. Локальные объяснения фокусируются на отдельных случаях, в то время как глобальные охватывают всю выборку данных. Например, с одной стороны, в задаче прогнозирования временных рядов локальное объяснение предоставляет информацию о вкладе признаков на каждом временном шаге, что дает возможность реальному времени контролировать данные. С другой стороны, глобальное объяснение оценивает среднее влияние каждого признака за определенный период, что дает возможность осуществить всеобъемлющую оценку «черного ящика» модели, которая может служить вспомогательной технологией для сравнения производительности моделей. С точки зрения принципов объяснения методы искусственного интеллекта с объяснениями делятся на встроенные и пост-хок. Встроенные методы относятся к врожденной объяснимости модели, таким как «белые ящики», например линейная регрессия и деревья решений. Для «черных ящиков» встроенные подходы стремятся интегрировать возможности объяснения непосредственно в модель, что приводит к уникальным объяснениям, специфичным для этой модели или типа моделей. Пост-хок методы фокусируются на разработке модельно-агностических объяснений, применимых к любому «черному ящику» моде-

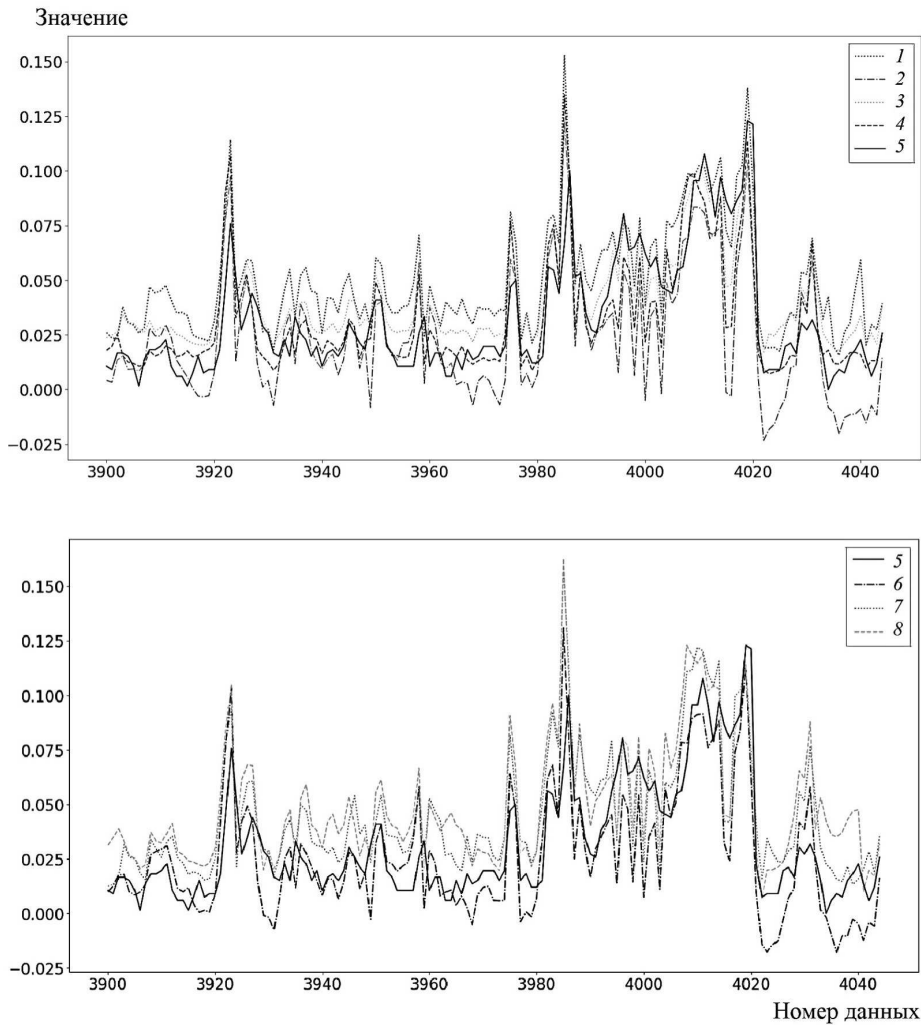


Рис. 3. Оригинальные данные против прогнозов разных моделей: 10-модельный сравнительный анализ

1 — ANN; 2 — RNN; 3 — LSTM; 4 — GRU; 5 — тест; 6 — Bi-RNN; 7 — Bi-LSTM; 8 — Bi-GRU.

ли. Обычно они включают изменение каждого признака в наборе данных и измерение соответствующих изменений в прогнозных результатах для вычисления вклада признаков, не принимая во внимание основную модель.

С учетом цели объяснения существующих прогностических моделей в качестве объяснительной основы использовался метод SHAP. Этот выбор обусловлен устойчивым теоретическим фундаментом, предоставляемым теорией кооперативных игр, а также наличием комплексных инструментов программирования, которые облегчают практическую реализацию.

5.2. Интерпретация результатов с учетом объяснений на основе SHAP. SHAP — это метод объяснения, разработанный на основе значений Шепли [13]. Он использует теорию кооперативных игр в качестве теоретической основы,

рассматривает «черный ящик» модели как «игру» и каждый признак как «игрока». Путем вычисления значения Шепли можно определить степень вклада каждого игрока (признака) в процесс игры (работу «черного ящика» модели), а затем узнать степень его вклада [14].

Можно создать информативную столбчатую диаграмму, которая дает возможность продемонстрировать общую важность признаков, конкретно отображая средний вклад каждого признака в результаты прогнозирования модели. Полученные графики-столбцы предоставляют краткий обзор важности признаков и помогают понять, как модель делает прогнозы на основе признаков, и, следовательно, руководят решениями, связанными с выбором и инженерией признаков. В конечном итоге эта техника применяется к отдельным значениям во всех наблюдениях. Определяются абсолютные значения этих вкладов, затем суммируются и усредняются, чтобы получить полное представление о ранжировке переменных (рис. 4). В целом признак PM10 демонстрирует доминирующую ранжировку вклада, за ним следуют CO и DEWP. Эти ранжировки указывают на ведущие признаки, влияющие на результаты прогнозирования модели.

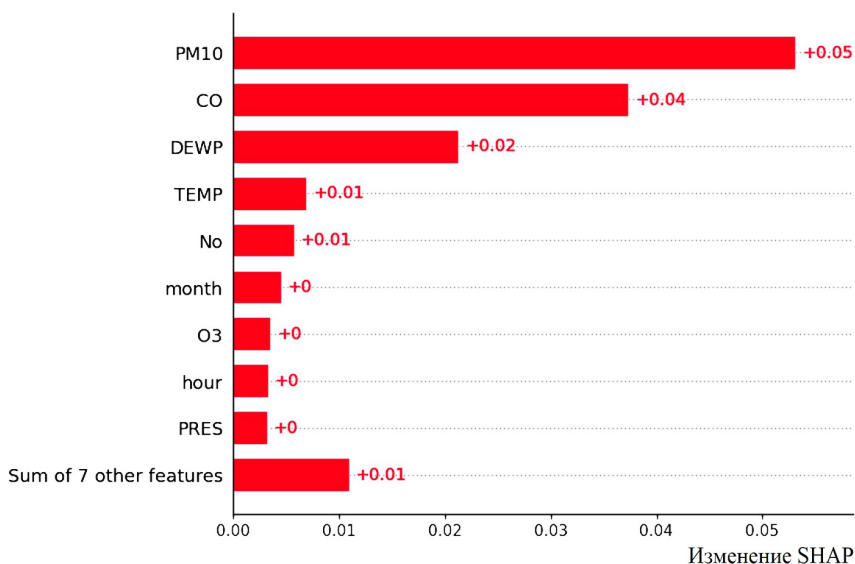


Рис. 4. Пример результатов глобального объяснения

Кроме того, возможно провести детальное исследование влияющих факторов, анализируя отдельные точки наблюдений. Это можно сделать путем создания водопадной диаграммы, которая ясно показывает влияние модели на прогнозируемые результаты для конкретной точки данных. Водопадная диаграмма визуально представляет величину вклада каждого признака в прогнозы модели.

Для иллюстрации этой методологии на рис. 5 показан пример наблюдения (2 октября 2016 г., 6:00), демонстрирующего локализованное рассмотрение переменных, от которых зависит прогностическая модель, в отличие от общего подхода, представленного на рис. 5. В данном случае прогностическая модель выдала постоянный вывод в размере 0.119 в 6:00. Отмечается, что PM10 оказала отрицательное воздействие в размере -0.05 на прогноз, CO — отрицательное воздействие в размере -0.04 и дру-

гие переменные следовали подобной тенденции. В результате окончательное предсказанное значение составило 0.005. Этот анализ подчеркивает ценность изучения отдельных точек данных, обогащая понимание функционирования модели и информируя последующие решения или корректировки в отношении конкретных значений признаков.

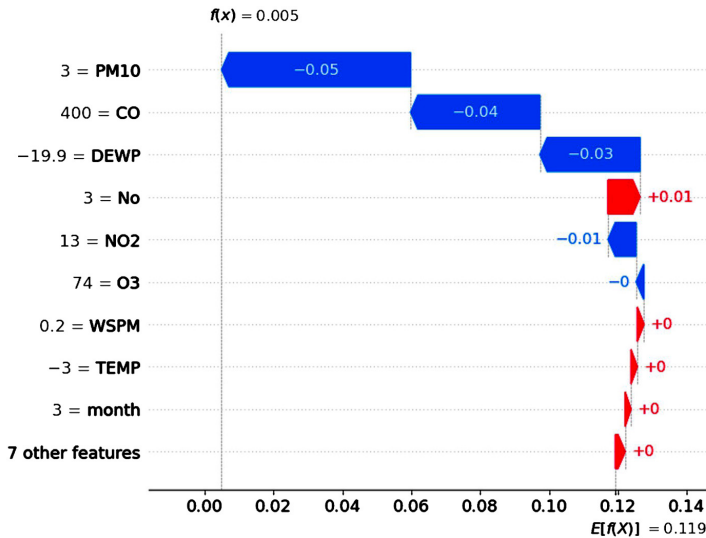


Рис. 5. Пример результатов локального объяснения

6. Заключение. Основной вклад настоящей работы выходит за рамки простого сравнения моделей прогнозирования временных рядов на различных данных. Мы также стремимся предложить понятные на уровне интерпретации идеи для повышения прозрачности «черных ящиков» прогностических моделей и помощи пользователям в определении достоверности изучаемых моделей признаков. На основе традиционных показателей производительности был сделан вывод, что XGBoost в настоящее время является превосходной прогностической моделью для задач прогнозирования временных рядов энергии [15, 16].

Результаты интерпретации, полученные с использованием SHAP, позволяют оценить вклад отдельных признаков при использовании XGBoost в конкретных контекстах. Это понимание помогает выделить важные признаки в определенных обстоятельствах.

В настоящем исследовании мы сосредоточились на глобальной интерпретации с целью анализа вклада признаков на протяжении продолжительного периода времени. В дальнейшем планируем уделить внимание локальной интерпретации для задач прогнозирования временных рядов, где результаты интерпретации генерируются для каждого момента времени. Такой подход будет способствовать реальному времени интерпретации и, возможно, выявлению проблем, таких как концептуальные изменения, которые могут негативно влиять на прогнозы временных рядов.

В работе не только сравниваются различные модели прогнозирования временных рядов, но и предоставлены понятные на уровне интерпретации идеи об их внутреннем устройстве. Используя объяснения на основе SHAP, можно установить вклад отдельных признаков, это поможет настраивать модели для конкретных приложений.

В будущем планируем исследовать реальное время и локальные интерпретации, чтобы справиться с потенциальными вызовами в задачах прогнозирования временных рядов.

Литература

1. Zhan D., Kwan Mei-Po, Zhang W., Yu X., Meng B., Liu Q. The driving factors of air quality index in China // Journal of Cleaner Production. 2018. Vol. 197. P. 1342–1351.
2. Saad M. Sh., Melvin A., Md Sh., Yeon A., Saad M., Rahman A., Kamarudin Yu. M. Classifying sources influencing indoor air quality (IAQ) using artificial neural network (ANN) // Sensors. 2015. Vol. 15. N 5. P. 11665–11684.
3. Navares R., Aznarte J. L. Predicting air quality with deep learning LSTM: Towards comprehensive models // Ecological Informatics. 2020. Vol. 55. P. 101019.
4. Athira V., Geetha P., Vinayakumar R., Soman K. P. Deepairnet: Applying recurrent networks for air quality prediction // Procedia Computer Science. 2018. Vol. 132. P. 1394–1403.
5. Di Q., Amini H., Shi L., Kloog I. An ensemble-based model of PM_{2.5} concentration across the contiguous United States with high spatiotemporal resolution // Environment International. 2019. Vol. 130. Iss. 0160–4120. P. 104909. <https://doi.org/10.1016/j.envint.2019.104909>
6. Fann N., Risley D. The public health context for PM_{2.5} and ozone air quality trends // Air Qual Atmos Health. 2013. Vol. 6. P. 1–111. <https://doi.org/10.1007/s11869-010-0125-0>
7. Wang K., Yin H., Chen Y. The effect of environmental regulation on air quality: A study of new ambient air quality standards in China // Journal of Cleaner Production. 2019. Vol. 215. Iss. 0959–6526. P. 268–279. <https://doi.org/10.1016/j.jclepro.2019.01.061>
8. Palani S., Liong Shie-Yui, Tkalich P. An ANN application for water quality forecasting // Marine Pollution Bulletin. 2008. Vol. 56. N 9. Iss. 0025–326X. P. 1586–1597. <https://doi.org/10.1016/j.marpolbul.2008.05.021>
9. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network // Physica D: Nonlinear Phenomena. 2020. Vol. 404. Iss. 0167–2789. P. 132306. <https://doi.org/10.1016/j.physd.2019.132306>
10. Dey R. S., Fathi M. Gate-variants of Gated Recurrent Unit (GRU) neural networks // 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). 2017. P. 1597–1600. <https://doi.org/10.1109/MWSCAS.2017.8053243>
11. Chen T., Guestrin C. Xgboost: A scalable tree boosting system // Proceedings of the 22nd International Conference on knowledge discovery and data mining. 2016. P. 785–794.
12. Essam Al. D. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset // International Journal of Computer and Information Engineering. 2019. Vol. 13. N 1. P. 6–10.
13. Meng Y., Yang N., Qian Z., Zhang G. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values // Journal of Theoretical and Applied Electronic Commerce Research. 2020. Vol. 16. N 3. P. 466–490.
14. Merrick L. The explanation game: Explaining machine learning models using shapley values // Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference. CD-MAKE 2020. Dublin, Ireland. August 25–28. 2020. Vol. 4. P. 17–38.
15. Якушев В. П., Буре В. М., Митрофанова О. А., Митрофанов Е. П. Теоретические основы вероятностно-статистического прогнозирования неблагоприятных агрометеоусловий // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2021. Т. 17. Вып. 2. С. 174–182. <https://doi.org/10.21638/11701/spbu10.2021.207>
16. Егоров Н. В., Виноградова Е. М., Доронин Г. Г. Математическое моделирование полевого катода лезвийной формы с диэлектрическим покрытием // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 1. С. 65–71. <https://doi.org/10.21638/11701/spbu10.2023.106>

Статья поступила в редакцию 9 января 2024 г.

Статья принята к печати 12 марта 2024 г.

Контактная информация:

Ци Дуифан — аспирант; st073409@student.spbu.ru

Буре Владимир Мансурович — д-р техн. наук, проф.; vlb310154@gmail.com

Explanatory comparative analysis of time series forecasting algorithms for air quality prediction

D. Qi, V. M. Bure

St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Qi D., Bure V. M. Explanatory comparative analysis of time series forecasting algorithms for air quality prediction. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2024, vol. 20, iss. 2, pp. 206–219. <https://doi.org/10.21638/spbu10.2024.206> (In Russian)

This study explores the effectiveness of time series forecasting models for predicting air quality using datasets from a Purple Air Dual Laser Air Quality Sensor and the Kaggle Online platform. These datasets contain reliable and real sensor records, ensuring the richness of information required for environmental protection. The research focuses on identifying suitable forecast models for environmental analysis, including popular algorithm structures such as neural network models and ensemble models. Moreover, the study introduces the Explainable artificial intellect method to provide explanations for models with excellent performance indicators, thereby enhancing their trust and transparency. The performance of the models was evaluated using metrics such as mean absolute error, root mean square error, and coefficient of determination (R -squared). Results indicate that the neural network and ensemble models are effective in forecasting air quality time series. The study contributes to the body of knowledge on time series forecasting models and provides insights for future research in air quality prediction.

Keywords: air quality, time series forecasting, neural networks, ensemble models, explainable artificial intellect.

References

1. Zhan D., Kwan Mei-Po, Zhang W., Yu X., Meng B., Liu Q. The driving factors of air quality index in China. *Journal of Cleaner Production*, 2018, vol. 197, pp. 1342–1351.
2. Saad M. Sh., Melvin A., Md Sh., Yeon A., Saad M., Rahman A., Kamarudin Yu. M. Classifying sources influencing indoor air quality (IAQ) using artificial neural network (ANN). *Sensors*, 2015, vol. 15, no. 5, pp. 11665–11684.
3. Navares R., Aznarte J. L. Predicting air quality with deep learning LSTM: Towards comprehensive models. *Ecological Informatics*, 2020, vol. 55, p. 101019.
4. Athira V., Geetha P., Vinayakumar R., Soman K. P. Deepairnet: Applying recurrent networks for air quality prediction. *Procedia Computer Science*, 2018, vol. 132, pp. 1394–1403.
5. Di Q., Amini H., Shi L., Kloog I. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, 2019, vol. 130, iss. 0160–4120, pp. 104909. <https://doi.org/10.1016/j.envint.2019.104909>
6. Fann N., Riskey D. The public health context for PM2.5 and ozone air quality trends. *Air Qual Atmos Health*, 2013, vol. 6, pp. 1–111. <https://doi.org/10.1007/s11869-010-0125-0>
7. Wang K., Yin H., Chen Y. The effect of environmental regulation on air quality: A study of new ambient air quality standards in China. *Journal of Cleaner Production*, 2019, vol. 215, iss. 0959–6526, pp. 268–279. <https://doi.org/10.1016/j.jclepro.2019.01.061>
8. Palani S., Liong Shie-Yui, Tklich P. An ANN application for water quality forecasting. *Marine Pollution Bulletin*, 2008, vol. 56, no. 9, iss. 0025–326X, pp. 1586–1597. <https://doi.org/10.1016/j.marpolbul.2008.05.021>
9. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 2020, vol. 404, iss. 0167–2789, pp. 132306. <https://doi.org/10.1016/j.physd.2019.132306>
10. Dey R. S., Fathi M. Gate-variants of Gated Recurrent Unit (GRU) neural networks. *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600. <https://doi.org/10.1109/MWSCAS.2017.8053243>

11. Chen T., Guestrin C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd International Conference on knowledge discovery and data mining*, 2016, pp. 785–794.
12. Essam Al. D. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. *International Journal of Computer and Information Engineering*, 2019, vol. 13, no. 1, pp. 6–10.
13. Meng Y., Yang N., Qian Z., Zhang G. What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values. *Journal of Theoretical and Applied Electronic Commerce Research*, 2020, vol. 16, no. 3, pp. 466–490.
14. Merrick L. The explanation game: Explaining machine learning models using shapley values. *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020*. Dublin, Ireland, August 25–28, 2020, vol. 4, pp. 17–38.
15. Iakushev V. P., Bure V. M., Mitrofanova O. A., Mitrofanov E. P. Teoreticheskie osnovy veroiatnostno-statisticheskogo prognozirovaniia neblagopriiatnykh agrometeoslovii [Theoretical foundations of probabilistic and statistical forecasting of agrometeorological risks]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2021, vol. 17, iss. 2, pp. 174–182. <https://doi.org/10.21638/11701/spbu10.2021.207> (In Russian)
16. Egorov N. V., Vinogradova E. M., Doronin G. G. Matematicheskoe modelirovanie polevogo katoda lezviinoi formy s dielektricheskim pokrytiem [Blade-like field cathode with a dielectric coating mathematical modeling]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2023, vol. 19, iss. 1, pp. 65–71. <https://doi.org/10.21638/11701/spbu10.2023.106> (In Russian)

Received: January 9, 2024.

Accepted: March 12, 2024.

A u t h o r s' i n f o r m a t i o n :

Dongfang Qi — Postgraduate Student; st073409@student.spbu.ru

Vladimir M. Bure — Dr. Sci. in Technics, Professor; vlb310154@gmail.com