УДК №

Богатов Д.В.,

Студент, II курс, МГТУ им. Н.Э. Баумана, Информатика и системы управления, Москва

Научный руководитель: Березкин Д.В.,

Доцент каф.ИУ6 МГТУ им. Н.Э. Баумана,, Москва

ИНТЕРПРЕТИРУЕМОСТЬ И ОБЪЯСНИМОСТЬ В АНАЛИЗЕ ВРЕМЕННЫХ РЯДОВ

Статья посвящена анализу проблемы интерпретируемости моделей машинного обучения в задачах обработки временных рядов. Рассматриваются ключевые понятия интерпретируемости и объяснимости, их различия и особенности применения в условиях временной зависимости данных. Особое внимание уделяется специфике интерпретации временных рядов, включая нестационарность, многослойную природу признаков и временную динамику их влияния. Проведен обзор подходов к интерпретации, включая модельно-специфичные и модельно-независимые методы, а также глобальные и локальные стратегии объяснения. Рассматриваются ante-hoc и post-hoc подходы, их преимущества и ограничения в контексте временных данных. Отдельно анализируются методы оценки важности признаков, включая временную, динамическую и причинную важность, а также роль визуализации в интерпретации результатов моделей.

КЛЮЧЕВЫЕ СЛОВА: временные ряды, интерпретируемость, объяснимость, машинное обучение, LIME, SHAP, важность признаков, визуализация, ante-hoc, post-hoc, причинный анализ

Введение. Современные методы машинного обучения демонстрируют впечатляющие результаты в задачах анализа временных рядов, включая прогнозирование будущих значений и обнаружение аномальных паттернов. Однако по мере усложнения моделей и повышения их точности возникает

критическая проблема - снижение интерпретируемости получаемых результатов. Эта проблема особенно остра в областях, где принятие решений на основе модельных предсказаний требует понимания причинно-следственных связей и возможности объяснить логику работы алгоритма.

Актуальность темы интерпретируемости в анализе временных рядов обусловлена несколькими факторами. Во-первых, регуляторные требования в таких областях, как финансы и медицина, часто предписывают необходимость объяснения алгоритмических решений. Европейский регламент GDPR, например, закрепляет право граждан на получение объяснений автоматизированных решений, что делает интерпретируемость не только желательной, но и юридически обязательной характеристикой систем. Во-вторых, доверие пользователей к системам искусственного интеллекта напрямую зависит от их способности понимать логику принятия решений. В-третьих, интерпретируемость критически важна для выявления и устранения систематических ошибок, смещений и артефактов в работе моделей.

Понятие интерпретируемости и объяснимости. Интерпретируемость модели машинного обучения определяется как степень, в которой человек может понять причины принятия конкретного решения. Это понятие тесно связано, но не идентично объяснимости, которая относится к способности предоставить понятное человеку описание внутренней логики модели. В литературе эти термины часто используются взаимозаменяемо, хотя некоторые исследователи проводят между ними различие: интерпретируемость рассматривается как внутреннее свойство модели, в то время как объяснимость - как внешнее представление работы модели.

Для временных рядов интерпретируемость приобретает дополнительные измерения. Помимо понимания, какие признаки важны для предсказания, необходимо понимать временную динамику их влияния. Например, в задаче прогнозирования спроса важно знать не только то, что цена товара влияет на спрос, но и как это влияние распределено во времени, есть ли запаздывание эффекта, как долго сохраняется влияние изменения цены.

Особенности интерпретации для временных рядов. Временные ряды обладают рядом характеристик, которые делают задачу интерпретации особенно сложной. Во-первых, наличие временных зависимостей означает, что важность признака может варьироваться в зависимости от временного контекста. Признак, критически важный для краткосрочного прогноза, может быть несущественным для долгосрочного.

Во-вторых, концепция "признака" во временных рядах сама по себе неоднозначна. В классическом машинном обучении признаки обычно представляют собой отдельные измеримые характеристики объекта. Во временных рядах признаками могут быть как исходные временные точки, так и производные характеристики: скользящие средние, разности, автокорреляционные функции. Интерпретация должна учитывать эту многоуровневую природу признакового пространства.

В-третьих, нестационарность многих реальных временных рядов означает, что паттерны и зависимости могут изменяться со временем. Метод интерпретации должен быть способен выявлять и объяснять эти изменения. Например, в финансовых временных рядах корреляции между активами могут усиливаться в периоды кризисов — факт, который должен быть отражен в интерпретации модельных предсказаний.

Специфичные для модели и общие методы. Фундаментальное разделение методов интерпретации проходит между подходами, специфичными для конкретных типов моделей, и универсальными методами, применимыми к любым моделям.

Специфичные методы разработаны с учетом особенностей архитектуры конкретных моделей. Для линейных моделей это коэффициенты регрессии, для деревьев решений — правила разбиения, для нейронных сетей — веса связей и активации нейронов. В контексте временных рядов примером может служить интерпретация коэффициентов ARIMA модели, где каждый коэффициент имеет четкую временную интерпретацию (влияние предыдущих наблюдений или ошибок прогноза).

Преимущество специфичных для модели методов заключается в их способности использовать внутреннюю структуру модели для более точной и детальной интерпретации. Например, в рекуррентных нейронных сетях можно анализировать динамику скрытых состояний, чтобы понять, какая информация сохраняется и передается между временными шагами. Механизмы внимания в трансформерах предоставляют естественный способ визуализации того, на какие временные точки модель "обращает внимание" при формировании предсказания.

Общие методы работают с моделью как с черным ящиком, анализируя только соотношение между входами и выходами. Ключевые представители этого класса — LIME (Local Interpretable Model-agnostic Explanations) и SHAP (SHapley Additive exPlanations). Эти методы строят локальные аппроксимации сложной модели с помощью интерпретируемых суррогатных моделей или используют теоретико-игровые концепции для атрибуции важности признаков.

Для временных рядов общие методы требуют адаптации. Например, при применении LIME необходимо определить, что означает "локальность" во временном контексте — это может быть окрестность конкретной временной точки или схожесть временных паттернов. SHAP для временных рядов должен учитывать временные зависимости при расчете вкладов признаков, что приводит к модификациям типа TimeSHAP.

Глобальная и локальная интерпретация. Глобальная интерпретация направлена на понимание общей логики работы модели вдоль всего пространства признаков и временного диапазона. Для временных рядов это может включать выявление общих временных паттернов, которые модель использует для предсказаний, идентификацию доминирующих частотных компонент или определение характерных временных масштабов влияния различных факторов.

Методы глобальной интерпретации включают анализ важности признаков построение частичных зависимостей, и извлечение правил. В

контексте временных рядов особый интерес представляют методы декомпозиции, которые разделяют предсказания модели на интерпретируемые компоненты: тренд, сезонность, влияние внешних факторов.

Локальная интерпретация фокусируется на объяснении отдельных предсказаний или поведения модели в конкретных временных точках. Это особенно важно для понимания аномалий или неожиданных предсказаний. Локальные методы должны ответить на вопросы: почему модель сделала именно это предсказание в данный момент времени? Какие исторические наблюдения наиболее повлияли на это решение?

Для временных рядов граница между глобальной и локальной интерпретацией может быть размытой. Например, объяснение предсказания на конкретный день может потребовать понимания недельного цикла (локально-глобальный аспект) или долгосрочного тренда (глобальный контекст для локального объяснения).

Post-hoc и Ante-hoc подходы. Ante-hoc подходы предполагают построение изначально интерпретируемых моделей. Дизайн модели с самого начала учитывает требования интерпретируемости, встраивая их в архитектуру и процесс обучения. Примеры включают:

- Аддитивные модели (GAM - Generalized Additive Models), где общее предсказание является суммой интерпретируемых функций отдельных признаков
- Модели с явными ограничениями монотонности или гладкости
- Архитектуры нейронных сетей с интерпретируемыми модулями (например, отдельные подсети для тренда и сезонности)

В области временных рядов ante-hoc подход может выражаться в использовании моделей состояния пространства, где каждое состояние имеет физическую или бизнес-интерпретацию. Например, в модели для

прогнозирования продаж состояния могут соответствовать уровню базового спроса, эффекту маркетинговых кампаний, сезонным факторам.

Post-hoc подходы применяются к уже обученным моделям для объяснения их поведения. Эти методы не влияют на процесс обучения и могут применяться к любым, даже самым сложным моделям. Ключевые категории post-hoc методов:

- Методы возмущения: анализируют изменение выхода модели при изменении входов
- Градиентные методы: используют градиенты для оценки чувствительности предсказаний к входным признакам
- Методы на основе примеров: объясняют предсказания через схожие исторические случаи

Для временных рядов post-hoc методы сталкиваются со сложностями сохранения временной согласованности при возмущениях. Случайное изменение отдельных временных точек может создать нереалистичные временные ряды, что приведет к неверным интерпретациям. Поэтому разрабатываются специализированные методы возмущения, которые сохраняют временную структуру данных.

Визуальные методы интерпретации. Визуализация играет критическую роль в интерпретации моделей временных рядов, используя естественную способность человека воспринимать временные паттерны визуально. Эффективные визуальные методы должны одновременно представлять данные, предсказания модели и объяснения в понятной и не перегруженной форме.

Статические визуализации включают:

- Графики важности признаков во времени, показывающие как изменяется влияние различных факторов
- Тепловые карты внимания для моделей с механизмами внимания

- Декомпозиционные графики, разделяющие предсказание на составляющие компоненты
- Контрфактические визуализации, показывающие как изменилось бы предсказание при других условиях

Интерактивные визуализации позволяют пользователю исследовать модель более глубоко:

- Что-если анализ с возможностью изменения входных данных и наблюдения эффекта
- Детализированная функциональность для перехода от глобального к локальному уровню интерпретации
    - Временная навигация для изучения как модель работает в разные периоды времени

Особое внимание в визуальной интерпретации временных рядов уделяется представлению неопределенности. Предсказания модели часто сопровождаются доверительными интервалами, и визуализация должна эффективно коммуницировать эту неопределенность вместе с объяснениями.

Анализ важности признаков временных рядов. В контексте многомерных временных рядов критически важно понимать, какие переменные и в какой степени влияют на прогноз. Классические методы важности признаков требуют адаптации для временных рядов.

Временная важность признаков расширяет концепцию важности, учитывая временное измерение. Вместо единственного значения важности для каждого признака, эти методы предоставляют временной профиль важности, показывающий как влияние признака меняется в зависимости от горизонта прогнозирования. Например, для прогнозирования спроса текущая цена может быть критична для краткосрочного прогноза, но маркетинговые расходы прошлого месяца - для среднесрочного.

Динамическая важность признаков фокусируется на том, как важность меняется во времени в зависимости от контекста. В периоды экономической стабильности макроэкономические индикаторы могут иметь низкую важность для финансовых прогнозов, но становиться критическими во время кризисов. Методы типа LIME-TS или временное SHAP вычисляют локальную важность признаков для каждого временного окна.

Причинная важность признаков идет дальше корреляционного анализа, пытаясь выявить причинно-следственные связи. Методы на основе причинности по Грейджеру или более современные подходы с использованием причинного анализа алгоритмов помогают отличить истинные предикторы от ложных корреляций. Это особенно важно для временных рядов, где временные зависимости могут создавать иллюзию причинности.

Заключение. Таким образом, интерпретируемость моделей машинного обучения является ключевым аспектом анализа временных рядов, обеспечивающим прозрачность, надежность и практическую применимость получаемых результатов. Рассмотренные подходы - от модельно-специфичных до универсальных, а также ante-hoc и post-hoc методы - позволяют по-разному раскрывать внутреннюю логику моделей и учитывать временную природу данных. Их комбинированное использование, вместе с анализом важности признаков и визуализацией, способствует более глубокому пониманию моделей и повышает доверие к их прогнозам в реальных прикладных задачах.

ЛИТЕРАТУРА:

1. Adadi A., Berrada M. *Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)* // IEEE Access. 2018. Vol. 6. P. 52138–52160.
2. Assaf R., Schumann A. *Explainable deep neural networks for multivariate time series predictions* // Proceedings of IJCAI. 2019. P. 6488–6490.

3. Barredo Arrieta A. et al. *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI* // Information Fusion. 2020. Vol. 58. P. 82–115.
4. Bento J., Saleiro P., Cruz A.F., Figueiredo M.A., Bizarro P. *TimeSHAP: Explaining recurrent models through sequence perturbations* // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. P. 2565–2573.
5. Braei M., Wagner S. Anomaly detection in univariate time-series: *A survey on the state-of-the-art* // arXiv preprint arXiv:2004.00433. 2020.
6. Cleveland R.B., Cleveland W.S., McRae J.E., Terpenning I. *STL: A Seasonal-Trend Decomposition Procedure Based on Loess* // Journal of Official Statistics. 1990. Vol. 6, №1. P. 3–73.
7. Hyndman R.J., Athanasopoulos G. *Forecasting: Principles and Practice (2nd ed.)* // OTexts, 2018.
8. Lundberg S.M., Lee S.I. *A unified approach to interpreting model predictions* // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774.
9. Miller T. *Explanation in artificial intelligence: Insights from the social sciences* // Artificial Intelligence. 2019. Vol. 267. P. 1–38.