Миронов Артем Юрьевич, магистрант
Российский новый университет
Mironov Artem Yuryevich**,** Master's student
Russian New University

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ ЛОКАЛЬНЫХ ОБЪЯСНЕНИЙ ПРОГНОЗОВ СПРОСА ПРИ ИЗМЕНЕНИИ БАЗОВЫХ ЗНАЧЕНИЙ ПРИЗНАКОВ
STABILITY ANALYSIS OF LOCAL DEMAND FORECAST EXPLANATIONS UNDER CHANGES IN FEATURE BASELINES

Аннотация. Исследуется устойчивость локальных объяснений прогнозов спроса при изменении базовых значений признаков. Для линейной регрессии и случайного леса сопоставлены три варианта базы: средние значения обучающей выборки, медианы и средние значения последних 30 обучающих дней. Оценены устойчивость ранжирования факторов, совпадение трех наиболее значимых признаков и диагностический остаток. Дополнительно выполнены групповые замены связанных лаговых, скользящих и сезонных признаков. Показано, что точность аддитивного разложения линейной модели сохраняется при любой базе, однако состав ведущих факторов может изменяться. Для случайного леса устойчивость зависит от выбора базы и взаимодействий между связанными признаками.

Abstract. The paper studies the stability of local demand forecast explanations under changes in feature baselines. Three baselines are compared for linear regression and random forest models: training means, training medians, and the mean values of the last 30 training days. Feature-ranking stability, exact top-three overlap, and diagnostic residuals are evaluated. Group replacements of related lag, rolling, and seasonal features are also analyzed. The additive decomposition of the linear model remains numerically exact for every baseline, while the ranking of influential factors may change. For the random forest, explanation stability depends on the selected baseline and on interactions among related features.

Ключевые слова: Объяснимый искусственный интеллект, устойчивость объяснений, прогнозирование спроса, базовое значение, интерпретируемость, временные ряды.

Keywords: Explainable artificial intelligence, explanation stability, demand forecasting, baseline, interpretability, time series.

Введение

Локальное объяснение должно помогать пользователю понять конкретный прогноз, а не только показывать общую важность признаков. На практике результат объяснения зависит не только от обученной модели, но и от того, каким значением заменяется отсутствующий или нейтрализуемый признак. Для методов, использующих фоновые данные или базовую точку, этот выбор фактически задает вопрос, относительно какого состояния объясняется прогноз [1-4].

Проблема особенно заметна во временных рядах. Лаги, скользящие средние и сезонные компоненты формируются из связанных наблюдений, поэтому независимая замена одного признака может создавать комбинацию, которая редко встречается в данных [4, 7]. В современных исследованиях устойчивость объяснений рассматривается как самостоятельное свойство: оценивается повторяемость ранжирования факторов, чувствительность к небольшим изменениям данных и способность сохранять содержательно близкие объяснения для близких ситуаций [5, 6, 8].

Цель статьи – проверить, насколько локальные объяснения прогнозов спроса изменяются при выборе разных базовых значений и при совместной замене связанных признаков. Исследование продолжает разработку программного модуля интерпретации прогнозов, но рассматривает другой вопрос: не точность прогноза, а устойчивость самого объяснения.

Постановка эксперимента

Эксперимент выполнен на том же воспроизводимом синтетическом ряде ежедневных продаж, в котором учитываются сезонность, промоакции, особые и выходные дни, индекс цены и случайный шум. После построения лагов и скользящих средних сформировано 700 наблюдений: 560 обучающих и 140 тестовых. Модели получают 12 признаков. Использованы линейная регрессия и случайный лес из 200 деревьев.

Для каждого тестового наблюдения объяснение рассчитывалось относительно трех базовых профилей. Первый профиль образован средними значениями признаков на всей обучающей выборке. Второй использует медианы. Третий представляет локальный по времени профиль – средние значения последних 30 дней обучающего периода.

Таблица 1

Использованные базовые представления

База	Расчет	Интерпретация
Среднее	Среднее по 560 обучающим наблюдениям	Глобальный типичный уровень признаков
Медиана	Медиана по 560 обучающим наблюдениям	Устойчивая к выбросам центральная точка
Последние 30 дней	Среднее последних 30 обучающих дней	Недавний временной режим перед тестовым периодом

Для линейной регрессии при любой выбранной базе вклад i-го признака определялся как произведение коэффициента модели на отклонение текущего значения от базы. Поэтому сумма базового прогноза и вкладов должна точно восстанавливать результат модели. Для случайного леса использовалась локальная замена одного признака базовым значением; полученные эффекты не обязаны быть аддитивными.

Устойчивость оценивалась тремя показателями. Коэффициент Спирмена рассчитывался для рангов абсолютных вкладов всех 12 признаков. Для трех наиболее сильных факторов дополнительно определялись средний коэффициент Жаккара и доля наблюдений, в которых состав top-3 совпал полностью. Отдельно измерялась доля совпадения знаков влияния.

Устойчивость при изменении базовой точки

Для линейной регрессии смена базы не нарушила аддитивность. Максимальный остаток на 140 тестовых наблюдениях для всех трех вариантов не превысил $2{,}84 \cdot 10^{-14}$. Одновременно изменилось содержание объяснений. При сравнении среднего и медианного профилей средний коэффициент Спирмена по абсолютным вкладам составил 0,535, а полный состав трех ведущих факторов совпал только в 22,9 % наблюдений.

Причина состоит в том, что медиана по бинарным признакам промоакции и выходного дня равна нулю. Если в анализируемый день соответствующее значение также равно нулю, вклад такого признака относительно медианной базы исчезает. Средняя база, напротив, содержит дробную долю промо- и выходных дней и отвечает на другой вопрос: насколько текущий день отличается от среднего режима обучающего периода.

Сравнение среднего профиля с базой последних 30 дней дало более высокую ранговую устойчивость линейной модели: средний коэффициент Спирмена 0,883, а полный состав top-3 совпадал в 60,7 % наблюдений. При этом базовый прогноз изменился с 113,43 до 91,99 единицы. Следовательно, математическая точность разложения сама по себе не гарантирует неизменность смысловой интерпретации.

Таблица 2

Устойчивость локальных объяснений при смене базы

Модель и сравнение	Спирмен	Жаккар top-3	Полное совпадение top-3, %	Совпадение знака, %
Линейная: среднее – медиана	0,535	0,584	22,9	78,0
Линейная: среднее – 30 дней	0,883	0,804	60,7	82,6
Случайный лес: среднее – медиана	0,980	0,982	96,4	96,1
Случайный лес: среднее – 30 дней	0,870	0,690	39,3	81,0

У случайного леса картина оказалась иной. Средняя и медианная базы дали почти одинаковое ранжирование: средний коэффициент Спирмена равен 0,980, а состав трех ведущих факторов полностью совпал в 96,4 % тестовых наблюдений. При переходе к базе последних 30 дней полное совпадение top-3 снизилось до 39,3 %, хотя общий коэффициент Спирмена остался достаточно высоким – 0,870.

Диагностический остаток также изменялся вместе с базой. Среднее абсолютное значение остатка для случайного леса составило 3,77 единицы при средней базе, 3,92 при медианной и 2,96 при базе последних 30 дней. Меньший остаток для недавнего профиля не следует автоматически трактовать как лучшее объяснение: такая база просто ближе к состоянию ряда перед тестовым периодом и задает иной контекст сравнения.

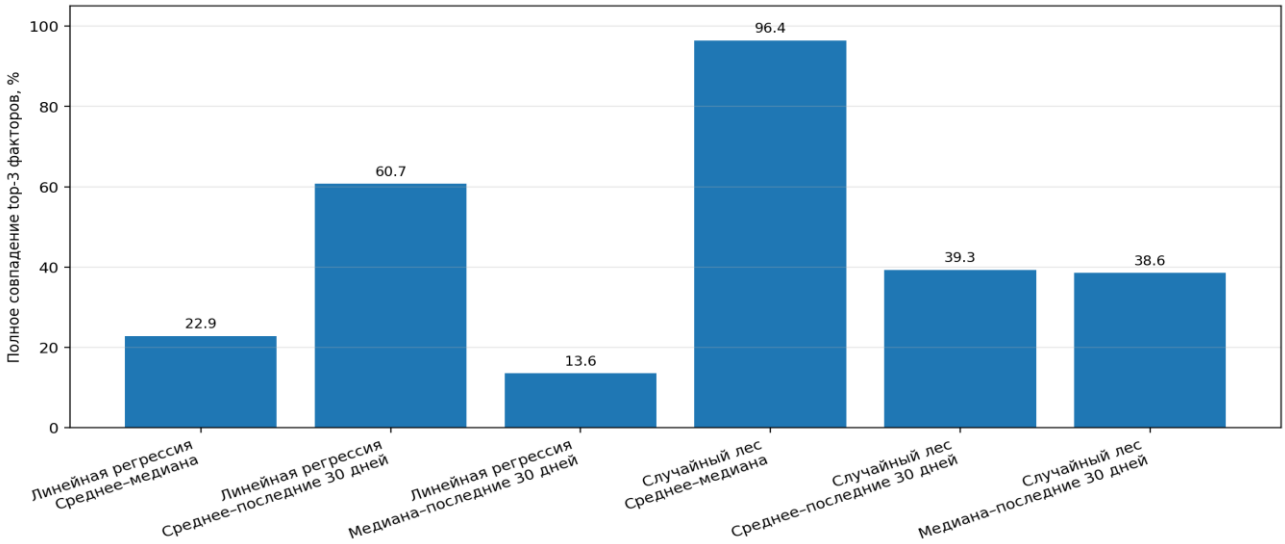

Рисунок 1. Доля полного совпадения трех наиболее значимых факторов при смене базы

Совместная замена связанных признаков

Отдельный эксперимент проведен для трех групп признаков: четырех лагов продаж, двух скользящих средних и пары сезонных компонент. Сначала вычислялись локальные эффекты каждого признака по отдельности, затем вся группа заменялась базовыми значениями одновременно. Разность между совместным эффектом и суммой индивидуальных эффектов использовалась как показатель взаимодействия.

Для линейной регрессии групповая и покомпонентная замены практически совпали: средние абсолютные расхождения находились на уровне 10^{-15}. Это соответствует аддитивной структуре модели. У случайного леса различия были заметны, особенно для лаговых признаков.

Таблица 3

Расхождение между совместной и покомпонентной заменой признаков

Группа признаков	Линейная модель, среднее |разрыва|	Случайный лес, среднее |разрыва|	Случайный лес, максимум |разрыва|
Лаги 1/7/14/30 дней	$5{,}59 \cdot 10^{-15}$	0,643	3,090
Скользящие 7/30 дней	$5{,}95 \cdot 10^{-15}$	0,169	0,747
Сезонные sin/cos	$5{,}67 \cdot 10^{-15}$	0,283	1,526

Наибольший средний разрыв для случайного леса получен в группе лагов – 0,643 единицы, максимальный достиг 3,09 единицы. Это означает, что влияние нескольких связанных прошлых значений нельзя всегда восстановить сложением результатов их независимых замен. Модель использует комбинации признаков, и совместное изменение может проходить по другим ветвям деревьев.

Связанность признаков подтверждается обучающей выборкой. Корреляция между скользящими средними за 7 и 30 дней равна 0,964, между средней за 7 дней и сезонной синусной компонентой – 0,975, между лагом в один день и средней за 7 дней – 0,878. Для временного ряда такие зависимости естественны, поэтому локальная интерпретация должна учитывать не только отдельные признаки, но и их группы.

Обсуждение результатов

Эксперимент показывает два разных вида устойчивости. Первый – вычислительная устойчивость: может ли объяснение точно восстановить прогноз. У линейной регрессии это свойство сохраняется при любой базовой точке, поскольку следует из самой формулы модели. Второй вид – смысловая устойчивость: сохраняются ли ведущие факторы и их знаки. Она уже зависит от выбранного контекста сравнения.

Эти результаты согласуются с исследованиями, где базовая выборка рассматривается как часть постановки объяснения, а не как техническая деталь [4]. Современные работы по устойчивости локальных объяснений также показывают, что близкие предсказания не гарантируют одинакового ранжирования признаков [5, 6]. Для временных рядов дополнительно возникает риск нарушить структуру данных при независимом маскировании или замене значений [7, 9].

Для прикладного модуля из этого следует практическое правило: вместе с локальным объяснением необходимо сохранять тип базы и способ замены признаков. Без этой информации два набора вкладов могут выглядеть противоречащими друг другу, хотя фактически они отвечают на разные вопросы. Для связанных временных признаков целесообразно дополнительно выводить групповые эффекты либо предупреждение о наличии сильных зависимостей.

Ограничения исследования

Эксперимент проведен на синтетическом наборе и двух моделях. Полученные численные значения нельзя переносить на реальные продажи без повторной проверки. Три выбранных базовых профиля не исчерпывают возможные варианты: в промышленной системе могут использоваться сезонные, товарные или сегментные эталоны.

Корреляция использована как простой индикатор связи и не описывает все формы зависимости. Групповой разрыв также не является универсальной мерой взаимодействия, но позволяет обнаружить случаи, когда сумма независимых локальных замен заметно отличается от совместного воздействия. Дальнейшая проверка должна включать реальные данные и сопоставление с SHAP, TimeSHAP или другими методами, учитывающими структуру временного ряда [10-12].

Заключение

Проведено исследование устойчивости локальных объяснений прогнозов спроса при изменении базовых значений признаков. Показано, что точное аддитивное разложение линейной регрессии сохраняется для разных баз, однако ранжирование и знак отдельных вкладов могут существенно меняться. Для случайного леса объяснения относительно средней и медианной базы оказались близкими, тогда как переход к профилю последних 30 обучающих дней заметно изменил состав ведущих факторов.

Совместная замена связанных признаков выявила дополнительное отличие нелинейной модели: для лаговой группы средний абсолютный разрыв между совместным эффектом и суммой индивидуальных замен составил 0,643 единицы, а максимальный – 3,09. Полученные результаты показывают, что для практического использования объяснений необходимо фиксировать базовый профиль, различать аддитивное разложение и локальную чувствительность и отдельно контролировать группы зависимых временных признаков

Список литературы:

1. Бирюков Д. Н., Дудкин А. С. Объяснимость и интерпретируемость – важные аспекты безопасности решений, принимаемых интеллектуальными системами // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 373-386. DOI: 10.17586/2226-1494-2025-25-3-373-386.

2. Рогулин Р. С. Обзор прикладных основ использования аналитики данных и машинного обучения в прогнозировании спроса // Экономические и социально-гуманитарные исследования. 2023. № 3. С. 115-126. DOI: 10.24151/2409-1073-2023-3-115-126.

3. 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. DOI: 10.1145/2939672.2939778.

4. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30.

5. Sepulveda E., Vandervorst F., Baesens B., Verdonck T. Enhancing explainability in real-world scenarios: Towards a robust stability measure for local interpretability // Expert Systems with Applications. 2025. Vol. 274. Article 126922. DOI: 10.1016/j.eswa.2025.126922.

6. Kelodjou G., Roze L., Masson V., Galarraga L., Gaudel R., Tchuente M., Termier A. Shaping Up SHAP: Enhancing Stability through Layer-Wise Neighbor Selection // Proceedings of the AAAI Conference on Artificial Intelligence. 2024. Vol. 38, No. 12. P. 13094-13103. DOI: 10.1609/aaai.v38i12.29208.

7. Meng H., Wagner C., Triguero I. SEGAL time series classification – Stable explanations using a generative model and an adaptive weighting method for LIME // Neural Networks. 2024. Vol. 176. Article 106345. DOI: 10.1016/j.neunet.2024.106345.

8. Aas K., Jullum M., Loland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values // Artificial Intelligence. 2021. Vol. 298. Article 103502. DOI: 10.1016/j.artint.2021.103502.

9. Chen H., Lundberg S. M., Lee S.-I. Explaining a series of models by propagating Shapley values // Nature Communications. 2022. Vol. 13. Article 4512. DOI: 10.1038/s41467-022-31384-3.

10. Chen Y., Kuang X., Zhang S. Robust time series explanations via information-theoretic temporal structural priors // Pattern Recognition Letters. 2026. Vol. 205. P. 37-43. DOI: 10.1016/j.patrec.2026.04.017.

11. Hertel M., Putz S., Mikut R., Hagenmeyer V., Schafer B. Explainable time-series forecasting with sampling-free SHAP for Transformers // Nature Communications. 2026. Vol. 17. Article 4723. DOI: 10.1038/s41467-026-73243-5.

12. Mishra L. N., Senapati B., Nayak S. K., Rangari A. J. Explainable Retail Demand Forecasting: A Hybrid LightGBM-SHAP Framework for Inventory Optimization // IEEE Access. 2026. Vol. 14. P. 43613-43633. DOI: 10.1109/ACCESS.2026.3675540