Автономная некоммерческая образовательная организация
высшего образования
«НАУЧНО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «СИРИУС»

Научный центр информационных технологий и искусственного интеллекта

направление «Математическое моделирование в биомедицине и геофизике»

К ЗАЩИТЕ ДОПУСТИТЬ

Научный руководитель направления
«Математическое моделирование
в биомедицине и геофизике»
чл.-к. РАН, д.ф.-м.н., профессор,
Ю.В. Василевский
«03» июля 2024 г.

ПРОГНОЗИРОВАНИЕ ПОТРЕБЛЕНИЯ КОНТРАГЕНТОВ ООО «ГАЗПРОМНЕФТЬ – РЕГИОНАЛЬНЫЕ ПРОДАЖИ»

Магистерская диссертация
по направлению подготовки 01.04.02 Прикладная математика и информатика
(направленность (профиль) «Математическое моделирование в биомедицине и нефтегазовом инжиниринге»)

Студент гр. М01ММ-22
А.Е. Донской
«03» июля 2024 г.

Научный руководитель магистерской диссертации
Профессор направления
«Математическое моделирование в биомедицине и геофизике» научного центра информационных технологий и искусственного интеллекта, к.ф.-м.н, Ph.D.
С.Ю. Малясов
«03» июля 2024 г.

Федеральная территория «Сириус», 2024

Автономная некоммерческая образовательная организация
высшего образования
«НАУЧНО-ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ «СИРИУС»

Научный центр информационных технологий и искусственного интеллекта

направление «Математическое моделирование в биомедицине и геофизике»

УТВЕРДИТЬ
Научный руководитель направления
«Математическое моделирование
в биомедицине и геофизике»
чл.-к. РАН, д.ф.-м.н., профессор,
Ю.В. Василевский
«30» января 2024 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ
на выполнение выпускной квалификационной работы
обучающегося по направлению подготовки
01.04.02 Прикладная математика и информатика
направленность (профиль)
«Математическое моделирование в биомедицине и нефтегазовом инжиниринге»

Донского Андрея Евгеньевича

1. Тема: «Прогнозирование потребления контрагентов ООО «Газпромнефть – Региональные продажи»
2. Цель: реализовать сценарное прогнозирование потребления нефтепродуктов в рамках проекта компании ООО «Газпромнефть – Региональные продажи»
3. Задачи:
Разработать и протестировать подходы к сценарному прогнозированию продаж. Оценить использование экзогенных данных для прогнозирования. Подготовить архитектуру проекта для дальнейшего масштабирования в рамках компании.
4. Рабочий график (план) выполнения выпускной квалификационной работы:

№	Перечень заданий	Сроки выполнения
1	Анализ специфики доступных данных, получение доступов, первичное общение с заказчиками.	30.01.2024 – 14.02.2024
2	На основе открытых статистических моделей Theta Model, Prophet разработать базовые модели прогнозирования в проекте.	14.02.2024 – 28.02.2024
3	Проработка стандартных помесячных и подневных моделей прогнозирования.	28.02.2024 – 14.03.2024

№	Перечень заданий	Сроки выполнения
4	Проработка модели сценарного прогнозирования – модели дополнения неполного месяца.	14.03.2024 – 28.03.2024
5	Проработка модели сценарного прогнозирования – иерархическая разбивка месяца.	28.03.2024 – 14.04.2024
6	Учет влияния экзогенных данных (выходные и праздничные дни, ценовые, температурные и другие факторы)	14.04.2024 – 01.05.2024
7	Оптимизация прогнозирования и загрузки данных.	01.05.2024 – 23.05.2024

Дата выдачи: «30» января 2024 г.
Руководитель ВКР:
С.Ю. Малясов

Задание принял к исполнению:
Студент группы М01ММ-22
А.Е. Донской
«30» января 2024 г.

Реферат

Выпускная квалификационная работа «Прогнозирование потребления контрагентов ООО «Газпромнефть – Региональные продажи» содержит: 57 страниц, 13 рисунков, 3 таблицы, 24 источника.

Ключевые слова: ВРЕМЕННЫЕ РЯДЫ, ПОСТАВКА ТОПЛИВА, СЦЕНАРНОЕ ПРОГНОЗИРОВАНИЕ, ИЕРАРХИЧЕСКАЯ РАЗБИВКА

Объектом исследования являются стратегии сценарного прогнозирования в контесте поставок топлива на АЗС.

Цель работы:

Реализовать сценарное прогнозирование потребления контрагентов в рамках проекта для ООО «Газпромнефть – Региональные продажи».

Задачи работы:

Разработать и протестировать подходы к сценарному прогнозированию продаж. Оценить использование экзогенных данных для прогнозирования. Подготовить архитектуру проекта для дальнейшего масштабирования в рамках компании.

Результаты работы:

В данной работе удалось разработать несколько моделей сценарного прогнозирования, основанные на базовых моделях Theta Model и Prophet. Были рассмотрены влияние факторов на результаты прогноза, а также подходы к ускорению получения планов.

На основе полученной работы уже согласован автоматический прием дальних планов, а также расширены исходные данные на новые сегменты.

The abstract

The final qualifying work «Customer demand forecasting at LLC «Gazprom Neft – Regional Sales» comprises: 57 pages, 13 figures, 3 tables, 24 references.

Keywords: TIME SERIES, FUEL SUPPLY, SCENARIO FORECASTING, HIERARCHICAL DECOMPOSITION

The subject of the Study is the development of forecasting strategies for the supply of fuel to gas stations.

The purpose of the Study:

To implement customer demand scenario forecasting within the project for LLC «Gazprom Neft – Regional Sales».

The objectives of the Study:

To develop and test approaches to scenario forecasting of sales. In addition, the study will evaluate the use of exogenous data for forecasting. Finally, the project architecture will be prepared for further scaling within the company.

The results of the Study:

This work successfully developed several models of scenario forecasting based on the Theta Model and Prophet. The influence of factors on the forecasting results and approaches to accelerating the acquisition of plans were examined.

Based on the results, the automated reception of long-term plans has already been agreed upon, and the original data has been expanded to new segments.

Сокращения, обозначения, термины и определения

В настоящей работе применяют следующие сокращения и обозначения:

1. АБ – Автомобильный бензин;
2. БД – База данных;
3. Вид топлива АСКУ – Тип продаваемого продукта (например, «Бензин 92»);
4. ГАЗ – Газовое топливо;
5. Группа НП – Группа нефтепродукта, более точная классификация топлива (например, «АИ-92»);
6. ДТ – Дизельное топливо;
7. ИНН – Индивидуальный номер контрагента;
8. Категория НП – Категория нефтепродукта (АБ, ГАЗ, ДТ);
9. Классификация АЗС – Виды АЗС на рынке (например, «Сеть АЗС ГПН»);
10. Номенклатура – То же, что и Вид топлива АСКУ;
11. НКП – Сегмент юридических лиц;
12. НП – Нефтепродукт;
13. Озеро Данных – общее хранилище неструктурированных данных разного формата;
14. Отделение – Регионы потребления (например, «Отделение Центр»);
15. План – Прогноз потребления на будущее (например, «План-1»);
16. Разрез – Гранулярность данных;
17. Сегмент – Группа контрагентов;
18. Тип ТО – Тип точки обслуживания, то же, что и Классификация АЗС;
19. Факт – Фактические данные;

20. Фреймворк – Широкий набор инструментов программного обеспечения для решения определенной задачи;

21. Baseline – Простой и быстрый для реализации подход к прогнозированию;

22. B2G – Предприятия госсектора;

23. CRT – Крупные корпоративные клиенты;

24. Fleet – Мелкие корпоративные клиенты;

25. Prophet – Открытый пакет для статистического предсказания временных рядов (в рамках проекта – реализация на Python);

26. Theta Model – Название алгоритма простой статистической модели (в рамках проекта – реализация пакета DARTS на Python).

Оглавление

Введение

Прогнозирование потребления контрагентов в нефтегазовой отрасли является сложной, но важнейшей задачей, позволяющей обеспечивать эффективное управление ресурсами компании и адаптацию к динамично изменяющимся условиям мирового рынка. Способность предвидеть изменения в спросе на нефтепродукты позволяет не только строить долгосрочные стратегические планы, но и оперативно реагировать на краткосрочные колебания рынка, что является критически важным для поддержания конкурентоспособности и финансовой стабильности [1].

В данной работе основное внимание уделяется изучению и анализу различных математических подходов к прогнозированию временных рядов, которые могут быть применены для оценки будущего спроса. Среди наиболее распространенных методов – статистические модели, а также более современные подходы, которые позволяют обрабатывать большие объемы данных для более точного прогнозирования.

Важность данного исследования также подчеркивается взаимосвязью прогнозов спроса с другими аспектами операционной деятельности компании, включая управление запасами, планирование закупок, определение оптимальных маршрутов транспортировки и распределение ресурсов. Эффективное прогнозирование способствует снижению затрат, оптимизации складских запасов и уменьшению потерь, связанных с неоптимальной логистикой.

Кроме того, рассматривается влияние внешних экономических, политических и экологических факторов на потребление нефтегазовой продукции, что также должно быть учтено при разработке прогнозных моделей. Эти факторы могут оказывать значительное воздействие на мировые цены и спрос, а их анализ помогает предсказать потенциальные рыночные изменения, которые могут повлиять на стратегическое планирование и операцион-

ную деятельность компании.

Ввиду вышесказанного, сценарное прогнозирование играет важную роль в управленческом арсенале нефтегазовых компаний, поскольку оно позволяет не только оценить вероятные будущие состояния рынка, но и подготовиться к различным возможным исходам [2]. Использование сценарного анализа предоставляет возможность моделировать разные экономические, политические и технологические условия, влияющие на спрос и цены на нефтегазовую продукцию. Это особенно важно в условиях, когда рыночная среда характеризуется высокой степенью неопределенности и динамичностью изменений.

Для каждого сценария могут быть разработаны специализированные прогнозные модели, которые учитывают определенные предположения о будущем. Имея несколько различных моделей для различных условий, компания может быстрее реагировать на изменения, что помогает в принятии обоснованных управленческих решений [1]. Такой подход не только способствует снижению потенциальных убытков, но и помогает оптимизировать операционные процессы и стратегическое планирование, повышая тем самым адаптивность компании к изменяющимся внешним и внутренним условиям.

Таким образом, настоящая работа представляет собой обзор предлагаемых инструментов и методик прогнозирования, которые могут помочь компаниям оставаться на шаг впереди рыночных изменений, поддерживать устойчивость в условиях неопределенности и максимально эффективно использовать доступные ресурсы для достижения стратегических целей.

1 ОБЗОР ЛИТЕРАТУРЫ

Прогнозирование временных рядов является важной областью анализа данных, которая занимается изучением и моделированием временных данных для предсказания будущих значений на основе прошлых и настоящих наблюдений. Эта дисциплина находит применение в самых разных сферах, от финансов и экономики до метеорологии и здравоохранения [3].

1.1 Специфика прогнозирования временных рядов

Прогнозирование временных рядов включает анализ и использование исторических данных для предсказания будущих значений. Эта задача имеет несколько специфических аспектов, которые делают её отличной от других видов аналитических задач.

1.1.1 Основные компоненты временного ряда

Временной ряд — это последовательность точек данных, индексированных во временном порядке. Основная задача анализа временных рядов — моделирование и прогнозирование этих данных. Временные ряды характеризуются наличием четырёх основных компонентов: тренда, сезонности, цикличности и остаточных (или случайных) компонентов [4]. Эти компоненты могут быть изучены и моделированы с использованием различных методов машинного обучения и статистических подходов.

1. **Тренд**

Тренд отражает долгосрочное направление временного ряда. Он может быть восходящим, нисходящим или стационарным. Математически тренд часто моделируется как линейная функция (1.1),

$$T_t = a + b \cdot t \tag{1.1}$$

где T_t – значение тренда в момент времени, а a и b – параметры, описывающие уровень и угол наклона тренда соответственно.

2. **Сезонность**

Сезонность описывает повторяющиеся краткосрочные колебания в данных, которые обычно связаны с временем года, месяца или недели. Сезонный компонент может быть определен выражением (1.2),

$$S_t = A \cos\left(\frac{2\pi t}{P} + \phi\right) \tag{1.2}$$

где P – период сезонности, A – амплитуда, а ϕ – фаза колебаний.

3. **Цикличность**

Цикличные колебания – это колебания, длительность которых более долгая, чем сезонные колебания, и которые не связаны с фиксированным календарным периодом. Эти колебания могут быть вызваны экономическими, политическими или другими факторами. Модель может быть представлена аналогично уравнению (1.2).

4. **Остаточные (случайные) компоненты**

Остаточные компоненты представляют собой шум, который остается после учета всех других компонентов. Это непредсказуемые случайные изменения, которые не описываются трендом, сезонностью или цикличностью.

1.1.2 Типы временных рядов

Временные ряды можно классифицировать по различным критериям, в зависимости от их характеристик и природы данных. Основные типы временных рядов включают стационарные, нестационарные, детерминированные и стохастические временные ряды [5].

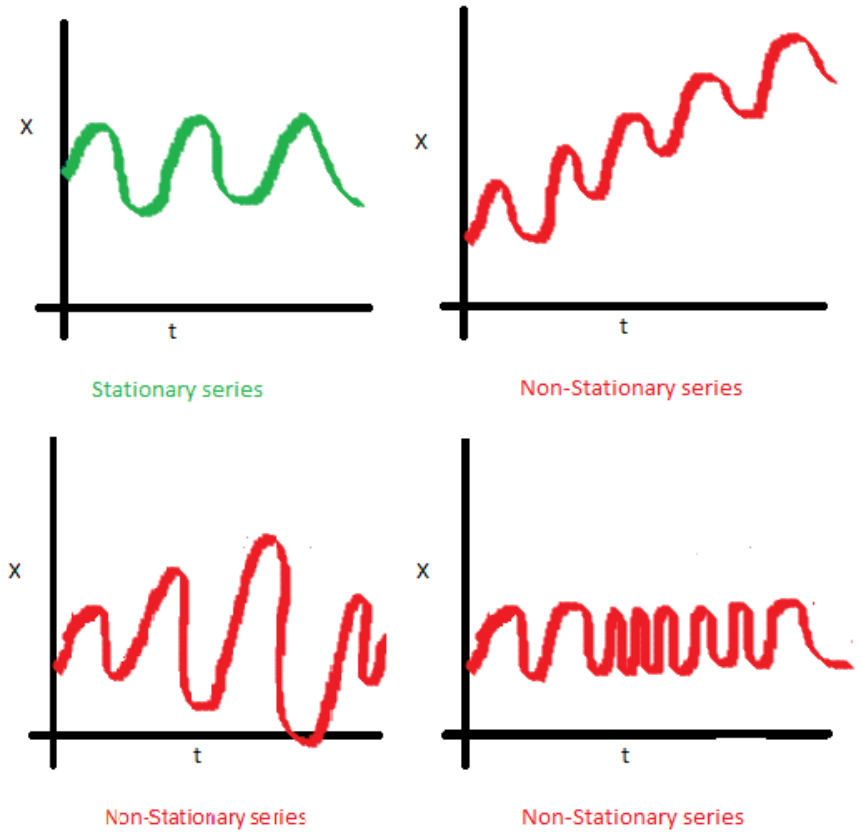

Рисунок 1.1 – Примеры стационарных и нестационарных рядов

Различение этих типов критически важно для выбора подходящих методов их анализа и прогнозирования.

1. **Стационарные и нестационарные временные ряды**

Стационарные временные ряды обладают свойствами, которые не зависят от времени. Для стационарного временного ряда следующие статистические характеристики остаются постоянными: среднее значение $E[X_t] = \mu$; дисперсия $Var(X_t) = \sigma^2$; ковариация $Cov(X_t, X_{t+k}) = C(k)$, где $C(k)$ зависит только от лага k, а не от времени t.

Нестационарные временные ряды демонстрируют изменения в статистических характеристиках во времени. Эти изменения могут быть вызваны трендами, сезонностью или другими структурными изменениями в данных. Нестационарные ряды часто требуют предварительной обработки, такой как

дифференцирование или преобразование данных, чтобы сделать их подходящими для статистического анализа.

Примеры рядов приведены на рисунке 1.1.

2. **Детерминированные временные ряды**

Детерминированные временные ряды определяются точными функциональными отношениями и не содержат случайных компонентов. Примеры включают математические формулы или алгоритмически сгенерированные последовательности. Такие временные ряды полностью предсказуемы, если известны их начальные условия и правила генерации.

3. **Интегрированные временные ряды**

Интегрированный временной ряд обозначает, что ряд становится стационарным после одной или нескольких операций дифференцирования. Эти ряды часто моделируются с помощью моделей ARIMA [6].

Например, если ряд X_t не стационарен, но $\Delta X_t = X_t - X_{t-1}$ стационарен, то ряд X_t является интегрированным рядом порядка 1 и обозначается как $I(1)$.

4. **Сезонно интегрированные временные ряды**

Сезонно интегрированные временные ряды требуют сезонного дифференцирования для достижения стационарности. Это особенно важно в случаях, когда временные ряды демонстрируют сильные сезонные колебания. Примером может служить применение сезонной разности в сезонных моделях ARIMA [7], где сезонное дифференцирование учитывает периодичность данных.

1.2 Подходы к прогнозированию

Существует множество программных пакетов и библиотек, которые могут быть использованы для анализа и прогнозирования временных рядов:

• R и его пакеты, такие как forecast, fable [8], и prophet, предоставляют мощные инструменты для статистического анализа и моделирования временных рядов;

• Python предлагает библиотеки, такие как statsmodels для традиционных статистических методов, scikit-learn для машинного обучения, и pytorch или tensorflow для глубокого обучения;

• SAS и SPSS также предлагают развитые средства для анализа временных рядов [9], хотя они могут быть менее доступны из-за лицензионных ограничений.

Глобально методы прогнозирования временных рядов можно разделить на 3 группы:

1. Локальные (чаще всего статистические модели семейства ARIMA [7] или похожие) – построение модели под каждый временной ряд в отдельности;

2. Глобальные (нейронные сети, бустинговые модели [10; 11]) – построение одной модели на все имеющиеся временные ряды, за счет того, что ищутся общие особенности между группой временных рядов сразу;

3. Ансамблевые – объединение предыдущих способов через одну модель. Чаще всего применяются различные методы агрегации и усреднения результатов модели, а также поиска наиболее сильных (уверенных) предсказаний среди них.

В рамках реализованного проекта в разделе 3.2 будут рассмотрены локальные, а также ансамблевые подходы в разделах 3.3.3 и 3.3.4.

2 ПОСТАНОВКА ЗАДАЧИ

2.1 Постановка задачи от заказчика и практическая значимость

Необходимо реализовать модель планирования потребления по клиенту в разрезе месяца, номенклатуры, отделения и классификации АЗС для осуществления закупочных мероприятий.

Исходная гранулярность данных – по месяцам, но в процессе необходим переход на подневные данные, в том числе с возможностью осуществлять прогноз в разрезе АЗС для каждого отделения.

Текущий подход заказчика строится вручную на основании фактической реализации за несколько последних полных недель без учета праздничных дней.

Также, необходимо встраивание учета внешних факторов в модель. В контуре данных компании доступна полезная информация, рассматриваемая как её потенциальный фактор: цены, погодные условия, праздничные дни и прочее.

2.2 Базовые понятия в задаче

В рамках проекта изначально рассматривается три сегмента корпоративных клиентов (контрагентов):

1. **Fleet** – «мелкие» корпоративные клиенты, средний пролив в месяц по данным последнего года менее 10 тысяч литров;
2. **CRT** – «крупные» корпоративные клиенты, средний пролив в месяц по данным последнего года более 10 тысяч литров;
3. **B2G** – предприятия госсектора (муниципальные коммунальные службы, общественный транспорт и прочее).

Для каждого из имеющихся сегментов формируется до пяти видов месячных бюджетов разной дальности планирования:

1. **ТП-90** – План закупки на 4 месяца вперед (90 дней: например, в мае формируется ТП-90 на сентябрь);
2. **ТП-60** – План закупки на 3 месяца вперед (60 дней: например, в мае формируется ТП-60 на август);
3. **ТП-35** – План закупки на 2 месяца вперед (35 дней: например, в мае формируется ТП-35 на июль);
4. **ТП-10** – План закупки на следующий месяц (10 дней: например, ТП-10 на июнь формируется в середине мая);
5. **ТП-0** – План закупки на следующий месяц в конце текущего месяца (0 дней: например, ТП-0 на июнь формируется 31 мая - 1 июня).

В процессе переговоров с заказчиком, наименования для удобства были изменены, потому что не отражали действительность. Например, ТП-35, несмотря на свое название, подразумевающее прогноз за 35 дней, на самом деле не производился ровно за 35 дней (аналогично с другими планами). Ввиду этого, существует другое обозначение прогнозов, называемое «упрощенной терминологией относительно терминологии заказчика»:

1. **ТП-90** – План +5 (План 5);
2. **ТП-60** – План +4 (План 4);
3. **ТП-35** – План +3 (План 3);
4. **ТП-10** – План +2 (План 2);
5. **ТП-0** – План +1 (План 1).

В такой интерпретации, например, План +1 на июнь можно получить только на основе полного месяца мая.

Кроме этого, после введения иерархического сценарного прогнозирования из раздела 3.3.4, а также модели дополнения неполного месяца из раздела 3.3.3, появилась необходимость достраивания текущего месяца до конца. Такое прогнозирование назовем «положительными планами». Чаще всего, реализуются 3 вида положительных планов:

1. **ТП+5** – Положительный план +5 (прошло 5 дней текущего месяца), необходимо спрогнозировать 25 оставшихся дней;
2. **ТП+15** – Положительный план +15 (прошло 15 дней текущего месяца), необходимо спрогнозировать 15 оставшихся дней;
3. **ТП+25** – Положительный план +25 (прошло 25 дней текущего месяца), необходимо спрогнозировать 5 оставшихся дней;

2.3 Оценка качества прогнозов

Основная выбранная оптимизируемая метрика в задаче предсказания объемов на временных рядах – средняя абсолютная ошибка прогноза (2.1), где Y_i – фактическое значение временного ряда во время i, а $\hat{Y}_i$ – прогноз модели в момент времени i.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{2.1}$$

Целесообразность выбора метрики объясняется её интепретируемостью для заказчика.

Эта метрика считается для каждого временного ряда в нужном разрезе. Далее полученные предсказанные объемы складываются для каждого отделения относительно разреза номенклатуры продукта (например, АИ-92, АИ-95 и так далее для АБ). Впоследствии, ошибка взвешивается на объемы по отделениям и категориям НП и усредняется, так как важнее получить меньшую ошибку именно на тех отделениях, на которых ожидается большой объем

продаж, а для отделений (или нефтепродуктов) с низким ожидаемым объемом, ошибка менее значительна.

2.4 Исходные данные

Преимущественно основные данные для осуществления прогноза находятся в «Озере данных» (общее место хранение неструктурированных данных разного формата) и могут быть получены через прямой SQL–запрос к базе данных. Верхнеуровнево, существует три вида данных:

1. **Оперативный факт** – текущие (оперативные) значения по транзакциям. Непосредственно используются внутри прогнозной модели;
2. **Закрытый факт** – транзакционные данные по закрытому периоду, то есть с учетом корректировок после закрытия периода;
3. **План** – план поставок, текущее прогнозирование, которое предстоит превзойти по качеству.

Таким образом, одновременно существуют два вида фактических данных – оперативный и закрытый. Закрытый факт выгружается значительно реже, поэтому основная модель строится именно на открытом факте, который обновляется ежедневно. Однако, ввиду того, что период может еще являться не закрытым, значения между фактами могут незначительно отличаться. В контексте задачи это отличие считается несущественным. На рисунке 2.1 отображены отличия, фиксируемые при выгрузке данных. Рисунок получен с помощью реализованного дашборда (подробнее в разделе 5.2).

В таблице 2.1 отражен синтетический пример фактических данных, на основе которого можно понять структуру временных рядов.

Все данные собраны в хранилище данных и включают информацию по продажам с 2019-01-01 по текущий день.

Рисунок 2.1 – Динамика в различных фактических данных, конфиденциальная информация скрыта

Таблица 2.1 – Пример информации по поставкам топлива, конфиденциальная информация скрыта

year	month	Сегмент	Отделение	Вид топлива АСКУ	Категория НП	Группа НП
2020	12	CRT	Отделение Урал	Топливное с присадками зимнее ДТ	ДТЗ	ДТ
2020	3	Fleet	Отделение Урал	Бензин 92	АБ	Аи-92
2022	5	Fleet	Отделение Тюмень	Топливное с присадками летнее	ДТ	ДТЛ
2023	5	B2G	Отделение Новосибирск	Бензин 95	АБ	Аи-95
2021	8	Fleet	Отделение Центр	Бензин 95 бренд	АБ	Аи-95

Таблица 2.2 – Сводная таблица по используемым данным

Название объекта данных	Цель использования	Выгрузка
Транзакционные данные по клиентам	Для оценки транзакционной активности клиента за год	Прямое подключение к БД
Данные о сегменте клиента	Для добавления сегмента к транзакционным данным	Прямое подключение к БД
Закрытый факт по транзакциям клиентов	Для оценки точности прогноза и плана	Выгрузка от контактного лица
Данные по ценам (биржа, мелкий опт, конкуренты)	Для добавления ценовых факторов в модель	Выгрузка из BI-приложений
Данные по договорам клиентов	Для детекции уходящих клиентов	Прямое подключение к БД
Справочник клиентов	Для присоединения наименований клиентов к прогнозам	Прямое подключение к БД
Данные об объектах управления	Для присоединения пятизначных номеров АЗС к прогнозам	Прямое подключение к БД

В таблице 2.2 приведена сводная таблица по всем используемым данным и их источникам выгрузки. Получение прогнозов напрямую зависит от актуальности этих данных, ввиду чего была реализована функциональность ежедневного автоматического обновления и сохранения данных в нужном разрезе.

Помимо описанных выше данных, в проекте используются как внешние источники информации (например, производственный календарь), так и внутренние, но не относящиеся непосредственно к работе проекта (например, температурные условия в городе). Эта информация чаще всего используется в качестве дополнительного фактора моделей и корректируется, например, в выходные дни. Более подробно применение факторных моделей рассмотрено в разделе 3.5.

3 СТРАТЕГИИ ПРОГНОЗИРОВАНИЯ

3.1 Оценка внедряемых изменений

В разделе 2.3 ранее была описана основная метрика, относительно которой происходит оценка точности модели. Однако в контексте прогнозирования временных рядов важна не столько фактическая ошибка, сколько стабильность модели в целом – то есть, насколько ожидаема аналогичная ошибка на будущем периоде. Ввиду этого, для оценки проводимых экспериментов используется подход, который называется «Обратное тестирование» (чаще на английском – «Back–Testing») [12].

На рисунке 3.1 приведена схема обратного тестирования. Она схожа с классическим подходом кросс-валидации при оценке моделей машинного обучения, но имеет некоторые отличия.

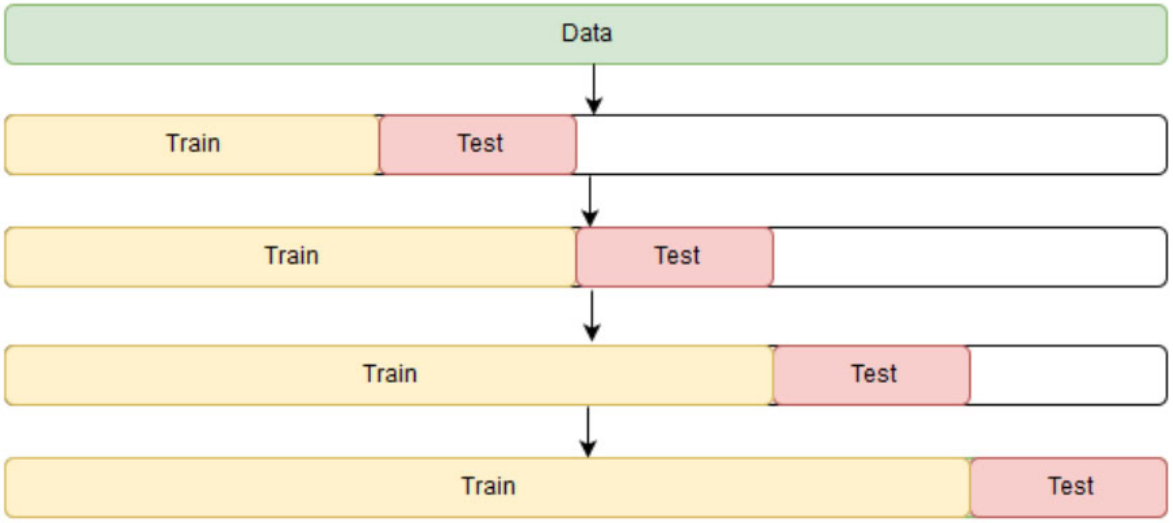

Рисунок 3.1 – Иллюстрация обратного тестирования

Аналогично обычной кросс-валидации, вся выборка данных разбивается на K-частей (фолдов, от англ. fold), но размеры этих частей неодинаковы друг относительно друга. Кроме этого, ввиду того, что временные данные имеют явную структуру, зависящую от времени, предсказание всегда ведется

вперед.

Таким образом, на основе одной выборки данных мы можем получить несколько предсказаний, при этом не допуская «заглядывания в будущее», то есть ситуации, когда мы учимся на данных, которые впоследствии будем предсказывать.

После получения предсказаний, для каждого из результатов считается ошибка по формуле (2.1) и далее усредняется по дате. Это дает нам взвешенную оценку на ошибку модели и позволяет достоверно судить, насколько эффективны изменения, предлагаемые в процессе разработки модели. Именно на основе такой методики производился дальнейший отбор используемых моделей.

3.2 Базовые используемые модели

В качестве основы для прогнозирования на текущий момент выступают две статистические модели – Theta Model и Prophet, каждая из которых проявляет себя лучше в определенных условиях, например, на конкретном сегменте и отделении. Детальное описание структуры моделей приведено в текущих подразделах.

На основе простых моделей строятся сценарные подходы, которые имеют большую ценность для заказчика, потому что позволяют учесть особенности продаж на рынке. Эти модели будут рассмотрены в разделе 3.3.

3.2.1 Theta Model

Theta Model является методом прогнозирования временных рядов, который получил известность благодаря работе [13] Спироса Макридакиса и его коллег. Эта модель была предложена в рамках соревнования «M3-Competition» [14], где она показала выдающиеся результаты, опередив мно-

гие другие методы прогнозирования.

Theta Model основывается на идее, что временной ряд можно представить как сумму двух или более компонент, каждая из которых может быть представлена через преобразования с определенным параметром тета θ. Суть метода заключается в декомпозиции исходного временного ряда и применении различных преобразований, направленных на сглаживание или уточнение данных.

Итоговое предсказание $\hat{Y}_t$ на момент времени t является взвешенной суммой так называемых тета-линий $Z_t(\theta_i)$. Полная формула прогноза приведена на (3.1). Чаще всего коэффициенты для взвешивания равны $\omega_1 = \omega_2 = 0.5$.

$$\hat{Y}_t = \omega \cdot Z_t(\theta_1) + (1 - \omega) \cdot Z_t(\theta_2) \tag{3.1}$$

$$Z_t(\theta) = \theta \cdot Y_t + (1 - \theta) \cdot (\alpha + \beta t) \tag{3.2}$$

Тета-линия задается через выражение (3.2) [13]. В оригинальной статье используются 2 линии с коэффициентами $\theta_1 = 0$ и $\theta_2 = 2$ [15]. Первая тета-линия уменьшает вариативность и сохраняет уровень данных, а вторая приводит к противоположному эффекту – усиление тренда и сезонности.

$$\hat{y}_{t+1} = \alpha \cdot y_t + (1 - \alpha) \cdot \hat{y}_t \tag{3.3}$$

$Z_t(\theta_1)$ прогнозируется линейно вперед и фактически отражает уровень временного ряда, а $Z_t(\theta_2)$ экстраполируется через экспоненциальное сглаживание ряда согласно уравнению (3.3) [15].

Преимущества модели:

1. Простота и понятность метода;
2. Высокая эффективность на множестве различных временных рядов;
3. Скорость работы алгоритма;

4. Гибкость в применении к разным типам данных.

Недостатки модели:

1. Модель может быть не очень эффективна для данных с сильными нелинейными зависимостями или в ситуации, когда временные ряды содержат сложные сезонные паттерны.

2. Необходимость выбора оптимального значения θ для каждого конкретного случая, что может потребовать дополнительных исследований и экспериментов.

3. Нельзя добавить внешние факторы

Theta Model, разработанная для прогнозирования временных рядов, получила несколько значимых модификаций, которые улучшают её применимость и точность в различных условиях и для разных типов данных. Одной из таких модификаций является Dynamic Optimized Theta [16].

Dynamic Optimized Theta Model вводит концепцию динамического подбора параметра θ на основе данных, что позволяет адаптироваться к изменениям в данных временных рядов. Чаще всего оптимизация выполняется неградиентными методами, такими как «Метод Нелдера-Мида» [17], он же «метод деформируемого многогранника».

За счет алгоритмов оптимизации, модель автоматически выбирает наиболее подходящее значение θ, что может значительно улучшить точность прогноза в сравнении с традиционной моделью, где θ фиксировано или выбирается исходя из эмпирического анализа.

Несмотря на свою легкость в реализации, Theta Model остается одним из популярных статистических методов прогнозирования временных рядов благодаря своей эффективности и универсальности. В рамках проекта эта модель выступает в качестве базовой.

3.2.2 Prophet

Пакет Prophet, представляет собой инструмент с открытым исходным кодом для прогнозирования временных рядов, который оптимизирован для работы с данными, демонстрирующими сильные сезонные эффекты и различные тренды на длинных временных интервалах. Эта модель удобна в использовании и эффективна в сценариях, где данные часто содержат пропуски или имеют значительные изменения в трендах, что обычно сложно моделировать с помощью традиционных методов временных рядов.

Модель Prophet представляет временной ряд $y(t)$ как комбинацию четырёх основных компонентов: тренд, сезонность, регрессоры и немоделируемый шум [18]. Используются следующие разновидности модели: аддитивная (3.4) и мультипликативная (3.5). Их различие состоит в том, насколько сильно изменения в тренде временного ряда влияют на общую динамику, например, сезонности. В случае мультипликативной модели влияние кратно больше, а в аддитивной остается постоянным.

$$y(t) = trend(t) + seasonality(t) + \beta \cdot regressor(t) + \epsilon_t \tag{3.4}$$

$$y(t) = trend(t) \cdot (1 + seasonality(t) + \beta \cdot regressor(t)) + \epsilon_t \tag{3.5}$$

Тренд $trend(t)$ – этот компонент моделирует изменения значения временного ряда во времени, что часто представляет основное направление данных. Обычно это кусочно-линейная или логистическая кривая, что позволяет модели гибко адаптироваться к изменениям тренда на различных временных интервалах.

Сезонность $seasonality(t)$ – описывает повторяющиеся короткие колебания в данных, которые могут быть связаны с дневной, недельной, месячной или годовой периодичностью. Prophet использует ряды Фурье (3.6) для моде-

лирования сезонных изменений [19], что позволяет точно улавливать сложные сезонные паттерны.

$$seasonality(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right)\right) \tag{3.6}$$

В формуле (3.6) N – количество гармонических составляющих (принято по умолчанию 10 для годовой сезонности и 3 для недельной), a_n и b_n – коэффициенты амплитуды косинусной и синусной составляющей, P – период сезонности (например, для годовой сезонности равен 365), t – индекс времени.

Регрессоры $regressor(t)$ – внешние переменные или факторы, которые могут влиять на изучаемый временной ряд. Формула приведена в (3.7). Примерами могут служить индикаторы праздников или маркетинговых акций. Эта часть модели позволяет включать в анализ информацию, которая известна заранее и может значительно повлиять на результаты. Важно уточнить, что функции $x_j(t)$ могут быть разрывными.

$$regressor(t) = \sum_{j=1}^{J} \beta_j x_j(t) \tag{3.7}$$

Ошибки (немоделируемый шум) $\epsilon_t \sim N(0, \sigma^2)$ – непредсказуемые случайные компоненты, представляющие собой остаточную разницу между моделируемыми значениями и реальными данными.

Прогнозирование временных рядов с помощью Prophet имеет ряд преимуществ, особенностей и недостатков, которые делают его подходящим или менее подходящим в зависимости от конкретной ситуации.

Преимущества модели:

1. Простота использования – Prophet разработан так, чтобы быть до-

ступным для аналитиков, которые не специализируются на статистическом моделировании временных рядов;

2. Обработка отсутствующих данных – Prophet способен обрабатывать пропуски в данных без предварительной обработки или заполнения пропусков, что является большим преимуществом при работе с реальными, часто неполными данными;

3. Гибкость компонентов – модель позволяет автоматически учитывать многие типы сезонности и подстраивать прогноз под них;

4. Включение внешних регрессоров;

5. Работа с изменениями тренда – модель может адаптироваться к изменениям тренда через включение точек изменения, что делает её эффективной для прогнозирования в условиях резких рыночных изменений или неожиданных событий.

Недостатки модели:

1. Производительность – Prophet требует достаточно большого количества данных для эффективного обучения, что может быть ограничением в случаях, когда доступные данные ограничены по времени;

2. Предположения о нормальности ошибок.

3.3 Модельные сценарии прогнозирования

На основе базовых моделей из раздела 3.2 были построены отдельные сценарные подходы, позволяющие учитывать особенности рыночной ситуации, а также использовать больше данных.

3.3.1 Помесячные модели

CustomTheta и **CustomProphet** – самые простые модели из представленных в категории сценарных. Фактически, они отличаются от базовых

незначительно. Основные особенности – необходимая предобработка данных и заполнение пропущенных значений в ряде. Сезонность моделей – 12 точек назад (то есть динамика текущих продаж совпадает с динамикой аналогичных месяцев в предыдущие года).

Эти модели используют исключительно помесячные данные. Согласно разделу 2.4, доступная для прогноза история начинается с 2019-01-01, что значительно сокращает длину временного ряда и не требует больших нагрузок на оперативную память.

Тем не менее, в данных есть некоторое количество коротких временных рядов. Это связанно либо с особенностями при записи продаж на конкретном отделении, либо с потерей информации. Часто для таких временных рядов доступно менее чем 24 точки для прогноза модели. Однако в таких условиях невозможно корректно оценить сезонную компоненту модели, либо она будет вносить большой вклад по ошибке. В связи с этим, было принято решение использовать в таких ситуациях наивное прогнозирование – то есть, заполнять прогноз последним известным значением. Такой подход не увеличивает глобальную ошибку в предсказании, так как часто такие ряды имеют маленький объем продаж, например, «Газ» в сегменте B2G.

Модели являются статистическими, поэтому в случае резко спадающего тренда прогноз модели может выдать отрицательные числа, что является неестественным. Такие случаи исправляются в постобработке предсказаний аналогично для всех сценариев.

3.3.2 Подневные модели

Согласно требованиям из раздела 2.1, необходима реализация прогнозов в разрезе дней, а впоследствии – и в разрезе АЗС. В проекте для этого написаны соответствующие базовым моделям классы **DailyCustomTheta** и

DailyCustomProphet.

Структура и логика модели совпадает с месячными моделями, но они используют подневные данные. Важно, что подобный переход требует значительно большей свободной оперативной памяти, особенно при использовании Prophet согласно методике раздела 3.2.2. Поэтому реализация подобных структур требовала в первую очередь именно оптимизации архитектуры проекта и потребления памяти при прогнозировании и выгрузке данных. Применяемые для этого подходы описаны в разделе 6.2.

Значительное преимущество таких моделей в том, что точек во временном ряде становится значительно больше, что позволяет более точно учитывать сезонность при прогнозе. Таким образом, при таком прогнозе возникает сразу несколько сезонностей – 12 месяцев (годовая) и 7 дней (недельная). Месячная сезонность (~ 30 дней) значительного вклада не вносит.

Для модели CustomProphet учет нескольких сезонностей встроен автоматически на основе базовой модели из раздела 3.2.2. Стандартная же модель Theta не позволяет учитывать две сезонности одновременно, поэтому реализация DailyCustomTheta сохраняет только недельную сезонность и является сильно упрощенной. Тем не менее, даже такая упрощенная модель в некоторых ситуациях показывает себя лучше, чем Prophet на основе обратного тестирования. Чаще всего такое заметно на временных рядах с плохо выраженной сезонностью или при её полном отсутствии, например, как у продукта «Газ» некоторых сегментов.

3.3.3 Модель дополнения неполного месяца

В процессе сравнения автоматических прогнозов с референсными результатами было выявлен существенный недостаток помесячных моделей – они не могут улавливать резкую динамику на текущем рынке, например, если

произошли какие-то значительные изменения в налоговом законодательстве. Невозможность учета таких данных связана с «неполным месяцем».

Например, если мы находимся в дате 2024-06-15 и хотим сделать прогноз на июнь до конца месяца, то используя помесячные модели, этот прогноз не будет ничем отличаться от прогноза, сделанного 2024-06-01, так как данные с 2024-06-01 по 2024-06-15 нельзя учитывать в прогнозе. Эти данные составляют неполный месяц и в **помесячной модели** могут использоваться только в том случае, если будут сложены до месяца 2024-06. Но полученный объем будет неполным относительно фактического, что значительно ухудшит прогноз.

В связи с этим была реализована MonthCompletionModel – модель дополнения неполного месяца, которая использует одновременно и помесячный разрез данных, и подневный. Иллюстрация работы представлена на рисунке 3.2.

Стратегия прогнозирования:

1. Строится прогноз вперед на 1 точку с помощью помесячной модели и помесячных данных (красная точка на рисунке);
2. Подневные данные прогнозируются с помощью подневной модели до конца месяца (зеленая линия на рисунке);
3. Подневный объем последнего полного месяца складывается. В его составе одновременно и прошедшие дни, и предсказанные до конца неполного месяца (зеленая точка на рисунке);
4. Помесячный и подневный прогноз на шаг вперед взвешиваются с определенными весами (оранжевая точка на рисунке);
5. Полученный взвешенный прогноз добавляется к помесячному временному ряду (синяя линяя);
6. На основе дополненного временного ряда делается предсказание

вперед на необходимое количество точек вперед.

Рисунок 3.2 – Механизм работы MonthCompletionModel, конфиденциальная информация скрыта

На примере работы алгоритма на Отделении Кузбасс (Бензин 100) видно, что полученный подход подхватывает изменение в подневном тренде и снижает помесячный прогноз на одну точку вперед. То есть, если сравнивать дополнение месяца в этой ситуации, без подневных данных динамика идет вверх, а с подневными корректируется на снижение. Благодаря этому итоговый прогноз на дальние планы также становится сниженным.

Аналогично работает при повышенных продажах в процессе месяца, что позволяет оперативно среагировать на это и скорретировать планы.

3.3.4 Иерархическая разбивка месяца

Второй комплексный подход, который используется в проекте – иерархическая разбивка месяца, схема которой представлена на рисунке 3.3. Предложенный подход проявил себя при необходимости реализации точных подневных прогнозов. Это связано с тем, что подневные модели из раздела 3.3.2

несмотря на хорошее качество по некоторым сегментам помесячно, не предоставляли достаточно хорошую метрику в подневном разрезе.

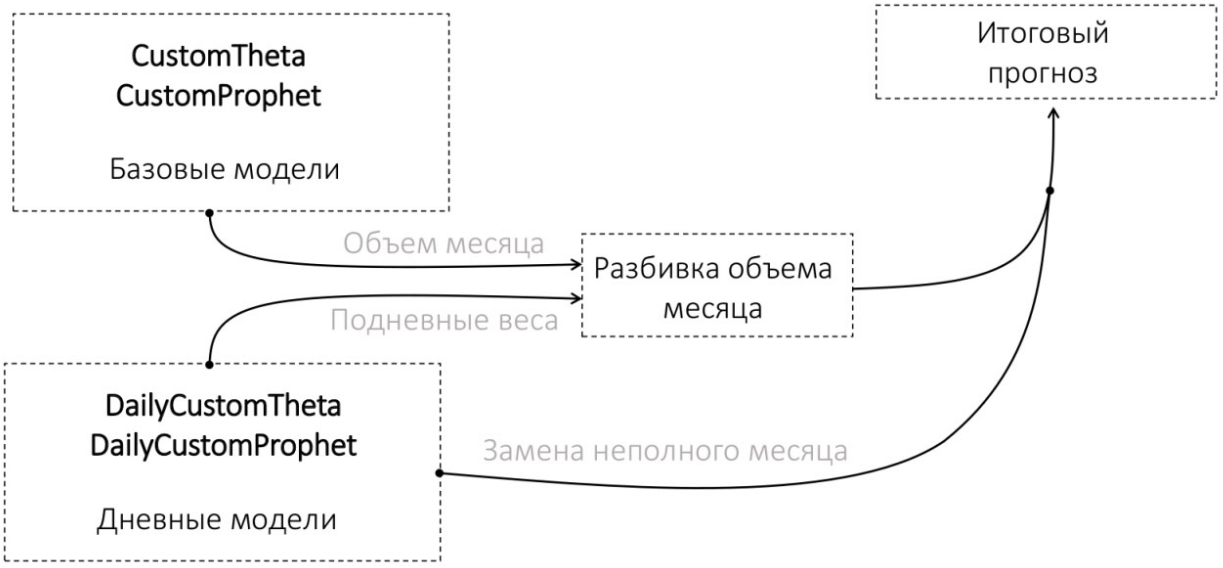

Рисунок 3.3 – Механизм работы HierarchicalModel

Было замечено, что классические базовые модели CustomProphet и CustomTheta дают более точные результаты по объемам месяца, а подневные DailyCustomTheta и DailyCustomProphet несмотря на хорошую динамику, ошибаются в суммарном объеме. Исходя из этого, была реализована новая стратегия прогнозирования, включающая в себя получение прогноза через помесячную базовую модель и разбивающая каждый прогнозируемый месяц по дневным долям, полученным через подневные модели.

Стратегия прогнозирования:

1. Получение помесячного прогноза на нужный план;
2. Получение подневного прогноза на нужный план;
3. Расчет долей подневного прогноза (доля каждого дня в разрезе месяца прогноза);
4. Разбивка помесячного прогноза через полученные доли;
5. Замена неполного месяца.

Кроме этого, последним шагом в прогнозировании является замена известных (прошедших) дней из прогноза на фактические для увеличения точности разбивки.

Несмотря на значительное увеличение количества необходимых операций при прогнозе, данный подход отлично проявляет себя при необходимости строить точные подневные планы, а также имеет удобное расширение при необходимости выйти на разрез АЗС–день.

3.4 Вспомогательные модели

Из-за необходимости учета метрики (2.1) при любом разрезе данных, важно уметь конвертировать любой прогноз в помесячный, а любой помесячный – в подневный. Последнее возможно с помощью реализации иерархической разбивки, которая напрямую дает возможность получить точный подневный прогноз при наличии хорошего предсказания объема на месяц. Для получения помесячного прогноза в случае подневных моделей, результаты за месяц складываются автоматически через вспомогательную модель ModelSum, которая также будет участвовать в оптимальном правиле, о котором будет рассмотрено в разделе 4.

3.5 Учет факторов

Исходя из преимуществ использования Prophet как базовой модели раздела 3.2.2, любые комплексные подходы, использующие эту модель, позволяют дополнительно включать внешние (экзогенные) регрессоры, описывающие данные. Подход по учету экзогенных данных широко распространен с момента появления моделей семейства ARIMA, в частности – SARIMAX [17].

Эта особенность наиболее важна при прогнозировании экономических

данных, потому что зачастую динамика продаж в определенные периоды напрямую зависит от внешних условий – косвенные продажи, привлечение покупателей через акционные кампании, температура на улице, выходные и праздничные дни. В связи с этим использование любой доступной внешней информации является важной гипотезой при необходимости увеличить точность прогноза.

3.5.1 Выходные и праздничные дни

В процессе реализации моделей и сравнения прогнозов с неавтоматическими планами было замечено, что прогнозы на праздничные месяцы (декабрь-январь, апрель-май) значительно выбиваются из общей динамики потребления.

Ввиду этого, из открытых источников был выгружен полный производственный календарь с указанием все выходных дней в году, чтобы использовать это как фактор модели. В случае подневных моделей раздела 3.3.2, фактором являлось бы, является ли день выходным (принимает значения 1 или 0), а в помесячных моделях раздела 3.3.1 фактором бы являлось количество выходных дней в рассматриваемом месяце.

Выгрузка подобных данных важна ежегодно, потому что количество выходных дней (и их фактическое расположение в календаре) отличается год к году. Например, в 2024 году значительно отличался апрель от предыдущего года. В 2023 году все выходные дни – только суббота и воскресенье, 10 штук, а в 2024 – 9 выходных, причем есть нерабочие понедельник и вторник, а также рабочая суббота. Естественно, рабочие понедельник и вторник ломали классическую сезонность и без дополнительных факторов точное прогнозирование здесь невозможно, ввиду чего без фактора праздников прогноз значительно отличался от фактических данных.

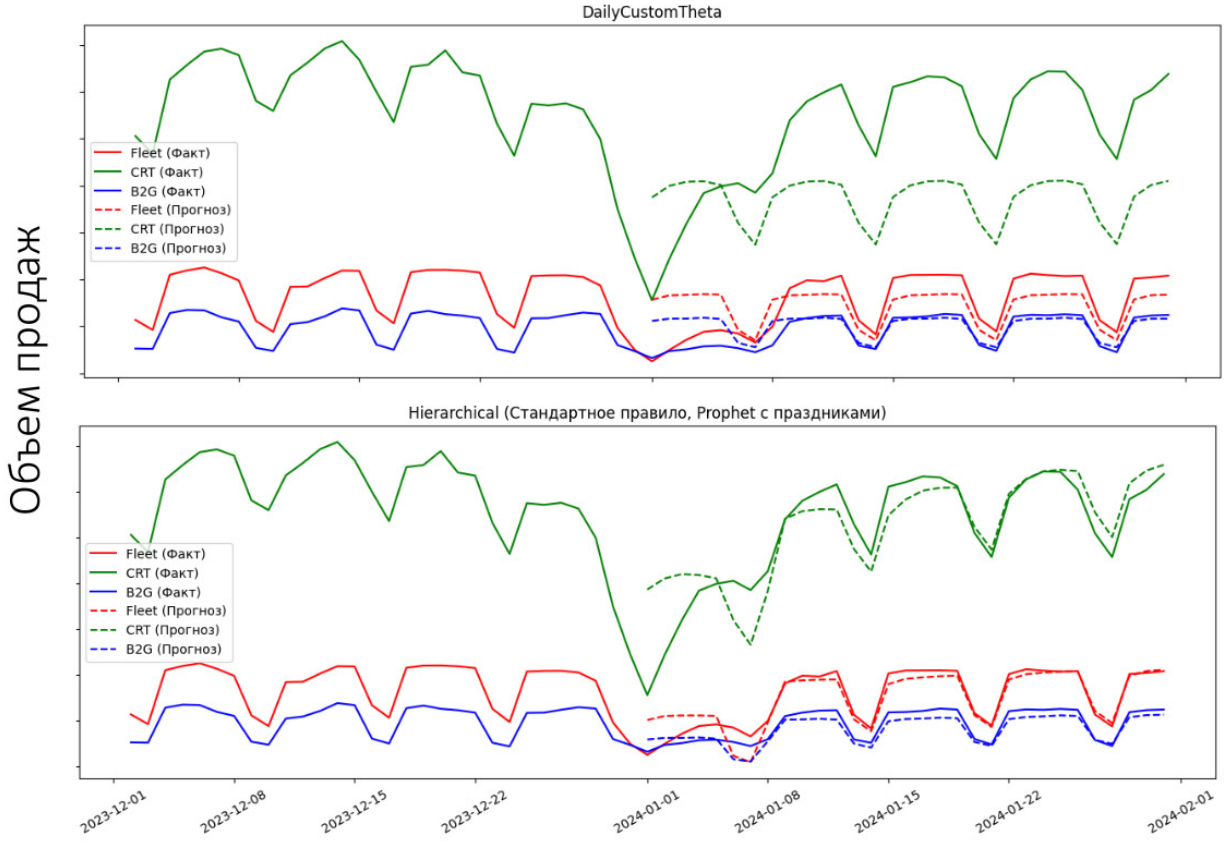

Рисунок 3.4 – Различие в динамике с учетом фактора выходных дней, конфиденциальная информация скрыта

Если ссылаться на пример прогнозов во время январских праздников 2024 года (см. рисунок 3.4), видно, что потребление в окрестности 1 января сильно отличается от общего уровня временного ряда и продажи там значительно ниже. При этом заметно, как подстраивается модель Prophet при использовании этого регрессора (нижний график) и снижает ожидаемые продажи в этот период. Ввиду хороших результатов на обратном тестировании, описанном в разделе 3.1, фактор был встроен во все сценарные модели.

3.5.2 Ценовые факторы

Внутри данных компании есть дополнительная информация о различных ценах нефтепродуктов (розничная цена, цена конкурентов и прочее). По рекомендациям заказчика, эти факторы также дополнительно были про-

тестированы для учета влияния прогнозов. Сущность используемых факторов определена бизнес-логикой по советам коллег, которые занимаются неавтоматическими прогнозами. Примеры рассматриваемых ценовых факторов приведены в таблице 3.1.

Таблица 3.1 – Примеры рассматриваемых ценовых факторов

Фактор	**Простая разность**	**Относительное % отклонение**
Цена розницы относительно цены мелкого опта	Чистая Цена ГПН Розница− Чистая Цена ГПН Мелкий Опт	$\left(\frac{\text{Чистая Цена ГПН Розница}}{\text{Чистая Цена ГПН Мелкий Опт}} - 1\right) \times 100\%$
Цена ГПН относительно конкурентов для Розницы	Чистая Цена ГПН Розница− Чистая Цена конкурентов Розница	$\left(\frac{\text{Чистая Цена ГПН Розница}}{\text{Чистая Цена конкурентов Розница}} - 1\right) \times 100\%$
Цена ГПН относительно конкурентов для Мелкого опта	Чистая Цена ГПН Мелкий Опт− Чистая Цена конкурентов Мелкий Опт	$\left(\frac{\text{Чистая Цена ГПН Мелкий Опт}}{\text{Чистая Цена конкурентов Мелкий Опт}} - 1\right) \times 100\%$
Розничная цена относительно биржевой	Чистая Цена ГПН Розница− Цена биржа	Не использовалось

Несмотря на некоторые приросты в качестве прогнозов по отдельным Отделениям и Нефтепродуктам после проведения обратного тестирования, описанного в разделе 3.1, было решено не использовать фактор в лучшей версии сценарной модели, так как в общем прирост в качестве был незначительный.

Важно уточнить, что для использования фактора необходимо знать его значение еще и на предсказываемый период. Поэтому при необходимости реализовать встраивание этого фактора, нужно было бы учесть в качестве важной задачи и реализацию модели по предсказанию факторов. При тестировании же моделей информация о будущем значении факторов бралась из исторических данных, подразумевая, формально, идеальное их прогнозирование. Но даже с таким подходом ценовые компоненты не давали значительного прироста в качестве.

3.5.3 Температурные факторы

Заказчик в процессе разработки модели высказывал гипотезу о том, что погода также может влиять на продажи. Это имело под собой и логическое

обоснование – в плохую погоду потребитель менее вероятно захочет заправлять транспортное средство. Из доступных погодных факторов среди данных компании была доступна информация о температуре в различных городах на каждый день и ночь.

Исходя из этого, для подневных моделей раздела 3.3.2 проверялся фактор температуры в городе днем и ночью за конкретный день, а для помесячных моделей раздела 3.3.1 – медианная температура днем и ночью в конкретном городе за месяц.

На основе годового обратного тестирования было выяснено, что фактор не дает значительного прироста в точности, поэтому на текущий момент не используется в лучшей реализации прогнозирования.

3.5.4 Другие факторы

Кроме приведенных выше факторов, также рассматривались бинарные регрессоры о различных крупных событиях во внутренней и внешней политике. В частности, например, этим является эпидемия COVID и экстренное введение нерабочих дней, что значительно изменяло динамику общей сезонности в 2020 году.

3.5.5 Общие выводы по факторам

Как было сказано ранее, не каждый из тестируемых факторов даёт хороший прирост в точности модели. Чаще происходит совершенно наоборот – фактор либо не меняет точность, либо вовсе только ухудшает прогнозирование модели. Несмотря на это, работа по поиску эффективных факторов ведется и по сей день с целью уточнения сценарных моделей как минимум на отдельных сегментах.

Важно уточнить, что все факторы проверяются на годовом обратном те-

стировании, описанном в разделе 3.1, чтобы снизить влияние неопределенности и оценить стабильность модели. Но ввиду того, что некоторые факторы могут проявлять себя положительно только на определенных месяцах (например, влиять исключительно на сильно-праздничный январь), работа по уточнению уже проверенных факторов не остановлена. Напротив, реализация этих факторов встроена в общую архитектуру прогнозирования для оперативного изменения прогноза по запросу заказчика.

3.6 Високосный год

В 2024 году команда проекта также столкнулась с проблемой заниженных прогнозов относительно фактических данных в феврале. В ходе анализа предсказаний впоследствии, было выявлено, что различие в прогнозах составляло продажи примерно за один рабочий день.

Аналогичное поведение не было замечено на других месяцах, поэтому была выдвинута гипотеза об отличии прогнозов от ожидаемых из-за високосного года и удлиненного февраля в нём. Действительно, оказывается, ввиду того, что сезонность в помесячных моделях раздела 3.3.1 учитывает аналогичные периоды в прошлом (то есть февраль – 2023, февраль – 2022 и т.д.), а история данных начинается с 2019 года согласно разделу 2.4, то для учёта корректной сезонности для високосных лет просто не хватает достаточного количества данных. Это непосредственно влияет на то, что февраль 2024 года подстраивается под «невисокосную» динамику и снижает прогнозы как раз на один рабочий день.

Чтобы избежать этого негативного эффекта, было принято простое, но эффективное решение о пост-корректировке прогноза через умножение месячного объема на $\frac{29}{28}$, чтобы добавить недостающий день високосного года. Предполагается, что этот подход не сможет себя оправдать только в 2032 го-

ду, когда 29 февраля выпадет на выходной день, а значит его объемы продаж будут значительно отличаться от продаж за обычный день. Это повлечет за собой необходимость пересмотра пост-корректировки прогноза на более универсальный метод.

Обратное тестирование подхода показало увеличение качества на любых сегментах и планах.

4 РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

На основе добавляемых изменений в модель формируется так называемое «лучшее правило» в разрезе Сегмент – Нефтепродукт (иногда и в разделе Отделение). Для этого проводится «глобальное обратное тестирование» и на основе его результатов определяется нужная модель в нужном разрезе.

1. Начало обратного тестирования – 2022-11-01;
2. Конец обратного тестирования – 2023-11-01;
3. Начало доступной истории – 2019-01-01;
4. Конец доступной истории – 2023-11-01;
5. Замеряемая ошибка – MAPE;
6. Планы для тестирования: 1, 2, 3, 4, 5;
7. Положительные планы для тестирования: +5, +15, +25;
8. Количество дней неполного месяца (при тестировании отдельных сценариев): 5, 15, 25;
9. Веса в модели дополнения неполного месяца: 1.0/0.0, 0.75/0.25, 0.5/0.5, 0.25/0.75, 0.0/1.0;
10. Тестируемые базовые модели: CustomTheta, CustomProphet
11. Тестируемые модели с дополнением месяца или разбивки: CustomTheta + DailyCustomTheta, CustomTheta + DailyCustomProphet, CustomProphet + DailyCustomTheta, CustomProphet + DailyCustomProphet;
12. Прочие параметры отдельных моделей.

Подобное тестирование проводится примерно раз в квартал с целью корректировки лучшего правила под накопленные изменения в структуре прогнозирования.

На текущий момент выработано оптимальное правило, дающее оптимальные прогнозы на имеющихся данных. Оно приведено ниже.

1. **Сегмент B2G, Нефтепродукт ГАЗ на любых планах:**

Иерархическая сценарная модель от MonthSum(DailyCustomTheta) с разбивкой по DailyCustomProphet с праздниками;

2. **Сегмент Fleet, Нефтепродукт ДТ (кроме отделений Тюмень и Челябинск) на любых планах:**

Иерархическая сценарная модель от CustomProphet с разбивкой по DailyCustomProphet с праздниками;

3. **Остальные случаи:**

Иерархическая сценарная модель от CustomTheta с разбивкой по DailyCustomProphet с праздниками;

Это оптимальное правило дает лучшие результаты по дневному прогнозированию. Причем важно то, что непосредственно подневная модель используется только в разрезе B2G – Газ, а потом результаты суммируются до месяца (через вспомогательную модель MonthSum). В остальных же случаях лучшее значение объема по месяцу дают именно помесячные модели, а прогноз подневно получается через иерархическую разбивку по лучшей подневной модели – DailyCustomProphet. Любая упомянутая выше модель, основанная на Prophet, также учитывает количество выходных и праздничных дней в месяце для лучшего результата согласно разделу 3.5.1.

Также замечено, что если в текущем месяце известно малое количество дней (например, 5), то использовать эту информацию для корректировки прогноза через модель дополнения неполного месяца нецелесообразно. Чаще всего в таких ситуациях подневное прогнозирование ухудшает будущий прогноз. Тем не менее, все сценарные опции сохранены в отдельных фабриках модели (структура данных, позволяющая вызывать необходимую модель исходя из параметров ряда) и могут быть вызваны в зависимости от ситуации на рынке и запроса заказчика.

5 АРХИТЕКТУРА ПРОЕКТА

Изначально проект развивался из так называемой Ad – Hoc задачи, то есть задачи, которая требует быстрого и индивидуального подхода, а также оперативного исправления и тестирования гипотез при возникновении соответствующих требований у заказчика.

Ввиду быстрого тестирования и развития проекта в рамках выделенных задач, вопрос о корректности выстраиваемой архитектуры поднялся лишь через полгода. В это время особо остро встал вопрос масштабируемости и скорости тестирования гипотез, однако архитектура проекта была негибкой в нужной степени и требовала значительной автоматизации.

Исходя из этого, был согласован так называемый ежеквартальный «технический спринт», который не включал тестирование никаких бизнес-гипотез и основывался только на создании гибкой архитектуры и выполнения технических задач из запланированного списка. Это позволило в достаточно кратчайшие сроки перейти с Ad – Hoc формата на полноценный формат действующего проекта, актуальный по сей день.

5.1 Текущая архитектура

На рисунке 5.1 приведена упрощенная схема текущей архитектуры проекта. Основная задача, которая стояла в процессе разработки – разделить все прогнозирование на независимые блоки, которые действуют независимо друг от друга и могут быть расширены через новые классы при необходимости добавить новую функциональность.

Исходя из этого, в процессе обсуждения были архитектурно проработаны и в итоге разработаны следующие блоки с соответствующими функциями и особенностями:

1. **Блок данных (data):**

Включает в себя три части (Загрузчик, Адаптер, Контейнер). Контейнер – сущность, которая хранит полную информацию про конкретный временной ряд (Департамент, Продукт, Сегмент, Тип отделения), его разрез (дневной, помесячный) и его историю продаж. Загрузчик – выгружает данные нужного формата (.parquet, .xlsx, прочее) и хранилища (сырые файлы, БД и прочее). Алаптер – класс, взаимодействующий с Загрузчиком для превращения сырых выгруженных данных в данные нужного разреза (например, сжимает подневные данные до месячного формата, используя необходимую обработку).

2. **Модели (models):**

Включает в себя реализацию базовых моделей, описанных общим протоколом класса и создает соответствующие фабрики моделей на их основе, то есть правила выбора модели в зависимости от конкретных характеристик ряда. Пример оптимальной фабрики приведен в разделе 4.

3. **Конфигурационные файлы (config):**

Все внешние файлы, константы и значения, которые могут использоваться в одном из блоков.

4. **Прогнозирование (forecasting):**

Блок, позволяющий динамически формировать батчи временных рядов для их параллельной обработки на соответствующих моделях и необходимом разрезе.

5. **Обратное тестирование (experiments):**

Надстройка над блоком прогнозирования, реализующая расширяющееся окно для обратного тестирования моделей на исторических данных.

6. **БД (team folder, sandbox):**

Локально хранимые файлы и внешнее Озеро Данных.

Помимо вышеперечисленных блоков, для автоматизации работы реа-

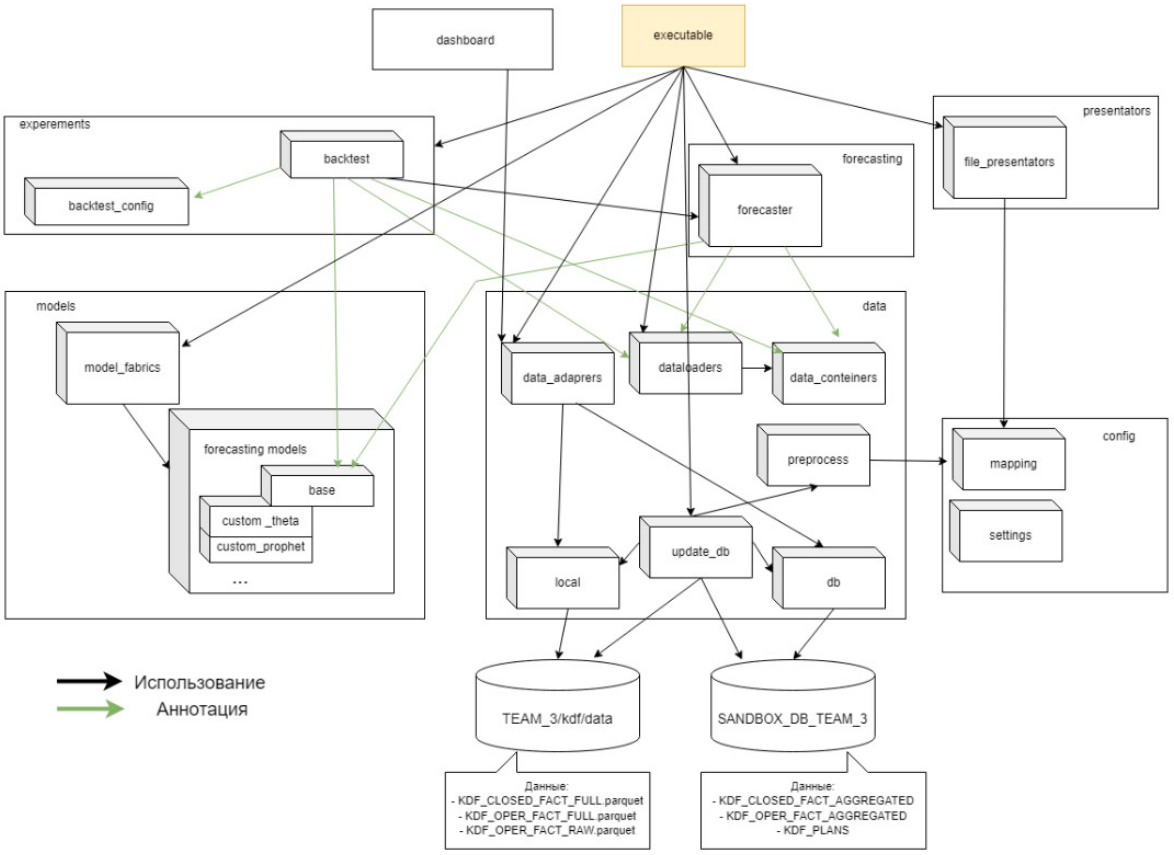

Рисунок 5.1 – Упрощенная архитектура проекта

лизован ряд исполняемых файлов:

1. Обновление и агрегация данных;
2. Прогнозирование нужного плана;
3. Обратное тестирование;
4. Развертывание дашборда;
5. Подбор гиперпараметров.

Описанная выше архитектура позволила значительно увеличить скорость тестирования масштабных гипотез, а также увеличить производительность проекта в целом, так как были локализованы и переписаны узкие места с точки зрения кода.

5.2 Дашборд

Кроме этого, согласно запросу заказчика был реализован дашборд. Дашборд – это информационная панель, отражающая динамику текущих данных из разных источников, а также дает возможность получать необходимые прогнозы без непосредственного участия команды разработки.

В качестве фреймворка был выбран Streamlit ввиду скорости создания страниц в нем [20]. Пример взаимодействия с дашбордом показан на рисунках 5.2, 2.1.

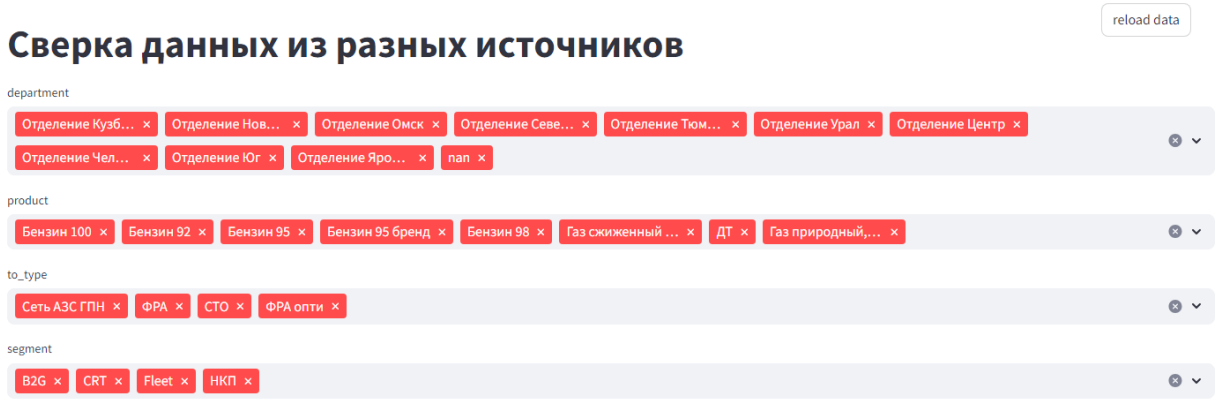

Рисунок 5.2 – Пример взаимодействия с дашбордом

В дашборде доступна следующая функциональность:

1. Выгрузка и сравнение ценовых факторов;
2. Анализ проведенных обратных тестирований между собой;
3. Просмотр выгруженных подневных данных;
4. Сравнение данных из разных источников;
5. Получение любых стандартных прогнозов на любой план;
6. Сравнение отчетов о прогнозах;
7. Анализ сезонности по сегментам.

Помимо очевидных преимуществ в виде пользы заказчику, дашборд

также помогает команде разработки следить за обновлением данных и случайно не пропустить какие-то резкие изменения на рынке.

Кроме этого, периодически происходят ошибки с третьей стороны при выгрузке данных, что сразу видно из-за резкого снижения объемов в определенном сегменте или отделении. В такой ситуации возникает срочная задача, требующая быстрого реагирования, так как проблема влияет на текущие эксперименты ввиду нерепрезентативности данных.

Пример подобной ошибки показан на рисунке 5.3.

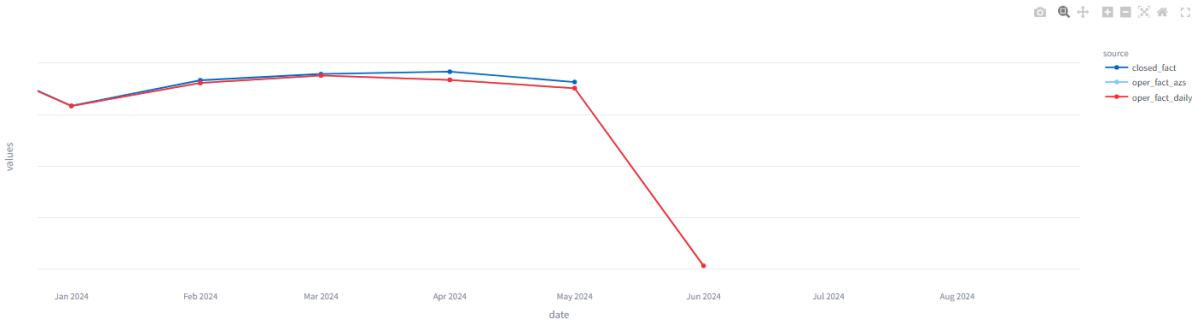

Рисунок 5.3 – Резкое изменение динамики ряда

6 БУДУЩЕЕ ПРОЕКТА

6.1 Дальнейшее развитие проекта

Уже на данный момент проект показывает значительно более высокое качество прогнозирование по дальним планам – 60 и 90, из-за чего был введен их автоматический приём. Кроме этого, изначальный список сегментов расширился: к CV сегменту был добавлен сегмент НКП (отдельные юридические лица).

В качестве дальнейших целей развития проекта можно выделить следующее:

1. Интеграция во внешний блок прогнозирования FIRST;
2. Расширение сегментов и доступных данных;
3. Автоматическое тестирование;

Таким образом, проект показывает отличные результаты и удовлетворяет запросы заказчика, а ближайшие задачи в основном сосредоточены на проектировании архитектуры, позволяющей встраиваться во внешнюю систему.

6.2 Снижение потребления памяти

Неизбежное увеличение количества данных и необходимость оценки затрачиваемых ресурсов для интеграции в FIRST потребовало оптимизации используемой оперативной памяти внутри прогнозируемого модуля.

6.2.1 Категориальные переменные

Одним из узких мест, которое было найдено в процессе работы над проектом, оказалось неэффективное хранение данных. Несмотря на то, что данные хранятся в сжатом формате – parquet, их формирование, разархивация

и прочие действия все равно сильно замедленны в производительности [21], так как информации очень много.

Было замечено, что существует более оптимальный метод хранения строковых значений в массиве (пример массива – в таблице 2.1). Это связано с тем, что многие колонки в массиве (Департамент, Тип отделения, Тип продукта, Сегмент) хранят незначительное количество уникальных элементов и эффективнее закодировать эти уникальные значения числами, где каждое число соответствует уникальному элементу столбца.

Концепция такого хранения реализована в рамках такого типа данных, как category (категория) в Python [22]. Проект был дополнен этапом перевода всех возможных данных в этот формат, чтобы сэкономить на использовании оперативной памяти. Результаты замеров нововведения отражены в следующем подразделе.

6.2.2 Мемопрофилирование

Мемопрофилирование — это процесс анализа использования памяти программой, особенно в области программирования и разработки программного обеспечения. Этот процесс включает в себя идентификацию и оптимизацию частей кода, которые неэффективно используют память, что может привести к утечкам памяти, снижению производительности или другим проблемам [23].

Мемопрофилирование помогает разработчикам улучшать эффективность и стабильность программных продуктов, обеспечивая более рациональное распределение ресурсов памяти и улучшение общей производительности приложений.

Внедрение категориальных переменных привело к значительному изменению профиля потребления памяти. Если сравнивать исходный стандарт-

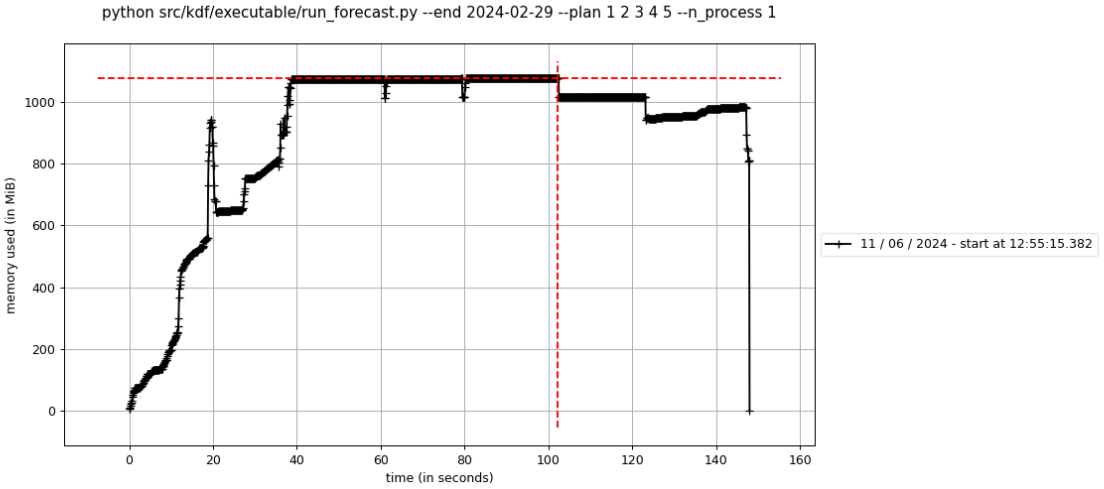

Рисунок 6.1 – Исходное потребление памяти

ный прогноз на 5 планов (рисунок 6.1) с обновленным (рисунок 6.2), заметно снижение общего потребления оперативной памяти на 20%, снижение пика потребления на 20%, а также ускорение прогнозирования на 38%.

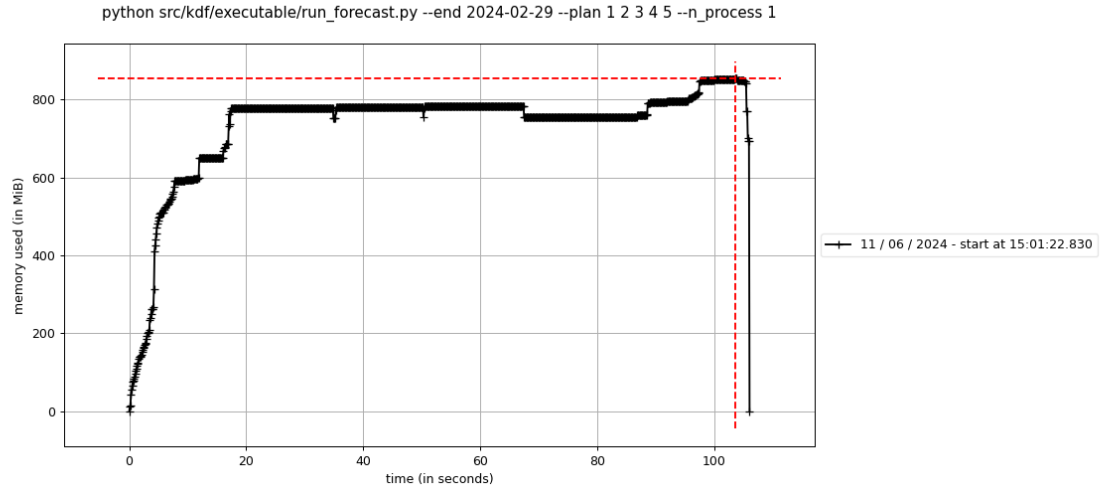

Рисунок 6.2 – Обновленное потребление памяти

6.3 Ускорение прогнозирования

Как было сказано в предыдущем разделе, использование категориальных переменных позволяет ускорить прогноз (в зависимости от типа прогноза) на 20-30%.

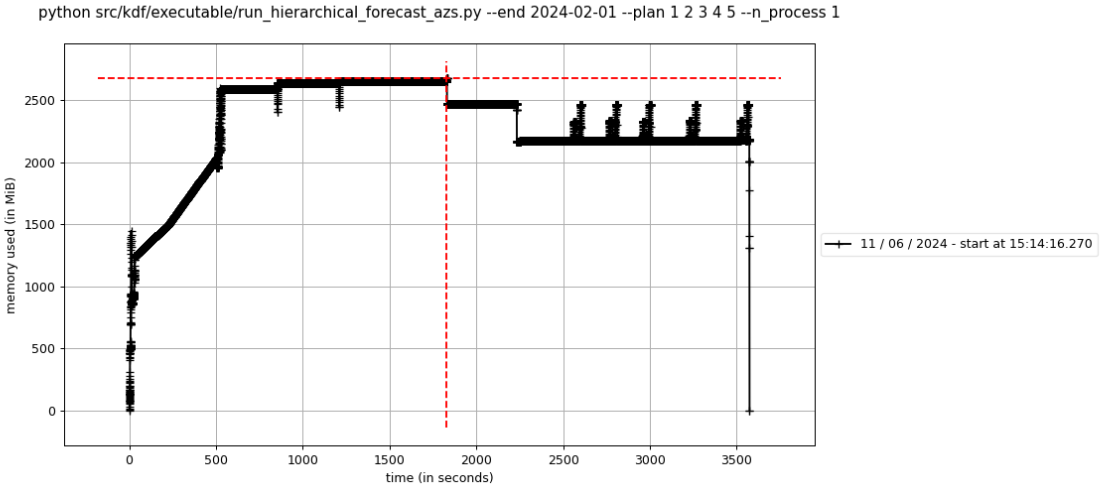

Рисунок 6.3 – Исходное потребление памяти в разрезе АЗС

Однако основным источником потребеления оперативной памяти остаётся выгрузка результатов в удобный для заказчика формат – Excel. Это особенно видно при мемопрофилировании прогнозов в разрезе АЗС. Исходное потребление (рисунок 6.3) и обновленное потребление (рисунок 6.4) отличаются после введения категориальных переменных значительно: общее потребление оперативной памяти упало на 40%, время прогноза сократилось на 30%. Однако 5 пиков (5 этапов сохранения каждого плана) остались без изменения по количеству потребляемой памяти.

Поэтому для дальнейшего ускорения в рамках задачи определено тестирование открытых пакетов для ускоренной сборки Excel – файлов, например, через PyExcelerate [24].

Тем не менее, для ускорения прогнозов на текущий момент уже применяются некоторые подходы:

1. **Кэширование данных**

Выгрузка и агрегация данных в нужном формате через Загрузчик и Адаптер требует некоторого количества времени, что мешает производить быстрые эксперименты независимо друг от друга. Однако ввиду того, что выгрузка данных не изменяется в течение дня, было принято кэшировать каж-

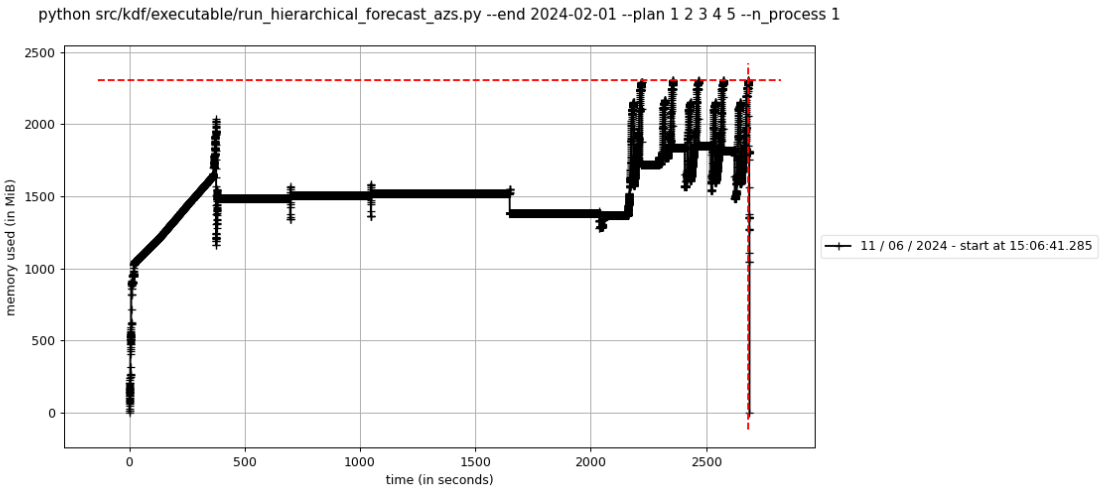

Рисунок 6.4 – Обновленное потребление памяти в разрезе АЗС

дый результат Адаптера и при необходимости догружать только новые данные.

Важно уточнить, что при таком подходе неожиданно могут возникнуть расхождения в данных, например, если значения Оперативного Факта обновятся за последние два месяца, а эти данные уже загружены в кэш. Чтобы избежать этого, была добавлена автоматизация удаления устаревшего кэша раз в 3 дня.

2. **Распараллеливание прогнозирования**

Использование сценарных моделей требует большого количества времени, особенно для проведения обратного тестирования на медленных моделях типа Prophet. Для ускорения процесса, прогнозирование осуществляется через запуск дополнительных процессов с помощью пакета multiprocessing. Это возможно ввиду того, что каждый временной ряд прогнозируется независимо друг от друга.

Заключение

В данном проекте были успешно реализованы и апробированы модели сценарного прогнозирования, ориентированные на оптимизацию процессов поставки топлива на АЗС. Несмотря на кажущуюся простоту работы статистических моделей, на их основе можно разработать сложные и мощные системы прогнозирования. Использование базовых моделей ThetaModel и Prophet позволило получить несколько сценариев прогнозирования (подневные и помесячные модели, дополнение неполного месяца, иерархическая разбивка), которые демонстрируют устойчивую точность и адаптивность к изменениям на рынке.

Дополнительно, в проекте был особо выделен учет внешних факторов, таких как выходные и праздничные дни, ценовые регрессоры и температурные условия в городах. Некоторые из этих факторов оказывают существенное влияние на предсказательные способности моделей. Их интеграция в сценарные модели прогнозирования позволила повысить точность и адаптивность, а также добавила больше возможностей при реагировании на изменения на рынке для заказчика.

Архитектура проекта была разработана с учетом модульности и масштабируемости, что обеспечивает гибкость при внедрении новых функций и удобство в управлении данными. Это включает в себя использование современных технологий обработки и хранения данных, а также оптимизацию обучения и прогнозирования.

Важным аспектом проекта является внедрение автоматизированной системы для оперативного обновления данных и получения прогнозов, что позволило уже на текущем этапе получить одобрение заказчика на автоматический прием долгосрочных планов на 60 и 90 дней, а также расширить имеющиеся данные новыми сегментами НКП.

Список использованных источников

1. Time series forecasting: applications to the upstream oil and gas supply chain / L. B. Sheremetov, A. González-Sánchez, I. López-Yáñez, A. V. Ponomarev // IFAC Proceedings Volumes. — 2013. — Т. 46, № 9. — С. 957—962.

2. Cerqueira V., Torgo L., Mozetič I. Evaluating time series forecasting models: An empirical study on performance estimation methods // Machine Learning. — 2020. — Т. 109, № 11. — С. 1997—2028.

3. Shumway R. H., Stoffer D. S., Stoffer D. S. Time series analysis and its applications. Т. 3. — Springer, 2000.

4. Chatfield C. Time-series forecasting. — Chapman, Hall/CRC, 2000.

5. Chatfield C. The analysis of time series: theory and practice. — Springer, 2013.

6. ARIMA models / R. H. Shumway, D. S. Stoffer, R. H. Shumway, D. S. Stoffer // Time series analysis and its applications: with R examples. — 2017. — С. 75—163.

7. Webby R., O'Connor M. Judgemental and statistical time series forecasting: a review of the literature // International Journal of forecasting. — 1996. — Т. 12, № 1. — С. 91—118.

8. Shmueli G., Polak J. Practical time series forecasting with R: A hands-on guide. — Axelrod schnall publishers, 2024.

9. Yaffee R. A., McGee M. An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®. — Elsevier, 2000.

10. Connor J. T., Martin R. D., Atlas L. E. Recurrent neural networks and robust time series prediction // IEEE transactions on neural networks. — 1994. — Т. 5, № 2. — С. 240—254.

11. Boosted embeddings for time-series forecasting / S. R. Karingula, N. Ramanan, R. Tahmasbi, M. Amjadi, D. Jung, R. Si, C. Thimmisetty, L. F. Polania, M. Sayer, J. Taylor // International Conference on Machine Learning, Optimization, and Data Science. — Springer. 2021. — С. 1—14.

12. Campbell S. D. A review of backtesting and backtesting procedures. — 2005.

13. Spiliotis E., Assimakopoulos V., Makridakis S. Generalizing the theta method for automatic forecasting // European Journal of Operational Research. — 2020. — Т. 284, № 2. — С. 550—558.

14. Makridakis S., Hibon M. The M3-Competition: results, conclusions and implications // International journal of forecasting. — 2000. — Т. 16, № 4. — С. 451—476.

15. Assimakopoulos V., Nikolopoulos K. The theta model: a decomposition approach to forecasting // International journal of forecasting. — 2000. — Т. 16, № 4. — С. 521—530.

16. The optimised theta method / J. A. Fioruci, T. R. Pellegrini, F. Louzada, F. Petropoulos // arXiv preprint arXiv:1503.03529. — 2015.

17. Alharbi F. R., Csala D. A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach // Inventions. — 2022. — Т. 7, № 4. — С. 94.

18. Taylor S. J., Letham B. Forecasting at scale // The American Statistician. — 2018. — Т. 72, № 1. — С. 37—45.

19. Rafferty G. Forecasting Time Series Data with Facebook Prophet: Build, improve, and optimize time series forecasting models using the advanced forecasting tool. — Packt Publishing Ltd, 2021.

20. DeFalco D. J., DeFalco J. A. First measure everything: engineering trends in data visualization // Design Recommendations for Intelligent Tutoring Systems. — 2020. — Т. 8. — С. 47.

21. Rey A., Freitag M., Neumann T. Seamless Integration of Parquet Files into Data Processing. — 2023.

22. Heydt M. Learning pandas. — Packt Publishing Ltd, 2017.

23. Asking and Answering Questions During Memory Profiling / A. F. Blanco, A. Q. Córdova, A. Bergel, J. P. S. Alcocer // IEEE Transactions on Software Engineering. — 2024.

24. Zimmerman K. PyExcelerate [Электронный ресурс]. — 2023. — https://github.com/kz26/PyExcelerate (Дата обращения: 03.06.2024).