Geekbrains

Исследование особенностей прогнозирования временных рядов на примере платёжного календаря Иркутской нефтяной компании

Программа:
Разработчик — Искусственный интеллект.
Климов Александр Сергеевич

Иркутск
2025

Оглавление

Введение

Данная работа посвящена решению прикладной задачи на реальных данных и процессах Группы компаний Иркутская нефтяная компания, а именно – изучение возможностей повышения точности планирования расходов группы компаний с целью повышения эффективности управления денежными средствами и снижения риска возникновения кассового разрыва.

В основу работы легла гипотеза о том, что применение машинного обучения для анализа временных рядов, позволит с большей точностью строить платёжный календарь. Модели машинного обучения могут использоваться в качестве альтернативного плана платежей для выявления периодов платёжного календаря, для которых вероятность ошибки наиболее высока с целью корректировки плана и выявления системных ошибок и недостатков существующего процесса.

Актуальность темы обоснована экспертной оценкой размера убытков от неэффективного отвлечения денежных средств из оборота, обслуживание кредитов и займов, для недопущения кассовых разрывов. Эти суммы измеряются десятками миллионов рублей ежегодно.

Большая часть данных, используемых в данной работе, является конфиденциальными и представляют коммерческую тайну. Однако представленные в работе оценки эффективности и отклонений полностью соответствуют реальным данным.

Целью проекта является повышение точности прогноза выбытий денежных средств группы компаний Иркутская нефтяная компания.

Достижение цели проекта осуществляется через решение задачи о повышении точности планирования выбытий для воздействия на два основных фактора снижающих эффективность управления финансами группы компаний, а именно:

- Чрезмерное резервирование денежных средств под завышенные планы расходов
- Несвоевременное прогнозирование кассовых разрывов

Повышение точности планирования выбытий денежных средств достигается за счёт внедрения моделей машинного обучения и делегирования им функции прогнозирования временных рядов (Далее ВР) выбытий денежных средств (Далее ДС).

Формулировка и конкретизация цели проекта по SMART

Цель:

Уменьшить среднеквадратичное отклонение между прогнозными и фактическими значениями выбытий денежных средств на 40%.

S – Уменьшаем среднеквадратичное отклонение между прогнозными и фактическими значениями выбытий ДС

M – 40% снижение отклонения считается хорошим результатом, 60% - превосходным. Оценка на основе среднеквадратичного отклонения позволяет сравнивать эффективность как действующего процесса, так и эффективность моделей прогнозирования. Более того, такой подход оценки точности учитывает наличие крупных ошибок, что является полезным свойством при работе со снижением рисков кассового разрыва.

A – Существующие методики не только не позволяют повысить точность прогнозирования, наоборот - наблюдается увеличение отклонения с каждым кварталом, это связано с многофакторностью взаимосвязей и отсутствия комплексного подхода к решению проблемы, поэтому внедрение машинного обучения будет серьёзным шагом на встречу оптимизации, однако из-за многофакторности нет однозначной уверенности что решить проблему удастся в одним лишь внедрением прогнозной модели. 40% снижение кажется реалистичным показателем, который можно достичь на основе изменения алгоритмов прогнозирования и внедрения машинного обучения

R – Решение данной проблемы позволит значительно повысить эффективность управления капиталом, от чего зависит около 40% прибыли компании в рамках долгосрочных инвестиционных циклов и существующих ограничений на объём заимствований.

T – трёх месяцев будет достаточно, чтобы собрать данные и создать датасет, а так же обучить несколько моделей машинного обучения и выбрать оптимальную модель.

Основные этапы.

В ходе работы над проектом мной реализованы несколько ключевых мероприятий:

- Консолидированы данные о финансах группы компаний за последние 9 лет
- Разработан датасет платежей и финансовых обязательств группы компаний
- Исследованы временные ряды платежей по текущей, инвестиционной и финансовой деятельности компании
- Проведён сравнительный анализ эффективности применения моделей машинного обучения для прогнозирования временного ряда выбытий группы компаний.

Обзор используемого инструментария

1. **Библиотеки анализа данных:**
   - **Pandas**: Используется для манипуляции и анализа данных, включая обработку временных рядов.
   - **NumPy**: Обеспечивает поддержку многомерных массивов и матричных операций, что полезно для математических вычислений.
   - **Scikit-learn**: Предоставляет широкий спектр алгоритмов машинного обучения для классификации, регрессии и кластеризации.
   - **TensorFlow/Keras**: Используется для построения и обучения нейронных сетей.
2. **Инструменты визуализации:**
   - **Matplotlib**: Основной инструмент для создания графиков и визуализации данных.
   - **Seaborn**: Расширяет возможности Matplotlib, позволяя создавать более сложные визуализации.
   - **Plotly**: Используется для создания интерактивных графиков.
3. **Среда разработки, ЯП:**
   - **Jupyter Notebook:** Является одновременно и средой разработки (Python 3.13/3.11.9) и средой документирования кода и инструментом визуализации результатов исполнения кода, обеспечивающей совместный доступ к данным для группы заинтересованных лиц
   - **SQL:** использование библиотеки pyodbc для подключения и направления SQL запросов к SQL серверу первичных данных.

Состав команды:

Климов Александр Сергеевич - разработчик, аналитик данных, специалист по машинному обучению. Задачи: сбор и обработка данных, создание датасета, анализ временных рядов, обучение и тестирование моделей машинного обучения

Усова Екатерина Сергеевна - финансовый аналитик, казначей. Задачи: конкретизация целей и задач, верификация полученных результатов.

Глава 1.

Теоретические основы прогнозирования временных рядов

1.1. Понятие и классификация временных рядов:

Временной ряд — это последовательность наблюдений, упорядоченная по времени. Каждое наблюдение в этой последовательности называется уровнем временного ряда и представляет собой значение определённой переменной в конкретный момент времени. Временные ряды широко используются в различных областях, таких как экономика, финансы, метеорология, медицина и т. д., для анализа и прогнозирования будущих тенденций.

Временной ряд может состоять из нескольких компонентов, которые определяют его структуру и поведение. К основным компонентам временного ряда относятся:

- Тренд (T) — это долгосрочная тенденция изменения уровня временного ряда. Тренд может быть возрастающим, убывающим или постоянным.
- Сезонность (S) — это периодические колебания уровня временного ряда, которые повторяются через определённые промежутки времени. Сезонность может быть вызвана различными факторами, такими как погодные условия, праздники, сезонные изменения спроса и т. д.
- Цикличность (C) — это более длительные колебания уровня временного ряда, которые не имеют строго периодического характера. Цикличность может быть связана с экономическими циклами, политическими событиями и другими факторами.
- Случайные колебания (E) — это нерегулярные и непредсказуемые изменения уровня временного ряда, которые не могут быть объяснены другими компонентами. Случайные колебания могут быть вызваны ошибками измерения, непредвиденными событиями и другими факторами.

1.1.1 Классификация временных рядов.

Временные ряды можно классифицировать по различным признакам. Рассмотрим основные классификации временных рядов:

1. Классификация по частоте наблюдений:
    - Дискретные временные ряды — это ряды, в которых наблюдения проводятся через равные промежутки времени (ежедневно, ежемесячно, ежегодно).

- Непрерывные временные ряды — это ряды, в которых наблюдения проводятся непрерывно в течение определённого периода времени.

2. По количеству переменных:
   - Одномерные временные ряды — это ряды, которые содержат наблюдения одной переменной (например, курс валюты, температура воздуха).
   - Многомерные временные ряды — это ряды, которые содержат наблюдения нескольких переменных (например, курс нескольких валют, температура и влажность воздуха).

3. По характеру данных:
   - Количественные временные ряды — это ряды, в которых данные представлены в виде чисел (например, объём продаж, цена акции).
   - Качественные временные ряды — это ряды, в которых данные представлены в виде категорий или классов (например, уровень удовлетворённости клиентов, рейтинг продукта).

4. По наличию тренда:
   - Стационарные временные ряды — это ряды, в которых среднее значение и дисперсия остаются постоянными во времени.
   - Нестационарные временные ряды — это ряды, в которых среднее значение и/или дисперсия изменяются во времени.

5. По наличию сезонности:
   - Сезонные временные ряды — это ряды, в которых наблюдаются периодические колебания уровня временного ряда.
   - Несезонные временные ряды — это ряды, в которых не наблюдается сезонности.

6. По наличию цикличности:
   - Циклические временные ряды — это ряды, в которых наблюдаются более длительные колебания уровня временного ряда, не имеющие строго периодического характера. Ациклические временные ряды — это ряды, в которых не наблюдается цикличности.

1.2 Методы анализа временных рядов.

Для анализа временных рядов используются различные методы, которые позволяют выявить основные компоненты временного ряда, оценить их влияние на уровень ряда и прогнозировать будущие тенденции. К основным методам анализа временных рядов относятся:

Метод скользящего среднего — это метод, который позволяет сгладить случайные колебания временного ряда и выявить тренд.

Метод экспоненциального сглаживания — это метод, который позволяет учитывать более свежие данные с большим весом и выявить тренд и сезонность.

Метод разложения временного ряда на компоненты — это метод, который позволяет выделить тренд, сезонность и случайные колебания из временного ряда.

Метод авторегрессии и скользящего среднего (ARIMA) — это метод, который позволяет моделировать временные ряды с учётом тренда, сезонности и случайных колебаний. Выбор метода анализа временного ряда зависит от его характеристик и целей исследования. Например, для прогнозирования будущих тенденций может быть использован метод ARIMA, а для выявления сезонности — метод разложения временного ряда на компоненты.

Метод регрессии: Используются для анализа зависимости временного ряда от других переменных. В том числе с помощью моделей регрессионного анализа, таких как случайные леса, XGBoost, нейронные сети и другие алгоритмы, которые могут быть применены для прогнозирования временных рядов.

Кросс-корреляционный анализ: Используется для изучения взаимосвязей между несколькими временными рядами.

1.3 Особенности прогнозирования экономических временных рядов

Прогнозирование экономических временных рядов представляет собой сложный и многогранный процесс, который требует учета специфических характеристик экономических данных. В отличие от других типов временных рядов, экономические временные ряды часто подвержены влиянию множества факторов, таких как сезонные колебания, экономические циклы, политические события и изменения в законодательстве. В этом разделе мы рассмотрим ключевые особенности, которые необходимо учитывать при прогнозировании экономических временных рядов.

Сезонность и цикличность

Одной из основных характеристик экономических временных рядов является наличие сезонных и циклических колебаний. Сезонные колебания могут быть связаны с определенными периодами года, когда спрос на товары и услуги изменяется (например, увеличение продаж в преддверии праздников). Циклические колебания, в свою очередь, отражают более длительные экономические циклы, такие как подъемы и спады в экономике. При прогнозировании необходимо использовать методы, которые позволяют выделять и корректировать эти колебания, такие как декомпозиция временных рядов и использование сезонных индексов.

Влияние внешних факторов

Экономические временные ряды часто подвержены влиянию внешних факторов, таких как изменения в экономической политике, колебания валютных курсов, инфляция и другие макроэкономические показатели. Эти факторы могут значительно влиять на динамику временного ряда и должны быть учтены при построении прогнозов. Для этого могут использоваться регрессионные модели, которые позволяют включать в анализ дополнительные переменные, влияющие на целевой показатель.

Нелинейность и асимметрия

Экономические временные ряды могут демонстрировать нелинейные зависимости и асимметрию, что делает их прогнозирование более сложным. Например, в условиях экономического кризиса изменения в спросе могут быть более резкими и непредсказуемыми, чем в период стабильного роста. Для учета этих особенностей могут применяться модели, основанные на методах машинного обучения, такие как деревья решений, нейронные сети и другие алгоритмы, способные выявлять сложные зависимости в данных.

Долгосрочные и краткосрочные прогнозы

При прогнозировании экономических временных рядов важно различать долгосрочные и краткосрочные прогнозы. Долгосрочные прогнозы могут быть более неопределенными и подвержены большему количеству факторов, в то время как краткосрочные прогнозы могут быть более точными, но также требуют учета текущих тенденций и событий. Выбор подходящего метода прогнозирования зависит от горизонта прогнозирования и целей анализа.

Оценка точности прогнозов

Оценка точности прогнозов является важным этапом в процессе прогнозирования экономических временных рядов. Для этого используются различные метрики, такие как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (RMSE) и другие. Важно не только оценить точность модели на исторических данных, но и проводить тестирование на новых данных, чтобы убедиться в ее надежности и применимости.

Заключение

Прогнозирование экономических временных рядов требует комплексного подхода и учета множества факторов, влияющих на динамику данных. Успешное прогнозирование может значительно повысить эффективность принятия решений в бизнесе и экономике в целом. Важно использовать разнообразные методы и подходы, адаптируя их к специфике анализируемых данных и целям исследования.

Глава 2.

Анализ платёжного календаря как объекта прогнозирования

2.1 Понятие платёжного календаря и его роль в управлении финансами предприятия

Платёжный календарь является одним из важнейших инструментов управления финансами предприятия. Он представляет собой документ, в котором отражаются все предстоящие денежные потоки предприятия, включая доходы и расходы, с указанием сроков их осуществления. Платёжный календарь позволяет предприятию эффективно планировать свои финансовые ресурсы, оптимизировать денежные потоки и обеспечивать своевременное выполнение обязательств перед контрагентами и государством, прогнозируя и избегая кассовых разрывов и нарушений в исполнении финансовых обязательств.

2.1.1 Определение платёжного календаря

Платёжный календарь – это документ, который содержит информацию о предстоящих денежных потоках предприятия на определённый период времени. Он включает в себя следующие основные элементы:

Дата осуществления денежного потока: указывается дата, когда предприятие планирует получить доход или осуществить расход.

Сумма денежного потока: указывается сумма, которую предприятие планирует получить или потратить.

Контрагент: указывается наименование контрагента, с которым предприятие осуществляет сделку.

Цель денежного потока: указывается цель, для которой предприятие осуществляет денежный поток (например, оплата товаров, выплата заработной платы, уплата налогов и т. д.).

Платёжный календарь может быть составлен на различные периоды времени: на неделю, месяц, квартал или год. Выбор периода зависит от специфики деятельности предприятия и его финансовых целей.

2.1.2 Роль платёжного календаря в управлении финансами предприятия

Платёжный календарь играет важную роль в управлении финансами предприятия. Он позволяет планировать денежные потоки, учитывая предстоящие доходы и расходы. Это позволяет предприятию заранее подготовиться к возможным финансовым трудностям и принять меры по их предотвращению.

Кроме того, платёжный календарь позволяет предприятию оптимизировать свои денежные потоки, перераспределяя их во времени. Это может помочь предприятию избежать кассовых разрывов и обеспечить своевременное выполнение обязательств.

Обеспечивать своевременное выполнение обязательств: платёжный календарь помогает предприятию обеспечить своевременное выполнение своих обязательств перед контрагентами и государством. Это способствует укреплению деловой репутации предприятия и повышению его кредитоспособности.

Платёжный календарь позволяет предприятию контролировать свои финансовые ресурсы, отслеживая поступление и расходование денежных средств. Это помогает предприятию избежать нецелевого использования средств и обеспечить их эффективное использование, выстраивая процесс верификации платежей перед их непосредственной оплатой.

2.1.3 Принципы составления платёжного календаря

При составлении платёжного календаря необходимо соблюдать следующие принципы:

- Принцип реальности: платёжный календарь должен быть основан на реальных данных о предстоящих денежных потоках. Он должен учитывать все возможные доходы и расходы предприятия, а также изменения в экономической ситуации.
- Принцип точности: платёжный календарь должен быть составлен с максимальной точностью. Ошибки в составлении платёжного календаря могут привести к неправильному планированию денежных потоков и, как следствие, к финансовым трудностям.
- Принцип своевременности: платёжный календарь должен быть составлен своевременно, чтобы предприятие могло принять необходимые меры по оптимизации денежных потоков и обеспечению своевременного выполнения обязательств.
- Принцип гибкости: платёжный календарь должен быть гибким и допускать возможность внесения изменений в случае изменения экономической ситуации или возникновения непредвиденных обстоятельств, то есть быть основанным на изменяемых макропараметрах

2.1.4 Методы составления платёжного календаря

Существует несколько методов составления платёжного календаря:

1. Метод прямого счёта: основан на прямом подсчёте всех предстоящих денежных потоков предприятия. Этот метод является наиболее точным, но требует большого объёма информации и времени на составление.

2. Метод коэффициентов: основан на использовании коэффициентов, которые учитывают зависимость между различными видами денежных потоков. Этот метод позволяет сократить время на составление платёжного календаря, но может быть менее точным.
3. Метод экспертных оценок: основан на оценке предстоящих денежных потоков экспертами. Этот метод может быть использован в случае отсутствия достаточной информации для составления платёжного календаря другими методами.

Выбор метода составления платёжного календаря зависит от специфики деятельности предприятия, его финансовых целей и наличия необходимой информации. Организации, имеющие сложный состав поступлений и выбытий, или сложную структуру (Холдинги, группы компаний) применяют одновременно несколько подходов к построению платёжных календарей, например для краткосрочного планирования на 7-10 платёжных дней используется метод прямого счёта, для обеспечения точности прогноза, основанного на утверждённых графиках платежей по действующим контрактам. Для среднесрочного прогнозирования платежей (1-2 месяца), используются методы коэффициентов, учитывающие сезонность, данные аналогичных периодов в прошлом и данные о действующих контрактах с определёнными графиками платежей. Для построения прогноза на более длительные периоды, прибегают к экспертным оценкам изменения макропараметров, таких как индекс спроса, курсы валют и прочих, чтобы на основании этих параметров строить прогноз, который может производиться по нескольким сценариям: «Оптимистичному», подразумевающему комфортные макропараметры и «Пессимистичному», при котором макропараметры не благоприятны для компании.

2.1.5 Анализ платёжного календаря как объекта прогнозирования

Платёжный календарь является важным объектом прогнозирования, поскольку он позволяет предприятию предсказать свои будущие денежные потоки и принять необходимые меры по их оптимизации. Анализ платёжного календаря позволяет выявить возможные проблемы с денежными потоками и разработать меры по их решению, к примеру внешние заимствования, продажа активов, изъятие денежных средств из оборота.

Анализ платёжного календаря проводится с целью:

- Выявления возможных проблем с денежными потоками: анализ платёжного календаря позволяет выявить возможные кассовые разрывы, нехватку денежных средств для выполнения обязательств и другие проблемы.
- Определения потребности в дополнительных источниках финансирования: анализ платёжного календаря позволяет определить потребность предприятия в дополнительных источниках финансирования для покрытия дефицита денежных средств.

- Оценки эффективности управления денежными потоками: анализ платёжного календаря позволяет оценить эффективность управления денежными потоками предприятия и выявить возможности для их оптимизации.

2.1.6 Методы анализа платёжного календаря

Для анализа платёжного календаря используются следующие методы:

Горизонтальный анализ: позволяет сравнить показатели платёжного календаря за разные периоды времени. Это помогает выявить тенденции в изменении денежных потоков и определить возможные проблемы.

Вертикальный анализ: заключается в изучении структуры денежных потоков предприятия. Он позволяет определить долю каждого вида денежных потоков в общем объёме и выявить наиболее значимые источники доходов и расходов.

Коэффициентный анализ: основан на расчёте и анализе финансовых коэффициентов, которые характеризуют эффективность управления денежными потоками. Например, коэффициент текущей ликвидности показывает способность предприятия своевременно выполнять свои обязательства.

Факторный анализ: позволяет определить влияние различных факторов на изменение денежных потоков. Например, можно проанализировать влияние изменения объёмов продаж, цен на товары или процентных ставок на денежные потоки предприятия.

2.1.8 Этапы анализа платёжного календаря

Анализ платёжного календаря включает в себя следующие этапы:

1. Сбор и подготовка данных: на этом этапе собираются все необходимые данные о предстоящих денежных потоках предприятия. Данные должны быть точными и актуальными.

2. Расчёт показателей платёжного календаря: на этом этапе рассчитываются основные показатели платёжного календаря, такие как общий объём денежных потоков, структура денежных потоков, коэффициенты ликвидности и т. д.

3. Анализ показателей платёжного календаря: на этом этапе проводится анализ рассчитанных показателей с целью выявления возможных проблем и определения потребности в дополнительных источниках финансирования.

4. Разработка рекомендаций по оптимизации денежных потоков: на основе результатов анализа разрабатываются рекомендации по оптимизации денежных потоков предприятия. Например, можно предложить меры по ускорению поступления денежных средств или снижению расходов.

5. Мониторинг и контроль: после реализации рекомендаций необходимо осуществлять мониторинг и контроль за выполнением платёжного календаря. Это позволит своевременно выявлять отклонения от плана и принимать необходимые меры.

2.1.9 Практические аспекты использования платёжного календаря для прогнозирования

Платёжный календарь является не только инструментом для управления текущими денежными потоками, но и мощным инструментом для прогнозирования финансового состояния предприятия на будущее.

Для краткосрочного прогнозирования платёжный календарь позволяет:

- Оценить предстоящие доходы и расходы на ближайший период.
- Выявить возможные кассовые разрывы и принять меры по их предотвращению.
- Оптимизировать платежи для обеспечения ликвидности.

В контексте долгосрочного прогнозирования платёжный календарь может быть использован для:

- Планирования крупных инвестиций и финансовых проектов.
- Оценки потребности в дополнительном финансировании.
- Прогнозирования финансовых результатов и разработки финансовых стратегий.

2.1.10 Проблемы и ограничения платёжного календаря

Несмотря на все преимущества, платёжный календарь имеет свои ограничения и проблемы, которые необходимо учитывать при его использовании для прогнозирования. Наиболее чувствительным ограничением, является отсутствие точного представления о сроках окончания подрядных или строительных работ, приводящих к отсрочкам платежей и/или возникновению новых неучтённых трат. Курсовые разницы, для организаций, владеющих или использующих валютные денежные средства, представляют так же серьёзный фактор риска, так как резкие скачки курса валют к национальной валюте приводят к необходимости уплаты дополнительного налога на прибыль, без учёта ликвидности валютного актива. Нестабильность рыночных цен на продукцию предприятия, а так же основные средства предприятия, приводит к существенной разнице между доходами предприятия и расходами на инвестиционную деятельность, способную привести к кассовому разрыву. Риски и ограничения, можно классифицировать следующим образом:

• **Неопределённость будущих денежных потоков** - платёжный календарь основан на прогнозах и предположениях, которые могут меняться в зависимости от экономической ситуации, действий конкурентов, изменений в законодательстве и других факторов.

• **Необходимость постоянного обновления данных.** Для эффективного использования платёжного календаря необходимо постоянно обновлять данные о предстоящих денежных потоках. Это требует времени и ресурсов, а также системы для сбора и анализа данных.

Платёжный календарь является важным инструментом для управления финансами предприятия и прогнозирования его финансового состояния. Он позволяет планировать денежные потоки, оптимизировать платежи, обеспечивать своевременное выполнение обязательств и контролировать финансовые ресурсы. Однако для эффективного использования платёжного календаря необходимо учитывать его ограничения и проблемы, а также постоянно обновлять данные и анализировать результаты.

2.2 Особенности платёжного календаря Иркутской нефтяной компании

2.2.1 Основные этапы формирования платёжного календаря Иркутской нефтяной компании

В основу планирования платежей компании ложится Бюджет доходов и расходов компании (Далее БДР) на предстоящий календарный год. БДР является базой для долгосрочного планирования платежей на период от месяца до конца года.

В среднесрочной перспективе, от 3 платёжных дней до конца текущего месяца, платёжный календарь опирается на текущий мониторинг бюджета движения денежных средств (Далее БДДС), скорректированный на обновлённые макропараметры, относительно утверждённой годовой версии, такие как: курс доллара, курс юаня, цена нефти ESPO, цена нефти Urals, индекс потребительских цен.

Краткосрочная перспектива платёжного календаря (на три платёжных дня) формируется на основе заявок на платежи, подаваемых заявителями в системе 1С ERP Управление холдингом, на основании графиков платежей по действующим контрактам, актам выполненных работ, а так же в соответствии с графиком исполнения обязательств по налогам и займам.

Важным фактором является то, что Иркутская нефтяная компания представлена группой компании, в которую входит 13 юридических лиц, что приводит к необходимости управления системой банковских счетов, консолидации платёжных данных и данных о БДДС и БДР всей группы компаний в единую систему.

Таким образом, казначейство обладает точной информацией об объёме платежей и их назначении на 3 дня вперёд, далее следует экстраполяция БДДС, с учётом актуальных макропараметров.

2.2.2 Сильные и слабые стороны существующего платёжного календаря

Подход к планированию платежей, описанный в предыдущем разделе, обладает рядом недостатков, среди которых:

1. Большое количество рисков, заложенных в среднесрочной перспективе, а именно – рыночные (ценовые) риски, риски смещения сроков платежей на неопределённый срок, валютные риски.

2. Отсутствие гибкости планирования и сценарного планирования, особенно при изменении производственной программы предприятия. Слишком сложно и долго строить новый платёжный календарь, в том случае если руководством компании принято решение о пересмотре коммерческой и производственной программы.

3. Трудоёмкость ведения и актуализации платёжного календаря, вызванная необходимостью подготовки текущего мониторинга бюджета компании, лежащего в основе платёжного календаря.

4. Высокая сложность факторного анализа при прогнозировании кассовых разрывов и расчёте необходимых заёмных денежных средств

К плюсам такого подхода можно отнести высокую детализацию и прозрачность структуры выбытий и поступлений денежных средств, ведь все данные в платёжном календаре происходят из БДДС, подробно раскрывающего природу денежных потоков.

2.2.3 Структура платёжного календаря

Платёжный календарь делится на пять категорий платежей: Платежи по текущей деятельности, платежи по инвестиционной деятельности, платежи по финансовой деятельности, курсовые разницы и остатки

Каждая из категорий представлена двумя видами денежных потоков: поступления и выбытия, которые определяют баланс и остатки денежных средств.

Платежи по текущей деятельности – поступления и выбытия, непосредственно проистекающие из операционной деятельности предприятия, такой как реализация товаров и

услуг, выплата заработной платы, закупка расходных материалов и услуг по поддержанию текущей деятельности. Данная категория платежей является наиболее прогнозируемой, так как подчинена циклам и происходит из оперативного денежного потока компании, то есть денежный поток ограничен в каждый определённый момент времени, а ограничения происходят из производственной программы компании, определяемой на один календарный год вперёд.

Поступления по текущей деятельности являются основным источником поступлений группы компаний, и хоть размер этих поступлений сильно зависит от макропараметров, таких как стоимость нефть или курс доллара к рублю, перечень этих макропараметров известен и неизменен с течением времени. Таким образом размер поступлений денежных средств по текущей операционной деятельности группы компаний является, пожалуй, самым простым в прогнозировании денежным потоком, определяющий ограничения на размер выбытий по текущей деятельности для всей группы компаний.

Выбытия по текущей деятельности определены двумя важными факторами:

1. Безубыточностью деятельности группы компаний, что означает что выбытия по текущей деятельности не могут превосходить объём поступлений, так как задачей финансистов является обеспечение безубыточности группы компаний на всём промежутке жизненного пути компаний. Исходя из принципа безубыточности формируется максимальная сумма выбытий по текущей деятельности, за счёт корректирования сроков оплаты, распределения платёжной нагрузки в зависимости от поступлений и в редких случаях за счёт привлечения дополнительного финансирования во избежание кассовых разрывов.

2. Высокой долей платежей с закреплёнными сроками оплаты, такие как заработная плата, транспортные расходы за транзит через трубопроводы, налоги (НДС, НДПИ, налог на прибыль), оплата процентов по кредитам и займам.

Эти два фактора приводят к отсутствию выбросов во временном ряду и сглаживают ежедневные значения платежей, что упрощает процесс прогнозирования будущих платежей, однако даже в этих условиях имеется существенная доля платежей, зависящих срока поставки на склад материалов или сроков выполнения работ по текущему ремонту и обслуживанию. Первые приводят к неопределённостям уже на дистанции в 30 дней, вторые к неопределённостям на дистанции 60 и более дней. Следствием этого может являться избыточное удержание денежных средств в резерве или нецелесообразный вывод денежных

средств из оборота. Снижение объёма денежных средств, удерживаемых под оплату расходов по текущей деятельности за счёт более точного планирования графика платежей – одна из основных целей данного исследования.

Платежи по инвестиционной деятельности – поступления и выбытия, проистекающие из инвестиционных денежных потоков, таких как продажа основных средств, возврат займов на инвестиционные проекты и вложений в дочерние общества, проценты по депозитам, оплата приобретения основных средств, капитальный ремонт, импортное оборудование, услуги по инжинирингу, строительному контролю и строительным подрядам. Размеры денежных потоков по инвестиционной деятельности уступают операционной деятельности, однако прогнозирование данного вида платежей – более сложная задача и вот почему:

1. Высокая доля контрактных платежей с неопределённым графиком оплаты, высокая доля не спрогнозированных платежей, требующих пересмотра бюджета доходов и расходов

2. Выбытия по инвестиционной деятельности обладают более непредсказуемыми размерами из-за существенных различий в порядке оплаты, нежели по операционной деятельности

3. Сроки оплат могут существенно изменяться на протяжении всего жизненного цикла проекта или контракта, а сами размеры выплаты авансирования и окончательного платежа могут существенно отличаться в разных контрактах

4. Высокая степень риска отложенного срока окупаемости инвестиционных проектов, приводящего к отсутствию поступлений по инвестиционной деятельности на всём горизонте планирования

Целесообразность прогнозирования поступлений и выбытий по инвестиционной деятельности с помощью машинного обучения требует отдельного анализа, так как уверенности в эффективности такого подхода нет

Платежи по финансовой деятельности – категория денежных потоков, которые проистекают от деятельности, направленной на привлечение заёмных средств, либо получение дивидендов от выдачи займов и вкладов. Данная деятельность подчинена строго определённым графикам платежей, либо напрямую проистекает из эффективного управления финансами при осуществлении текущей и инвестиционной деятельности. Задачи по прогнозированию денежных потоков финансовой деятельности с помощью машинного

обучения не стоит, однако именно эти денежные потоки помогут в будущем оценить эффективность внедряемых решений за счёт снижения объёмов заимствований для покрытия кассовых разрывов и общем снижении количества подобных операций.

Исследование особенностей структуры платёжного календаря Иркутской нефтяной компании, позволяет уточнить и конкретизировать задачи проекта с учётом специфики деятельности организации и дочерних обществ, а так же с учётом особенностей и характера временных рядов выбытий по виду деятельности.

Уточнённая формулировка задач:

1. Консолидировать данные о платежах группы компаний, макропараметрах и законтрактованных обязательствах
2. Разработать решение по формированию датасетов, характеризующих временные ряды платежей по отдельным организациям группы, поддерживающих гибкую настройку, в том числе по видам деятельности, временным промежуткам и статьям бюджета.
3. Сформировать временной ряд и соответствующий ему датасет по текущей деятельности головной организации ООО «ИНК»
4. Определить основные характеристики временного ряда выбытий по текущей деятельности головной организации ООО «ИНК»
5. Провести сравнительный анализ моделей машинного обучения, обученных прогнозированию временного ряда на сформированном датасете
6. Определить модели, подходящие под сформулированные критерии достижения цели проекта (снижение MSE на 40 и более процентов)

Глава 3.
Разработка дата сета и формирование временных рядов денежных потоков ГК ИНК

3.1 Обзор имеющейся информации и источников данных

Данные о бюджетах и финансовой деятельности ГК ИНК хранятся в нескольких разрозненных информационных системах (Далее ИС). За прошедшие 15 лет менялся как состав группы компаний, так и состав используемых инструментов – бухгалтерии 1С, система документооборота DocsVision, CRM система управления предприятием 1С УПП, нормативно-справочная система 1С НСИ и внедряемая в текущий момент времени CRM система 1С ERP Управление Холдингом. Для анализа данных и построения временных рядов, было необходимо консолидировать эти данные в единый дата сет.

Первичный обзор доступных систем и их возможностей по обмену данными, позволил сформировать понимание того, каким образом можно объединить архивные и актуальные данные, а так же определил возможный состав данных, подлежащих анализу.

Из доступных источников информации можно выделить следующие категории:

1. Платежи:

1.1. Размеры платежей – их сумма является значением целевой функции прогнозирования.

1.2. Даты платежей – основа для построения временного ряда

1.3. Контрагенты – атрибут, необходимый для выделения внутригрупповых операций, а так же для проверки корректности отнесения тех или иных платежей к той или иной деятельности (инвестиционная, финансовая, текущая операционная)

1.4. Назначение платежа, статья расходов, статья БДДС – атрибуты для классификации платежей внутри категорий и видов деятельности

1.5. Виды деятельности – отделение выбытий от поступлений

1.6. Валюта платежей – дополнительный признак, позволяющий сформировать представление о соотношении платежей в национальной и иностранной валюте

2. Макропараметры:

2.1. Курсы валют – дополнительный признак при анализе временного ряда

2.2. Стоимость нефти – один из главных макропараметров, определяющих сумму поступлений

2.3. Индекс потребительских цен – для косвенной характеристики инфляции и колебания цен.

3. Контракты:

3.1. Дата заключения контракта

3.2. Ориентировочная сумма контракта – сумма может изменяться, однако для исключения дублирования сумм, принято решение брать лишь первичную сумму контракта

3.3. Валюта контракта

3.4. Ожидаемая дата завершения контракта

3.5. Статья бюджета, статья БДДС

4. Производственный календарь

4.1. Признак праздничного или рабочего дня на каждый день дата сета

3.2 Формирование датасета платежей

Как показал обзор источников данных, формирование датасета – это задача унификации и консолидации данных из разных систем в единый формат. Во-первых, были объединены данные о поступлениях и выбытиях денежных средств из 1С УПП (с 2016 года по сентябрь 2023 года) и 1С ERP Управление холдингом (с сентября 2023 года, по н.в.).

Во-вторых, были объединены данные о контрактуемых обязательствах группы компаний из систем DocsVision (данные о контрактах до 2019 года), 1С НСИ (с 2019 года по н.в.) и дополнены данными об изменениях размеров обязательств из 1С УПП Бюджетирование, для учёта изменений сумм первичных договоров, изменяемых дополнительными соглашениями. Сумма законтрактованных обязательств определяет объём платежей, которые группа компаний юридически обязана погасить в обозримом будущем. Большая часть всех выбытий денежных средств имеют контрактное происхождение.

В третьих, необходимо собрать данные за рассматриваемый период (с 2016 года по н.в.) об экономических макропараметрах, оказывавших влияние на экономику группы компаний, такие как курс доллара, стоимость нефти марки BRENT и ESPO, и прочих макропараметрах. Источниками этих данных являются сторонние сервисы и сайты, в том числе использующие API интерфейс для обработки внешних обращений.

Одним из первых препятствий при объединении данных о выбытиях и поступлениях, стали различные справочники статей расходов и доходов в системах управления казначейством 1С УПП и 1С ERP. Создание мэппнга статей оказалось не тривиальной задачей, так как каждая статья доходов и расходов в 1С УПП соответствовала сочетанию статьи БДДС и номенклатур-

ной группы в 1С ERP, таким образом мэппинг справочников статей расходов и доходов представлен словарём, где каждому ключу – статье, соотвтетствует массив из двух элементов: статья БДДС и родительская категория номенклатуры, присвоенной платежу в 1С ERP.

Вторым препятствием при консолидации данных о законтрактованных обязательствах группы компаний, являлось отсутствие требования об обязательном заполнении некоторых полей договора, например сумма договора в рублях. Для решения этой задачи, была реализована функция по восстановлению курса валюты на дату заключения контракта и пересчёт суммы контракта в рублях, там, где эта сумма отсутствовала. Решение этой задачи было реализовано с помощью библиотеки forex_python, а так же альтернативным способом через библиотеку requests и динамические запросы к сайту «https://www.cbr.ru/scripts/XML_daily.asp?».

На этом трудности по консолидации данных о заключаемых договорах не заканчивались, ведь помимо корректной суммы контракта в рублях, необходимо было определить дату начала действия контракта и его плановую дату окончания. К большому сожалению, журналы изменения статусов договора далеко не всегда обладали исчерпывающей информацией, позволяющей точно установить дату вступления контракта в силу, для определения этой даты пришлось сопоставлять имеющиеся данные о контрактах с данными об авансировании из казначейства, а так же анализировать даты загрузки вложений в карточку договора, соответствующих по формальным признакам сканированным версиям договоров. Определить дату планового окончания договора удалось только для полностью оплаченных контрактов, впрочем этого было вполне достаточно для поставленной дели – определения финансовой законтрактованной нагрузки на предприятия группы компаний.

Учитывая сложную структура группы компаний, возникла ещё одна проблема – внутригрупповые операции и реализация через комитента. Часть расходов возникали при оплате материалов, сырья и оказания услуг внутри самой группы компаний. Эти транзакции не влияли на группу компаний в целом, однако были отражены в расходах и доходах каждого отдельного предприятия. В тоже самое время выручка комитента приводила к задвоению доходов, если не идентифицировать и не исключить данные транзакции. Анализ ситуации привёл к возникновению дополнительных отборов данных при формировании датасета, а именно:

- Контрагенты не должны содержаться в списках организаций
- Статьи «агентское вознаграждение» и «выручка комитента» проходили дополнительную проверку на наличие аналогичных сумм поступлений в группе компаний в отчётном периоде

Кроме того, неочевидным источником долговых обязательств оказались налоговые льготы и вычеты, данные о которых не содержатся в реестре договоров, а рассчитывались вручную на основе объёма добычи углеводородов на разных участках недр. Позже эти данные

добавлялись к итоговому датасету. В общей сложности правки и корректировки затронули 14% от общего объёма данных за 9 лет, с 2016 года.

Все упомянутые данные являются коммерческой тайной и не будут предоставлены в качестве приложений, только в формате графиков и описаний датасетов в ходе выполенния программного кода.

Исходные данные, на основе которых формировался датасет для обучения и тестирования моделей машинного обучения, представляют собой три таблицы:

1. **Платежи:**
    a. Дата оплаты
    b. Сумма платежа (тыс. рублей)
    c. Валюта
    d. Организация
    e. Контрагент
    f. Статья БДДС
    g. Номенклатура
    h. Деятельность (текущая/инвест/финансовая)
    i. Направление (выбытия/поступления)
    j. День недели
    k. Номер рабочего дня в году
    l. Номер недели
    m. Год
2. **Договоры:**
    a. Номер
    b. Общая сумма договора (тыс. рублей)
    c. Организация
    d. Контрагент
    e. Дата заключения
    f. Ожидаемая дата окончания
    g. Бюджетная статья
3. **Макропараметры:**
    a. Дата
    b. Курс доллара
    c. Курс Юаня
    d. Стоимость нефти марки Brent
    e. Стоимость нефти USPO (по данным реализации)

Применяя отборы и группируя данные по дате и иным аналитикам, итоговый датасет платежей за текущую деятельность имеет следующий вид:

Фильтры: организация «Общество с ограниченной ответственностью Иркутская нефтяная компания», деятельность «текущая», направление «выбытия», исключая внутригрупповые операции (Контрагенты отсутствуют в списке организаций группы компаний)

Состав датасета:

1. Дата
2. Агрегированная сумма платежей (тыс. рублей)
3. День недели
4. Номер рабочего дня в году
5. Номер недели
6. Год
7. Накопительная сумма платежей с начала недели
8. Прошлогодняя сумма на дату
9. Прошлогодняя сумма на день
10. Прошлогодняя недельная сумма платежей
11. Курс доллара
12. Стоимость нефти Brent
13. Накопленная сумма законтрактованных обязательств

Подробная реализация формирования датасета, представлена в Приложении 1.

Визуализация временного ряда представлена на рисунке 1.

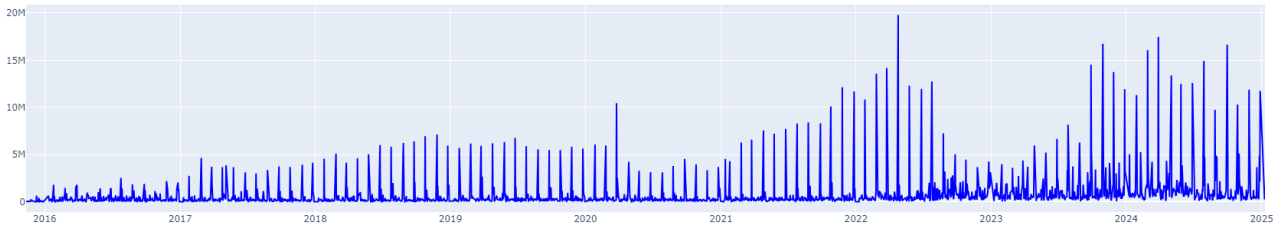

Рис. 1 Временной ряд выбытий ООО «ИНК» по текущей деятельности

Визуализация временного ряда законтрактованных платёжных обязательств представлена на на рисунке 2.

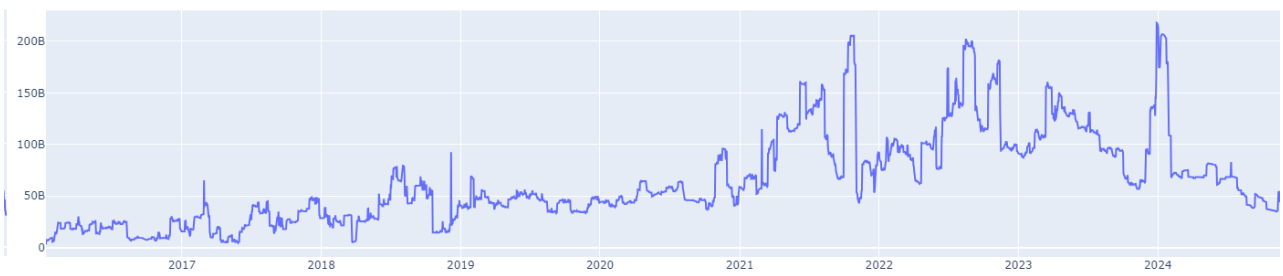

Рис. 2 график контрактования платёжных обязательств ООО «ИНК»

Дальнейшая работа по обучению и тестированию моделей прогнозирования будут производиться на описанном выше датасете, оценка эффективности прогноза так же будет проводиться для данного датасета.

3.3 Исследование характеристик временных рядов ГК ИНК

Выбор инструментов для дальнейшего анализа временного ряда (Далее ВР) зависит от основных характеристик этого ВР, таких как:

1. Тренды - долгосрочные тенденции, показывающие общее направление изменения данных (например, устойчивый рост или падение).
2. Сезонности - повторяющиеся изменения, связанные с календарными периодами, такими как времена года или праздники.
3. Цикличность ВР - колебания, обусловленные экономическими циклами, которые могут длиться от нескольких месяцев до нескольких лет.
4. Наличие случайных колебаний - непредсказуемых изменений, вызванных случайными событиями или аномалиями

Типы временных рядов так же влияют на применимость тех или иных инструментов и подходов к анализу и прогнозированию ВР. По типам ВР можно разделить на две группы:

1. **Стационарные ряды**: Ряды, которые сохраняют постоянные статистические свойства (среднее, дисперсия) со временем. Их анализ проще, так как они не содержат трендов или сезонности.
2. **Нестационарные ряды**: Ряды, которые имеют переменные статистические свойства, на которые влияют тренды или сезонные эффекты. Для их анализа часто требуется предварительная обработка, чтобы устранить эти эффекты.

Для более глубокого понимания структуры данных ВР, особенно для нестационарных рядов, проводится дополнительный анализ – декомпозиция и автокорреляция.

1. **Декомпозиция**: Разложение временного ряда на его компоненты (тренд, сезонность, случайные колебания)
2. **Автокорреляция**: Изучение зависимости между значениями ряда в разные моменты времени, что помогает выявить наличие трендов или сезонности.

С программной реализацией исследования ВР ГК ИНК можно ознакомиться в **Приложении 2** к данной работе, графическое описание и результаты анализа представлено ниже:

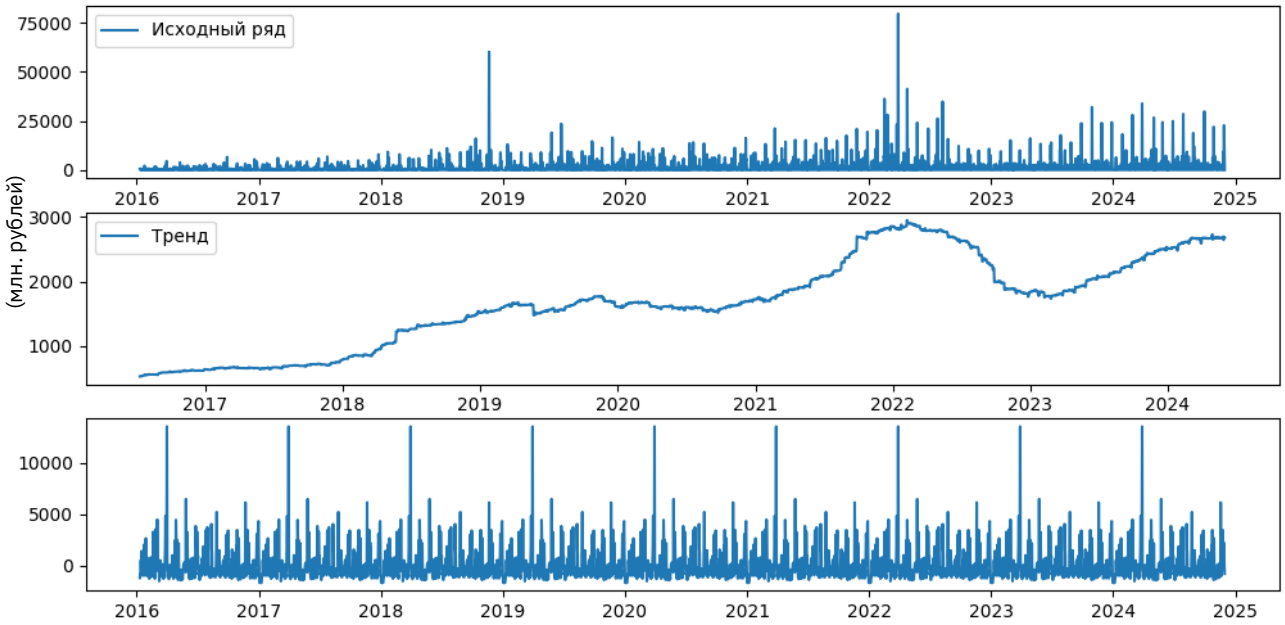

Рис. 3 Визуализация временного ряда, тренда и сезонности

Визуальный анализ представленных графиков позволяет установить наличие общего тренда – компания разрастается на протяжении всего периода датасета, вместе с этим растут выбытия – расходы на все виды деятельности, а так же налоги. При детальном рассмотрении графиика, можно установить что на конец месяца, а именно на 27-29 числа каждого месяца выпадает пик выбытий – выплата налогов и долговых обязательств. Сезонность присутствует и равна одному календарному месяцу.

Наиважнейшим свойством временного ряда является его стационарность. Чтобы подтвердить стационарность ряда, был проведён тест Дики-Фуллера, результаты которого позволяют отвергнуть нулевую гипотезу о не стационарности ВР:

Результаты теста Дики-Фуллера

- **Статистика теста**: -5.589267891856472
- **p-значение**: 1.340667659884753e-06
- **Число лагов**: 29
- **Число наблюдений**: 3219

Критические значения

- **1%**: -3.4323830917455003

- **5%**: -2.8624382973530444
- **10%**: -2.5672481834828367

Интерпретация результатов

- **Статистика теста**: Значение -5.5892 значительно меньше критических значений на уровнях 1%, 5% и 10%, что указывает на сильные доказательства против нулевой гипотезы.
- **p-значение**: Значение 1.3407e-06 также указывает на то, что нулевая гипотеза о нестационарности может быть отвергнута, так как оно значительно меньше 0.05.

Стационарность ВР означает, что его статистические свойства, такие как среднее и дисперсия, не изменяются со временем. Это является требованием к ВР при анализе и прогнозировании его с помощью разнообразных моделей машинного обучения. Если ВР не является стационарным, возникает потребность в преобразовании данных и приведении ВР к стационарности. Однако в нашем случае ряд стационарен.

Следующим важным исследованием ВР, является выявление автокорреляции – зависимости значений целевой переменной от прошлых значений этой переменной. Автокорреляция и частичная автокорреляция временного ряда используются для анализа зависимости между текущими и прошлыми значениями ряда. Они помогают выявить периодичности и структуру данных, что важно для построения моделей прогнозирования.

Автокорреляция

- **Определение**: Автокорреляция измеряет степень зависимости между значениями временного ряда на разных временных интервалах (лагах).
- **Применение**:
    - Позволяет определить, насколько текущее значение ряда связано с его предыдущими значениями.
    - Используется для выявления сезонных эффектов и трендов.
    - Помогает в выборе параметров для моделей временных рядов, таких как ARIMA.

Частичная автокорреляция

- **Определение**: Частичная автокорреляция измеряет зависимость между текущим значением и значениями на предыдущих временных интервалах, устраняя влияние промежуточных значений.
- **Применение**:
    - Позволяет более точно определить, какие лаги действительно влияют на текущее значение, исключая влияние других лагов.
    - Используется для выбора порядка авторегрессионных моделей, что важно для построения точных прогнозов.
    - Помогает в анализе структуры временного ряда и выявлении скрытых зависимостей.

Графики автокорреляции и частичной автокорреляции

Коррелограммы: Графики автокорреляции (ACF) и частичной автокорреляции (PACF) визуализируют зависимости и помогают в интерпретации данных. Эти графики позволяют быстро оценить, какие лаги имеют значительное влияние, что упрощает процесс построения моделей прогнозирования.

ACF (Autocorrelation Function) показывает, как значения временного ряда коррелируют с его лагированными версиями. Если бары на графике ACF превышают границы значимости, это указывает на наличие значительной автокорреляции на соответствующем лаге. Постепенное снижение высоты баров может указывать на долгосрочные зависимости в данных. Регулярные пики на определенных лагах могут свидетельствовать о сезонности в данных.

PACF (Partial Autocorrelation Function)

PACF измеряет корреляцию между наблюдениями на разных временных интервалах, устраняя влияние всех более коротких лагов. Значительный пик на определенном лаге указывает на необходимость включения этого лага в авторегрессионную модель. Лаг, на котором график PACF обрывается, помогает определить максимальный лаг для включения в модель.

Применение в моделировании

- **Выбор модели**: ACF и PACF помогают определить порядок авторегрессионных (AR) и скользящих средних (MA) моделей, что критично для построения точных прогнозов.

- **Анализ структуры**: Эти графики позволяют выявить скрытые зависимости и паттерны, что улучшает понимание динамики временного ряда.

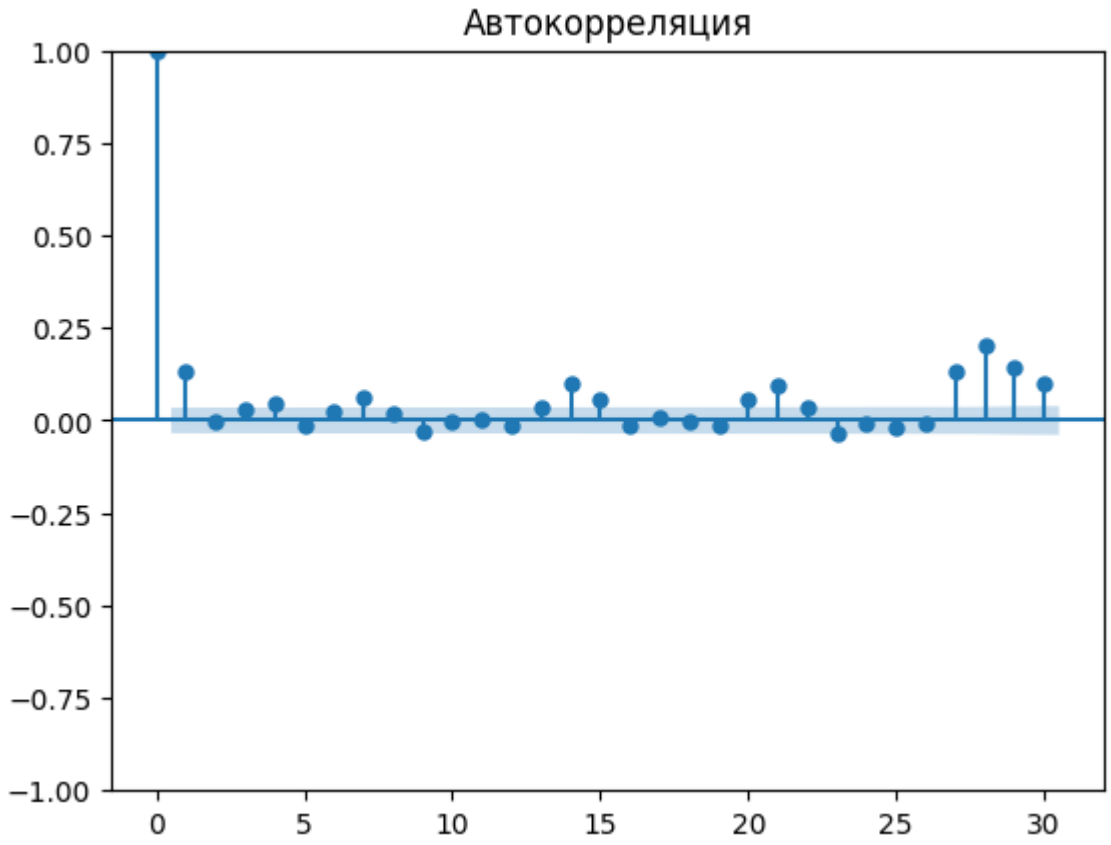

Рис. 4 Визуализация коэффициента автокорреляции временного ряда

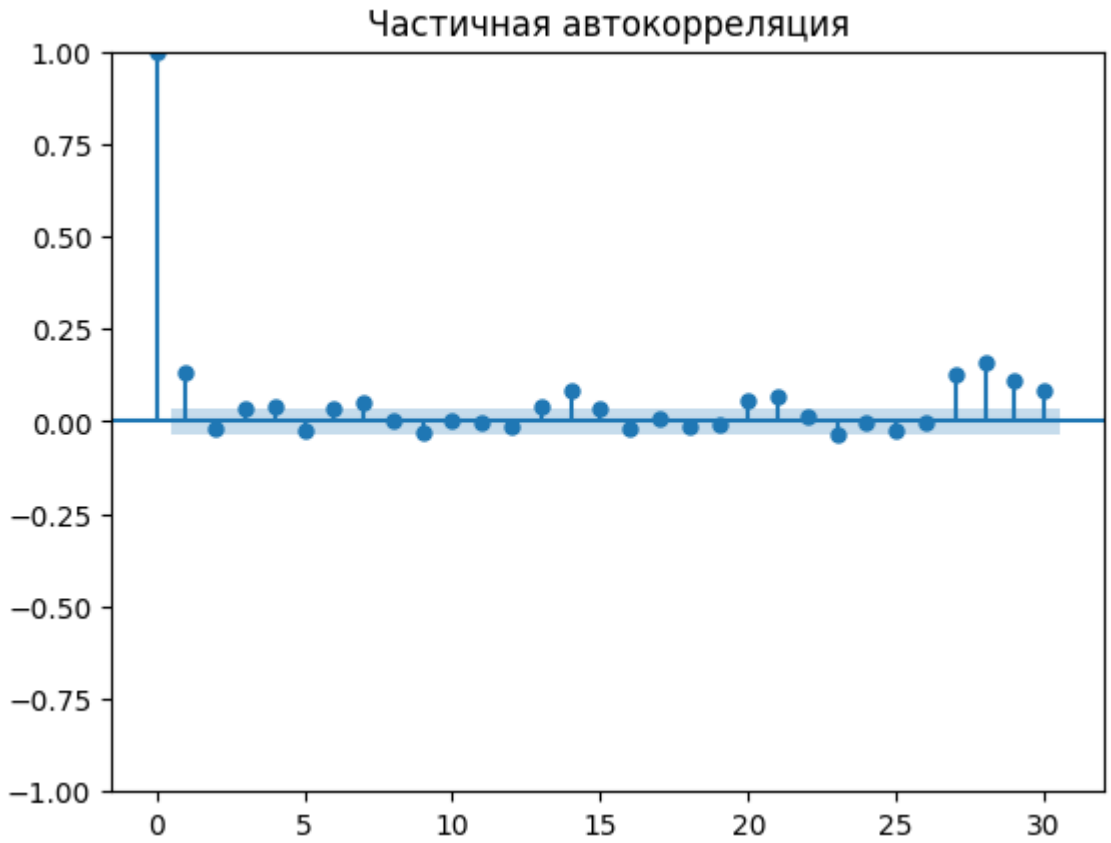

Рис. 5 Визуализация коэффициента частичной автокорреляции временного ряда

ВР выбытий ГК ИНК демонстрирует отсутствие автокорреляции, то есть значения целевой переменной практически не зависит от значений этой переменной в прошлом и не обладает долго периодичными корреляциями.

Результаты анализа графиков ACF и PACF свидетельствуют о том, что модели машинного обучения, полагающиеся на выявление сезонных трендов и автокорреляцию, такие как ARIMA, SARIMAX и ряд нейросетей могут не продемонстрировать хороших результатов по точности построения прогноза, однако при должном описании каждого значения целевой переменной набором атрибутов в дата сете, может позволить строить прогнозы простыми моделями линейной регрессии, моделями скользящего среднего, регрессионными моделями с использованием внешних факторов, например деревья решений и случайные леса. Важным элементом верификации прогноза модели в этом случае будет анализ остатков (ошибок) на наличие автокорреляции и нормальности распределения для того, чтобы убедиться в адекватности того, каким образом модель описывает ВР.

Глава 4. Анализ временных рядов

с помощью моделей машинного обучения

4.1 Применение модели ARIMA и SARIMAX для прогнозирования временного ряда платёжного календаря

В данной главе рассматриваются результаты прогнозирования временного ряда платёжного календаря Иркутской нефтяной компании с использованием моделей ARIMA и SARIMAX. Целью данного анализа является оценка точности прогнозов, полученных с помощью этих моделей, и их сравнение между собой.

4.1.1 ARIMA (AutoRegressive Integrated Moving Average) — это классическая модель для анализа временных рядов, которая учитывает автокорреляцию и тренды. Для данной модели были выбраны параметры (p, d, q), где:

- p — порядок авторегрессии, определяется лагом на графике PACF, на котором корреляция становится незначительной

Рис. 6 Визуализация функции PACF временного ряда

- d — порядок интегрирования ВР для приведения его к стационарному. Наш ВР стационарен.
- q — порядок скользящего среднего. Определяется лагом на графике ACF, на котором корреляция становится незначительной

Учитывая результаты анализа временного ряда и график функции PACF, можно определить параметры p=1, d=0, q=0. Обучение модели проводилось на 80% временного ряда и предсказанием на 20% тестовых данных.

В качестве предсказания, модель ARIMA предложила константу в качестве прогноза на весь период.

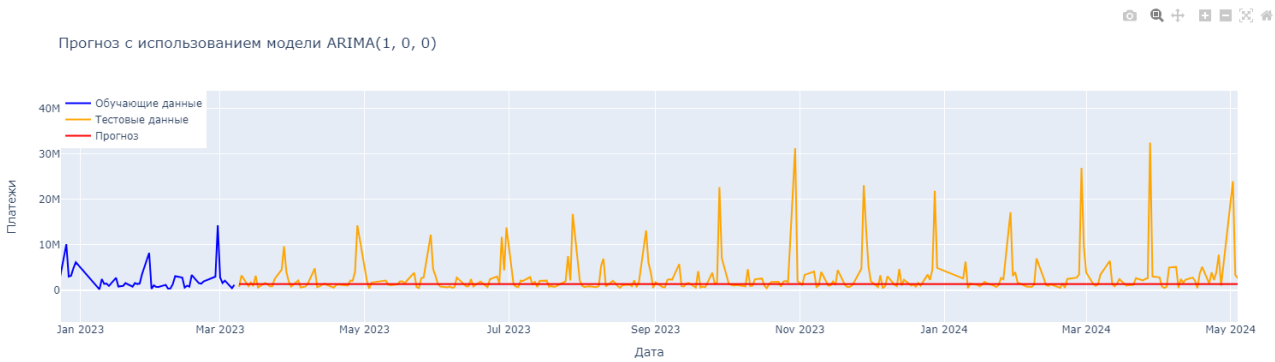

Рис. 7 Визуализация временного ряда и прогнозных значений модели ARIMA

Таким образом применение данной модели не целесообразно для прогнозирования стационарного ВР без автокорреляции. Расчёт метрик MAE, RMSE, MAPE не производился в виду отсутствия адекватного прогноза, применимого на практике.

4.1.2 **Модель SARIMAX** (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) расширяет модель ARIMA, добавляя возможность учитывать сезонные компоненты и экзогенные переменные. Данная модель настраивается параметрами (p, d, q), как у ARIMA и сезонными параметрами (P, D, Q, s), где s — период сезонности.

Для данной модели была предпринята попытка найти наилучшее сочетание параметров с помощью кросс-валидации. Подробное описание работы с моделями ARIMA и SARIMAX изложено в **Приложении 2**. В нашем случае были выбраны значения p=0, d=1, q=1, P=0, D=0, Q=0, s=12.

Прогнозирование с использованием модели SARIMAX дало оценку mse = 28240719950117.688 однако графический анализ результатов предсказания продемонстрировал непригодность применения и данной модели, отсутствие автокорреляции не позволяет строить хоть сколько-нибудь точные прогнозы.

4.2 Анализ временного ряда с помощью Xgboost

XGBoost (eXtreme Gradient Boosting) — это модель машинного обучения, которая относится к классу алгоритмов градиентного бустинга. Она используется для решения задач классификации и регрессии.

Основные принципы работы XGBoost:

1. **Бустинг (Boosting):** XGBoost строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Каждая последующая модель пытается исправить ошибки предыдущей, тем самым улучшая общую точность предсказаний.
2. **Градиентный спуск:** для минимизации функции потерь (например, среднеквадратичной ошибки) XGBoost использует метод градиентного спуска. Это позволяет алгоритму эффективно находить оптимальные параметры модели.
3. **Регуляризация:** XGBoost включает механизмы регуляризации, такие как ограничение глубины деревьев, минимальное количество объектов для разделения узла и максимальное количество листьев в дереве. Это помогает предотвратить переобучение модели и улучшить её обобщающую способность.
4. **Параллелизация:** XGBoost может эффективно использовать многопоточность и распределённые вычисления, что позволяет ускорить процесс обучения модели и сделать его более масштабируемым.
5. **Гибкость:** XGBoost поддерживает различные функции потерь и метрики оценки, что делает его универсальным инструментом для решения широкого спектра задач машинного обучения.

Модель XGBoost достаточно требовательна к настройке гиперпараметров, которые могут существенно влиять на качество модели, к тщательной предобработке данных, включая обработку пропущенных значений и нормализацию признаков.

Архитектура XGBoost

XGBoost строит деревья решений поэтапно, добавляя каждое новое дерево для коррекции ошибок предыдущих. Основные этапы работы XGBoost:

1. **Инициализация**:
- Начинается с предсказания среднего значения целевой переменной.
2. **Обучение деревьев**:
- На каждом шаге добавляется новое дерево, которое минимизирует функцию потерь. Для этого вычисляются градиенты и гессианы, которые используются для построения дерева.
3. **Обновление предсказаний**:
- После добавления нового дерева обновляются предсказания модели.
4. **Регуляризация**:
- В процессе обучения учитываются штрафы за сложность модели, что помогает избежать переобучения.

При прогнозировании ВР платежей по текущей оперативной деятельности, данная модель продемонстрировала достойный результат. Модель точно предсказывает момент возникновения экстремумов во временном ряду, однако не достаточно точно предсказывает значение этих самых экстремумов. Так как работа модели не основана на автокорреляции и вычислении сезонных трендов, точность прогноза демонстрирует стабильные показатели, как при прогнозировании на 45 дней, так и на 90, плавно снижаясь к горизонту планирования до 365 дней по MSE от впечатляющих 258 245 до 1 543 322. При сравнении с существующим процессом построения прогноза, модель превосходит точность планирования текущего календаря на порядок, более чем в 10 раз, даже при построении прогноза на год вперёд

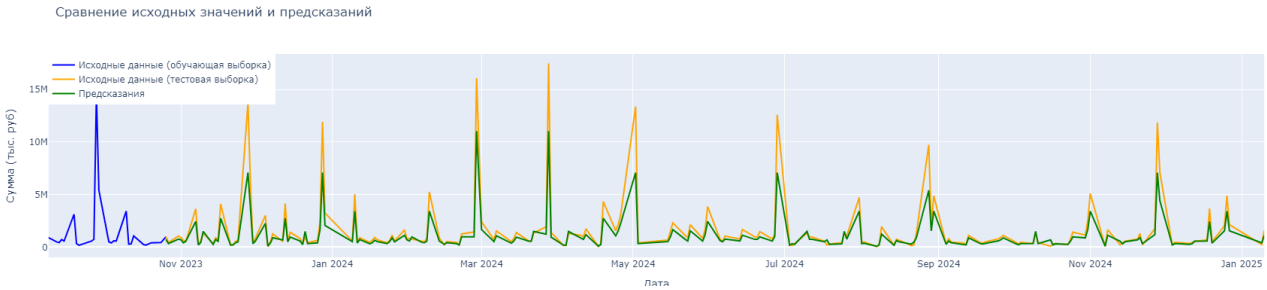

Рис. 8 Визуализация временного ряда и прогнозных значений на 390 дней, модели XGBoost

При прогнозировании на 90 дней вперёд, MSE = 258 245, что само по себе является превосходным результатом, достаточным, чтобы констатировать достижение поставленной цели

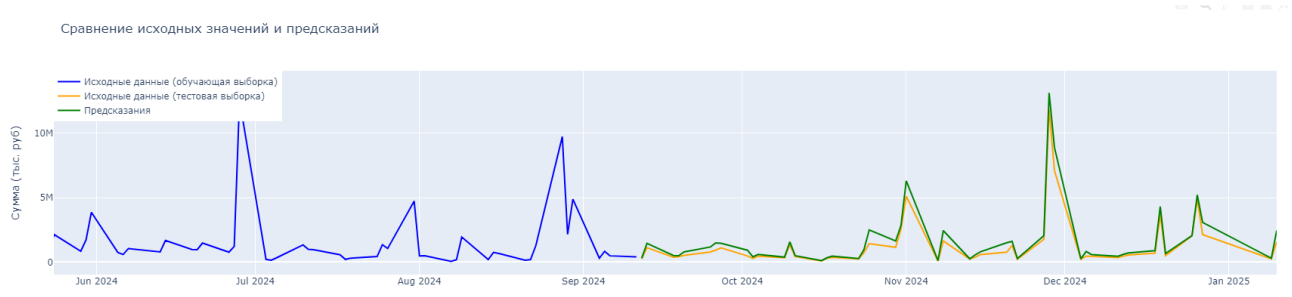

Рис. 9 Визуализация временного ряда и прогнозных значений на 45 дней, модели XGBoost

Для повышения точности прогнозирования экстремальных значений выбытий, особенно в критичных периодах, таких как конец календарного года, была предпринята попытка подобрать наиболее оптимальные макропараметры модели с помощью кросс-валидации, используя GridSearchCV.

Опытным путём были определены диапазоны поиска макропараметров, таких как величина ансамбля деревьев, глубины деревьев решений, шага градиентного спуска, доли случайных объёмов для построения деревьев, доля случайных признаков.

Однако, модель продемонстрировала склонность к переобучению, из-за чего не смотря на прекрасные показатели точности на обучающей выборке, превзойти имеющийся результат на тестовой выборке не удалось.

Подробная реализация прогноза ВР с помощью XGBoost, изложена в **Приложении 3.**

4.3 Анализ временного ряда с помощью модели случайного леса

Не смотря на успех модели XGBoost, останавливаться на этом рано, и мы исследуем работу другой модели, основанной на ансамблировании деревьев решений, а именно модель Случайного леса. В отличии от XGBoost, который последовательно строит ансамбль деревьев, постепенно снижая ошибки, модель Случайного леса полагается на множество независимых деревьев решений и объединяет их предсказания, Это помогает уменьшить вариативность модели и предотвратить переобучение. Кроме того модель Случайного леса не использует оптимизацию с помощью градиентного спуска. Вместо этого она строит каждое дерево решений с максимальной точностью на своей подвыборке данных, используя для этого бутстрэпинг.

Бутстрэпинг (от англ. "bootstrapping") — это статистический метод, который используется для оценки распределения выборочных статистик (например, среднего, медианы, стандартного отклонения и т.д.) путем многократного повторного выборки из исходного набора данных с возвращением. Этот метод позволяет оценить неопределенность и вариабельность статистик, а также строить доверительные интервалы.

Для построения предсказаний ВР было решено использовать ансамбль из 100 независимых деревьев.

Рис. 10 Визуализация ВР и прогнозных значений на 45 дней, модели случайного леса

Модель достойно справляется с прогнозированием на 90 дней, с не большим увеличением MSE, до 469 091, и на один календарный год, при этом точно предсказывая не платёжные дни, когда платежи отсутствуют. На дистанциях менее 6 месяцев модель уступает предыдущей, демонстрируя MSE в два раза выше.

Рис. 11 Визуализация ВР и прогнозных значений на 90 дней, модели случайного леса

Подробная реализация прогноза ВР с помощью модели случайного леса, изложена в **Приложении 3.**

4.4 Применение нейросетевых моделей для анализа временных рядов

Следующим этапом исследования, является прогнозирование ВР с помощью нейросетевых моделей. Для этой цели были составлены прогнозы с помощью двух видов нейросетевых моделей:

- Рекуррентной нейронной сети LSTM (Long Short-Term Memory)
- Сверточной моделью Conv1D с одним сверточным слоем

4.4.1 Нейросетевая модель LSTM (Long Short-Term Memory)

LSTM - это специализированный тип рекуррентной нейронной сети (RNN), разработанный для решения проблемы долгосрочных зависимостей в последовательных данных. Он был предложен в 1997 году Зеппом Хохрайтером и Юргеном Шмидхубером.

Рекуррентные нейронные сети (RNN, Recurrent Neural Networks) — это класс нейронных сетей, специально разработанных для обработки последовательных данных. Они отличаются от традиционных нейронных сетей тем, что имеют циклические соединения, позволяющие им сохранять информацию о предыдущих входах и использовать её для обработки текущих данных. Это делает RNN особенно подходящими для задач, где порядок данных имеет значение, таких как временные ряды, текст, аудио и другие последовательные структуры.

Механизм работы LSTM

LSTM использует несколько ключевых компонентов для управления потоком информации:

- **Состояние ячейки:** Это основная структура, которая хранит информацию на протяжении всей последовательности. Состояние ячейки проходит через всю сеть, подвергаясь минимальным изменениям.
- **Гейты:** LSTM имеет три типа ворот, которые регулируют, какая информация будет сохранена, забыта или передана дальше:

- Гейт забывания (Forget Gate): Определяет, какую информацию из состояния ячейки следует забыть. Он использует сигмовидный слой, который выдает значения от 0 до 1 для каждого элемента состояния ячейки.
- Гейт входа (Input Gate): Определяет, какую новую информацию следует добавить в состояние ячейки. Он также использует сигмовидный слой для принятия решения и слой для создания вектора новых значений-кандидатов.
- Гейт выхода (Output Gate): Определяет, какая информация из состояния ячейки будет передана на выход. Он использует сигмовидный слой для фильтрации состояния ячейки и слой tanh для нормализации значений.

Пошаговый процесс работы LSTM:

1. Определение текущего состояния ячейки
2. Обработка входных данных
3. Изменение информации о состоянии ячейки за счёт удаления (забывания) избыточной информации (гейт забывания) и добавления новых данных о состоянии ячейки (гейт входа)
4. Обновление состояния ячейки
5. Генерация выхода - гейт выхода формирует выходное значение на основе обновленного состояния ячейки.

Эффективность прогнозирования моделью LSTM

Данная модель продемонстрировала достойный результат, схожий с моделью Случайного леса. Подробное сравнение эффективности предсказания моделей будет представлено в следующем разделе данной главы.

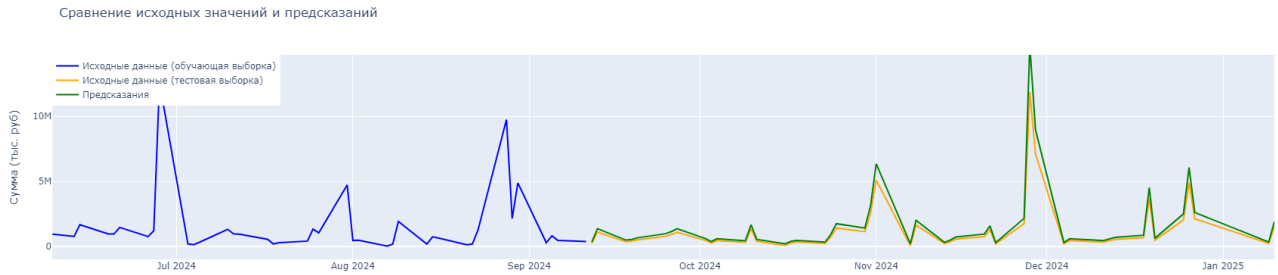

Рис. 12 Визуализация ВР и прогнозных значений на 90 дней, модели LSTM (нормализация)

MSE: 466 809

MAE: 363

4.4.2 Модель Conv1D с одним свёрточным слоем

Conv1D — это одномерная свёртка, применяемая к одномерным данным, таким как временные ряды или последовательности. Она используется для извлечения локальных признаков из входных данных.

Механика работы модели Conv1D с одним свёрточным слоем

1. **Входные данные:** входные данные для модели Conv1D представляют собой одномерный массив или последовательность значений. Например, это может быть временной ряд, где каждое значение соответствует определённому моменту времени. При этом одномерность входных данных не означает собой отказ от признаков,

описывающих значение целевой переменной, в нашем случае на вход подаётся одномерный массив признаков датасета, описывающих целевую переменную в каждый момент времени.

2. **Свёрточный слой:** свёрточный слой Conv1D состоит из нескольких фильтров (ядра свёртки), которые применяются к входным данным для извлечения локальных признаков. Каждый фильтр имеет определённый размер и количество выходных каналов.

3. **Применение фильтров:** фильтры применяются к входным данным путём скользящего окна. Для каждого положения окна фильтр умножается на соответствующие значения входных данных, и результаты суммируются. Это операция свёртки.

4. **Извлечение признаков:** результатом применения фильтров является карта признаков, которая представляет собой набор значений, полученных в результате свёртки. Карта признаков содержит информацию о локальных признаках, извлечённых из входных данных.

5. **Активация:** после свёртки обычно применяется функция активации, такая как ReLU, для введения нелинейности в модель. Это позволяет модели лучше обучаться на сложных данных.

6. **Выходные данные:** выходные данные модели Conv1D представляют собой карту признаков, полученную после применения свёрточного слоя и функции активации. Эти данные могут быть использованы для дальнейшей обработки или классификации.

Требования к входным данным

Входные данные для модели Conv1D должны соответствовать следующим требованиям:

- **Одномерность:** входные данные должны быть одномерными, то есть представлять собой последовательность значений.
- **Формат данных:** входные данные должны быть представлены в виде массива или тензора с определённым количеством измерений. Например, для временных рядов это может быть массив значений, где каждое значение соответствует определённому моменту времени.

• **Размерность:** размерность входных данных должна соответствовать размеру фильтров, используемых в свёрточном слое. Например, если фильтры имеют размер 3, то входные данные должны иметь размерность, кратную 3.

Эффективность прогнозирования моделью Conv1D с одним свёрточным слоем.

Данная модель продемонстрировала самые высокие показатели точности и самую лучшую динамику сохранения точности прогноза при увеличении дистанции прогноза. Ближе всего к данной модели находится модель XGBoost.

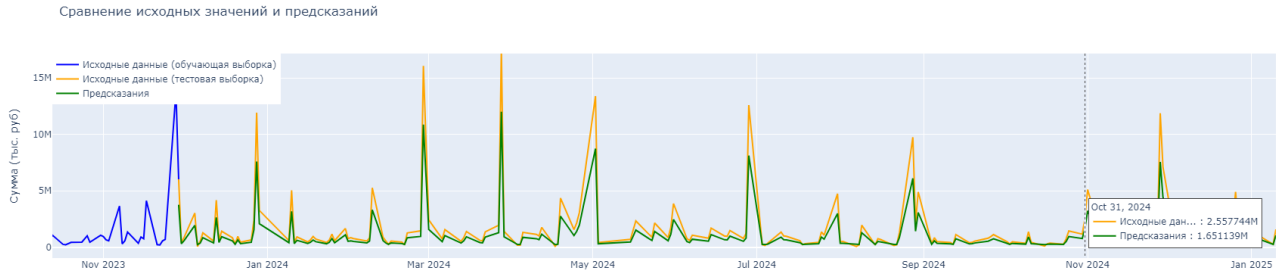

Рис. 14 Визуализация ВР и прогнозных значений на 365 дней, модели Conv1D

MSE: 1 224 534

MAE: 569

Подробная реализация прогноза ВР с помощью нейросетевых моделей, изложена в **Приложении 3.**

4.5 Сравнительный анализ моделей и фактически применяемого подхода

Для оценки эффективности прогнозирования ВР были рассчитаны метрики MAE и MSE для существующего платёжного календаря за период с 2019 года до настоящего времени, в двух вариантах. Первый вариант – прогноз на 3 месяца вперёд (Мониторинг бюджета М03, М07) в сравнении с фактом и второй вариант – прогноз на 45 дней вперёд (текущий мониторинг) с ежемесячной актуализацией в сравнении с фактом. Таким образом, можно сравнивать между собой прогнозы, составленные на 45 и 90 дней вперёд. Для существующего процесса формирования платёжного календаря, метрики отклонений составляют

- MSE для прогноза на 90 дней и факт: 185 538 578
- MAE для прогноза на 90 дней и факт: 8 173
- MSE для прогноза на 45 дней и факт: 43 233 078
- MAE для прогноза на 45 дней и факт: 5694

Данные показатели были рассчитаны для единиц измерения – миллионы рублей, поэтому метрики для моделей при расчётах так же приводились к данным единицам измерения.

Все исследуемые модели, кроме ARIMA и SARIMAX показали впечатляющие результаты при планировании временного ряда выбытий по текущей деятельности. Точность прогноза превосходит существующий процесс на порядок, поэтому сравнение между существующим процессом и результатами работы моделей бессмысленно.

Между собой, точность моделей машинного обучения отличается не существенно и разница становится видна лишь на дистанции в один календарный год. Нейросетевая модель с одним свёрточным слоем показала наилучшие результаты, на всех исследованных горизонтах планирования, демонстрируя наилучшее сохранение точности прогноза при увеличении дистанции планирования, впрочем, для достижения поставленной цели подходят как модели случайного леса, так и XGBoost, так и исследованные нейросетевые модели, а результат превзошёл все, даже смелые ожидания.

Сравнительная таблица показателей MSE и MAE

Модель/алгоритм	**Период прогнозирования**	**MSE**	**MAE**
Действующий процесс создания ПК	90	185 538 578	8 173
Действующий процесс создания ПК	45	43 233 078	5 694
Модель SARIMAX	365	28 240 719 950	-
XGBoost	365	1 136 768	501
XGBoost	90	258 245	335
Модели случайного леса	365	1 415975	641
Модели случайного леса	90	469 091	315
Рекуррентной нейронной сети LSTM	365	1 069 401	482
Рекуррентной нейронной сети LSTM	90	466 809	363
Модель Conv1D	365	1 224 534	569
Модель Conv1D	90	230 467	234

Подводя итоги проделанной работы, можно с уверенностью утверждать, что внедрение машинного обучения для прогнозирования выбытий денежных средств полностью оправдано и может быть рекомендовано в целях оптимизации существующего процесса и повышения эф-

фективности использования финансовых средств. Оценить экономический эффект от внедрение, можно лишь после опытной эксплуатации, однако никаких сомнений в том что эффект будет ощутимым при грамотном внедрении, не возникает.

Заключение

В результате проделанной работы, удалось достичь поставленной цели и выполнить все задачи. Более того, изначальная цель в повышении точности прогноза на 40% была перевыполнена, так как точность полученных прогнозов превосходит текущий прогноз на порядок (точность увеличена более чем в 20 раз).

Более того, прогнозирование ВР с помощью моделей не нуждается в наличии актуальных бюджетов, коммерческих и производственных программ, а процесс построения нового прогноза занимает от 5 до 10 минут.

Процесс поддержания датасетов в актуальном состоянии может быть полностью автоматизирован посредством WSL утилиты, а новые прогнозы могут генерироваться каждый день, предоставляя пользователям актуальные данные на ежедневной основе. Данный аспект является перспективным направлением улучшения и расширения достигнутых результатов.

Затраты на работу таких моделей не превосходят средней заработной платы одного ведущего специалиста, что так же является преимуществом перед аналогичными процессами, существующими в компании.

Датасет, полученный в ходе выполнения проекта может быть дополнен дополнительными признаками и атрибутами и повторно использован для решения широкого спектра задач в области финансового анализа, а так же для оценки эффективности внедряемых решений.

Полученный в ходе выполнения проекта опыт может быть применён и в других областях, таких как прогнозирование изменений характеристик скважин (дебеты, коэффициенты отвлекаемости), при решении логистических вопросов (сроки контрактации, сроки поставок, оказания услуг) и при решении иных задач.

Список используемой литературы

- Книга: «Анализ временных рядов» Авторы: Катаргин Н.В., Качалина Е.А. Город: Санкт-Петербург, 2024 год, Издательство: Лань, 180 страниц.
- Книга: «Машинное обучение с PyTorch и Scikit-Learn» Авторы: Себастьян Рашка, Юси Лю, Вахид Мирджалили, 2024 год, Издательство: Фолиант, 688 страниц.

Приложения

1. Загрузка первичных данных – документ в формате .ipynb с программным кодом по формированию датасета временного ряда
2. Исследование статистических характеристик временного ряда, построение прогноза ARIMA, SARIMAX – документ в формате .ipynb демонстрирующий ход исследований временного ряда и результаты построения прогноза с помощью моделей машинного обучения.
3. Сравнительный анализ эффективности моделей машинного обучения – документ в формате .ipynb с программным кодом построения прогноза моделей XGBoost, Случайного леса, нейросетевых моделей LSTM и Conv1D с демонстрацией результатов.