Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ

Школа профессионального и академического образования

ДОПУСТИТЬ К ЗАЩИТЕ ПЕРЕД ГЭК

Директор ШПиАО
_____Д.В. Денисов_
(подпись) (Ф.И.О.)
«_____»_____2024 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

ПРОЕКТИРОВАНИЕ АЛГОРИТМА ПРОГНОЗИРОВАНИЯ СЫРЬЯ

Научный руководитель: Кислицин Евгений Витальевич _____
канд. экон. наук, доцент
подпись

Нормоконтролер: Огуренко Егор Владимирович _____
подпись

Студент группы: РИМ-220963 Михайличенко
Людмила Александровна

подпись

Екатеринбург
2024

РЕФЕРАТ

Выпускная квалификационная работа магистра 98 стр., 36 рис., 7 таблиц, 47 источников, 7 прил.

ПРОГНОЗИРОВАНИЕ ЗАКУПОК, ВРЕМЕННЫЕ РЯДЫ, МАШИННОЕ ОБУЧЕНИЕ, ПРОГНОЗИРОВАНИЕ СПРОСА, ПРОГНОЗИРОВАНИЕ СЫРЬЯ, ГРАДИЕНТНЫЙ БУСТИНГ, XGBOOST, ЭКСТРЕМАЛЬНЫЙ ГРАДИЕНТНЫЙ БУСТИНГ, LSTM.

Цель работы – проектирование алгоритма прогнозирования сырья, состоящего из нескольких этапов: прогнозирование спроса и расчет сырья на основании рассчитанного спроса. Рассматриваются процесс сбора, предобработки набора данных и предварительный анализ статистических моделей, моделей машинного обучения для выбора лучшей на основании метрик точности и качества. Предложен алгоритм прогнозирования сырья с использованием градиентного бустинга над решающими деревьями.

Для достижения поставленной цели были решены следующие **задачи**:

– исследование существующих методов и систем прогнозирования;

– анализ предприятия и системы прогнозирования сырья на предприятии;

– проектирование алгоритма прогнозирования сырья на предприятии.

Предметом исследования является процесс прогнозирования закупки сырья для производственного предприятия ООО «Правильное решение». В качестве объекта исследования выступает система планирования сырья производственного предприятия.

В результате спроектирован алгоритм прогнозирования сырья, включающий две составляющие: прогнозирование спроса с помощью модели экстремального градиентного бустинга над решающими деревьями (XGBoost) и расчет сырья исходя из прогноза и собранных наборов данных об остатках сырья и продукции.

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ПО – программное обеспечение.

RNN – рекуррентные нейронные сети.

AI – искусственный интеллект.

МО – машинное обучение.

MAE (Mean Absolute Error) – средняя абсолютная ошибка.

MAPE (Mean Absolute Percentage Error) — средняя абсолютная ошибка.

MSE (Mean Squared Error) — среднеквадратичная ошибка.

RNN (Recurrent neural network) — рекуррентные нейронные сети.

ИИ — искусственный интеллект.

LSTM (Long short–term memory) — долгая краткосрочная память.

ARIMA (autoregressive integrated moving average) — модель авторегрессии скользящего среднего.

SARIMA (season autoregressive integrated moving average) – модель, которая является расширением ARIMA, включающая сезонные компоненты.

СУБД — система управления базами данных.

XGBoost – библиотека с открытым исходным кодом, используемая в машинном обучении.

KNN (k–nearest neighbors’ algorithm) — метод k–ближайших соседей.

ВВЕДЕНИЕ

В конце 2023 года в России зарегистрировано более 6 млн. предприятий и 176 крупных производственных предприятий, что на 7% и 3.5% больше, чем годом ранее. Это связано с уходом некоторых иностранных производственных компаний, что способствует развитию отечественного производства.

Актуальность темы исследования обусловлена ростом косметического отечественного производства, высокой динамичностью рынка сырья, изменчивостью потребительского поведения, а также стремительным развитием отечественного рынка программного обеспечения в сфере анализа, производства и прогнозирования. Применение передовых методов прогнозирования с использованием машинного обучения дает предприятиям конкурентное преимущество, способствуя рациональному использованию ресурсов, снижению издержек и улучшению финансовых показателей.

Проблемы с которыми сталкиваются малые и средние предприятия при выборе существующих программных решений прогнозирования спроса и расчета потребностей в сырье заключаются в высокой стоимости покупки и обслуживания программного обеспечения, отсутствие персонализации по предметной области, часто медленная поддержка, сложность интеграции с другими модулями и системами, безопасность финансовых данных. Еще важной комплексной проблемой выступает отсутствие настроенного автоматизированного процесса подготовки, сбора и обработки данных, а также быстрое старение модели прогнозирования спроса в стороннем программном обеспечении.

Прогнозирование – ключевой элемент стратегии бизнеса, влияющий на финансовые показатели. Системы прогнозирования помогают оптимизировать затраты и повысить эффективность производства, обеспечивая правильное сбалансированное количество сырья [1].

Классические методы прогнозирования имеют ряд недостатков: если данных много или мало, то страдает точность, учитываются только продажи,

но не учитываются другие факторы, влияющие на спрос, так как, например, портрет потребителя или его поведение, методы прогнозирования устаревают в связи с отсутствием учета тенденций развития политической, экономической ситуаций и активного развития искусственного интеллекта [2].

Машинное обучение способно помочь в прогнозировании потребности в сырье на высоком уровне и как результат предприятие получает повышенную точность прогноза, выявление новых тенденций, закономерностей в поведении потребителей и данных.

Предметом исследования является процесс прогнозирования закупки сырья для производственного предприятия ООО «Правильное решение».

В качестве **объекта** исследования выступает система планирования сырья производственного предприятия.

Цель исследования: спроектировать алгоритм прогнозирования сырья для производственного предприятия.

Для достижения поставленной цели исследования необходимо решить следующие **задачи**:

– исследование существующих методов и систем прогнозирования;

– анализ предприятия и системы прогнозирования сырья на предприятии;

– проектирование алгоритма прогнозирования сырья на предприятии.

Методы исследований. При проведении исследований использовались методы моделирования, анализа данных, временные ряды и алгоритмы машинного обучения, а также методы качественного, количественного и сравнительного анализа, контекстного анализа, проверка гипотез.

Научная новизна: в работе предложен алгоритм прогнозирования сырья для производственного косметического предприятия, который основан на прогнозировании спроса и анализе производственных данных о сырье. Такой подход позволит повысить точность прогнозов закупки сырья, сократить издержки и улучшить управление производственными процессами на предприятии. Алгоритм позволит учитывать множество переменных и их

взаимосвязи, что существенно повысит точность прогнозирования по сравнению с традиционными методами.

Практическая значимость полученных результатов состоит в том, что предложен алгоритм прогнозирования сырьевых запасов для ООО «Правильное решение», который позволяет с помощью метода градиентного бустинга над решающими деревьями улучшить финансовое планирование закупки сырья и сократить издержки, а также повысить производственную эффективность.

Результаты данного исследования представляют практическую значимость для прогнозирования сырья и могут быть применены в различных областях, где эффективное управление сырьевыми ресурсами играет ключевую роль. Эти области включают в себя производство косметики, фармацевтики и металлургию.

1 ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ УПРАВЛЕНИЯ ЗАПАСАМИ И ПРОГНОЗИРОВАНИЯ

1.1 Понятие прогнозирование спроса и модели управления запасами

Процесс управления запасами – это комплексная задача, направленная на эффективное использование запасов, сбалансированное с потребностями клиентов, заказчиков и минимизациями затрат. Это включает в себя оптимизацию процесса транспортировки, хранения, распределения, покупки, контроля расходов, а также расчет потребностей материалов, сырья и запасов для производства необходимого объема продукции. Нарушения приводят к избытку или дефициту [3].

Эффективное управление производством и логистикой базируется на взаимосвязи трех процессов: прогнозировании спроса, управление запасами и управление поставками. Прогнозируя спрос, компании определяют объем производства и необходимый уровень запасов, чтобы затем оптимизировать поставки и обеспечить бесперебойное выполнение заказов.

Существует множество политик управления поставками (Приложение А), в которые входит управление закупками. Ключевыми компонентами управления запасами являются ABC–анализ, прогнозирование спроса, своевременная инвентаризация (JIT), экономичное количество заказа (EOQ), MRP, бережливое производство, Шесть Сигм, теория ограничений (ToC) и DDMRP – все это является частями комплексного подхода к управлению запасами. ABC–анализ направлен на разделение товаров по степени их ценности, что позволяет приоритезировать управление наиболее значимыми из них.

Все чаще исследования сосредоточены на разработке более сложных моделей управления запасами [4], которые учитывают неопределенность и различные факторы, влияющие на управление запасами. Исследование [5] показало, что модель управления для запасов, которые имеют высокую степень изменчивости превзошла другие по затратам на хранение и дефициту.

Все больше исследований [6] в последнее десятилетие фокусируется на разработке математических моделей и использовании методов искусственного интеллекта для оптимизации запасов и ценовых решений для максимизации прибыли.

Ильенкова Н.Д. рассматривала спрос как отображение объема (количества) продукции, которую покупатель и заказчик может приобрести в течение определенного времени и периода [7].

Спрос на запасы – сумма величины расхода по факту запасов за промежуток времени и величина дефицита запасов за тот же период.

Прогнозирование – вероятностное предположение, исследование перспектив развития изучаемого процесса, объекта в будущем. Прогнозирование выступает неотъемлемой частью качественного планирования запасов сырья [8]. Прогнозирование как предсказание будущего по мнению авторов [9] предназначено для принятия решения в бизнесе, которое будет обосновано и подкреплено статистически.

Прогнозирование спроса – это прогнозные оценки, моделирование будущей структуры спроса, основанной на исследованиях, отношениях, причинах, тенденциях и обоснованных научно моделях прогнозирования [10].

Спрос тесно связан с прогнозированием запасов. Это процесс оценки будущего спроса на производственные ресурсы, играющий ключевую роль в обеспечении эффективности, рентабельности и конкурентоспособности предприятия.

По исполняемым функциям запасы подразделяются: текущие, подготовительные (буферные), гарантийные (страховые или резервные), сезонные, переходящие.

Необходимость хранить и обеспечивать определенный товарно–материальный запас продиктован регулярными колебаниями спроса, контролем производственных мощностей, создание страхового запаса в связи с возможными проблемами с логистикой, возможностью сокращения малых заказов.

Сырье представляет собой ключевые материалы для производства товаров, включая как натуральные ресурсы (минералы, нефть, дерево), так и обработанные продукты (металлы, пластик, химикаты, растительные компоненты, эмульгаторы, поверхностно–активные соединения, красители и силиконы). Основываясь на прогнозах спроса, который формируется с помощью моделей и методологий прогнозирования, предприятия могут формировать стратегию закупок, улучшая производство и сокращая расходы.

1.2 Модели и методологии прогнозирования спроса

Методика прогнозирования представляет собой приемы и специальный комплекс правил для подготовки и реализации прогноза. Модели и методологии прогнозирования давно изучаются исследователями по всему миру российскими исследователями: Мазманова Б.Г., Гаврилов Н.П., Бушуева Л.И., Анискина Ю.П., Беляевский И.К., Анискина Ю.П., а также иностранными исследователями: Джон Шрайбфедер, Chen Wei, Maria Garci, John Doe, Armstrong J.S., Baker M.J. и другие.

Представленная Мазмановой Б.Г. типология прогнозов часто используется как основа методологического деления прогнозов по признакам (рисунок 1) [6].

Рисунок 1 – Типология прогнозирования

Прогнозирование варьируется от краткосрочного для оперативных задач до среднесрочного для стратегического планирования и долгосрочного для адаптации к будущим инновациям.

Экстраполятивное планирование предсказывает будущее на основе прошлых данных. Вариантное прогнозирование готовит к разным ситуациям, а инвариантное выявляет устойчивые шаблоны для долгосрочных стратегий. Точечное прогнозирование дает конкретный ответ, интервальное – диапазон вероятных результатов с учетом неопределенности. Точечное прогнозирование полезно для планирования производства и закупок, так как предоставляет конкретные значения спроса.

Также прогнозирование может быть типизировано по масштабу: частные, местные, региональные; отраслевые: по стране, мировые и по авторству: личные, на уровне предприятия, на уровне государственных органов.

Существует разнообразие методов и подходов к формулированию моделей прогнозирования [11]. Но по мнению автора исследования Михайловой Е. Б. [12] ряд исследователей смешивают инструменты и методы прогностической методологии и деления, а также не исследуется вся широта существующих методов. В общем случае деление методов прогнозирования представлено на рисунке 2.

Рисунок 2 – Классификация методов и моделей прогнозирования

Качественные методы способны помочь в изучении объекта с помощью неточной экспертной вариативной деятельности. Основная сложность использования данных методов – это формализация полученных субъективных результатов.

Метод сценариев предполагает создание нескольких возможных сценариев развития событий для анализа и прогнозирования. Метод интервью включает проведение бесед с экспертами или участниками исследования для получения ценных данных и мнений. Аналитические доклады позволяют изучить информацию и отчеты для выявления тенденций и выводов. Метод анкетирования основан на сборе структурированных ответов на вопросы для анализа мнений и предпочтений опрошенных. Метод «Дельфи» представляет собой процесс коллективного экспертного прогнозирования через серию итераций. Метод мозгового штурма способствует генерации новых идей и решений через коллективное обсуждение. Метод «комиссий» предполагает формирование специальных групп для принятия решений и разработки рекомендаций по заданной проблеме. Каждый из этих методов обладает своими особенностями и применим в зависимости от поставленной задачи и целей исследования [13].

Комбинированные методы представляют собой сочетание качественных и количественных методов, а также использование экспертных оценок для определения весов факторов в каузальной модели. Адаптивный подход предполагает частую корректировку прогнозов с учетом новых данных и изменения факторов.

Количественные методы представляют собой способы изучения объекта, товара или услуги для подготовки предположения о его будущем поведении, состояни. Эти способы строятся на основе моделирования прогноза и с использованием предиктивного моделирования, математических различных моделей.

Можно также обратить особое внимание на классификацию методов и моделей прогнозирования, представленную в работе [14] и, объединив с

исследованием Мамонтовым Д.В. и Селезневым С.В. [15], получается схема представленная на рисунке 3. Она выглядит обобщенно и научно обоснованно.

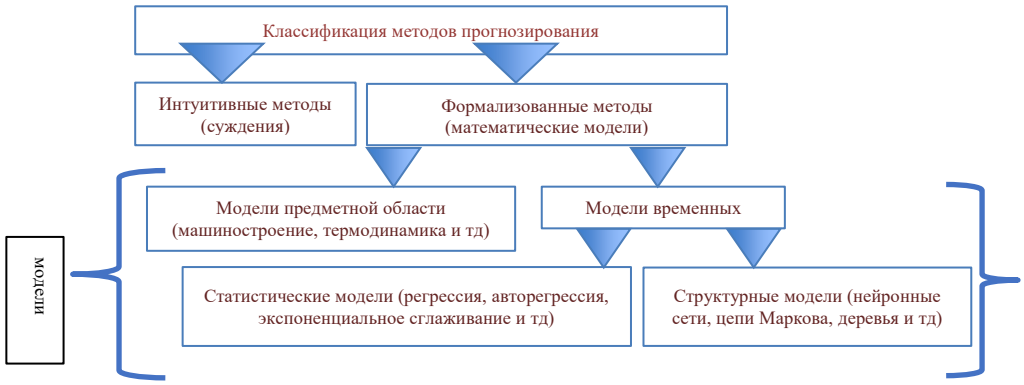

Рисунок 3 – Классификация методов прогнозирования

Интуитивные методы суждения были описаны ранее и включают в себя метод «Дельфи», сценарный метод, метод анкетирования и многие другие.

Формализованные методы включают в себя модели предметной области и временные ряды, которые играют важную роль в построении будущих прогнозов. В свою очередь временные ряды условно делят на два вида: статистические модели и структурные модели. Важно отличать модель от метода, так как метод – это последовательность шагов для построения модели и для прогностического моделирования [16]. Статистические модели делятся на регрессионные, авторегрессионные и модели экспоненциального сглаживания.

Временной ряд представляет собой последовательность измерений, полученных в различные временные точки, организованных в хронологическом порядке. Временные ряды делятся по времени – моментные и интервальные; по форме представления уровней – ряды абсолютных, относительных и средних величин; по расстоянию между датами или интервалами времени – полные и неполные; по содержанию показателей – частных и агрегированных показателей [17].

Существует большое количество моделей и подходов в прогнозировании. Простые статистические модели прогнозирования

основаны на предположении, что будущее будет повторять прошлое. Они подходят для стабильных рынков, примеры таких моделей арифметическое среднее, медиана, мода. Несмотря на свою наивность, такие модели могут быть полезны. Она предполагает сохранение текущей тенденции и использует модели, такие как линейная, логарифмическая регрессия и трендовый анализ [18].

Корреляционный метод используется для установления наличия и силы взаимосвязи между двумя и более переменными. Он позволяет определить, какие переменные влияют на изменения других переменных, и может использоваться для прогнозирования значений переменных. Однако, он не позволяет установить причинно–следственные отношения между переменными. К основным моделям относится коэффициент корреляции Пирсона, Спирмена, Кендалла и так далее.

Метод регрессионного анализа используется для построения уравнения регрессии, которое позволяет оценить влияние независимых переменных на изменения зависимой переменной. Он позволяет построить модель для прогноза значений переменных. Здесь могут быть применены модели схожие с моделями экстраполяции, а также модели многомерных и иерархических регрессий.

В анализе временных рядов существует множество статистических моделей для прогнозирования. Простые модели, такие как SMA (Simple Moving Average) и EMA (Exponential Moving Average), подходят для краткосрочных прогнозов и сглаживания рядов. Более сложные модели, такие как AR (Autoregressive) и MA (Moving Average), учитывают прошлые значения ряда и ошибки прогнозирования соответственно.

Для нестационарных временных рядов (с трендом) используется модель ARIMA (Autoregressive Integrated Moving Average), которая включает этап интегрирования для устранения нестационарности. Модель SARIMA (Seasonal ARIMA) расширяет ARIMA, добавляя сезонные компоненты для рядов с периодическими колебаниями.

Модели ARIMAX и SARIMAX учитывают влияние экзогенных переменных (внешних факторов) на ряд. ARIMAX подходит для рядов без сезонности, а SARIMAX – для сезонных рядов. В статье [19] авторы исследования показывают как ARIMA может быть эффективна в прогнозировании спроса.

Отдельное внимание хочется остановить на модели Prophet, которая предназначена для прогнозирования временных рядов с помощью разложения ряда на сезонность, тренд, праздничные и выходные события логистическими кривыми с настраиваемыми параметрами и отлично работает с большими наборами данных.

Модели экспоненциального сглаживания (ETS) представлены простым экспоненциальным сглаживанием для рядов без тренда и сезонности, двойным экспоненциальным сглаживанием (метод Хольта) с учетом тренда, тройным экспоненциальным сглаживанием (метод Хольта–Винтерса) – учитывает как мультипликативную, аддитивную сезонность, так и тренд.

Машинное обучение является подмножеством искусственного интеллекта. Прогнозирование на основе машинного обучения получило первое упоминание в 1964 году и сейчас получило одно из базовых мест в прогнозировании [20]. Прогнозирование спроса является важной задачей, для решения которой были исследованы различные методы и алгоритмы. Среди них изучены многие алгоритмы машинного обучения, такие как K–ближайший сосед, гауссов наивный Байес, деревья решений и многие другие. Эти алгоритмы позволяют строить прогнозы будущего спроса, основываясь на анализе исторических закономерностей в данных временных рядов [21].

Павлышенко Б.М. в своем исследовании [22] изучил возможность использования машинного обучения в прогнозировании и аналитике продаж, показывая, что временные ряды не всегда обоснованный выбор для небольшого объема исторических данных.

Чаще всего прогностическая модель, построенная на методах машинного обучения, учитывает сезонность, цену товара, спрос клиентов, и

его скидка в рамках акции, остатки на складе, и будущее спланированные поставки товара в сети магазинов [23].

Линейная регрессия представляется как метод для моделирования линейных отношений между зависимыми и независимыми переменными. Процесс построения модели включает в себя подгонку прямой линии, чтобы минимизировать сумму квадратов расстояний каждой точки данных к этой линии, то есть после построения модели можно прогнозировать, например, количество продаж (спрос) на основе независимых переменных, например, сезонность, рекламный бюджет [24].

Деревья решений представлены иерархической структурой, которая используется для принятия решений на основе различных условий. Они могут применяться в экономике для классификации и прогнозирования. Ансамбль из деревьев представляет собой случайный лес, в котором используется совокупность деревьев и чем больше правил, тем глубже дерево.

Метод К–ближайший сосед (KNN) использует методологию поиска сходства путем создания индекса и поиска близких k кортежей основному и делит на кластеры исходя из сходства [21].

Ансамблевое обучение комбинирует прогнозы нескольких моделей для повышения точности и надежности результатов. Это полезно для прогнозирования экономических переменных и управления рисками.

Среди ансамблевых моделей с последовательным или параллельным обучением алгоритмов выделяют бэггинг, бустинг, стекинг и блендинг. Бэггинг предполагает обучение множества базовых моделей на разных подвыборках данных, полученных с помощью бутстрэпа, и усреднение их предсказаний, что снижает дисперсию и помогает бороться с переобучением. Бустинг итеративно обучает последовательность слабых моделей, каждая из которых старается исправить ошибки предыдущей, формируя взвешенную комбинацию для снижения смещения предсказаний. Стекинг и блендинг комбинируют предсказания разнородных базовых моделей с помощью мета–

модели, но в стекинге она обучается на предсказаниях базовых, а в блендинге непосредственно на признаках, что снижает риск переобучения [17].

Градиентный бустинг представляет собой обобщение других методов бустинга, так как позволяет оптимизировать любую дифференцируемую функцию потерь. Алгоритм базируется на методе градиентного спуска, широко используемый в задачах оптимизации. В основе подхода лежит концепция последовательной оптимизации, где каждый последующий алгоритм в наборе моделей (ансамбле) фокусируется на улучшении точности, исправляя погрешности, допущенные его предшественниками. Этот процесс осуществляется путем обучения каждой новой модели на остаточных ошибках, полученных от предыдущих моделей в цепочке. Метод градиентного спуска, лежащий в основе градиентного бустинга, наглядно демонстрирует, как следующая модель корректирует ошибки предыдущих, постепенно улучшая общее предсказание ансамбля (рисунок 4) [25].

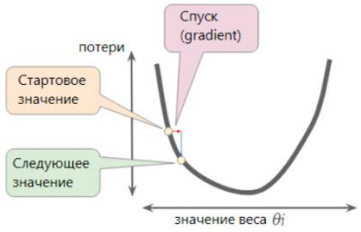

Рисунок 4 – Принцип работы градиентного спуска

У каждого из приведенных методов машинного обучения есть свои преимущества и свои недостатки, они описаны в Приложении Б. Глубокое обучение использует нейронные сети с большим количеством слоев для автоматического извлечения сложных признаков из данных и скрытых представлений. Одной из основных проблем последние десятилетия стало включение трендов и сезонных колебаний в модели прогнозирования. Традиционные методы, такие как ARIMA, часто не справлялись с нелинейными данными, поэтому в поисках решения исследователи обратили внимание на машинное обучение [26].

Глубокое обучение, часть машинного обучения, позволяет выявлять скрытые представления в данных. Рекуррентная нейронная сеть (RNN), разработанная в 1980–х, широко используется в прогнозировании спроса. За последние два десятилетия нейронные сети претерпели значительные изменения в области прогнозирования, причем Transformer стал одной из наиболее успешных архитектур согласно многим исследованиям [27].

На рисунке 5 представлены основные архитектуры нейронных сетей, используемых в современном прогнозировании временных рядов и их развитие от RNN до новых архитектур, базирующихся на Transformer.

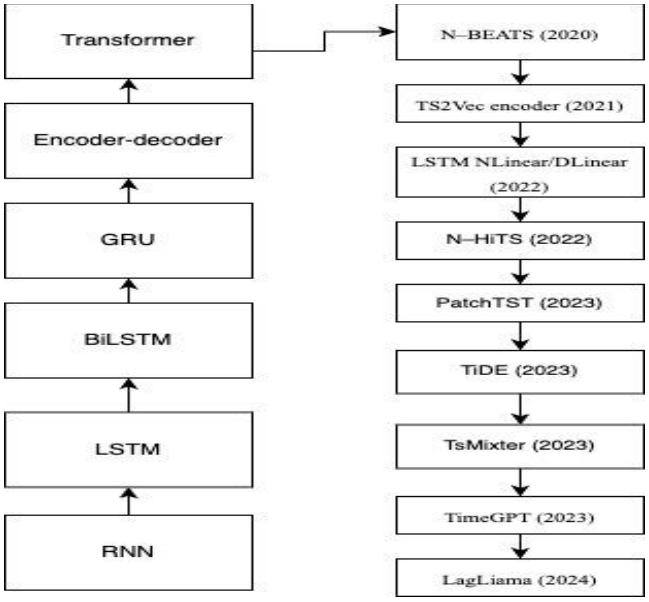

Рисунок 5 – Модели архитектур нейронных сетей для временных рядов

В последнее время большой интерес исследователей привлекают подходы глубокого обучения. К ним относятся рекуррентные нейронные сети (RNN), архитектуры с длинной краткосрочной памятью (LSTM) и закрытыми рекуррентными единицами (GRU), автокодировщики, а также сверточные нейронные сети (CNN). Применение этих методов позволяет добиться более высокой точности прогнозов по сравнению с традиционными алгоритмами машинного обучения [28]. Архитектура RNN базируется на обработке последовательностей и одна из основных проблем RNN является исчезающий или взрывающийся градиент, что особенно заметно при работе с длинными

последовательностями. LSTM изначально был базовым блоком для работы рекуррентных нейронных сетей (RNN) для временных рядов и имел 6 обучаемых функций в каждой из которых было по 2 матрице параметров, в итоге получалось 12 матриц для обучения. Такая сеть не быстро учится, и она забывает быстро, подходит для небольшого набора данных [29].

LSTM превосходит по производительности ARIMA, если используется двунаправленная архитектура BiLSTM за счет двойного перемещения всех данных, которые поступают на вход слева направо и справа налево [30]. На смену архитектуре долгой краткосрочной памяти (LSTM) пришел механизм вентилей для рекуррентных нейронных сетей (GRU). Имеет всего 4 обучающие функции, за счет этого быстрее учится и использует меньше памяти при схожем качестве прогнозирования с LSTM.

Трансформеры (Transformers) – это архитектура, представленная в 2017 году, которая позволяет учитывать долгосрочные зависимости, автоматически извлекает признаки и хорошо обрабатывает внешние факторы. Архитектура состоит из следующих шагов: предобработка данных, позиционное кодирование, механизм внимания, полносвязные слои, повторение слоев, предсказание и обучение [31].

В исследовании [32] авторами предлагается новая модель LSTM DLinear и Nlinear, взамен популярным RNN. Эта архитектура базируется на том, что вместо традиционной линейной операции входного, забывающего и выходного вентилей LSTM использует нелинейную активационную функцию, при этом забывающий вентиль остается линейным, тем самым улучшена способность работы со сложными нелинейными зависимостями.

Модель TS2Vec encoder подход к векторному представлению временных рядов, основан на архитектуре Transformer. Из основных преимуществ стоит выделить универсальность, так как работает архитектура с различными типами данных, а также контекстуальность и масштабируемость модели [33].

Модель прогнозирования временных рядов N–BEATS, разработанная в 2020 году Element AI состоит из двух блоков преобразование и линейной комбинации. Преобразование включает нелинейные слои для распознавания сложных зависимостей в данных. Комбинация применяет взвешенные параллельные блоки преобразования для уменьшения ошибки прогнозирования [34].

Модель N–HiTS является иерархической моделью прогнозирования спроса, представленная в 2022 году, которая учитывает взаимосвязи между различными уровнями иерархии и использует механизмы внимания для передачи информации между уровнями [35].

PatchTST является моделью, основанной на трансформерах с использованием патчей. Она разбивает временные ряды на патчи фиксированной длины и обучает трансформеры на этих патчах, позволяя модели учитывать локальные паттерны и контекст временных рядов [36].

Модель нейронной сети TiDE объединяет дифференциальные уравнения, которые используются для моделирования динамики временных рядов и трансформеры для учета контекстуальной информации [37].

Модель нейронной сети TsMixter объединяет различные типы временных рядов с помощью миксерной архитектуры, позволяя моделировать зависимости между разными типами временных рядов и учитывать их взаимное влияние на прогнозирование [38].

Модель нейронной сети LagLiama использует архитектуру Llama и включает механизм задержки (lag). Она учитывает зависимости между временными рядами с различными временными задержками и позволяет моделировать долгосрочные зависимости [39].

Модель TimeGPT представляет собой генеративную модель, которая уже обучена на большом количестве разных временных рядов. Она разработана Azul Garza и Max Mergenthaler в 2023 году и является прорывом в области прогнозирования. На текущий момент она работает только по API, для коммерческих целей только платная версия [40].

Модель нейронной сети TimesNet базируется на временных рядах с множественной периодичностью, модульной архитектуре и возможностью фиксации в 2D–пространстве с первичным настраиваемым блоком параметров с преобразованием Фурье для поиска периодов и разделении внутрипериодных и межпериодных вариаций. Одномерные временные ряды преобразуются в набор двумерных тензоров и с использованием нескольких отрезков времени и отправляются в блок изучения и прогнозирования, потом аддитивной агрегацией делается обратное преобразование [41].

Недостатком в использовании нейронных сетей остается необходимость большого количества исторических данных, положительным является то, что они способны находить скрытые зависимости, надежные и стабильны по отношению к шумам [42].

Основными моделями прогнозирования широко используемыми на предприятиях являются экспоненциальное среднее, модель Хольта–Винтерса, парная регрессия, множественная регрессия, метод k–ближайших соседей, модель Бокса–Дженкинса (ARIMA) и SARIMA [43]. Было даже предложено прогнозирование с помощью изображения временного ряда и обоснована эффективность данного метода в статье [44].

Особо место занимает вероятностное прогнозирование (методология DeepAR) для большого количества временных рядов, оно позволяет получать оценку распределений во временном ряду, учитывая его факторы и исторические данные.

Каждая модель имеет программную оболочку в виде модуля, программы или иной реализации для взаимодействия с пользователем, сервером и другими системами предприятия. Далее уделим внимание изучению рынка цифровых сервисов и программ, представленных на рынке на текущий момент.

1.3 Цифровые сервисы и программы прогнозирования спроса и запасов

Малые предприятия долго использовали MS Excel для прогнозирования потребностей в сырье, но технологическое развитие толкает их к созданию более специализированных и гибких систем. Каждое предприятие организует собственную архитектуру в соответствии с финансовыми возможностями, предметной областью и производственными процессами. Далеко не все компании начинают свой путь с использования специализированных программных продуктов и пакетов, таких как системы управления цепочками поставок (SCM) или вспомогательные системы планирования и составления расписаний (APS) [45].

В качестве инструментов прогнозирования могут выступать электронные таблицы (Excel и другие офисные пакеты), статистические пакеты (SPSS), среды имитационного моделирования (AnyLogic, Matlab Simulink), а также алгоритмы машинного и глубокого обучения. Как правило, система прогнозирования является частью единого программного обеспечения или системы управления предприятием и взаимодействует как модуль или подсистема с другими компонентами, передавая результаты в другую систему или модель [46].

Информационные системы, предназначенные для прогнозирования спроса и планирования материальных потребностей, можно условно разделить на несколько категорий:

- системы управления ресурсами предприятия (ERP) с функцией прогнозирования спроса;
- вспомогательные системы планирования и составления расписаний (APS), которые интегрируются в ERP-системы;
- системы планирования производственных ресурсов (MRP), включающие в себя прогнозирование спроса и расчет необходимых материалов;

- системы управления цепочками поставок (SCM), которые обеспечивают управление потоками на предприятии и создание актуальных планов;
- информационно-аналитические пакеты и интегрированные системы бизнес-планирования, предоставляющие настраиваемые сервисы для прогнозирования спроса.

Сегментация программного обеспечения для прогнозирования спроса и потребности в сырье может осуществляться по следующим критериям:

– компонентность (предоставление как услуги или готового решения);
– размер предприятия и тип конечного пользователя;
– регион использования;
– развертывание (локальные решения или облачные сервисы).

Среди наиболее популярных зарубежных программных решений можно выделить: Oracle Demantra, SAP Integrated Business Planning (IBP), SAP Analytics Cloud, SAS Demand Planning, Kinaxis RapidResponse, Anaplan, JDA Inventory Optimization, JDA Software, LOKAD, Logility Solutions, Demand Management, GAINSystems, Arkieva, John Galt Solutions, Gong.io, Aviso, BoostUp.ai, Alteryx.

Существует множество отечественных реализаций в сфере ПО по управлению запасами: Knoweledge Space (ООО «Интегрированные системы управления»), Loginom Planiqum Suite (ООО «Решейп Аналитикс»), Novo Forecast Enterprise (ООО «Ново Биай»), Optimacros (ООО «Оптимакрос»), In.Plan (ООО «Акстим Тех»), Система интегрированного планирования IPS (АО «Северсталь–Инфоком»), GoodsForecast Integrated Planning Platform (ООО «Гудфокаст»), Форсайт Аналитическая платформа (ООО «Форсайт»), «Deductor», «Прогноз» и написано ряд статей о представленных на рынке решениях авторами Н. Б. Паклина, В. И. Орешкова, Ш. Акобира, Бариновой О. В., А. А. Грицая и многих других.

Решения компаний Norbit, Галактика АММ, Forecast Now!, Корус Консалтинг, Planetra и 1С предлагают функциональность для прогнозирования спроса и управления запасами на базе статистических и машинно–обучаемых моделей, ориентированную на российские предприятия. Сравнение систем и цифровых сервисов приведено в Приложении В. Цифровые сервисы предлагаются как готовые решения и облачные платформы, модульные, кастомизированные или монолитные.

Последнее время наблюдается значительный рост интереса бизнес–сообщества к отечественным разработкам в области прогнозирования и планирования, обусловленный приостановкой продаж и использования иностранного ПО в России.

Несмотря на широкий выбор, для малых предприятий и стартапов стоимость часто оказывается высокой. Рынок программных продуктов растет более чем на 11% в год, активно развиваются новые модели прогнозирования, включая искусственный интеллект, например, модель TimeGPT. В следующем разделе рассматривается процесс производства и система прогнозирования сырья в ООО «Правильное решение», сбор наборов данных, анализ наборов данных, подбор моделей для краткосрочного прогнозирования спроса и метрик для анализа эффективности данных моделей.

Выводы по первому разделу:

В главе рассмотрено понятие прогнозирования, существующие методики и модели прогнозирования, включая виды и типы программного обеспечения, цифровых сервисов для прогнозирования спроса и потребности в сырье, представленных на рынке. Для прогнозирования продаж используются разные методы и модели, включая традиционные статистические модели, такие как ARIMA, SARIMA и современные модели машинного и глубокого обучения, такие как проверенные LSTM, Fb–Prophet и более современные новые N–Beats, Transformers, LagLiama.

2 КОМПЛЕКСНЫЙ АНАЛИЗ И ПРОГНОЗИРОВАНИЕ СЫРЬЯ НА ПРЕДПРИЯТИИ ООО «ПРАВИЛЬНОЕ РЕШЕНИЕ»

2.1 Анализ предприятия ООО «Правильное Решение»

Ежегодный темп роста мирового рынка косметической продукции и сырья оценивается в 5,1 %. Категория «Красота и здоровье» входит в топ–5 российского рынка по обороту и составила 383,5 млрд. руб. в 2022 году. Из–за падения импорта косметической продукции на 80% наблюдается стремительный рост российских производителей, которые наращивают производственные мощности, расширяя ассортимент.

Общество с ограниченной ответственностью «Правильное решение» (далее предприятие) создано в 2014 году. Предприятие является современным и высокопроизводительным производством косметических средств. На текущий момент предприятием производится более 700 видов косметических средств: шампуни, гели, кремы, бальзамы, сыворотки, скрабы и многое другое, в портфеле более 6 брендов. Основной зарегистрированный вид деятельности: розничная и оптовая торговля фармацевтической продукцией, изделиями, применяемыми в медицинских целях, парфюмерными и косметическими товарами, включая мыло, и чистящие средства. Организационная структура предприятия представлена в рисунке 6.

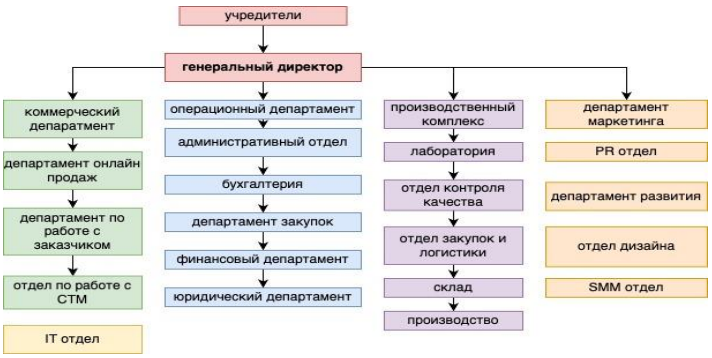

Рисунок 6 – Организационная структура предприятия ООО «Правильное решение»

Все отделы предприятия подчиняются генеральному директору, средняя численность персонала около 100 человек. Общество реализует готовую

продукцию в оптово–розничном сегменте через компании региональной дистрибуции и сети магазинов, на торговые площадки и маркетплейсы и на заказ. Продукция реализуется на территории России, с помощью разных каналов сбыта: маркетплейсы, сеть магазинов «Магнит», «Золотое яблоко», интернет–магазин.

Важный показатель, характеризующий деятельность предприятия – это валовая прибыль. Планируемая валовая прибыль предприятия в 2024 году 600 млн. руб., в 2025 году более 1 млрд. рублей. Пополнение оборотного капитала планируется посредствам расширения каналов сбыта, повышения эффективности маркетинговых кампаний, создание и автоматизация процессов, включая прогнозирование спроса для повышения производственной эффективности и расчета сырья.

Производство предприятия расположено в Московской области, пос. Часцы. В январе 2024 года предприятие переехало на новую производственную площадку размером 3700 кв.м., это позволит увеличить объем производства в связи с ростом спроса на продукцию. Предприятие включает 4 цеха производства и 5 производственных линий. Ежемесячная производственная мощность составляет до 3 млн. единиц продукции при участии более 50 сотрудников производственного комплекса и склада.

Процесс производства косметики начинается с разработки и тестирования новых формул в лаборатории. После утверждения и сертификации, рецептура переходит в производство для запуска партии и в отдел закупок для обеспечения сырьем. Затем следует проверка и подготовка сырья, его варка и превращение в полуфабрикат (ангро), который тестируется на соответствие стандартам. Далее ангро упаковывается, этикетируется и проверяется на финальном контроле качества. Готовая продукция поступает на склад, откуда распределяется по заказам клиентов.На всех стадиях производства проводится строгий контроль качества, обеспечивающий соответствие продукции санитарно–гигиеническим нормам и требованиям (Приложение Г).

В представленном процессе производства обязательным остается прогнозирование потребности в сырье для производства как первый производственный этап. Далее будет проанализирован данный процесс на предприятии.

2.2 Анализ процесса прогнозирования сырья на предприятии

Прогнозирование объемов продаж не производится, хотя является первым этапом планирования потребности в оборотных активах, основываясь на прогнозе спроса и производственных объемах.

Определение потребности в материальных ресурсах производится руководителем производства после получения информации об объеме продаж за последнюю неделю, а также учитываются поступившие уже заказы потребителей и заказчиков по средствам выгрузки отчета из 1С:Предприятие.

Руководитель производства каждую среду получает также от финансового отдела информации о наличии денежных средств для покупки сырья и упаковочного материала.

Расчет необходимого сырья и материалов основан на уже существующих заказах в 1С:Предприятие и отгрузок предыдущей недели, далее рассчитывается на следующую неделю необходимое количество для производства, учитывая норму расхода материалов, сырья и размер остатков. Этот оформляется как отчет в 1С: Предприятие и служит как план по закупке сырья на следующую неделю.

На текущий момент компания использует стратегию управления запасами – периодическое пополнение с экстренными поставками, что связано с отсутствием достаточного количества денежных средств для формирования страхового запаса и отсутствие системы прогнозирования.

Основной текущей проблемой на предприятии в области планирования и прогнозирования производственного процесса и управления закупками является отсутствующий процесс прогнозирования спроса для расчета

потребности сырья для будущих производственных нужд, как следствие, дефицит финансирования и сложности в планировании производственных процессов, включая загрузку производственных мощностей.

Для оценки возможности решения текущих проблем по прогнозированию сырья посредствам покупки программного обеспечения были исследованы программные решения на рынке, представленные в Приложении В и произведено сравнение их с потребностями предприятия.

Все представленные решения являются дорогими и не учитывают предметную область предприятия, не включают реализацию конвейера по сбору, очистке и подготовке данных для последующей подачи в модель, работающую в модуле для прогнозирования потребности в сырье. Но и без программного решения предприятию не обойтись в 2024 году так как разрастается цепочка поставок и растет спрос на продукцию, и текущая схема закупок не учитывает чувствительность факторов, которые влияют на спрос.

Исходя из анализа текущего процесса закупок сырья на предприятии необходимо спроектировать и предложить алгоритм прогнозирования сырья, начиная с создания модели прогнозирования спроса и заканчивая краткосрочным расчетом потребности в сырье для производственных целей. Данное решение после проектирования алгоритма прогнозирования сырья должно быть автоматизировано на всех этапах.

2.3 Подготовительные этапы для прогнозирования сырья

Первоначально задача состоит в сборе набора данных и подборе наилучшей модели на основании метрик оценки качества моделей для последующего проектирования алгоритма с целью краткосрочного одношагового прогнозирования спроса и расчета потребности сырья для отдела закупок и производственного отдела на представленном предприятии.

Важно отметить, что основа для прогнозирования спроса – это качественные данные, правильно подобранная методология и модель

прогнозирования спроса. Далее будет рассмотрен этап сбора и предобработки наборов данных для построения и выбора модели.

2.3.1 Сбор и предобработка данных для прогнозирования

Источником данных послужила внутренняя бухгалтерская и оперативная отчетность, экспортированная из системы 1С:Предприятие в формате Excel. Данные прошли предварительную обработку, обеспечивая базис для последующего аналитического процесса.

Для сбора набора данных были выгружены из 1С:Предприятие и иных источников следующие отчеты за 2021 (с 15.06.2021) – 2024 год (по 12.03.2024) представленные на рисунке 7.

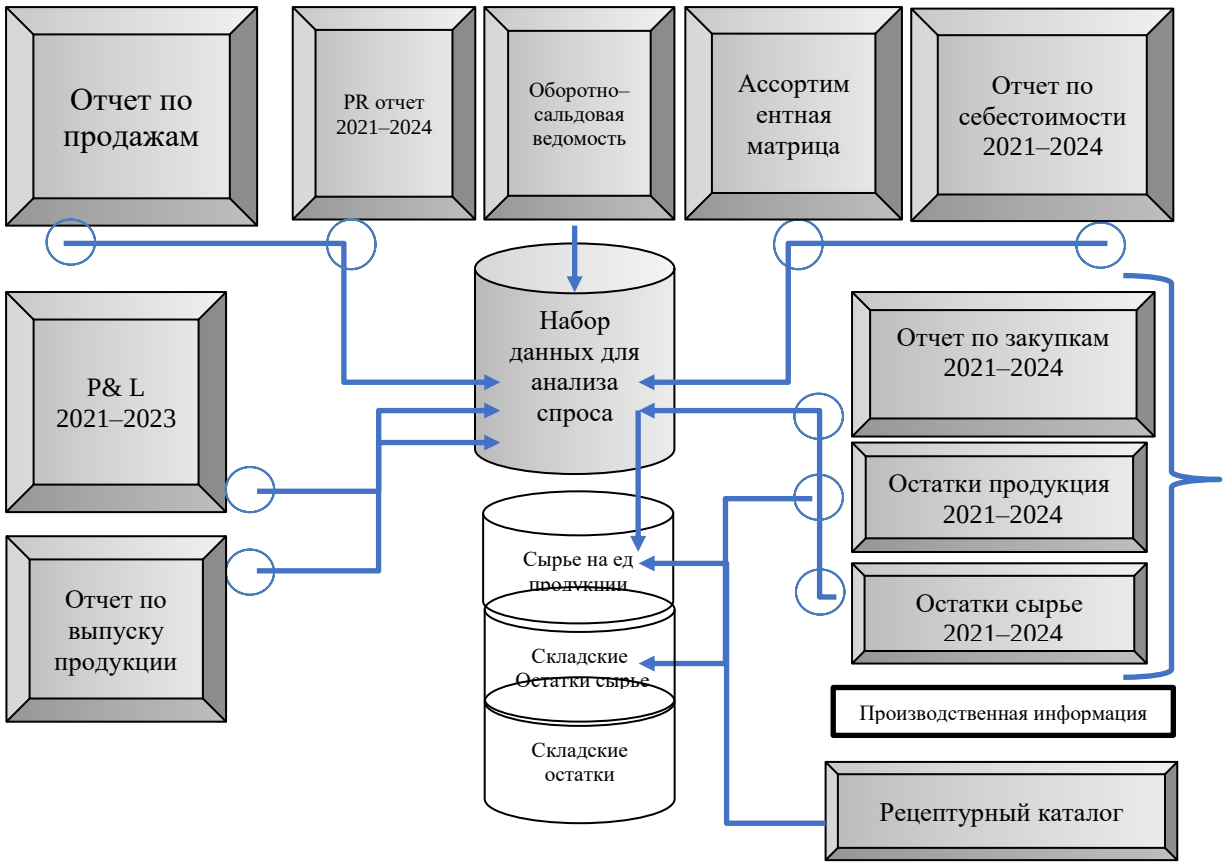

Рисунок 7 – Источники данных для сбора набора данных

Первый этап состоит из сбора набор данных для прогнозирования спроса. Для прогнозирования спроса данные были собраны из разных отчетов,

их описание раскрыто подробнее в Приложении Д. Сопоставление источников данных представлено на рисунке 8.

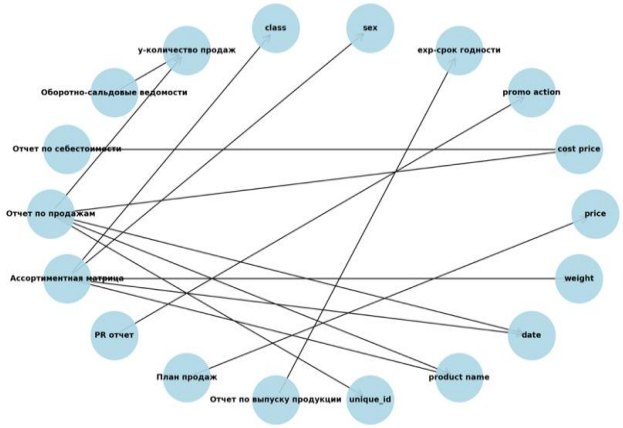

Рисунок 8 – Схема связей между источником информации и колонкой в наборе данных

В рамках формирования структуры набора данных для прогнозирования и анализа спроса, следует отметить, что массив информации обладает достаточной полнотой и репрезентативностью. Обрабатывая и добавляя информацию о продукции из разных источников, таких как отчеты по продажам, ассортиментные матрицы и данные о себестоимости, получается набор данных, который позволит проводить глубокий анализ факторов, влияющих на динамику сбыта и подбирать, моделировать прогнозирование спроса лучшим образом.

Уникальные идентификаторы продуктов (unique_id) служат связующим звеном между различными массивами данных. Важным аспектом является тайм–код, представленный в виде дат продаж, который позволяет исследовать сезонные колебания и тенденции.

Набор данных включает наименование продукции (product name) и уникальный идентификатор (unique_id), извлекаемые из отчетов по продажам, ассортиментной матрицы и ежегодного отчета по себестоимости. Для анализа товаров представлены «сегмент рынка» и «категория товара», а также временные элементы, такие как «date» (дата), «месяц» и «day_of_week» (день недели). Атрибуты товара, включая «weight» (вес), «price» (цена продукции),

«Себестоимость» и «Срок годности продукции», помогают понять ценообразование и стоимость товара.

Промо–активности представлены в колонке «promo action», а демографические факторы, такие как «sex» (пол) и «class» (класс продукта), дают представление о поведении потребителя. Дополнительные атрибуты, включая «количество компонентов», «Сложность изготовления» и «рейтинг товаров», дают представление о сложности производства и популярности продукции. «Количество просмотров» и «конкуренция» взяты из отчетов анализа продаж и маркетинговых отчетов.

Средние продажи за 3, 6 и 12 месяцев (sales_3m_avg, sales_6m_avg, sales_12m_avg), скидки (discount, seasonal_discount) позволяют оценить влияние различных факторов на объем продаж за недельный период, который является целевой переменной (y). «Функция» товара поступает из ассортиментной матрицы и представляет основную характеристику продукции. Этот набор данных способствует эффективной реализации моделей машинного обучения.

По итогу сбора набор данных, составлен краткий мета–отчет с информацией о дате и времени сбора данных, их целостности, структуре.

Второй этап состоит из сбора информации для расчета сырья на основании сделанного моделью прогноза из следующих источников:

– отчет по закупкам сырья. отчет содержит историческую информацию о сырье и закупках: цена, дата, объем закупки, потребление;

– остатки продукция. этот отчет содержит информацию об остатках продукции на складе;

– остатки сырья, этот отчет содержит информацию об остатках сырья на складе. всего на предприятии используется около 781 наименования сырья;

– рецептурный каталог, который содержит информацию о количестве и массе компонентов в единице продукции, сложности изготовления, длительности производства.

Набор данных по содержанию сырья в единице продукции содержится в рецептурной карте и состоит из артикула, наименования продукции, номенклатуры сырья, количества сырья и единицы измерения. Набор данных остатки сырья содержит уникальный идентификатор сырья, наименование сырья, остаток, выраженный количеством и единицу измерения сырья. Набор данных остатки продукции содержит уникальный идентификатор продукции, наименование продукции, количество.

Ожидается, что после внедрения системы расчета потребности в сырье наборы данных будут регулярно обновляться перед расчетом сырья и модифицироваться в соответствии с изменяющимися условиями рынка и подбором новых признаков или данных.

Одним из преимуществ предложенной схемы и подготовки данных является панельный тип данных, который характеризуется как данные о 212 объектах (продуктах), измеренные в течение 144 недель и может быть использован в исследованиях и подборе моделей прогнозирования спроса для временных рядов и методов машинного обучения. Данная структура двунаправленная, что позволяет расширять исследования в глубь и в ширь.

Анализ данных, выбор моделей и метрик будет изложено в следующем разделе и является ключевым фактором в успешном проектировании алгоритма прогнозирования сырья для предприятия.

2.3.2 Анализ набора данных для прогнозирования

Методология анализа полученного набора данных начинается с подготовки среды для исследования, загрузки набора данных, исследования общей информации и типах данных. С кодом можно ознакомиться в репозитории [47].

По результату идентификации полученного набора данных в нем 26 столбцов, 30528 строк, размер 6.1 мб+. Основные характеристики представлены на рисунке 9.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30528 entries, 0 to 30527
Data columns (total 26 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   product name            30528 non-null  object
 1   unique_id               30528 non-null  int64
 2   Сегмент рынка           30528 non-null  object
 3   Категория товара        30528 non-null  object
 4   date                    30528 non-null  datetime64[ns]
 5   weight                  30528 non-null  int64
 6   price                   30528 non-null  int64
 7   cost price              30528 non-null  float64
 8   promo action            30528 non-null  object
 9   exp                     30528 non-null  int64
 10  sex                     30528 non-null  object
 11  class                   30528 non-null  object
 12  y                       30528 non-null  int64
 13  функция                 30528 non-null  object
 14  количество компонентов  30528 non-null  int64
 15  сложность изготовления  30528 non-null  object
 16  рейтинг товаров         30528 non-null  float64
 17  количество просмотров   30528 non-null  int64
 18  конкуренция             30528 non-null  object
 19  month                   30528 non-null  int64
 20  day_of_week             30528 non-null  int64
 21  sales_3m_avg            30528 non-null  float64
 22  sales_6m_avg            30528 non-null  float64
 23  sales_12m_avg           30528 non-null  float64
 24  discount                30528 non-null  int64
 25  seasonal_discount       30528 non-null  int64
dtypes: datetime64[ns](1), float64(5), int64(11), object(9)
memory usage: 6.1+ MB
```

Рисунок 9 – Информация о собранном наборе данных для прогнозирования спроса

Приступая к анализу собранного набора данных необходимо провести ABC анализ (рисунок 10). Он позволяет оценить оборачиваемость продукта и выделить наиболее важную группу товаров для концентрации должного внимания при проектировании алгоритма прогнозирования спроса.

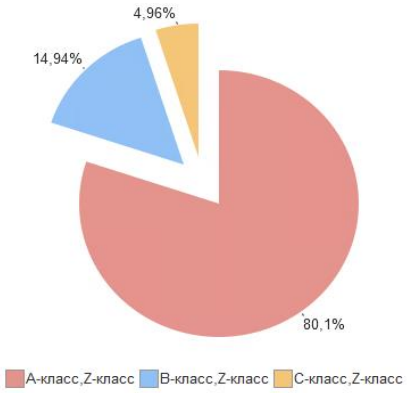

Рисунок 10 – ABC анализ ассортиментной матрицы предприятия ООО «Правильное решение»

Согласно анализу в группе А лидером является набор для тела «Груша и сандал», продукт пользуется большим спросом благодаря своему сочетанию ароматов и высокому качеству составляющих. На втором месте располагается Флюид для тела «Ши и Горький миндаль» объемом 250 мл, продукт известен своими питательными и увлажняющими свойствами, делая кожу мягкой и

гладкой. Третью позицию занимает Набор для лица «Гибискус и эстрагон», включающий тоник 200 мл, крем 50 мл и маску 50 мл, далее идет сыворотка с DMAE «Пион и Ежевика» 30 мл. Завершает пятерку лидеров Набор для кожи вокруг глаз «Сакура и кардамон», состоящий из жидких патчей 50 мл и крема 50 мл. По итогу ABC анализа подбор модели и прогнозирование будет для группы высокооплачиваемых продуктов, состоящий из 212 наименований и 4 брендов.

Итоговый набор данных для датасета состоит из разнообразных атрибутов продукта, таких как название продукта (product name), уникальный идентификатор (unique_id), сегмент рынка, категория товара, дата продажи (date), вес (weight), цена (price) и себестоимость (cost price). Кроме того, присутствуют дополнительные параметры, включая информацию о промоакциях (promo action), сроке годности (exp), рейтинге товаров (рейтинг товаров), количестве просмотров (количество просмотров) и уровне конкуренции (конкуренция). Для более глубокого анализа продаж также доступны такие атрибуты, как месяц, день недели консолидации продаж за неделю (day_of_week), средние продажи за различные периоды (sales_3m_avg, sales_6m_avg, sales_12m_avg), скидки (discount) и сезонные скидки (seasonal_discount).

Оценка качества данных показала, что в данных отсутствуют пропущенные значения во всех колонках, что упрощает дальнейшую работу с набором данных. Типы данных включают в себя подготовленную заранее временную метку, корректно распознанную как календарная дата и время (datetime64), а также количественные значения, представленные в виде целого числа (int64) и числа с плавающей точкой (float64), что подходит для статистического прогноза и машинного обучения. Некоторые атрибуты, такие как наименование продукции (product name), сегмент рынка, категория товара, конкуренция и другие представлены в текстовом формате (object), поэтому данные требуют преобразования для некоторых моделей прогнозирования.

Обзор основных статистик (Приложение Е) указывает на различное распределение значений по атрибутам, таким как цена продукции (price) (рисунок 11), себестоимость продукции (cost price) (рисунок 12), рейтинг товара (рисунок 13) и количество просмотров продукции на маркетплейсе (рисунок 14) в соотношении с количеством продаж (y). Стандартное отклонение и среднее значение показывают значительный разброс в данных, что важно учитывать при моделировании и прогнозировании.

Средняя цена товара составляет 592,2 рубля, однако наблюдается значительная вариативность, с ценами от 108 до 1999 рублей. Аналогично, вес товаров варьируется от 15 мл/гр до 5000 мл/гр, со средним значением 278,3 мл/гр. Это указывает на разнообразие реализуемых косметических товаров предприятием. Количество компонентов в товарах относительно стабильно, от 5 до 8 компонентов имеют 106 продуктов, более 8 компонентов содержат 53 продукта.

Рейтинг товаров в среднем составляет 3,5 балла, что говорит о высоком качестве продукции. Количество просмотров товаров колеблется от 4000 до 11000, со средним значением 5971,7. Это свидетельствует о различном уровне интереса к разным товарам. Стандартное отклонение даты (date) равно нулю, что подтверждает, что все продажи по наименованиям продукции относятся к одному временному периоду. Исследование взаимосвязи между рейтингом товаров и объемами продаж показало отсутствие значимой корреляции, что свидетельствует о том, что рейтинг, вероятно, не играет решающей роли в определении спроса на товар.

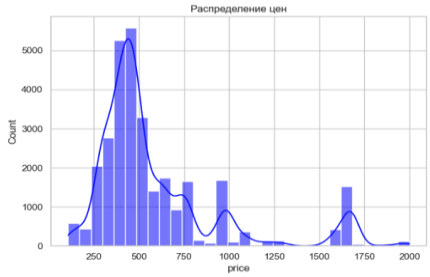

Рисунок – 11. Распределение цен

Распределение цен характеризуется положительным смещением, что означает наличие товаров с высокими ценами, но большинством товаров с низкими ценами. Это может требовать нормализации данных до их применения в машинном обучении. Средние продажи за 3, 6 и 12 месяцев указывают на наличие определенных трендов и сезонности спроса.

Рисунок – 12. Распределение себестоимости

Цены и себестоимость похожи по своей структуре и накапливают положительное смещение. Многие товары имеют низкие цены и себестоимости, но существуют также те, чьи цены и себестоимости значительно превышают средние цены.

Рисунок – 13. Распределение количества просмотров

Распределение количества просмотров продуктов демонстрирует три пика, что может указывать на два различных типа продуктов: те, что часто просматривают, и те, что просматривают реже. Это может быть связано с популярностью или доступностью товара.

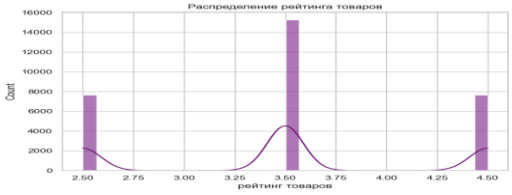

Рисунок 14 – Распределение рейтинга товаров

Распределение рейтинга товаров отражает наличие трех пиков: 2,5, 3,5 и 4,5 баллов. Это можно интерпретировать как признак наличия трех типов товаров: тех, которые часто покупают и товар удовлетворяет ожиданиям, и тех, которые в меньшей степени удовлетворяют ожидания потребителя, а также тех, которые по какой–то причине не удовлетворяют потребности клиента.

Агрегирование данных и их анализ предоставляют ценные инсайты, поэтому были рассчитаны агрегированные данные, такие как цена, себестоимость, рейтинг товаров, количество просмотров и средние продажи за 12 месяцев для каждого из 212 продуктов.

Среднее значений статистик для продукции показало, что в анализе агрегированных данных по уникальным идентификаторам (unique_id) товаров обнаружены различия в динамике цен и продаж. Рассматривая средние показатели цен на товары (price_mean), наблюдается, что большинство товаров имеют среднее ценовое значения в узком диапазоне, но есть некоторые товары с очень высокой средней ценой, видимые в выбросах (рисунок 15).

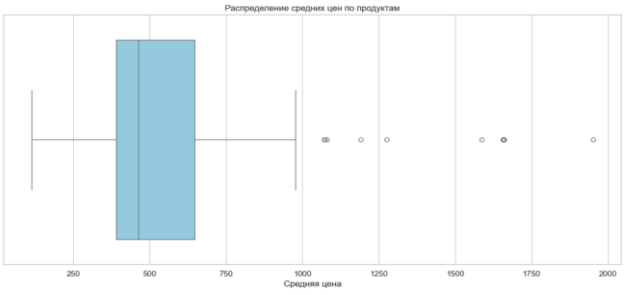

Рисунок 15 – Распределение средних цен

Себестоимость продукции (cost price_mean) также варьируется в узком диапазоне, что может свидетельствовать о стабильности производственных затрат, а также присутствуют выбросы с высокими значениями (рисунок 16).

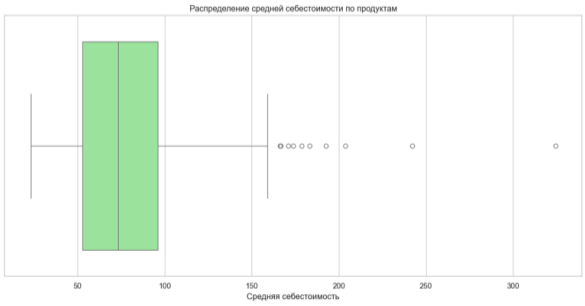

Рисунок 16 – Распределение средней себестоимости

Средние продажи большинства продуктов находятся в нижнем диапазоне, что указывает на ограниченный спрос или меньшую популярность некоторых товаров. Продукты с высоким уровнем продаж выделяются как выбросы и могут быть интересны для более детального анализа (рисунок 17).

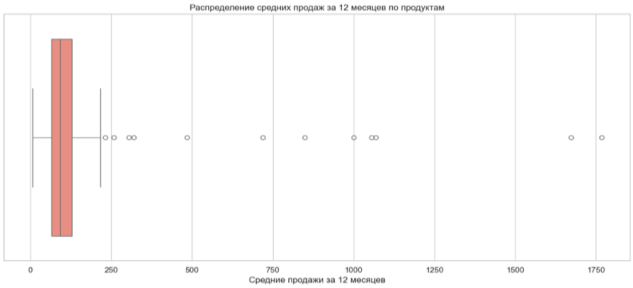

Рисунок 17 – Средние продажи за 12 месяцев

Среднемесячное количество продаж за последние 12 месяцев (sales_12m_avg_mean) демонстрирует в среднем 128 единиц, при этом максимальные значения доходят до 1768 единиц, что указывает на значительные колебания популярности товаров на протяжении года. Возможно, влиянии сезонности или маркетинговых кампаний.

Если проанализировать, сгруппировав продукцию и подсчитать количество продаж, отсортировать по убыванию, то получится результат о продажах топ–5 продукции: на первом месте скраб кофейный, на втором флюид, далее соль для ванной.

Самым распространенным является сегмент уход за кожей лица – 61 продукт, наименьшее количество уникальных значений приходится на дезодоранты – 19 продуктов, что может указывать на более узкую номенклатуру в этом сегменте.

В наборе данных присутствует 43 уникальных категории товара, что говорит о широком ассортименте продукции (рисунок 18). Шампуни занимают наибольшую долю, что делает их важной категорией для бизнеса.

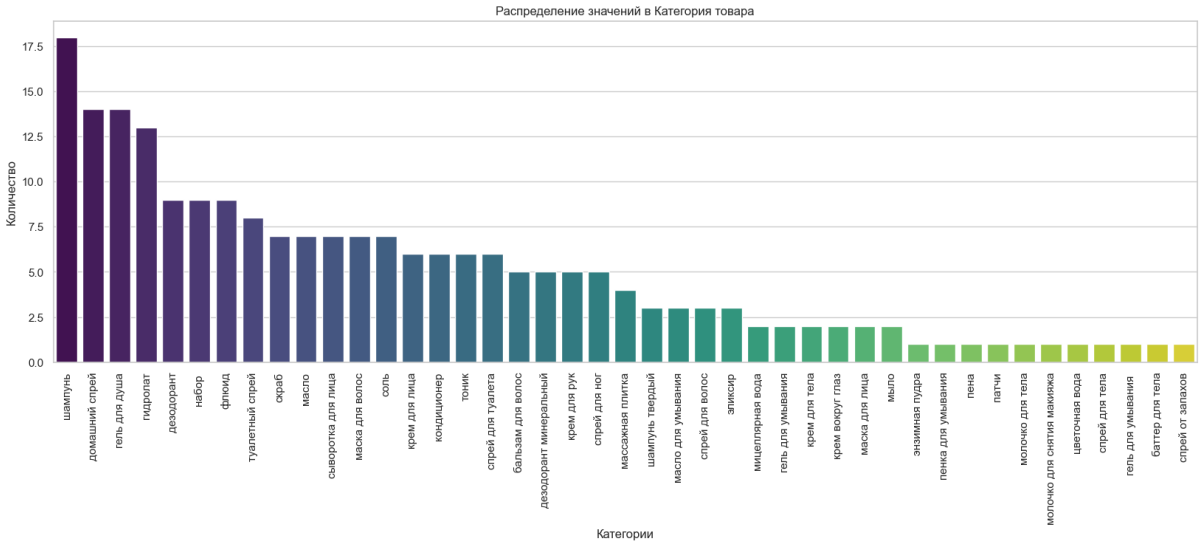

Рисунок 18 – Распределение количества продаж по категории товара

Есть только два уникальных значения в колонке промо акции: да и нет. Большее количество продукции участвовало в промоакциях (119 против 93), что может свидетельствовать о широком использовании маркетинговых кампаний для увеличения продаж.

В признаке–колонке половая принадлежность продукции (sex) есть 3 уникальных значения: унисекс, женский и мужской. Продукция, классифицированная как унисекс, встречается чаще всего (138 наименования), что может указывать на универсальность или широкий рыночный охват. Продукция для женщин (70 продуктов) встречается чаще, чем продукция для мужчин (4 единицы).

В признаке–колонке класс продукта («class») есть 2 уникальных значения: «мидл–маркет» и «масс–маркет». Большинство продукции (185 наименований) относится к категории мидл–маркет. Продукция, классифицированная как масс–маркет, встречается реже (22).

В признаке–колонке функция есть 8 уникальных значений. Самая распространенная функция продукции «увлажнение» (114 наименований).

Функции «освежение» и «очищение» также встречаются часто (38 и 26 единицы соответственно). Функции «питание» и «расслабление» встречаются реже (22 и 7 единиц соответственно) (рисунок 19).

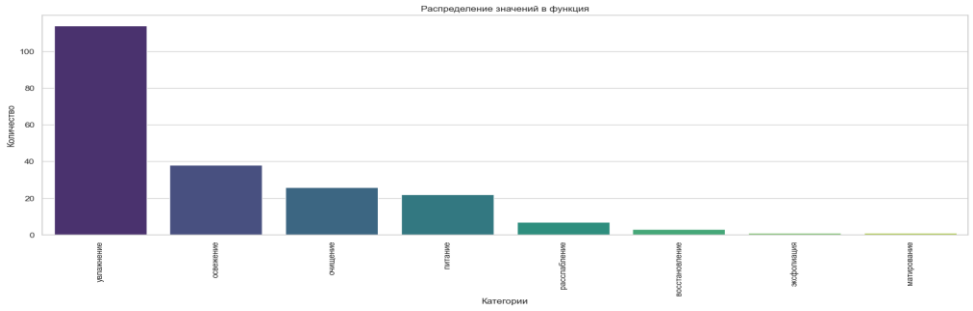

Рисунок 19 – Распределение значений по функции продукции

В колонке набора данных «сложность изготовления» есть 3 уникальных значения: «средняя», «низкая» и «высокая». Большинство продукции имеет «среднюю» сложность изготовления (106 наименований). Остальные продукты имеют либо низкую, либо высокую сложность изготовления (по 53 единицы). Это напрямую зависит от производственного процесса и от используемого количества компонентов сырья.

В колонке набора данных «конкуренция» есть также 3 уникальных значения: «средняя», «низкая» и «высокая». Большинство продукции испытывает «среднюю» конкуренцию на рынке (106). Остальные продукты испытывают либо «низкую», либо «высокую» конкуренцию (по 53 наименования). Большинство наименований продукции предприятия на рынке испытывает среднюю конкуренцию.

Скидки на продукцию предоставляют клиентам от 2 до 13 процентов, но большинство продукции имеют 7 и 8 процентов скидки в разные периоды времени (87 и 77 продуктов соответственно).

Для более точного прогнозирования спроса необходимо провести корреляционный анализ, чтобы выявить взаимосвязи между ценой, рейтингом, количеством просмотров, скидками. Также важно учесть внешние факторы, такие как конкурентная среда, и провести анализ по категориям товаров (рисунок 20).

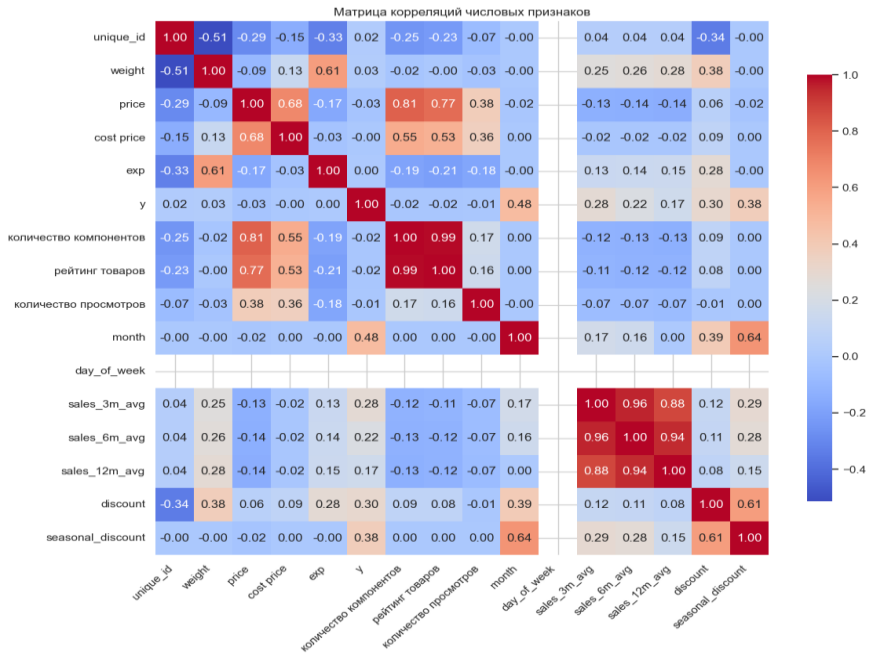

Рисунок 20 – Матрица корреляций количественных признаков

Коррелограмма позволяет проанализировать связь между различными количественными признаками и выявить ключевые взаимосвязи. Из матрицы видно, что значительные корреляции существуют между количеством компонентов продукта, его ценой, себестоимостью и рейтингом, что указывает на важность этих характеристик в прогнозировании продаж. Взаимосвязь между ценой (price) и рейтингом товаров (рейтинг товаров) демонстрирует корреляцию на уровне 0,77, что предполагает, что более высокие цены могут ассоциироваться с более высоко оцениваемыми товарами, либо же что товары с высоким рейтингом обладают более высокой ценой из–за своей воспринимаемой ценности или качества. Связь между количеством компонентов и ценой (корреляция 0,81), а также между количеством компонентов и рейтингом товаров (0,99), подсказывает, что более сложные в изготовлении товары, возможно, оцениваются выше и стоят дороже. Это может быть признаком того, что потребители готовы платить больше за товары, которые воспринимаются как более сложные в производстве. Интересное наблюдение касается количества просмотров, которое также демонстрирует среднюю корреляцию (0,64) с месяцем, что может указывать на сезонные колебания интереса к определенным товарам. Это подчеркивает

важность учета сезонных факторов при прогнозировании продаж. Анализ автокорреляции продаж (y) по различным средним показателям (3–месячным, 6–месячным и 12–месячным) подтверждает наличие сильной связи между текущими и прошлыми значениями продаж.

В исследовании мы будем рассматривать всю продукцию и топ–5 продуктов по количеству продаж, которые генерируют основную прибыль предприятию. Первичный визуальный анализ для топ–5 по количеству продаж временных рядов продемонстрировал различную динамику продаж и объем, тренд и сезонность, каждый ряд уникальный (рисунок 21).

Рисунок 21 – Временные ряды продаж для 5 уникальных товаров

Вычисление и анализ p–значений, критических значений и проведение статистического теста Дики–Фуллера позволило выявить, что 134 ряда в наборе данных нестационарны, а проведенная декомпозиция для каждого ряда показала, что ряды содержат трендовые и сезонные компоненты [47]. Это указывает на то, что для моделирования и анализа этих рядов потребуется дополнительное дифференцирование или преобразование данных для достижения стационарности перед применением моделей ARIMA или других.

Далее произведено удаление тренда и сезонности с помощью сезонного декомпозирования и произведено преобразование Бокса–Кокса, которое позволяет оптимизировать распределение данных, стабилизируя дисперсию в зависимости от значения лямбды (λ). Найдены оптимальные значения параметра лямбда для каждого ряда, значение колеблется в диапазоне 1,06 – 1,30 и преобразование привело к нормальному распределению данных.

На рисунке 22, 23 приведена декомпозиция ряда с уникальным идентификатором 16, а также визуализация автокорреляционной и частичной автокорреляционной функции (ACF, PACF). Данный временной ряд представлен выраженной повторяющейся сезонностью и постепенно восходящим трендом, остатки имеют сравнительно низкую вариабельность, предполагающую, что модель хорошо улавливает тренд и сезонность.

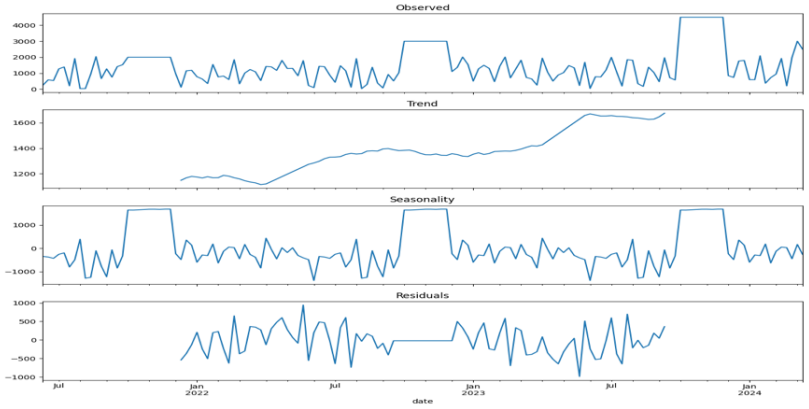

Рисунок 22 – Визуализация декомпозиции ряда для id 16 из топ–5 по количеству продаж

Анализ ACF (рисунок 23) для временного ряда показывает затухание после первого лага, но остаются значительные автокорреляции на уровнях, соответствующих сезонным интервалам, что подтверждает сезонную природу данных. На графике PACF заметен сильный пик на первом лаге, что характерно для данных, хорошо моделируемых авторегрессионным процессом первого порядка (AR(1)). Последующие лаги находятся в пределах статистической незначимости, что предполагает отсутствие дополнительных AR–эффектов.

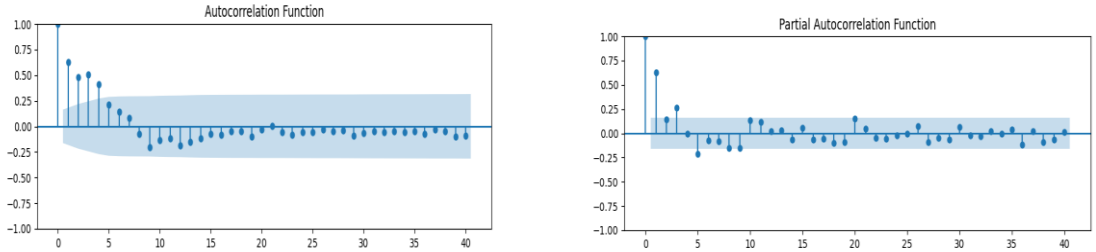

Рисунок 23 – Визуализация автокорреляции и частичной автокорреляции ряда для id 16 из топ–5 по количеству продаж

Наличие сезонности в данных указывает на возможность изменений в потребительских предпочтениях перед новым годом, а наличие тренда как результат эффективности маркетинговых стратегий на протяжении времени. Сезонность может быть связана с периодически повышающимся спросом в зимний период, октябрь–ноябрь, что типично для косметической индустрии. Для проверки автокорреляции на первом порядке в наборе временных рядов также применен тест Дарбина – Уотсана, используя среднее коэффициентное смещение, по результатам которого выявлено, что данные в рядах имеют автокорреляцию, что может стать проблемой для временных рядов, так как повлияет на качество прогнозирования.

Остатки, или случайные колебания в данных подчеркивают наличие факторов, не объясненных текущей моделью декомпозиции. Эти остатки могут включать в себя внешние события или изменения в поведении потребителей, которые не были учтены трендом и сезонностью.

Исходя из проведенного анализа набора данных можно предложить модели и метрики, которые наиболее эффективно могут позволить предсказывать продажи на предприятии, учитывая, что проведенная попытка кластеризации продукции методом локтя позволила разделить ассортиментную матрицу на 4 кластера с распределением 178, 2, 27 и 5 продуктов, что не является информативным, поэтому важно учитывать особенности каждого ряда для базисных статистических моделей. С кодом анализа можно ознакомиться в репозитории [47].

2.3.3 Описание моделей и метрик для прогнозирования

Рассмотрим модели, которые будут использованы в исследовании, следуя общим этапам для моделирования (рисунок 24).

Рисунок 24 – Этапы исследования моделей

При таком количестве недельных наблюдений (144) для каждого продукта и большом количестве продуктов (212), может возникнуть проблема недостаточности данных для построения надежных моделей.

Временной ряд можно записать с помощью формулы:

$$Xt = x1, x2, \cdots, xn, \quad (1)$$

где x – значение наблюдения;

t – индекс времени.

Он состоит из сочетания составляющих: тренд, сезонность, цикл, случайные колебания.

Из традиционных статистических моделей для анализа временных рядов выбраны:

ARIMA (Autoregressive Integrated Moving Average) является моделью прогнозирования временных рядов, которая сочетает три основных компонента: авторегрессию (AR), интеграцию (I) и скользящее среднее (MA). Она предназначена для анализа одномерных временных рядов, которые являются стационарными или которые можно привести к стационарному состоянию путем дифференцирования. AR(p) – часть авторегрессии предполагает, что текущее значение ряда можно объяснить через его предыдущие значения. I(d) – раздел интеграции означает, что данные могут потребовать дифференцирования d раз, чтобы стать стационарными. MA(q) – компонент скользящего среднего предполагает, что текущее значение ряда можно объяснить через прошлые ошибки прогноза.

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

где Y_t – оператор лага;

c – константа;

ϕ_i – коэффициенты авторегрессии;

θ_i – коэффициенты скользящего среднего;

d – порядок дифференцирования;

ε_t – ошибка прогноза в момент времени t.

SARIMA (Seasonal ARIMA) является расширением ARIMA модели для учета сезонности в данных. Так как временные ряды имеют сезонные паттерны, то SARIMA добавляет сезонные элементы в ARIMA для моделирования временных рядов.

SARIMA обозначается как ARIMA(p, d, q)(P, D, Q)[S], где P, D, Q – сезонные аналоги параметров p, d и q, а S – длина сезонного периода.

$$\Phi(B^m)\phi(B)(1-B)^d(1-B^m)^D y_t = \Theta(B^m)\theta(B)\varepsilon_t \qquad (3)$$

где y_t – временной ряд;

$\Phi(B^m)$ – полином сезонной авторегрессии (SAR) порядка P;

$\phi(B)$– полином авторегрессии (AR) порядка p;

B – оператор сдвига назад (лаг оператор), такой, что $B^k y_t = y_{\{t-k\}}$;

$\Theta(B^m)$ – полином сезонной скользящей средней (SMA) порядка Q;

$\theta(B)$ – полином скользящей средней (MA) порядка q;

d – порядок несезонного дифференцирования;

D – порядок сезонного дифференцирования;

ε_t – белый шум (случайная ошибка).

Модель Хольта–Винтерса является методом экспоненциального сглаживания для анализа временных рядов с сезонными компонентами. Она расширяет модель экспоненциального сглаживания Хольта, добавляя компонент для сезонности. Модель включает три уравнения: уровень, тренд и сезонность. У каждого компонента свой коэффициент сглаживания: α (уровень), β* (тренд) и γ (сезонность). Модель может быть адаптирована для

сложения (аддитивной) или умножения (мультипликативной) сезонности в зависимости от характера данных.

Формула (4) для сглаженного значения тренда представлена ниже:

$$l_t = \alpha(y_t - s_{t-S}) + (1-\alpha)(l_{t-1} + b_{t-1}) \quad (4)$$

где y_t – фактическое значение в момент времени (t);

s_{t-S} – сезонный компонент в момент времени (t–S) (где (S) – период сезонности);

l_{t-1} – сглаженное значение тренда в предыдущий момент времени;

b_{t-1} – скорость изменения тренда в предыдущий момент времени;

α– параметр сглаживания для тренда.

Формула (5) для сглаженного значения сезонного компонента представлена ниже:

$$s_t = \gamma(y_t - l_{t-1}) + (1-\gamma)s_{t-S} \quad (5)$$

где s_{t-S} – сезонный компонент в момент времени;

γ – параметр сглаживания для сезонного компонента, $0 \leq \gamma \leq 1$.

Формула (6) для прогноза (m) периодов вперед $\widehat{y_{t+m}}$:

$$\widehat{y_{t+m}} = l_t + m\,b_t + s_{t+m-S} + (m-1)b_S \quad (6)$$

где l_t – сглаженное значение тренда в момент времени (t);

b_t – скорость изменения тренда в момент времени (t);

s_{t+m-S} – сезонный компонент для момента времени (t+m–S);

b_S – среднее изменение тренда за период сезонности (S).

Из методов машинного обучения будут исследованы:

Линейная регрессия один из самых основных и широко используемых типов прогнозных моделей из–за ее простоты и эффективности в определенных случаях. Она строится на предположении о линейной зависимости между одной или несколькими независимыми переменными (предикторами) и зависимой переменной (объемом продаж). В случае одной независимой переменной она называется простой линейной регрессией, а при нескольких – множественной линейной регрессией.

$$y = \beta 0 + \beta 1x + \varepsilon, \quad (7)$$

где x_1 до x_n – независимые переменные;

β_1 до β_n – коэффициенты для каждой из независимых переменных;

n – количество независимых переменных.

Метод K–ближайших соседей в контексте временных рядов использует сходство между различными отрезками временных рядов для прогнозирования. Временные ряды сначала разбиваются на сегменты, после чего для каждого нового сегмента находятся K наиболее похожих сегмента из обучающего набора данных. Прогноз осуществляется путем усреднения или взвешивания значений этих K соседей.

Существуют несколько популярных мер расчета расстояния между объектами, такие как манхэттенское, евклидово, косинусное и расстояние Минковского. Например, для Евклидова расстояния между двумя временными рядами X и Y, каждый из которых имеет длину n, используется формула (8).

$$d(X,Y) = \sqrt{\sum_{\{t=1\}}^{\{n\}} (x_t - y_t)^2} \quad (8)$$

где $d(X, Y)$ – Евклидово расстояние между векторами X и Y;

Σ – символ суммирования;

n – количество элементов в векторах X и Y;

x_t – элемент вектора X в позиции t;

y_t – элемент вектора Y в позиции t;

Основная формула (9) для K соседей представлена ниже.

$(x_t - y_t)^2$ – квадрат разности между соответствующими элементами векторов X и Y.

$$y^{=} \left(\frac{1}{K}\right) \sum y_i \quad (9)$$

где y – прогнозируемый ответ;

K – число ближайших соседей;

y_i –ответы ближайших соседей.

Деревья принятия решений могут быть адаптированы для прогнозирования временных рядов, где каждый узел в дереве представляет решение пошагово и нелинейно, основанное на значении или изменении значений во времени. Это может включать различные статистики, такие как среднее значение или стандартное отклонение определенного окна временного ряда. Формула (10) используется как критерий для измерения качества разбиения узла в дереве принятия решений, существуют и другие критерии, такие как информационная энтропия и среднеквадратичная ошибка.

$$I_G(p) = 1 - \Sigma(p_i)^2 \quad (10)$$

где $I_G(p)$ –индекс Джини для измерения частоты разбиения;

p_i – доля объектов класса i в узле.

Градиентный бустинг над решающими деревьями создает последовательность моделей, где каждая следующая модель корректирует ошибки предыдущей. Это усиливает слабые обучающие алгоритмы, делая их более мощными.

AdaBoost (Adaptive Boosting) последовательно добавляет учащихся, выбирая те, которые лучше всего исправляют ошибки предыдущих моделей.

Gradient Boosting использует градиентный спуск для минимизации ошибок, делая это более гибким и мощным. Формула (11) представляет основную идею алгоритма.

$$y(x)^{=\Sigma}\left(\alpha_b * h_b(x)\right) \quad (11)$$

где $y(x)$ – прогнозируемый отклик для входных данных;

α_b – вес b–го слабого обучающего алгоритма;

$h_b(x)$ – прогноз от b–го слабого обучающего алгоритма.

Бэггинг работает путем создания нескольких подвыборок из исходного набора данных с помощью бутстрапа (повторной выборки с заменой) и обучения модели на каждой подвыборке. Прогнозы от каждой модели затем усредняются (для регрессии) для получения окончательного прогноза. Схематично данный метод изображен на рисунке 25.

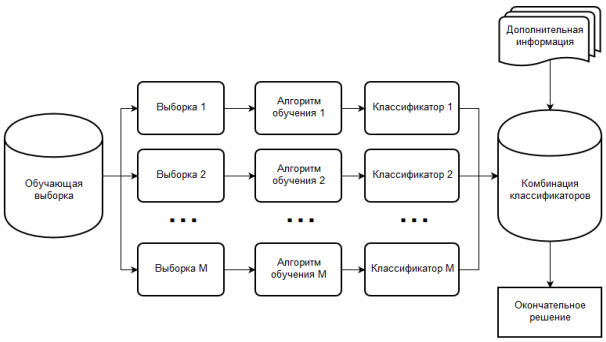

Рисунок 25 – Схема работы бэггинга

Случайный лес (Random Forest) представляет собой ансамбль деревьев решений, обученных на разных подвыборках данных с использованием бэггинга. В контексте временных рядов случайный лес может помочь уменьшить переобучение, которое часто встречается при использовании одиночных деревьев решений, и улучшить точность прогнозирования за счет ансамблевого подхода. Формула (13) отображает предсказание по случайному лесу.

$$\hat{y} = \frac{1}{N}\sum y_n \tag{12}$$

где $\widehat{y_n}$ – прогнозируемый отклик;

N – количество моделей;

y_n – прогноз n–й модели.

Нейронная сеть с архитектурой LSTM является одним из вариантов рекуррентных нейронных сетей и бывает однонаправленной и двунаправленной (BiLSTM). В Приложении Ж представлено сравнение двух архитектур для обучения.

Основным в моделировании является выбор количества слоев и нейронов в каждом слое, наличия связей между нейронами, а также тип функции активации, количество эпох. Также будет предложена архитектура TimesNet. Архитектура TimesNet основана на архитектуре Transformer и использует nn.TransformerEncoder для обработки последовательностей данных, а линейный слой (nn.Linear) используется для получения

окончательного прогноза. В данной архитектуре используется только энкодер Transformer, а декодер отсутствует.

Prophet представляется моделью из одноименной библиотеки временных рядов для одномерного прогнозирования, разработанная командой Facebook. Она является линейной модель (GAM), также она аддитивная и объединяет компоненты для анализа закономерностей в данных. Одной из особенностей является возможность модели не только производить прогноз, но генерировать новые признаки (поведение стекинг модели). Еще одной из особенностей использования Prophet может стать использование встроенных настроек, касающихся праздничных, выходных дней.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \quad (13)$$

g(t) – кусочно–линейный или логистический тренд кривой роста, который отражает долговременную тенденцию развития;

s(t) – периодические колебания (например, недельная и годовая сезонность);

h(t) – влияние праздников;

ε(t) представляет любые изменения, которые не учитываются моделью; предполагается, что ε(t) нормально распределено.

При выборе модели для прогнозирования спроса необходимо учитывать особенности полученных данных, технические характеристики моделей, доступность тренировочных данных и ожидаемые результаты.

Для оценки эффективности прогнозирования спроса можно использовать различные метрики, которые позволяют сравнивать прогнозируемые значения с реальными данными.

Для оценки качества моделей будут использованы: среднеквадратичная ошибка (MSE), средняя абсолютная ошибка (MAE), коэффициент детерминации (R^2) и MAPE.

MAPE (Mean Absolute Percentage Error) - это метрика оценки ошибки прогноза, которая измеряет среднюю абсолютную процентную ошибку между фактическими и прогнозируемыми значениями, что помогает оценить,

насколько модель точно прогнозирует относительно фактического значения. MAPE будет использована как бизнес–метрика для предоставления аналитику предприятия относительного взгляда на точность модели, насколько прогнозируемые значения соответствуют реальному показателю. Формула данной метрики представлена ниже:

$$\text{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{y_{pred,i} - y_{true,i}}{y_{true,i}} \right|}{n} \times 100\% \quad (15)$$

где y_{pred} – предсказанные значения модели;

y_{true}– реальные значения (истинные метки);

n – общее количество наблюдений и прогнозов;

i – индекс, конкретное наблюдение.

Средняя абсолютная ошибка (MAE) представляет собой среднее абсолютное значение разности между прогнозируемыми и реальными значениями. Эта метрика позволяет оценить величину ошибки прогноза в исходных единицах измерения.

Среднеквадратичная ошибка (MSE) является средним квадратом разности между прогнозируемыми и реальными значениями. Она учитывает не только величину ошибки, но и ее вариабельность.

Коэффициент детерминации (R^2) является основной статистической метрикой оценки точности и объясняемой части изменчивости в модельном прогнозе. Используется для оценки насколько хорошо модель объясняет зависимость между предсказаниями и исходными или контролирующими данными (фактическим спросом).

Будут также рассчитываться две метрики оценки производительности модели с помощью разделения данных на обучающие и тестовые наборы несколько раз и рассчитываться среднее значение R^2 (Avg, R^2 (Cross–Validation)) и среднее значение MSE (Avg, MSE (Cross–Validation)). При расчете Avg, MSE, R^2 в контексте кросс–валидации, значения этих метрик

считаются для каждого фолда (части данных) и затем усредняются по всем фолдам для получения общего значения оценки производительности модели. Это позволяет получить более надежные результаты оценки модели на основе различных частей данных.

Для повышения качества анализа и моделирования временных рядов будут применены различные методы и критерии. Информационный критерий Акаике послужит инструментом для сравнительной оценки относительного качества моделей и выбора наиболее подходящей, учитывая как точность подгонки, так и сложность модели. Преобразование Бокса–Кокса позволит стабилизировать дисперсию и приблизить распределение данных к нормальному, что положительно скажется на качестве модели и точности прогнозирования. Логарифмирование данных будет использовано для уменьшения разброса значений и линеаризации трендов, особенно в случаях экспоненциального роста или большого разброса значений. Тест Дарбина–Уотсона поможет выявить наличие автокорреляции первого порядка в остатках регрессионной модели, что может указывать на неадекватность модели и необходимость ее доработки.

В следующем разделе после анализа набора данных и временных рядов, собранных для прогнозирования спроса, подбора моделей и метрик, будут исследованы результаты обучения моделей и выбрана наилучшая подходящая модель и построен алгоритм прогнозирования сырья.

Выводы по второму разделу:

В разделе анализируется ООО «Правильное решение», включая экономические показатели и производственный процесс. На предприятии отсутствует систематизированное планирование сырья, из–за чего возникают проблемы с прогнозирование закупок сырья. Для решения этой проблемы собраны данные о продажах за 144 недели с 15 июня 2021 по 12 марта 2024, покрывающие 212 наименований продукциии выбраны методологии, модели для разработки качественной модели прогнозирования спроса.

3 ПРОЕКТИРОВАНИЕ АЛГОРИТМА ПРОГНОЗИРОВАНИЯ СЫРЬЯ

3.1 Выбор и анализ результатов обучения моделей для прогнозирования спроса

В исследовании приняли участие статистические модели ARIMA, SARIMA, метод Хольта–Винтерса, Prophet для каждого ряда, модели машинного обучения линейная регрессия, градиентный бустинг, деревья решений, к–ближайших соседей, бэггинг, случайный лес, включая модели нейронных сетей архитектуры LSTM, BiLSTM и TimesNet.

В исследовании моделей временных рядов с использованием статистических моделей будет представлен общий средний результат по метрикам для сводной таблицы, а также результат для топ–5 продукции по количеству продаж. Оценка моделей включает вычисление MSE, MAE, R^2 и MAPE.

Для модели ARIMA применено 2 подхода, которые позволяют оценить эффективность модели. В первом подходе каждый временной ряд анализируется на стационарность, и если необходимо, выполняется разностное преобразование. Затем для данных применяется декомпозиция, выделяя тренд и сезонность, и строится модель ARIMA, параметры которой подбираются с минимизацией AIC. Информация о качестве и эффективности прогнозных моделей для топ–5 продукции представлен в таблице 1.

Таблица 1 – Результат метрик модели ARIMA для топ–5 продукции

ID	AIC	Order	MSE	MAE	R^2	SMAPE	MAPE	Trend	Seasonality
16	2328,265	(2, 2, 1)	715453,5	606,16	0,40	0,56	2,53	TRUE	TRUE
17	2323,089	(0, 2, 2)	719890,1	592,90	0,38	0,56	2,51	TRUE	TRUE
352	2105,516	(0, 1, 1)	154394,5	105,07	0,03	1,78	0	TRUE	TRUE
373	2169,592	(0, 2, 2)	255627,7	164,25	0,08	1,78	0	TRUE	TRUE
376	2110,042	(0, 2, 2)	172176,7	143,81	0,11	1,77	0	TRUE	TRUE

MSE и MAE у ID 16, 17 высокие, что указывает на наличие ошибок в предсказании модели. Низкий R^2 указывает на то, что модели объясняют часть вариации временных рядов, но не идеально. Далее было использовано

логарифмирование для набора данных в статистических моделях, чтобы стабилизировать дисперсию, так как данные имеют ненормальное распределение и колебания в данных. Особое внимание при выборе модели будут обращено на коэффициент детерминации как основной метрики.

Значения AIC (Akaike Information Criterion) в моделях без логарифмирования варьируются от 2071,53 до 2528,26, что указывает на относительно небольшие различия в качестве моделей. Модель с наименьшим AIC (2071.53) является предпочтительной. Модель требует доработки и применения других подходов.

Второй подход аналогичен первому, но включает преобразование Бокса–Кокса для стабилизации дисперсии временных рядов, так как данные имеют изменяющуюся амплитуду. Информация о качестве и эффективности прогнозной модели для топ–5 продукции представлен в таблице 2.

Таблица 2 – Результат метрик модели ARIMA для топ–5 продукции с преобразованием Бокса–Кокса

ID	AIC	Order	MSE	MAE	R^2	SMAPE	MAPE	Trend	Seasonality
16	1101,68	(2, 1, 1)	790651,44	623,32	0,342	58,23	152,70	TRUE	TRUE
17	1144,19	(2, 1, 1)	780845,44	685,59	0,332	57,67	178,49	TRUE	TRUE
352	1265,27	(0, 2, 1)	156618,52	120,35	0,923	8,89	14,01	TRUE	TRUE
373	1044,42	(0, 1, 0)	236592,33	111,69	0,838	8,56	13,17	TRUE	TRUE
376	1255,41	(0, 2, 1)	161543,86	123,70	0,916	9,09	14,22	TRUE	TRUE

Некоторые модели, например, используемые для продуктов с ID 373 и 352 демонстрируют низкие показатели ошибок и высокий коэффициент объяснения вариабельности данных. Модели для продукции с ID 16, 17 демонстрируют высокие значения ошибок и низкие показатели R^2. Это указывает на необходимость пересмотра их параметров или перенастройки.

Среднеквадратичная ошибка (MSE) варьируется от 11 000 до 858 000, что говорит о значительном различии в точности предсказаний между разными моделями. Точно также и абсолютная ошибка (MAE) колеблется от 51 до 166 000, указывая на разную степень точности различных моделей в прогнозировании данных. Коэффициент детерминации (R^2), который оценивает долю вариации в зависимой переменной, объясненную независимыми переменными, находится в пределах от 0,130 до 0,925,

показывая значительные различия в объясняющей способности моделей. Некоторые значения MAPE достигают 383%, что может указывать на значительные проблемы в некоторых моделях при работе с данными, склонными к ошибкам или аномалиям.

Модель ARIMA прогнозируют спрос плохо, но неплохо учитывает сложность данных и дополнительно требуется различная обработка для каждого временного ряда, модель работает долго и это указывает на невозможность ее использования в будущей системе без построения системы параллельных вычислений.

В исследовании модели SARIMA первым подходом были данные без преобразования Бокса-Кокса. Результат обучения моделей первых топ пять временных рядов представлен в таблице 3.

Таблица 3 – Результат метрик модели SARIMA для топ–5 продукции

id	aic	order	seasonal order	mse	mae	R^2	MAPE	training_ time
16	1462	(0, 1, 1)	(1, 1, 1)	610894	607,6	0,49	1,4	21,8
17	1456	(1, 1, 1)	(1, 1, 1)	581489	626,4	0,50	1,8	22,0
352	1275	(0, 1, 1)	(0, 1, 1)	262416	190,7	0,75	1,3	23,6
373	1314	(0, 1, 1)	(1, 1, 1)	232231	204,3	0,02	1,7	24,3
376	1279	(0, 1, 1)	(0, 1, 1)	257747	189,3	0,66	1,5	24,4

Метрики у этого подхода разные для временных рядов, так AIC варьируется от 942,92 до 1462,66, MSE от 2749 до 610 894, MAE от 22,93 до 626,49 и показывают значительный разброс. R^2 от 0.01 до 0.85, что тоже свидетельствует о возможном наличии шума и аномалий в данных. Проведя анализ полученных результатов, выявлена лучшая модель с ID 9 относительно низкими значениями MSE (2749,63) и MAE (22,93), а также сравнительно низким значением AIC (942,92), R^2 этой модели составляет 0,69, что указывает на удовлетворительное объяснение изменчивости ответов с помощью модели.

Далее был применен подход с преобразованием Бокса-Кокса данных, исходя из результатов низкое значение R^2 (0,020 – 0,956), MAPE с широкой вариативностью, а также по-прежнему высокий MSE от 2432 до 7 млн. позволяет оценить объясняющую способность моделей, процент ошибок прогнозирования как неудовлетворительную.

Лучшая модель с номером 11 показывает очень низкий MSE 2434,60 и высокий R^2 0,85. Модель с номером 253 является худшей, так как имеет высокие значения MSE 1 645 129 и низкий R^2 0,02.

Анализируя результаты, становится очевидным, что возможно присутствуют аномалии или высокий уровень шума в данных, не улавливаемых моделью. В то же время наблюдается низкий коэффициент детерминации R^2, что сигнализирует о недостаточности модели для объяснения вариативности наблюдаемых значений. Главным недостатком SARIMA является время обучения всех 212 моделей для прогнозирования спроса, которое составляет более 90 минут.

При анализе модели тройного экспоненциального сглаживания (методика Хольта–Винтерса) для 212 временных рядов также использовали для сравнения качества и точности модели с логарифмированием и без него. Результаты показывают, что R^2 в обоих случаях колеблется около 90%, что показывает высокую способность моделей объяснять вариативность наблюдаемых данных.

MSE без логарифмирования данных колеблется от 10 000 до 89 000, что свидетельствует о низкой точности предсказательной способности моделей. MAPE имеет большую вариативность от 0,031 до 132,19, то есть модель в различной степени переобучается или не до обучается Результат работы моделей для топ–5 продукции в таблице 4.

Таблица 4 – Результат работы метода Хольта–Винтерса для топ–5 продукции с логарифмированием

ID	Smoothing level	Smoothing seasonal	MSE	MAE	R^2	MAPE	Trend	Seasonality
16	0,427	0,0327	32272	1574,3	0,8	84,24	TRUE	TRUE
17	0,494	0,0001	19557	1181,7	0,9	74,13	TRUE	TRUE
352	0,95	0,0027	61929	2083,9	0,9	83,72	TRUE	TRUE
373	0,935	0,0050	92722	2803,7	0,8	108,41	TRUE	TRUE
376	0,971	0,001	65987	2202,3	0,7	86,99	TRUE	TRUE

Полученные результаты демонстрируют неудовлетворительные метрики с логарифмированием данных, по-прежнему сохраняется высокое MSE для моделей.

Далее было проведено обучение с помощью Prophet и смоделировано три различные конфигурации: модели с автоматическим учетом сезонности, с учетом недельной сезонности и с подбором гиперпараметров и кросс валидацией. Наибольший интерес представляет конфигурация с кросс валидацией и подбором гиперпараметров, результаты для топ–5 продуктов представлены в таблице 5.

Таблица 5 – Результат метрик модели Prophet

ID	MSE	MAE	R^2	MAPE
16	4140	53,5	0,65	1,0
17	3789	48,6	0,67	1,5
352	2546	44,6	0,87	2,4
373	1706	34,0	0,88	1,4
376	2330	42,4	0,87	2,2

Среднеквадратичная ошибка (MSE) показывает вариацию от 1700 до 7200 между различными идентификаторами уникальных объектов, что свидетельствует о разнообразии точности прогнозов. Средняя абсолютная ошибка (MAE) варьируется в диапазоне от 31,7 до 68,67, что указывает на колебания точности отдельных моделей в разных сценариях использования. Коэффициент детерминации (R^2) колеблется от 0,650 до 0,890, что отражает разную степень соответствия моделированных данных наблюдаемым значениям. Сравнение усредненных метрик для исследованных статистических моделей представлены в таблице 6.

Таблица 6 – Результат метрик по статистическим моделям

Модель	R^2 (Test)	MSE (Test)	MAE (Test)	MAPE (Test)
ARIMA	0,515	161543	123	14,22
SARIMA	0,695	257747	189	5,47
Holt–Winters	0,963	55352	71,85	156,91
Prophet	0,752	6338	60,86	6,86

Сравнение показало значительные различия в их точности и способности объяснять вариацию данных. Модель ARIMA продемонстрировала коэффициент детерминации R^2 на уровне 0,515, что указывает на умеренное качество предсказания, при этом среднеквадратическая ошибка (MSE) составила 161 543. Средняя абсолютная

ошибка (MAE) достигла 123, а средний абсолютный процент ошибки (MAPE) был 14,22%, что указывает на значительные отклонения в предсказаниях.

Модель SARIMA улучшила объясняющую способность с R^2 до 0.695, однако MSE увеличилась до 257747, что указывает на значительные отклонения в предсказаниях по сравнению с фактическими значениями. MAE возросла до 189, несмотря на то, что MAPE снизилась до 5.47%, демонстрируя улучшение относительной точности.

Модель Holt-Winters показала высокий R^2 на уровне 0.963, что свидетельствует о высоком качестве предсказания, и значительно уменьшенное значение MSE до 55352. MAE снизилась до 71,85, однако MAPE оказалась чрезвычайно высокой – 156,91%, указывая на проблемы модели с относительными ошибками при прогнозировании данных с малыми значениями.

Модель Prophet продемонстрировала наиболее сбалансированные результаты. Коэффициент детерминации R^2 составил 0,752, что выше, чем у ARIMA и SARIMA. MSE была значительно ниже и составила 6338, указывая на высокую точность модели. MAE достигла 60,86, а MAPE составила 6,86%, что свидетельствует о высокой абсолютной и относительной точности модели.

Модель Prophet оказалась наиболее точной и устойчивой среди рассмотренных, благодаря низким значениям MSE, MAE и MAPE, а также высокому R^2. Модель Holt-Winters также продемонстрировала высокую точность, но её высокая относительная ошибка (MAPE) ограничивает её применение. Модели SARIMA и ARIMA показали умеренные результаты и требуют дальнейшей настройки для повышения точности прогнозов.

Далее рассмотрим результаты построения моделей с помощью методов машинного обучения (Таблица 7). Для сравнения результатов обучения статистических моделей с иными методами и моделями все метрики и результаты усреднены по 212 временным рядам и представлены в логарифмическом представлении. С реализацией моделей можно ознакомиться в репозитории [47].

Таблица 7 – Результат метрик по моделям машинного обучения

Модель	R^2 **(Test)**	**MSE (Test)**	**MAE (Test)**	**MAPE (Test)**	**Avg, MSE (Cross–Validation)**	**Avg,** R^2 **(Cross–Validation)**
Градиентный бустинг (XGBoost)	0,769	1859,26	20,69	0,47	1983,54	0,775
Деревья решений	0,683	2553,83	23,13	0,35	3046,00	0,710
Бэггинг	0,699	2427,46	21,93	0,36	3476,04	0,715
Случайный лес	0,695	2456,77	22,19	0,35	3108	0,702
К–ближайших соседей	0,953	83672,70	95,70	0,17	148242,92	0,910
Линейная регрессия	0,890	4300,48	7,76	0,39	5580,07	0,790
BiLSTM	0,620	2908,53	25,68	0,59	20870,22	0,620
LSTM	0,989	17445,72	47,28	1,25	19849,93	0,989
TimesNet	0,552	3431,14	34,51	2,28	3081,19	0,639

Проанализировав полученные результаты, учитывая, что также были использованы модели нейронных сетей, можно сделать вывод, что модель градиентного бустинга (XGBoost) продемонстрировала наилучшие показатели с точки зрения коэффициента детерминации (0,769) и сравнительно низкой среднеквадратической ошибки (1859,26) на тестовых данных. Также, градиентный бустинг показал наиболее стабильные результаты на кросс-валидации с Avg. MSE равным 1983,54 и Avg. R^2 равным 0,775. Это свидетельствует о высокой способности модели к генерализации и точному прогнозированию спроса.

Деревья решений и метод бэггинга продемонстрировали сопоставимые результаты, но заметно уступают градиентному бустингу. Деревья решений показали коэффициент детерминации на тестовых данных равный 0,683 и среднеквадратическую ошибку 2553,83. Бэггинг продемонстрировал коэффициент детерминации 0,699 и среднеквадратическую ошибку 2427,46. Несмотря на их простоту и интерпретируемость, эти модели менее точны в прогнозировании по сравнению с XGBoost.

Случайный лес, еще одна популярная ансамблевая техника, также показала результаты, близкие к деревьям решений и бэггингу, но с несколько худшими показателями на кросс-валидации. Его среднеквадратическая

ошибка на кросс-валидации составила 3108, что значительно выше по сравнению с XGBoost.

Метод K-ближайших соседей продемонстрировал крайне высокие значения ошибок, что свидетельствует о его несоответствии для данной задачи прогнозирования спроса. MSE составила 83672,70 и MAE равна 95,70. Высокая ошибка может быть обусловлена сложностью временных рядов и наличием большого количества продуктов, что делает KNN неприменимым для этой задачи.

Линейная регрессия, хотя и продемонстрировала высокий коэффициент детерминации (0,890), показала значительные ошибки на тестовых данных (MSE равная 4300,48). Этот метод прост и эффективен, но его линейная природа может ограничивать точность прогнозов в случае сложных временных зависимостей.

Рекуррентные нейронные сети (RNN), такие как LSTM и BiLSTM, не показали высокого уровня точности. LSTM имеет высокий коэффициент детерминации (0,989), однако значение MAPE равное 1,25 указывает на значительные ошибки в прогнозах. BiLSTM также показала слабые результаты по большинству метрик, с коэффициентом детерминации 0,620 и среднеквадратической ошибкой 2908,53. Это вероятно связано с недостаточным количеством исторических данных для моделирования и предсказания. BiLSTM является хорошим компромиссом между объяснением вариации и умеренными ошибками. Модель BiLSTM имеет умеренные результаты по метрике R^2 в кросс–валидации в сравнении с TimesNet и LSTM. TimesNet показала самые низкие показатели по коэффициенту детерминации (0,552) и высокие значения MAPE (2,28), что делает ее наименее точной моделью для данного набора данных.

Поэтому градиентный бустинг (XGBoost) является наиболее предпочтительной моделью для прогнозирования спроса в данном контексте, сочетая в себе высокую точность и стабильность на кросс-валидации. Сравнительные результаты других моделей подчеркивают важность выбора

подходящего метода машинного обучения, учитывая специфику задачи и характеристики данных.

На рисунке 26 представлены графики предсказанных значений и реальных значений, доверительные интервалы, а также график остатков базовой модели экстремального градиентного бустинга. Графики иллюстрируют результаты модели XGBoost, обученной с использованием лагов временного ряда без удаления выбросов в данных.

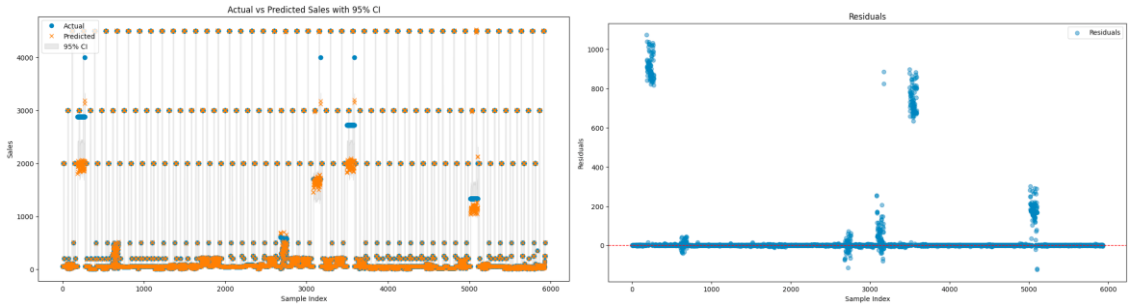

Рисунок 26 – График предсказаний и остатков для базовой модели экстремального градиентного бустинга

Лаги представляют собой значения временного ряда, сдвинутые на определенное количество временных шагов назад. Введение лагов позволяет модели захватывать автокорреляционные зависимости данных, что важно для точного прогнозирования. В данном исследовании использованы лаги от 1 до 6 для целевой переменной количества продаж (y). Это означает, что для каждого наблюдения были созданы дополнительные признаки, содержащие значения y за 1, 2, 3, 4, 5 и 6 временных шагов до текущего наблюдения.

Можно сделать вывод, что исходя из того, что доверительные интервалы узкие, модель уверена в своих предсказаниях, однако в некоторых областях графика, особенно в диапазоне высоких продаж, доверительные интервалы становятся шире, указывая на некоторую неопределенность.

На рисунке 27 представлено распределение целевой переменной – продаж.

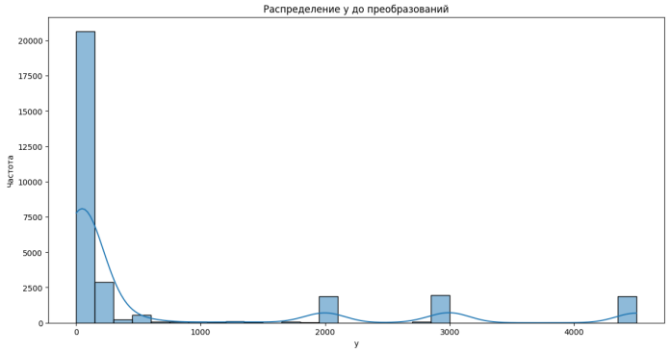

Рисунок 27 – Распределение целевой переменной y (продажи)

Распределение имеет значительное количество значений около нуля, а также присутствуют редкие выбросы на высоких значениях, данные имеют высокую степень скошенности. Поэтому будет продолжена работа по улучшению базовой модели.

ARIMA и SARIMA, хотя и не достигают таких высоких значений R^2, представляют собой интересные варианты для серий с сильной структурой и стационарностью, учитывая их исключительно высокие значения MSE и MAE на тестовых данных, что делает их не подходящими для прогнозирования в стабильных временных рядах с меньшей частотой и амплитудой изменений, но они требуют большего времени на обучения, что влияет на выбор модели для дальнейшего использования.

Модель LSTM работает лучше, чем модели K–Nearest Neighbors и Linear Regression, но не так эффективно, как модели экстремального градиентного бустинга, случайный лес и бэггинг. LSTM показывает хорошие результаты, учитывая временные зависимости, но сталкивается с высокими значениями MSE и MAE на тесте. Среднеквадратическая ошибка (MSE) модели LSTM на тестовых данных составляет 17445,72, а средняя абсолютная ошибка (MAE) — 47,28. Несмотря на высокий коэффициент детерминации (R^2 = 0,989), значительные ошибки показывают, что модель не справляется с задачей прогнозирования так хорошо, как это делает, например, градиентный бустинг с MSE равным 1859,26 и MAE равным 20,69. Главным преимуществом модели LSTM является способность учитывать сложные временные зависимости в данных, что важно для долгосрочных прогнозов.

После ряда экспериментов была подобрана улучшенная модель экстремального градиентного бустинга. Модель настроена с использованием следующих гиперпараметров: количество деревьев 173, максимальная глубина деревьев 7, скорость обучения 0,196, доля выборки для обучения 0,74, доля признаков для обучения каждого дерева 0,52, параметр регуляризации гамма 0,00012, минимальный вес ребенка 1, регуляризационный параметр альфа 0,303, регуляризационный параметр лямбда 0,665. При оценке качества модели на тестовой выборке средняя абсолютная ошибка составила 1,12, среднеквадратическая ошибка 2,23, среднеквадратическое отклонение 1,49, средняя абсолютная процентная ошибка 0,091, коэффициент детерминации 0,98. Доверительные интервалы для метрик бутстрэппинга показали, что средняя абсолютная ошибка варьируется от 1,015 до 1,819, среднеквадратическая ошибка от 1,951 до 5,704, среднеквадратическое отклонение от 1,397 до 2,388, средняя абсолютная процентная ошибка от 0,080 до 0,191, а коэффициент детерминации от 0,948 до 0,982. Эти результаты демонстрируют высокую точность и надежность модели, подтверждая её эффективность для решения поставленной задачи прогнозирования.

Сравнение фактических и предсказанных продаж с 95%-ми доверительными интервалами для каждой точки представлено на рисунке 28.

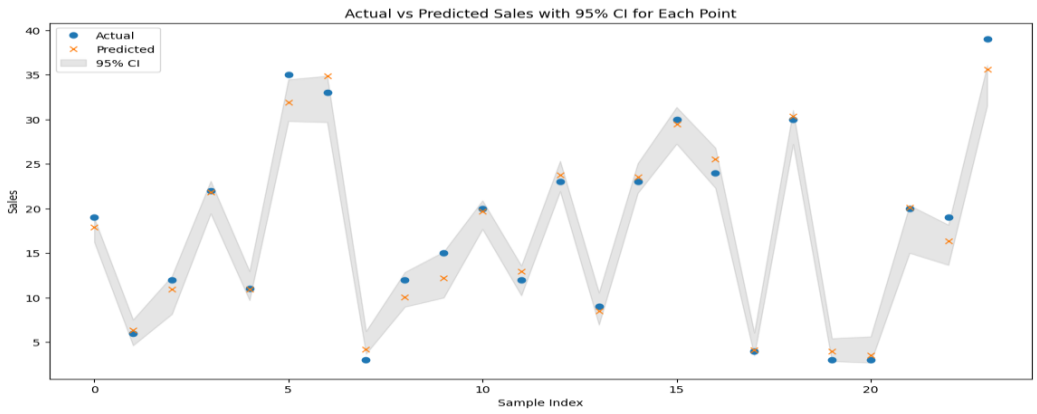

Рисунок 28 - Сравнение фактических и предсказанных продаж с 95%-ми доверительными интервалами для каждой точки

Из графика видно, что в большинстве случаев фактические значения продаж находятся внутри 95%-х доверительных интервалов предсказанных

значений, что свидетельствует о надежности предсказаний модели в рамках указанного уровня доверия.

Различные модели машинного обучения и статистические методы имеют разную степень адаптации к набору данных для прогнозирования спроса и результаты моделирования свидетельствуют о высокой прогнозной мощности алгоритма экстремального градиентного бустинга согласно результирующим метрикам. Далее будет предложен алгоритм на основе модели градиентного бустинга (с помощью библиотеки Xgboost), потому что данная модель быстро обучается и имеет высокую точность.

3.2 Проектирование алгоритма прогнозирования сырья

Инструментарий для проектирования алгоритма прогнозирования сырья включает:

– язык программирования Python, так как это динамичный язык общего назначения, который используется для анализа данных, выдвижения гипотез, проведения и подбора модели, а также для последующего расчет сырья;

– microsoft Excel для первичного хранения и организации собранных, выгруженных данных;

– библиотеки для обработки больших данных, машинного обучения, работы с временем и статистических расчетов и моделирования такими как pandas, NumPy, statsmodels, matplotlib, OpenPyXL, datetime;

– интегрированная среда разработки VScode для эффективного кодирования и отладки;

– 1С: Предприятие – корпоративная информационная система для агрегации и подготовки исходных данных.

Основные этапы проектирования алгоритма прогнозирования сырья:

– сбор требований, написание технического задания;

– сбор и предобработка отдельных наборов данных для расчета потребности сырья, такие как, остатки на складе сырья и продукции,

рецептурный каталог, необходимы для расчета сырья на единицу продукции, включая очистку от ошибок и пропущенных значений, приведение к стандартному формату и кодирование категориальных признаков, таких как наименование продукта, участие в промо акции, сегмент рынка, класс;

– проектирование алгоритма прогнозирования спроса на косметическую продукцию предприятия;

– сбор и предобработка информации об остатках сырья на складе. сбор набор данных, затем на уровне реализации на предприятии: интеграция с системой управления складскими запасами (wms) или базой данных для получения актуальных данных об остатках сырья;

– сбор и предобработка информации об остатках продукции на складе для расчета потребности в производстве недостающего количества;

– сбор и предобработка информации о рецептурах косметической продукции. получение данных о составе и рецептурах косметических продуктов из внутренних систем или баз данных и пересчет из процентного содержания в единице продукции в количественное представление;

– проектирование расчета потребности в сырье. на основе прогноза спроса, остатков сырья и продукции на складе и данных о содержании сырья в рецептуре косметической продукции на единицу производится расчет необходимого количества сырья для производства;

– документирование архитектуры алгоритма и системы для последующего развертывания и автоматизации на предприятии, чтобы создать конвейер данных для автоматизации процесса прогнозирования и обновления данных.

Разберем подробнее шаги для разработки модели прогнозирования спроса, изображенной на рисунке 29.

Рисунок 29 – Этапы прогнозирования спроса

Подготовка и написание технического задания для реализации поставленной задачи. Формулируются требования и задачи для модели, определяются ограничения, последовательность этапов разработки, инструменты разработки, требования к данным и процессу.

Сбор данных представляет собой методологическую процедуру извлечения, направленную на получение релевантной информации для формирования полноценного и согласованного набора данных для последующего использования.

Предварительная обработка данных и анализ для прогнозирования спроса. К анализу данных также важно подойти на этапе создания модели, чтобы правильно разложить временные ряды на тренд, сезонность, а также выявить закономерности и зависимости, потому что подбор модели, параметров модели, напрямую зависит от качества и количества данных, предметной области и специфики задачи.

Подбор метрик для оценки и выбора модели, а также для оптимизации модели. Для этих целей будут использованы статистические модели, машинного обучения, нейронные сети.

Документальное оформление архитектуры алгоритма и тестирование для последующего проектирования модуля расчета потребности сырья.

Ниже представлена общая блок–схема алгоритма прогнозирования сырья на основе данных о спросе моделью градиентного бустинга (XGBoost) (рисунок 30).

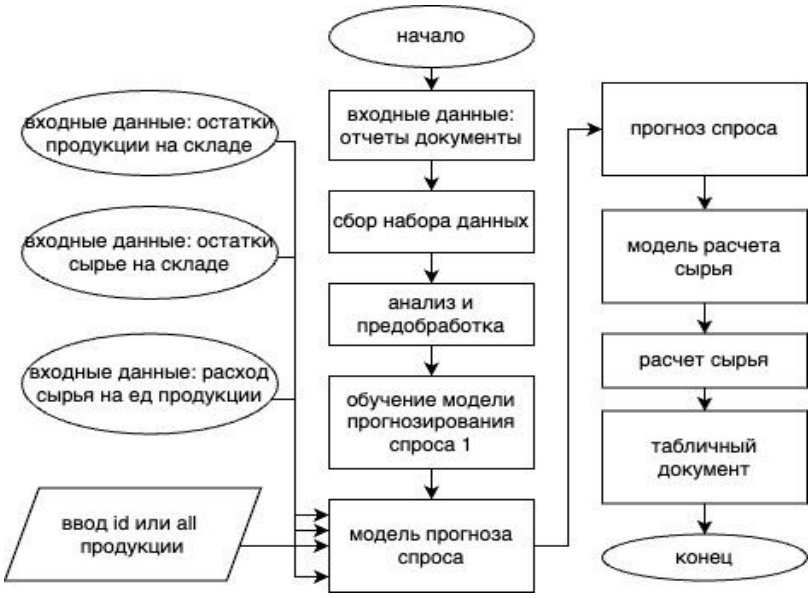

Рисунок 30 – Алгоритм прогнозирования сырья

Разберем подробнее и поэтапно предлагаемый алгоритм прогнозирования сырья (рисунок 31). Используется модель XGBoost, известная своей высокой производительностью и точностью, особенно в задачах регрессии, где необходимо предсказать количественное значение на основе множества входных признаков.

Для первой базовой модели изначально загружаются необходимые библиотеки, включая pandas для обработки данных, numpy для числовых операций, XGBoost для построения модели, библиотеки sklearn для оценки качества модели и кросс-валидации, а также matplotlib для визуализации результатов.

Затем загружаются данные из файла Excel, проверяются на наличие пропусков, которые затем заполняются, числовые значения заполняются средними значениями, а категориальные — наиболее частыми значениями (модой). Далее данные преобразуются, дата переводится в формат datetime, из неё извлекаются год, месяц и день для дальнейшего использования в модели.

Устанавливается многомерный индекс, чтобы обеспечить уникальность данных для каждого продукта и даты.

Выполняется Label Encoding для колонок promo action, sex, сложность изготовления, конкуренция. Оставшиеся необработанные категориальные переменные преобразуются в числовые значения с помощью OneHotEncoding, что позволяет модели обрабатывать их корректно.

Создаются лаговые переменные, представляющие собой значения целевой переменной y (количество продаж) за предыдущие периоды. Это помогает модели учитывать временные зависимости и улучшает точность прогнозирования.

После подготовки данных проверяются и обрабатываются бесконечные значения и пропуски, возникающие из-за добавления лагов. Далее из данных удаляются выбросы с использованием межквартильного размаха (IQR), чтобы исключить экстремальные значения, которые могут исказить результаты модели. Данные разделяются на признаки (X) и целевую переменную (y), а затем на тренировочные и тестовые наборы, чтобы обеспечить объективную оценку модели.

Для кросс-валидации используется TimeSeriesSplit, что особенно важно для временных рядов, чтобы избежать утечек данных между тренировочными и тестовыми наборами. Модель XGBoost создаётся с заранее подобранными гиперпараметрами, которые подбираются с помощью библиотеки Optuna и оценивается с использованием кросс-валидации. Вычисляются метрики MSE и R^2 для оценки качества модели.

Модель обучается на тренировочных данных и используется для прогнозирования на тестовых данных. Вычисляются метрики качества, такие как MAE, MSE, RMSE, MAPE и R^2. Дополнительно рассчитываются индивидуальные метрики для каждого уникального продукта, что позволяет детально оценить точность модели для различных продуктов.

Для прогнозирования продаж на следующую неделю используются последние доступные данные. Чтобы оценить доверительные интервалы

метрик модели, применяется бутстрэппинг. Проводится параллельное выполнение бутстрэппинга, что позволяет получить распределения метрик и оценить 95% доверительные интервалы. Также проводятся t-тесты для оценки значимости различий метрик и для MAPE и коэффициента детерминации.

Визуализация результатов включает построение графиков, сравнивающих реальные и предсказанные значения, а также графиков остатков модели. Это помогает наглядно оценить точность модели и выявить возможные области для улучшения.

Итоговые прогнозы и метрики сохраняются в файлы для дальнейшего анализа и отчетности. Этот процесс позволяет не только построить модель для прогнозирования продаж на будущую неделю, но и тщательно оценить её качество, обеспечивая возможность дальнейшего улучшения модели. Далее сохраняется лучшая модель для последующего использования.

Основная улучшенная модель после ряда исследований и с наиболее высокой точностью представляет собой несколько этапов. Первым шагом является загрузка данных из файла Excel и преобразование столбца с датой в формат datetime. Далее добавляются временные признаки, такие как год, месяц, день, день недели, квартал и другие, которые могут помочь модели лучше понять сезонные и временные зависимости в данных. Также добавляются лаговые признаки, скользящие средние и стандартные отклонения, а также экспоненциальное сглаживание для улучшения временной динамики данных.

После добавления временных признаков, данные очищаются от выбросов с использованием межквартильного размаха (IQR). Затем целевая переменная (продажи) логарифмируется для уменьшения влияния сильных выбросов и приведения данных к более нормальному распределению. Далее выполняется OneHotEncoding для категориальных признаков, что позволяет модели работать с ними.

Модель XGBoost используется для прогнозирования временных рядов. Для оптимизации гиперпараметров модели используется библиотека Optuna,

которая позволяет найти наилучшие параметры путем выполнения кросс-валидации на тренировочных данных. После нахождения наилучших параметров модель обучается на тренировочных данных и оценивается на тестовых данных.

Также используются методы бутстрэппинга для вычисления доверительных интервалов для прогнозов и метрик модели. Результаты прогнозирования отображаются на графиках, включая фактические и прогнозные значения, остатки и распределение остатков. Дополнительно строится график важности признаков для понимания вклада каждого признака в модель.

Прогнозы и метрики сохраняются в CSV-файлы для дальнейшего анализа и использования. Конечным результатом этой части алгоритма является сохранение модели.

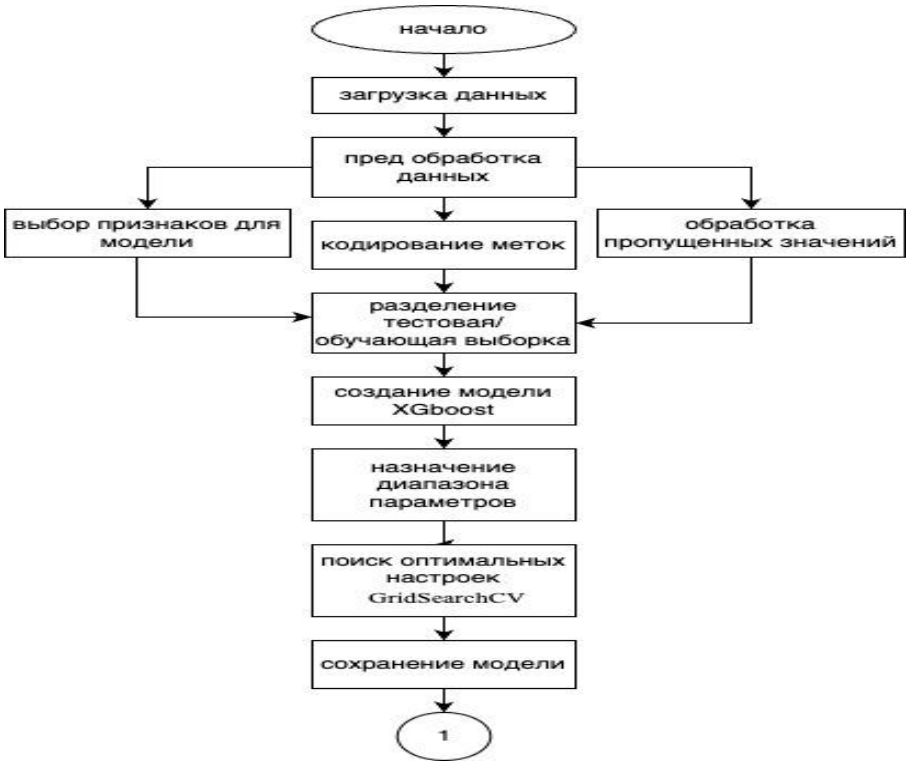

Рисунок 31 – Часть алгоритма проектирования сырья

Прогнозирование спроса на следующую неделю состоит из ввода необходимых продуктов для планирования или выбор всей ассортиментной матрицы как результат вывод данных о продажах для выбранных продуктов.

Для взаимодействия с пользователем разработана функцию для получения ввода от пользователя, позволяющую выбрать все продукты или указать конкретные продукты по их ID. Далее преобразовываются данные о продажах в формат, подходящий для модели XGBoost, и делается прогноз продаж на следующую неделю для выбранных продуктов с помощью сохраненной модели XGBoost. Дополнительно округляются спрогнозированные продажи до целых чисел и сохраняется результат прогноза в файл для последующего планирования сырья (рисунок 32).

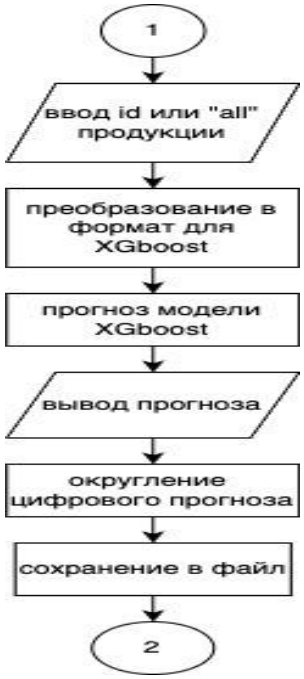

Рисунок 32 – Часть алгоритма проектирования сырья

Алгоритм планирования сырья на основе прогнозируемого спроса состоит из этапов (рисунок 33), первым шагом является объединение информации о прогнозируемом количестве продаж с текущими запасами на складе по каждому уникальному идентификатору товара. Далее производится подсчет, сколько единиц товара потенциально не хватит на складе для удовлетворения прогнозируемого спроса в следующий период. Если дефицит товара на складе меньше или равен нулю, дополнительное производство не требуется. В противном случае количество, необходимое к производству, равно величине дефицита.

Далее делается рассчет необходимого количества сырья для изготовления, учитывается выборка продукции, где потребность в производстве положительна (production_needed > 0), и дальше происходит слияние их с данными о рецептуре сырья. После получения информации о необходимом сырье для производства продуктов, данные сливаются с информацией о текущих запасах каждого сырьевого материала в raw_material_stock. Использование raw_material_ID как ключа позволяет точно определить, какие запасы доступны для каждого типа сырья. Вычисление дефицита сырья производится как разница между необходимым и имеющимся количеством сырья.

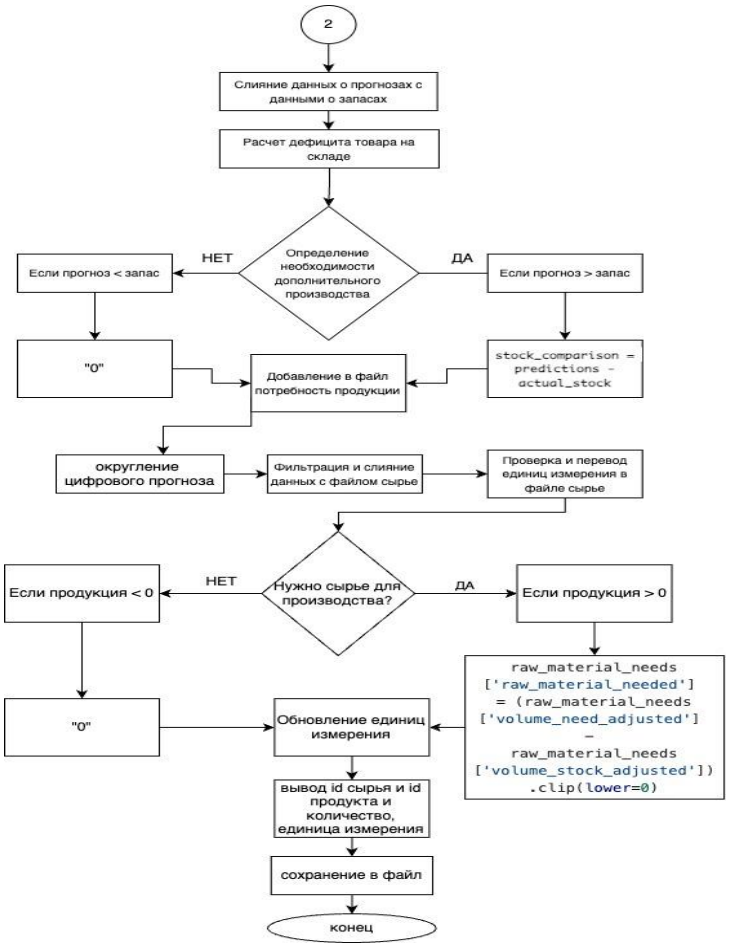

Рисунок 33 – Часть алгоритма проектирования сырья

Как итог, выводятся результаты, включающие название и уникальный идентификатор продукта, идентификатор и название сырья, а также необходимое количество сырья и единицу измерения для закупки сырья для

производственных нужд. Заканчивается алгоритм сохранением в файл информации о потребности в сырье для закупок и производства.

Для тестирования и демонстрации работы спроектированного алгоритма проектирования сырья был использован фреймворк Streamlit, фреймворк с открытым исходным кодом и созданный для команд, которые занимаются машинным обучением, для создания минимально жизнеспособного продукта в виде веб–приложения, которое позволяет продемонстрировать и протестировать работу модели.

Была написана программа для реализации веб–приложения, которое прогнозирует продажи и определения потребности в сырье на основе данных о продажах, сырье, запасах продукции и сырьевых материалах с помощью модели градиентного бустинга, модель сохраняется как finalized_model_xgb.pkl и готова к дальнейшему использованию. Приложение начинается с загрузки необходимых данных пользователем через интерфейс, где он может загрузить четыре файла в формате Excel, каждый из которых содержит необходимую информацию. Реализация продемонстрирована на рисунке 34. С кодом можно ознакомиться в репозитории.[47]

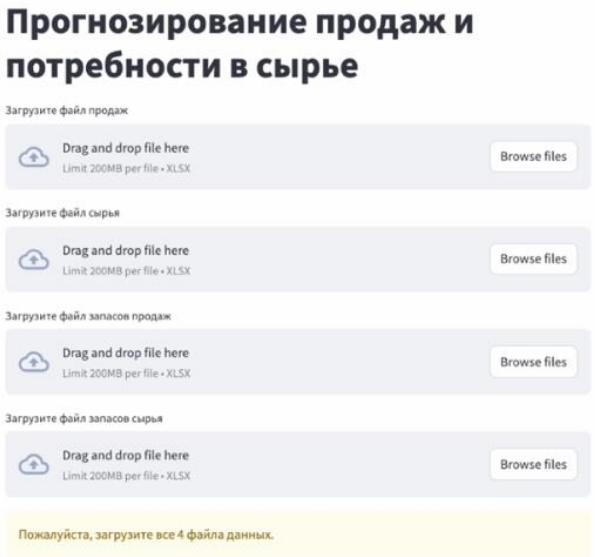

Рисунок 34 – Стартовая страница веб–приложения для прогнозирования сырья

В начале работы приложения пользователь может ввести данные через текстовое поле, где можно указать либо все продукты для анализа, напечатав «all», либо выбрать конкретные ID продуктов. Введенные данные затем используются для фильтрации необходимой информации.

Основной функционал приложения включает прогнозирование продаж на следующую неделю для выбранных продуктов и расчет необходимого количества сырья для закупки (рисунок 35).

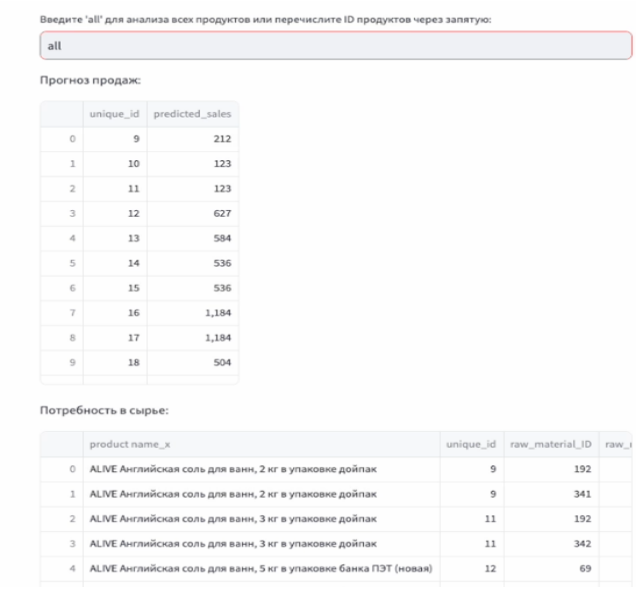

Рисунок 35 – Пример прогноза продаж и расчета сырья в приложении Streamlit

Эти прогнозы затем используются для оценки потребности в сырье, где рассчитывается, сколько материала требуется на основе предполагаемого объема продаж и текущих запасов. Программа также выполняет корректировку объемов сырья, пересчитывая их из миллиграммов в граммы при необходимости, и предоставляет подробный отчет о потребностях в сырье, включая тип и количество требуемого материала. Дополнительно есть возможность сохранения результатов прогнозирования спроса и расчета потребности сырья. Представленный алгоритм прогнозирования сырья перед внедрением в архитектуру предприятия должен пройти тщательное тестирование результатов прогнозирования и тестирование расчетной части сырья.

3.3 Рекомендации и расчет экономической эффективности по внедрению алгоритма в архитектуру предприятия

Перед внедрением важно определить в какую систему предприятия будет встроен алгоритм прогнозирования сырья. На текущий момент предприятие располагает одной базой данных MySQL и системой 1С: Предприятие, поэтому планируется, что модуль выступит как отдельная аналитическая компонента с последующей интеграцией в 1С. Часть работ заключается в проектировании конфигурации 1С, автоматизирующей бизнес–процессы предметной области. Важно также отметить, что XGBoost может быть интегрирована в рамах потока данных, таких как Apache Spark, Apache Hadoop, и Apache Flink с использованием абстрактных Rabit и XGBoost4J, что делает выбранный алгоритм более привлекательным.

На стадии проектирования алгоритма были точно определены автоматизируемые процессы. Необходимо разработать методы предварительной обработки данных, а также стратегии обучения и проверки моделей для прогнозирования спроса и последующего расчета сырья. Гарантирование согласованности и качества интегрируемых данных является приоритетом.

Архитектура предлагаемого решения для прогнозирования сырья на основании прогнозирования спроса включает интеграцию различных компонентов и использование технологий машинного обучения. Основные данные о продажах, запасах и заказах поступают из системы 1С, откуда они через модуль интеграции передаются в облачное хранилище Яндекс.Облака и базу данных на SQL Server. Использование базы 1С и модуля интеграции обеспечивает централизованный сбор данных о продажах, запасах и заказах. Использование Яндекс.Облака обеспечит отсутствие значительных инвестиций в локальную инфраструктуру. Облако также обеспечивает резервное копирование и защиту данных.

На следующем этапе автоматизированные SQL-запросы извлекают необходимые данные из хранилища для дальнейшего анализа. Центральным элементом системы является программный модуль машинного обучения, который создает несколько моделей прогнозирования спроса на основе полученных данных. Данные из базы данных будут проходить предварительную обработку в модуле по средствам введения необходимого запроса сотрудником и конвертироваться в формат .csv для последующего анализа в модуле, созданном для анализа и визуализация с помощью библиотек scikit-learn, statsmodels, pmdarima, prophet, TensorFlow и Keras, Matplotlib и Seaborn. В модуле анализа и визуализации будет осуществляться детальный анализ данных, на основе которого строятся прогнозные модели. На основании прогнозов спроса будет производиться расчет потребности в сырье. Результаты прогнозов оформляются в виде отчетов JSON, которые затем проходят этап тестирования и контроля для проверки их точности и надежности. Это необходимо перед внедрением прогнозов в операционную деятельность компании. Рекомендуемая схема изображена на рисунке 36.

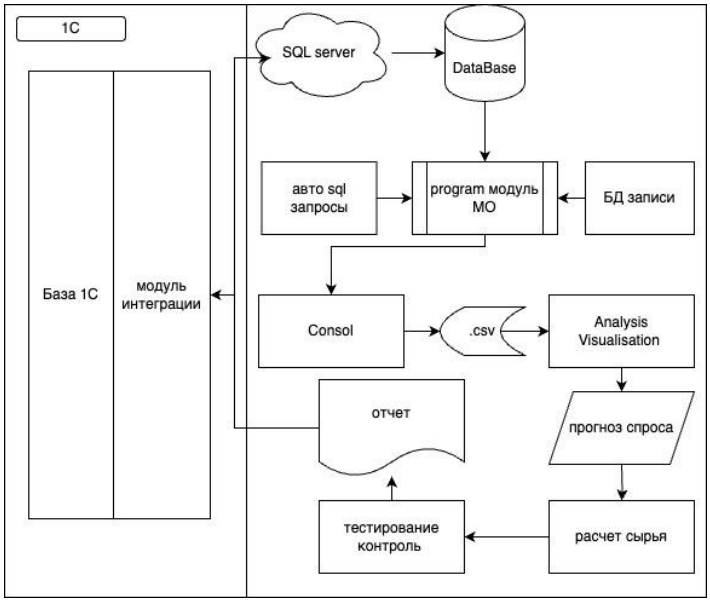

Рисунок 36 – Схема интеграции алгоритма прогнозирования сырья в общую архитектуру предприятия

Интеграция системы прогнозирования потребности в сырье в структуру предприятия начинается с тщательного планирования архитектуры данных. Первично необходимо создать надежный пайплайна, который будет заниматься сбором, хранением, извлечением и обработкой информации. Основной задачей является интеграция многообразия источников данных в единую реляционную базу данных, что подразумевает разработку точной схемы организации данных и определение взаимосвязей между элементами базы.

Обязательно нужно проводить тестирование и оптимизацию, они играют ключевую роль в подтверждении функциональности системы. Нагрузочное тестирование и интеграционное тестирование обеспечивают выявление и устранение потенциальных проблем производительности и совместимости компонентов модуля для прогнозирования сырья.

Оформление документации по архитектуре системы важный этап для дальнейшего использования модуля для поддержки разработчиками и эксплуатацией пользователей. Наличие документации упрощает процесс обучения и внедрения системы в текущую инфраструктуру предприятия.

Финальный этап включает развертывание системы и настройку мониторинга. Особое внимание должно уделяться не только техническим аспектам, но и организационным, таким как обучение сотрудников и поддержка пользователей. Эти меры гарантируют, что система останется функциональной и эффективной, соответствуя развивающимся потребностям бизнеса.

Для оценки экономической эффективности внедрения модуля машинного обучения, который позволит рассчитывать потребность в сырье для закупок на производство, необходимо учитывать несколько ключевых факторов. При годовом объеме закупок сырья в 44 089 750 рублей, ожидаемое сокращение излишков сырья составит 15% от годового объема закупок, что приведет к уменьшению затрат на хранение излишков, составляющих 5% от

их объема, и затрат на утилизацию просроченного сырья, составляющих 2% от объема излишков.

До внедрения модуля годовые излишки сырья составляли 6,613,462.5 рублей. Сокращение излишков на 15% снизит их до 5 621 443 рублей, что приведет к экономии от сокращения излишков в размере 992 019 рублей. Также будет сэкономлено на хранении и утилизации излишков 462 942,37 рублей в год, что в общей сложности даст годовую экономию в размере 1 454 961,75 рублей.

Инвестиционные затраты на внедрение системы включают разработку и внедрение модуля стоимостью 2 000 000 рублей, обучение персонала за 50 000 рублей и ежегодную поддержку и обслуживание системы за 200 000 рублей. Общие затраты на внедрение и поддержку системы за три года составят 2 650 000 рублей.

Годовая экономия от внедрения системы составит 1 454 961,75 рублей, а общая экономия за три года достигнет 4 364 885,25 рублей. Таким образом, чистый экономический эффект за этот период будет равен 1 714 885,25 рублей, что свидетельствует о положительной экономической эффективности внедрения модуля машинного обучения для расчета потребности в сырье. При значительном объеме закупок сырья внедрение такого модуля становится экономически целесообразным.

Выводы по третьему разделу:

В разделе рассмотрены используемые подходы в реализации моделей, проанализированы результаты полученных метрик выбранных моделей и методов. Выбрана модель для прогнозирования спроса градиентный бустинг над решающими деревьями (XGBoost) как наиболее быстрая и точная. Разработан алгоритм прогнозирования сырья на одну неделю вперед и сформированы рекомендации по внедрению модели в архитектуру предприятия и рассчитан предполагаемый экономический эффект от внедрения данного модуля. Рекомендуется также в будущем рассмотреть применением моделей для панельных данных.

ЗАКЛЮЧЕНИЕ

В рамках магистерской диссертации решается актуальная современная задача проектирование алгоритма прогнозирования на базе машинного обучения, нейронных сетей и статистических моделей.

Решены следующие задачи:

– исследованы методологии и модели прогнозирования, а также актуальные цифровые сервисы и системы, включающие функцию прогнозирования спроса и сырья. Проведен обзор существующих методов и систем прогнозирования спроса и сырья, включая традиционные статистические модели (ARIMA, SARIMA) и современные методы машинного обучения (XGBoost, LSTM). Пронализированы различные цифровые сервисы и программные решения для прогнозирования спроса и управления запасами, выявлены их сильные и слабые стороны;

– проведен детальный анализ текущих процессов прогнозирования сырья на предприятии. Установлено, что на данный момент предприятие не использует автоматизированные системы управления запасами и прогнозирования сырья. Выявлены основные проблемы, связанные с отсутствием автоматизации, а именно сложности в планировании производственных процессов, дефицит финансирования и нехватка системы прогнозирования спроса, что приводит к неэффективному управлению запасами;

– исследованы статистические модели, модели машинного обучения и нейронные сети и проведена оценка точности моделей с использованием различных метрик (MAE, RMSE, MAPE) для прогнозирования спроса на базе собранных наборов данных и различных источников. Каждая из рассмотренных статистических моделей имеет свои сильные и слабые стороны. ARIMA и SARIMA показали умеренные результаты, причем SARIMA продемонстрировала лучшее объяснение сезонных колебаний. Holt-Winters достигла высокой точности по абсолютным показателям, но имела

проблемы с малыми значениями. Prophet показала наиболее сбалансированные результаты, демонстрируя хорошую точность как по абсолютным, так и по относительным показателям, что делает её наиболее предпочтительной моделью для прогнозирования спроса на сырья, если необходимо использовать статистическую модель. У всех статистических моделей есть существенный минус, они время затратны. Базовая модель градиентного бустинга (XGBoost) продемонстрировала высокую точность и стабильность с коэффициентом детерминации R^2 0,769, средней квадратичной ошибкой (MSE) 1859,26, средней абсолютной ошибкой (MAE) 20,69 и средним абсолютным процентом ошибки (MAPE) 0,47%. Эти показатели делают её оптимальной для прогнозирования спроса на сырье. Модели на основе деревьев решений, такие как сами деревья решений, бэггинг и случайный лес, показали схожие результаты, однако уступают XGBoost по точности. Метод K-ближайших соседей показал высокий R^2, однако его MSE было крайне высоким, а MAE составила 95,70, что указывает на значительные ошибки в прогнозах и делает модель менее надежной, несмотря на высокие показатели R^2. Линейная регрессия продемонстрировала хорошие результаты, уступая только базовой модели XGBoost по точности. Модели на основе нейронных сетей, такие как LSTM и BiLSTM, показали противоречивые результаты. LSTM имела высокий R^2, но с высокими ошибками MSE и MAE, что указывает на проблемы с точностью. BiLSTM продемонстрировала слабые результаты с R^2, MSE и MAE. Модель TimesNet оказалась наименее эффективной среди рассмотренных нейронных моделей.

– выбрана наиболее производительная улучшенная модель экстремального градиентного бустинга над решающими деревьями, реализованная с помощью библиотеки XGBoost, которая имеет лучшие метрики качества, ошибка предсказания на уровне 9% и точности на уровне 97%, имеет высокую скорость работы. Также исследованы возможности использования нейронных сетей (LSTM) для улучшения точности прогнозов,

однако они показали менее стабильные результаты по сравнению с XGBoost для данного набора данных.

– разработан комплексный алгоритм прогнозирования сырья, включающий два основных этапа: прогнозирование спроса и расчет потребности в сырье на основании прогнозов спроса. Прототип алгоритма реализован с использованием фреймворка Streamlit, что позволяет обеспечить гибкость и удобство в использовании для предприятия. Проведен предварительный анализ экономической эффективности внедрения разработанного алгоритма. Результаты показали значительное сокращение издержек и улучшение финансовых показателей за счет более точного прогнозирования и эффективного управления запасами. Предложены рекомендации по внедрению разработанного алгоритма на предприятии, включая этапы интеграции с существующими системами и обучение персонала. Сделан акцент на необходимости регулярного обновления моделей и данных для поддержания высокой точности прогнозов. Планируется проводить дальнейшие исследования и тестирование новых моделей машинного обучения и нейронных сетей для возможного повышения точности прогнозов в будущем так как будет увеличиваться количество исторических данных. На текущий момент система сбора данных урегулирована и должна проходить на необходимом уровне для будущего построения анализа и прогнозирования.

Таким образом, поставленная цель достигнута, спроектирован алгоритм прогнозирования сырья и передан для внедрения на предприятие. Данный алгоритм готов к внедрению и дальнейшему использованию для повышения эффективности работы предприятия.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ганичев Н. А. Новый цифровой разворот - от дискурса экономического роста к технологическому расколу мира и принудительной рационализации / Н.А. Ганичев // Вопросы теоретической экономики. – 2022. – № 4. – С. 7-24.

2. Behera G. A Comparative Analysis of Weekly Sales Forecasting Using Regression Techniques / G. Behera, A. Bhoi, A.K. Bhoi // Intelligent Systems / eds. S.K. Udgata, S. Sethi, X.-Z. Gao. – Singapore: Springer Nature. – 2022. – P. 31-43.

3. Гаджинский А. М. Проектирование товаропроводящих систем на основе логистики: учебник. — 3-е изд., стер. / А. М. Гаджинский. – Москва: Дашков и К, 2021. – 322 с.

4. Тедеев К. С. Модель управления запасами для повышения эффективности деятельности предприятия розничной торговли / К.С. Тедеев, Л.Г. Протасова // Управленец. –2017. –№ 5 (69). – С. 98-103.

5. Уразбахтин И. Р. Прогнозирование динамики спроса на запасы промышленного предприятия с высокой степенью изменчивости / И.Р. Уразбахтин // Управление экономическими системами: электронный научный журнал. –2015. –№ 10 (82). – С. 39-47.

6. Мазманова Б. Г. Методические вопросы прогнозирования сбыта / Б.Г. Мазманова // Маркетинг в России и за рубежом. 2000. –№ 1. – С. 15-35.

7. Потапова И. И. Теоретические и практические аспекты анализа и прогнозирования потребительского спроса на рынке продовольственных товаров / И.И. Потапова, О.К. Минева // Нефтегазовые технологии и экологическая безопасность. –2004. –№ 3. – С. 52-57.

8. Светуньков И. С. Методы социально-экономического прогнозирования: учебник и практикум для вузов: в 2 т. Т. 1. Теория и методология / И. С. Светуньков, С. Г. Светуньков – Москва: Юрайт, 2015. – 351 с.

9. Рон Хайндман, Джордж Атанасопулос. Прогнозирование: принципы и практика / Рон Хайндман, Джордж Атанасопулос. – Москва: ДМК Пресс, 2023. – 458 с.

10. Амирханова П. М. Методы прогнозирования спроса / П.М. Амирханова // Вестник науки. – 2020. Т. 4. – № 4 (25). – С. 40-42.

11. Андриевская Н. К. Применение статистических методов, кластерного анализа и нейро-сетевых технологий при прогнозировании закупочных цен лекарств / Н.К. Андриевская, Т.В. Мартыненко, Т.А. Васяева // Проблемы искусственного интеллекта. – 2023. – № 4 (31). – С. 41-55.

12. Михайлова Е. Б. Проблема классификации моделей и методов прогнозирования / Е.Б. Михайлова // Учет и статистика. – 2017. – № 1 (45). – С. 75-81.

13. Прокофьев О. В. Методы и модели прогнозирования временных рядов / О.В. Прокофьев, А.Е. Савочкин // Современные информационные технологии. – 2018. – № 28. – С. 40-43.

14. Шелест А. В. Обзор методов и моделей прогнозирования временных рядов / А.В. Шелест, К.А. Пархоменко // Компьютерное проектирование и технология производства электронных систем: сборник тезисов 54 научной конференции аспирантов, магистрантов и студентов, Минск, 23–27 апреля 2018 г. – Белорусский государственный университет информатики и радиоэлектроники ; отв. ред. Раднёнок А. Л. – Минск, 2018. – С. 112 - 113.

15. Мамонтов Д. В. Классификация методов и моделей прогнозирования / Д.В. Мамонтов, С.В. Селезнев // Устойчивое развитие горных территорий. – 2014. – № 1. – С. 51-55.

16. Селиверстова А. В. Сравнительный анализ моделей и методов прогнозирования / А.В. Селиверстова // Современные научные исследования и инновации. – 2016. – № 11. – С.241-248.

17. Нильсен Э. Практический анализ временных рядов: прогнозирование со статистикой и машинное обучение / Э. Нильсен – СПб: ООО «Диалектика», 2021. – 544 с.

18. Канаева Н. Н. Сравнительный анализ моделей прогнозирования спроса на гостиничные услуги / Н.Н. Канаева, В.Е. Степанова, П.А. Новгородова // Экономические исследования и разработки. – 2018. – № 4. – С. 142-154.

19. Fattah J. Forecasting of demand using ARIMA model/ J. Fattah// International Journal of Engineering Business Management. 2018. Т. 10. – P. 184797901880867.

20. Benhamida F. Z. Demand Forecasting Tool For Inventory Control Smart Systems / F. Z. Benhamida // Journal of Communications Software and Systems. 2021. – Vol. 17. № 2. – P. 185-196.

21. Seyedan M. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities / M. Seyedan, F. Mafakheri // Journal of Big Data. – 2020. – Т. 7. Predictive big data analytics for supply chain demand forecasting. –P. 53-76.

22. Pavlyshenko B. Machine-Learning Models for Sales Time Series Forecasting / B. Pavlyshenko // Data. – 2019. – Т. 4. – P. 15-24.

23. Zohdi M. Demand forecasting based machine learning algorithms on customer information: an applied approach / M. Zohdi // International Journal of Information Technology. – 2022. – Т.14. – № 4. – P. 1937-1947.

24. Lazzeri F. Machine Learning for Time Series Forecasting with Python / F. Lazzeri. – John Wiley & Sons, 2020. – 224 p.

25. Dairu X. Machine Learning Model for Sales Forecasting by Using XGBoost / X. Dairu, Z. Shilong // 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). – 2021. – P. 480-483.

26. Niu Y. Data Prediction Based on Support Vector Machine (SVM)—Taking Soil Quality Improvement Test Soil Organic Matter as an Example / Y. Niu, S. Ye // IOP Conference Series: Earth and Environmental Science. – 2019. –Vol. 295. № 2. – P. 012021.

27. Vaswani A. et al. Attention is all you need/ A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, I. Polosukhin//Advances in neural information processing systems. – 2017. – T. 30. – P. 1-4.

28. Joseph R. V. A hybrid deep learning framework with CNN and Bi-directional LSTM for store item demand forecasting / R. V. Joseph // Computers and Electrical Engineering. –2022. –T. 103. – P. 108358.

29. Phyu M. M. Retail Demand Forecasting Using Sequence to Sequence Long Short-Term Memory Networks / M.M. Phyu, M.T. Khine // 2023 IEEE Conference on Computer Applications (ICCA). –2023. – P. 208-213.

30. Dai Y., Huang J. A sales prediction method based on lstm with hyper-parameter search/ Y. Dai, J. Huang //Journal of Physics: Conference Series. IOP Publishing. –2021. – №. 1. – T. 1756. – P. 012015.

31. Wen Q. et al. Transformers in time series: A survey/ Q. Wen //arXiv preprint arXiv:2202.07125. – 2022. – P. 6778-6786.

32. Zeng A. et al. Are transformers effective for time series forecasting? / A. Zeng, M. Chen, L. Zhang, Q. Xu // Proceedings of the AAAI conference on artificial intelligence. – 2023. – T. 37. – №. 9. – P. 11121-11128.

33. Yue Z. et al. Ts2vec: Towards universal representation of time series/ Z. Yue, Z. Yue, Y. Wang, J. Duan, T. Yang //Proceedings of the AAAI Conference on Artificial Intelligence. – 2022. – T. 36. – №. 8. – P. 8980-8987.

34. Oreshkin B. N. et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting / B. N. Oreshkin, D. Carpov, Y. Bengio, N. Chapados //arXiv preprint arXiv:1905.10437. – 2019. – P. 5442-5447.

35. Challu C. et al. Nhits: Neural hierarchical interpolation for time series forecasting / C. Challu, Kin G. Olivares, B. N. Oreshkin, F. Garza // Proceedings of the AAAI Conference on Artificial Intelligence. – 2023. – T. 37. – №. 6. – P. 6989-6997.

36. Zhang Y. et al. Multi-resolution time-series transformer for long-term forecasting/ Y. Zhang // International Conference on Artificial Intelligence and Statistics. – PMLR, 2024. – P. 4222-4230.

37. Das A. et al. Long-term forecasting with tide: Time-series dense encoder / A. Das // arXiv preprint arXiv:2304.08424. – 2023. – P. 1-21.

38. Chen S. A. et al. Tsmixer: An all-mlp architecture for time series forecasting/ S. A. Chen // arXiv preprint arXiv:2303.06053. – 2023. – P.1-24.

39. Rasul K. et al. Lag-llama: Towards foundation models for time series forecasting/ K. Rasul // arXiv preprint arXiv:2310.08278. – 2023. – P. 1-23.

40. Dutt A. A Comparative Analysis of TimeGPT and Time-LLM in Predicting ESP Maintenance Needs in the Oil and Gas Sector / A. Dutt, A. Chotrani // International Journal of Computer Applications. –2024. –Т. 186. – P. 975-8887.

41. Wu H. et al. Timesnet: Temporal 2d-variation modeling for general time series analysis / H. Wu // The eleventh international conference on learning representations. – 2022. – P. 1-23.

42. Гильгенберг Е. О. Прогнозирование спроса с помощью нейросетей / Е.О. Гильгенберг, Я.И. Никонова // Всероссийская научно-практическая конференция. – Хабаровск, 2023. – С. 30-35.

43. Ramos P., Santos N., Rebelo R. Performance of state space and ARIMA models for consumer retail sales forecasting / P. Ramos, N. Santos, R. Rebelo // Robotics and computer-integrated manufacturing. – 2015. – Т. 34. – P. 151-163.

44. Li X., Kang Y., Li F. Forecasting with time series imaging / Li X., Y. Kang, F. Li // Expert Systems with Applications. – 2020. – Т. 160. – P. 113680.

45. Сергеев В. И., Эльяшевич И. П. Управление закупками и запасами в цепях поставок: учебник / В. И. Сергеев, И. П. Эльяшевич - Москва: НИЦ ИНФРА-М., 2023. – 402 с.

46. Singh K. E-commerce system for sale prediction using machine learning technique/ K. Singh, P. Booma, M., U. Eaganathan // Journal of Physics: Conference Series. – IOP Publishing, 2020. – Т. 1712. – №. 1. – P. 012042.

47. Mihaylichenko L.A. Repository. – URL: https:// https://github.com/alduinsh/Ml_analysis_forecast_system (дата обращения: 24.05.2024).

ПРИЛОЖЕНИЕ А

Таблица А.1 – Стратегии управления поставками

Стратегия	Описание	Преимущества	Недостатки
Just–in–Time (JIT)	Эта политика направлена на минимизацию уровня запасов путем поставки и использования материалов перед их необходимостью.	– Снижение затрат на хранение.	– Риск прерывания производства при задержках поставок. –Требует высокой координации с поставщиками.
Безопасный запас	Политика предусматривает поддержание минимального запаса для обеспечения непрерывности производства в случае задержек.	– Обеспечение непрерывности производства.	– Дополнительные затраты на удержание запасов. –Риск образования излишних запасов.
Оптимальный размер заказа	Цель – определить оптимальный размер заказа для минимизации издержек на удержание запасов и размещение заказов.	– Минимизация издержек на удержание запасов и размещение заказов.	– Проблемы с точностью прогнозирования и объемом заказа.
Циклическое управление запасами	Заказы размещаются периодически с фиксированными интервалами для упрощения управления запасами.	– Простота управления запасами. –Возможность планирования и оптимизации заказов.	– Возможно образование избыточных запасов. –Риск недостачи товаров при высоком спросе.
Динамическое управление запасами	Уровень запасов корректируется в реальном времени в зависимости от изменений внешних условий.	– Гибкость в реагировании на изменения внешних условий. –Минимизация рисков при изменениях спроса.	– Требуется более сложное и автоматизированное управление запасами. – Проблемы с точностью данных и прогнозированием.

ПРИЛОЖЕНИЕ Б

Таблица Б.1 – Преимущества и недостатки методов машинного обучения

Метод машинного обучения	**Преимущества**	**Недостатки**
Линейная регрессия	Простота и интерпретируемость	Предполагает линейную зависимость
Методы деревьев решений	Обработка нелинейных зависимостей	Склонность к переобучению при большом количестве признаков
Методы кластеризации	Выявление скрытых структур в данных	Не всегда применимы к экономическим данным и прогнозированию
Методы нейронных сетей	Способность обрабатывать сложные данные	Требуют большого объема данных и вычислительных ресурсов
Методы машинного обучения с подкреплением	Учитывают взаимодействие агента с окружающей средой	Требуют длительного обучения и настройки параметров
Методы ансамблевого обучения	Улучшение обобщающей способности модели	Требуют большого объема данных и вычислительных ресурсов
Глубокое обучение	Высокая гибкость и способность к обработке больших данных	Требуют большого объема данных и вычислительных ресурсов

ПРИЛОЖЕНИЕ В

Таблица В.1 – Сравнение ПО и цифровых сервисов

ПО	Описание	Преимущества	Недостатки
GoodsForecast Integrated Planning Platform (ООО «Гудфокаст»)	Комплексное решение (платформа) для планирования и прогнозирования, которое использует машинное обучение для оптимизации запасов и производственных процессов.	Интегрированное планирование цепочек поставок и продаж. Гибкая настройка под бизнес–процессы предприятия. Машинное обучение для прогнозирования спроса. Поддержка различных уровней детализации данных.	Высокая стоимость внедрения и обслуживания. Требуется существенная подготовка данных. Ограниченная интеграция с некоторыми ERP–системами. Дороговизна.
Галактика AMM	Российская система автоматизированного материально–технического обеспечения. Включает в себя управление закупками, учет материалов, планирование и оптимизацию запасов.	Широкий функционал для управления складскими запасами и планирования производства. Гибкие настройки отчетности и аналитики. Интеграция с ERP–системой «Галактика».	Сложный интерфейс для новых пользователей. Ограниченная поддержка прогнозирования спроса. Трудоемкое внедрение. Дороговизна.
In.Plan (ООО «Акстим Тех»)	Система гибких инструментов для планирования и оптимизации производственных процессов с динамическим планированием	Интегрированное управление производством и снабжением. Планирование производства на основе прогноза спроса. Гибкая настройка бизнес–процессов.	Ограниченная аналитика и отчетность. Ограниченные возможности интеграции с внешними системами. Требуется обучение пользователей. Дороговизна.
1С: Предприятие	Программы 1С предлагают широкий спектр решений для различных отраслей и включают модули для HR, CRM, ERP и других нужд бизнеса, также имеет встроенное моделирование прогнозирования	Широкий спектр решений для управления предприятием. Интеграция с другими продуктами 1С. Гибкость в настройке под конкретные бизнес–процессы. Доступная стоимость для малых и средних предприятий.	Ограниченные возможности прогнозирования спроса. Трудоемкая интеграция с внешними системами. Ограниченная аналитика в стандартных конфигурациях. Не используют МО.

Продолжение ПРИЛОЖЕНИЯ В

ПО	Описание	Преимущества	Недостатки
Knoweledge Space (ООО «Интегрированные системы управления»)	Система управления цепочками поставок и производственными процессами. Поддерживает прогнозирование спроса, управление запасами и оптимизацию производства.	Интегрированное управление цепочками поставок. Гибкая настройка бизнес–процессов. Поддержка различных моделей прогнозирования спроса.	Ограниченная интеграция с ERP–системами. Сложный интерфейс для новых пользователей. Высокие требования к качеству данных.
Loginom Planiqum Suite (ООО «Решейп Аналитикс»)	Аналитическая платформа для прогнозирования и планирования производственных процессов. Поддерживает машинное обучение и интеграцию с ERP–системами.	Широкий функционал для прогнозирования и планирования. Поддержка различных моделей и алгоритмов машинного обучения. Гибкая настройка бизнес–процессов.	Высокая стоимость внедрения. Требуется обучение пользователей. Ограниченная интеграция с некоторыми ERP–системами. Дороговизна.
«Большая птица»	Программное решение для управления цепочками поставок и планирования производства	Простая и интуитивно понятная система. Быстрое внедрение и настройка. Поддержка основных моделей прогнозирования спроса.	Ограниченная функциональность. Ограниченная аналитика. Требуется подготовка данных для качественного прогнозирования.
Корус управление запасами	Решение для управления запасами. Интегрируется с ERP–системой «Корус» и позволяет прогнозировать спрос и оптимизировать запасы.	Интеграция с ERP–системой «Корус». Гибкие настройки для управления запасами. Интеграция с системами прогнозирования спроса.	Ограниченные возможности прогнозирования спроса. Ограниченная аналитика и отчетность. Трудоемкое внедрение.
Optimacros (ООО «Оптимакрос»)	Платформа для прогнозирования и оптимизации цепочек поставок. Использует модели машинного обучения для прогнозирования спроса и управления запасами.	МО для прогнозирования спроса. Гибкая настройка под бизнес–процессы предприятия. Интеграция с ERP–системами.	Высокая стоимость внедрения. Требуется подготовка данных для прогнозирования. Ограниченные возможности отчетности и аналитики.

Продолжение ПРИЛОЖЕНИЯ В

ПО	Описание	Преимущества	Недостатки
Microsoft Azure	Облачная платформа для прогнозирования спроса и управления цепочками поставок. Поддерживает машинное обучение и интеграцию с ERP–системами.	Интегрированное планирование цепочек поставок и продаж. Широкий функционал для прогнозирования спроса. Интеграция с ERP–системой SAP.	Высокая стоимость внедрения и обслуживания. Трудоемкое обучение пользователей. Требуется существенная подготовка данных. Дороговизна. Не поддерживается в РФ
Amazon Forecast	блачный сервис прогнозирования спроса на основе машинного обучения, предоставляемый Amazon Web Services. Интегрируется с другими облачными сервисами AWS.	Простое и быстрое внедрение. Поддержка различных моделей прогнозирования спроса. Интеграция с облачными сервисами AWS.	Ограниченная аналитика и отчетность. Требуется подготовка данных для качественного прогнозирования. Ограниченная интеграция с ERP–системами.

ПРИЛОЖЕНИЕ Г

Таблица – Этапы и содержание производственного процесса

=> 1 Лаборатория	1 Лабораторные изыскания	2 Тестирование лабораторных образцов, проверка стабильности формулы	3 Передача документов на сертификацию декларирование соответствия	4 Передача рецептуры и технологической производственной карты в производство
=> 2 Цех развески	1 Приемка и тестирование поступившего сырья	2 Подготовка сырья в производство	3 Передача готовой развески в цех варки или в цех мыловарения	
=> 3 Цех Варки или Мыловарения	1 Приемка сырья	2 Процесс варки – изготовление англо полуфабрикат	3 По окончании варки англо передается в карантин для проведения тестирования на соответствие продукции (проводит ОТК)	4 Составление акта соответствия продукта, выпуск протокола качества – передача в цех фасовки
=> 4 Цех фасовки	1 Приемка тары со склада для фасовки (флакон/банка)	2 Фасовка продукции в финальную тару, производство укупоривания и этикетирования	3 Выборочная проверка готового продукта службой ОТК	
=> 5 Склад	1 Прием и пересчет продукции поступившей из зоны фасовки на склад	2 Перемещение продукта в зону хранения до востребования	3 Распределение продукта по цехам сборки	

ПРИЛОЖЕНИЕ Д

Таблица Д.1 – Список информационного содержания в P&L 2021–2023

Колонка	Описание
Свод бизнес план 24/25	Краткая сводка планов и прогнозов на 2024 и 2025 годы.
Monthly Pack PL	Ежемесячный план, связанный с пакетами услуг или продуктов.
План 2024 Ideal	Идеальный план на 2024 год.
Plan 2024 Optimal	Оптимальный план на 2024 год.
P&l 2023	Отчет о прибылях и убытках за 2023 год.
P&L 2022	Отчет о прибылях и убытках за 2022 год.
P&L 2021	Отчет о прибылях и убытках за 2021 год.
CF2023	Денежный поток за 2023 год.
CF2022	Денежный поток за 2022 год.
CF2021	Денежный поток за 2021 год.
План СТМ/ОПТ 2024	План продаж по каналам СТМ (сетевая торговля) и ОПТ (оптовые продажи) на 2024 год.
WB2024	План продаж для Warner Bros на 2024 год.
OZON2024	План продаж для OZON на 2024 год.
ДДС 2023 массив	Массив данных о декларациях по налогу на добавленную стоимость за 2023 год.
ДДС 2022 массив	Массив данных о декларациях по налогу на добавленную стоимость за 2022 год.
ДДС 2021 массив	Массив данных о декларациях по налогу на добавленную стоимость за 2021 год.
ВБ 21–23	Данные о выручке за период с 2021 по 2023 год.
Озон 21–23	Данные о выручке от продаж на платформе «Озон» за период с 2021 по 2023 год.
Интернет–магазин 21–23	Данные о выручке от продаж в интернет–магазине за период с 2021 по 2023 год.

Таблица Д.2 – Отчет по продажам

Колонка	Описание
Период	Временной период, к которому относятся данные.
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Номенклатура	Наименование товара или услуги.
Категория	Категория товара отражает группу товаров, в которую входит номенклатура
Количество	Количество проданных единиц товара или услуги.
Выручка без НДС	Сумма выручки без учета налога на добавленную стоимость.
Цена реализации без НДС	Цена товара или услуги без налога на добавленную стоимость.
Валовая прибыль	Прибыль, оставшаяся после вычета всех себестоимостей.

Таблица Д.3 – PR отчет

Колонка	**Описание**
Параметр	Описание
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Наименование	Наименование товара или услуги.
Период – 7 дней до рекламы	Временной период за 7 дней до начала рекламной акции.
Период – 5 день акции/рекламы	Временной период на 5–й день акции/рекламы.
Период – 5 дней после акции/рекламы	Временной период за 5 дней после завершения акции/рекламы.
Канал акции/рекламы	Канал, через который проводится акция или реклама.
Стоимость акции	Стоимость товара или услуги во время акции.
Стоимость полная	Полная стоимость товара или услуги без учета скидки.
Скидка %	Процентная скидка на товар или услугу.
Скидка руб/ед	Размер скидки на одну единицу товара или услуги.
Затраты на акцию/рекламу на ед	Затраты на проведение акции или рекламы на одну единицу товара или услуги.
Заказы шт	Количество заказанных единиц товара или услуг.
Заказы руб	Общая сумма заказов в рублях.
Конкуренция	Количество конкурентов на рынке

Таблица Д.3 – Оборотно–сальдовые ведомости

Колонка	**Описание**
Наименование	Наименование косметического товара.
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Счет	Бухгалтерский счет
Показатели количество	Показатели, связанные с количеством средств, находящихся на счете.
Показатели бухгалтерский учет	Показатели, отражающие финансовое состояние счета в бухгалтерском учете.
Кредит	Сумма денежных средств на счете, предоставленных банком или другим кредитором.

Таблица Д.4 – Ассортиментная матрица 2021–2024

Колонка	Описание
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Наименование	Наименование косметического товара.
Сегмент	Сегмент, к которому относится товар (например, уход за кожей, макияж и т.д.).
Вес	Вес товара.
Упаковка	Тип упаковки товара (например, банка, тюбик, флакон и т.д.).
Класс	Класс товара в рамках ассортиментной матрицы (например, эконом, стандарт, премиум и т.д.).
Пол	Предполагаемая аудитория товара по половому признаку (например, женский, мужской, унисекс).

Таблица Д.5 – Ежегодный отчет по себестоимости

Колонка	Описание
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Наименование	Наименование продукции или услуги.
Себестоимость производственная плановая	Плановая себестоимость производства товара или услуги.
Себестоимость производственная фактическая	Фактическая себестоимость производства товара или услуги.
Себестоимость производственная сметная	Сметная себестоимость производства товара или услуги.
Себестоимость продуктовая плановая	Плановая себестоимость продукции (включая себестоимость материалов).
Себестоимость продуктовая фактическая	Фактическая себестоимость продукции (включая себестоимость материалов).
Себестоимость продуктовая сметная	Сметная себестоимость продукции (включая себестоимость материалов).

Таблица Д.6 – Отчет по выпуску продукции – оптовые отгрузки:

Колонка	Описание
Уникальный идентификатор	Уникальный номер или код, идентифицирующий запись.
Наименование	Наименование продукции или услуги.
Дата	Дата отгрузки продукции.
Срок годности	Срок годности продукции в месяцах.
Количество	Количество единиц продукции, отправленных в оптовую отгрузку.
Вид упаковки	Тип упаковки продукции (например, коробка, паллет, контейнер и т.д.).
Счет учета	Счет, на котором отражается отгрузка продукции в бухгалтерском учете.

ПРИЛОЖЕНИЕ Е

Таблица Е1 – Статистики набора данных

Статистика	unique_id	date	weight	price	cost price	exp	y	кол–во компонентов	рейтинг товаров
count	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0
mean	276,5	44862,5	278,3	592,2	82,9	24,8	672,4	5,3	3,5
min	9,0	44362,0	15,0	108,0	23,2	24,0	1,0	3,0	2,5
25%	234,8	44612,3	50,0	391,0	52,8	24,0	20,0	4,5	3,3
50%	287,5	44862,5	100,0	465,0	73,4	24,0	50,0	5,0	3,5
75%	340,3	45112,8	250,0	653,0	96,0	24,0	250,0	5,8	3,8
max	423,0	45363,0	5000,0	1999,0	324,4	36,0	4500,0	8,0	4,5
std	85,0	0,0	687,2	362,7	43,1	3,1	1292,5	1,8	0,7
Статистика	**кол–во просмотров**	**month**	**day_of_week**	**sales_3m_avg**	**sales_6m_avg**	**sales_12m_avg**	**discount**	**seasonal_discount**	
count	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	30528,0	
mean	5971,7	6,7	1,0	129,6	128,9	127,6	1,9	0,9	
min	4000,0	1,0	1,0	0,0	0,0	0,0	0,0	0,0	
25%	4000,0	3,0	1,0	20,0	20,0	20,6	0,0	0,0	
50%	4500,0	7,0	1,0	36,0	40,0	41,0	1,0	0,0	
75%	7500,0	10,0	1,0	66,0	82,0	118,7	2,0	0,0	
max	11000,0	12,0	1,0	2881,0	2881,0	2881,0	13,0	5,0	
std	2273,3	3,5	0,0	310,1	295,5	275,6	2,3	1,9	

ПРИЛОЖЕНИЕ Ж

Таблица – Сравнение используемой архитектуры BiLSTM и LSTM

Критерий	BiLSTM	LSTM
Входные данные	Временные ряды и дополнительные признаки обрабатываются отдельно	Все признаки, включая временные, объединяются в один входной вектор
Архитектура	Модель строится с использованием Sequential API из Keras. Состоит из LSTM слоев с регуляризацией L2 и Dropout слоем для предотвращения переобучения.	Слой LSTM. Первый слой модели – это LSTM слой. Он принимает на вход последовательности данных и обрабатывает их, учитывая временные зависимости.
	Дополнительные признаки обрабатываются отдельно с помощью Input и Dense слоев.	Полносвязные слои. За LSTM слоем следуют два полносвязных слоя. Первый слой использует функцию активации ReLU, второй – линейную функцию активации.
	Результаты объединяются с помощью Concatenate слоя. В конце добавляется выходной Dense слой.	Выходной слой. Выходной слой содержит один нейрон и выдает предсказанное значение целевой переменной.
Слои LSTM	Используются двунаправленные слои LSTM с регуляризацией L2 и выпадением (dropout)	Используется несколько слоев LSTM
Выход	Полносвязные слои с активацией ReLU и выпадением, выходной слой с одним нейроном	Полносвязные слои с активацией ReLU, выходной слой с одним нейроном
Обработка временных рядов	Отдельная ветвь для временных рядов	Временные ряды объединяются с другими признаками
Сложность модели	Более сложная, больше параметров	Проще, меньше параметров
Объединение признаков	Используется Concatenate	Не используется
Интерпретируемость	Сложнее интерпретировать влияние отдельных признаков	Проще интерпретировать влияние признаков