Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт экономики, государственного управления и финансов
Базовая кафедра цифровых финансовых технологий Сбербанка России

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Д. В. Солнцев
подпись
«____» _____2024 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Разработка модели прогнозирования продаж в розничной торговле на основе методов машинного обучения

38.04.01 «Экономика»
38.04.01.17 «Финансово-экономическая аналитика и принятие решений в цифровой среде»

Научный руководитель	_____	__к.э.н., доцент_	Ю.И. Черкасова
	подпись, дата	должность, ученая степень	инициалы, фамилия
Выпускник	_____		В.Р. Рассохин
	подпись, дата		инициалы, фамилия
Рецензент		рук. отдела экономики	
	_____	ООО «Командор-Холдинг»	Л.П. Привалихина
	подпись, дата	должность, ученая степень	инициалы, фамилия
Нормоконтролер	_____		Э.Ф. Мамедова
	подпись, дата		инициалы, фамилия

Красноярск 2024

РЕФЕРАТ

Выпускная квалификационная работа по теме «Разработка модели прогнозирования продаж в розничной торговле на основе методов машинного обучения» содержит 78 страниц текстового документа, 21 иллюстрации, 6 таблиц, 5 формул, 1 приложение, 52 использованных источников.

ПРОГНОЗИРОВАНИЕ ПРОДАД, МОДЕЛЬ ВРЕМЕННОГО РЯДА, РОЗНИЧНАЯ ТОРГОВЛЯ, РОГНОЗНАЯ АНАЛИТИКА, ОЦЕНКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ, МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Цели данной диссертационной работы – разработка масштабируемой для множества номенклатур модели прогноза продаж на основе алгоритмов машинного обучения.

В задачи исследования входит:

1. Провести литературный актуальных исследований по заданной теме;
2. Провести сбор и предварительную обработку исторических данных о продажах;
3. Обучить модели машинного обучения для прогнозирования продаж товара с возможностью масштабирования;
4. Оценить производительность модели с использованием метрик точности прогнозирования

Актуальность темы заключается с те, что современные ритейлеры сталкиваются с огромными объемами данных, поступающих из различных источников. Это создает необходимость в использовании методов машинного обучения, которые способны эффективно анализировать данные.

Результатом полученных исследования является протестированная на трех наборах данных модель прогноза продаж.

АННОТАЦИЯ

Крупнейшие розничные компании в мире такие как Walarant, X5 Group в поисках решения задачи увеличения точности прогноза продаж, данное исследования направлено на рассмотрения проблематики данной задачи с учетом современного развития предлагаемых решений в данной области бизнес-задач. Крупным розничным компаниям необходимо иметь методологию, программное обеспечение, группу специалистов по прогнозированию количества продаж многономенклатурной ассортиментной матрицы. Так как это часть крупной системы поддержки бизнеса, от которой зависят другие части структурированной системы бизнес-процессов, позволяющей компаниям оставаться на рынке и получать прибыль. Данная ячейка системы помогает правильно спроектировать систему расчета запасов продукции, минимизировав риски дефицита товаров и скопления их излишков на складах. Связанные с этой структурой отделы логистики и снабжения получают необходимую информацию, позволяющую оптимизировать план поставок и спланировать транспортировку товаров. Методология исследования включает поэтапное построение многономенклатурной модели прогноза продаж с реальными данными о продажах региональной розничной сети и использованием методов машинного обучения. Результатами исследования являются выстроенная система прогноза, товаров, применимая на практике. Модель дает возможность масштабировать решение задачи прогнозирование, которое можно применять к нескольким продуктам и торговым точкам с возможность последующего расчёта планируемого объема заказа поставщику. Предложенная модель может быть адаптирована для различных рыночных условий и типов продукции, обеспечивая гибкость учёта факторов влияния, и как следствие повышенную точность прогноза.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Возможность точно прогнозировать продажи имеет важное значение для компаний розничной торговли стремящихся развиваться и быть устойчивыми в конкурентной среде. Причиной этого является то, что для того, чтобы принимать обоснованные решения о закупе товара или его производстве, управлении запасами и распределении ресурсов необходимы оценочные прогнозы продаж, чем выше точность таких прогнозов, тем более эффективны бизнес-процессы использующие данную информацию. Прогнозирования продаж – это в первую очередь анализ временного ряда исторических данных, то есть выявление закономерностей в последовательных данных связанных с продажами и распределенных во времени с определенным интервалом между наблюдениями. С развитием электронной коммерции и сопутствующим изменением повеления потребителей традиционные методы прогнозирования спроса часто оказываются неоптимальными, что вынуждает ритейлеров обращаться к актуальным методам аналитике данных. Частично это связано с доступностью больших наборов данных и увеличением мощности вычислительных ресурсов, что позволило обучать сложные модели на основе больших наборах данных. Однако разработка точных моделей прогнозирования спроса остается сложной задачей, и аналитикам приходится тщательно оценивать сильные и слабые стороны различных методов машинного обучения [1]. Это исследование направлено на изучение использования моделей машинного обучения для прогнозирования продаж и оценку их эффективности с возможностью масштабирование для практического применения в розничной торговле. В отличие от традиционных методов, машинное обучение позволяет учитывать большое количество факторов и анализировать исторические данные для выявления скрытых закономерностей и трендов.

Объектом исследования является региональная розничная компания «Командор».

Предметом исследования является модель прогнозирования продаж на основе машинного обучения

Цели данной диссертационной работы – разработка масштабируемой для множества номенклатур модели прогноза продаж на основе алгоритмов машинного обучения.

В задачи исследования входит:

1. Провести литературный обзор актуальных исследований по заданной теме;
2. Описать объект исследования со стороны рассматриваемое проблемой прогнозирования продаж;
3. Провести сбор и предварительную обработку исторических данных о продажах;
4. Разработать и обучить модели машинного обучения для прогнозирования продаж товара;
5. Оценить производительность моделей с использованием метрик точности прогнозирования и с учетом масштабируемости использованных данных;

Теоретическую основу работы составляют труды отечественных и зарубежных авторов в области анализа данных, прогнозирования временных рядов, прогнозирования на основе алгоритмов машинного обучения. В данной работе использовались статьи, таких научных деятелей: Пивкин К.С. [2] исследующий применение методов машинного обучения для кластеризации групп товаров и поиска факторов, влияющих на продажи Э. Котович [3] с исследованием на тему прогнозирования спроса; У. Рен [4] Л. Менкулини [5] со сравнением эффективности использования моделей машинного обучения для прогнозирования временных рядов; труд по анализу временных рядов П. Гудвина [6]; У. Маккинни [7], Д. Плас [8], Дж. Грус [9] опубликовавшие учебники по языку программирования python и машинному обучению.

Базовые методики, принципы, приемы, которые отражены в этих работах легли в основу построения моделей прогнозирования и оценки их

эффективности использования на реальных данных. Так же автором данной работы были опубликованы две научные статьи на темы прогнозирования временных рядов, схожие с темой диссертационной работы.

Для достижения целей представленной работы используется методы логарифмирование, дифференцирования, метод Хольта-Винтерса, алгоритмы машинного обучения, аддитивные модели: SARIMAX, prophet. Для оценки точности прогнозирования исследуются такие показатели оценки, как средняя абсолютная ошибка, среднеквадратическая ошибка и средняя абсолютная процентная ошибка.

Научная новизна данной работы заключается в следующем:

1. Адаптация и масштабирование прогнозных моделей для больших данных. Представленное исследование вносит вклад в научное знание через разработку масштабируемых моделей прогнозирования, способных эффективно обрабатывать многотысячные наборы данных, отражающие продажи тысяч товаров. Это позволяет значительно улучшить управление запасами и оптимизировать бизнес-процессы в крупных розничных компаниях.

2. Практическое внедрение и оценка модели в реальных условиях. В работе проводится детальный анализ внедрения разработанных моделей на практике, включая этапы предварительной обработки данных, выбора и настройки моделей, их обучения и оценки производительности. Это обеспечивает возможность реального применения модели и её адаптации к изменениям во внешней среде.

Практическая значимость работы состоит в том, что построенная модель прогнозирования также является масштабируемой для применения на многотысячных данных отражающих продажи тысячи товаров. Эффективная модель прогнозирования может помочь оптимизировать бизнес-процессы, связанные с управлением запасами, расчетом необходимых логистических издержек, издержек, связанных с хранением и сортировкой товаров.

Диссертационная работа состоит из введения, заключения, списка использованной литературы и приложения и трех глав. Во введение читатель знакомится с актуальностью и значимостью выбранной темы, поставленными целями и задачами. Первая глава посвящена теоретическим основам прогнозирования, описанию различных алгоритмов машинного обучения, используемых для прогнозирования временных рядов, литературному обзору актуальных исследования на заданную тему. Во второй главе дан обзор объекта исследования – розничной компании, ее основные финансово-экономические показатели, приведен общий обзор основных направлений деятельности компании, основные товарные группы. Также дана общая статистика спроса на товарные группы, тренды, сезонность, обоснование выбранных факторов для дальнейшего моделирования. Третья глава сосредоточена на разработке и построении многономенклатурной модели прогнозирования продаж с использованием алгоритмов машинного обучения. Эта глава включает в себя описание процесса предварительной обработки данных, выбора и настройки моделей, обучения и оценки их производительности. Особое внимание уделяется выбору подходящих алгоритмов, сравнению их результатов и внедрению на практике. В заключении представлен краткое изложение основных выводов и результатов исследования разработки модели, оценка практической значимости разработанной модели, предоставлены возможные направлений для дальнейших исследований и улучшения модели.

Данное исследование рассчитано на структурирование имеющихся методов прогнозирования временных рядов и исследование возможностей их применения в практической задаче прогнозирования многомерных временных рядов.

1 Основы теории прогнозирования и применяемые методы машинного обучения

1.1 Теория прогнозирования, терминология, методы прогноза

Первыми исследованиями на тему прогнозирования временного ряда начались с 19 века. Впервые попытка разложить временной ряд на тренд и сезонность представлена в статье метеоролога Байса-Баллота, который выполнил разложение между трендом и сезонностью, моделируя тренд полиномом, а сезонность – фиктивными переменными. Затем в 1884 году Э. Пойнтинг [10] предложил метод анализа временных рядов, который стал известен как "метод декомпозиции временных рядов". Этот метод позволяет разделить временной ряд на компоненты, такие как трендовая составляющая, сезонная составляющая, чтобы лучше понять их поведение и особенности. Дальнейший толчком в развитии методов прогнозирования временных рядов стала работа Г. Эндрьюса и Д. Бокса 1970г. [11], где были разработаны первые модели авторегрессии и скользящего среднего и расширение этих моделей до моделей ARIMA. Параллельно с этим вышла работа К. Грэнджера и П. Ньюболда [12] ставшая классикой в эконометрики. В ней авторы рассматривают методы прогнозирования экономических временных рядов, применение авторегрессивных (AR) и-скользящих средних (MA) моделей которые используются для анализа и предсказания поведения экономических переменных, таких как ВВП, инфляция, уровень безработицы. В 1980-е годы началось активное развитие нейронных сетей. Одна из первых работ в которой описывается применения метода нейронной сети для прогнозирования временного ряда стала модель с использованием нейронной сети прямого распространения (FNN) диссертационное исследование П. Вербоса 1974 г. [13]. Две ключевые работы впоследствии оказавшие значительно влияние на эту область исследования были публикации Д.

Хопфилд 1982 г. [14] и Д. Румельхарт 1986 г. [15], впоследствии ставшие основой для разработки моделей прогнозирования с помощью рекуррентных нейронных сетей (RNN) впервые появившихся в статье Дж. Уильямса 1987 г.[16] и LSTM (долговременная кратковременная память) в работе 1997 г. Ю. Шмид-Хубера [17]. Хотя в этом исследовании не приводились конкретные примеры применения LSTM к временным рядам оно заложило основу для будущего исследования этих алгоритмов вышеупомянутого автора с использованием LSTM для прогноза временных рядов в публикации 1999г. [18]. Эти работы стали основами для применения нейронных сетей, как инструмента для моделирования временных зависимостей. Современная популярность применение этих моделей в задачах прогнозирования подтверждает это [19].

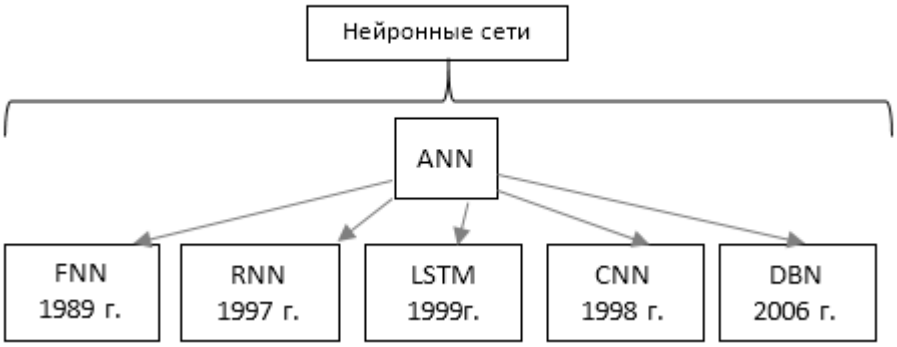

Рисунок 1 – Виды нейронных сетей используемые для прогнозирования временных рядов

На Рисунке 1 сгруппированы ранее упомянутые алгоритмы с использованием нейронных сетей.

В исследовании 1986 г. были впервые предложены Р. Куинланом [20] деревья решений – это один из основных методов машинного обучения, использующий древовидную структуру для представления решений и их возможных последствий. Метод деревьев решений стал важным инструментом в арсенале методов машинного обучения благодаря своей интерпретируемости, гибкости и эффективности при решении различных

задач, включая классификацию и регрессию. В 1995 г. В. Вапник [21] опубликовал работу по теоретическим основам статистического обучения, где впервые упоминается метод опорных векторов (SVM). Эта книга объединила теоретические аспекты и практические методы для решения задач классификации и регрессии, предоставив мощные инструменты для анализа данных. Дальнейшему развитию нейронный сетей их применение в прогнозирование временны рядов помогли работы Я. Лекуна, Дж. Бенджио и Дж. Хинтона в 1998 г. их публикация [22] описывает сверточные нейронные сети (CNN), которая произвела революцию в анализе изображений и нашла широкое применение в других областях. Сверточные нейронные сети стали основой для многих современных архитектур, используемых для прогнозирования временных рядов, таких как модели, комбинирующие CNN и LSTM. Эти модели способны учитывать, как пространственные, так и временные зависимости в данных, что значительно повышает точность прогнозов.

Следующим этапом в расширяющим методики прогнозирования является появление моделей бэггинга в опубликованной статье Л. Бреймана в 1996г. [23] бэггинг был предложен, как метод повышения точности предсказаний моделей машинного обучения. Основная идея заключается в том, чтобы создать несколько версий одной и той же модели, обучая их на различных подвыборках исходных данных, а затем комбинировать результаты, например, путем усреднения для регрессии или голосования для классификации для получения окончательного предсказания. Каждый бутстреп-набор данных содержит часть исходных наблюдений, возможно с повторениями, что позволяет моделям, обученным на этих наборах, захватывать различные аспекты данных и улучшать общую устойчивость предсказаний. На каждом бутстреп-наборе данных обучается отдельная модель, и все модели объединяются для получения итогового прогноза. Объединение предсказаний отдельных моделей, обученных на различных бутстреп-наборах данных, снижает дисперсию итоговой модели. В результате,

бэггинг делает предсказания более стабильными и устойчивыми к случайным шумам в данных, что особенно важно для высоко вариативных моделей, таких как деревья решений. Важно отметить, что бэггинг не всегда приводит к улучшению результатов. В некоторых случаях, особенно когда исходная модель уже имеет низкую дисперсию, применение может не дать значительного улучшения. Однако, в большинстве практических сценариев, особенно при работе с высоко вариативными данными, он демонстрирует свою эффективность [24].

В этом же журнале в 2001 г. Л. Бреймана опубликовал работу с методом случайного леса [25]. метод ансамблевого обучения, который строит множество деревьев решений во время обучения и выводит класс или среднее значение (в случае регрессии) отдельных деревьев для улучшения точности предсказаний. Основная идея заключается в использовании множества моделей (деревьев решений) для снижения риска переобучения и повышения общей точности модели. Преимуществом этой модели стало улучшенная точность благодаря снижению дисперсии и переобучении по сравнению с существующими в то время моделями. Метод также позволяет предоставлять информацию о важности различных признаков, что полезно для интерпретации модели. Однако у модели также недостатки, связанные с большими требованиями к вычислительному устройству из-за занимаемой при расчёт оперативной памяти при обучении с большими объемами данных. И особенностями отсутствия возможности учета трендовой составляющей в задачах прогнозирования [26].

Еще одним ставшим впоследствии популярным методом прогнозирования стал градиентный бустинг, появившийся в работе Д. Фридмана 2001 г. [27]. В этой статье Фридман представил алгоритм градиентного бустинга, описывая его как метод, который можно использовать для построения ансамбля слабых моделей, таких как упомянутые до этого деревья решений, чтобы улучшить их общую производительность. Эта работа заложила основы для широкого применения этого метода в задачах

машинного обучения и прогнозирования временных рядов. Впоследствии метод стали очень популярными благодаря своей эффективности и высокой точности различных задачах прогнозирования и классификации. В 2006 г. Д. Хинтон, С. О. и выше упоминавшийся Ю. Шмид-Хубер публикуют работу [28] в которой представлен алгоритм обучения «Глубокая сеть доверия» (DBN), которая предоставляет возможность реализации способа обучения без учителя так и дальнейшее обучение с учителем**.** В 2016 году было опубликовано исследование Т. Чен и К. Гуестрин [29]. Авторы описали систему масштабируемого бустинга деревьев, которая получила широкое признание среди специалистов по машинному обучению за достижение передовых результатов во многих задачах. XGBoost использует новаторский алгоритм, учитывающий разреженные данные, а также методы ускорения вычислений, такие как компрессия данных, что позволяет обрабатывать данные объемом более миллиарда примеров с меньшими ресурсами по сравнению с существующими системами. Схожая модель LightGBM, высокоэффективная реализация метода градиентного бустинга деревьев решений (GBDT), была представлена в статье, опубликованной в 2017 году [30].

Еще одной важной вехой в развитии нейронных сетей стало создание архитектуры трансформеров, представленной в статье А. Васвани в 2017 году [31]. Трансформеры значительно улучшили возможности обработки временных рядов, предоставив более гибкий и мощный инструмент для моделирования долгосрочных зависимостей. Эта архитектура, изначально разработанная для задач обработки естественного языка, быстро нашла применение в анализе временных рядов благодаря своей способности эффективно обрабатывать большие объемы данных [32]. Прогресс в области вычислительных мощностей и доступность больших данных способствовали развитию новых моделей и алгоритмов. Современные исследователи продолжают работать над улучшением архитектур нейронных сетей, делая их более точными и эффективными для решения различных задач. Например, комбинация методов глубокого обучения с классическими статистическими

моделями позволяет создавать гибридные модели, которые объединяют преимущества обоих подходов.

Кроме того, разработки в области объяснимости моделей стали важным направлением исследований. Понимание того, как и почему модель делает определенные прогнозы, имеет критическое значение для многих областей, таких как медицина и финансы. Новые методы интерпретации и визуализации помогают исследователям и практикам лучше понимать поведение моделей и доверять их прогнозам.

Прогнозирование - это процесс предсказания будущих событий или тенденций на основе анализа прошлых данных. Это позволяет организациям и индивидуумам принимать обоснованные решения, управлять рисками, и планировать свои действия.

Временной ряд – это последовательность наблюдений, выполненных последовательно во времени [33].

Метрика – это количественная мера, используемая для оценки производительности модели, алгоритма или системы. В контексте машинного обучения и анализа данных метрики используются для оценки точности, эффективности и результативности модели или алгоритма [34].

Сезонность – это повторяющиеся колебания или паттерны в данных, которые происходят с регулярными интервалами времени, такими как дни, недели, месяцы или годы. Примером могут быть ежемесячные продажи, которые возрастают в праздничные месяцы.

Тренд – это долгосрочное направление или общее движение данных в определенном направлении, которое может быть восходящим, нисходящим или горизонтальным. Тренд может указывать на общий рост или падение показателя со временем.

Стационарность – это свойство временного ряда, означающее, что его статистические характеристики (например, среднее значение и дисперсия) остаются постоянными во времени. Модели для прогнозирования временных рядов часто требуют стационарных данных для правильного анализа.

Автокорреляция – это корреляция временного ряда с его собственными прошлыми значениями. Это помогает выявить зависимости между текущими и прошлыми значениями данных.

Лаг (или задержка) – это временной интервал между значениями временного ряда, используемыми для прогнозирования. Например, лаг 1 означает использование предыдущего значения для прогнозирования текущего значения.

1.2 Характеристика моделей машинного обучения для прогнозирования продаж

Машинное обучение описывает способность систем учиться на обучающих данных по конкретной проблеме для автоматизации процесса построения аналитических моделей и решать сопутствующие задачи [35]. Машинное обучение – это область искусственного интеллекта, которая изучает методы и алгоритмы, позволяющие компьютерам извлекать закономерности из данных и использовать их для принятия решений или делать прогнозы на новых данных. Все классификаторы, а также регрессионные модели, относятся к методам машинного обучения. В контексте задач машинного обучения, как правило, присутствуют несколько условий по которым мы можем квалифицировать рассматриваемый алгоритм как метод машинного обучения. Эти условия существования данных для обучения, алгоритм обучения, целевая переменная, оценка эффективности, присутствие итерационного процесса в обучении. Рассмотрим подробнее каждое условие. В машинном обучение всегда присутствуют данные, которые содержат информацию о наблюдениях или объектах. Эти данные могут включать в себя различные признаки или характеристики. В машинном обучение присутствуют алгоритмы обучения для извлечения закономерностей из обучающих данных, которые позволяют делать прогнозы или

классифицировать новые данные. В данных всегда присутствует целевая переменная, которую необходимо предсказать или классифицировать. Например, в задачах регрессии целевая переменная может быть непрерывной, а в задачах классификации – категориальной. Классификаторы в машинном обучении обучаются на наборе обучающих данных, где каждый пример имеет метку класса или категории, и их задача состоит в том, чтобы научиться предсказывать класс или категорию для новых данных на основе их признаков. Это может включать в себя задачи, такие как определение категории электронной почты (спам или не спам), классификация изображений или текстов, и многое другое. Регрессионные модели также относятся к методам машинного обучения и используются для предсказания непрерывных значений, таких как цены на недвижимость, температура или стоимость акций, на основе входных признаков.

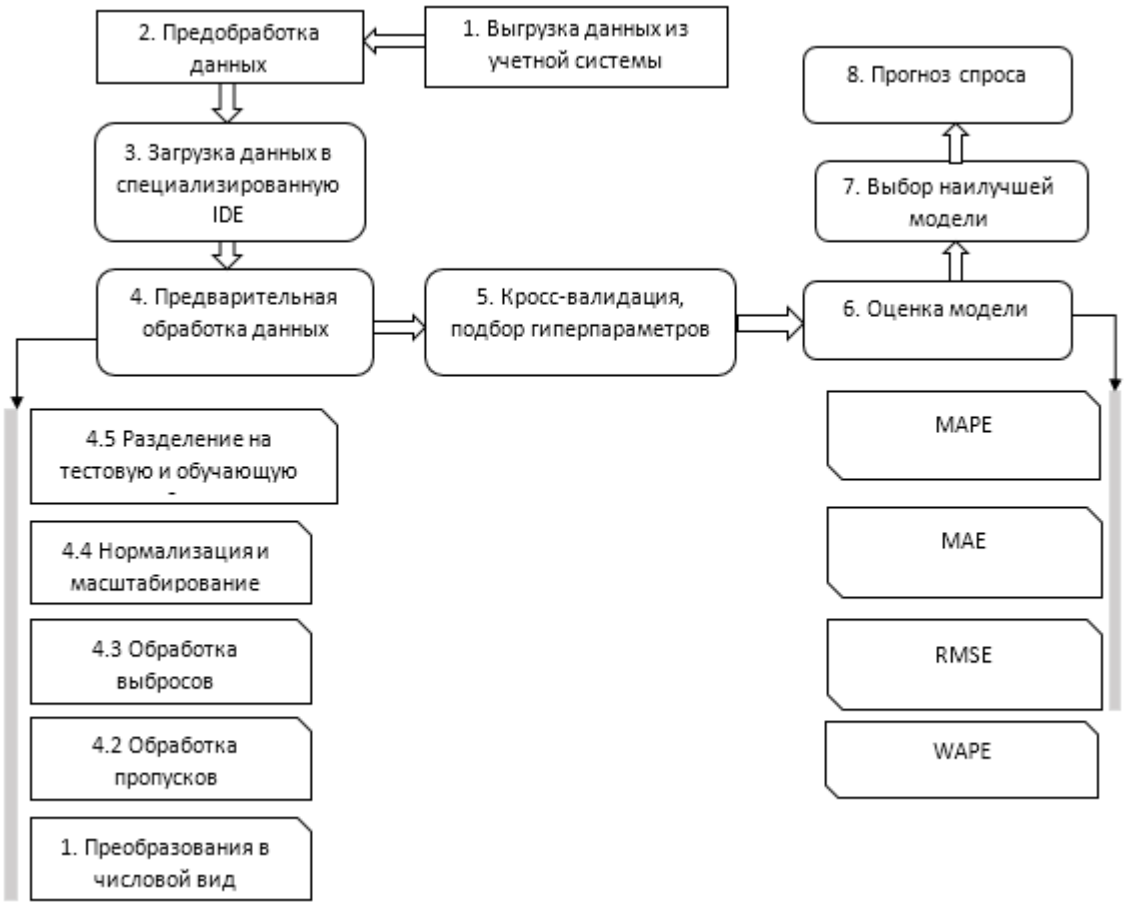

Рисунок 2 – Схема применения модели прогнозирования

На Рисунке 2 представлена общая схема построения модели прогнозирования товаров далее рассмотрим каждый блок отдельно.

Данные должны загрузиться в интегрированную среду разработки (IDE), такие, как Jupyter Notebook, VSCode.

Для оценки качества метода и его способности обобщения на новых данных есть возможность использовать различные метрики оценки эффективности такие как среднеквадратичная ошибка (MSE). Машинное обучение является итерационным процессом, который включает в себя обучение модели на обучающих данных.

Процесс предобработки данных необходим для того, чтобы свести нужный набор данных, со всеми признаками и целевой переменной в вид необходимый для максимальной точности прогноза модели. Для этого может потребоваться разные методы предобработки данных, выбор которых зависит от используемой модели машинного обучения в общем случаем можно выделить несколько групп методов. Для стабилизации дисперсии и устранения гетероскедастичности используется логарифмирование и «коробка-кокс» преобразования [36]. Это важно, когда дисперсия меняется с уровнем значения переменной, что может мешать корректному моделированию. Стабильная дисперсия улучшает результаты моделей, таких как линейная регрессия, которая предполагает постоянную дисперсию ошибок. Для удаления трендов и сезонности используется дифференцирование и сглаживание, например, скользящие средние, делая данные более пригодными для анализа и моделирования. Для масштабирования данных используют методы стандартизации и нормализации, помогающие привести данные к одному масштабу. Это означает предотвращение излишне больших значений признаков в сравнении с другими. Для моделей, чувствительных к масштабу данных, данный фактор влияет на результаты. такие алгоритмы на основе расстояния, например, k-ближайших соседей, K-средних, и градиентные методы оптимизации, например, линейная регрессия, нейронные сети.

Для удаления или замены выбросов используются следующие методы:

1. Метод процентилей – это один из способов удаления выбросов из набора данных. Он основан на использовании процентильных значений для определения и последующего удаления или замены экстремальных значений (например, 1-й и 99-й процентили);

2. Метод изоляционного леса – этот метод является алгоритмом машинного обучения, который специально разработан для обнаружения выбросов. Он строит несколько случайных деревьев, где выбросы изолируются на ранних этапах построения дерева.

Для заполнения пропусков используют несколько методов, однако, есть специфика обработки временных рядов, когда, помимо, пропущенных данных регрессоров есть пропуски даты. Их можно либо удалить, либо использовать один из методов:

1. Заполнение пропусков фиксированным значением;

2. Заполнение предыдущими или последующими значениями;

3. Использования скользящего среднего или экспоненциальное сглаживание;

4. Использование дополнительной модели машинного обучения для нахождения пропусков.

Выбор метода обработки пропусков зависит от характера данных и целей анализа. Простые методы, такие как удаление или заполнение фиксированными значениями, могут быть подходящими для небольших наборов данных с редкими пропусками. Для более сложных временных рядов, особенно с сезонностью или трендами, рекомендуется использовать интерполяцию, сглаживание или методы машинного обучения.

Кросс – валидация позволяет оценить, насколько хорошо модель будет работать на тренировочных данных, и предотвратить переобучение. Он заключается в разделении доступных данных на несколько непересекающихся подмножеств, называемых "фолдами", обучении модели на одной части данных и оценке её производительности на оставшихся данных. Этот процесс

повторяется несколько раз, пока каждый фолд не будет использован как для обучения, так и для оценки [37].

Для подбора гиперпараметров каждой модели используется инструмент библиотеки Scikit-learn - GridSearchCV [38]. Оценка модели производилась на тестовых данных, которые не использовались для расчета и обучения моделей. Оценка модели производилась с помощью метрик, описанных ниже.

Оценка модели производилась с помощью метрик, описанных ниже. WAPE (абсолютная ошибка в процентах) - это метрика, используемая для оценки точности прогнозной модели (1). Она рассчитывается как отношение суммы абсолютных ошибок к сумме реальных значений. WAPE измеряет среднюю абсолютную ошибку в процентах от общего объема продаж [39].

$$WAPE = \frac{\sum_{i=1}^{n}|y_i - \widehat{y_i}|}{\sum_{i=1}^{n} y_i}, \quad (1)$$

где $\widehat{y_i}$ – прогнозное значение;

y_i – фактическое значение;

WAPE имеет ряд преимуществ, включая чувствительность к большим отклонениям и возможность учитывать различную значимость каждого наблюдения. WAPE = 20.40% означает, что суммарные абсолютные ошибки составляют 20.40% от общего объема продаж. Это средний процент отклонения прогнозов от реальных значений. Суммарная абсолютная ошибка может быть использована для количественной оценки точности модели, но это отклонение не указывает на направление ошибки (переоценка или недооценка), а только на её величину.

MAPE, или средняя абсолютная процентная ошибка (2), является одной из метрик оценки точности прогнозов. Она вычисляется как среднее значение абсолютных процентных ошибок между фактическими и прогнозируемыми значениями. MAPE удобна тем, что она выражается в процентах, что делает ее легкой для интерпретации и сравнения между различными моделями

прогнозирования. Однако, MAPE может быть чувствительной к нулевым значениям в данных [40].

$$MAPE = 100\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|, \quad (2)$$

где n – размер выборки;

y_i – то же, что и в формуле (1);

$\widehat{y_i}$ – то же, что и в формуле (1).

Среднеквадратическая ошибка (RMSE) - это метрика оценки качества модели или алгоритма, которая измеряет среднюю величину ошибки между предсказанными и прогнозными значениями [41]. Чем меньше значение метрики, тем лучше модель предсказывает значения. RMSE рассчитывается как квадратный корень из среднего значения квадратов ошибок между предсказанными и прогнозными значениями (3). Формула для расчета RMSE выглядит следующим образом:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (3)$$

где n – размер выборки;

y_i – то же, что и в формуле (1);

$\widehat{y_i}$ – то же, что и в формуле (1).

MAE (Средняя Абсолютная Ошибка) - это метрика, которая измеряет среднюю величину ошибок в предсказаниях модели [42]. Она рассчитывается как среднее значение абсолютных разностей между предсказанными и прогнозными значениями (4).

Формула для расчета MAE выглядит следующим образом:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \quad (4)$$

где n – размер выборки;

y_i – то же, что и в формуле (1);

$\widehat{y_i}$ – то же, что и в формуле (1).

MAE является популярной метрикой, потому что она легко интерпретируется и дает представление о среднем размере ошибки. MAE менее чувствительна к экстремальным значениям (выбросам), чем RMSE, который может быть искажен несколькими крупными ошибками.

Bias (смещение) в прогнозировании — это мера того, насколько средние прогнозы отклоняются от средних реальных значений [43]. Положительное смещение указывает на систематическое завышение прогнозов, а отрицательное — на систематическое занижение. Bias рассчитывается как средняя разница между прогнозами и фактическими значениями (5).

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i), \quad (5)$$

где n – размер выборки;

y_i – то же, что и в формуле (1);

$\widehat{y_i}$ – то же, что и в формуле (1).

Bias представляет собой среднее арифметическое всех разностей между фактическими и прогнозируемыми значениями. Положительное значение метрики указывает на систематическое смещение прогнозов вверх относительно фактических данных, в то время как отрицательное значение указывает на систематическое смещение вниз.

Первая рассматриваемая модель подтип модели ARIMA. Модель является мощным инструментом для моделирования временных рядов благодаря своей гибкости и способности описывать широкий спектр их характеристик. В основе ARIMA лежат три ключевых компонента:

1. Авторегрессия (AR): Каждый элемент временного ряда линейно зависит от нескольких предыдущих значений. Компонент авторегрессии

определяется параметром, который указывает на количество прошлых значений, используемых для прогнозирования текущего значения.

2. Проинтегрированный (I): Этот компонент указывает на количество дифференцирований, необходимых для преобразования временного ряда в стационарный. Параметр d определяет степень интеграции.

3. Скользящее среднее (MA): Ошибки модели зависят от предыдущих ошибок. Параметр q указывает на количество предыдущих ошибок, которые учитываются при прогнозировании.

Модель ARIMA обозначается как ARIMA(p,d,q), где p — порядок авторегрессии, d — порядок интеграции и q — порядок скользящего среднего.

Модель SARIMA является расширением модели ARIMA и включает дополнительные параметры для учета сезонных эффектов. Эта модель описывается параметрами:

1. p, d, q – параметры несезонной части модели;
2. P, D, Q – параметры сезонной части модели;
3. S – длина сезонного периода.

Сезонные параметры учитывают повторяющиеся паттерны, характерные для временного ряда, что позволяет модели более точно отражать сезонные вариации.

Модель SARIMAX расширяет возможности SARIMA путем добавления экзогенных переменных (регрессоров), которые могут влиять на зависимую переменную временного ряда. Это позволяет модели учитывать внешние факторы, которые могут улучшить точность прогнозирования. Модель SARIMAX описывается как SARIMA с добавлением экзогенных переменных X.

Метод Auto-ARIMA представляет собой автоматизированный подход к выбору оптимальных параметров модели ARIMA на основе исторических данных временного ряда. Этот метод использует информационный критерий Акаике (AIC) для минимизации и определения наилучших параметров p, d и q. Модель полезена в ситуациях, когда структура временного ряда неизвестна

или сложна для ручного анализа. Однако его эффективность может быть ограничена при наличии сильной сезонности или высокой дисперсии данных.

В данной диссертации рассматривается модель Prophet, разработанная для прогнозирования временных рядов, особенно тех, которые содержат сложные сезонные и трендовые компоненты. Prophet является мощным и гибким инструментом для анализа временных рядов и часто используется для бизнес-аналитики и прогнозирования. Prophet моделирует временной ряд как сумму нескольких компонентов: тренд, сезонность и влияние праздников, а также случайный шум.

Модель включает в себя аддитивные компоненты, что позволяет учитывать долгосрочные изменения уровня временного ряда, повторяющиеся сезонные паттерны и эффекты отдельных событий. Тренд в модели Prophet может быть, как линейным, так и логистическим, что позволяет ей моделировать как линейные изменения, так и насыщение.

Prophet автоматически обрабатывает выбросы в данных, интерполирует пропущенные значения и предоставляет пользователям возможность интерпретировать и настраивать параметры модели, что делает ее прозрачной и гибкой в настройке. Это особенно важно для временных рядов с неполными данными или аномалиями. Модель Prophet используется для прогнозирования временных рядов в различных областях, включая бизнес-аналитику, финансы и маркетинг [44]. В данной диссертации Prophet будет применен для прогнозирования продаж, что позволит учесть сезонные, трендовые и праздничные эффекты, улучшая точность прогнозирования.

1.3 Обзор исследований в области анализа данных для прогнозирования продаж

Одной из наиболее серьезных проблем в прогнозировании спроса является работа со сложными временными рядами данных, на которые может

влиять множество факторов, таких как сезонность, недельные и месячные тренды, праздничные дни. Примером сложности применения и использования моделей для прогнозирования временного ряда, является статья К. Пивкина [45], которая описывает использования моделей: авторегрессии – скользящего среднего, модель экспоненциального сглаживания использованных для предсказания количества чеков с последующей оценкой и сравнением прогноза со средним арифметическим значение, в котором результаты моделей показывают меньший результат, чем среднее значение тестовых данных. Еще одним примером использования такого подхода приводится в статье Лясковской Е.А. [46]. Автор определяет наиболее влиятельные факторы, относящиеся к спросу с помощью построения деревьев решений, результатом исследования являлся тот факт, что наибольшее влияния на будущий спрос оказывает временный сроки выполнения прошлых заказов на рынке дорожно-строительной техники. В работе Maryam Zohdi [47] было проведено исследование данных розничного магазина о клиентах такие как: возраст, пол. Использовались различные модели: k-ближайших соседей (KNN), многослойный перцептрон (MLP), дерево решений. Автор статьи [48] акцентирует внимание пользы прогноза спроса на планирования запасов и выстраивания оптимальной цепочки поставок.

Примером использования CNN в качестве метода в современных исследованиях является работа У. Лекун [49] используют доработанную архитектуру CNN, которая включает сглаживания для уменьшения шума, используя временной ряд данных посещения сайта и волатильность валюты. Используя тренировочные данные в размере 70% (256 дней) и 80% (292 дня) и тестовые 30% (109 дней) и 20% (73 дня) от набора данных. При этом прогнозирования происходит с разными параметрами, с коэффициентом сглаживания и без. Исследователи получают результат в 9,29% MAPE при прогнозе на 73 дней и 10,44 % на 190 днях. Выводом становится, то что использования коэффициента сглаживания улучшает качество прогнозирования. Применение алгоритма машинного обучения DBN для

прогнозирования временного ряда представлено в статье Ю. Рен [50]. Для оценки результатов авторы использую три временных ряда: потребление энергии (476 месяцев), курс валюты (296 месяцев) и фондовый индекс (260 месяцев) с временными наблюдениями раз в месяц. 70 % данных используется для обучения, а 30% для тестирования. В сравнении с моделями FNN, GDBM+FNN результаты DBN были на 0,2-0,5 % лучше остальных по метрике MAPE.

2 Показатели, характеризующие объем продаж товаров и факторы, оказывающие на них влияние

2.1 Организационно-экономическая характеристика предприятия

В данной работе объектом исследования является компания «Командор». В этом параграфе представлена краткая организационная характеристика. Первый магазин ООО «Торговая сеть Командор» появился в 1999 г., компания работает на рынке розничный продаж более 25 лет. Основная деятельность компании по ОКВЭД: 47.11. Торговля розничная преимущественно пищевыми продуктами, включая напитки, и табачными изделиями в неспециализированных магазинах. Компания включает в себя 450 торговых точек, в 105 населенных пунктах 5 регионах (Красноярский край, Республика Хакасия, Республика Тува, Кемеровская область, Иркутская область). По мимо основной деятельности компания занимается розничной продажей всего ассортимента продовольственных товаров, а также оптовыми продажами и производством собственной продукции.

Таблица 1 – Общая характеристика ООО «Торговая сеть Командор»

Характеристика	Значение
1.Организационно-правовая форма	Общество с ограниченной ответственностью
2.Полное фирменное наименование Общества	Общество с ограниченной ответственностью «ТОРГОВАЯ СЕТЬ КОМАНДОР"
3.Уставный капитал	131 млн. руб.
4.Дата создания (ЕГРЮ)	18.12.1991
5.Режим налогообложения	Общий режим налогообложения
6.Вид деятельности	47.11. Торговля розничная преимущественно пищевыми продуктами, включая напитки, и табачными изделиями в неспециализированных магазинах.

Количество дополнительных видов деятельности составляет 69, количество основной деятельности – 1. Предприятие применяет общую систему налогообложения (таблица 1).

Органами управления Общества являются:

а) Совет директоров.

б) Общее собрание акционеров

Руководство текущей деятельностью Общества осуществляется Генеральным директором, избранным голосованием Советом директоров. Срок полномочий также определяется Советом директоров. Генеральный директор подотчетен Общему собранию акционеров и Совету директоров. К компетенции Генерального директора относятся все вопросы руководства текущей деятельность компании, за исключением вопросов, отнесенных к компетенции собрания акционеров, Совета директоров.

Компания подчиняется ряду нормативным документам, регулирующим деятельность общества с ограниченной ответственность, включающих в себя [51]:

1. Конституция РФ определяет основные принципы государственного устройства и закрепляет основные права и свободы граждан. Компания обязана соблюдать конституционные права своих сотрудников, клиентов и партнеров, а также выполнять свои обязанности перед государством**;**

2. Трудовой кодекс РФ регулирует отношения между работниками и работодателями. Компания должна соблюдать права и гарантии работников, оплату труда, рабочее время, отпуска, условия труда и другие требования, предусмотренные Трудовым кодексом;

3. Налоговый кодекс РФ: Определяет порядок и ставки налогообложения, а также правила ведения налогового учета. Компания обязана правильно и своевременно уплачивать налоги, а также подавать соответствующие налоговые декларации;

4. Федеральный закон «Об обществах с ограниченной ответственностью» от 08.02.1998 N 14-ФЗ. Компания должна соблюдать требования, предусмотренные законом для ООО, включая установление учредительных документов, регистрацию, учет финансовой отчетности и т.д.;

5. Федеральный закон «О государственной регистрации юридических лиц и индивидуальных предпринимателей» от 28.08.2001 № 129-ФЗ (ред. от 31.12.2017); Устанавливает правила и процедуры государственной регистрации юридических лиц. Компания должна правильно произвести регистрацию, чтобы иметь законный статус юридического лица**.**

6. Федеральный закон «О лицензировании отдельных видов деятельности» от 04.05.2011 N 99-ФЗ (ред. от 31.12.2017). Определяет виды деятельности, требующие специальной лицензии. Если компания занимается лицензируемой деятельностью, она обязана получить соответствующую лицензию и соблюдать условия её действия;

7. Федеральный закон «О несостоятельности (банкротстве)» от 26.10.2002 № 127-ФЗ (ред. от 23.04.2018). Определяет виды деятельности, требующие специальной лицензии. Если компания занимается лицензируемой деятельностью, она обязана получить соответствующую лицензию и соблюдать условия её действия.

Организационная структура холдинга представлена на рисунке А.1 и А.2 «Организационная структура» в приложении А. Она является линейно-функциональной и делиться на подразделения по функциональным признакам (финансы, управление персоналом, маркетинг, недвижимость, логистика, служба контроля, юридический отдел). Компания является холдингом.

Холдинг – это компания, которая владеет контрольными пакетами акций других компаний (дочерних компаний), но не занимается непосредственным производством товаров или оказанием услуг. Головной компанией является – «Командор-Холдинг» осуществляющая руководящую, финансовую, маркетинговую функции. Дочерними компаниями являются: «Торговая сеть Командор», «ДМ Трейдинг», «Ит-Левел». «Торговая сеть Командор» является

компанией непосредственно занимающаяся организацией розничных продаж, «ДМ Трейдинг» компания выполняет функцию организации помещений и аренды необходимых площадей. «ИТ-Левел» исполняет функции по обеспечению холдинга программными продуктам, их функционированию, внедрению, обновлением и поддержкой.

Преимущества холдинговой структуры:

1. Холдинг позволяет диверсифицировать риски, владение компаниями в разных отраслях помогает снизить общий риск;

2. Холдинговая структура дает возможность перераспределять ресурсы между дочерними компаниями для обеспечения их роста и стабильности;

3. Стратегическое управление и контроль на уровне головной компании позволяют эффективно координировать деятельность дочерних компаний Организационная структура холдинга позволяет сочетать автономность дочерних компаний с централизованным стратегическим управлением.

Дирекция по финансам и экономике в компании имеет решающее значение для управления финансовыми аспектами бизнеса и обеспечения финансовой стабильности. Её функциями является финансовый анализ и планирование, бухгалтерская и финансовая отчетность, управление инвестициями и анализ рисков.

Дирекция по информационным технологиям выполняет ряд критически важных функций, связанных с разработкой, внедрением и поддержкой информационных технологий, которые поддерживают и оптимизируют бизнес-процессы компании.

Дирекция по управлению персоналом в крупной компании розничного ритейла занимается рядом задач, связанных с управлением персоналом компании. В основе этой дирекции лежит стратегическое планирование и реализация политики компании в области управления персоналом.

Служба контроля в компании занимается всем, что связано с обеспечением безопасности и защитой имущества компании. К их обязанностям входит найм и обучение охранников, установка систем видеонаблюдения и других систем безопасности, а также организация перевозки денежных средств и их защита.

Юридический отдел в компании играет критическую роль в обеспечении законности, соответствия нормативным требованиям и защите прав и интересов компании. Его главная функция – предоставление правовой поддержки и консультаций всем отделам и руководству компании.

Дирекция по логистике выполняет ключевые функции, направленные на эффективное управление цепочками поставок, оптимизацию транспортировки товаров и обеспечение бесперебойного снабжения. Вот основные направления её деятельности: управление цепочкой поставок, контроль запасов, организация перевозок, обработка заказов, координация с другими отделами.

Данная исследовательская работа сделана в функциональных рамках отдела информационного сопровождения поставок. Данный отдел занимается координацией и управлением всеми информационными потоками, связанными с процессом поставок. Специалисты отдела занимаются сбором, обработкой и анализом данных о движении товаров от поставщиков до складов и магазинов, что позволяет своевременно выявлять и устранять возможные проблемы. Они поддерживают связь с поставщиками и транспортными компаниями, обеспечивая корректность и актуальность информации о поставках.

Конкретные функции данного исследования исполняет менеджер сектора анализа и планирования выполняет важные функции, связанные с прогнозированием и планированием логистических операций. Он анализирует исторические данные о продажах и поставках, чтобы определить тенденции и разработать точные прогнозы спроса. Эти прогнозы помогают компании лучше планировать объемы закупок, избегая как излишков, так и дефицита товаров. Менеджер также оценивает эффективность текущих планов и

стратегий, предлагая изменения и улучшения для оптимизации процессов. Одной из ключевых задач менеджера является разработка долгосрочных и краткосрочных планов поставок, которые учитывают сезонные колебания спроса, маркетинговые акции и другие факторы.

Магазины компании подразделяются на три формата: суперстор, дискаунтер, фрешмаркет. Суперстор – это крупный магазин, занимающий большую площадь и предлагающий широкий ассортимент товаров. В супермаркете можно найти не только продукты питания, но и бытовую технику, электронику, одежду и товары для дома. Супермаркеты обычно предлагают конкурентоспособные цены за счет объемов продаж и высокой оборачиваемости товара. Дискаунтер также известный как магазин у дома, характеризуется меньшей торговой площадью и ограниченным ассортиментом товаров повседневного спроса, таких как продукты питания, напитки и бытовая химия. Цены в дискаунтерах обычно ниже за счет экономии на декоре, обслуживании и расположении в менее престижных районах. Эти магазины удобно расположены в жилых районах, что позволяет покупателям быстро приобрести необходимые товары. Фрешмаркет, также называемый супермаркетом, занимает среднюю площадь по сравнению с супермаркетами и дискаунтерами. В супермаркете можно найти широкий ассортимент продуктов питания, включая свежие фрукты, овощи, мясо, рыбу, молочные продукты, а также готовую еду и бакалею. Эти магазины часто располагаются в районах с высокой плотностью населения и могут быть частью торговых центров или находиться в центральных районах города. Цены в фрешмаркетах могут быть выше, чем в дискаунтерах, но ниже, чем в специализированных магазинах, при этом акцент делается на свежесть и качество продукции.

Среднесписочная численность персонала «ТС Командор» за 2023 год составила – 7257 человек.

Динамика показателей капитала, внеоборотных активов и общей величины активов изображена на Рисунке 3, источником данных является сайт аудита [52].

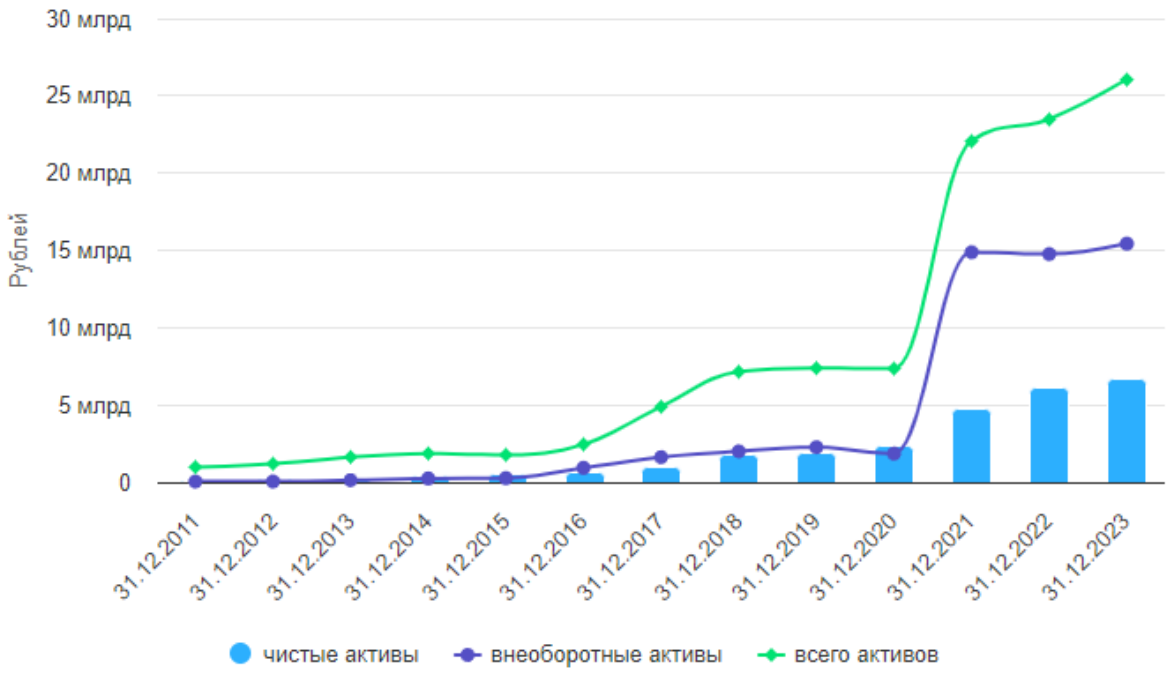

Рисунок 3 – Динамика активов компании за период 2011-2023 гг.

Анализ динамики показателей капитала, внеоборотных активов и общей величины активов на Рисунке 3 показывает значительные изменения за период с 2017 по 2023 годы. До 2020 года внеоборотные активы не превышали 1,9 млрд, что свидетельствует о стабильном уровне инвестиций в долгосрочные активы. Однако в 2021 году наблюдается резкий скачок внеоборотных активов до 14,9 млрд руб. Этот скачок может указывать на крупные инвестиции в недвижимость, оборудование и другие долгосрочные активы, связанные с расширением производства. На конец 2023 года величина внеоборотных активов оставалась на уровне около 15,5 млрд. руб., что демонстрирует устойчивость и стабильность в управлении долгосрочными активами. Это может свидетельствовать о высоком уровне доверия инвесторов и хорошей стратегической позиции компании на рынке.

Чистые активы компании также показали положительную динамику. До 2020 года их величина колебалась в пределах 1-2 млрд. руб., что было сопоставимо с уровнем внеоборотных активов. Однако после 2020 года чистые активы значительно увеличились и достигли 6,6 млрд. руб. к концу 2023 года.

Увеличение чистых активов может быть связано с ростом прибыли, эффективным управлением затратами и увеличением капитала компании. Рост чистых активов также свидетельствует о повышении финансовой устойчивости и ликвидности компании. Это означает, что компания способна покрывать свои обязательства и имеет значительные резервные средства для дальнейшего развития.

Основные показатели рентабельности, а также показатель EBIT (прибыль до вычета налогов и процентов к уплате), за последние годы можно проследить на Рисунке 4, источником данных является сайт аудита [52].

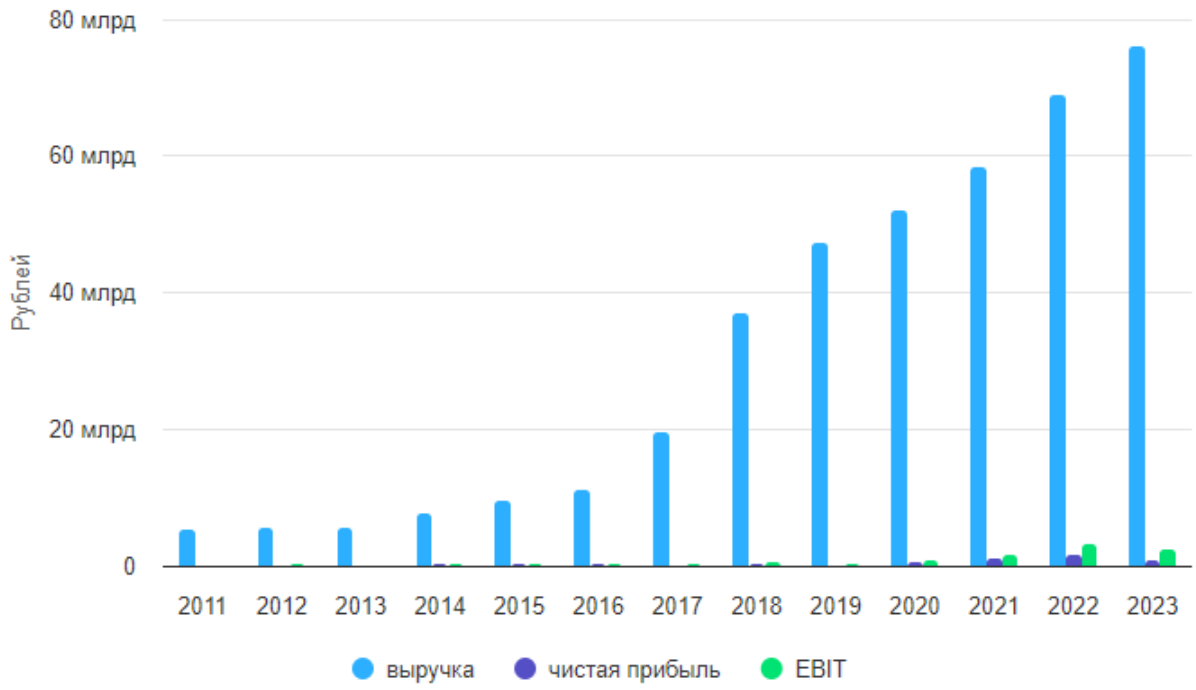

Рисунок 4 –Динамика финансовых результатов за период 2011-2023 гг.

Анализируя Рисунок 4, где отражается информация о выручки компании с 2011 по 2023 годы, можно отметить значительный рост показателя за период с 2016 г. по 2023г. с 11,1 млрд руб. до 76 млрд руб. на 548%. Такой рост обусловлен различными факторами, включая расширение ассортимента продукции, открытия новых магазинов, улучшение качества обслуживания клиентов и эффективные маркетинговые кампании, за данные период было

открыто более 200 магазинов, компания расширила свою деятельность увеличив количество регионов присутствия. Рост выручки до 76 млрд руб. к 2023 году свидетельствует о высокой эффективности операционной деятельности компании, результатом успешного внедрения цифровых технологий и улучшения логистических процессов, а также высокий уровень адаптивности компании к изменяющимся условиям рынка и способности использовать возможности для роста. Для более полного понимания динамики развития компании и выявления факторов, способствующих дальнейшему увеличению выручки, необходимо детально рассмотреть спрос на продукцию в различных сегментах. Анализ спроса позволит выявить предпочтения потребителей, определить наиболее востребованные товарные категории.

2.2 Анализ показателей, характеризующих динамику продаж компании «Командор»

Как было сказано ранее компания представляет собой многоструктурный механизм, где главные функции разделены на структурные дирекции и отделы. Суть деятельности в перепродажи товаров с наценкой, в ходе данного процесса необходимо контролировать логистические процессы, иметь площадь для хранения и продажи, необходимо планировать денежные потоки, место хранения, перевозку и спрос для постоянного процесса продаж, выбирая надежных поставщиков. Планирование продаж занимает немаловажную часть успеха деятельности, так как продовольственные товары имеют возможность портиться и в большинстве своей не могут продолжительное время хранится на складах, даже с учетом камер заморозки. Управление запасами является одним из ключевых бизнес-процессов в розничном ритейле. Для успешной работы магазина необходимо иметь правильный баланс между запасами и спросом. Слишком высокие запасы могут привести к увеличению затрат на хранение товаров, а слишком низкие

На Рисунке 5 представлено распределение прибыли полученной с продаж групп товаров за 2023. На рисунке видно, что большинство значений прибыли сконцентрированы в диапазоне от 0 до 2 %. Этот факт указывает на то, что основная масса товаров приносит относительно небольшой процент от общей прибыли. Значение превышающие 2 % это группы товары, которые приносят значительно больше прибыли по сравнению с другими. Анализируя процентное распределение прибыли, можно выделить категории товаров, которые приносят наибольший вклад в общую прибыль компании: хлеб собственного производства – 7,41 %, сладости из сахара и шоколада – 5,43 %, фрукты – 4,72 %. Производство хлеба в собственных условиях позволяет контролировать качество и себестоимость продукции, что способствует высокой маржинальности, также хлеб является основным продуктом в рационе многих людей и обладает стабильным спросом. Сладости, такие как конфеты и шоколад, обычно имеют высокий уровень спроса, особенно в периоды праздников и подарочных сезонов. Спрос на фрукты может варьироваться в зависимости от сезона и урожайности, но даже вне сезона они остаются востребованными. Общая прибыль от каждой категории может быть обусловлена не только спросом, но и факторами, такими как стоимость производства, конкуренция на рынке, маркетинговые стратегии и локального рынка. Исследование прогнозирования продаж для этих товарных групп также может включать анализ внешних факторов, таких как сезонные колебания, конкурентная среда, экономические тенденции и изменения в потребительском поведении. Важно также учитывать, что необходимо регулярно обновлять модели прогнозирования, чтобы учитывать новые данные и изменения на рынке, обеспечивая точные и актуальные прогнозы продаж. Прогнозирование продаж для ключевых товарных групп является важной составляющей стратегического планирования и позволяет компаниям эффективно выстраивать свои бизнес-процессы и стратегии роста. Анализируя прогнозы, компания может также решить, на какие новые товарные категории стоит сосредоточить внимание для дальнейшего расширения.

года, что указывает на их постоянную популярность. Фрукты постоянного ассортимента также приносят стабильную прибыль, с пиками в зимние месяцы. Необходимо выделить факторы, влияющие на количество продаж для каждого наблюдения.

Как уже было сказано в первой главе использование современных технологий, таких как большие данные и аналитика, позволяет значительно улучшить точность прогнозов и эффективность управления спросом. Анализ данных о покупках с использованием машинного обучения позволяет выявлять скрытые паттерны и прогнозировать изменения в спросе с высокой точностью. Одним из главных преимуществ машинного обучения является его адаптивность и способность улучшать точность прогнозов по мере накопления новых данных. Одним из ключевых этапов при использовании машинного обучения для прогнозирования спроса является подготовка данных.

2.3 Выбор и анализ факторов, влияющих на продажи

Одним из первых шагов является выбор подходящих источников данных. Обычно используются исторические данные о продажах, данные о маркетинговых акциях, сезонные факторы и экономические показатели. Эти данные необходимо интегрировать с данными продаж для каждого наблюдения и подготовить для дальнейшего анализа.

Исторические данные о продажах предоставляют информацию о прошлых тенденциях и объемах продаж, что позволяет выявить сезонные колебания и долгосрочные тренды. Данные о маркетинговых акциях, такие как рекламные кампании, скидки и промоакции, помогают понять, как различные маркетинговые усилия влияют на объемы продаж. Сезонные факторы, включая праздники и изменения погоды, также играют важную роль, так как определенные товары могут быть более востребованы в разные времена года. Экономические показатели, такие как уровень инфляции, безработица и

покупательная способность, влияют на общую экономическую ситуацию и потребительское поведение.

Выбор товара "бананы" для прогнозирования продаж обоснован рядом причин, связанных с его характеристиками и рыночным поведением. Бананы являются одним из самых популярных и часто покупаемых фруктов по всему миру. Их стабильный спрос делает их идеальным объектом для анализа и прогнозирования. Постоянная популярность бананов обусловлена их доступностью, полезными свойствами и универсальностью в кулинарии, что способствует регулярному и предсказуемому потреблению. Кроме того, бананы, как один из наиболее продаваемых товаров в категории фруктов, имеют обширные и детализированные исторические данные о продажах. Хотя бананы доступны круглый год, их продажи могут подвержены сезонным колебаниям и влиянию трендов. Для этого необходимо посмотреть график продаж. На Рисунке 8 использованы данные продаж десяти магазинов временного ряда 2023 г., использовались продажи бананов, измеряемых в килограммах.

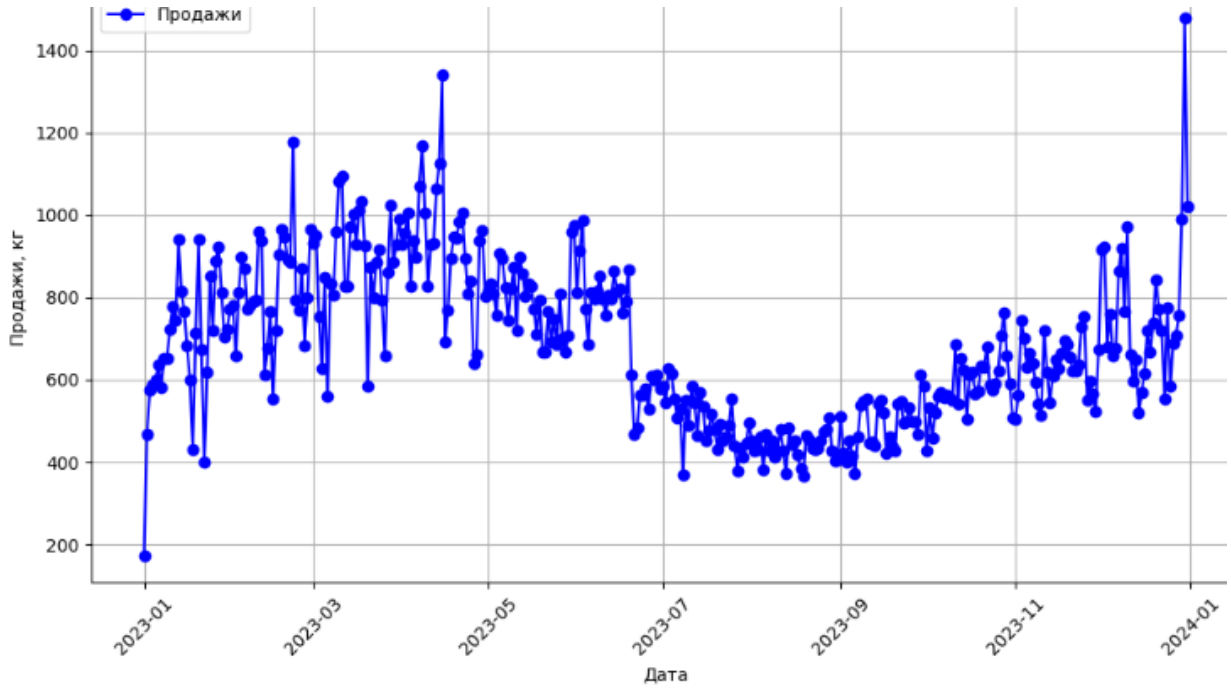

Рисунок 8 – Продажа бананов за 2023 г., кг

Исторические данные продаж также могут включать информацию о ценовые изменения, сезонные колебания и маркетинговые акции. Анализ этих факторов позволяет выявить сезонные паттерны и тренды, что улучшает точность прогнозов. Из данных Рисунка 8 можно видеть повышенные продажи в начале года, и постепенной снижения объемов с апреля до конца августа. Для более детального исследования квартальной сезонности воспользуемся коррелограммой. Автокорреляция помогает исследовать структуру временного ряда и выявлять наличие корреляций между значениями ряда на различных временных отсчетах (лагах). Это важно для понимания того, как текущие значения могут зависеть от предыдущих значений, что может указывать на наличие сезонности, цикличности или трендов в данных. Для анализа автокорреляции временного ряда с 365 наблюдениями были выбраны лаги до 180, что соответствует практическому правилу использования лагов, не превышающих половину длины ряда. Это позволяет сохранить достаточное количество точек данных для надежного анализа

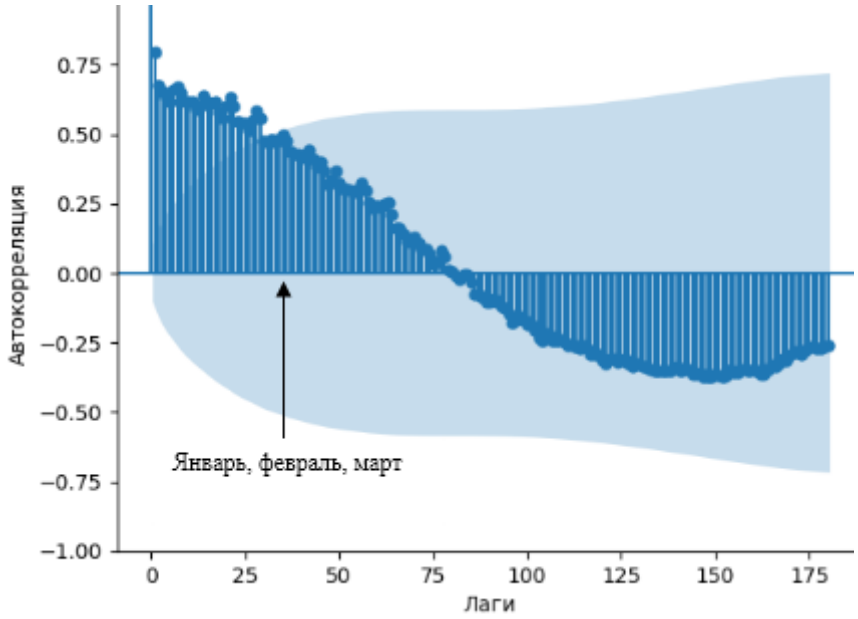

Рисунок 7 – Функция автокорреляции с лагом 180

График автокорреляции данных на Рисунке 7 указывает на сильную зависимость текущих значений временного ряда от предыдущих значений.

Это говорит о том, что данные имеют сильную краткосрочную зависимость. Снижение автокорреляции от 0,75 до 0 на 85 лаге указывает на то, что с увеличением лага зависимость между текущими и прошлыми значениями ослабевает. Значение 0,25 на 60 лаге все еще свидетельствует о наличии заметной зависимости, хотя и слабее, чем на коротких лагах. Постепенное снижение автокорреляции может указывать на наличие долгосрочного тренда в данных. Долгосрочные тренды могут привести к высокой автокорреляции на коротких интервалах, которая затем постепенно уменьшается по мере увеличения лага. Это может свидетельствовать о том, что значения временного ряда повторяются с определенной периодичностью, хотя и с ослаблением зависимости на больших интервалах. Стационарные ряды часто показывают быстрое снижение автокорреляции до нуля. В данном случае снижение происходит медленнее, что может указывать на нестационарность ряда. Отрицательная автокорреляция на 85 лаге указывать на сезонные эффекты, где пики в данных наблюдаются каждые три месяца, а провалы - в промежутках. Для более точной интерпретации и моделирования временного ряда следует учитывать эти особенности. Следующим этапом поиска сезонности являются анализ данных представленных на Рисунке 10 – распределение продаж по дням неделям. Продолжая анализ, важно отметить, что сезонность может проявляться не только на уровне годовых циклов, но и внутри недели. Например, если распределение продаж по дням недели показывает явные пики в определенные дни и спады в другие, это указывает на недельную сезонность. Рисунок 10 позволяет нам визуализировать эти тенденции и лучше понять, как продажи изменяются в течение недели. Если в начале недели продажи ниже, а в выходные наблюдаются пики, это типичный пример недельной сезонности. Чтобы подтвердить и количественно оценить эту сезонность, можно использовать методы спектрального анализа или применить сезонную декомпозицию временного ряда.

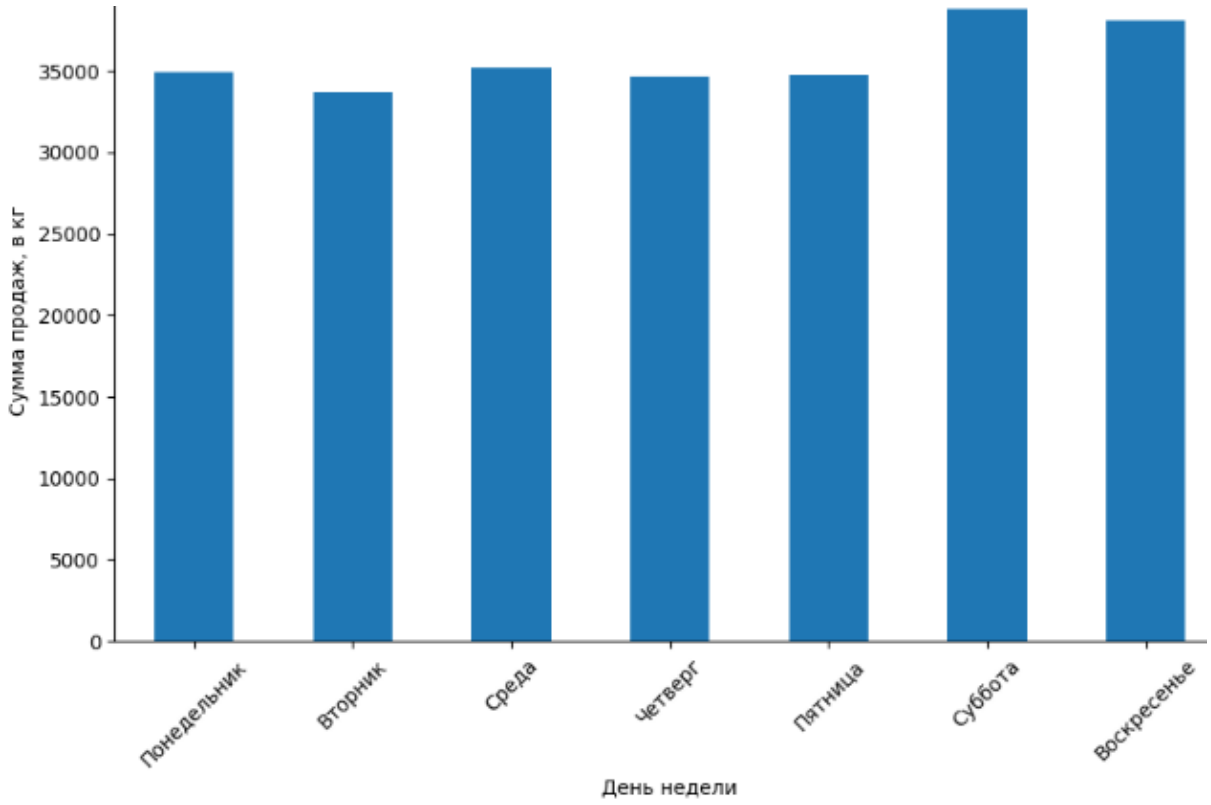

Рисунок 10 – Продажи бананов по дням неделям за 2023 г. 10 магазинов, кг

Гистограмма на Рисунке 10 показывает суммарные продажи в килограммах для каждого дня недели. В воскресенье и субботу происходят самые высокие продажи среди всех дней недели. Это может быть связано с тем, что большинство людей совершает покупки в выходные дни, когда у них больше свободного времени. Понедельник имеет наименьшие объемы продаж, что может быть связано с тем, что люди менее склонны к покупкам после выходных. Данная гистограмма демонстрирует присутствия недельной сезонности. Выраженная цикличность продаж по дням недели свидетельствует о поведенческих особенностях потребителей. Таким образом, гистограмма на Рисунке 10 наглядно демонстрирует наличие недельной цикличности в данных о продажах, что имеет важное значение для принятия стратегических и операционных решений в управлении розничной компанией.

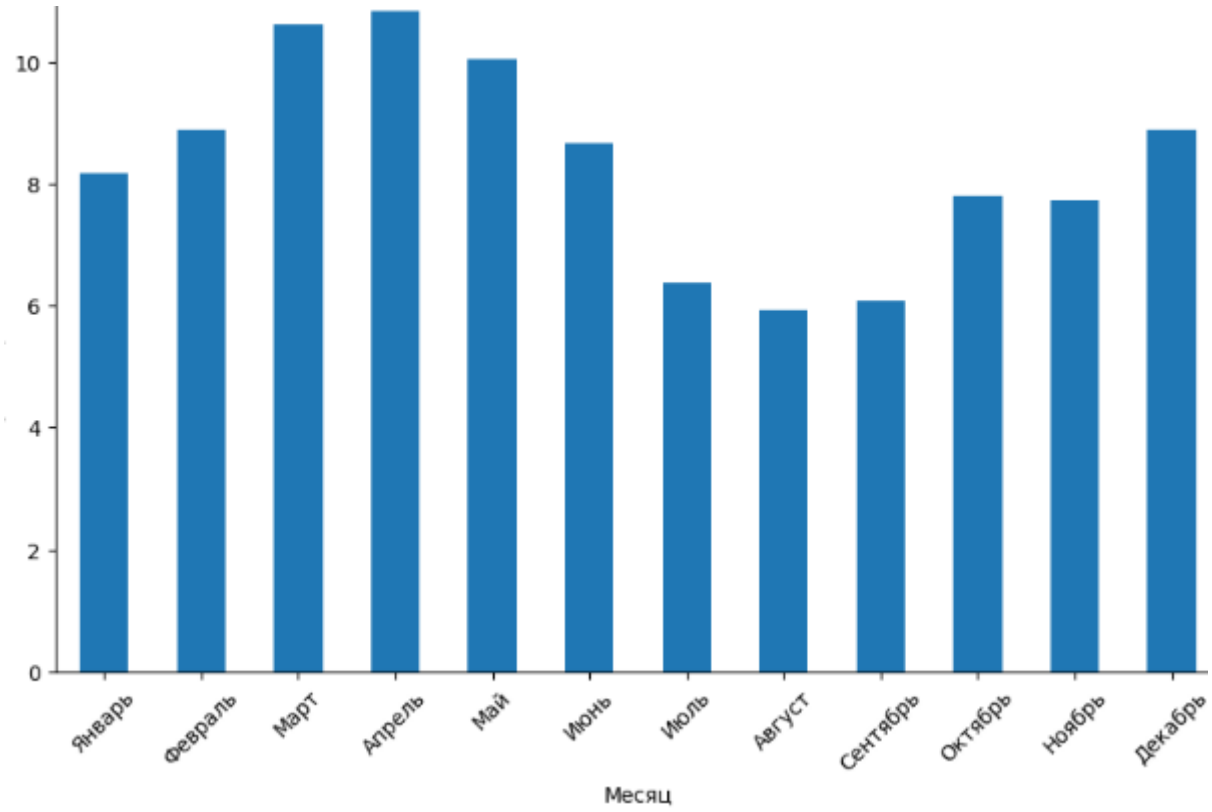

Рисунок 11 – Продажи бананов по месяцам за 2023 г., в % от года.

Гистограмма на Рисунке 11 показывает суммарные продажи в % для каждого месяца. Видно, что максимальные продажи приходятся на апрель затем продажи снижаются ежемесячно до минимума в августе и возрастают помесячно до декабря. Это означает присутствие сезонности продаж, данный признак необходимо будет внести в наборы данных для построения модели. Включение сезонных данных в модели прогнозирования помогает компании лучше распределять ресурсы и планировать рабочую силу. В пиковые месяцы могут потребоваться дополнительные сотрудники, увеличенные запасы товаров и усиленная поддержка клиентов. Важно также учитывать внешние факторы, такие как погодные условия, экономические изменения и другие события, которые могут влиять на сезонные колебания. Например, необычно жаркое или холодное лето может значительно изменить привычные тренды продаж.

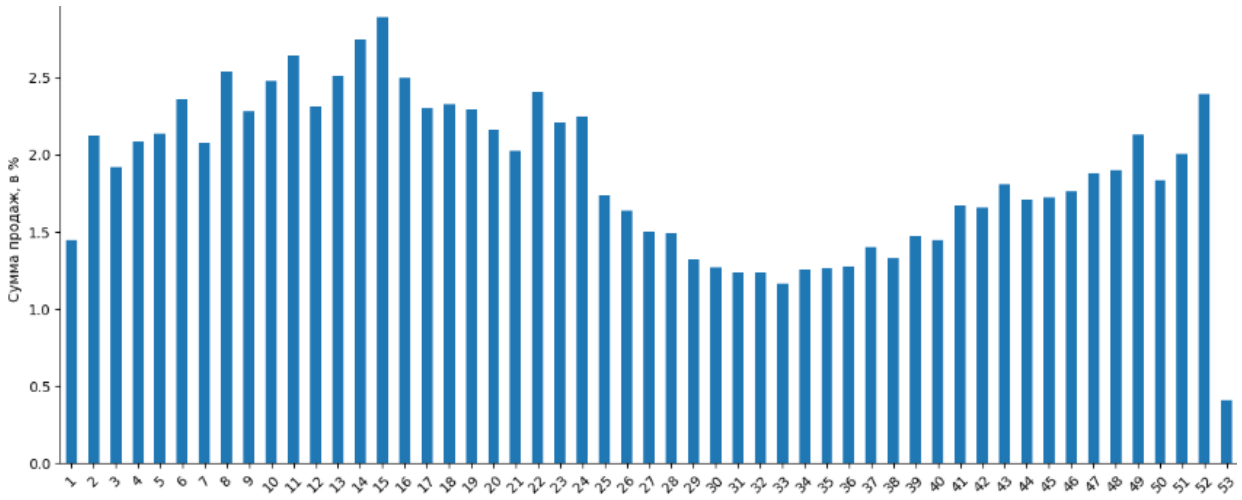

Рисунок 12 – Распределение продаж по неделям 2023 г., %

Информация о сезонности также можно рассмотреть на Рисунке 12. Минимальный процент продаж приходится на 32, 33, 34 неделю года – август

Метки праздников могут быть значимым признаком в анализе продаж по нескольким причинам. Во-первых, праздники часто связаны с увеличением потребительского спроса на определенные товары или услуги. Например, праздничные распродажи или специальные предложения могут стимулировать потребителей к покупкам. Во-вторых, выходные и праздничные дни могут влиять на торговую активность в розничных точках, так как многие потребители имеют свободное время для посещения магазинов. Третье важное соображение заключается в том, что праздники часто влияют на психологическое состояние потребителей, стимулируя их к потреблению благ и услуг. Например, в периоды праздников люди могут быть более склонны к щедрым тратам на подарки или украшения, что приводит к повышению продаж в определенные периоды года.

Кроме того, праздники часто сопровождаются специальными мероприятиями и массовыми мероприятиями, такими как парады, фестивали или ярмарки, которые также могут стимулировать торговлю и повышать объемы продаж. Например, местные праздники или события могут привлечь большое количество посетителей, что может стать дополнительным фактором роста продаж в близлежащих магазинах и заведениях. С другой стороны,

необходимо учитывать, что влияние праздников на продажи может различаться в зависимости от отрасли и типа товаров или услуг. Например, продажи продуктов питания могут значительно возрастать перед праздниками, в то время как продажи электроники или бытовой техники могут быть более стабильны или даже снижаться в эти периоды. Таким образом, использование меток праздников как признаков в анализе данных позволяет учитывать сезонные факторы и изменения в потребительском поведении, что может существенно повысить точность и прогностическую способность моделей. В Таблице 2 перечислены все отмеченные признаки, используемые при обучении моделей.

Таблица 2 – Признаки модели прогнозирования

№	Используемый фактор построения модели
1	9 мая
2	8 марта
3	1 мая
4	23 февраля
5	Предновогодние дни
6	Выходные дни после нового года
7	Пасха
8	Масленица
9	Квартал
10	Время года
11	Дни недели
12	№ недели в году
13	День года
14	Месяц

Факторы, перечисленные в Таблице 2 для построения модели, представляют собой временные и сезонные переменные, которые могут значительно влиять на анализ и прогнозирование различных продуктов, таких

как бананы или лимоны. Они включают в себя как ежегодные праздники (например, 9 мая, 8 марта, 1 мая, 23 февраля, Пасха, Масленица), так и временные рамки (предновогодние дни, выходные дни после нового года), а также структурные характеристики времени (квартал, время года, дни недели, номер недели в году, день года и месяц). Их учет в модели позволяет улавливать сезонные и временные изменения, оптимизировать прогнозирование, основанные на цикличности и изменчивости потребительского рынка в разные периоды года.

3 Модель прогнозирования продаж товара на основе машинного обучения

3.1 Реализация модели прогнозирования продаж

В данной главе будут подробно рассмотрены все этапы создания модели, начиная с предварительной подготовки данных и заканчивая оценкой точности прогноза. Прежде чем приступить к непосредственной реализации модели, необходимо провести анализ исходных данных.

В конце прошлой главы были обозначены выбранные факторы влияния (x), следующим этап включает в себя сбор и предварительную обработку данных, которые будут использоваться для обучения модели. Важно убедиться, что данные не содержат пропусков, выбросов и ошибок формата входных данных, которые могут негативно повлиять на качество прогноза.

В ходе использования различных инструментов и технологий для построения модели машинного обучения с целью прогнозирования продаж было принято решение воспользоваться языком программирования Python. Этот выбор основывается на ряде ключевых факторов, которые делают Python наиболее подходящим для реализации задач машинного обучения, являясь высокоуровневым языком программирования общего назначения, известным своей динамической строгой типизацией и автоматическим управлением памятью. Эти характеристики значительно упрощают процесс разработки, повышают производительность и обеспечивают высокую читаемость и качество кода. Python способствует переносимости программ, что делает его идеальным для разработки масштабируемых и поддерживаемых моделей машинного обучения. Практически все современные модели машинного обучения поддерживаются на нем. Это обеспечивает легкость интеграции и адаптации моделей в различных средах и платформах. Это позволяет быстро

адаптировать и масштабировать решения в зависимости от требований бизнеса, обеспечивая при этом высокую производительность и надежность.

Python обладает обширной экосистемой библиотек, которые активно используются в машинном обучении:

1. NumPy и Pandas библиотеки для работы с массивами данных и анализом данных;
2. Scikit-learn популярная библиотека для машинного обучения, предоставляющая широкий набор алгоритмов и инструментов для построения и оценки моделей;
3. Matplotlib и Seaborn: библиотеки для визуализации данных, что важно для анализа и интерпретации результатов моделей.
4. SciPy: библиотека для научных и инженерных вычислений, дополняющая функционал NumPy.
5. Statsmodels: библиотека для статистического анализа данных, предоставляющая модели и инструменты для работы с временными рядами.

Помимо библиотек для работы с данными и построения моделей, Python имеет мощные средства для визуализации и анализа результатов. Это позволяет не только строить точные модели, но и эффективно интерпретировать их выводы, что важно для принятия обоснованных решений на основе данных. Визуализация помогает выявлять скрытые закономерности и тренды, а также представлять результаты работы моделей в понятной и наглядной форме для бизнеса.

Следующий этап заключается в разделении данных на тренировочную и тестовую выборки. Это позволяет оценить качество модели и избежать проблемы переобучения. Тренировочная выборка используется для обучения модели, а тестовая – для проверки ее точности на новых данных. В данном исследовании будет использоваться три набора данных представленных на рисунке 1. Тренировочные данные взяты за период с 2022.01.01 по 2024.03.23 и включают в себя 813 дней. Тестовые данные с 2022.03.24 по 2024.04.20 – 28 дней.

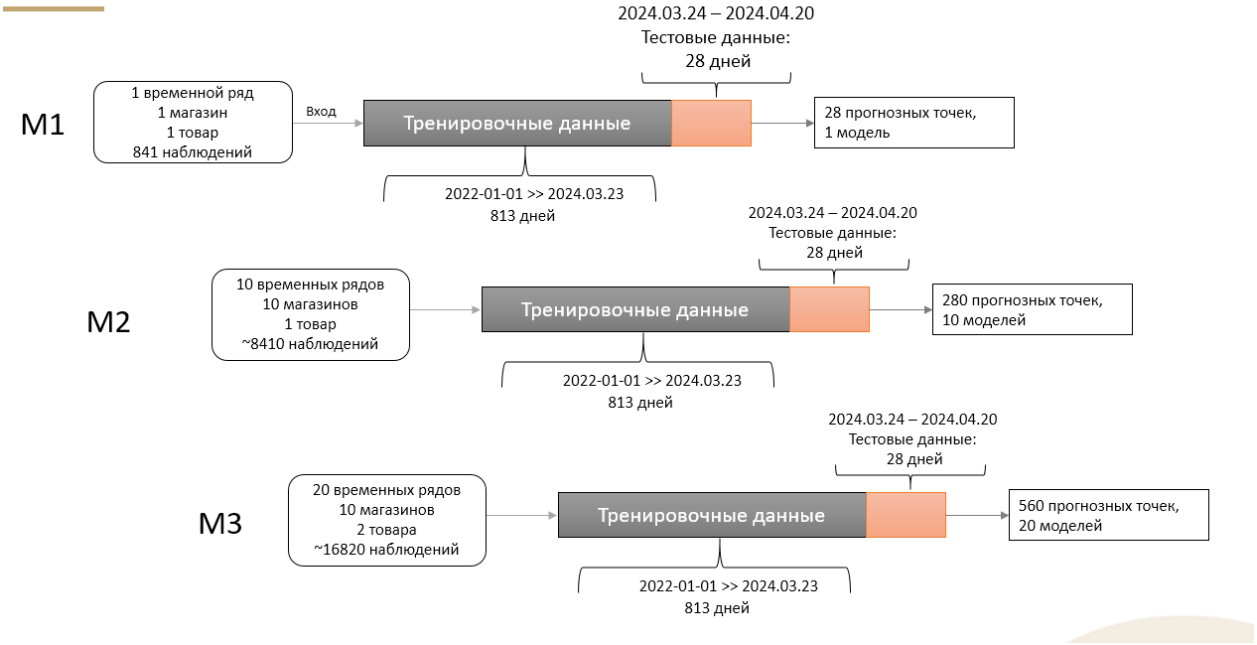

Рисунок 13 – Наборы данных для обучения моделей

Рисунок 13 отражает три набора данных**.** Каждый из наборов отличается количеством временных рядов. Первый набор «M1» имеет данные 1-го временного ряда, включает данные о продажах 1-го товара (бананы) и 1-го магазина, состоит из 841 наблюдения. Второй набор данных «M2» состоят из десяти наборов временных рядов, 8410 наблюдений, которые включают данные о 10 магазинах и одном продаваемом товаре (бананы). Третий набор данных имеет 20 временных рядов, 16820 наблюдений, которые включат данные о продажах двух товаров (бананы, лимон) в десяти разных магазинах. В качестве исходных данных были использованы данные учетных систем в формате .xlsx (Microsoft Excel).

В Таблице 3 представлены характеристики первого набора данных.

Большой диапазон значений в Таблице 3 от 20,7 до 230,4 указывает на разнообразие наблюдений. Это может означать, что данные охватывают широкий спектр случаев или измерений. Дисперсия и стандартное отклонение показывают, что данные имеют значительный разброс относительно среднего значения. Среднее значение 52,8 с достаточно высоким стандартным

отклонением (22,4) предполагает, что, хотя данные в среднем сосредоточены около 52,8, существует значительное отклонение от этого среднего, что может быть вызвано значительными колебаниями или изменчивостью в наблюдениях. Высокий эксцесс (3,8) указывает на наличие выбросов или экстремальных значений, которые могут значительно отклоняться от среднего значения. Это следует учитывать при анализе данных, так как наличие экстремальных значений может искажать анализ и приводить к неверным выводам, если не принять это во внимание.

Таблица 3 – Описательные характеристики целевой переменной

Показатель	**Значение**
Минимальное значение	20,7
Максимальное значение	230,4
Мат. ожидание:	52,8
Дисперсия:	505,1
Стандартное отклонение	22,4
Эксцесс	3,8

Следующим этапом станет предобработка данных, которая будет состоять из следующих этапов:

1. Обработка пропусков;
2. Обработка выбросов;
3. Нормализация и масштабирование данных;
4. Логарифмирование;
5. Дифференцирование;

После завершения подготовки данных следует этап выбора и применения методов машинного обучения. Существует множество алгоритмов, которые можно использовать для прогнозирования продаж, включая линейную регрессию, решающие деревья, метод ближайших соседей и нейронные сети. Выбор алгоритма зависит от особенностей данных и целей исследования.

Линейная регрессия позволяет выявить зависимость между продажами и одним или несколькими факторами, что делает её подходящей для простых моделей с линейными связями. Решающие деревья полезны для создания более сложных моделей, способных учитывать нелинейные зависимости и взаимодействия между переменными. Метод ближайших соседей, основанный на сходстве наблюдений, позволяет делать прогнозы на основе близких по характеристикам примеров из исторических данных. Нейронные сети, обладая способностью обучаться сложным паттернам и нелинейным зависимостям, особенно эффективны для анализа больших объемов данных с множеством факторов. Каждый из этих методов имеет свои преимущества и ограничения, поэтому важно провести экспериментальное сравнение их эффективности на реальных данных компании. На основе результатов этого сравнения можно выбрать оптимальный метод или комбинировать несколько подходов для достижения эффективных результатов в прогнозировании продаж.

В данной диссертации модель SARIMAX и метод Auto-ARIMA будет протестирована для прогнозирования временных рядов продаж. Использование экзогенных переменных позволит учесть внешние факторы

Также рассматривается модель Prophet, разработанная для прогнозирования временных рядов, особенно тех, которые содержат сложные сезонные и трендовые компоненты. Prophet является мощным и гибким инструментом для анализа временных рядов и часто используется для бизнес-аналитики и прогнозирования.

Использование Prophet в данной работе обусловлено его способностью моделировать сложные временные ряды с учетом различных факторов, что особенно важно в условиях динамичных и изменяющихся данных. Применение Prophet включает создание модели, обучение на исторических данных, прогнозирования точек.

Одним из аспектов реализации модели является выбор метрик для оценки точности прогноза. Используемые метрики оценки: средняя

абсолютная ошибка (MAE), средняя абсолютная процентная ошибка (MAPE), взвешенная абсолютная процентная ошибка (WAPE), Среднеквадратическая ошибка (RMSE), смещение (Bias) указанные в первой главе. Эти метрики вместе дают всестороннюю оценку точности прогнозной модели, позволяя понять ее поведение с разных сторон и улучшить качество прогнозирования.

Использования модели SARIMAX при 28 дней прогноза дали следующие результаты, указанные на Рисунке 14.

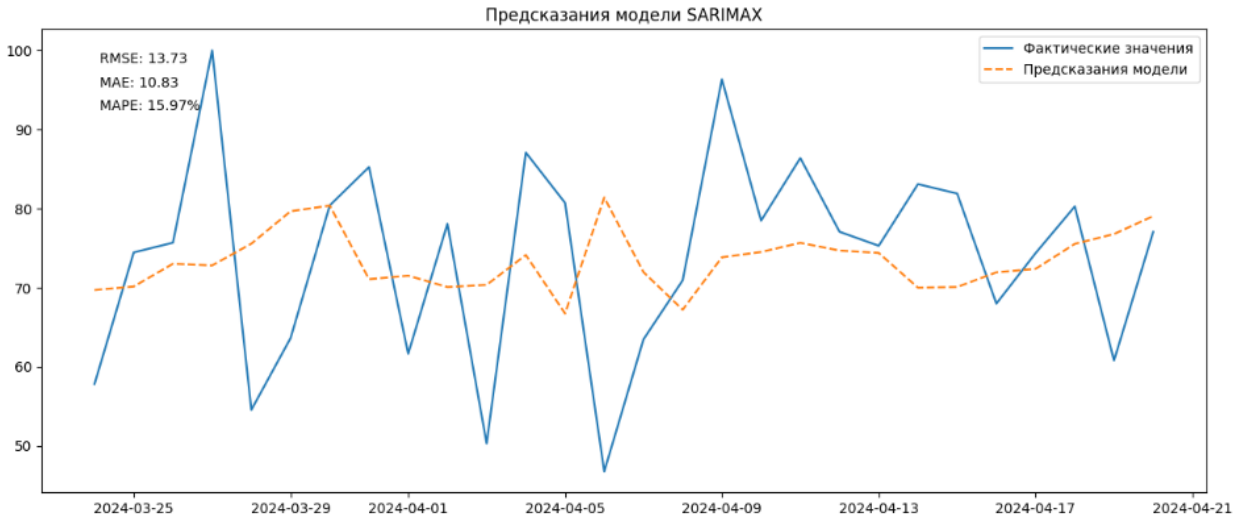

Рисунок 14 – Прогнозные значения модели SARIMAX на данных M1

SARIMAX позволяет моделировать сложные временные ряды с учетом как обычной динамики, так и сезонности, а также с учетом внешних регрессоров. Использованное интегрирование порядка 1, указывает на однократное дифференцирование временного ряда для обеспечения стационарности. Параметр сезонной составляющей определился 12 (годичной сезонностью). Для построения модели использовались регрессоры: цена, номер дня недели, номер недели в году. Коэффициент признака «Цена» составил -0,1693, что означает отрицательную связь между ценой и продажами, коэффициент «Номер дня недели» имеет значение равное 1.9501 указывает на положительную связь между окончанием недели и продажами.

Далее были использованы еще 3 модели прогнозирования: Auto-ARIMA с возможность автоматического подбора гиперпараметров стремящаяся минимизировать критерий Акаике, значение которого может указывает на то, что если к расчету модели прибавить свободный коэффициент, то ее прогнозные значения улучшатся. Модель экспоненциальное сглаживание Холта Уинтерса использует экспоненциальное сглаживание для обновления и прогнозирования трех компонентов (тренда, уровня и сезонности). Оно учитывает веса для последних значений временного ряда, придавая больший вес более поздним данным. Однако в данном исследовании исходя из метрик, эта модель показала наихудший результат.

Рисунок 15 – Прогноз модели Prophet набора данных M1 с доверительным интервалом

Рисунок 15 представляет прогноз временного ряда с помощью модели Prophet. На графике отображается прогнозируемое значение вместе с доверительным интервалом, который отражает уровень неопределенности в прогнозе. Это позволяет оценить надежность прогноза риски, связанные с предсказанными значениями временного ряда. Чтобы оценить является ли данная модель наилучшей необходимо сравнить метрики моделей.

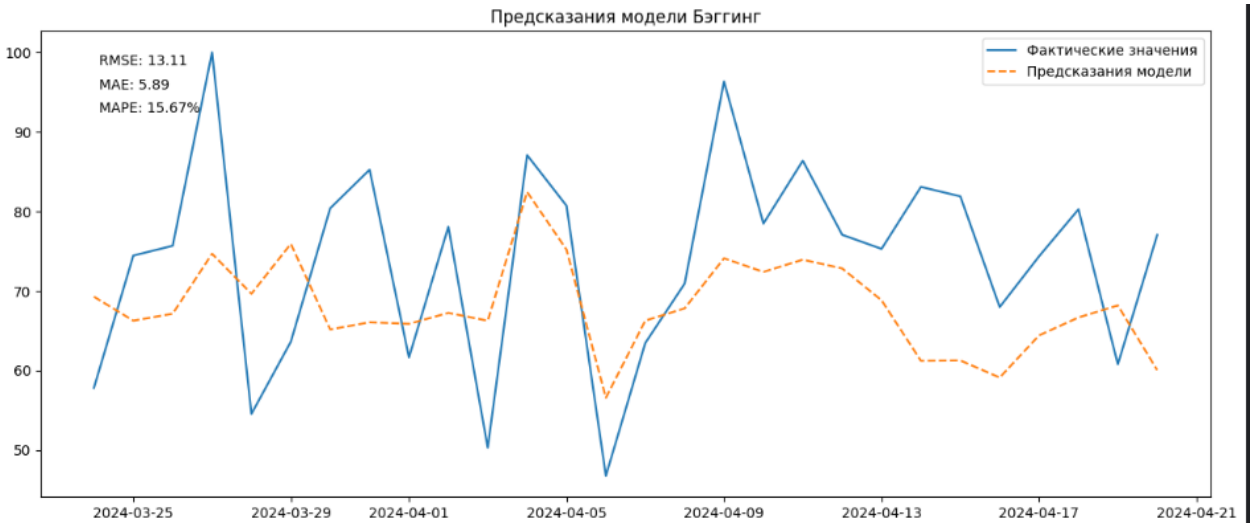

Рисунок 16 – Прогноз тестовых данных модели Бэггинг набора данных M1

На рисунке 16 представлены метод бэггинга (Bootstrap Aggregating), который является одной из ключевых техник ансамблевого обучения, направленной на улучшение точности и устойчивости моделей машинного обучения. Бэггинг базируется на использовании бутстреп-выборок и агрегации результатов нескольких моделей для уменьшения дисперсии и предотвращения переобучения, более подробно работа модели описано в параграфе 1.1.

Агрегированные результаты моделей для набора данных M1 представлены в Таблице 4.

Таблица 4 – Оценка моделей на данных M1

Метрика	**Lightgbm (градиентный бустинг)**	**SARIMAX (с учетом регрессоров)**	**Auto-ARIMA**	**Экспоненциальное сглаживание Холта Уинтерса**	**Prophet(с учетом регрессоров)**
RMSE	13.67	13.8	12.7	21,3	13,9
MAE	10.71	10.6	9.9	17,2	11,35
MAPE, %	16,14	16	14.9	24	15,9

Результаты различных моделей прогнозирования продаж, представленных в Таблице 4 демонстрируют различия в их точности и

пригодности для задачи. Модель Lightgbm (градиентный бустинг) показывает хороший баланс между ошибками, с RMSE равным 13.67 и MAE равным 10.71, а MAPE составляет 16.14%. Это указывает на хорошую точность модели, особенно при прогнозировании данных с сезонностью и трендами. Тем не менее, Auto-ARIMA демонстрирует наилучшие результаты среди всех рассмотренных моделей, с наименьшими значениями RMSE и MAE (12.7 и 9.9 соответственно), а также с MAPE равным 14.9%. Эти показатели свидетельствуют о высокой точности модели и ее способности адекватно учитывать временные зависимости и сезонные колебания.

SARIMAX, учитывающая регрессоры, показывает RMSE 13.8 и MAE 10.6, что сопоставимо с Lightgbm, однако она уступает Auto-ARIMA по точности прогнозирования. MAPE для SARIMAX составляет 16%, что также находится в пределах допустимого, но не лучшего показателя. Это говорит о том, что хотя SARIMAX хорошо справляется с задачей, она не является лучшей моделью в данном наборе.

Экспоненциальное сглаживание Холта-Уинтерса показывает наихудшие результаты с RMSE 21.3, MAE 17.2 и MAPE 24%. Высокие значения ошибок указывают на значительные отклонения прогноза от фактических данных, что делает эту модель наименее предпочтительной для данной задачи. Это может быть связано с тем, что данная модель менее способна учитывать сложные временные зависимости и регрессоры по сравнению с другими моделями.

Модель Prophet, учитывающая регрессоры, демонстрирует результаты, схожие с Lightgbm и SARIMAX, с RMSE 13.9, MAE 11.35 и MAPE 15.9%. Эти результаты показывают, что Prophet достаточно эффективен в прогнозировании временных рядов с учетом сезонности и внешних регрессоров, однако все же уступает по точности модели Auto-ARIMA. В заключение, анализ результатов показывает, что Auto-ARIMA является наиболее точной моделью для прогнозирования в данном случае, демонстрируя наименьшие значения ошибок. Lightgbm и SARIMAX также показывают хорошие результаты и могут быть полезны в зависимости от

специфических требований и условий задачи. Экспоненциальное сглаживание Холта-Уинтерса, напротив, демонстрирует низкую точность и не рекомендуется для использования в данном контексте. Prophet также представляет собой жизнеспособный вариант, особенно при необходимости учета дополнительных регрессоров, однако уступает Auto-ARIMA по ключевым метрикам точности.

Использование модели Lightgbm (градиентного бустинга) более применимо для работы с множественными временными рядами, особенно когда необходимо справляться с выбросами и пропусками в данных. Одним из ключевых преимуществ является способность модели эффективно обрабатывать большие и сложные наборы данных, включая те, которые содержат неоднородности и пропуски. Благодаря своей архитектуре, Lightgbm позволяет выделять значимые паттерны и взаимосвязи в данных, даже при наличии аномалий. Кроме того, Lightgbm предоставляет мощные механизмы обработки выбросов, что делает его подходящим для использования в условиях, где данные могут содержать экстремальные значения. Модель автоматически регулирует вес этих выбросов, минимизируя их негативное влияние на качество прогнозов. Это особенно важно для временных рядов, где аномалии могут сильно исказить результаты традиционных моделей. Lightgbm также поддерживает параллельные вычисления и эффективное использование памяти, что делает его пригодным для масштабирования на большие наборы данных. Это позволяет использовать модель для анализа множественных временных рядов одновременно, обеспечивая точные и устойчивые прогнозы для каждого из них. При расширении модели на данные с множественными временными рядами, модель демонстрирует высокую гибкость и адаптивность, позволяя легко интегрировать дополнительные признаки и регрессоры.

Таким образом, Lightgbm является предпочтительным выбором для задачи прогнозирования следующего набора данных M2, связанных с анализом и прогнозированием множества временных рядов, особенно в

условиях наличия выбросов и пропусков данных. Его высокая точность, устойчивость к аномалиям и способность к масштабированию делают его незаменимым инструментом для получения надежных прогнозов в сложных и динамичных условиях.

3.2 Масштабирование модели прогнозирования на уровне торговой сети

Набор данных M2 отличается от M1 количеством временных рядов. Их увеличение связано с количеством магазинов в данных равное 10 в отличие от одного магазина в прошлом наборе данных. Для разработки модели прогнозирования M2 использовался «словарь» python, который оказался эффективным инструментом для работы с множественными временными рядами. Словарь позволил организовать данные таким образом, что каждый магазин имел свой собственный временной ряд, что упростило доступ к данным и их обработку, пример указан на Рисунке 17.

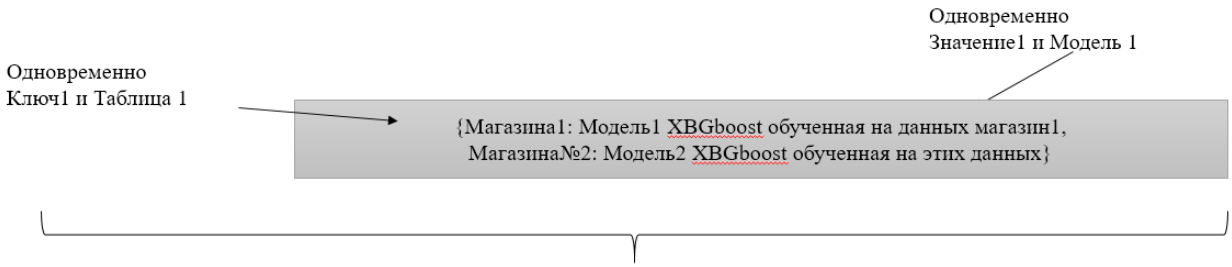

Рисунок 17 – Применение словаря для хранения данных о продажах

Словарь в Python работает следующим образом: словарь – это пара ключ-значение, где ключ – это ссылка на значение, все элементы словаря являются объектами, которыми может быть что угодно. В нашем случае использование словаря подразумевает хранение в значение таблицы данных о продажах каждого магазина отдельно и соответствующие им обученные модели в форме «Таблица1: Модель1, Таблица2: Модель2». Использование

словаря способствовало улучшению качества прогнозов, так как каждый временной ряд можно было анализировать и обрабатывать отдельно.. Кроме того, словарь облегчил интеграцию модели Lightgbm, так как данные были четко структурированы и легко доступны для обучения и тестирования. При разработке модели прогнозирования M2 было также учтено влияние сезонности и трендов, которые могут различаться для разных магазинов. Словарь Python позволил легко применять различные модели и методы обработки данных к каждому отдельному временному ряду, что сделало модель более гибкой и адаптивной. Это особенно важно для масштабируемости решения, так как с увеличением числа магазинов модель можно легко адаптировать без значительных изменений в коде. Также это упростило процесс валидации и тестирования модели, так как данные каждого магазина можно было обрабатывать и анализировать независимо. В целом, применение словаря Python для работы с множественными временными рядами в наборе данных M2 оказалось успешным решением, которое позволило значительно улучшить качество прогнозов и упростить процесс разработки модели.

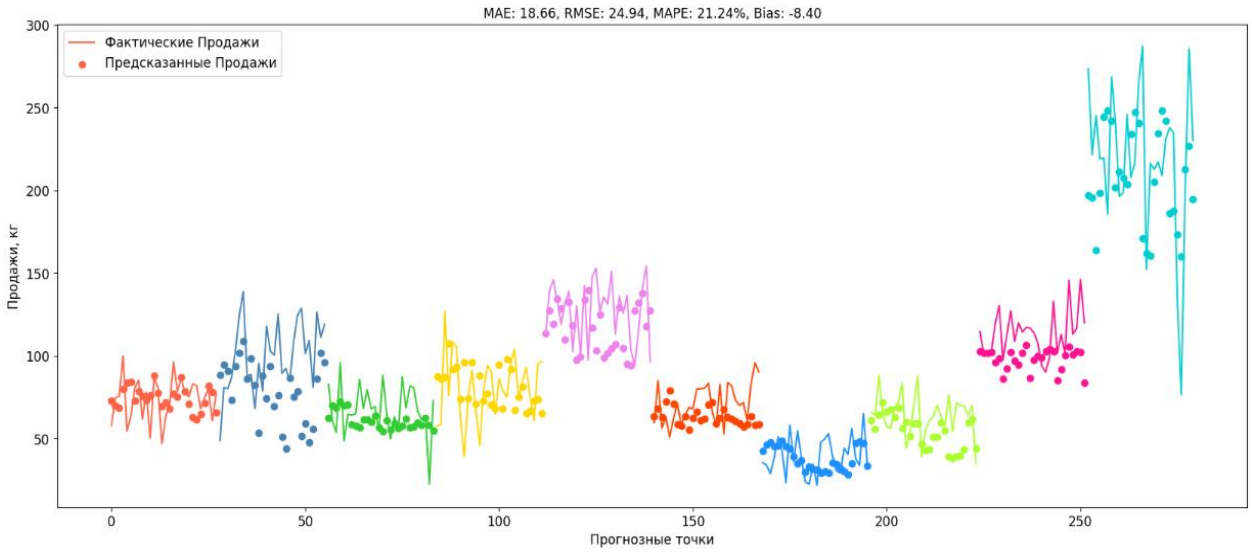

Рисунок 18 – Результаты прогнозирования набора данных M2 моделью Lightgbm

На Рисунке 18 указаны результаты прогнозирования данных M2. MAE в размере 18,66 означает, что модель в среднем ошибается на 18,66 единиц в своих предсказаниях. Это мера абсолютной ошибки. RMSE равный 24,94 означает, что среднеквадратичное отклонение модели от фактических данных составляет 24,94 единицы. MAPE в размере 21,24% показывает среднюю процентную ошибку модели относительно фактических данных. Она показывает, насколько процентов модель ошибается в своих предсказаниях. Bias (смещение) равное -8,4 указывает на то, что в среднем предсказанные значения модели ниже фактических на 8,4 единицы.

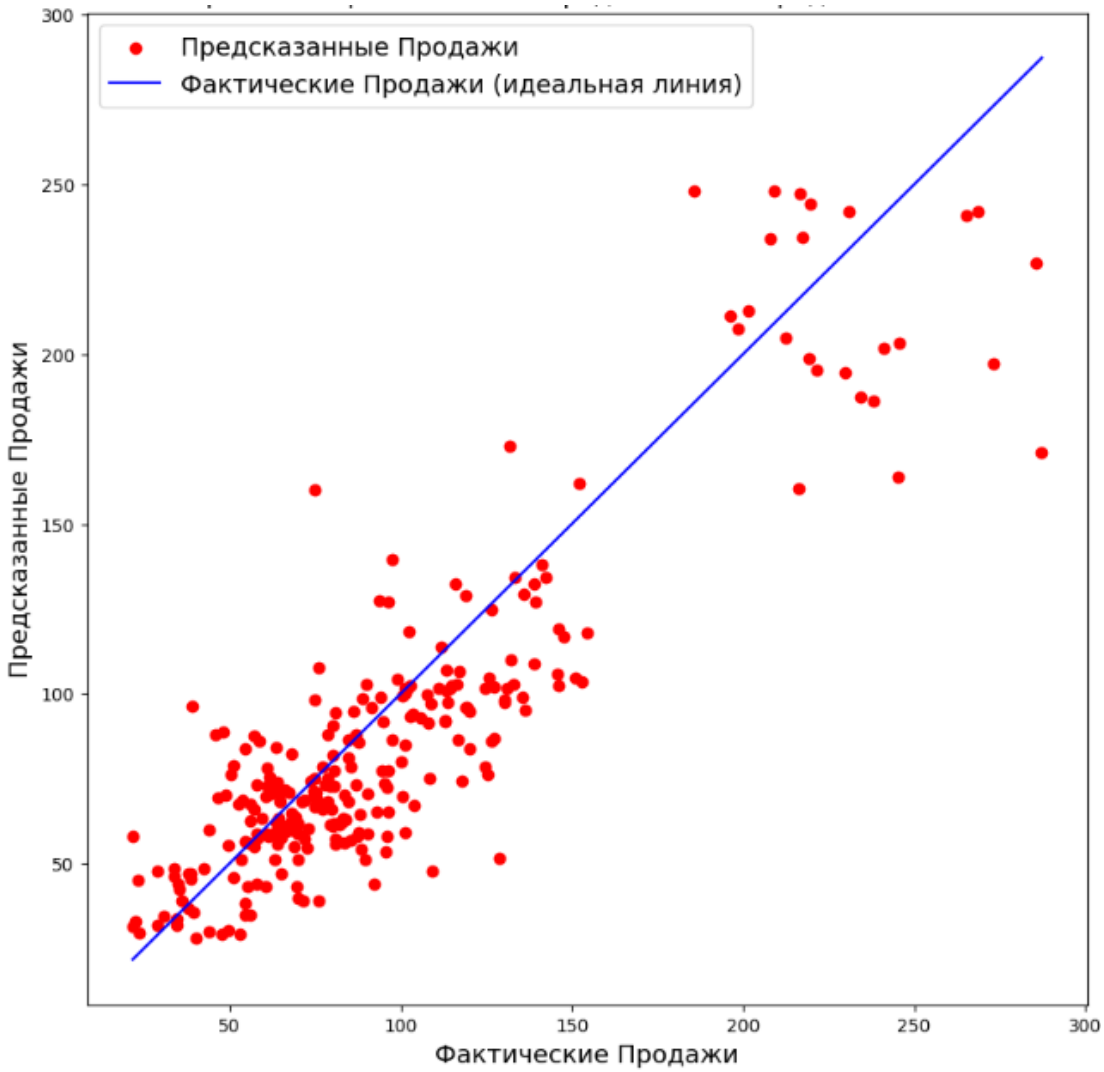

Рисунок 19 – Сравнение прогнозных и фактических точек модели Lightgbm набора данных M2

Разброс точек на Рисунке 19 отражает неточность прогнозных точек в сравнении с фактическим выраженных в значениях метрик, указанных выше. Однако мы также можем посмотреть результаты данного прогноза для каждого магазина в Таблице 5.

Таблица 5 представляет собой результаты оценки прогнозных моделей для продаж бананов в различных магазинах. Метрики MAE (средняя абсолютная ошибка), WAPE (взвешенная абсолютная процентная ошибка) и Bias (смещение) используются для оценки точности прогнозов моделей. Магазин № 11 показал наилучшую точность с MAE 10.71 и WAPE 14.49%, что может указывать на небольшие отклонения прогнозов от фактических данных. Однако наблюдается небольшое положительное смещение (Bias 0.552), что может указывать на недооценку прогнозов по сравнению с реальными продажами. Магазин № 44 и № 77 показали наихудшие результаты среди всех магазинов.

Таблица 5 – Оценка прогноза набора данных M2

Товар	Магазин	MAE	WAPE, %	Bias
Бананы	Магазин № 1	10,71	14,49	0,552
Бананы	Магазин № 2	18,28	22,63	-0,06
Бананы	Магазин № 3	20,72	16,32	-9,88
Бананы	Магазин № 4	39,04	17,84	-11,6
Бананы	Магазин № 5	13,71	19,18	-8,20
Бананы	Магазин № 6	17,05	15,02	-15,3
Бананы	Магазин № 7	28,49	28,44	-21,3
Бананы	Магазин № 8	13,48	20,02	-6,61
Бананы	Магазин № 9	14,23	22,21	-9,84
Бананы	Магазин № 10	11,00	27,90	-1,52

В магазине № 44 высокое значение MAE (39.04) и WAPE (17.84%) свидетельствуют о значительных ошибках в прогнозах, что может быть вызвано сложностью моделирования или специфическими характеристиками продаж в этом магазине. Отрицательное смещение (Bias -11.6) указывает на систематическую переоценку прогнозов. Магазин № 77 также показал

высокие значения MAE (28.49) и WAPE (28.44%), что указывает на значительные ошибки в прогнозировании продаж. Отрицательное смещение (Bias -21.3) также говорит о том, что модель склонна к переоценке прогнозов. Магазины № 33 и № 66 имеют средние значения ошибок и смещения. Ошибка WAPE для магазина № 66 составляет 15.02%, что отражает умеренное согласование между прогнозами и фактическими данными. Общий анализ показывает, что точность прогнозов значительно варьируется от магазина к магазину. Это может быть связано с различными факторами, включая уникальные характеристики каждого магазина, разные потребительские предпочтения. Для дальнейшего улучшения точности прогнозов может потребоваться более тщательная настройка моделей для каждого магазина

Набор данных M3 представляет собой расширенную версию предыдущих наборов M1 и M2, включающую информацию о продажах из 10 магазинов и для двух различных товаров – бананы и лимоны, что формирует в сумме 20 временных рядов. Этот набор данных стал возможным благодаря расширению и углублению аналитики по продажам, включая большее количество переменных и сценариев. Каждый из 10 магазинов в наборе данных M3 представлен отдельным временным рядом для каждого из двух товаров. Это означает, что для каждого магазина существует по два временных ряда, отражающих продажи каждого из товаров. Такая структура данных позволяет анализировать влияние различных факторов на продажи в разных магазинах и для разных товаров. Каждый временной ряд в наборе M3 содержит информацию о продажах товаров в разные временные точки, что делает возможным анализ трендов, сезонных колебаний и влияния маркетинговых мероприятий на продажи. Учет двух различных товаров дает возможность сравнивать и анализировать их производительность в различных условиях и магазинах.

Для разработки моделей прогнозирования на базе набора данных M3 может использоваться подход, аналогичный примененному для M2, с использованием словаря Python. Каждый магазин и каждый товар могут

рассматриваться как отдельные ключи в словаре, связанные с соответствующими временными рядами продаж. Это позволяет удобно управлять и анализировать большой объем данных, обеспечивая точные и надежные прогнозы. Расширение набора данных до M3 открывает новые возможности для масштабирования в масштабах торговой розничной сети с тысячами номенклатур в различных магазинах и для различных товарных групп.

Так как модели с параметрами остались прежними, то графики по продажам бананов и их оценка осталась прежней. Для прогноза продажи лимонов представлен Рисунок 20.

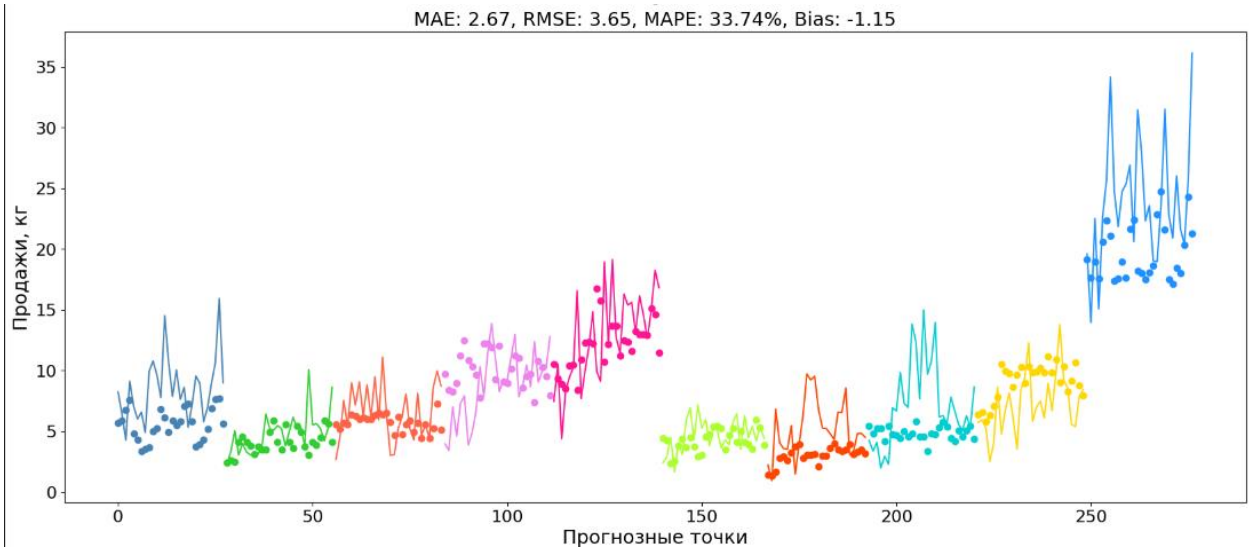

Рисунок 20 – Результаты прогнозирования набора данных M3 продажи 2-го товара моделью Lightgbm

Значение MAE составляет 2.67. Это означает, что средняя абсолютная ошибка модели составляет приблизительно 2.67 кг. Чем ниже значение MAE, тем лучше модель справляется с прогнозированием. RMSE равен 3.65. RMSE является стандартным отклонением ошибок модели и в данном случае составляет примерно 3.65 кг. По сравнению с MAE, RMSE учитывает большие ошибки с большим весом, что может быть полезно в случае наличия значительных выбросов в данных. MAPE составляет 33.74%. Этот показатель

показывает среднюю абсолютную процентную ошибку модели. В данном случае, в среднем модель ошибается на 33.74% при прогнозировании продаж. Смещение (bias) равно 1.15. Это означает, что среднее смещение между предсказанными и фактическими значениями составляет 1.15 кг. MAPE в 33.74% указывает на наличие значительного процентного отклонения в предсказаниях, что может требовать дополнительной настройки модели.

Таблица 6 – Оценка прогноза набора данных M3 (лимоны)

Магазин	**MAE**	**WAPE**	**Bias**
Магазин № 1	10,71	14,48	0,552
Магазин № 2	28,46	28,41	-21,3
Магазин № 3	13,45	19,96	-6,62
Магазин № 4	18,26	22,60	-0,05
Магазин № 5	20,71	16,31	-9,88
Магазин № 6	13,70	19,17	-8,19
Магазин № 7	10,99	27,88	-1,51
Магазин № 8	14,23	22,20	-9,84
Магазин № 9	17,05	15,01	-15,3
Магазин № 10	38,98	17,81	-11,6

Из представленного анализа можно сделать общие выводы об оценки модели. Средняя абсолютная ошибка (MAE) и процентная ошибка (WAPE) значительно различаются в зависимости от магазина, что указывает на разную точность модели для разных участников. Например, магазины № 2 и № 10 показывают высокие значения MAE и WAPE, что свидетельствует о необходимости дополнительной оптимизации модели для улучшения точности предсказаний. Смещение (bias) также имеет значительные колебания между магазинами. Например, магазин № 2 имеет значительное отрицательное смещение, что может свидетельствовать о систематической недооценке предсказаний моделью. Третий общий вывод заключается в необходимости индивидуального подхода к каждому магазину при

построении модели прогнозирования. Учет специфических особенностей и факторов, влияющих на продажи в каждом магазине, может помочь улучшить точность предсказаний и снизить ошибки.

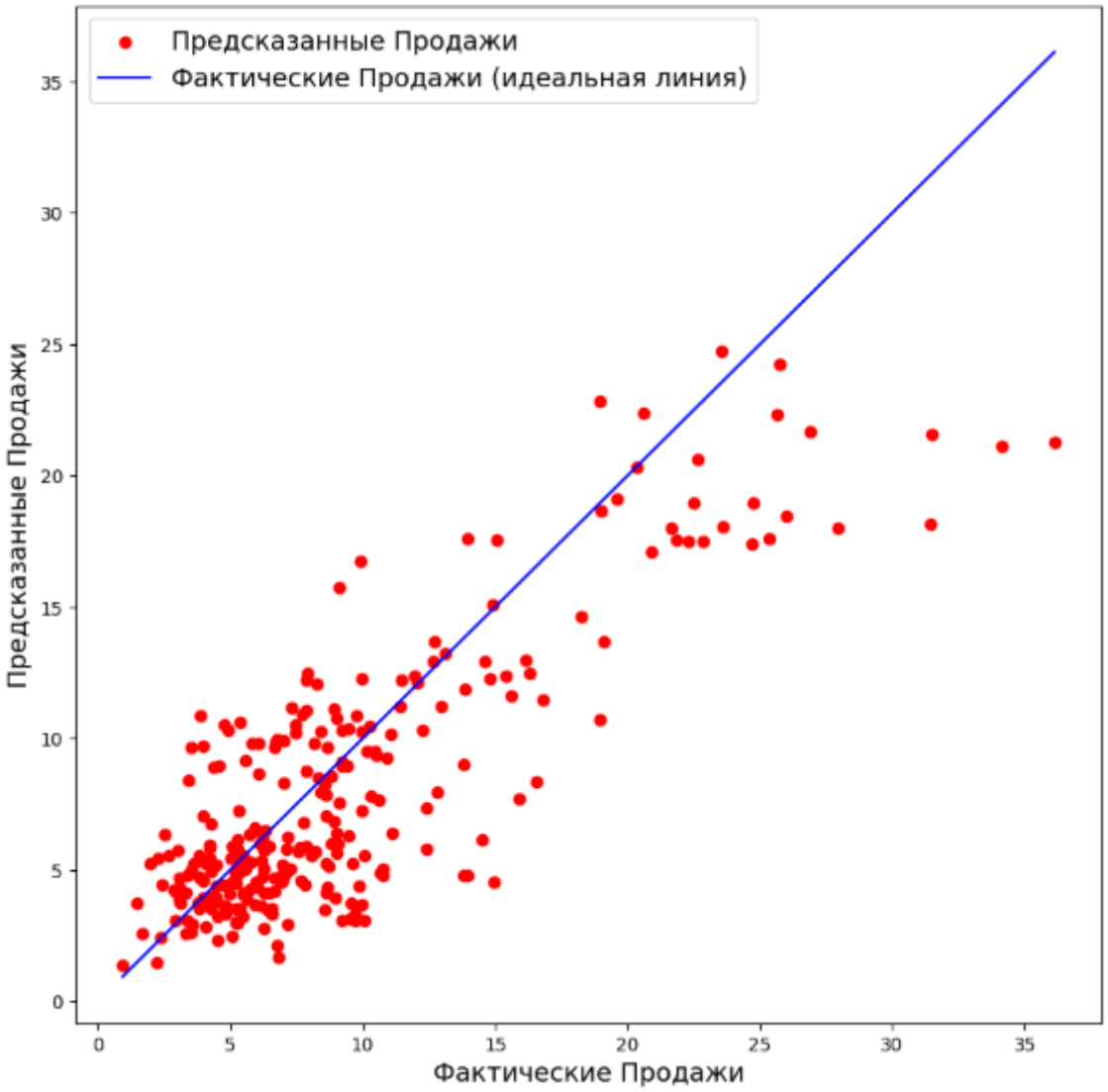

Рисунок 21 – Сравнение прогнозных и фактических точек модели Lightgbm набора данных M2

Рисунок 21 позволяет визуализировать и оценить качество прогнозирования продаж товара лимоны набора данных M3.

Выводы и рекомендации по применению модели прогнозирования спроса в розничной торговле следующие. Модель LightGBM продемонстрировала высокую эффективность в прогнозировании временных

рядов, что подтверждается точностью её предсказаний для различных магазинов и товаров. Важно отметить, что качество прогнозов значительно варьировалось в зависимости от магазина, что подчеркивает необходимость индивидуальной настройки модели для каждого магазина. Рекомендуется использовать методику, основанную на словаре Python, для хранения данных и моделей, что облегчает управление множеством временных рядов и повышает эффективность их обработки. Это особенно актуально для крупных розничных сетей с большим количеством торговых точек и разнообразием ассортимента. Модель должна учитывать сезонные колебания и тренды, которые могут существенно влиять на продажи. Это позволит повысить точность прогнозов и избежать значительных ошибок. При построении модели необходимо также учитывать специфические факторы, такие как региональные особенности, маркетинговые кампании и экономические условия, которые могут влиять на поведение покупателей.

Анализ метрик ошибок, таких как MAE, RMSE, MAPE и Bias, позволяет выявить сильные и слабые стороны модели, что помогает в её дальнейшем улучшении. Для магазинов с высокими значениями ошибок требуется дополнительная настройка модели и возможно использование дополнительных данных для повышения точности прогнозов. Рекомендуется проводить тестирование модели на новых данных, чтобы своевременно выявлять и исправлять возможные проблемы. Для повышения эффективности модели можно использовать комбинацию нескольких методов машинного обучения и статистических моделей, что позволит учесть различные аспекты временных рядов и улучшить общую точность прогнозов. Внедрение автоматизированной системы прогнозирования на базе модели LightGBM может значительно сократить затраты времени и ресурсов на планирование и управление запасами, что приведет к повышению общей эффективности бизнеса. Применение модели прогнозирования спроса на основе LightGBM позволяет более точно планировать закупки, минимизировать избыточные запасы и избегать дефицита товаров, что положительно скажется на

финансовых показателях компании. Регулярное обновление модели с учетом новых данных и изменений на рынке поможет поддерживать её актуальность и точность. Это особенно важно в условиях быстро меняющегося потребительского спроса.

ЗАКЛЮЧЕНИЕ

Исследование и разработка моделей прогнозирования временных рядов продаж в розничной торговле представляют собой актуальную задачу, требующую комплексного подхода и применения современных методов машинного обучения. В данной диссертационной работе был использован набор данных M2, включающий информацию о продажах из нескольких магазинов. Основной целью было разработать и оценить эффективность модели прогнозирования с использованием LightGBM, а также расширить анализ до набора данных M3, включающего два товара (бананы и лимоны) и данные из 10 магазинов. Модель LightGBM проявила себя как мощный инструмент для прогнозирования временных рядов, показав разнообразные результаты точности и ошибок для различных магазинов и типов товаров. Анализ метрик MAE, RMSE, MAPE и bias позволил глубже понять, как модель справляется с прогнозированием и какие аспекты требуют дальнейшей оптимизации Использование словаря Python для хранения и управления данными о продажах каждого магазина и соответствующими моделями обучения стало ключевым элементом исследования. Этот подход обеспечил структурирование данных и гибкость в адаптации моделей к уникальным особенностям каждого временного ряда. Благодаря этому удалось эффективно управлять разнообразием данных и обеспечить точность прогнозов на уровне отдельных магазинов.

Расширение исследования до набора данных M3 позволило включить данные о продажах двух товаров в 10 магазинах, что представляет собой более глубокий и детализированный анализ. Анализ MAE, WAPE и bias для лимонов подтвердил необходимость подбора гиперпараметров к каждому магазину при построении модели прогнозирования. Например, высокое значение MAE и RMSE указывает на необходимость дополнительной настройки и улучшения прогнозов, особенно в магазинах с большими колебаниями продаж. Основываясь на данных набора M3, можно сделать вывод о значительном

влиянии уникальных характеристик каждого магазина на точность прогнозов. Это подчеркивает важность индивидуального подхода к каждому временному ряду при разработке и адаптации моделей. Применение словаря Python для организации данных оказалось критически важным для эффективного управления и анализа множества временных рядов. Оно не только улучшило структурирование данных, но и ускорило процесс обучения и тестирования моделей.

Для дальнейших исследований рекомендуется использовать группу схожих товаров для одной модели прогнозирования и исследовать возможность сравнение двух моделей для одного набора тренировочных данных с применением подбора гиперпараметров для каждой модели. Добавление доверительного интервала для каждой модели, добавление зависимых факторов с помощью моделей, добавление API для комфортного использования без специальных знаний программирования, рекомендуется исследовать возможность использования дополнительных источников данных, таких как погодные условия, социальные медиа и экономические индикаторы. Также рекомендуется интегрировать модель с системами управления запасами и заказами, что позволит автоматизировать процессы планирования и снизить вероятность человеческих ошибок.

В заключение, результаты данной работы имеют практическую значимость для улучшения управленческих решений в розничной торговле и подтверждают перспективность применения современных методов анализа данных и машинного обучения для прогнозирования временных рядов. Это открывает новые возможности для повышения эффективности бизнес-процессов, улучшения оптимизации запасов, снижений списаний непроданной продукции с истекшим сроком годности.

СПИСОК ИСПЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Хак, М. С. Прогнозирование розничного спроса: сравнительное исследование многомерных временных рядов / М. С. Хак, Ш. Амин, Д. Миах // Cornell University. – Август, 2023. – URL: https://arxiv.org/abs/2308.11939 (дата обращения: 12.04.2024).

2. Пивкин, К. С. Прогнозирование ключевых показателей розничной сети во времени / К. С. Пивкин // Вестник ПГУ. Серия: Экономика. – 2017. – № 4. – URL: https://cyberleninka.ru/article/n/prognozirovanie-klyuchevyh-pokazateley-roznichnoy-seti-vo-vremeni (дата обращения: 13.05.2024).

3. Шкор, О. Н. Использование big data and advanced analytics, а также data science для оптимизации производственных и бизнесрешений в сфере розничной торговли / О. Н. Шкор, Э. В. Котович // Big data и анализ высокого уровня : сборник научных статей IX международной научно-практической конференции. В 2х частях / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2023. – С. 454–458.

4. Рен, Ю. Новая модель dbn для прогнозирования временных рядов / Ю. Рен // Международный журнал компьютерных наук IAENG. – 2017. – Т. 44, №. 1. – С. 79–86.

5. Сравнение Prophet и глубокого обучения с ARIMA при прогнозировании оптовых цен на продукты питания / Л. Менкулини и др. // Прогнозирование. – 2021. – Т. 3, №. 3. – С. 644–662. – URL: https://arxiv.org/abs/2107.12770 (дата обращения: 13.05.2024).

6. Гудвин, П. Предупрежден: скептическое руководство по прогнозированию / П. Гудвин. – Biteback Publishing, 2017. – С. 304.

7. Маккинни, У. Python и анализ данных : практическое пособие / У. Маккинни ; пер. с анг. А. А. Слинкина. – Москва : ДМК Пресс, 2020. – 540 с. – URL: https://znanium.ru/catalog/product/2012523 (дата обращения: 10.06.2024).

8. Плас, В. Д. Python для сложных задач: наука о данных и машинное обучение / В. Д. Плас. – Москва, 2023. — 576 с.

9. Грас, Дж. Data Science. Наука о данных с нуля / Дж. Грас. – О Рейли Медиа, Инк., 2019.

10 Пойнтинг, Дж. Х. (1884 г.). Сравнение колебаний цен на пшеницу и импорта хлопка и шелка в Великобританию / Дж. Х. Пойнтинг // Журнал Лондонского статистического общества. – 2012. – № 47(1). – С. 34–74. – URL: https://doi.org/10.2307/2979211 (дата обращения: 10.06.2024).

11. Анализ временных рядов: прогнозирование и контроль / Д. Э. П. Бокс, Г. М. Дженкинс, Г. К. Рейнсел, Г. М. Льюнг // Журнал анализа временных рядов. – 2016. – №. 37 (5). – С. 712. – URL: https://www.researchgate.net/publication/299459188 (дата обращения: 10.06.2024).

12. Грейнджер, К. Прогнозирование экономических временных рядов / К. Грейнджер. – Орландо : Академическая пресса, 1986. – xiv, 338. – ISBN 0122951832.

13. Вербос, П. Д. За пределами регрессии: новые инструменты прогнозирования и анализа в поведенческих науках : неопубликованный доктор философии / П. Д. Вербос ; Гарвардский университет, факультет прикладной математики. – Кембридж, Массачусетс, 1974. – URL: https://gwern.net/doc/ai/nn/1974-werbos.pdf (дата обращения: 10.06.2024).

14. Хопфилд, Дж. Нейронные сети и физические системы с возникающими коллективными вычислительными способностями / Дж. Хопфилд // Доклад Национальной академии наук США. – Апр. 1982.

15 Румельхарт, Д. Э. Изучение представлений с помощью ошибок обратного распространения / Д. Э. Румельхарт, Д. Э. Хинтон, Р. Дж. Уильямс // Природа. – 1986. – № 323. – С. 533–536.

16. Уильямс, Р. Дж. Алгоритм обучения для непрерывной работы полностью рекуррентных нейронных сетей / Р. Дж. Уильямс, Д. Зипсер // Нейронные вычисления. – 1989. – С. 270–280.

17. Хохрейтер, С. Длинная кратковременная память / С. Хохрейтер, Дж. Шмидхубер // Neural Computation. – 1997. – № 9. – С. 1735-1780.

18. Герс, Ф. А. Непрерывное прогнозирование с использованием LSTM с воротами забывания / Ф. А. Герс, Дж. Шмидхубер, Ф. Камминс // Нейронные сети WIRN Vietri-99. Перспективы нейронных вычислений : доклад конференции / ред В. Маринаро, Р. Тальяферри. – Лондон, 1999. – С. 133–138. – URL: https://doi.org/10.1007/978-1-4471-0877-1_10 (дата обращения: 10.06.2024).

19. Лю, Цзяньюй. Комбинированная модель для прогнозирования временных рядов с использованием LSTM посредством извлечения динамических функций на основе пространственного сглаживания и последовательного общего вариационного разложения / Лю, Цзяньюй и др. // Cornell University – 2024.

20. Куинлан, Дж. Р. Индукция деревьев решений / Дж. Р. Куинлан // Mach Learn. – 1986. – № 1. – С. 81–106. – URL: https://doi.org/10.1007/BF00116251(дата обращения: 10.06.2024).

21. Вапник, В. Н. Введение: четыре периода исследования проблемы обучения / В. Н. Вапник // Природа статистической теории обучения. Статистика для инженерии и информатики. – Спрингер, Нью-Йорк, штат Нью-Йорк, 2000. – С. 1–15. – URL: https://doi.org/10.1007/978-1-4757-3264-1_1 (дата обращения: 10.06.2024).

22. Градиентное обучение, применяемое для распознавания документов / Ю. Лекун, Л. Ботту, Ю. Бенджио, П. Хаффнер // Proceedings of the IEEE. – 1998. – Т. 86, № 11. – С. 2278–2324.

23. Брейман, Л. Предикторы бэггинга / Л. Брейман // Машинное обучение. – 1996. – Т. 24, № 2. – С. 123–140.

24 Валахович, Э. Периодически коррелированные временные ряды и периодический блочный бутстрап с переменной полосой пропускания / Э. Валахович // Cornell University. – 2024.

25 Брейман, Л. Случайные леса / Л. Брейман // Машинное обучение. – 2001. – № 45. – С. 5–32. – URL: https://doi.org/10.1023/A:1010933404324 (дата обращения: 10.06.2024).

26. Ваго, Н. О. П. Прогнозирование отказов машин на основе многомерных временных рядов: промышленное исследование / Н. О. П. Ваго, Ф. Форбичини, П. Фратернали // Cornell University. – 2024.

27. Фридман, Х. Аппроксимация жадной функции: машина повышения градиента / Х. Фридман // Энн. Статист. – 2001. – № 29 (5). – С. 1189–1232. – URL: https://doi.org/10.1214/aos/1013203451 (дата обращения: 10.06.2024).

28. Хинтон, Г. Е. Алгоритм быстрого обучения для глубоких сетей доверия / Г. Е. Хинтон, С. Осиндеро, Ю. В. Тех // Нейронный компьютер. – 2006. – № 18(7). – С. 1527-54.

29. Чен, Т. Xgboost: Масштабируемая система повышения древовидности / Т. Чен, К. Гестрин // Материалы 22-й международной конференции ACM по обнаружению знаний и интеллектуальному анализу данных. – 2016. – С. 785–794.

30. Гуолинь, К. LightGBM: высокоэффективное дерево решений для повышения градиента / К. Гуолинь // Нейронные системы обработки информации. – 2017.

31. Внимание — это все, что вам нужно / Васвани А. и др. // Достижения в области нейронных систем обработки информации. – 2017. – Т. 30.

32. Фредформер: преобразователь со смещением частоты для прогнозирования временных рядов / Пяо, Сихао и др. // Cornell University. – июнь 2024. – URL: https://arxiv.org/abs/2406.09009 (дата обращения: 10.06.2024).

33. Зохди, М. Алгоритмы машинного обучения на основе прогнозирования спроса на информации о клиентах: прикладной подход / М. Зохди // Международный журнал информационных технологий. – 2022.

34. Вианти, Д. Т. Алгоритм машинного обучения для задач прогнозирования солнца / Д. Т. Вианти, И. Харисудин // Физический журнал:

серия конференций. – 2021. – URL: https://iopscience.iop.org/article/10.1088/1742-6596/1918/4/042012/pdf (дата обращения: 12.05.2024).

35. Яниш, К. Машинное обучение и глубокое обучение / К. Яниш, П. Зшех, К. Хайнрих // Электронные рынки. – 2021. – Т. 31, № 3. – С. 685–695.

36. Блюм, Л. Влияние преобразования Бокса-Кокса на алгоритмы машинного обучения / Л. Блюм, М. Эльгенди, К. Менон // Передний Артиф Интел. – Апр. 2022.

37. Серкейра, В. Оценка моделей прогнозирования временных рядов: эмпирическое исследование методов оценки эффективности / В. Серкейра, Л. Торго, И. Мозетич // Маха Жира. – 2020. – Т. 109, № 2. – С. 1997–2028. – URL: https://doi.org/10.1007/s10994-020-05910-7 (дата обращения: 12.05.2024).

38. Анафиев, А. С. Обзор подходов к решению задач оптимизации гиперпараметров для алгоритмов машинного обучения / А. С. Анафиев, А. С. Карюк // ТВИМ. – 2022. – № 2. – С. 30–37.

39. Хевамалаж, Х. Оценка прогнозов для ученых, работающих с данными: распространенные ошибки и лучшие практики / Х. Хевамалаж, К. Аккерманн, К. Бергмейр // Интеллектуальный анализ данных и обнаружение знаний. – 2023. – Т. 37, №. 2. – С. 788–832. – URL: https://arxiv.org/pdf/2203.10716 (дата обращения: 10.05.2024).

40. Тофаллис, К. Показатель относительной точности прогнозирования для выбора и оценки моделей / К. Тофаллис // Журнал организаций операционных исследований. – 2015. – № 66(8). – С. 1352–1362. – URL: https://doi.org/10.1057/jors.2014.103 (дата обращения: 10.05.2024).

41. Ходсон, Т. О. Среднеквадратическая ошибка (RMSE) или средняя абсолютная ошибка (MAE): когда их использовать или нет / Т. О. Ходсон // Геонауки Модель Дев. – 2022. – № 15. – С. 5481–5487. – URL: https://doi.org/10.5194/gmd-15-5481-2022 (дата обращения: 10.05.2024).

42. Бикель, П. Дж. Математическая статистика: основные идеи и избранные. Том 1. / П. Дж. Бикель, К. А. Доксум // Журнал Американской

статистической ассоциации. – 2007. – № 56. – С. 21. – URL: https://www.researchgate.net/publication/236736826_Mathematical_Statistics_Basic_Ideas_and_Selected_Topics (дата обращения: 10.05.2024).

43. Патернейн, Д. Метрики демографической предвзятости набора данных: пример распознавания выражений лица / Д. Патернейн // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2024. – С. 3. – URL: https://www.researchgate.net/publication/377981226_Metrics_for_Dataset_Demographic_Bias_A_Case_Study_on_Facial_Expression_Recognition doi: 10.1109/TPAMI.2024.3361979 (дата обращения: 10.05.2024).

44. Применение алгоритма пророка Facebook для успешного прогнозирования продаж на основе реальных данных / Э. Зуник и др. // Cornell University. – 2020.

45. Пивкин, К. С. Корреляционный анализ факторов влияния на покупательский спрос розничного магазина как этап формирования модели прогнозирования и управления запасами / К. С. Пивкин // Вестник УдГУ. Сер. Экономика и право. – 2016. – № 3. – С. 40–50. – URL: https://cyberleninka.ru/article/n/korrelyatsionnyy-analiz-faktorov-vliyaniya-na-pokupatelskiy-spros-roznichnogo-magazina-kak-etap-formirovaniya-modeli-prognozirovaniya (дата обращения: 10.05.2024).

46. Лясковской, Е. А. Прогнозирование спроса на рынке дорожно-строительной техники с использованием инструментов интеллектуального анализа данных / Е. А. Лясковской // Вестник ЮУрГУ. Серия : Компьютерные технологии, управление, радиоэлектроника. – 2022. – Т. 22, № 3. – С. 117–131. – URL: https://sciup.org/vestnik-susu-ctcr/2022-3-22 (дата обращения: 12.05.2024).

47. Зохди, М. Алгоритмы машинного обучения на основе прогнозирования спроса на основе информации о клиентах: прикладной подход / М. Зохди // Международный журнал информационных технологий. – 2022.

48. Модель прогнозирования спроса на основе возраста продукта в модной рознице / Р. Вашиштха и др. // Cornell University. – июль 2020. – URL: https://arxiv.org/abs/2007.05278 (дата запроса: 12.05.2024).

49 Градиентное обучение, применяемое для распознавания документов / Ю. Лекун, Л. Ботту, Ю. Бенджио, П. Хаффнер // Proceedings of the IEEE. – 1998. – Т. 86, № 11. – С. 2278-2324.

50 Новая модель dbn для прогнозирования временных рядов / Ю. Рен и др. // Международный журнал компьютерных наук IAENG. – 2017. – Т. 44, № 1. – С. 79–86.

51. Конституция Российской Федерации. – Москва : Издательство Юрайт, 2023. – 82 с. // Образовательная платформа Юрайт : [сайт]. – URL: https://urait.ru/bcode/530439 (дата обращения: 18.06.2024).

52. ООО "Торговая сеть Командор": бухгалтерская отчетность и финансовый анализ // Бухгалтерская отчетность за 2011-2023 гг. – URL: https://www.audit-it.ru/buh_otchet/2465008567_ooo-ts-komandor (дата обращения: 18.06.2024).

Федеральное государственное автономное образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт экономики, государственного управления и финансов
Базовая кафедра цифровых финансовых технологий Сбербанка России

УТВЕРЖДАЮ
Заведующий кафедрой
_____ Д.В. Солнцев
подпись
«17» июня 2024 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ ПРОДАЖ В РОЗНИЧНОЙ ТОРГОВЛЕ
НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ
тема

38.04.01 «Экономика»
(код и наименование направления)

38.04.01.17 «Финансово-экономическая аналитика и принятие решений в цифровой среде»
код и наименование магистерской программы

Научный руководитель _____ 18.06.24 _____ к.э.н., доцент _____ Ю.И. Черкасова
подпись, дата должность, ученая степень инициалы, фамилия

Выпускник _____ 18.06.24 _____ В.Р. Рассохин
подпись, дата инициалы, фамилия

Рецензент _____ рук. отдела экономики ООО «Командор-Холдинг» Л.П. Привалихина
подпись, дата должность, ученая степень инициалы, фамилия

Нормоконтролер _____ 13.06.24г. _____ Э.Ф. Мамедова
подпись, дата инициалы, фамилия

Красноярск 2024