УДК 004.8 **Информационные технологии**

Афанасьев Г.И., *кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)*

Афанасьев А. Г., *ассистент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)*

Бурмистрова М.В., *ассистент, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)*

Тэт Вей Ян Со, *магистрант, Московский государственный технический университет им. Н.Э. Баумана (национальный исследовательский университет)*

ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ЭФФЕКТИВНЫХ БИЗНЕС-РЕШЕНИЙ В СИСТЕМАХ ЭЛЕКТРОННОЙ КОММЕРЦИИ

Аннотация: Электронная коммерция — это платформа, на которой люди могут покупать или продавать товары. Основная цель электронной коммерции — предоставить покупателям удобство, когда им не нужно будет идти в физический

магазин, чтобы совершить покупку. Таким образом можно будет сделать покупку онлайн и товар будет доставлен на дом заказчику. В 2019 году общая сумма продаж через электронную коммерцию в США составила 603 миллиарда долларов по сравнению с 3,17 миллиардами розничных продаж. Целью этой работы было исследование алгоритмов машинного обучения, способных прогнозировать продажи платформ электронной коммерции. Были проанализированы модели машинного обучения, которые использовалась в других исследованиях, чтобы решить вопрос о выборе моделей для определения лучших моделей машинного обучения для этого исследования. После того, как были выбраны модели, они были протестированы на точность, погрешность и производительность. Результаты тестирования на конкретном наборе исходных данных показали незначительное преимущество Случайного леса над Повышением градиента, а также преимущество SARIMA над ARIMA.

Ключевые слова: Электронная коммерция, машинное обучение, прогнозирование продаж, ARIMA, SARIMA.

Abstract: E-commerce is a platform where people can buy or sell products. The main purpose of e-commerce is to provide customers with the convenience of not having to go to a physical store to make a purchase. Thus, it will be possible to make a purchase online and the goods will be delivered to the customer's home. In 2019, total U.S. e-commerce sales were $603 billion, compared to $3.17 billion in retail sales. The aim of this work was to investigate machine learning algorithms capable of predicting

the sales of e-commerce platforms. Machine learning models that have been used in other studies were analyzed to address the issue of model selection to determine the best machine learning models for this study. After the models were selected, they were tested for accuracy, error, and performance. The results of testing on a specific set of initial data showed the light advantage of Random Forest over Gradient Boosting, as well as the advantage of SARIMA over ARIMA.

Key words: E-commerce, machine learning, sales forecasting, ARIMA, SARIMA

Введение

Областью данного исследования является электронная коммерция. Запуск нового метода сбора и анализа данных может оказать огромное влияние на организацию, поскольку результат может быть положительным или наоборот. Платформы электронной коммерции собирают большие объемы данных и хранят их в своих центрах обработки данных. Они не рассматривают это как преимущество для своих деловых возможностей, таких как анализ данных и их моделей за последние годы. Например, все данные клиентов из регистрации, истории поиска, продаж, чатов хранятся на сервере и будут использоваться только в случае возникновения проблемы с существующими данными. Понятно, почему они не хотят, чтобы другая компания анализировала их данные из-за проблем с конфиденциальностью, но они также могут создать свою собственную команду для анализа данных, которые могут быть им выгодны.

Показано, что электронная коммерция является одним из 10 крупнейших предприятий с очень большим объемом хранения данных. Имея доступ к объему данных, они смогут изменить правила игры для индустрии электронной коммерции. Эта отрасль тратит сотни миллионов долларов на рекламу, социальные сети, сортировку защищенных данных и многое другое, чтобы увеличить продажи, но они не понимали, что с помощью машинного обучения они смогут опередить своих конкурентов. Машинное обучение — это большая древовидная ветвь, которая имеет множество специализаций, таких как данные

майнинг, искусственный интеллект, дополненная реальность и прогнозирование [1; 2; 14]. В этом исследовании рассматривается только аспект прогнозирования с использованием машинного обучения [3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14].

Благодаря возможности прогнозирования с использованием алгоритма машинного обучения для электронной коммерции можно выявлять любые скрытые шаблоны, выбросы, точки интереса (POI) и многое другое. Это позволяет электронной коммерции иметь возможность правильно определить важные детали в каждом аспекте. Они смогут использовать все свои данные, такие как количество приобретенного продукта, категории продуктов, способ оплаты, процентная ставка, продолжительность доставки и местонахождение клиента, чтобы лучше понять, как улучшить свои продажи и управлять ими.

Если платформа электронной коммерции сможет прогнозировать свои продажи на предстоящий год, месяц или день, они смогут принимать более

эффективные бизнес-решения. Они также смогут отслеживать тенденции в своих продажах, если ежегодно проводится какой-либо фестиваль или мероприятие. Они также смогут отслеживать свои запасы, чтобы все товары были в достаточном количестве, что позволит избежать затоваривания или нехватки товара, поскольку они смогут получить приблизительную оценку покупок, которые могут произойти. Мало того, они также смогут лучше отслеживать свои финансы и делать разумные покупки, а также иметь надлежащий бюджет на протяжении всей коммерческой деятельности.

2. Методы машинного обучения

Прогнозирование продаж является важной задачей, которую должна выполнять электронная коммерция, и прогнозирование может оказать решающее влияние на процесс принятия бизнес-решений. Мало того, что по имея прогноз продаж для платформы электронной коммерции, они могут лучше понять свое финансовое положение для управления рабочей силой и дальнейшего улучшения управления своей цепочкой поставок.

Основываясь на [5] и [6], прогноз продаж позволяет платформе электронной коммерции иметь более точные и надежные прогнозы, которые помогут им в планировании запасов, конкурентоспособных ценах истратегии своевременного продвижения.

Согласно [7], прогноз продаж электронной коммерции позволяет понять жизненный цикл платформы электронной коммерции как ее продажи и рост, стабильность, спад и то, как на продажи влияют краткосрочные цели продукта,

такие как продвижение, ценообразование, сезон и рейтинг в Интернете.

Согласно исследованию, проведенному [5], использовался алгоритм сверточной нейронной сети (CNN) для прогнозирования продаж в электронной коммерции. Это исследование было проведено для устранения выявленного ограничения, которое заключалось в том, что метод требовал индивидуальной ручной разработки функций для конкретных сценариев, что сложно, отнимает много времени и требует большого количества экспертных знаний. Однако цель этого исследования заключалась в том, чтобы определить, может ли этот подход автоматически извлекать эффективные функции и обеспечивать прогнозирование продаж на основе извлеченных функций, как упоминалось в [5]. Основным алгоритмом, который использовался для этого исследования, был алгоритм CNN для прогнозирования продаж. Однако для целей сравнения в исследовании были выбраны также алгоритмы ARIMA, DNN, TL и WD, чтобы получить наиболее точные результаты для прогнозирования продаж. Использовались также уменьшение и перенос веса образца, методы обучения для дальнейшего повышения точности прогнозирования, которые доказали свою высокую эффективность в экспериментах. Результаты исследования показали, что MST модель ARIMA имеет самое высокое среднее значение, однако алгоритм CNN достиг своей цели, поскольку он может автоматически извлекать эффективные функции и выполнять прогнозирование продаж с использованием извлеченных функций.

Основываясь на исследованиях, проведенных [6] и [7], можно сделать

вывод, что использовался также алгоритмы нейронной сети. Оба этих алгоритма нейронной сети имеют свой собственный подход, в котором в исследовании 2018 года используется нелинейная авторегрессивная нейронная сеть (NARNN), а в исследовании 2019 года проводилось использование рекуррентных нейронных сетей (RNN) и сетей с долговременной кратковременной памятью (LSTM). Алгоритм работы реализовывался в рамках специальной нейронной сети. Эти исследователи использовали собственный алгоритм этого подхода для прогнозирования продаж и спроса в электронной коммерции. Проблема, о которой говорится в исследованиях, аналогична: это сложность в определении различных моделей спроса/продаж между продуктами и имеющихся корреляций. Цель обоих исследований заключалась в том, чтобы создать систематическую структуру предварительной обработки для преодоления проблем в условиях электронной коммерции, а также создать основу для прогнозирования. Алгоритм, который использовался для сравнения в этих исследований был ARIMA (анализ временных рядов). Обсуждение результатов исследования 2018 года показало, что ошибка прогноза для NARNN составляет 0,1016, а для ARIMA — 0,1389, что показывает, что NARNN имеет более низкую частоту ошибок по сравнению с ARIMA. Результаты исследования 2019 года также показывают, что LSTM имеет более низкое среднее значение и медиану по сравнению с ARIMA.

Прогнозирование продаж обычно делается с использованием наиболее распространенного метода — анализа временных рядов. Анализ временных рядов включает функцию авторегрессии, которая помогает любому типу

прогнозного анализа.

Согласно [8] исследованию – использовались модели машинного обучения для прогнозирования временных рядов продаж. В нем упоминается, что прогнозирование продаж является современным методом бизнес-аналитики.

Также упоминается [7], что модель ARIMA имеет лучший подход к производительности прогнозирования при анализе временных рядов. Основная проблема, заявленная в этом исследовании [8], заключается в том, что для данных временных рядов требуются большие данные для отражения сезонности, а большие данные о транзакционных продажах могут иметь много отсутствующих данных и выбросов. Затем эти данные необходимо будет учитывать множество различных факторов, которые могут повлиять на продажи. Цель этого анализа временных рядов состоит в том, чтобы объединить различные алгоритмы временных рядов, чтобы повысить точность этого прогноза. В исследовании было выбрано пять алгоритмов: ExtraTree, ARIMA, RandomForest, Lasso и Neural Network, которые являются алгоритмами временных рядов и находятся под наблюдением. Основываясь на результатах тестирования ошибок прогнозирования, ExtraTree имеет самую высокую ошибку проверки по сравнению с остальными, а нейронная сеть имеет самую низкую ошибку проверки, что делает его одним из лучших алгоритмов для прогнозирования.

Анализ другого исследования [9] показал, что было проведено исследование по прогнозированию продаж Walmart с использованием алгоритмов машинного обучения. Ключ к этому исследованию был сделан путем

реализации нескольких различных алгоритмы классификации в данных о продажах из всех магазинов Walmart по всей территории Соединенных Штатов. Проблема, которая была выделена в этом исследовании, заключалась в создании конкурентного сравнительного анализа, чтобы найти лучший алгоритм. Исследователь выбрал 3 различных алгоритма для сравнения и протестировал их, используя оценку MAE R^2 Score. Целью этого исследования является определение точности алгоритма с использованием различных гиперпараметров каждой модели для получения наилучшей средней абсолютной ошибки (MAE) и оценки R^2. Методы, которые использовались для этого исследования, были Случайный лес и Повышение градиента. Результаты этого исследования показывают, что случайный лес является лучшим алгоритмом, который набрал минимальное количество баллов в оценке MAE (1979,4) и высокий балл R ^ 2 (0,94), который показал высокую точность по сравнению с другими.

3. Обнаружение знаний в базах данных (KDD)

Для этого проекта и исследования была выбрана методология KDD, поскольку она в основном соответствует требованиям этого проекта. Эта методология широко используется в области машинного обучения для распознавание образов, статистика, базы данных, искусственный интеллект (ИИ) и визуализация данных (обычно результат) (DBD, 2019). В методологии KDD есть 5 шагов, которые будут обсуждаться.

Для этого прогнозирования продаж в электронной коммерции с использованием машинного обучения для данной работы, первым шагом был

поиск подходящего набора данных электронной коммерции, который является бесплатным и с открытым исходным кодом, затем предварительная обработка выбранного набора данных, и далее затем преобразование набора данных (удаление того, что не требуется, или объединение нескольких наборов данных), моделирование с использованием другого алгоритма для достижения более высокой точности, и, наконец, оценка результатов этапов моделирования.

3.1. Выбор данных

Набор данных для этого исследования был набором транзакционных данных. Это связано с тем, что мы будем делать прогноз продаж, для которого требуются все прошлые транзакционные данные для прогнозирования будущих продаж. Набор данных транзакционный будет из одной из электронной коммерции с открытым исходным кодом и может использоваться без каких-либо ограничений. Набор данных, который был получен, получен от Kaggle.com, который перечислил бразильский общедоступный набор данных электронной коммерции от Olist Store (https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce). Предоставляется около 100 000 данных истории транзакционных заказов. Все эти предоставленные данные относятся к периоду с 2016 по 2018 год и предоставляют восемь различных наборов данных, которые содержат разные наборы данных, такие как набор данных о продуктах, набор данных о заказах, набор данных о клиентах и наборы данных об элементах заказа.

3.2. Предварительная обработка данных

Предварительная обработка данных — это процесс, который объясняет,

как выбранные данные будут очищены от всего шума или выбросов. Это означает, что очистка данных, которые имеют огромное количество дополнительного значения, означает меньше и не требуется. Например, в этом наборе данных есть обзор продукта, и в прогнозе продаж нет необходимости в обзоре продукта, поэтому шум — это обзор продукта, и ему нужно быть удалены. Мало того, если в наборе данных есть пропущенные значения для значений продаж и цен, нам нужно будет обработать его соответствующим образом, заменив пропущенное значение средним значением или используя среднее значение или медианное вменение для обеспечения согласованности данных. На этом этапе мы также можем учитывать временную последовательность данных и известные изменения.

3.3. Преобразование данных

Преобразование данных — это процесс преобразования данных из одного формата в другой формат для удовлетворения потребностей. Этот процесс также можно отнести к процессу ETL, что означает извлечение, преобразование и загрузку. Преобразование стало действительно важной задачей, поскольку объем данных значительно увеличился. Таким образом, надежное преобразование данных позволит пользователю сосредоточиться на данных, которые удовлетворяют потребности бизнеса. То же самое касается и этого проекта: весь набор данных будет подвергнут серьезному преобразованию, и только важные данные будут объединены в новый формат данных. Это поможет сосредоточить внимание исследователя на важных данных и сможет построить лучшую модель

прогнозирования, которая обеспечит лучшие результаты по точности. Исследователь должен будет проанализировать все восемь различных наборов данных, определить все важные значения и информацию и преобразовать их в один или несколько различных форматов файлов. Это значительно облегчит исследовательскую работу на этапе моделирования, поскольку им не придется просматривать ненужную информацию.

3.4. Моделирование

Моделирование — это процесс, в котором оцениваются алгоритмы, которые необходимо использовать для целей исследования проекта. Для этого исследования были использованы два разных алгоритма: Повышение градиента и Случайный лес. Эти алгоритмы выбираются потому, что они обычно связаны с прогнозным анализом.

Повышение градиента — это метод машинного обучения, который включает в себя классификацию и регрессию продуктов на основе ансамбля слабых моделей прогнозирования, таких как дерево решений.

Случайный лес содержит огромное количество деревьев решений, которые работают в группе, и каждое отдельное дерево предоставит прогноз класса. Модели, которые были тестированы, были основаны на анализе временных рядов. Для этого исследования были протестированы два метода моделирования: авторегрессионное интегрированное скользящее среднее (ARIMA) и сезонное авторегрессионное интегрированное скользящее среднее (SARIMA). Эти модели были выбраны, поскольку они являются одними из лучших моделей, которые

могут обеспечить надлежащую точность прогноза.

4. Результаты и обсуждение

Для этих результатов и обсуждения необходимо показать всю оценочную матрицу, которая использовалась в соответствии с их алгоритмами. Поскольку этот проект был посвящен прогнозированию продаж, были использованы две регрессионные модели: «Случайный лес» и «Повышение градиента». Из анализа временных рядов были использованы еще две модели: ARIMA и SARIMA.

Для регрессионных моделей существует множество различных способов оценки модели с использованием различных оценочных матриц. Было выбраны точность, среднеквадратичная ошибка (RMSE), средняя Абсолютная ошибка (MAE) и метод R Squared для оценки моделей. Результаты, полученные от каждой модели, были сравнены с другой моделью, чтобы увидеть, какая модель имеет наименьшую ошибку.

Для моделей временных рядов количество способов оценки ограничено. Поэтому была выбрана средняя абсолютная ошибка в процентах (MAPE), среднеквадратичная ошибка (RMSE) и Средняя абсолютная ошибка (MAE) для оценки моделей временных рядов. Затем обе модели были сравнимы посредством оценочной матрицы, чтобы определить, какая модель лучше.

4.1. Модели

В таблице 1 показаны результаты прогнозирования продаж с использованием моделей Случайный лес и Повышение градиента.

Таблица 1. Результаты прогнозирования продаж с использованием выбранных моделей

	Точность обучения	Тестовая точность	RSME	MAE	R^2
Случайный лес	90.99%	87.39%	76 669.68	55133.85	0.87
Повышение градиента	99.99%	86.88%	82 230.19	61 188.42	0.85

Основываясь на результатах, приведенных выше в таблице 1, матрица оценки была указана как для Random Forest, так и для Gradient Boosting, было замечено, что точность обучения, достигнутая за счет повышения градиента, составила 99,99%, что выше по сравнению со Случайным лесом. Однако точность теста, полученная случайным лесом, составляет 87,39%, что выше по сравнению с Повышением градиента. Это уже показывает, что модель Случайного леса лучше соответствует данным, потому что оценки обучения и теста очень близки по сравнению с Повышение градиента. Здесь разница в точности обучения и теста для повышения градиента составляла около ~ 13%, что предполагало, что тестовые данные на самом деле не были приняты моделью, поэтому она имела более низкую точность теста. Переходя к другой оценочной матрице, было замечено, что среднеквадратическая ошибка (RSME), полученная Случайным лесом, меньше, что означает, что Случайный лес имеет более низкую невязку (ошибку прогноза) по сравнению с Повышением градиента. Средняя абсолютная ошибка, полученная при поиске градиента, выше, что означает, что усиление градиента имеет большую среднюю ошибку. Наконец, R^2,

полученный Случайным лесом, выше на 0,02 по сравнению с Повышением градиента.

Исходя из этого можно сделать вывод, что модель Случайного леса и Повышения градиента отлично справились с подгонкой данных к модели. Однако лучшей моделью для регрессионной модели является Случайный лес, поскольку он имеет меньшую ошибку точности по сравнению с Повышением градиента.

4.2. Модели временных рядов

Результаты прогнозирования продаж с использованием моделей временных рядов приведены в таблице 2.

Таблица 2. Результаты прогнозирования продаж с использованием моделей временных рядов

	RSME	MAE	MAPE
ARIMA	1806.35	1528.94	8.62%
SARIMA	1812.38	1521.32	6.69%

На основе приведенного выше в таблице 2 вся оценочная матрица была сведена в таблицу в соответствии с моделями временных рядов: ARIMA и SARIMA. Прежде всего, среднеквадратическая ошибка (RSME) для обеих моделей почти одинакова, но SARIMA имеет более высокое значение RSME, что означает, что она имеет небольшой остаток по сравнению с ARIMA. Далее идет средняя абсолютная ошибка (MAE). Это также был близкий вызов, поскольку разница составляла 7.62, что означает, что модель ARIMA имеет немного более

высокую частоту ошибок по сравнению с SARIMA. Наконец, MAPE — один из самых надежных методов прогнозирования временных рядов. Оценка SARIMA MAPE меньше, чем у модели ARIMA, что означает, что SARIMA имеет только 6,69% ошибки в модели, в то время как ARIMA имеет ошибку 8,62%.

Пройдя всю оценочную матрицу для моделей временных рядов, можно сделать вывод, что SARIMA более лучшая модель, поскольку она имеет более низкий показатель MAPE, что означает меньшую ошибку по сравнению с моделью ARIMA. SARIMA также позволяет увидеть сезонные тенденции в данных, что дает более полное представление.

5.Заключение

Таким образом, результаты исследования по использованию машинного обучения для прогнозирования бизнес-решений электронной коммерции показали, что существует множество различных методов машинного обучения, которые можно использовать при прогнозировании бизнес-решений в будущем. Для выбора наиболее подходящего метода машинного обучения для целей поставленной задачи целесообразно провести предварительно тестированные выбранных альтернатив на конкретных наборах исходных данных в целях определения предпочтительно варианта решений. В качестве решения выбирается модель, имеющая наилучший диапазон предсказания, где прогнозируемое значение и фактическое значение почти совпадают. Затем интегрировать выбранный метод в разработку рекомендательной системы. Результаты данного исследования на одном конкретном наборе исходных данных

показали преимущество Случайного леса над Повышением градиента, и модели SARIMA над ARIMA. Однако преимущество было незначительное, что говорит о возможности применения всех рассмотренных методов в общем случае.

Библиографический список:

1. Афанасьев Г.И., Абулкасимов М.М., Сурикова О.В. Алгоритмы оптимизации, используемые в нейронных сетях, и градиентный спуск // Аспирант и соискатель. 2019. № 6 (114). С.81- 86.
2. Галичий Д.А., Афанасьев Г.И., Нестеров Ю.Г., Распознавание эмоций человека при помощи современных глубокого обучения методов // E-SCIO. 2021. №5. С.317-331.
3. Чжан Чжибо, Афанасьев Г.И. Основные технологии и перспективы эволюции персонализированных рекомендательных систем // E-SCIO. 2024. №4(67). С.309-320.
4. Чжан Чжибо, Афанасьев Г.И. Применение моделей глубокого обучения в области рекомендательных систем, основанных на содержании //Искусственный интеллект в автоматизированных системах управления и обработки данных (ИИАСУ 22). Сборник статей Всероссийской научной конференции. В 2 т. (Москва. МГТУ им. Н.Э. Баумана. 27-28 апреля 2022 г.). - М.: МГТУ им. Н.Э. Баумана, Т.1. С. 278-284.
5. Zhao K., Wang C. Sales Forecast in E-commerce using Convolutional Neural Network. 2017. 8 p. https://doi.org/10.48550/arXiv.1708.07946.

6. Bandara, K. et al. Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology. 2019. 16p. https://doi.org/10.48550/arXiv.1901.04028.

7. Li M., Ji S., Liu G. Forecasting of Chinese E-Commerce Sales //An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model', Mathematical Problems in Engineering. 2018. 13p. https://doi.org/10.1155/2018/6924960.

8. Pavlyshenko B. Machine-Learning Models for Sales Time Series Forecasting. 2019. 11p. DOI:10.3390/data4010015.

9. Elias S., Singh S. Forecasting of WALMART sales using machine learning algorithms. 2018. Access mode: https://medium.com/@auggieheschmeyer/forecasting-walmart-sales-using-machine-learning-models-3bf38f6c533.

10. Yu J.. Le X. Sales Forecast for Amazon Sales Based on Different Statistics Methodologies // DEStech Transactions on Economics and Management. 2017. Access mode: https://www.semanticscholar.org/paper/Sales-Forecast-for-Amazon-Sales-Based-on-Different-Yu-Le/fb39929d6d4abcabaa3fb01ebdc05229167c99bb.

11. Madhuvanthi, K. et al. Car sales prediction using machine learning algorithmns, International Journal of Innovative Technology and Exploring Engineering. 2019. Access mode: https://research.vit.ac.in/publication/car-sales-prediction-using-machine-learning-algorithmns.

12. Xia G., He Q. The Research of Online Shopping Customer Churn

Prediction Based on Integrated Learning. 149(Mecae). 2018. Access mode: https://www.atlantis-press.com/proceedings/mecae-18/25893766.

13. Bohanec, M., Kljajić Borštnar M., Robnik-Šikonja M. Explaining machine learning models in sales predictions // Expert Systems with Applications. 2016. Access mode: https://www.researchgate.net/publication/309885039_Explaining_machine_learning_models_in_sales_predictions

14. Mohammed M., Khan M. B., Bashie E. B. M. Machine learning: Algorithms and applications, Machine Learning: Algorithms and Applications. CRC Press. 2016. 226p. https://doi.org/10.1201/9781315371658.