

УДК 004.852

<https://doi.org/10.33619/2414-2948/128/06>

АНСАМБЛЕВЫЕ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ РОЗНИЧНЫХ ПРОДАЖ В ЭЛЕКТРОННОЙ КОММЕРЦИИ

©Судаков В. А., ORCID: 0000-0002-1658-1941, SPIN-код: 1614-4760, Scopus ID: 7006296922, ResearcherID: I-2573-2018, д-р техн. наук, МИПЭА - Российский технологический университет, г. Москва, Россия, sudakov@ws-dss.com

©Заммам А., ORCID: 0009-0004-6213-8598, МИПЭА - Российский технологический университет, г. Москва, Россия, alizammam@yandex.ru

ENSEMBLE MACHINE LEARNING MODELS FOR RETAIL SALES FORECASTING IN E-COMMERCE

©Sudakov V.A., ORCID: 0000-0002-1658-1941, SPIN-код: 1614-4760, Scopus ID: 7006296922, ResearcherID: I-2573-2018, Dr. habil., Russian Technological University MIREA, Moscow, Russia, sudakov@ws-dss.com

©Zammam A., ORCID: 0009-0004-6213-8598, Russian Technological University MIREA, Moscow, Russia, alizammam@yandex.ru

Аннотация. Развитие электронной коммерции существенно влияет на традиции торговли, трансформируя розничную торговлю за счет использования интернета, данных и автоматизации. Несмотря на значительную простоту, индивидуализацию и доступность, сектор по-прежнему сталкивается с многочисленными проблемами, связанными с прогнозированием продаж и запасов. Эти проблемы связаны с поведением клиентов, огромными объемами данных о транзакциях, сезонностью и изменениями в потребительском спросе. Традиционные статистические модели не учитывают эти аспекты, что вызывает необходимость в сложных моделях машинного обучения. Используя набор данных о продажах Amazon, мы разрабатываем модель машинного обучения для электронной коммерции. Мы предварительно обрабатываем данные, а затем применяем методы адаптивного повышения, случайных лесов и ансамбля экстремального градиентного повышения. Мы оцениваем эти модели с помощью MAE, RMSE и R^2 . В наших экспериментах XGBoost дал наилучшие прогнозы при $R^2 = 0,8010$, MAE = 93,41 и RMSE = 15 084,84, опередив Random Forest и AdaBoost. Эти результаты демонстрируют полезность сложных моделей для больших наборов данных электронной коммерции при принятии обоснованных решений по инвентаризации и розничной торговле на основе данных.

Abstract. The rise of electronic commerce significantly impacts traditions of commerce by transforming retail through its use of internet connectivity, data, and automation. Despite offering substantial ease, customization, and access, the sector still experiences multiple challenges with sales and inventory predictions. These challenges are tied to the nature of the customer, vast amounts of transaction data, seasonality, and changes in customer demand. Traditional statistical models do not capture these aspects, creating the need for complex machine learning models. Using the Amazon Sales Dataset, we develop a machine learning model for electronic commerce. We preprocess the data and then apply Adaptive Boosting, Random Forests, and Extreme Gradient Boosting ensemble methods. We evaluate these models using MAE, RMSE, and R^2 . In our experiments, XGBoost gave the best predictions with $R^2 = 0.8010$, an MAE of 93.41, and an RMSE of 15,084.84, outperforming Random Forests and AdaBoost. These results demonstrate the usefulness of complex models for large e-commerce datasets in solving complex inventory and retail data-driven decisions.

Ключевые слова: прогнозирование, электронная коммерция, машинное обучение, ансамблевая модель, набор данных о продажах Amazon.

Keywords: forecasting, E-Commerce, Machine Learning, Ensemble Model, Amazon Sales Dataset.

Электронная коммерция описывает торговлю, связанную с покупкой и продажей товаров и услуг онлайн, и включает в себя коммерческие отношения и электронные платежи [1].

Из-за кажущейся простоты использования и глобального охвата многие клиенты предпочитают электронную коммерцию из-за широкого ассортимента товаров. Это идеальная платформа для разработки методов машинного обучения (ML) и глубокого обучения (DL) благодаря наличию больших наборов данных, необходимости персонализации, задачам обнаружения мошенничества и обеспечения безопасности, а также возможностям оптимизации цепочки поставок. Это включает в себя анализ настроений клиентов [2].

Многие из этих методов позволяют компаниям электронной коммерции повысить удовлетворенность клиентов, увеличить свою долю на рынке и выделиться среди конкурентов на рынке онлайн-продаж [3].

Для многих клиентов электронная коммерция, в частности, помогла изменить их опыт покупок и продаж [4].

В 2021 г мировой рынок электронной коммерции оценивался в 13 трлн долларов США и, как ожидалось, достигнет 55,6 трлн долларов США к 2027 году, что означает, что с 2022 г по 2027 г он будет расти на 27,4% в годовом исчислении. Впоследствии, по мере того как все больше и больше людей будут выходить в Интернет и пользоваться смартфонами, потребители будут получать более сложный опыт совершения покупок, который будет приносить им больше удовольствия. На самом деле, многие из этих технологий будут обладать сложным поведением, колеблющимся спросом и различным уровнем удовлетворенности клиентов. предприятия электронной коммерции будут по-прежнему сталкиваться с проблемой неустойчивого уровня спроса [5].

Автоматизация и оцифровка значительно повысили устойчивость, производительность и результативность в индустрии электронной коммерции. Большое количество как анонимных, так и идентифицированных пользователей, зарегистрировавшихся на платформах электронной коммерции, могут быть точно и легко доступны для таргетинга [6].

Технологии и электронная коммерция объединились, и компании поняли, что интеграция технологий искусственного интеллекта позволяет компаниям анализировать доступные данные о продуктах и клиентах с целью выявления маркетинговых улучшений для продуктов и услуг. Последние достижения в области информационно-коммуникационных технологий и искусственного интеллекта улучшили электронную коммерцию и упростили продвижение услуг и продуктов, оптимизировали работу бизнеса и позволили глубже понять целевых клиентов [7].

Искусственный интеллект предоставляет полный спектр технологий, которые направлены на решение различных задач и достижение успеха в сегменте электронной коммерции [8, 9].

Учитывая современное значение искусственного интеллекта и электронной коммерции, цель исследователей — объединить эти две области и объяснить возможное применение технологий искусственного интеллекта и машинного обучения в электронной коммерции.

В этом документе представлены многочисленные существенные достижения в области аналитики электронной коммерции. Разработка прогнозного моделирования на основе данных

о продажах Amazon позволяет улучшить структурирование управления запасами, что снижает риск как затоваривания, так и отсутствия на складе. Кроме того, выявление закономерностей закупок объясняет временный характер принятия решений при распределении судебных ресурсов в рамках разработки рекламных кампаний. Система позволяет принимать обоснованные решения о ценах и закупках на основе объема продаж. Будучи интегрированной в процессы управления, электронная коммерция способствует рациональному управлению затратами, повышению эффективности использования ресурсов и принятию управленческих решений на основе данных.

Связанные работы

Автор работы предоставляет обширную базу для сравнения классических статистических методов и алгоритмов машинного обучения в контексте электронной коммерции, специально ориентированных на прогнозирование продаж, а также всестороннюю интерпретацию полученных результатов. В ходе исследования была протестирована эффективность ARIMA, линейной регрессии (LR) и XGBoost с использованием двух различных наборов данных: набора данных о продажах Amazon, состоящего из временных рядов, и набора данных о продажах в электронной коммерции Walmart, не содержащих временных рядов [10]. Методология набора данных Amazon включала преобразование элементов данных о датах во временные ряды с последующей оценкой стационарности временных рядов с использованием расширенного теста Дики-Фуллера (ADF) и применение модели ARIMA для извлечения временных закономерностей. Набор данных Walmart был предварительно обработан путем заполнения пропущенных значений и кодирования категориальных атрибутов, за которыми следовала шкала, перед применением логистической регрессии и XGBoost, причем последний использовал байесовскую оптимизацию для настройки гиперпараметров модели.

Результаты показывают заметную разницу в эффективности моделей: было обнаружено, что XGBoost превосходит логистическую регрессию с меньшим уровнем погрешности: средняя абсолютная ошибка (MAE) составила 0,4294, среднеквадратичная ошибка (RMSE) - 0,5760, по сравнению с MAE 0,7163 и RMSE 0,9289, соответственно. Кроме того, ARIMA продемонстрировала более высокую производительность при обработке данных временных рядов (MAE 11,2906, RMSE 14,7900), в отличие от XGBoost, с MAE 216,4529 и RMSE 277,7282. Результаты демонстрируют доминирование ARIMA во времени и превосходную производительность XGBoost на более сложных непоследовательных данных, что указывает на важность выбора моделей на основе атрибутов конкретного набора данных. Исследование вносит существенный вклад в накопление знаний практиков о выборе методов прогнозирования для различных наборов данных.

Анализируется эффективность по одному примеру каждой из моделей при прогнозировании продаж Walmart, используя набор данных с еженедельными продажами и 16 атрибутами, включая, но не ограничиваясь этим, магазин, продажи и внешние переменные (например, экономику и т.д.). Процесс включает в себя комплексную разработку функций, например, нулевое вычисление, устранение отрицательных отклонений от объема продаж и применение логарифмического преобразования для устранения асимметрии зависимых переменных. Данные разбиваются на обучающие и тестовые наборы в соотношении 3:1 для оценки эффективности модели [11].

Анализ показывает, что XGBoost превосходит все другие модели, набрав наивысший балл по R2 (0,984) и наименьшее количество ошибок как для MSE (0,0655), так и для MAE (0,1246). За ним следует случайный лес (R2: 0,9781, MSE: 0,0887, MAE: 0,1432), а регрессия

OLS является наименее эффективной (R^2 : 0,7257, MSE: 1,1135, MAE: 0,6728). Исследователи объясняют положительные результаты XGBoost способностью модели фиксировать слабые связи и сложные взаимодействия в данных. Регрессия OLS не удается из-за наличия мультиколлинеарности и нелинейных зависимостей. XGBoost также учитывает самую высокую сложность данных в задаче прогнозирования продаж.

Полученные результаты демонстрируют важность моделей машинного обучения, в частности XGBoost, для улучшения прогнозирования продаж с точки зрения розничной торговли, в то время как Random Forest остается удобным выбором, учитывая простоту его модификации. В исследовании рассматриваются некоторые ограничения, такие как необходимость логарифмического преобразования и использования фиксированных гиперпараметров, а также указывается на потенциальное использование нейронных сетей и динамической оптимизации гиперпараметров в будущих исследованиях с целью повышения эффективности прогнозирования.

Разработана методология оценки транзакций электронной коммерции, основанную на моделях случайного леса (RF), деревьев принятия решений (DT) и XGBoost, используя данные о торговле Китая за 2005-2019 годы. В методологии особое внимание уделяется тщательной предварительной обработке данных, которая включала в себя вычисление пропущенных значений, идентификацию и устранение отклонений от нормы, кодирование меток категориальных предикторов и стандартизацию с помощью стандартного масштабирующего устройства. Набор данных был разделен на обучающий и тестовый наборы данных в соотношении 70:30. Результаты показали превосходную производительность модели XGBoost, поскольку она смогла достичь R^2 в 96,3% и RMSE в 7,7, в отличие от модели случайного леса (R^2 : 87) и модели дерева решений (R^2 : 54,76) [12].

Исследователи связывают высокую производительность модели XGBoost с ее способностью оптимизировать методологию коллективного обучения и взаимодействовать, выявлять и/или изучать сложные закономерности в данных транзакций. Авторы усиливают модель результатов обработки данных и тенденций XGBoost за счет использования визуальных данных с помощью графика временных рядов плотности продаж и прогноза. Авторы утверждают, что XGBoost служит мощным инструментом для оптимизации запасов и стратегических улучшений в планируемых впоследствии мероприятиях, связанных с транзакциями электронной коммерции, в то же время отмечая, что он не учитывает влияние социальных сетей. Поэтому авторы также выступают за включение искусственного интеллекта с глубоким обучением и расширение объема данных в будущие работы, чтобы улучшить аналитику и надежность прогнозируемых функций. Авторы доказывают превосходную производительность и действенность разнообразных и сложных алгоритмов искусственного интеллекта для решения динамических моделей прогнозирования электронной коммерции.

Справочник посвящен прогнозированию розничных продаж с использованием набора данных временных рядов от компании «1С», в частности прогнозированию ежемесячных продаж товаров для каждой пары «товар-магазин» [13].

В методологии используется исследовательский анализ данных (EDA), разработка функций и три модели: XGBoost, ARIMA и пользовательская нейронная сеть LSTM. Важные задачи предварительной обработки включали в себя сокращение целевых показателей, заполнение пробелов в наборе данных и реструктуризацию набора данных в виде таблицы. В разделе «Задержка по характеристикам» был создан ряд переменных с задержкой (с задержкой на 1, 3 и 12 месяцев), доход и цена, а также одноразовая кодировка переменных cat, таких как идентификатор магазина и тип товара.

Набор данных был разделен таким образом, что первые 32 месяца составляли обучающий набор, а 33-й месяц - проверочный набор, и для оценки модели использовался RMSE. Результаты показали, что XGBoost показал лучшие результаты среди всех методов, набрав валидационный балл 0,807 и тестовый балл 0,878. Это было связано с его способностью хорошо работать с разреженными данными и оптимизировать использование функций, особенно переменных с запаздыванием (например, важность функции занимает второе место по важности, `target_item_lag_1`). LSTM показал худший результат из трех методов, получив оценку RMSE 0,890, а ARIMA показала худший результат из трех методов, получив оценку RMSE 0,982, из-за своей одномерности. Эта работа демонстрирует масштабируемость и эффективность XGBoost, предполагая, что гиперпараметрическая оптимизация LSTM и другие стратегии, такие как ансамбли, могут улучшить результаты.

XGBoost превосходит все другие модели в работе с непоследовательными данными, что, вероятно, связано со сложными функциями и закономерностями и наличием слабых корреляций [10-13].

Для временных данных модели временных рядов (ARIMA) лучше подходят, но имеют недостаток многомерной зависимости. Пробелы и производительность сильно зависят от предварительной обработки и разработки функций, которые могут включать в себя функции задержки и манипуляции с функциями журнала [10-13].

Глубокое обучение с помощью LSTM вносит положительный вклад, но из-за дороговизны древовидные ансамбли (RF, XGBoost) обеспечивают лучшие результаты как в плане точности, так и в плане масштабируемости. Этот анализ также выявляет пробелы, такие как учет внешних переменных (например, экономических тенденций) и важность динамической корректировки параметров модели, что помогает рассмотреть контекст выбранных методов моделирования в следующих разделах [12, 13].

Методология

В этом разделе описывается метод анализа набора данных о продажах Amazon, который включает в себя подготовку данных, модели ML и показатели эффективности, как показано на Рисунке 1. В разделе «Набор данных» описываются поля данных, доступные в наборе данных, включая информацию о продажах, описания продуктов и промежуточные данные, что помогает понять данные и их ценность. В разделе «Предварительная обработка» описаны важные этапы, такие как сбор данных, преобразование типов, фильтрация данных о продажах и подготовка данных.

В разделе «Алгоритмы машинного обучения» описаны три продвинутых метода ансамбля — Экстремальный градиентный бустинг (XGBoost), адаптивный бустинг (AdaBoost) и случайный лес. Эти алгоритмы выбраны для данной методологии из-за их точности прогнозирования за счет непрерывного обучения и упрощенного улучшения функций. В разделе «Оценочные показатели» описаны важные показатели (MAE, RMSE и R2) и соответствующие им уравнения, которые определяют критерии построения и успешности моделей. Предлагаемый метод направлен на извлечение значений из набора данных путем успешной интерпретации для каждой модели, а также прогнозирования конечных значений интервалов.

Набор данных. Набор данных о продажах Amazon от Kaggle содержит данные о заказах на продажу для Amazon, соответствующий артикул, номер дизайна, наличие на складе, категорию, размер и цвет, что позволяет максимизировать прибыль от продукта. В наборе данных содержится подробная информация о более чем 128 000 записях данных о заказах на продажу, причем в каждой строке содержится подробная информация об одном заказе на

продажу. Каждый заказ описывается в наборе данных с помощью данных транзакции, атрибутов и типов платежей. Целью этого набора данных является анализ эффективности, моделей покупок, поведения в сфере продаж и процессов.

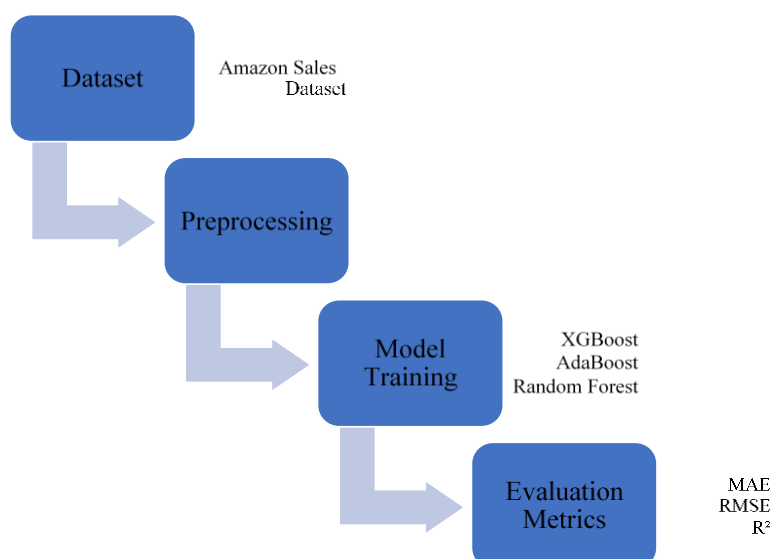


Рисунок 1. Этапы методологии

Этот набор данных содержит подробную информацию о транзакциях продажи и упорядочивает их по заказам, продуктам и отгрузкам. Заказам присваивается идентификатор заказа, а дата транзакции указывается в отдельном столбце Даты. Статус заказа определяется как зарегистрированный статус, в то время как статус отправки определяется в столбце Статус курьера. Информацию о выполнении заказа и ответственных лицах можно найти в столбцах «Способ выполнения» и «Исполнено по» соответственно. В столбце «Канал продаж» указывается платформа продаж заказа.

В нем есть столбцы с информацией о товаре, такие как стиль, артикул, категория, размер и ASIN, которые полезны для каталогизации и инвентаризации. Сведения об отправке описывают уровень обслуживания, город, штат, страну и почтовый индекс судна. Также учитываются такие элементы транзакций, как Количество товара, валюта, сумма и идентификаторы рекламных акций. В столбце B2B указано, какие транзакции являются продажами между предприятиями. Подробная разбивка этих столбцов приведена в Таблице, которая показывает, как структурированы данные.

Предварительная обработка данных. В этом разделе описывается метод обработки данных, который очищает и упорядочивает данные таким образом, чтобы прогнозируемые данные были сосредоточены на реализованных продажах и чтобы данные были хорошо структурированы и подходили для моделирования. Ключевыми задачами являются определение временного порядка продаж, обеспечение наличия данных и создание четко определенного набора функций, на основе которых можно обучить алгоритм.

Загрузка данных и приведение типов. Каталог данных создан на основе файла Amazon Sale Report.csv, и ключевые столбцы были преобразованы в соответствующие типы полей, которые являются согласованными и точными в числовом отношении. Столбец «Дата» оформлен как дата, в то время как столбцы «Сумма» и «Количество» оформлены как числовые типы, что позволяет выполнять фильтрацию, агрегацию и моделирование.

Фильтрация транзакций (Transaction Filtering). Набор данных был отфильтрован для учета только отгруженных и доставленных транзакций (например, «Отгружено», «Отгружено - доставлено покупателю», «Отгружено – получено», «Отгружено для доставки»), при этом

для каждой операции были оставлены значения «Количество» и «Сумма». Поставки, в которых отсутствуют поля Даты, суммы или количества, были проигнорированы, а транзакции, в которых количество или сумма равны нулю, были отфильтрованы, чтобы исключить заполнители, отмены и различные другие помехи. После фильтрации окончательный обработанный набор данных содержал 105 103 строки за период с 2022-03-31 по 2022-06-29 годы.

Таблица

СТОЛБЦЫ И ОПИСАНИЯ НАБОРА ДАННЫХ

<i>Имя</i>	<i>Описание</i>	<i>Тип</i>
<i>Order ID</i>	<i>Уникальный идентификатор для каждого заказа.</i>	<i>Нить</i>
Date	Дата продажи.	Дата
Status	Статус продажи.	Нить
Fulfilment	Метод исполнения.	Нить
Sales Channel	Платформа, через которую была осуществлена продажа.	Нить
ship-service-level	Уровень обслуживания доставки.	Нить
Style	Стиль продукта.	Нить
SKU	Подразделение по управлению запасами.	Нить
Category	Тип продукта.	Нить
Size	Размер продукта.	Нить
ASIN	Стандартный идентификационный номер Amazon.	Нить
Courier Status	Статус курьера.	Нить
Qty	Количество товара.	Целое число
Currency	Валюта, используемая для продажи.	Нить
Amount	Сумма продажи.	Плавать
ship-city ship-state ship-country	Географические данные о пункте назначения доставки.	Нить
ship-postal-code	Почтовый индекс города-корабля.	Целое число
promotion-ids	Любые рекламные предложения, применяемые к заказу.	Нить

Генерация признаков (Feature Engineering). Функции календаря были извлечены из даты транзакции, чтобы суммировать покупательское поведение клиента с течением времени. Эти функции включают в себя day_of_week, day_of_month и month. Были включены и сохранены категориальные характеристики, касающиеся продукта и характеристик исполнения, такие как категория, размер, стиль, качество исполнения, канал продаж, ship_service_level, ship_state, B2B и fulfilled_by. Такие идентификаторы, как SKU и ASIN, которые имеют высокую мощность, были опущены, чтобы предотвратить избыточное использование памяти при однократном кодировании и проблемы с разреженностью.

Работа с категориальными пропусками. Поскольку многие категориальные поля отсутствовали, при их заполнении в качестве заполнителя было использовано «Неизвестно», чтобы количество строк оставалось неизменным, а также чтобы убедиться, что модель обладает соответствующей надежностью для отсутствующих значений атрибутов.

Временное разбиение. Чтобы предотвратить временную утечку, было реализовано разделение по времени. Обучающий набор включает в себя первые 80% (84,082), а последние 20% (21,021) были использованы в качестве тестового набора. Это больше всего напоминает процесс после развертывания, при котором с помощью модели можно прогнозировать будущие тенденции..

Обучение моделей. Здесь описывается структура для обучения регрессионных моделей компонентов, включая настройки гиперпараметров, разделение по времени, конвейер предварительной обработки и структуру для настройки гиперпараметров. Цель состоит в том, чтобы обеспечить сопоставимое представление признаков и настройки оценки для всех моделей.

Случайная лесная регрессия. Случайный лес - это один из таких методов, который развился из метода суммирования, целью которого является уменьшение дисперсии статистической модели. Это достигается путем создания имитируемой вариативности данных путем взятия различных выборок начальной загрузки из одного обучающего набора и объединения прогнозов, как показано на Рисунке 2. Было показано, что этот метод улучшает качество прогнозирования многих методов обучения с наблюдением, в частности тех, которые относятся к генеалогическому древу и, в частности, тех, которые построены на сильно необрезанных деревьях. Конечным результатом является группа классификаторов с большой дисперсией и небольшим смещением [14].

Используется архитектура случайного леса, в которой категориальные переменные однократно кодируются при плотном выводе, а числовые объекты сохраняются в их первоначальном виде. Для этой модели было использовано 300 деревьев, максимальная глубина - 12, а размер листа - 3 (random_state=42). Этот дизайн направлен на то, чтобы охватить как можно больше нелинейной природы данных, насколько это возможно, и при этом свести к минимуму переобучение.

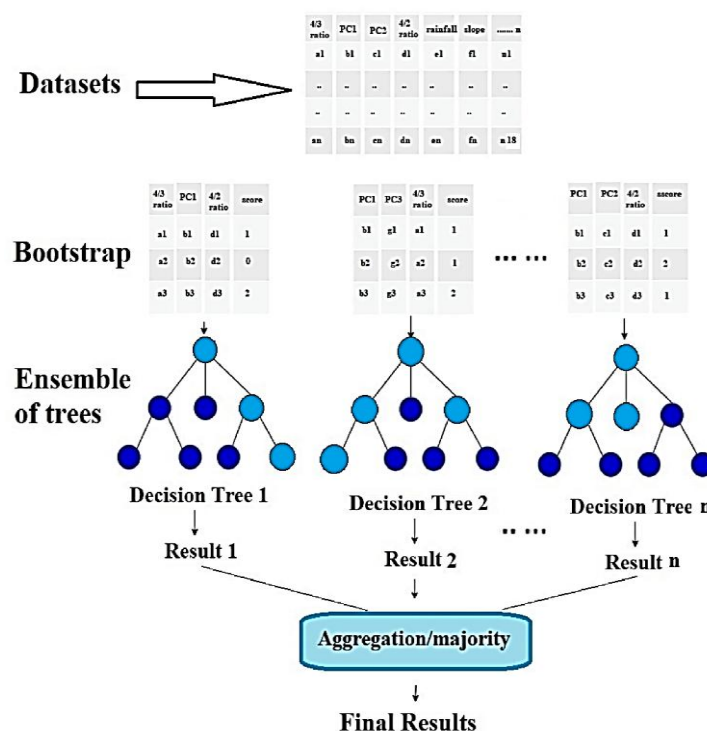


Рисунок 2. Структура Random Forest [15]

Метод AdaBoost - это один из методов машинного обучения, который направлен на улучшение его работы, поскольку он создает сильную классификацию путем суммирования линейной комбинации слабых классификаторов с учетом веса. Метод AdaBoost - это особый метод повышения эффективности, как показано на Рисунке 3. Природа слабых классификаторов оказывает большое влияние на эффективность результирующей

классификации. При выполнении процедуры был внесен ряд изменений в исходный алгоритм. AdaBoost является частным случаем метода групповой классификации. Объединение классификаторов осуществляется путем объединения различных типов классификаторов и методов прогнозирования. Таким образом, они повышают точность классификации. Базовые классификаторы формируются в процессе обучения. Классификация основана на голосовании [16].

Классификатор AdaBoost был создан с использованием того же метода предварительной обработки. Модель построила 300 оценок, используя скорость обучения 0,05 (random_state=42), что помогло классификатору модифицировать и сбалансировать слабых учащихся при переобучении в течение всего процесса обучения.

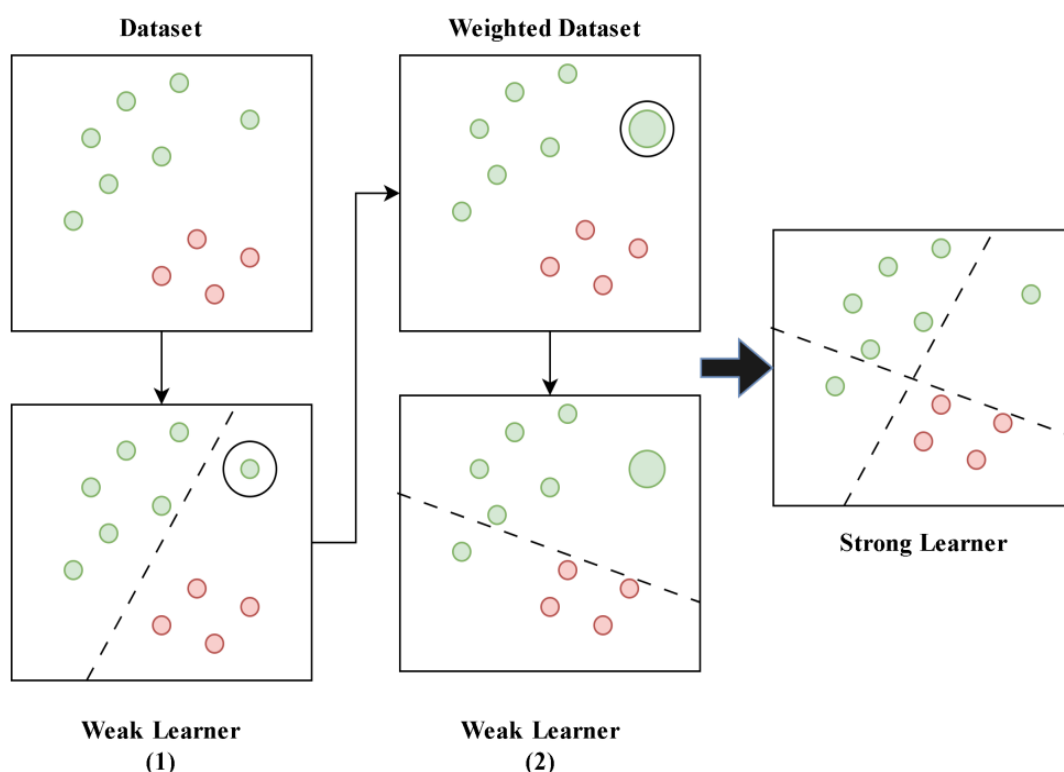


Рисунок 3. Структура Adaboost [17]

Экстремальное ускорение градиента (XGBoost). XGBoost - это высокоразвитая платформа для градиентного бустинга, разработанная для оптимизации эффективности, гибкости и общей простоты использования алгоритмов бустинга на основе градиента. Цель состоит в том, чтобы потенциально использовать современное аппаратное обеспечение для создания исключительно быстрой, масштабируемой и точной системы бустинга. XGBoost, усовершенствованная система регрессионных деревьев с градиентным ускорением (GBRT), представленная на рисунке, использует параллельное усиление дерева для решения наиболее сложных задач в области обработки данных [18].

Регрессор XGBoost был обучен при следующих параметрах: 500 деревьев, максимальная глубина 6, скорость обучения 0,05, подвыборка 0,8 и подвыборка по столбцам 0,8. Был оптимизирован целевой показатель квадратичной ошибки (случайное состояние = 42). В результате получается модель с широкими возможностями нелинейного моделирования и контролируемой доработкой на основе подвыборки.

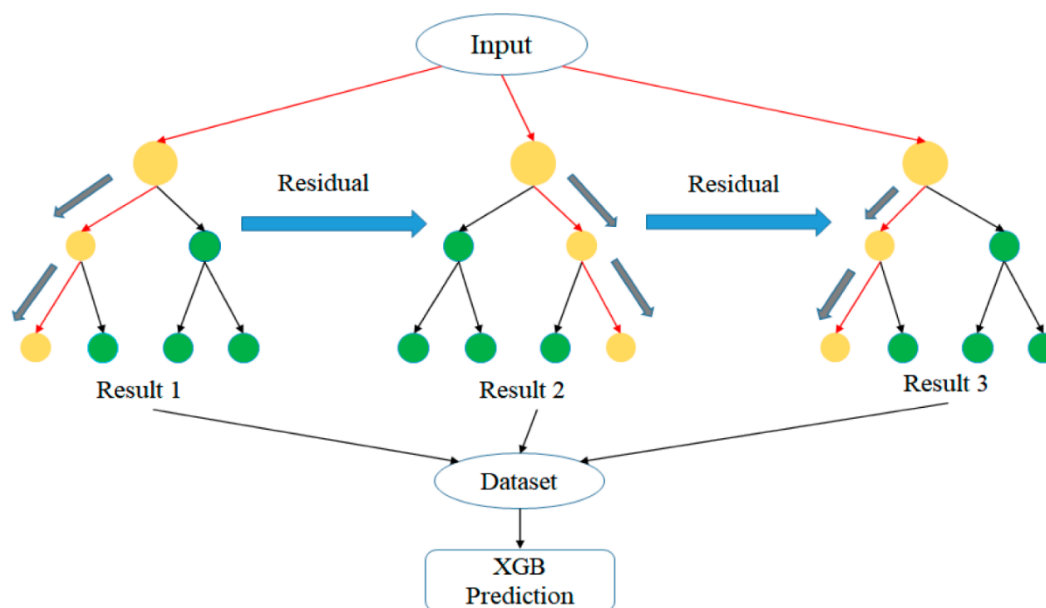


Рисунок 4. Концептуальное представление алгоритма XGBoost [19]

Показатели оценки. В регрессионном анализе различные оценочные показатели по-разному интерпретируют эффективность модели, некоторые из которых имеют более ограниченные математические выражения, чем другие. Средняя абсолютная ошибка (MAE) - это среднее значение общей абсолютной ошибки между прогнозируемыми и истинными значениями, как показано в уравнении (1).

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

где: n — количество наблюдений. y_i — фактические значения $\hat{y}_i$ — прогнозируемые значения.

Среднеквадратичная ошибка (MSE) возводит эти различия в квадрат, чтобы более строго наказывать за большие ошибки, как показано в уравнении (2):

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

Среднеквадратичная ошибка (RMSE) вычисляется как квадратный корень из MSE, возвращая единицы измерения к исходному масштабу, как показано в уравнении (3).

$$RMSE = \sqrt{MSE} \quad (3)$$

Между тем, показатель R-квадрат (R^2) количественно определяет долю дисперсии в зависимой переменной, объясняемой моделью, определенной в уравнении (4).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

где $\bar{y}$ — среднее значение наблюдаемых данных.

Хотя MAE и RMSE предлагают четкие интерпретации величины ошибки (RMSE придает больший вес более крупным ошибкам), R^2 указывает на сравнение модели с базовой моделью среднего уровня. MSE, хотя и является более неясным, имеет важное значение для оптимизации из-за своей плавности. Все эти показатели описывают производительность модели и объясняют, насколько она надежна и поддается обобщению.

Результаты

В следующем разделе приведены результаты работы трех моделей (базовые модели и запрошенные модели) в подготовленном тестовом наборе. Приведены показатели для количественной оценки точности и уровня объясняемой дисперсии для сравнения производительности каждой модели и применимости различных моделей.

Базовые показатели. Чтобы проиллюстрировать эффективность моделей, были созданы три базовые модели: глобальная медианная модель, медианная модель категории и медианная модель размера категории. Каждая из этих моделей обеспечивает низкий порог эффективности модели за счет использования непараметрических структур для данных возрастающей сложности. Глобальная медианная модель - это базовая постоянная гипотеза, медиана модели категории - это базовое представление о неоднородности продукта, а медиана размера категории - это модель базового представления о неоднородности продукта, основанная на категории и размере продукта. В то время как глобальная медианная модель, медиана модели категории и медиана размера категории достигают показателей MAE, RMSE и R2 для (216,15, 77,192,76, -0,0182), (157,26, 47,309,41, 0,3760), и (151,54, 44,442,64, 0,4138) соответственно, на в прошедшем тестировании другие модели показали гораздо худшие результаты. Сравнение базовых моделей (простых медианных моделей) направлено на то, чтобы предоставить доказательства эффективности продвинутых моделей ремоделирования. Все представленные модели работают значительно лучше, чем любые базовые модели и усовершенствованные модели модернизации, интегрирующие функции машинного обучения.

Результаты регрессии случайных лесов. Случайный лес имел значение MAE 130,32, RMSE 31,974,10 и R2 0,5783 в тестовом наборе. Модель демонстрирует приемлемую производительность при умеренно точных прогнозах и меньшей дисперсии ошибок прогнозирования по сравнению с базовой моделью.

Результаты регрессии AdaBoost. Результаты AdaBoost с коэффициентом полезного действия 176,29, RMSE 47,657,94 и R2 0,3714 не дотягивают до показателей Random Forest и XGBoost. По сути, использование бустинга для слабых учащихся для конкретного набора данных о транзакциях в этом исследовании недостаточно эффективно, учитывая его неоднородную структуру.

Результаты регрессии XGBoost. При использовании XGBoost показатели MAE, RMSE и R2, равные 93,41, 15,084,84 и 0,8010, соответственно, позволяют сделать вывод о том, что эта модель на сегодняшний день является более совершенной по сравнению с ее альтернативами и базовой моделью, благодаря значительным улучшениям в прогнозировании, которые она обеспечивает.

Обсуждение результатов. Базовые контрольные показатели указывают на то, что использование единой эвристики в отношении сумм транзакций позволяет прогнозировать только общий уровень. Например, использование глобальной медианной базовой модели показывает особенно низкие результаты (MAE = 216,15, RMSE = 77,192,76 и R2 = -0,0182) по сравнению с моделями для категории ~ и размера категории ~ с R2 = 0,3760 и 0,4138, соответственно, и с MAE = 157,26 и 151,54. Таким образом, базовые модели указывают на неоднородные суммы транзакций, при этом характеристики продукта являются значительной частью прогнозируемой дисперсии.

Основываясь как на исходных данных, так и на других методах обучения ансамблю, XGBoost добился наилучших результатов (MAE = 93,41; RMSE = 15,084,84; R2 = 0,8010) среди обученных моделей. По сравнению с базовыми показателями и другими методами, разработанными с использованием обученного ансамбля, XGBoost превзошел все из них. Как указано в соответствующей таблице работ, XGBoost также показал лучшие результаты среди

конкурирующих наборов данных для розничной торговли и электронной коммерции с различными функциями. XGBoost показал лучшие результаты среди прогнозов продаж Walmart, транзакций электронной коммерции в Китае и прогнозов продаж в магазинах товаров. Преимущество модели, в целом, объясняется ее способностью изучать нелинейные отношения и эффективно упорядочивать работу с гетерогенными пространствами признаков.

Случайный лес добился лишь незначительного улучшения производительности ($MAE = 130,32$; $RMSE = 31,974,10$; $R^2 = 0,5783$), но остался ниже XGBoost, что согласуется с предыдущими результатами, где RF показал конкурентоспособные показатели, но остался ниже методов повышения градиента. AdaBoost показал самые слабые результаты среди обученных моделей ($MAE = 176,29$; $RMSE = 47,657,94$; $R^2 = 0,3714$). AdaBoost не только не смог достичь наилучших результатов среди базовых показателей категории, но и показал самые слабые результаты среди базовых показателей категории + размер. Это доказывает, что фреймворк AdaBoost, в котором доминируют слабые учащиеся, плохо вписывается в закодированные транзакционные функции, которые имеют все размеры, разрежены и гетерогенны.

В целом, результаты подтверждают результаты ранее опубликованной литературы и подтверждают, что XGBoost является наиболее надежной моделью для структурированного прогнозирования розничных транзакций, особенно там, где категориальные и временные элементы взаимодействуют нелинейно. Таким образом, высокая производительность, достигнутая здесь, подтверждает пригодность модели для этого набора данных, а также помогает подтвердить выводы об обобщаемости преимуществ XGBoost, приведенных в исследованиях электронной коммерции и розничной торговли.

Заключение

В статье демонстрируется платформа машинного обучения прогнозированию продаж, использующая набор данных о продажах Amazon, в которой рассматриваются методы коллективного обучения, включающие Random Forest, AdaBoost и XGBoost. Применяя разработку временных характеристик в сочетании с предварительной обработкой и оценкой в хронологическом порядке, результаты показывают, что XGBoost превзошел остальные с коэффициентом полезного действия 0,8010. Тщательный анализ доступной литературы по табличному прогнозированию электронной коммерции подтвердил, что tabular XGBoost отлично справляется с этой задачей и является наиболее универсальной моделью. Полученные результаты также подтверждают, что модели повышения градиента позволяют изучать и прогнозировать нелинейные и сложные взаимосвязи в данных о крупномасштабных транзакциях электронной коммерции. Эта система применима в сфере онлайн-торговли и как таковая открывает новые возможности в области электронной коммерции для принятия управленческих решений на основе данных, а также в качестве прогностической модели, помогающей определить потребности в запасах и разработать надежные маркетинговые планы. Кроме того, модель может быть адаптирована таким образом, чтобы учитывать будущие сезоны, центральный план или ожидания, отдаленные социально-экономические переменные и даже отзывы клиентов и современные коммуникации.

Список литературы:

1. Laudon K. C., Traver C. G. E-Commerce 2018: business, technology, society, v. 14. 2019.
2. Nosratabadi S., Mosavi A., Duan P., Ghamisi P., Filip F., Band S. S., Gandomi A. H. Data science in economics: comprehensive review of advanced machine learning and deep learning methods // Mathematics. 2020. V. 8. №10. P. 1799. <https://doi.org/10.3390/math8101799>

3. Song X., Yang S., Huang Z., Huang T. The application of artificial intelligence in electronic commerce // *Journal of Physics: Conference Series*. IOP Publishing, 2019. V. 1302. №3. P. 032030. <https://doi.org/10.1088/1742-6596/1302/3/032030>
4. Srivastava A., Bala P. K., Kumar B. New perspectives on gray sheep behavior in E-commerce recommendations // *Journal of Retailing and Consumer Services*. 2020. V. 53. P. 101764. <https://doi.org/10.1016/j.jretconser.2019.02.018>
5. Al-Ebrahim M. A., Bunian S., Nour A. A. Recent machine-learning-driven developments in e-commerce: Current challenges and future perspectives // *Engineered Science*. 2023. V. 28. P. 1044. <https://doi.org/10.30919/es1044>
6. Zhou M., Zhao L., Kong N., Campy K. S., Xu G., Zhu G., Wang S. Understanding consumers' behavior to adopt self-service parcel services for last-mile delivery // *Journal of Retailing and Consumer Services*. 2020. V. 52. P. 101911. <https://doi.org/10.1016/j.jretconser.2019.101911>
7. Richard F., Štefan K., Lenka K. Role of artificial intelligence and machine learning in e-commerce: a literature review // *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 2025. V. 14. P. e31736-e31736. <https://doi.org/10.14201/adcaij.31736>
8. He X., Liu Y. Knowledge evolutionary process of Artificial intelligence in E-commerce: Main path analysis and science mapping analysis // *Expert Systems with Applications*. 2024. V. 238. P. 121801. <https://doi.org/10.1016/j.eswa.2023.121801>
9. Cubric M., Li F. Bridging the 'Concept-Product' gap in new product development: Emerging insights from the application of artificial intelligence in FinTech SMEs // *Technovation*. 2024. V. 134. P. 103017. <https://doi.org/10.1016/j.technovation.2024.103017>
10. Sun L. A comparative study of traditional and machine learning approaches for e-commerce sales forecasting // *Adv. Econ. Manag. Polit. Sci.* 2024. V. 135. №1. P. 46-51. <https://doi.org/10.54254/2754-1169/2024.18610>
11. Yang T. Sales prediction of walmart sales based on ols, random forest, and xgboost models // *Highlights Sci. Eng. Technol.* 2023. V. 49. P. 244-249.
12. Sarisa M., Boddapati V. N., Patra G. K., Kuraku C., Konkimalla S., Rajaram S. K. An Effective Predicting E-Commerce Sales & Management System Based on Machine Learning Methods // *Journal of Artificial Intelligence and Big Data*. 2020. V. 1. №1. P. 75-85. <https://doi.org/10.31586/jaibd.2020.1110>
13. Swami D., Shah A. D., Ray S. K. B. Predicting future sales of retail products using machine learning // *arXiv preprint arXiv:2008.07779*. 2020.
14. Aria M., Cuccurullo C., Gnasso A. A comparison among interpretative proposals for Random Forests // *Machine Learning with Applications*. 2021. V. 6. P. 100094. <https://doi.org/10.1016/j.mlwa.2021.100094>
15. Madani A., Niyazi B. Groundwater potential mapping using remote sensing and random forest machine learning model: a case study from lower part of Wadi Yalamlam, Western Saudi Arabia // *Sustainability*. 2023. V. 15. №3. P. 2772. <https://doi.org/10.3390/su15032772>
16. Hornyák O., Iantovics L. B. AdaBoost algorithm could lead to weak results for data with certain characteristics // *Mathematics*. 2023. V. 11. №8. P. 1801. <https://doi.org/10.3390/math11081801>
17. Al-Abdaly N. M., Seno M. E., Thwaini M. A., Imran H., Ostrowski K. A., Furtak K. Advanced ensemble machine-learning models for predicting splitting tensile strength in silica fume-modified concrete // *Buildings*. 2024. V. 14. №12. P. 4054. <https://doi.org/10.3390/buildings14124054>

18. Hakkal S., Ait Lahcen A. XGBoost to enhance learner performance prediction // Computers and Education: Artificial Intelligence. 2024. V. 7. P. 100254. <https://doi.org/10.1016/j.caeai.2024.100254>
19. Wang C. C., Kuo P. H., Chen G. Y. Machine learning prediction of turning precision using optimized XGBoost model // Applied Sciences. 2022. V. 12. №15. P. 7739. <https://doi.org/10.3390/app12157739>

References:

1. Laudon, K. C., & Traver, C. G. (2019). E-Commerce 2018: business, technology, society, vol. 14.
2. Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, 8(10), 1799. <https://doi.org/10.3390/math8101799>
3. Song, X., Yang, S., Huang, Z., & Huang, T. (2019, August). The application of artificial intelligence in electronic commerce. In *Journal of Physics: Conference Series* (Vol. 1302, No. 3, p. 032030). IOP Publishing. <https://doi.org/10.1088/1742-6596/1302/3/032030>
4. Srivastava, A., Bala, P. K., & Kumar, B. (2020). New perspectives on gray sheep behavior in E-commerce recommendations. *Journal of Retailing and Consumer Services*, 53, 101764. <https://doi.org/10.1016/j.jretconser.2019.02.018>
5. Al-Ebrahim, M. A., Bunian, S., & Nour, A. A. (2023). Recent machine-learning-driven developments in e-commerce: Current challenges and future perspectives. *Engineered Science*, 28, 1044. <https://doi.org/10.30919/es1044>
6. Zhou, M., Zhao, L., Kong, N., Campy, K. S., Xu, G., Zhu, G., ... & Wang, S. (2020). Understanding consumers' behavior to adopt self-service parcel services for last-mile delivery. *Journal of Retailing and Consumer Services*, 52, 101911. <https://doi.org/10.1016/j.jretconser.2019.101911>
7. Richard, F., Štefan, K., & Lenka, K. (2025). Role of artificial intelligence and machine learning in e-commerce: a literature review. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 14, e31736-e31736. <https://doi.org/10.14201/adcaij.31736>
8. He, X., & Liu, Y. (2024). Knowledge evolutionary process of Artificial intelligence in E-commerce: Main path analysis and science mapping analysis. *Expert Systems with Applications*, 238, 121801. <https://doi.org/10.1016/j.eswa.2023.121801>
9. Cubric, M., & Li, F. (2024). Bridging the 'Concept-Product' gap in new product development: Emerging insights from the application of artificial intelligence in FinTech SMEs. *Technovation*, 134, 103017. <https://doi.org/10.1016/j.technovation.2024.103017>
10. Sun, L. (2024). A comparative study of traditional and machine learning approaches for e-commerce sales forecasting. *Adv. Econ. Manag. Polit. Sci*, 135(1), 46-51. <https://doi.org/10.54254/2754-1169/2024.18610>
11. Yang, T. (2023). Sales prediction of walmart sales based on ols, random forest, and xgboost models. *Highlights Sci. Eng. Technol*, 49, 244-249.
12. Sarisa, M., Boddapati, V. N., Patra, G. K., Kuraku, C., Konkimalla, S., & Rajaram, S. K. (2020). An Effective Predicting E-Commerce Sales & Management System Based on Machine Learning Methods. *Journal of Artificial Intelligence and Big Data*, 1(1), 75-85. <https://doi.org/10.31586/jaibd.2020.1110>
13. Swami, D., Shah, A. D., & Ray, S. K. (2020). Predicting future sales of retail products using machine learning. *arXiv preprint arXiv:2008.07779*.

14. Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. *Machine Learning with Applications*, 6, 100094. <https://doi.org/10.1016/j.mlwa.2021.100094>
15. Madani, A., & Niyazi, B. (2023). Groundwater potential mapping using remote sensing and random forest machine learning model: a case study from lower part of Wadi Yalamlam, Western Saudi Arabia. *Sustainability*, 15(3), 2772. <https://doi.org/10.3390/su15032772>
16. Hornyák, O., & Iantovics, L. B. (2023). AdaBoost algorithm could lead to weak results for data with certain characteristics. *Mathematics*, 11(8), 1801. <https://doi.org/10.3390/math11081801>
17. Al-Abdaly, N. M., Seno, M. E., Thwaini, M. A., Imran, H., Ostrowski, K. A., & Furtak, K. (2024). Advanced ensemble machine-learning models for predicting splitting tensile strength in silica fume-modified concrete. *Buildings*, 14(12), 4054. <https://doi.org/10.3390/buildings14124054>
18. Hakkal, S., & Ait Lahcen, A. (2024). XGBoost to enhance learner performance prediction. *Computers and Education: Artificial Intelligence*, 7, 100254. <https://doi.org/10.1016/j.caeai.2024.100254>
19. Wang, C. C., Kuo, P. H., & Chen, G. Y. (2022). Machine learning prediction of turning precision using optimized XGBoost model. *Applied Sciences*, 12(15), 7739. <https://doi.org/10.3390/app12157739>

Поступила в редакцию
09.05.2026 г.

Принята к публикации
16.05.2026 г.

Ссылка для цитирования:

Судаков В. А., Заммам А. Ансамблевые модели машинного обучения для прогнозирования розничных продаж в электронной коммерции // Бюллетень науки и практики. 2026. Т. 12. №7. С. 55-69. <https://doi.org/10.33619/2414-2948/128/06>

Cite as (APA):

Sudakov, V., & Zammam, A. (2026). Ensemble Machine Learning Models for Retail Sales Forecasting in E-Commerce. *Bulletin of Science and Practice*, 12(7), 55-69. (in Russian). <https://doi.org/10.33619/2414-2948/128/06>