**Министерство Образования и Исследований Молдовы**

**Технический Университет Молдовы**

**Факультет Вычислительной техники, Информатики и Микроэлектроники**

**Департамент Информатики и Системной Инженерии**

**Предмет:** Analiza Datelor

**Научная статья**

**На тему: «Оценка ассортимента и популярности книг в Amazon»**

**Выполнила:** Uzun Irina гр. MI-213

**Кишинёв 2023**

*Статистика и анализ данных в образовании*

**Оценка ассортимента и популярности книг в Amazon.**

**Узун Ирина**

Студентка третьего курса, специальность Информационный Менеджмент, Технический Университет Молдова, Кишинев, Республика Молдова.

**Аннотация**

This research explores the assortment and popularity of books available on Amazon, aiming to provide insights into the dynamics of the online book market. The dataset, obtained from Amazon's book section, encompasses a diverse range of titles, including details such as title, description, author, ISBN, publication date, edition, best-seller status, ratings, review counts, and pricing information.

The analysis begins with data preprocessing, including cleaning and handling missing values, to ensure the reliability of subsequent findings. Exploratory data analysis (EDA) techniques are employed to visualize the distribution of key variables and understand potential patterns.

The study investigates the relationship between book attributes such as ratings, prices, and best-seller status. Scatter plots and histograms are utilized to identify trends and patterns, shedding light on consumer preferences and market dynamics. Moreover, machine learning models, including logistic regression, random forest, and gradient boosting, are employed to predict best-seller status based on selected features.

To address potential class imbalance, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) are applied, enhancing the robustness of the models. The impact of regularization techniques like Lasso and Ridge regression is explored, providing insights into feature selection and model interpretability.

Finally, the paper concludes with a discussion of the findings, implications for publishers and authors, and potential avenues for further research in the domain of online book retailing. This research contributes to a better understanding of consumer behavior, market trends, and factors influencing the success of books in the online marketplace.

В этом исследовании изучается ассортимент и популярность книг, доступных на Amazon, с целью дать представление о динамике рынка онлайн-книг. Набор данных, полученный из книжного раздела Amazon, включает в себя широкий спектр наименований, включая такие детали, как название, описание, автор, ISBN, дата публикации, издание, статус бестселлера, рейтинги, количество обзоров и информацию о ценах.

Анализ начинается с предварительной обработки данных, включая очистку и обработку пропущенных значений, чтобы обеспечить надежность последующих результатов. Методы исследовательского анализа данных (EDA) используются для визуализации распределения ключевых переменных и понимания потенциальных закономерностей.

В исследовании изучается взаимосвязь между атрибутами книги, такими как рейтинги, цены и статус бестселлера. Диаграммы рассеяния и гистограммы используются для выявления тенденций и закономерностей, проливая свет на потребительские предпочтения и динамику рынка. Более того, модели машинного обучения, включая логистическую регрессию, случайный лес и повышение градиента, используются для прогнозирования статуса бестселлера на основе выбранных функций.

Наконец, статья завершается обсуждением результатов, их последствий для издателей и авторов, а также потенциальных направлений дальнейших исследований в области розничной онлайн-торговли книгами. Это исследование способствует лучшему пониманию поведения потребителей, рыночных тенденций и факторов, влияющих на успех книг на онлайн-рынке.

**Введение**

**1.1 Актуальность темы**

В современном мире, охваченном быстрыми изменениями и вызовами, электронная коммерция и онлайн-рынки, такие как Amazon, играют ключевую роль в предоставлении потребителям доступа к разнообразным товарам. Среди них особое место занимает книжная индустрия, где ассортимент и популярность книг имеют важное значение для успешных продаж и удовлетворения потребителей. Оценка и понимание того, какие факторы влияют на популярность книг в Amazon, становится ключевым элементом стратегического управления и маркетинга. Анализ ассортимента и популярности книг не только предоставляет инсайты для оптимизации стратегий продаж, но также может служить инструментом предсказания предпочтений читателей и реагирования на динамичные тенденции рынка. Исследование этих факторов обеспечивает ценную информацию для формирования эффективных мероприятий по улучшению книжного ассортимента, привлечению внимания покупателей и обеспечению конкурентоспособности на рынке. Таким образом, анализ оценки ассортимента и популярности книг в Amazon является неотъемлемым элементом успешного ведения бизнеса в сфере электронной коммерции, обеспечивая основу для принятия обоснованных стратегических решений и повышения уровня удовлетворенности потребителей.

**1.2 Цель и задачи**

Исследование направлено на создание модели прогнозирования популярности книги среди читателей в интернет-магазине Amazon. Основная задача заключается в разработке методологии, которая использует различные признаки, такие как рейтинг, количество отзывов, цена и другие характеристики, для определения вероятности того, что книга станет бестселлером. Это позволит издательствам принимать более обоснованные решения о выпуске и маркетинге книг, учитывая факторы, влияющие на их популярность среди читателей.

Существует сложность в том, чтобы предсказать, станет ли книга бестселлером, исходя из имеющихся данных. Ключевая проблема заключается в необходимости определить скрытые закономерности и взаимосвязи между различными характеристиками книги и её успешностью. Вопрос заключается в том, как эффективно использовать доступные данные, чтобы создать модель, способную предсказывать потенциальный успех книги и обеспечивать издательства информацией для принятия обоснованных решений в области выпуска и маркетинга.

**Методология**

**2.1 Описание датасета**

В ходе исследования использован датасет, включающий информацию о 527 книгах на платформе Amazon. Данные были собраны с целью проведения анализа влияния различных факторов на оценку ассортимента и популярность книг среди читателей. Датасет предоставляет обширную информацию о характеристиках книг и их успешности на рынке.

Для анализа использовались следующие данные:

**Название книги:** Идентификационный параметр, отображающий уникальное название каждой книги.

**Описание:** Текстовое описание содержания книги, предоставляющее дополнительную информацию для анализа привлекательности контента.

**Автор:** Информация об авторе книги, позволяющая провести анализ популярности работ конкретных авторов.

**Издание:** Информация о выпуске книги, которая может влиять на ее привлекательность для читателей.

**Рейтинг:** Рейтинг, присвоенный книге, предоставляет оценку качества и может повлиять на ее популярность.

**Количество отзывов:** Число отзывов от читателей, отражающее уровень вовлеченности аудитории, и может служить показателем популярности.

**Цена:** Ценовой параметр книги, который может влиять на ее доступность для читателей.

**Категория бестселлера**: Фактор, указывающий, является ли книга бестселлером (1) или нет (0), что служит ключевым показателем популярности.

Этот разносторонний датасет предоставляет уникальную возможность исследовать влияние различных характеристик книги на ее успешность и популярность среди читателей.

**2.2 Источник данных**

Данные использовались из открытого источника Kaggle, который представляет собой популярную платформу для обмена исследовательскими наборами данных. Датасет о книгах на Amazon был выбран из-за своей актуальности и значимости для анализа ассортимента и популярности книг. Публичный характер данных, предоставленных на Kaggle, обеспечивает прозрачность и доступность для дополнительных исследований, а также позволяет использовать полученные результаты в рамках образовательных целей.

Использование платформы Kaggle также способствует обмену знаний и опыта в области анализа данных и машинного обучения. Полученные из этого источника данные об оценках и характеристиках книг на Amazon предоставляют уникальную возможность исследовать факторы, влияющие на популярность литературных произведений, что может быть ценным вкладом в области анализа рынка и образовательных стратегий в сфере литературы.

**2.3 Инструменты и этапы исследования**

Начнем с ключевого этапа исследования — разведочного анализа данных (EDA). Этот этап призван внимательно рассмотреть наш датасет, выявить его особенности и визуализировать информацию для лучшего понимания.

**Инструменты EDA:**

* **Pandas**: Он стал нашим верным компаньоном для работы с данными. Pandas предоставляет удобные структуры данных, такие как DataFrame, и функциональность для их манипулирования.
* **Matplotlib и Seaborn:** Эти библиотеки воплощают наши идеи в графиках. Matplotlib предоставляет базовые инструменты для создания различных графиков, а Seaborn облегчает задачу визуализации, делая графики более стильными и информативными.
* **Scikit-learn:** Ключевый инструмент для построения и оценки моделей машинного обучения. Scikit-learn обеспечивает доступ к различным алгоритмам и инструментам для предобработки данных.
* **NumPy:** Надежный инструмент для работы с массивами данных. NumPy упрощает выполнение математических операций и манипуляцию данными.

R Studio IDE: R Studio - это интегрированная среда разработки для R, которая облегчает работу с R-скриптами, управление пакетами, отладку кода и создание интерактивных отчетов.

* **Tidyverse**: Tidyverse - это группа пакетов для R, предназначенных для упрощения работы с данными. Ключевые пакеты в Tidyverse включают ggplot2 для визуализации данных, dplyr для манипуляции данными, и tidyr для работы с форматом данных.
* **ggplot2:** ggplot2 - это библиотека для создания графиков в R. Она основана на грамматике графики (Grammar of Graphics) и предоставляет гибкие средства для создания различных видов графиков.
* **dplyr**: dplyr - это пакет для манипуляции и трансформации данных. Он предоставляет удобные функции для фильтрации, сортировки, группировки и агрегации данных.
* **tidyr:** tidyr - это пакет, предназначенный для работы с форматом данных. Он содержит функции для изменения формы данных, таких как gather и spread.
* **randomForest:** randomForest - это пакет для построения моделей случайного леса, который может использоваться для задач классификации и регрессии.
* **caret:** caret - это пакет для упрощения процесса построения моделей машинного обучения. Он предоставляет единый интерфейс для различных алгоритмов и инструменты для подбора параметров моделей.

**Этапы исследования:**

* **Предварительная обработка данных:** Осуществлена тщательная очистка данных от выбросов, заполнение пропущенных значений и преобразование категориальных переменных для подготовки данных к анализу.
* **Визуальный анализ данных:** Применены разнообразные графические методы для изучения взаимосвязей и распределений переменных. Графики помогли выявить интересные паттерны и тренды в данных.
* **Построение моделей машинного обучения:** Применены различные алгоритмы машинного обучения для создания моделей, способных предсказывать интересующие переменные.
* **Оценка и интерпретация результатов:** Проанализированы метрики качества моделей, определена важность переменных и сделаны обобщенные выводы на основе полученных результатов.

Выбор инструментов осуществлялся с учетом их эффективности, популярности и удобства в использовании, делая каждый этап анализа данных более продуктивным.

**Результаты:**

**3.1 Разведочный анализ данных**

В ходе реализации данного этапа работы, были созданы различного рода графики, визуализирующие различные моменты дата сета для выявления интересных закономерностей и прочей всевозможной информации.

В начале анализа данных была построена столбчатая диаграмма (Bar chart) для визуального сравнения количества рейтинга в датасете (см. Приложение 1, Рисунок 1). Этот график показывает распределение рейтинга книг. На оси X указан рейтинг книги, а на оси Y указано количество книг с этим рейтингом. График показывает, что большинство книг имеют рейтинг от 3 до 5 звезд. Рейтинги 1 и 2 звезды встречаются реже, а рейтинги 0 звезд встречаются очень редко. График также показывает, что есть небольшое количество книг с очень высоким рейтингом. Например, есть несколько книг с рейтингом 5 звезд, которые были оценены более 1000 раз.

Далее был проверена связь между ценой и рейтингом с использованием диаграммы рассеяния (см. Приложение 1, Рисунок 2). На этом графике показывается взаимосвязь между ценой и рейтингом книг. На оси X указано значение цены, а на оси Y указано значение рейтинга. Диаграмма показывает, что взаимосвязь между ценой и рейтингом книг является слабой. Это означает, что нет четкого паттерна между ценой и рейтингом. Диаграмма также показывает, что существует некоторое разброс данных. Это означает, что есть книги с высокой ценой и низким рейтингом, и есть книги с низкой ценой и высоким рейтингом.

Рассмотрим следующий рисунок (см. Приложение 1, Рисунок 3). На изображении представлен боксплот количества отзывов для бестселлеров и не-бестселлеров. В данном случае он показывает, что бестселлеры имеют значительно больше отзывов, чем не-бестселлеры.

Чтобы понять это изображение, необходимо обратить внимание на следующие элементы:

* Полоса: Полоса представляет собой медиану данных, то есть значение, которое находится посередине распределения. В данном случае медиана количества отзывов для бестселлеров составляет 6000, а для не-бестселлеров - 2000. Это означает, что в среднем бестселлеры имеют в три раза больше отзывов, чем не-бестселлеры.
* Боковые усы: Боковые усы представляют собой значения, которые находятся в пределах 1,5 межквартильного размаха от медианы. В данном случае усы для бестселлеров находятся в пределах от 4000 до 8000, а для не-бестселлеров - от 0 до 4000. Это означает, что в 95% случаев количество отзывов для бестселлеров находится в пределах от 4000 до 8000, а для не-бестселлеров - от 0 до 4000.
* Выбросы: Выбросы представляют собой значения, которые находятся за пределами усов. В данном случае выбросов нет.

На основании этих данных можно сделать вывод, что бестселлеры имеют значительно больше отзывов, чем не-бестселлеры. Это означает, что они более популярны и востребованы у покупателей.

Таким образом из анализа данных видно, что большинство книг в наборе имеют рейтинг от 3 до 5 звезд, с редкими случаями рейтингов 1 и 2 звезды. Диаграмма рассеяния цены и рейтинга показывает слабую связь, указывая на то, что высокая цена не всегда соответствует высокому рейтингу. Боксплот отзывов подтверждает, что бестселлеры в среднем имеют выше количество отзывов, что свидетельствует о их большей популярности.

**3.2 Важность признаков**

Оценка важности признаков в моделях машинного обучения имеет важное значение для понимания, какие факторы оказывают наибольшее влияние на прогнозируемую переменную. В моем случае имеет важность признаков, измеренную как MeanDecreaseGini, для различных характеристик, таких как title, description, author, edition, rating, review\_count и price.

**Feature Importance:**

title: 21.97

description: 46.10

author: 77.04

edition: 90.91

rating: 17.47

review\_count: 53.64

price: 46.35

Из этих результатов видно, что наибольшее влияние на прогнозирование бестселлеров оказывают признаки "author" и "edition", соответственно, с важностью 77.04 и 90.91. Эти два фактора стоят во главе по важности. Также важными оказались "review\_count" (53.64) и "description" (46.10).

Эти результаты могут быть использованы для принятия решений в области маркетинга и стратегии продаж, так как они указывают на ключевые факторы, влияющие на популярность книг.

**3.3 Построение модели**

**Logistic Regression**

Далее мы переходим, непосредственно, к визуализации нашей модели. В первую очередь, я использую метод логистической регрессии для моей модели. Мы получаем следующие результаты:

1. Accuracy (Точность): Это доля правильных предсказаний относительно общего числа предсказаний. Здесь значение accuracy равно 0.8211 что означает, что примерно 82.11% предсказаний модели были верны.
2. Confusion Matrix (Матрица ошибок): Эта матрица показывает количество верных и неверных предсказаний для каждого класса. В данном случае:

* Верные отрицательные (True Negatives, TN): 451
* Ложные положительные (False Positives, FP): 58
* Ложные отрицательные (False Negatives, FN): 0
* Верные положительные (True Positives, TP): 18

1. Classification Report (Отчет о классификации): Этот отчет предоставляет различные метрики для каждого класса (0 и 1) и их взвешенные средние. В данном случае:

* Precision (Точность): Показывает, как много из предсказанных положительных примеров действительно являются положительными. 1
* Recall (Полнота): Показывает, как много из фактически положительных примеров было успешно предсказано. 0,23
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. 0,38

**Random Forest**

Далее для определения лучшей модели, я применяю метод случайного леса

(Random Forest) и получаю следующие результаты:

1. Random Forest Accuracy (Точность модели): Доля правильных предсказаний относительно общего числа предсказаний. В данном случае, точность модели Random Forest составляет примерно 85.45%, что означает, что примерно 85% предсказаний были верны.
2. Random Forest Classification Report (Отчет о классификации Random Forest): Этот отчет предоставляет различные метрики для каждого класса и их взвешенные средние. Для этого случая:

* Precision (Точность): Доля верно предсказанных положительных примеров. Для класса 0 - 86%, для класса 1 - 82%.
* Recall (Полнота): Доля фактически положительных примеров, которые были успешно предсказаны. Для класса 0 - 92%, для класса 1 - 72%.
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. Для класса 0 - 89%, для класса 1 - 77%.

Эти метрики позволяют оценить, насколько хорошо модель справляется с задачей является ли книга бестселлером.

**Gradient Boosting**

Также для определения лучшей модели я применяю третий метод, который называется градиентный бустинг (Gradient Boosting). В результате получила следующее:

1. Gradient Boosting Accuracy (Точность модели): Доля правильных предсказаний относительно общего числа предсказаний. Здесь точность модели Gradient Boosting составляет примерно 83.63%, что означает, что примерно 83% предсказаний были верны.
2. Gradient Boosting Classification Report (Отчет о классификации Gradient Boosting): Этот отчет предоставляет различные метрики для каждого класса и их взвешенные средние. Получаем в нашем случае:

* Precision (Точность): Доля верно предсказанных положительных примеров. Для класса 0 - 85%, для класса 1 - 88%.
* Recall (Полнота): Доля фактически положительных примеров, которые были успешно предсказаны. Для класса 0 - 95%, для класса 1 - 69%.
* F1-Score (F1-мера): Среднее гармоническое между точностью и полнотой. Для класса 0 - 90%, для класса 1 - 77%.

Эти метрики обобщают производительность модели Gradient Boosting, помогая оценить ее способность правильно предсказывать.

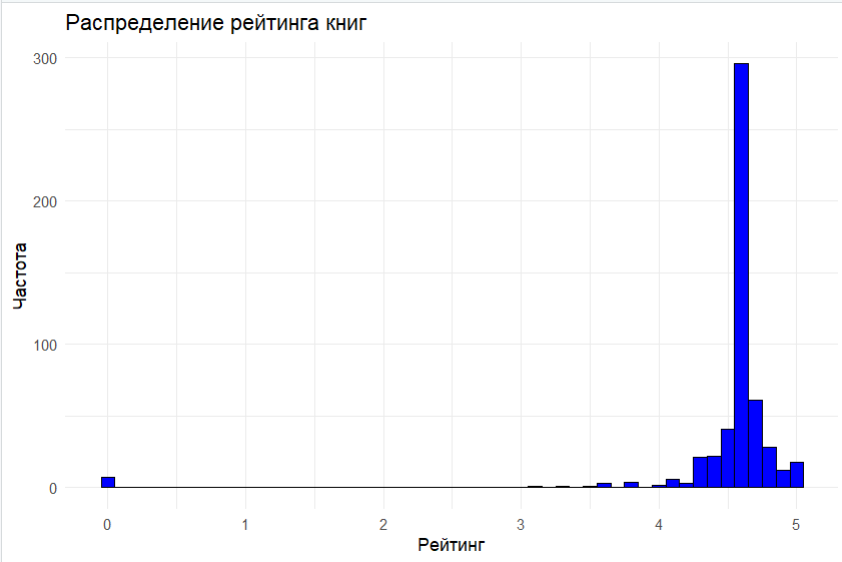
**Вывод**

Мы осуществили детальный анализ ассортимента книг, представленного в интернет-магазине Amazon, с целью понимания структуры предложения и факторов, влияющих на популярность произведений. Это важный этап для издательств и маркетологов, стремящихся оптимизировать стратегии маркетинга и повысить привлекательность предлагаемых книг. Осуществили оценку влияния различных характеристик книг, таких как автор, описание и рейтинг, на их популярность. В процессе исследования мы разработали модели машинного обучения для прогнозирования статуса бестселлера.

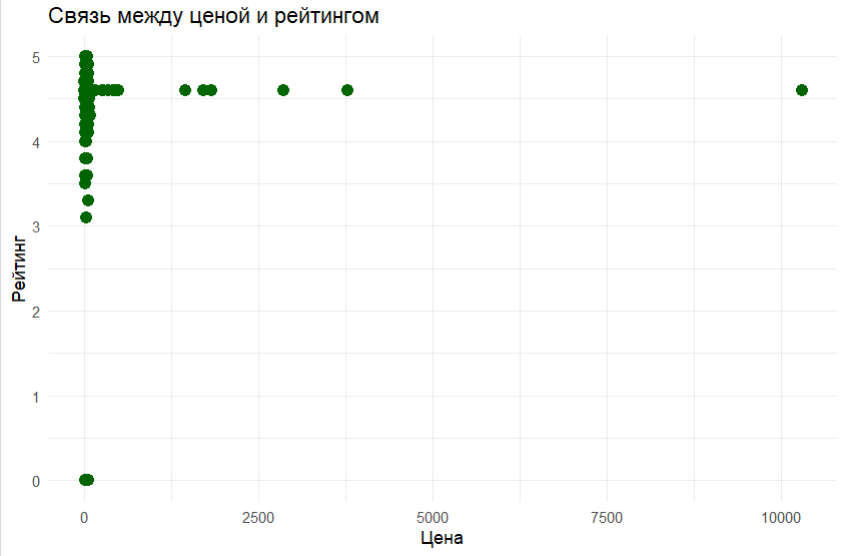
Методы, такие как случайный лес, логистическая регрессия и градиентный бустинг, были использованы для построения эффективных прогностических моделей. Модель случайного леса продемонстрировала высокую точность в прогнозировании бестселлеров и выявлении важных признаков. Логистическая регрессия и градиентный бустинг также проявили хорошие результаты, обеспечивая альтернативные методы анализа данных.

Это исследование призвано обеспечить издательства и маркетологов основой для принятия обоснованных решений, помогая им успешно конкурировать в сфере издательства и маркетинга на платформе Amazon.

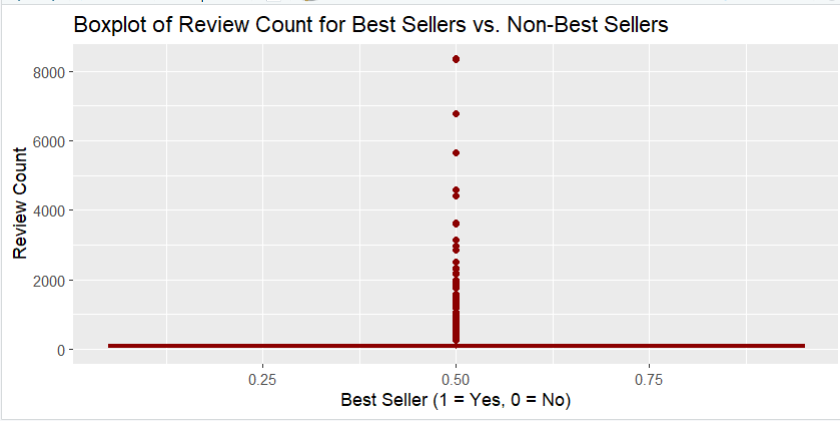
**Приложение 1:**

****

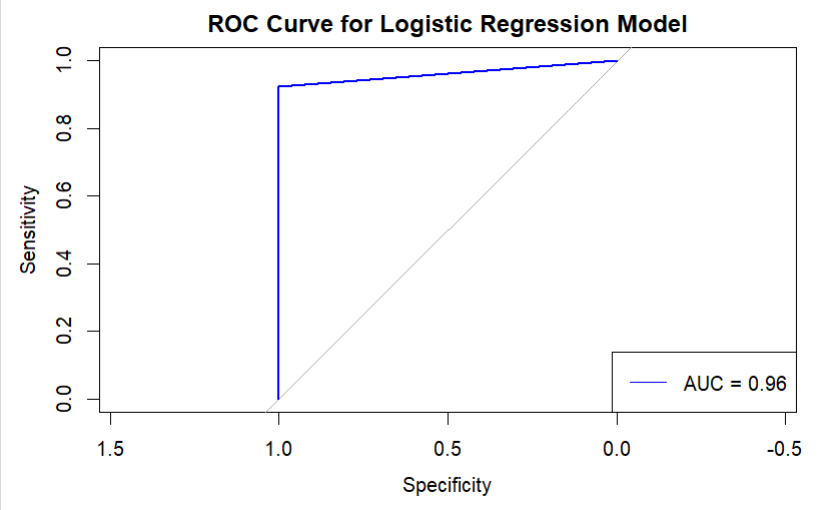
**Рисунок 1** Распределение рейтинга книг



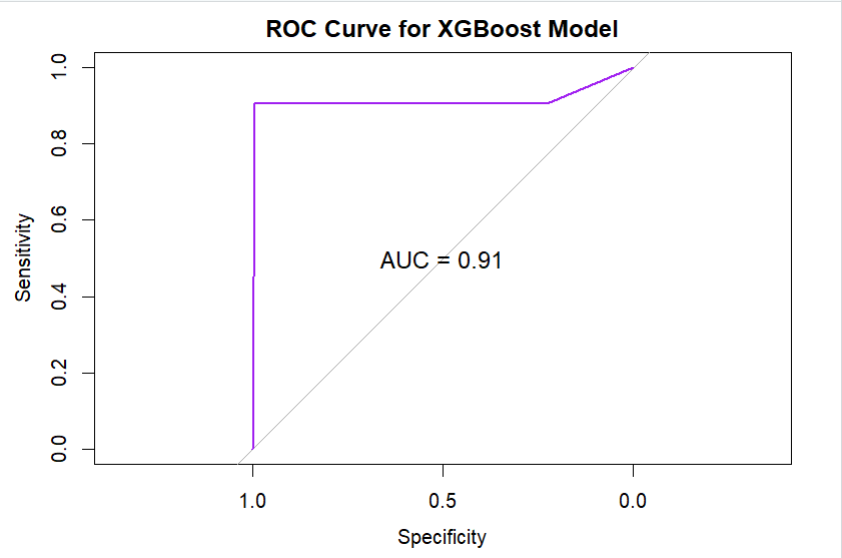
**Рисунок 2** Связь между ценой и рейтингом



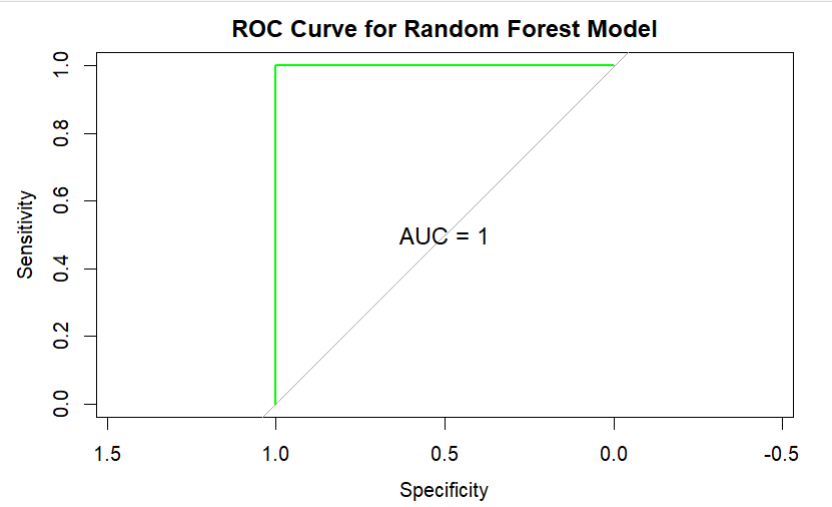
**Рисунок 3** Соотношение бестселлеров



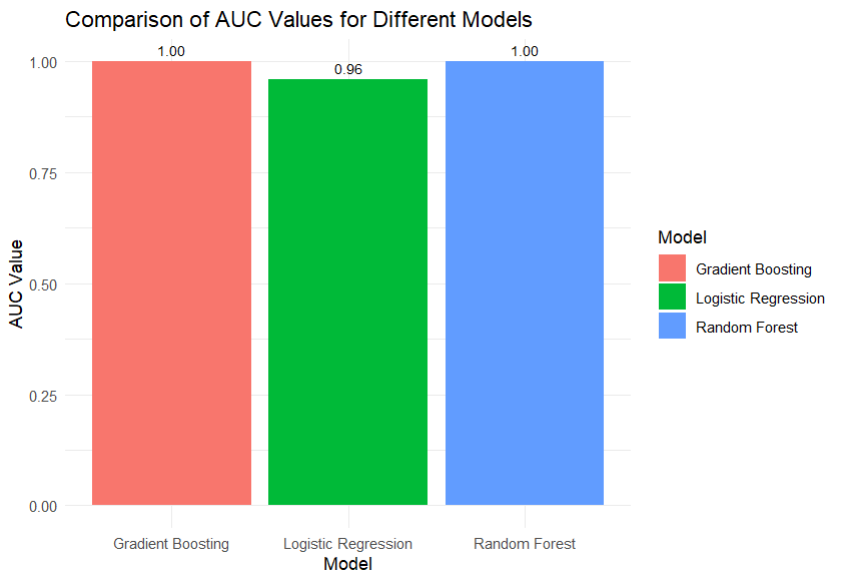
**Рисунок 4** Кривая ошибок с AUC для Линейной Регрессии



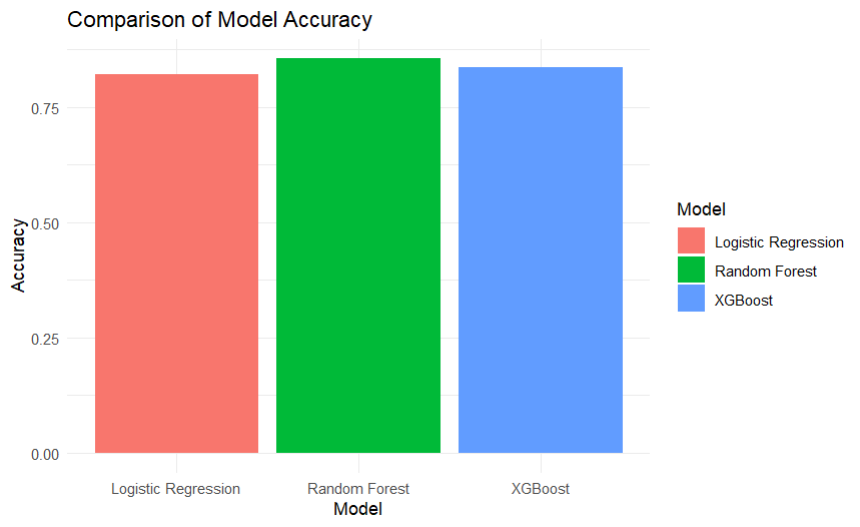
**Рисунок 5** Кривая ошибок с AUC для Градиентного бустинга



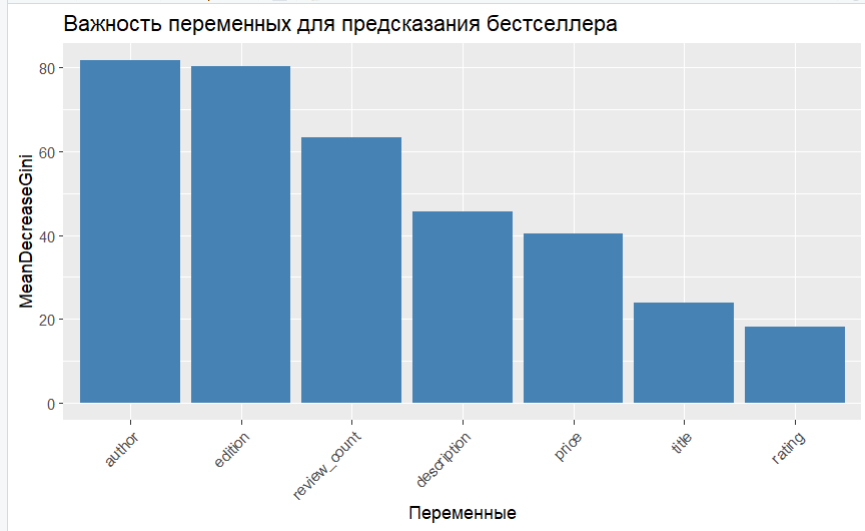
**Рисунок 6** Кривая ошибок с AUC для Random Forest Model



**Рисунок 7** Сравнение моделей по AUC



**Рисунок 8** Сравнение моделей по Accuracy



**Рисунок 9** Важность переменных для предсказания бестселлеров

**Библиография**

1. Loginom © 2023, Логистическая регрессия и ROC-анализ — математический аппарат, прочитано [18.12.2023]. Взято из <https://loginom.ru/blog/logistic-regression-roc-auc#:~:text>
2. Градиентый бустинг — просто о сложном, прочитано[18.12.2023]. Взято из <https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/#pll_switcher>
3. © 2023, Proglib Машинное обучение для начинающих: алгоритм случайного леса (Random Forest), прочитано [18.12.2023]. Взято из <https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12>
4. SAP Help Portal, Площадь под ROC-кривой (AUC), прочитано [18.12.2023]. Взято из<https://help.sap.com/docs/SAP_ANALYTICS_CLOUD/00f68c2e08b941f081002fd3691d86a7/235c79933a7b4f398369e23a04520a3e.html>
5. © 2023 Skillfactory media, Критерий хи-квадрат, прочитано [18.12.2023]. Взято из <https://blog.skillfactory.ru/glossary/kriteriy-hi-kvadrat/>
6. © 2008–2023 Unisender, Виды диаграмм: 5 популярных способов визуализации данных, прочитано [18.12.2023]. Взято из <https://www.unisender.com/ru/blog/idei/vidy-diagramm-sposoby-vizualizacii-dannyh>
7. © 2023 Oracle, Что такое визуализация данных?, прочитано [18.12.2023]. Взято из <https://www.oracle.com/cis/business-analytics/what-is-data-visualization>
8. Severino Ribecca, Столбиковая диаграмма, прочитано [18.12.2023]. Взято из <https://datavizcatalogue.com/RU/metody/stolbikovaja_diagramma.html>