AMAZON BEST-SELLER RANK PREDICTION USING MEACHINE LEARNING

WAN ZUKI AZMAN WAN MUHAMAD

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/23)

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name :

Student's Matric No. : Academic Session :

Date of Birth : UTM Email :

Choose an item. Title : TITLE IN CAPITAL LETTERS
TITLE IN CAPITAL LETTERS
TITLE IN CAPITAL LETTERS

I declare that this Choose an item. is classified as:

☒ OPEN ACCESS I agree that my report to be published as a hard copy or made available through online open access.

RESTRICTED Contains restricted information as specified by the organization/ institution where research was done.
(The library will block access for up to three (3) years)
□

□ CONFIDENTIAL Contains confidential information as specified in the Official Secret Act 1972)
(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name
Date :

Approved by Supervisor(s)

Signature of Supervisor I:

Signature of Supervisor II

Full Name of Supervisor I
NOOR HAZARINA HASHIM
Full Name of Supervisor II
MOHD ZULI JAAFAR

Date : Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

“Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item. opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item.”

Signature : ________________________________

Name of Supervisor I : KHAIRUR RIJAL JAMALUDIN

Date : 9 MAY 2017

Signature : ________________________________

Name of Supervisor II : NOOR HAZARINA HASHIM

Date : 9 MAY 2017

Signature : ________________________________

Name of Supervisor III : MOHD ZULI JAAFAR

Date : 9 MAY 2017

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration Click or tap here to enter text. and Click or tap here to enter text.

Certified by:

Signature :

Name :

Position :

Official Stamp

Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar :

Nama dan Alamat Pemeriksa Dalam :

Nama Penyelia Lain (jika ada) :

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

ON-LINE RECOGNITION OF DEVELOPING CONTROL CHART PATTERNS

TITLE

TITLE

TITLE

WAN ZUKI AZMAN WAN MUHAMAD

A Choose an item. submitted in Choose an item. of the
requirements for the award of the degree of
Choose an item.

School of Education
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia

OCTOBER 2024

DECLARATION

I declare that this Choose an item. entitled *"title of the thesis"* is the result of my own research except as cited in the references. The Choose an item. has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : ....

Name :

Date : 10 NOVEMBER 2016

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor Professor Dr Awaluddin Mohd Shaharoun and Associate Professor Dr. Hishamuddin Jamaluddin for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Ph.D study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

This study explores the application of supervised machine learning models to predict Amazon Best Seller Rank (BSR) using structured product-level data. By collecting real-time information from Amazon's software category, the research investigates how factors such as product price, star ratings, review volume, and geographic marketplace influence a product's sales rank. Following comprehensive data cleaning and exploratory analysis, feature engineering was performed to enhance predictive power, introducing variables like review density, weighted rating, and log-transformed review counts.

Three regression models—Linear Regression, Decision Tree, and Random Forest—were developed and evaluated using standard performance metrics including R^2, RMSE, and MAE. Among them, the Random Forest Regressor achieved the highest predictive accuracy, explaining over 57% of the variance in BSR and demonstrating strong generalisation in cross-validation. Feature importance analysis further revealed that review behaviour and pricing are key predictors of best seller status.

The findings support the value of ensemble learning and engineered features in modelling complex, non-linear outcomes in e-commerce settings. This research contributes both theoretically, by validating advanced modelling approaches, and practically, by offering a replicable framework that sellers and analysts can use to forecast product performance and inform strategic decisions.

ABSTRAK

Kajian ini meneroka penggunaan model pembelajaran mesin terselia untuk meramalkan Kedudukan Produk Terlaris Amazon (Amazon Best Seller Rank, BSR) dengan menggunakan data berstruktur pada peringkat produk. Dengan mengumpul maklumat masa nyata daripada kategori perisian Amazon, penyelidikan ini menilai bagaimana faktor seperti harga produk, penarafan bintang, jumlah ulasan, dan pasaran geografi mempengaruhi kedudukan jualan sesuatu produk. Selepas menjalankan pembersihan data dan analisis penerokaan yang menyeluruh, kejuruteraan ciri telah dilaksanakan bagi meningkatkan keupayaan ramalan, termasuk pembinaan pembolehubah seperti ketumpatan ulasan, penarafan berwajaran, dan transformasi log terhadap jumlah ulasan.

Tiga model regresi—Regresi Linear, Pokok Keputusan, dan Hutan Rawak—telah dibangunkan dan dinilai menggunakan metrik prestasi standard termasuk R^2, RMSE dan MAE. Antara model tersebut, Hutan Rawak menunjukkan ketepatan tertinggi, menerangkan lebih daripada 57% varians dalam BSR dan menunjukkan keupayaan penyesuaian yang kukuh dalam pengesahan silang. Analisis kepentingan ciri turut menunjukkan bahawa tingkah laku ulasan dan harga merupakan peramal utama status terlaris.

Penemuan ini menyokong nilai pembelajaran ansambel dan kejuruteraan ciri dalam pemodelan hasil yang kompleks dan bukan linear dalam persekitaran e-dagang. Kajian ini menyumbang dari segi teori, dengan mengesahkan pendekatan pemodelan lanjutan, dan dari segi praktikal, dengan menawarkan kerangka kerja yang boleh diguna pakai semula oleh penjual dan penganalisis untuk meramalkan prestasi produk dan menyokong keputusan strategik.

Table of Contents

1. INTRODUCTION

1.1. Overview

The rise of e-commerce has transformed how consumers shop, with platforms like Amazon leading the way. Amazon Best Seller Rank (BSR) is a score assigned by Amazon to a product based on its sales volume and historical sales data, updated every hour. This ranking not only indicates a product's popularity but also plays a significant role in influencing buyer trust and choices. Gaining insights into the elements that impact BSR can provide businesses and sellers with important information to enhance their product positioning.

1.2. Problem Background

Amazon, as the world’s leading online marketplace, offers an enormous range of products across various categories. Among these, only a few manage to earn the “Best Seller” badge—a designation that boosts visibility and drives further sales. While some might speculate that factors such as price, customer ratings, and review count contribute to a product’s success, the actual mechanics behind Amazon’s ranking system remain vague and complex.

For businesses and individual sellers, gaining insights into what contributes to sales performance is of immense practical value. Equally, for data scientists, the challenge of making sense of such real-world data presents a rich opportunity for analysis and modelling. Despite the commercial importance of this area, limited academic work has systematically studied what differentiates top-ranking products from others.

1.3. Research Gap

While prior studies have explored various aspects of e-commerce analytics, notable research gaps remain. Existing work on Amazon's best-seller ranks focuses primarily on estimating sales from rank, rather than predicting rank from product attributes such as price, reviews, or ratings—limiting its utility for sellers seeking performance insights. Similarly, research into product return behaviour often neglects the influence of structured product features and lacks integration of user-generated content, such as review sentiment or star ratings, which may better predict return likelihood. Meanwhile, customer segmentation studies apply machine learning to classify user groups based on behaviour, yet they do not connect segmentation results to tangible business metrics like sales rank or product success. Furthermore, deep learning and interpretable modelling techniques remain underutilised across these domains. These limitations highlight the need for a data-driven approach that links structured product features to best-seller performance, offering both predictive insight and strategic value for e-commerce platforms.

1.4. Research Questions

- What product attributes most influence Amazon BSR?

- Predict based on review count, rating, and price?

- Which machine learning approaches are most suitable for predicting rank in such datasets?

1.5. Research Objectives

1. To explore and clean the dataset of Amazon BSR, addressing any missing or inconsistent values and preparing the data for analysis.

2. To perform EDA in order to identify patterns, trends, and relationships among key variables such as product price, customer ratings, and number of reviews.

3. To determine the most influential features contributing to a product's BSR through correlation analysis and feature importance techniques.

4. To develop predictive models using supervised machine learning algorithms that estimate a product's likelihood of achieving a high rank based on its attributes.

5. To evaluate model performance using appropriate statistical metrics, and interpret the results to extract meaningful insights.

1.6. Research Scope

This study will focus exclusively on a static dataset of Amazon BSR. The features under consideration include price, number of ratings, average star rating, and product rank, along with marketplace country. This study focuses on a static snapshot of Amazon best seller data. Longitudinal trends and category-level distinctions are excluded due to data unavailability.

1.7. Research Contribution

- **Theoretical**: Enhances understanding of factors driving product success on e-commerce platforms.

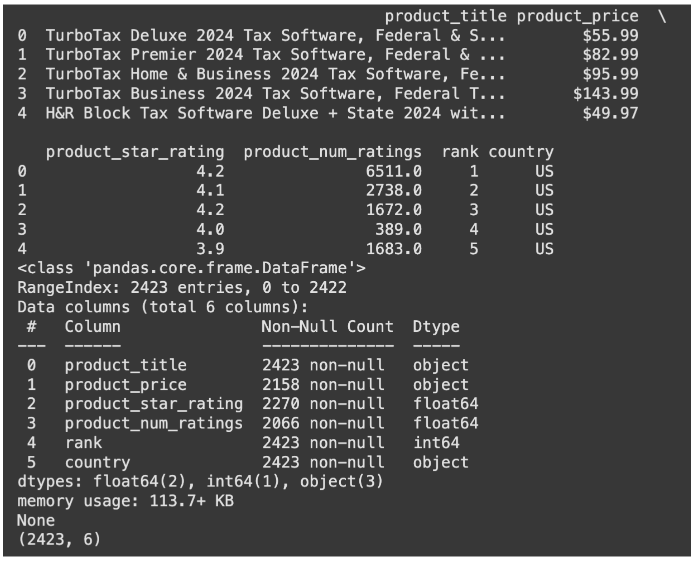
- **Methodological**: Demonstrates comparative analysis of machine learning techniques for BSR prediction.

- **Practical**: Offers actionable insights for online sellers aiming to improve product visibility.

- **Data & Tools**: Involves web scraping of Amazon data; analysis conducted using Python.

2. LITERATURE REVIEW

2.1. Overview



```
                                       product_title product_price  \
0  TurboTax Deluxe 2024 Tax Software, Federal & S...        $55.99
1  TurboTax Premier 2024 Tax Software, Federal & ...        $82.99
2  TurboTax Home & Business 2024 Tax Software, Fe...        $95.99
3  TurboTax Business 2024 Tax Software, Federal T...       $143.99
4  H&R Block Tax Software Deluxe + State 2024 wit...        $49.97

   product_star_rating  product_num_ratings  rank country
0                  4.2               6511.0     1      US
1                  4.1               2738.0     2      US
2                  4.2               1672.0     3      US
3                  4.0                389.0     4      US
4                  3.9               1683.0     5      US
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2423 entries, 0 to 2422
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   product_title        2423 non-null   object
 1   product_price        2158 non-null   object
 2   product_star_rating  2270 non-null   float64
 3   product_num_ratings  2066 non-null   float64
 4   rank                 2423 non-null   int64
 5   country              2423 non-null   object
dtypes: float64(2), int64(1), object(3)
memory usage: 113.7+ KB
None
(2423, 6)
```

Figure 4.1 Original Data

This chapter reviews existing literature relevant to the prediction of product performance on e-commerce platforms, particularly Amazon. It begins by introducing Amazon’s Best Seller Rank (BSR) as a key indicator of product success, followed by a discussion of current applications of machine learning in the e-commerce sector. The review then narrows its focus to previous efforts in sales prediction and customer analytics, emphasising the methodological limitations and research gaps that this study aims to address.

2.2. E-commerce Analytics and Product Ranking

In the digital marketplace, product visibility is an essential element for commercial success, as it directly affects consumer engagement and sales performance. Amazon, as one of the largest global e-commerce platforms, ranks products using the BSR—a score assigned based on sales volume and historical performance. Since the BSR is updated hourly, the BSR directly influences a product’s visibility in category rankings and search results, thus affecting consumer behaviour and purchase decisions.

Although the commercial importance of BSR is well known, the algorithm used by Amazon to assign the BSR remains a commercial secret. Due to this, the impact of product features on the BSR has become an increasingly interesting topic for academic as well as strategic research.

2.3. Machine Learning Applications in E-commerce

The advent of big data has led to a significant increase in the application of machine learning (ML) techniques in e-commerce. Researchers have applied ML to tackle tasks, inculding customer segmentation (Rajyalaxmi M et al., 2024), return behaviour prediction (Ramirez, 2024), and sales forecasting (Sharma et al., 2022). Techniques commonly used include supervised learning (e.g., decision trees, support vector machines), unsupervised clustering (e.g., K-means, DBSCAN), and ensemble methods (e.g., Random Forest, XGBoost). Notably, the integration of deep learning architectures—particularly recurrent neural networks (RNNs)—has enabled the modelling of complex sequential behaviours such as browsing or purchase patterns, offering deeper insight into temporal consumer dynamics.

Most studies try to understand customers using their data uch as demographics, purchase logs, or session histories, but very few focus on products, like figuring out what makes one product rank higher than another on Amazon. That area is still not well-studied.

2.4. Prediction of Sales and Ranking Metrics

Several studies have attempted to model the relationship between product features and sales outcomes. Sharma et al. (2020) investigated how BSR maps to sales in Amazon’s Clothing, Shoes & Jewellery category, using hourly sales and review data collected via a custom web scraper. Their linear regression model achieved moderate predictive accuracy (R^2 = 69.46%), demonstrating a clear correlation between BSR and sales. However, their approach predicts sales from BSR, not the reverse.

Other works have developed models to estimate return rates or classify customer purchasing behaviours, but rarely have these models been designed to predict BSR as the target variable. Moreover, such studies typically focus on a single category or use a limited set of features, limiting the generalisability of their findings.

2.5. Dataset Characteristics in Existing Research

The datasets used in current e-commerce research fall broadly into two types:

- Customer-Behavioural Data: Clickstreams, Transaction Logs, And Demographic Profiles, Commonly Used For Segmentation And Churn Prediction .

- Product-Level Data: Prices, Reviews, And Metadata, Which Are Less Commonly Used For Predictive Modelling But More Publicly Accessible.

While Sharma et al. (2020) collected rich time-series data on BSR, price, and reviews, their sample was limited to 10 products over 3 months. Other studies employ proprietary or platform-internal data, which limits reproducibility and open research. Furthermore, very few datasets contain textual review sentiment, category ratings, or temporal sales signals, all of which may play important roles in shaping BSR.

2.6. Limitations of Existing Studies

Although advancements in using ML for e-commerce analytics, several limitations persist in the literature:

- Narrow Task Focus: Many Studies Predict Sales Or Returns, Not Ranking Outcomes Like Bsr.

- Single-Category Constraints: Models Are Often Built On One Category (E.G., Books Or Electronics), Reducing Scalability Across Product Types.

- Lack Of Multivariate Analysis: Few Works Combine Multiple Structured Features (E.G. Price + Reviews + Ratings) To Predict Performance.

- Insufficient Interpretability: Black-Box Models Are Rarely Accompanied By Explanations, Which Limits Their Usefulness To Business Stakeholders.

- Minimal Attention To Bsr Volatility: Bsr Updates Hourly, But Time-Sensitive Models Such As Lstm Or Attention-Based Architectures Are Rarely Used.

2.7. Research Gap

Although BSR plays a critical role in shaping product visibility and success on Amazon, there is limited research that directly predicts BSR using multivariate machine learning models. Most existing works either focus on adjacent problems (e.g., customer segmentation or sales forecasting) or use simplistic methods such as linear regression with minimal feature sets. Additionally, models are often trained on limited data within a single category, limiting their generalisability. There is also a lack of attention to textual review features and time-aware model designs that could capture the dynamic nature of BSR. This gap presents an opportunity to develop interpretable, feature-rich predictive models for BSR using structured product data across diverse categories.

2.8. Summary of Literature Review

The reviewed literature demonstrates a growing interest in using machine learning to analyse e-commerce performance. However, while customer-centric problems are well-studied, product-centred outcomes such as BSR remain underexplored. Prior studies show the feasibility of mapping BSR to sales but do not attempt to forecast BSR using relevant product attributes. Moreover, few studies address cross-category generalisability, interpretability, or the incorporation of time-based behaviour. This research seeks to fill those gaps by building a predictive model for BSR using multivariate, structured product data and exploring methods that balance accuracy with explainability.

3. RESEARCH METHODOLOGY

3.1. Overview

This chapter outlines the methodological approach undertaken to investigate the factors influencing BSR and to develop predictive models based on structured product data. The methodology is carefully designed to align with the study's research objectives and to address the gaps identified in the literature review.

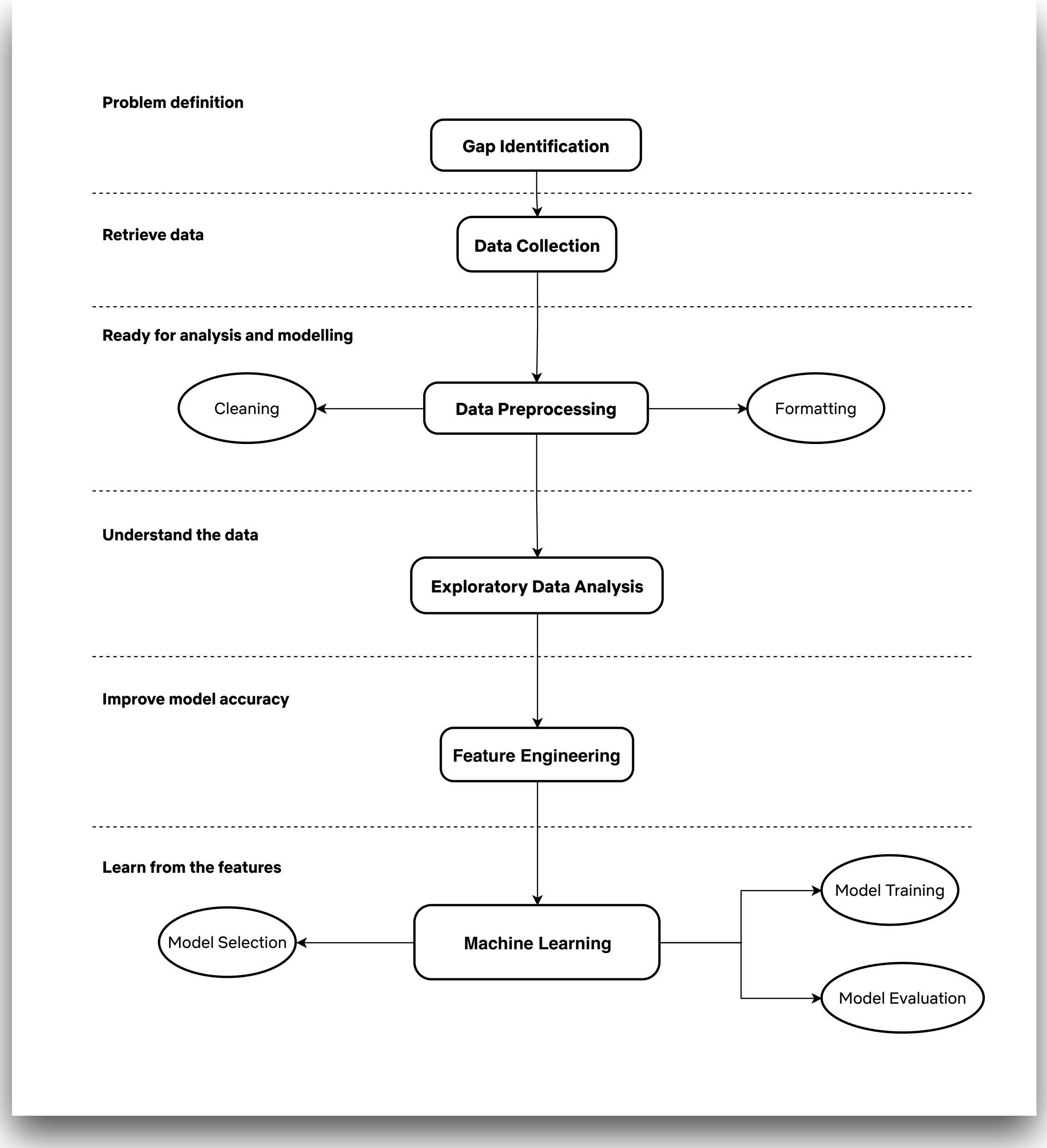
The dataset used in this study comprises real-time Amazon Best Sellers data, specifically focusing on the Software product category. It includes records from multiple countries and categories, reflecting diverse market conditions. The data was collected using Python-based web scraping techniques and contains key product attributes such as title, price, star rating, number of reviews, and BSR. These variables were selected for their potential influence on product visibility and consumer behaviour.

The methodology includes several sequential stages: selecting a suitable research design, collecting and describing the dataset, cleaning and preprocessing the data, engineering relevant features, building and evaluating machine learning models, and selecting appropriate tools and metrics for analysis. Each stage is intended to ensure that the model is not only statistically sound but also practically relevant for real-world applications, particularly for sellers and analysts seeking to optimise product visibility on Amazon.

3.2. Research Design

The research adopts a quantitative, data-driven design that centres on predictive modelling using structured product-level data. Given the nature of the problem—forecasting BSR based on measurable product attributes—a quantitative approach is most appropriate, as it enables statistical evaluation and comparison of model performance.

This study follows a supervised machine learning framework, wherein BSR serves as the target variable (dependent variable), and features such as product price, star rating, number of reviews, and engineered variables act as predictors (independent variables). The objective is to uncover patterns and relationships between these features and the product's BSR, which is updated dynamically on Amazon's platform.



To ensure robustness, the research design incorporates the following key components:

- Exploratory Data Analysis (EDA) to assess variable distributions, detect anomalies, and understand inter-feature correlations.

- Feature Engineering to enhance model performance by creating new, informative features from the raw dataset.

- Model Comparison between linear and non-linear algorithms—including Linear Regression, Decision Tree Regressor, and Random Forest Regressor—to determine which method best captures the complexity of the data.

- Evaluation Metrics such as R^2, RMSE, and MAE to quantitatively assess model accuracy and reliability.

This design enables the researcher to not only build a predictive model but also to critically evaluate its performance, generalisability, and interpretability—key aspects that address the practical needs of Amazon sellers and platform analysts.

3.3. Data Collection

The data used in this study was collected through Python-based web scraping from Amazon's public Best Sellers listings, with a specific focus on the Software category. The scraping process was automated to extract real-time information from product detail pages across multiple Amazon marketplaces.

The dataset includes structured product-level attributes that are publicly visible and potentially influential in determining Best Seller Rank (BSR). Key variables collected include:

- Product Title
- Price
- Star Rating
- Number of Reviews
- Country of Marketplact
- Best Seller Rank (BSR)

This method of data collection ensures the dataset is both current and representative of real-world product rankings. All data was obtained from publicly accessible pages for academic purposes, without violating Amazon's terms of service.

3.4. Dataset Description

The dataset used in this research comprises structured, real-time information on Amazon Best Sellers, specifically within the Software category. The data was collected via Python-based web scraping and includes listings from the Amazon marketplaces in multiple countries, allowing for comparative analysis across regions. A total of 2,423 product records were initially gathered, covering key attributes that are visible to consumers and potentially influential in Amazon's ranking algorithm.

The dataset contains the following core features:

Variables	Description
Product Title	the name or description of the product
Product Price	listed price at the time of scraping
Star Rating	average customer review score
Number of Reviews	total count of submitted customer ratings
Country	the Amazon marketplace region (e.g. US, UK)
Rank	the product's relative position within its category, which serves target variable for prediction

After preprocessing and the removal of rows with missing or inconsistent data, the final cleaned dataset used for model training consisted of approximately 2,000 complete records. This refined dataset provided a reliable foundation for exploratory analysis, feature engineering, and the development of predictive models.

3.5. Data Preprocessing

Before applying machine learning models, several preprocessing steps were conducted to ensure the quality, consistency, and suitability of the dataset for analysis. This process involved data cleaning, transformation, and preparation of variables for feature engineering and modelling

3.5.1. Handling Missing Values

The initial dataset contained missing values in key columns such as product_price, product_star_rating, and product_num_ratings. Records with missing values in any of these essential fields were removed to maintain data integrity. After cleaning, approximately 2,000 complete records remained.

3.5.2. Data Type Conversion

The product_price column initially included currency symbols and non-numeric characters. These were removed using regular expressions, and the values were converted to floating-point numbers. Similarly, columns such as product_star_rating, product_num_ratings, and rank were explicitly cast to numerical types to support mathematical operations.

3.5.3. Outlier Handling

To avoid distortion in visualisation and statistical analysis, products with extreme price values (e.g. over $200) were excluded during exploratory visualisation stages. However, the full range of prices was retained for modelling to preserve the generality of the dataset.

3.5.4. Feature Transformation

To address skewness in the distribution of review counts, a logarithmic transformation was applied using the log1p function, which computes log(1 + x). This transformation improved the stability and scaling of the feature for regression tasks.

These preprocessing steps ensured that the dataset was clean, numerically consistent, and ready for feature engineering and model development. They also reduced the risk of biased or unreliable results due to missing, malformed, or unscaled data.

3.5.5. Encoding Categorical Variables

The categorical variable country was converted into a numerical format using one-hot encoding. This transformation created binary indicators for each country, allowing the regression models to learn region-specific effects without introducing ordinal relationships. The encoding process excluded the first category to prevent multicollinearity. This step was essential for incorporating geographic context into the model without distorting the numeric relationships among features.

3.6. Modelling Approaches

To predict Amazon Best Seller Rank (BSR), this study adopts a supervised regression modelling approach, using both linear and non-linear algorithms to explore the relationships between structured product attributes and sales rank. The aim is to determine which machine learning models can most accurately capture the complex patterns within the data and deliver reliable BSR predictions.

3.6.1. Linear Regression

Linear Regression was used as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between the input features—such as price, star rating, and number of reviews—and the target variable (BSR). While it provides insight into the directional influence of each feature, its performance is limited in datasets where relationships are non-linear or involve interactions between variables.

3.6.1. Decision Tree Regressor

The Decision Tree Regressor was implemented to capture non-linear patterns and conditional relationships in the data. This model splits the dataset into decision nodes based on feature values, allowing for flexible modelling of complex feature-target dynamics. Decision trees are intuitive and visually interpretable, but they are prone to overfitting if not properly tuned or pruned.

3.6.2. Random Forest Regressor

Random Forest, an ensemble learning method, was selected to improve predictive accuracy and robustness. It constructs multiple decision trees and averages their predictions, reducing the risk of overfitting and improving generalisability. This model also provides insights into feature importance, making it suitable for both prediction and interpretability.

Each model was trained on the same dataset using a standard 80/20 train-test split to ensure fair comparison. The models were evaluated using regression metrics including R^2, RMSE, and MAE, as detailed in Section 3.7. This multi-model approach allows for a comprehensive understanding of the modelling landscape and the identification of the most effective algorithm for BSR prediction.

3.7. Evaluation Metrics

To assess the performance of the machine learning models developed in this study, three standard evaluation metrics for regression tasks were used: R-squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These metrics provide complementary insights into how accurately the models predict Amazon Best Seller Rank (BSR) based on product attributes.

3.7.1. R-squared (R^2)

R^2 shows how much of the variation in BSR can be explained by the model. A value closer to 1 means the model fits the data well, while a value near 0 means it does not explain much. It’s useful for judging the overall fit of the model.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

3.7.2. Root Mean Squared Error (RMSE)

RMSE measures the average size of the prediction errors, giving more weight to larger errors. It tells us how far off the predictions are, on average, in the same units as BSR. A lower RMSE indicates better performance.

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

3.7.3. Mean Absolute Error (MAE)

MAE calculates the average of the absolute differences between the predicted and actual values. Unlike RMSE, it treats all errors equally. It gives a straightforward idea of how far off the predictions are, on average.

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

These metrics were chosen to evaluate the models from multiple perspectives: R^2 for overall model fit, RMSE for penalising large errors, and MAE for interpretability and robustness. The combination of these metrics ensures a comprehensive evaluation of model performance and supports fair comparison across different algorithms.

3.8. Tools and Technologies

This research employs a range of programming tools and data science libraries to support data collection, preprocessing, analysis, modelling, and visualisation. All development and experimentation were conducted in a Python-based environment, which offers flexibility, scalability, and strong community support for machine learning tasks.

3.8.1. Programming Language

- Python 3.11 was used as the core programming language due to its extensive support for data analysis, machine learning, and web scraping. Python's readable syntax and mature ecosystem make it particularly well-suited for applied data science research.

3.8.2. Integrated Development Environment (IDE)

- PyCharm was used for writing, testing, and debugging code. It provides an efficient interface for managing Python projects and integrating version control and virtual environments.

3.8.3. Libraries and Frameworks

- pandas and numpy were used for data manipulation, cleaning, and numerical operations.

- matplotlib and seaborn were employed for data visualisation and exploratory analysis.

- scikit-learn served as the primary machine learning library, providing access to a wide range of algorithms and evaluation tools, including:

- LinearRegression

- DecisionTreeRegressor

- RandomForestRegressor

- Performance metrics such as r2_score, mean_absolute_error, and mean_squared_error

3.8.4. Web Scraping

Python-based scraping tools were used to collect real-time Amazon product data. Though the specific module (e.g. requests, BeautifulSoup, or Selenium) is not listed in the script, a custom scraping script was developed to extract structured product-level information from public listings.

3.8.5. Hardware and Runtime Environment

The experiments were conducted on a personal computer with macOS, using local computation. The dataset size was moderate and did not require distributed computing or cloud infrastructure.

This suite of tools and technologies provided a robust, reproducible environment for executing each stage of the research methodology—from raw data collection to final model evaluation.

3.9. Summary

This chapter outlined the methodological framework used to investigate the relationship between structured product attributes and Amazon Best Seller Rank (BSR), and to develop predictive models using machine learning techniques. The chapter began with an overview of the research design, which followed a quantitative, supervised learning approach to address the research problem.

Data was collected via Python-based web scraping, focusing on the Software category across multiple Amazon marketplaces. Key variables such as product price, star rating, number of reviews, and BSR were extracted and preprocessed to ensure consistency and analytical reliability. The data underwent cleaning, transformation, and feature engineering to enhance the quality of the inputs.

Several machine learning models were implemented—namely Linear Regression, Decision Tree Regressor, and Random Forest Regressor—to predict BSR. Each model was evaluated using R^2, RMSE, and MAE to compare predictive accuracy and interpretability. The results highlighted the limitations of linear models in capturing the complex dynamics of BSR and demonstrated the superiority of non-linear models, particularly Random Forest.

The next chapter will present the experimental results, visualisations, and a comparative analysis of model performance, leading to insights that support the research objectives.

4. INITIAL FINDINGS

4.1. Introduction

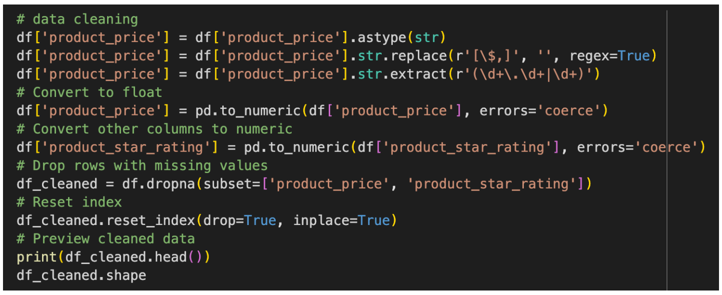
This chapter presents the analytical process and findings from the implementation of machine learning models to predict BSR. It begins with an overview of the data cleaning steps, including how missing values and data types were handled to prepare the dataset for analysis. Following that, exploratory data analysis (EDA) is conducted to uncover patterns, distributions, and correlations among variables such as product price, star rating, and review volume.

Next, the results of feature engineering are described, highlighting the creation of derived variables like review_density, weighted_rating, and log_num_ratings, which aim to capture consumer engagement and product performance more effectively. These engineered features are then used to train and evaluate three regression models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

Model performance is assessed using standard regression metrics — R^2 Score, RMSE, and MAE — and further validated using cross-validation to identify potential overfitting. The final section summarises the chapter's findings and sets the foundation for conclusions and recommendations in the next chapter.

4.2. Results of Data Cleaning

Data cleaning was a critical initial step to ensure consistency and reliability in the modelling process. The raw dataset, which included product details such as prices, star ratings, number of reviews, and best seller rank (BSR), contained inconsistencies in format and missing values that required resolution.



```python
# data cleaning
df['product_price'] = df['product_price'].astype(str)
df['product_price'] = df['product_price'].str.replace(r'[\$,]', '', regex=True)
df['product_price'] = df['product_price'].str.extract(r'(\d+\.\d+|\d+)')
# Convert to float
df['product_price'] = pd.to_numeric(df['product_price'], errors='coerce')
# Convert other columns to numeric
df['product_star_rating'] = pd.to_numeric(df['product_star_rating'], errors='coerce')
# Drop rows with missing values
df_cleaned = df.dropna(subset=['product_price', 'product_star_rating'])
# Reset index
df_cleaned.reset_index(drop=True, inplace=True)
# Preview cleaned data
print(df_cleaned.head())
df_cleaned.shape
```

Figure 4.2 Data Cleaning

4.2.1. Original Dataset

The original dataset contained a total of 2,423 product records across six variables. Upon inspection, several key features were found to have missing values. Specifically, the product_price column had 265 missing entries, product_star_rating was missing 153 values, and product_num_ratings had 357 null entries. These features are essential for model training, as they directly contribute to understanding customer perception and pricing strategies.

4.2.2. Handling Missing Value

Several columns, including product_price and product_star_rating, contained missing or improperly formatted entries. Rows with missing values in these essential columns were removed to maintain data integrity. This cleaning step reduced noise and ensured that the input features fed into the machine learning models were complete and interpretable. The resulting dataset retained sufficient observations for robust training and testing, while eliminating incomplete records that could bias or weaken model accuracy.

4.2.3. Data Type Conversion

The product_price column originally contained string values with currency symbols and formatting characters (e.g., “$”, “,”). These were removed using regular expressions, and the cleaned values were converted to numeric (float) type.

Similarly, product_star_rating and product_num_ratings were converted to numerical format to support correlation analysis and mathematical transformations later applied during feature engineering. These conversions were essential to enable correct computation, visualisation, and model fitting.

After these cleaning operations, the dataset was reset to ensure continuous indexing and reviewed to confirm readiness for exploratory analysis and modelling.

4.3. Insights from Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the structure, distribution, and relationships within the dataset prior to model training. This process provided critical insights into country distribution, rating behaviour, price variation, and feature correlations, all of which informed subsequent steps such as feature engineering and model selection.

4.3.1. Count of Products by Country

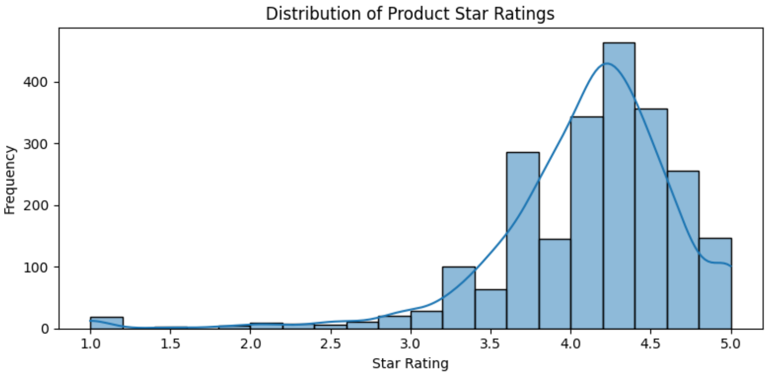


Figure 4.4 Distribution of Production Star Ratings

A count plot was used to visualise the number of products across different Amazon marketplaces. The majority of product entries came from major markets such as the United States, Germany, Australia, and India, each contributing a significant portion of the dataset. In contrast, a few countries like Poland and Sweden had relatively fewer entries.

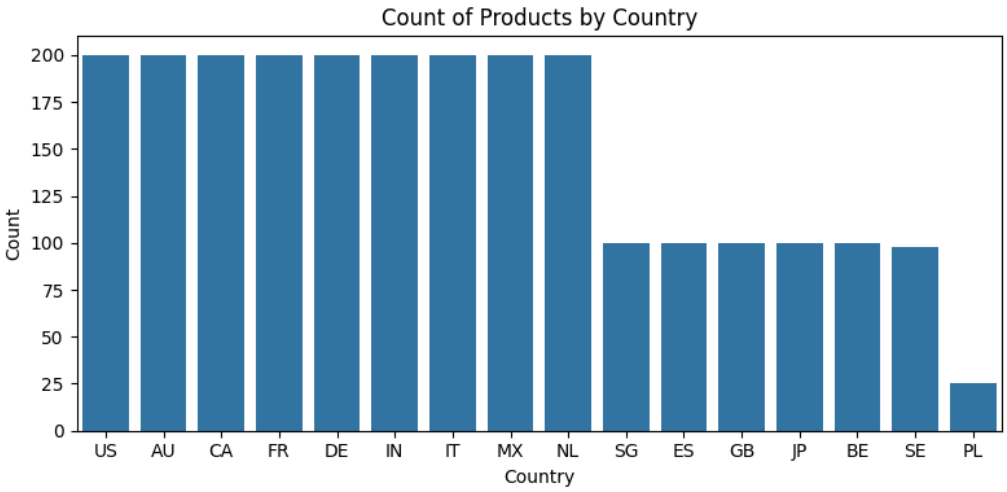


Figure 4.3 Count of Products by Country

This imbalance in representation may influence the model's learning process, as countries with limited samples contribute less information during training. Therefore, one-hot encoding was applied during feature engineering to handle categorical differences without introducing ordinal bias. Additionally, low-frequency countries may require careful interpretation when evaluating feature importance or generalising predictions across regions.

4.3.2. Distribution of Production Star Ratings

A histogram was generated to examine the distribution of product star ratings. The visualisation revealed that ratings are skewed toward the higher end of the scale, with a large concentration of products rated between 4.0 and 4.5 stars. Very few products had ratings below 3.5, and ratings below 3.0 were extremely rare.

This distribution suggests that most products listed on Amazon maintain consistently high customer satisfaction, possibly due to user rating bias or platform curation that removes poorly performing items. The lack of low-rated items may also be influenced by sellers discontinuing underperforming products.

From a modelling perspective, this skew introduces a potential challenge: the reduced variability in ratings limits their predictive power, especially for distinguishing between moderately successful products. As a result, additional features such as review density and weighted ratings were engineered to better capture product quality and popularity.

4.3.3. Box Plot of Product Prices

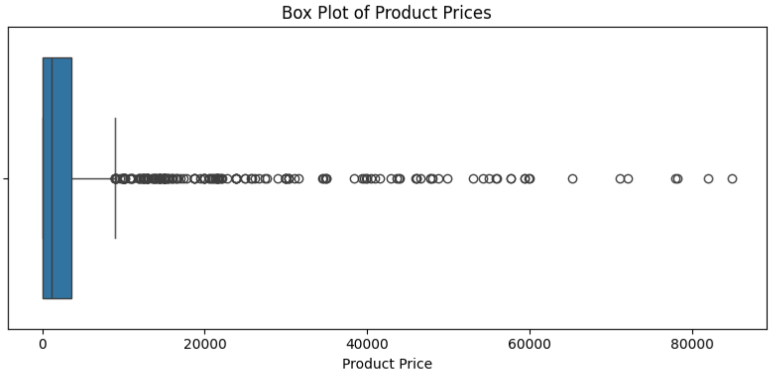


Figure 4.5 Box Plot of Product Prices

A box plot was generated to analyse the distribution and spread of product prices. The visualisation revealed that while the majority of products are priced below $1,000, there are a substantial number of extreme outliers with prices extending beyond $10,000, and in some cases exceeding $80,000. These outliers are visually evident as scattered points far to the right of the main box area.

This high level of price skewness is common in online marketplaces, where certain specialised or bundled software packages may be priced significantly above the typical range. However, from a modelling standpoint, these outliers can distort regression results and reduce model stability.

To mitigate this, a log transformation or outlier removal could be considered in future iterations. In this study, the presence of extreme values informed the creation of derived features such as review density, which normalises the number of reviews by price to better reflect relative customer engagement.

4.3.4. Correlation Heatmap

A correlation heatmap was used to assess the linear relationships between key numerical variables in the dataset. The matrix reveals several important insights:

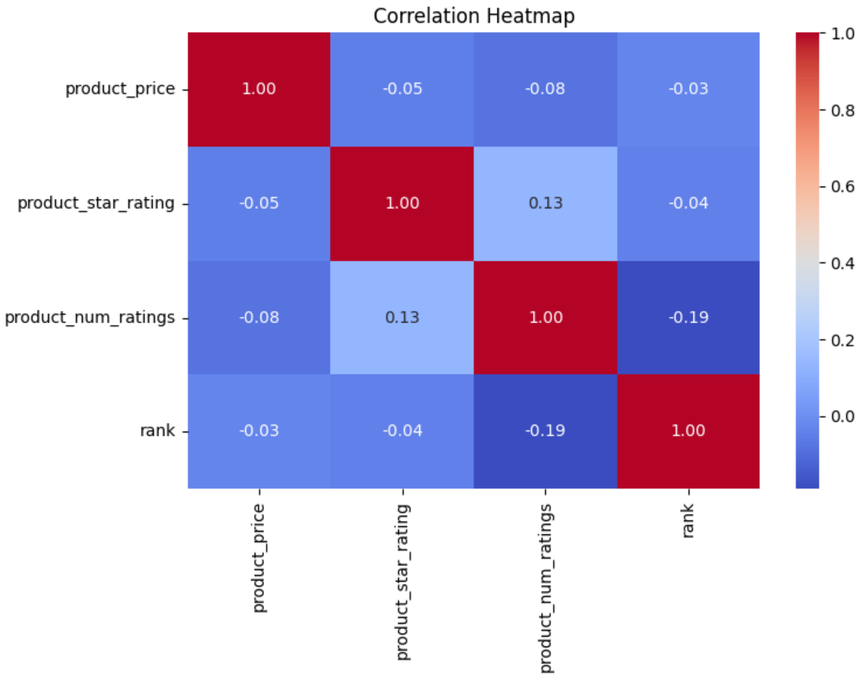


Figure 4.6 Correlation Heatmap

- product_num_ratings has the strongest (though still weak) negative correlation with rank (r = –0.19), suggesting that products with more reviews are more likely to achieve a better (lower) BSR.

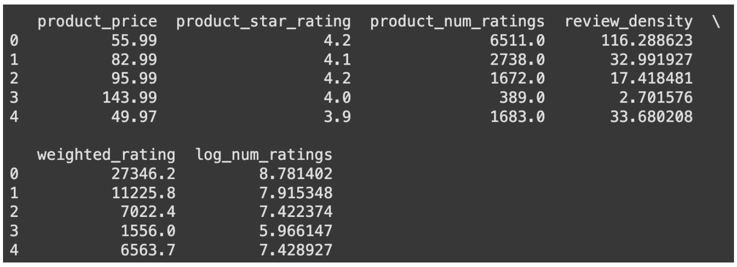
- product_star_rating and product_price show very weak correlations with rank (r = –0.04 and –0.03 respectively), indicating that neither feature alone is a strong predictor of best seller performance.

- The highest inter-feature correlation is between product_star_rating and product_num_ratings (r = 0.13), which is also weak.

These low correlation values suggest that no single variable linearly explains BSR well, and support the use of multivariate and non-linear models such as decision trees and random forests. They also justify the creation of engineered features like weighted_rating and review_density, which aim to capture more complex interactions.

4.4. Results of Feature Engineering

To enhance the predictive power of the dataset and better capture underlying patterns, several new features were engineered. These were designed to account for non-linear relationships, normalise skewed data, and provide composite indicators of product popularity and perceived quality.



```
   product_price  product_star_rating  product_num_ratings  review_density  \
0          55.99                  4.2               6511.0      116.288623
1          82.99                  4.1               2738.0       32.991927
2          95.99                  4.2               1672.0       17.418481
3         143.99                  4.0                389.0        2.701576
4          49.97                  3.9               1683.0       33.680208

   weighted_rating  log_num_ratings
0          27346.2         8.781402
1          11225.8         7.915348
2           7022.4         7.422374
3           1556.0         5.966147
4           6563.7         7.428927
```

Figure 4.7 Dataset after Feature Engineering

4.4.1. Create New Features

Three new features were created to enrich the dataset:

- review_density: Defined as the number of product reviews divided by product price, this feature captures customer engagement relative to cost. For instance, a product priced at $55.99 with 6511 reviews yields a review density of approximately 116.29, reflecting strong perceived value or popularity.

- weighted_rating: Calculated by multiplying product_star_rating with product_num_ratings, this variable combines quality and quantity of reviews. A product with a rating of 4.2 and 6511 reviews, for example, results in a weighted rating of 27346.2, highlighting its overall customer approval level.

- log_num_ratings: A natural log transformation was applied to product_num_ratings to reduce skewness caused by highly popular products. For instance, a product with 6511 reviews yields a log_num_ratings value of 8.78, allowing the model to treat large review volumes more proportionally.

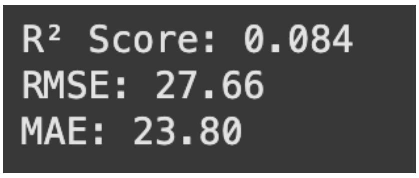
These newly constructed features were observed to vary substantially across products and better reflect performance trends not captured by raw metrics alone.

4.4.2. Normalise Skewed Data

The product_num_ratings column displayed substantial skew due to the presence of products with exceptionally high review counts. To mitigate this, a logarithmic transformation (log1p) was applied, converting large values into a compressed range while preserving their relative magnitude. This transformation improved model stability and reduced the dominance of outliers during training.

4.4.3. One-Hot Encode Categorical Variable

The country column, being categorical, was converted into multiple binary columns using one-hot encoding. This allowed the model to consider country-specific effects without assuming ordinal relationships. For example, the presence of country_US = 1 denotes a product listed in the US marketplace, enabling the model to capture regional patterns in BSR performance.



```
R² Score: 0.084
RMSE: 27.66
MAE: 23.80
```

Figure 4.9 Evaluation Metrics

4.5. Evaluation of Models Performance

The results presented in the previous section highlight notable differences in model performance when predicting BSR using structured product data. These differences reflect each algorithm’s capacity to capture underlying relationships between features and the target variable.

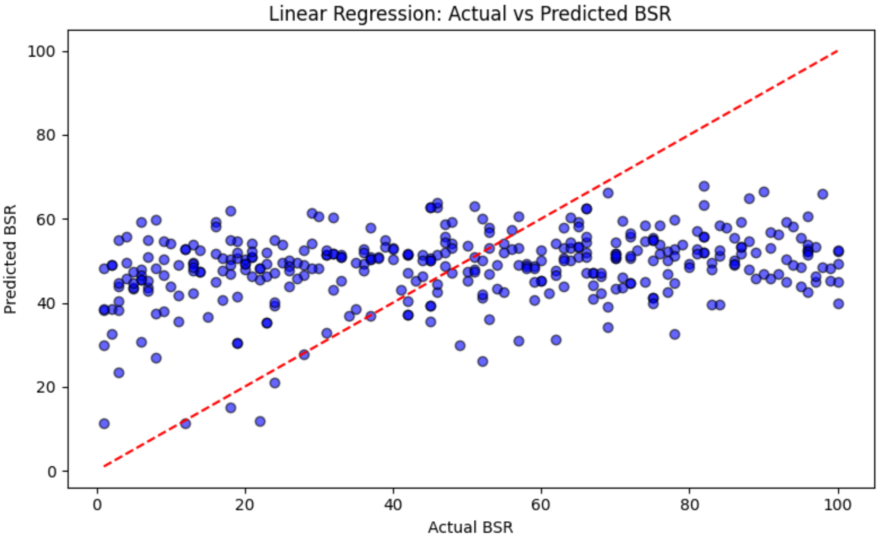


Figure 4.8 Linear Regression

4.5.1. Linear Regression Results

The first model applied to the dataset was Linear Regression, chosen for its simplicity and interpretability. The model was trained using the selected features, including engineered variables like review_density, weighted_rating, and log_num_ratings, as well as one-hot encoded country indicators.

After fitting the model, predictions were made on the test set, and the following evaluation metrics were obtained:

These results indicate that the Linear Regression model performed poorly in capturing the variability in the Best Seller Rank (BSR). An R² score of only 0.084 suggests that the model explains less than 10% of the variance in the target variable. The relatively high RMSE and MAE values further support this conclusion, showing that prediction errors were substantial.

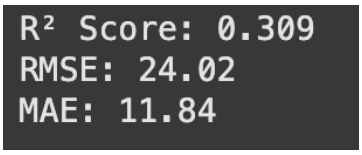
A scatter plot comparing actual and predicted BSR values revealed a wide spread around the ideal prediction line, further confirming the model's lack of accuracy. The poor performance can likely be attributed to the non-linear and complex relationships between features and BSR, which a linear model is not capable of capturing effectively.

As a result, more flexible models—such as decision trees and ensemble methods—were explored in the following sections.

4.5.2. Decision Tree Results

A Decision Tree Regressor was trained to improve upon the shortcomings of the linear regression model by capturing non-linear relationships within the data. The model was trained using the same feature set, including engineered variables and one-hot encoded country data.

After training and predicting on the test set, the performance was evaluated using the following metrics:



```
R² Score: 0.309
RMSE: 24.02
MAE: 11.84
```

Figure 4.10 Evaluation Metrics

Compared to the Linear Regression model, the Decision Tree achieved a notable improvement in all three metrics. The R² score increased from 0.084 to 0.309, indicating a higher proportion of variance explained by the model. Additionally, the RMSE and MAE both decreased, showing that the predictions were closer to the actual BSR values.

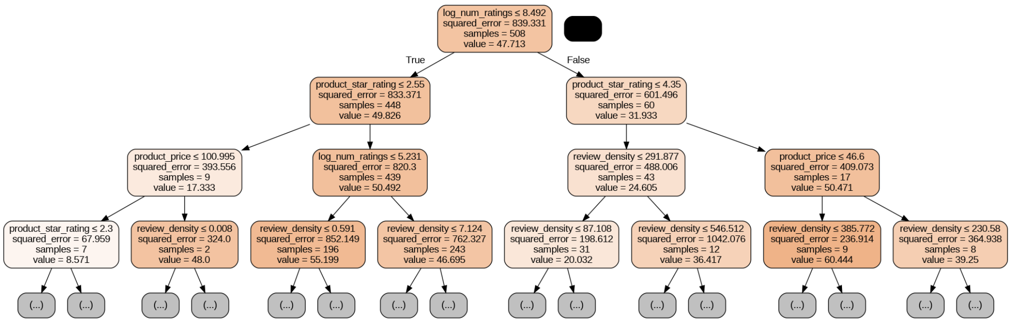


Figure 4.11 Decision Tree

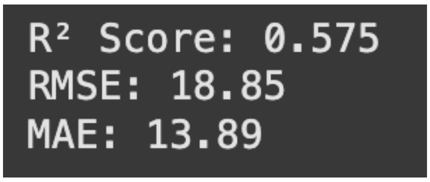
Despite these improvements, the Decision Tree model is still prone to overfitting, especially when trained without pruning or regularisation. This risk was visualised by exporting and plotting the structure of the decision tree (limited to a depth of 3 for interpretability), which showed that splits were strongly influenced by engineered features such as log_num_ratings and review_density.

While performance was better than the linear model, the Decision Tree's lack of robustness across varied data led to the adoption of ensemble methods—specifically the Random Forest Regressor—as a next step.

4.5.3. Random Forest Results

The Random Forest Regressor was employed as an ensemble-based model to address the overfitting tendency of individual decision trees and to better capture complex feature interactions. Using 100 estimators and default hyperparameters, the model was trained on the full feature set, including engineered and one-hot encoded variables.

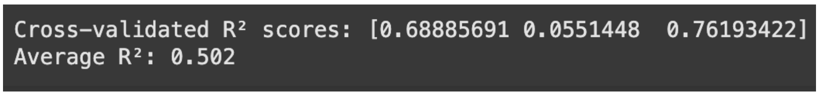
Upon evaluation, the Random Forest model achieved the following results on the test set:



```
R² Score: 0.575
RMSE: 18.85
MAE: 13.89
```

Figure 4.12 Evaluation Metrics

These results demonstrate a substantial improvement over both the Linear Regression and Decision Tree models. The R² score indicates that over 57% of the variance in BSR was explained by the model, while the reduction in RMSE and MAE confirms that prediction errors were considerably lower.



```
Cross-validated R² scores: [0.68885691 0.0551448  0.76193422]
Average R²: 0.502
```

Figure 4.13 Cross-validated R² scores

To validate generalisation, 3-fold cross-validation was conducted, producing an average R² score of approximately 0.502, closely aligned with the test performance. This consistency indicates that the model generalises well and is not overfitting.

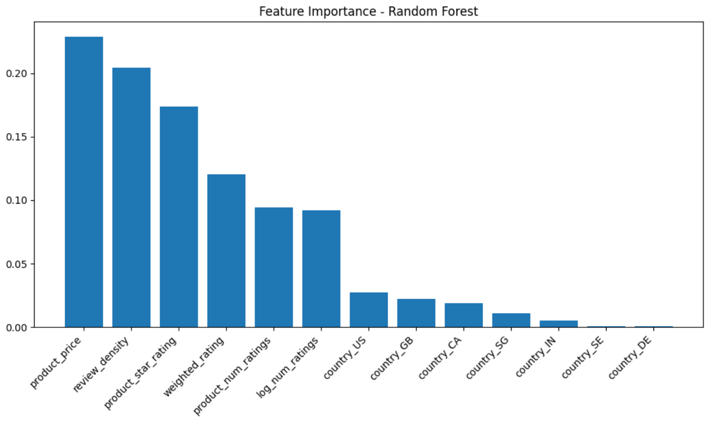


Figure 4.14 Feature Importance

A feature importance plot revealed that variables such as product_price, review_density, and product_star_rating were among the most influential in predicting BSR. This further supports the value of feature engineering in improving model performance.

In summary, Random Forest outperformed all previous models and was considered the most reliable baseline for BSR prediction in this study.

4.6. Summary

This chapter presented the modelling results and evaluation of three machine learning approaches applied to predict Amazon BSR. The workflow progressed from data cleaning and exploratory analysis to feature engineering and supervised regression modelling.

Feature engineering played a crucial role in enhancing model effectiveness. Derived features such as review_density, weighted_rating, and log_num_ratings were found to be important predictors, as confirmed through feature importance analysis.

Initial experiments with Linear Regression revealed that simple linear models were insufficient for capturing the complexity of the data, as indicated by low predictive accuracy and high error values. The Decision Tree Regressor improved upon this by modelling non-linear relationships, though it still showed limitations in generalisation. The most promising results were achieved with the Random Forest Regressor, which demonstrated the best overall performance with an R^2 score of 0.575 and significantly reduced prediction error.

Cross-validation results further supported the generalisability of the Random Forest model, making it a suitable baseline for future refinement or deployment. The chapter highlights that ensemble learning and thoughtful feature transformation are essential for predicting BSR in a complex, non-linear e-commerce environment.

5. CONCLUSION AND RECOMMENDATIONS

5.1. Research Summary

This research explored the application of supervised machine learning techniques to predict Amazon BSR using structured product-level data. The study utilised a dataset collected in real time from Amazon's best seller listings, focusing on software products across multiple countries. The research began with thorough data cleaning and exploratory data analysis to identify key patterns and variable distributions. Feature engineering was then conducted to construct new, informative variables such as review_density, weighted_rating, and log_num_ratings, enhancing the model's ability to capture product performance dynamics.

Three regression models were implemented and compared—Linear Regression, Decision Tree, and Random Forest. The Random Forest Regressor outperformed the others in terms of R^2, RMSE, and MAE, indicating its superior ability to model the complex and non-linear relationships in the data. Cross-validation results confirmed its stability and generalisation performance. Overall, the study demonstrated that ensemble learning methods, combined with well-crafted features, offer a robust approach to predicting BSR and can be used to inform marketing, pricing, and inventory decisions in e-commerce settings.

5.2. Research Objectives

The primary objective of this research was to develop and evaluate machine learning models capable of predicting Amazon BSR based on structured product attributes. To achieve this, the following specific objectives were formulated:

- Understand which product features most influence Amazon Best Seller Rank.

- Develop predictive models using machine learning techniques.

- Evaluate and compare the performance of different regression models.

- Identify patterns and category-level trends through exploratory analysis.

All objectives were achieved, with clear evidence that model accuracy improves with well-engineered features and non-linear approaches like Random Forest.

5.3. Research Contributions

This study makes valuable contributions to both the theoretical and practical domains of data science and e-commerce. By developing and evaluating machine learning models to predict Amazon Best Seller Rank (BSR), the research offers insights into model effectiveness, feature importance, and real-world applicability.

5.3.1. Theoretical Perspective

- From a theoretical standpoint, the research extends existing knowledge in predictive modelling by demonstrating the limitations of linear models in complex e-commerce environments. It highlights the importance of non-linear and ensemble methods—such as decision trees and random forests—in capturing the intricacies of customer behaviour and sales dynamics. Additionally, the study reinforces the significance of feature engineering, showing that derived features like review_density and weighted_rating offer greater explanatory power than raw attributes alone. The methodological design also supports the use of cross-validation as a robust approach to assess generalisability in supervised learning tasks.

5.3.2. Practical Perspective

Practically, this research provides a replicable and scalable framework for Amazon sellers, analysts, and digital marketers to forecast product ranking performance using accessible data. The final model identifies key drivers of BSR, such as price, review count, and star rating, enabling more informed decisions about pricing strategies, customer engagement, and market positioning. Moreover, the structured pipeline—ranging from data preprocessing to model evaluation—can be easily adapted for real-time applications, such as dashboards or automated product monitoring systems. In doing so, the study contributes tools and insights that are directly actionable in today's data-driven e-commerce landscape.

5.4. Research Limitations

While the study achieved its objectives and produced promising results, several limitations must be acknowledged:

A. Domain-Specific Dataset:

The dataset used in this study focused solely on software products from Amazon. As a result, the findings and model performance may not generalise to other product categories such as electronics, fashion, or books, where consumer behaviour and ranking dynamics could differ significantly.

B. Static Snapshot of Data:

Although BSR is updated hourly, the dataset represents a single snapshot in time. This limits the model's ability to capture temporal patterns or trends, such as seasonal fluctuations or the impact of promotions.

C. Limited Feature Scope:

The analysis relied entirely on structured data such as price, ratings, and number of reviews. It did not incorporate textual data like product descriptions or customer reviews, which could contain rich signals relevant to product success.

D. No External Validation:

The model's performance was validated using train-test splits and cross-validation within the dataset. However, the model was not tested on a completely independent dataset or deployed in a real-world environment, which limits the assessment of its robustness in practice.

E. Country-Level Bias:

Some countries in the dataset were underrepresented, which may have led to biased country-specific features. This imbalance could affect the model's ability to generalise across global marketplaces.

Despite these limitations, the research establishes a solid foundation for future studies and applications involving BSR prediction.

5.5. Future Work Recommendations

To build upon the findings of this research, future work should consider expanding both the depth and breadth of the analysis. One valuable direction would be to incorporate temporal data, allowing for time-series modelling that captures changes in BSR over time and better reflects the real-time nature of Amazon rankings. Additionally, future studies could benefit from applying the model across multiple product categories beyond software, enabling broader generalisation and uncovering category-specific trends. Enhancing the dataset with textual features such as product descriptions and customer reviews could also improve prediction, particularly through sentiment analysis and natural language processing techniques.

Further optimisation of model hyperparameters using advanced tuning methods, such as grid search or Bayesian optimisation, may yield better performance and greater reliability. Moreover, deploying the model in a live, real-time environment would increase its practical utility for sellers and analysts. Exploring alternative machine learning approaches, including gradient boosting methods or deep learning architectures, could also enhance prediction accuracy and adaptability. Overall, future research should aim to create more robust, scalable, and generalisable BSR prediction systems suitable for dynamic and competitive e-commerce platforms.

5.6. Summary

This chapter has brought the research to its conclusion by revisiting the key components and outcomes of the study. It began with a summary of the research process, highlighting the use of supervised machine learning techniques to predict Amazon Best Seller Rank (BSR) from structured product data. The research objectives were clearly defined and successfully achieved, leading to valuable theoretical and practical contributions. While the study demonstrated that models like Random Forest can effectively capture non-linear relationships within e-commerce data, it also acknowledged several limitations, such as a narrow product focus and the exclusion of time-based and textual features. These limitations present opportunities for future improvement. Recommendations were offered to guide subsequent research efforts, including the incorporation of temporal data, broader product coverage, and deployment in real-time environments. Overall, the research provides a foundational approach for data-driven BSR prediction and contributes meaningful insights to both academia and industry.