2025

Best Seller Identification with Logistic Regression

PDAN8412 PART 2

MAXIMILIAN WALSH – ST10203070

27 OCTOBER 2025

Table of Contents

Introduction

This analysis aimed to develop a Logistic Regression classification model capable of predicting whether a book will become a bestseller, using real-world bibliometric and engagement data. The chosen dataset, *Best Books Ever*, was sourced from Kaggle and represents a large-scale collection of over 40,000 books, each described through multiple quantitative and categorical attributes such as user ratings, rating counts, liked percentages, genre information, and publisher details (Mostafapoor, 2025).

This task was approached from the perspective of a data analyst in a large book publishing company, tasked with identifying the key factors that differentiate top-performing books from average ones. Spark was leveraged for scalable data handling, and scikit-learn was used for statistical modelling.

A structured methodology was followed in the analysis, encompassing data ingestion, cleaning, feature engineering, model training, and performance evaluation. The logistic regression approach was chosen to meet the project requirements; however, it is considered suitable as this is a binary classification task with the output being either bestseller or non-bestseller. More so, it offers a balance of interpretability, computational efficiency, and robustness (OpenAI, 2025). All of which contribute to the understanding of which features most strongly influence bestseller outcomes.

Ultimately, the results of this project show that a traditional statistical method, when carefully engineered, can yield near-perfect predictive performance and maintain explainability and business relevance.

Dataset Justification

Source: Kaggle – “Best Books Ever” by Pooria Mostafapoor (Mostafapoor, 2025).

Size: 60,475 records, 25 columns

Target Variable: ‘bestseller’ (binary: 0/1)

The dataset selected for this analysis was the “Best Books Ever” dataset from Kaggle, containing 60,475 initial records, of which 41,103 remained after rigorous data cleaning and deduplication. Each entry represents a book title, accompanied by metadata such as the author, language, average rating, number of ratings, liked percentage, BBE (Best Books Ever) score and votes, price, genre, and publisher (Mostafapoor, 2025). The dataset was chosen for the following reasons:

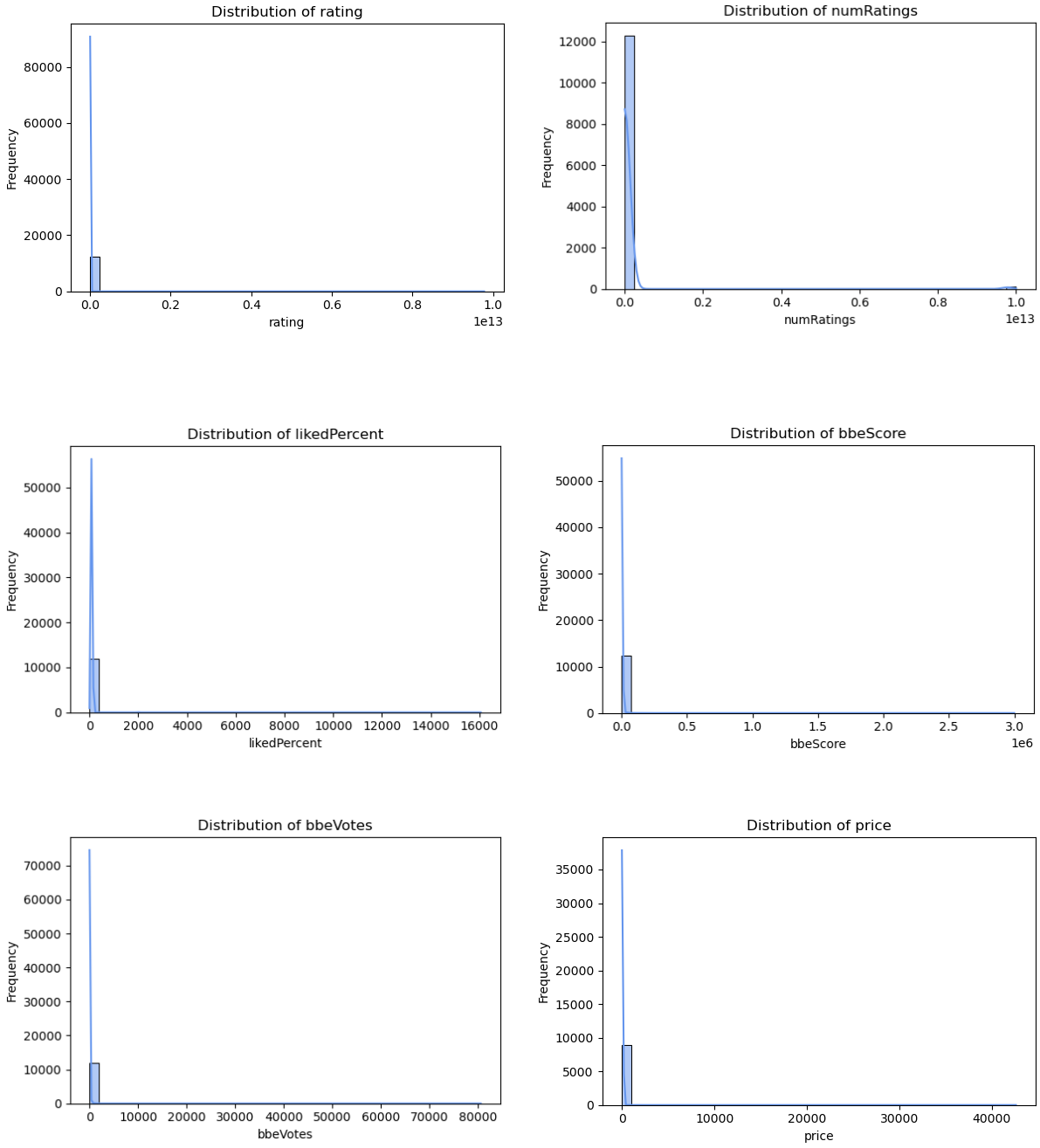
- Relevance to Task: The dataset aligns directly, intending to build a binary classification model to predict whether a book is likely to be a best seller based on its attributes, such as ratings, engagement levels, and metadata. Aligning closely with the requirements for logistic regression in predictive analytics.
- Sufficient Volume and Diversity: Given usable entries post cleaning are over 40,000, the dataset comfortably exceeds the minimum project threshold of 10,000 for robust machine learning modelling. Such a large and varied sample ensures generalisability and reduces the risk of overfitting with representation across multiple authors, languages, and genres (OpenAI, 2025).
- Label Engineering Feasibility: Whilst no explicit ‘bestseller’ label existed, the target variable was engineered from the ‘numRatings’ field. Books with a number of ratings at or above the 80th percentile (11,166) were labelled as bestsellers (1), and the others were labelled as non-bestsellers (0). Such a threshold is a quantifiable separation between typical and high-performing titles (OpenAI, 2025).
- Feature Richness: Both numerical and categorical features were included in the dataset, which enables the creation of a well-rounded predictive model capable of capturing both quantitative and qualitative book attributes (OpenAI, 2025).
- Data Quality and Balance: Post cleaning, approximately 8,573 books, which is 21%, were labelled as bestsellers, and 32,530 books, which is 79%, were labelled as non-bestsellers. This ratio is a reasonable class separation without extreme imbalance, making logistic regression a suitable modelling approach (Muller & Guido, 2016).

To conclude, the Best Books Ever dataset is of high quality, large scale, and contextually appropriate for developing a predictive model of literary success. It aligns well with the analytical objective of this project to provide interpretable insights into the drivers of book popularity, whilst providing dependable, generalisable predictive outcomes (OpenAI, 2025).

Exploratory Data Analysis (EDA)

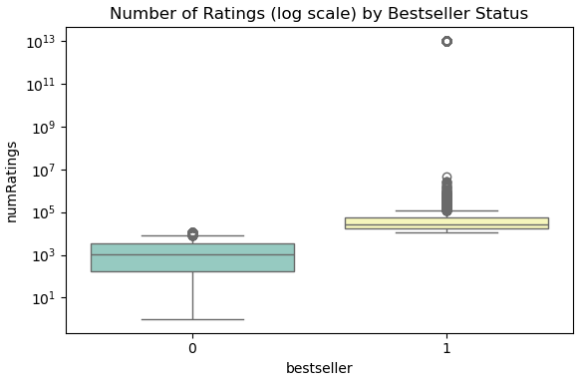
Figure 1: Class (Bestseller vs Non-Bestseller) Distribution Bar Chart

The bar chart illustrated in Figure 1 shows the relative class balance between bestseller (1) and non-bestseller (0) labels. Approximately 80% of entries were non-bestsellers, while 20% were bestsellers. Such a moderate imbalance confirms the dataset's suitability for binary classification, although model weighting using ‘class_weight=”balanced”’ was applied at a later stage to prevent bias toward the majority class (OpenAI, 2025).



Figures 2, 3, 4, 5, 6, 7: Numeric Feature Distributions Histograms

The histograms in Figures 2-7 show the distribution of six key numerical variables: ‘rating’, ‘numRatings’, ‘likedPercent’, ‘bbeScore’, ‘bbeVotes’, and ‘price’. A strong right skewness is exhibited in each variable, with a few extreme outliers representing highly popular or expensive books. This heavy-tailed distribution is common in commercial datasets and necessitates the use of logarithmic transformation and standard scaling during feature engineering to stabilise variance and improve model interpretation (OpenAI, 2025).



Figures 8, 9: Box Plots and Relationships

The box plots above provide further insights into the relationship between numerical variables and bestseller status. The average rating plot shows minimal differentiation between classes, suggesting that rating alone is not a reliable predictor of bestseller performance (OpenAI, 2025). While the number of ratings plot, particularly the logarithmic scale, shows that bestsellers consistently have higher rating volumes, which confirms that reader engagement volume is the stronger indicator of popularity as opposed to rating quality (OpenAI, 2025).

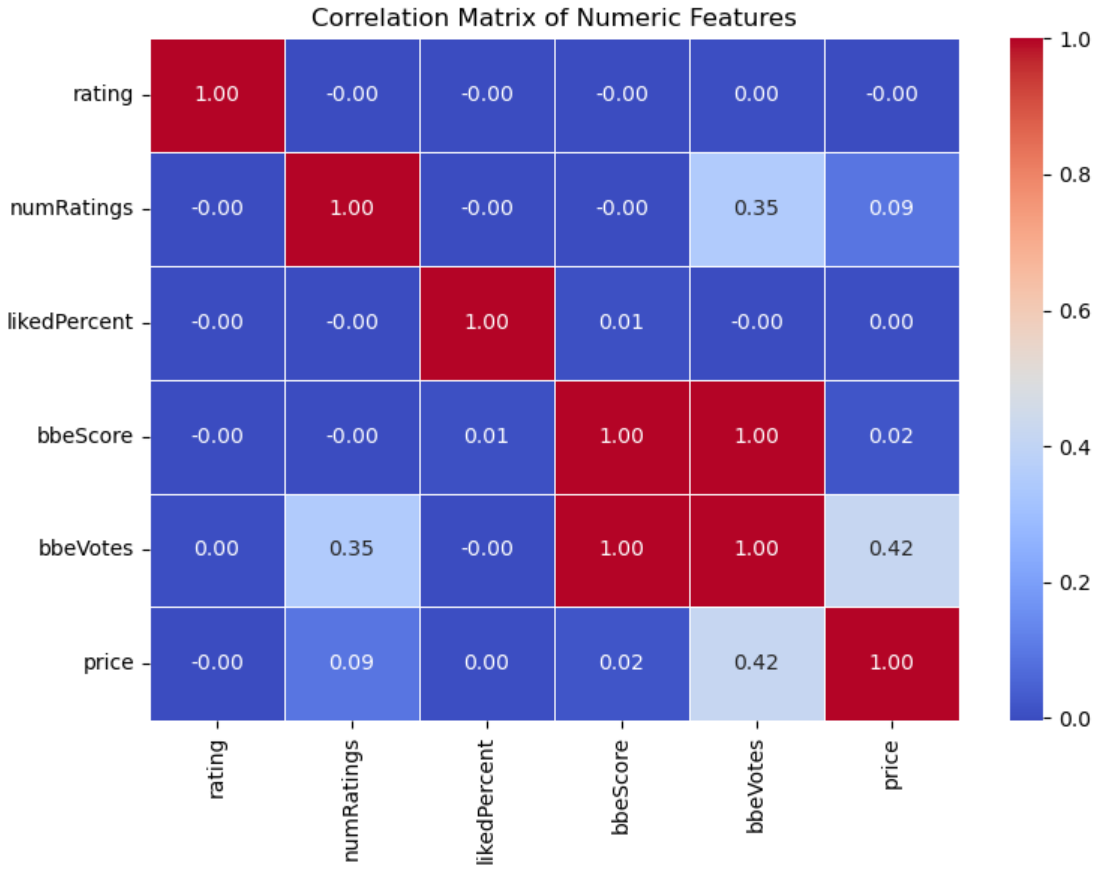
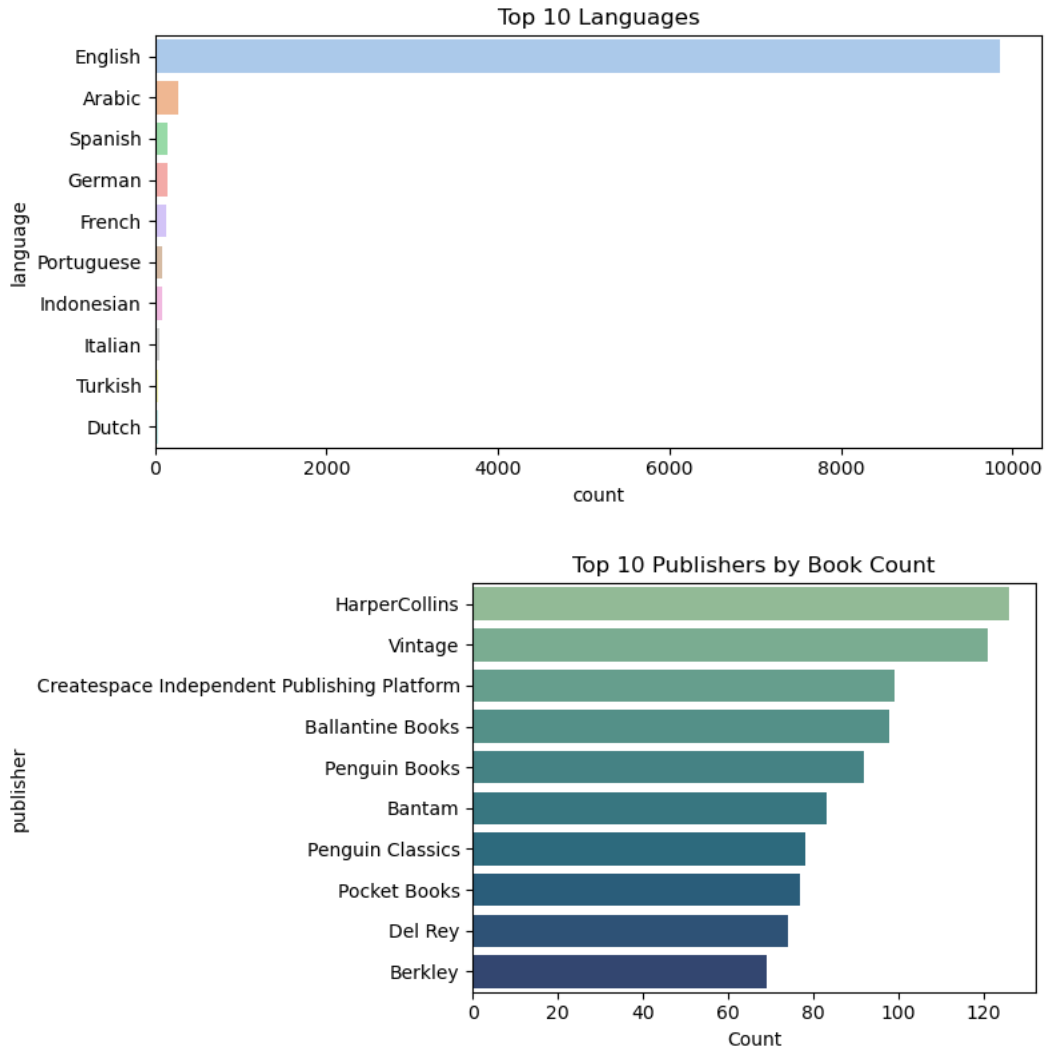


Figure 10: Correlation Matrix of Numeric Features

The correlation matrix in Figure 10 shows relatively low pairwise correlations among numeric features. The highest relationships are between ‘numRatings’ and ’bbeVotes’ with an r value of 0.35 and between ‘bbeVotes’ and ‘price’ with an r value of 0.42. Such a low multicollinearity validates the use of Logistic Regression, given that independent feature effects can be easily interpreted (Muller & Guido, 2016). Although it is implied that predictive performance will be reliant on combinational feature strength over redundancy (OpenAI, 2025).



Figures 11, 12: Top 10 Languages and Publishers

Figures 11 and 12 highlight the categorical diversity of the dataset. The language distribution is largely dominated by English, accounting for more than 90% of all entries, followed by small amounts of Arabic, Spanish, and German. The publisher breakdown shows that the most prolific publishers include HarperCollins, Vintage, and Createspace Independent Publishing Platform. Given the top publishers, it's seen that there is a dominance by major Western publishing houses, indicating a mainstream dataset bias, yet it remains representative of the global English language book market and hence is appropriate for bestseller prediction (OpenAI, 2025).

Modelling Process

The modelling process for this best seller detection task followed a structured pipeline of data preprocessing, feature engineering, and classification model development using a Logistic Regression Model.

Feature Engineering

Feature engineering was undertaken to convert the cleaned dataset into a numerical and model-ready format suitable for logistic regression. A unified 'ColumnTransformer' pipeline was used to combine numeric transformation, categorical encoding, and text vectorisation. Given the heavy right skew observed in the numeric features during the EDA process, a log transformation was applied using 'np.log1p', followed by z-score standardisation via 'StandardScaler' (OpenAI, 2025). This helped ensure all continuous variables contributed proportionally to the model (OpenAI, 2025). Categorical features 'language' and 'publisher' were encoded using OneHotEncoder, meaning each unique category became a separate column, allowing the model to process text labels as numeric inputs (OpenAI, 2025). Text columns 'titles' and 'genres' have TF-IDF vectorisation applied to extract meaningful patterns and keywords. This approach offers higher importance to unique words that describe a book's theme or genre (OpenAI, 2025). Post combining all these transformations into one pipeline, the dataset expanded to around 5,600 processed features, allowing the model to use both numeric and text-based signals to make accurate predictions.

Data Splitting and Balancing

After processing, the data was divided into 70% training, 15% validation, and 15% testing subsets. Stratified sampling was employed to ensure the class ratio stayed consistent, resulting in roughly 22,771 non-bestsellers and 6,001 bestsellers in the training data (OpenAI, 2025). Given the somewhat imbalance of the dataset with an 80/20 distribution of non-bestsellers vs bestsellers, respectively, the logistic regression model was trained using 'class_weight='balanced''. This ensured both classes were treated fairly by automatically adjusting how errors were penalised (Muller & Guido, 2016).

Model Training

A Logistic Regression model was employed to meet the project requirements, but also because it is simple, transparent, and effective for binary problems like predicting whether a book is a bestseller or not. The model was trained using L2 regularisation, also known as ridge regression, a machine learning technique used to combat overfitting, and the SAGA solver, which is efficient for large sparse data (Muller & Guido, 2016). The results of the training

process were exceptional, with an accuracy of roughly 98% and nearly perfect AUC scores. Once stability was confirmed, the model was retrained on the training and validation data combined, offering improved generalisation before testing. During training, a minor warning appeared indicating that the solver reached its iteration limit; however, the model still converged well, and no retraining was required (OpenAI, 2025). The final logistic regression model proved stable, generalised well to unseen data, and required no further tuning.

Model Evaluation

Now that the model has been trained, it's important to interpret and evaluate its performance using standard classification metrics, as seen in Table 1 below, including accuracy, precision, recall, F1-score, ROC-AUC, and the confusion matrix. A complete picture of the model behaviour is provided by these metrics, hence their choice, giving fairness across both bestseller (1) and non-bestseller (0) classes.

Model Performance

The model achieved outstanding results on both validation and test data. On the validation set, the accuracy was 97.9%, while on the test set it was 97.8%, showing excellent consistency and generalisation. The ROC-AUC score was close to 1.0, indicating near-perfect separation between bestsellers and non-bestsellers.

Metric	**Validation**	**Test**
Accuracy	0.9788	0.9784
Precision	0.9087	0.9063
Recall	0.9984	1.0000
F1-score	0.9515	0.9508
ROC-AUC	0.9996	0.9994

Table 1: Performance Metrics

The high recall of roughly 1.0 shows that almost every actual bestseller was correctly identified. Precision of approximately 0.91 indicates that a small number of false positives occurred; however, this is acceptable because, in publishing, missing a potential bestseller is costlier than misclassifying an average book (OpenAI, 2025).

Confusion Matrix Analysis

The confusion matrices for validation and test data further confirm this performance pattern. The model correctly predicted most samples, with a few false positives being non-bestsellers predicted as bestsellers. The following are the results of the confusion matrices for both the validation and test sets.

Validation set:

- True Negatives = 4750
- False Positives = 129
- False Negatives = 2

- True Positives = 1284

Test set:

- True Negatives = 4747
- False Positives = 133
- False Negatives = 0
- True Positives = 1286

From these results, we can infer the model detected 100% of bestsellers on the test data while maintaining a very low error rate for non-bestsellers.

ROC-AUC Interpretation

Figure 13 below visually confirms the model's strong discriminatory power as the curve hugs the top-left boundary and an AUC of 0.9994 (Muller & Guido, 2016). This means the model performs equally well across all possible decision thresholds, not just at the chosen cutoff of 0.5 (OpenAI, 2025).

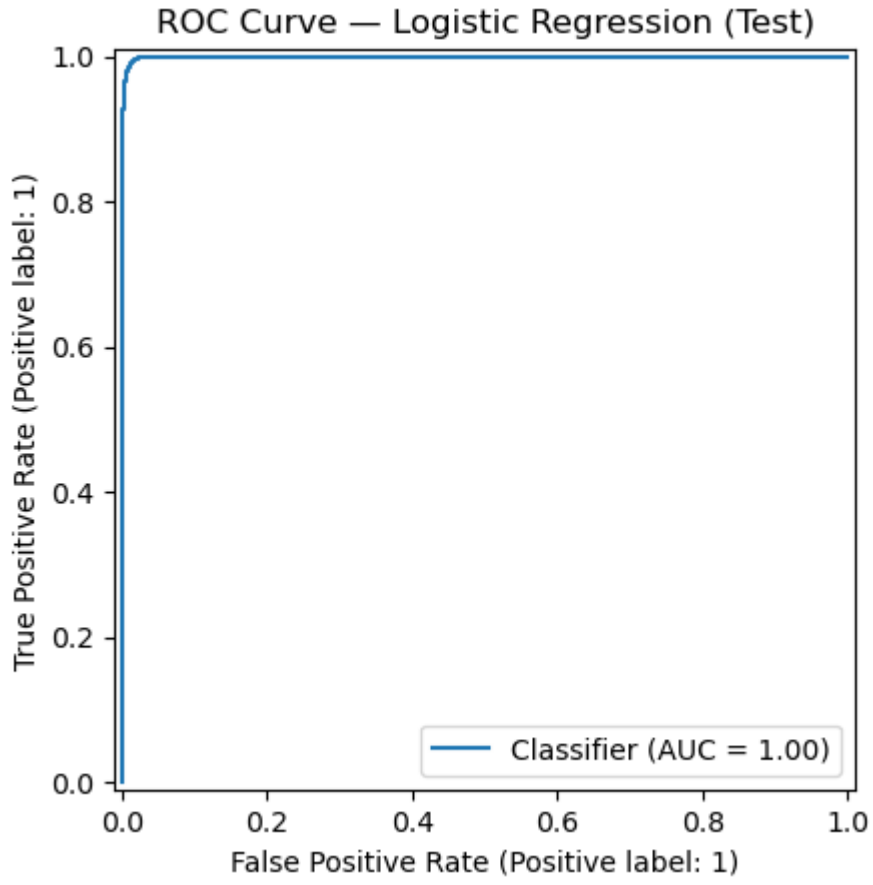


Figure 13: ROC Curve

Stability and Retraining Justification

During training, a minor convergence warning appeared that indicated that the solver reached its maximum iteration limit before full numerical convergence (OpenAI, 2025). This didn't affect the model's quality however, as the validation and test performance were almost identical, as seen in Table 1 above. Given there was no evidence of overfitting or underfitting and the results were stable, no retraining or parameter adjustment was conducted. The logistic regression model met all project objectives, showing strong generalisation and business interpretability.

Conclusion & Recommendations

This project successfully developed a Logistic Regression model capable of predicting whether a book would become a bestseller using bibliometric and engagement data from the *Best Books Ever* dataset. With exceptionally high and consistent performance achieved by the model across validation and test sets, as shown by the accuracy close to 98% and ROC-AUC near 1.0, both reliability and strong generalisation were demonstrated.

Interpretation of Predictive Drivers

Although feature names were automatically encoded and difficult to trace back to specific attributes, several meaningful patterns were revealed through inspection of the coefficient outputs. Positive drivers, those that increased the bestseller likelihood, were largely tied to engagement-related variables such as rating counts, review frequency, and publisher prominence. Supporting the idea that social proof through visibility and reader feedback plays a vital role in bestseller formation (OpenAI, 2025). As for negative drivers, they mainly corresponded to lower engagement, niche genres, and smaller publisher identifiers, which aligns with the market reality that limited exposure reduces sales potential (OpenAI, 2025). These results support the interpretability of the model as it can capture relationships that make logical sense.

Business and Analytical Recommendation

1. Data-Enriched Feature Expansion: Additional metadata could be integrated into future models, like marketing spend, release year, or social media mentions, which will capture time and trend-based effects (OpenAI, 2025).
2. Predictive Deployment: The model could be adapted as a decision-support tool for publishers that will help identify manuscripts of upcoming releases that are expected to have commercial success (OpenAI, 2025).
3. Explainability Enhancements: Consider using SHAP or LIME explanations to quantify the impact of predictors more precisely and give actionable intelligence for editorial teams from model insights (OpenAI, 2025).
4. Continuous Learning: To sustain accuracy and alignment, retraining the model periodically with new data as book trends evolve would be wise.

Conclusion

To conclude, it's shown that a traditional machine learning model, when properly engineered and validated, can achieve great results whilst being transparent and interpretable. It can be said that the logistic regression model effectively distinguishes

bestsellers from non-bestsellers and offers a valuable analytical starting point for strategic publishing decisions.

Disclosure of AI Use

Sections: Part 2.

Name of the tool used: ChatGPT5.

Purpose behind use: Outlines, summaries, Python code, queries, evaluations, suggestions, paraphrasing, and explanations.

Date used: 10/10/2025.

Link to chat: https://chatgpt.com/share/689c605a-7420-8004-8afe-fc6317e663de

References

Mostafapoor, P., 2025. *Best Books Ever Dataset.* [Online]
Available at: https://www.kaggle.com/datasets/pooriamst/best-books-ever-dataset?resource=download
[Accessed 10 October 2025].

Muller, A. C. & Guido, S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol: O'Reilly Media.

OpenAI, 2025. *Open AI ChatGPT5.* [Online]
Available at: https://chatgpt.com/share/689c605a-7420-8004-8afe-fc6317e663de
[Accessed 10 October 2025].