

PDAN8412 POE Part 2



October 30, 2025

Allana Morris

ST10204772

Contents

[Dataset Justification 4](#_Toc212756510)

[Overview of the Dataset 4](#_Toc212756511)

[Data Quality and Suitability 4](#_Toc212756512)

[Feature Availability and Relevance 4](#_Toc212756513)

[Suitability for Logistic Regression 4](#_Toc212756514)

[Analysis Planning 6](#_Toc212756515)

[Exploratory Data Analysis (EDA) 6](#_Toc212756516)

[Objective 6](#_Toc212756517)

[Data Cleaning and Inspection Steps 6](#_Toc212756518)

[Feature Selection 7](#_Toc212756519)

[Objective 7](#_Toc212756520)

[Feature Evaluation and Selection Strategy 7](#_Toc212756521)

[Feature Engineering and Transformation 7](#_Toc212756522)

[Train Model 8](#_Toc212756523)

[Objective 8](#_Toc212756524)

[Model Architecture 8](#_Toc212756525)

[Hyperparameter Selection and Tuning 8](#_Toc212756526)

[Class Imbalance Mitigation 8](#_Toc212756527)

[Data Partitioning Strategy 9](#_Toc212756528)

[Interpret and Evaluate Model 9](#_Toc212756529)

[Objective 9](#_Toc212756530)

[Performance Indicators 9](#_Toc212756531)

[Model Diagnostic Visualisations 9](#_Toc212756532)

[Success Criteria 10](#_Toc212756533)

[Write a Report 10](#_Toc212756534)

[Objective 10](#_Toc212756535)

[Report Section Structure 10](#_Toc212756536)

[Report Formatting and Presentation 11](#_Toc212756537)

[Model Evaluation 12](#_Toc212756538)

[Report 13](#_Toc212756539)

[Data Cleaning and Preparation 13](#_Toc212756540)

[Initial Data Overview 13](#_Toc212756541)

[Outlier Detection and Feature Transformation 13](#_Toc212756542)

[Feature Engineering and Standardisation 13](#_Toc212756543)

[Exploratory Data Analysis 14](#_Toc212756544)

[Numerical Feature Distributions 14](#_Toc212756545)

[Bestseller Class Distribution 15](#_Toc212756546)

[Feature Correlations 15](#_Toc212756547)

[Genre Distribution 16](#_Toc212756548)

[Feature Distributions Stratified by Target Class 17](#_Toc212756549)

[Comparative Box Plot Analysis 18](#_Toc212756550)

[Model Training 18](#_Toc212756551)

[Feature Engineering Strategy 18](#_Toc212756552)

[Initial Model Development and Optimization 19](#_Toc212756553)

[Class Imbalance Mitigation and Final Model Training 19](#_Toc212756554)

[Model Evaluation and Results 20](#_Toc212756555)

[Overall Performance Metrics and Comparative Analysis 20](#_Toc212756556)

[Classification Performance Analysis 21](#_Toc212756557)

[Receiver Operating Characteristic Analysis 24](#_Toc212756558)

[Precision-Recall Trade-off Analysis 25](#_Toc212756559)

[Feature Importance and Model Interpretability 26](#_Toc212756560)

[Decision Threshold Optimization 27](#_Toc212756561)

[Conclusions 28](#_Toc212756562)

# Table of Figures

[Figure 1: Distribution of Numerical Features 14](#_Toc212756469)

[Figure 2: Breakdown of Bestseller distribution. 15](#_Toc212756470)

[Figure 3: Correlation Matrix 15](#_Toc212756471)

[Figure 4: Genre Breakdown 16](#_Toc212756472)

[Figure 5: Violin Plot of Distribution of Numerical by Bestseller 17](#_Toc212756473)

[Figure 6: Box Plot of Numerical Features by Bestseller 18](#_Toc212756474)

[Figure 7: Training Loss Curve of Final Model 19](#_Toc212756475)

[Figure 8: Comparative Bar graph of Metrics between Initial and Final Models 20](#_Toc212756476)

[Figure 9: Change in Metrics from Initial to Final Model 21](#_Toc212756477)

[Figure 10: Confusion Matrix for Initial Model 22](#_Toc212756478)

[Figure 11: Confusion Matrix for Final Model 23](#_Toc212756479)

[Figure 12: ROC AUC Curve Comparison 24](#_Toc212756480)

[Figure 13: Precision-Recall Curve Comparison 25](#_Toc212756481)

[Figure 14: Top Features by Coefficient 26](#_Toc212756482)

[Figure 15: Threshold Optimization 27](#_Toc212756483)

# Dataset Justification

## Overview of the Dataset

The Goodreads Books 100K dataset from Kaggle by Dhamani (2021) was selected for this analysis. This dataset has more than one hundred thousand individual book entries and is thus very suitable for statistical and machine learning analysis. The dataset is very rich in features explaining each book in great detail. Such information ranges from the title of the book, the author's name, to the star rating out of 5, number of ratings, number of reviews, and number of pages. Such attributes provide an overview of the performance and characteristics of each book and hence make it useful for predictive modelling. The main goal of this analysis was to find out whether a book can be classified as a bestseller by its measurable attributes.

## Data Quality and Suitability

First, the quality of the dataset and its preparedness for analysis were checked. The data was fairly consistent and well-structured, though with a few missing values and some numeric variable outliers (Dhamani, 2021). There were missing values, but those were not in the important fields of the data; being field that were used to train the model. These were not very important to the predictive task, and thus such records were removed. It is acceptable to delete records in non-essential fields where missingness is limited and non-informative in data preparation (Little & Rubin, 2020). Outlier analysis showed extreme values in some variables related to the number of ratings and text reviews. The values were valid, but they really do need a transformation in order not to skew the model. This was done via a logarithmic transformation that reduces the influence of highly skewed data while preserving the relative ranking of books by popularity (Figiel, n.d.; Lakhloufi, 2025). The data was also checked for duplication, and nothing major was detected (Dhamani, 2021). Thereafter, the target variable, which should tell whether a book is a bestseller, was constructed from the available attributes so it would be balanced and meaningful for classification. Generally, after these cleaning and preparation steps, the dataset was adequate for modelling since it maintained integrity, diversity, and statistical reliability.

## Feature Availability and Relevance

The features in this dataset were richly varied and incorporated a range of variables relevant for the prediction of bestseller status. The average rating, number of ratings, and number of text reviews reflected the reception and popularity of a book among its readers, all of which would likely make up strong predictors of success (Dhamani, 2021). Further informed by other variables such as the number of pages, these give insight into the physical and temporal characteristics, respectively, of each of these books, which might influence their appeal to readers. Categorical features of author and genre provided context about authorship and audience but required numerical encoding before they could be used in the model. Each of these variables added insight into the characteristics of successful and less popular books. The continuous and categorical data combined to provide a multi-dimensional record of performance for each book, and as such, the features are both relevant and adequate for a robust predictive model.

## Suitability for Logistic Regression

The problem of predicting whether a book is a bestseller is inherently binary, since each book either meets or does not meet the bestseller threshold. Thus, logistic regression is an appropriate modelling choice since it estimates the probability that an observation belongs to one of two classes (Hilbe, 2009). It yields coefficients which are interpretable in terms of their relationship with the likelihood of a book becoming a bestseller. This interpretability is a key advantage when communicating results to stakeholders who may not have a technical background. Features in the dataset are numerical or can be transformed into numerical representations, thus fitting the assumptions of logistic regression. Relationships between predictors and the outcome are likely to be monotonic, so logistic regression is able to pick up the general direction of these relationships (Hilbe, 2009). In turn, the sample size of the dataset is large enough that any estimates of the model parameters are statistically stable and the model generalises well. These combined reasons constitute the justification for considering logistic regression a suitable and effective modelling technique for this classification task.

# Analysis Planning

## Exploratory Data Analysis (EDA)

### Objective

Exploratory data analysis is primarily aimed at ensuring the data is clean, well-understood, and ready for modelling through systematic exploration and cleaning. The phase identifies data quality issues, missing values, outliers, and underlying patterns that will inform subsequent modelling decisions.

### Data Cleaning and Inspection Steps

First, the dataset will be loaded and inspected for its basic properties. The pandas’ methods such as describe and info are used to understand the dimensions and data types in the dataset, respectively. Basic information regarding the number of records, features, and their compositions will be documented.

Missing value analysis will be done systematically across columns. The existence count and percentage in each feature will be identified. It will involve making decisions on each feature on whether to remove missing values or impute them using the mean, median, or some forward filling strategy. The records with critical missing information will be completely removed because ensuring the quality of the data outweighs quantity concerns.

The cleaning process will involve the detection and removal of duplicate entries to ensure that no repeated records compromise the analysis. Descriptive statistics will then be calculated for all numeric features, including mean, median, standard deviation, minimum, maximum, and quartile values. These statistics will then provide a thorough overview of the distribution of the features and possibly pinpoint anomalies and other potential data quality issues.

Visualizations of the distributions of all numerical features will be done through the use of histograms and boxplots to understand their distribution and skewness. Outliers will be identified using the interquartile range method, where values more than 1.5 times the IQR away from the quartiles will be identified as potential anomalies. The decision on the retention, transformation, or removal of outliers will depend on whether they are data errors or real extreme values.

Features that are not normally distributed in a symmetric manner will be transformed using appropriate mathematical functions such as log transformations to normalize their distributions and stabilize variance. The target variable will be created from the available features, defining the binary classification problem of predicting whether a book is a bestseller or not. Class balance will be assessed to understand the proportion of each target category.

The relationships between these different features will be explored using correlation analysis to check for potential multi-collinearity and redundancy between the features. Relationships between these will be visualized through a correlation heatmap, which flags features that might show high positive or negative correlation. Exploratory scatter plots and pair plots will be generated to visualize relationships between features and target variables to find out which predictors show the strongest associations with the outcome.

## Feature Selection

### Objective

The aim of feature selection is to select the most informative predictors and to remove the redundant or noisy variables. It involves critical assessment of each characteristic with regard to relevance with the target variable and its usefulness in improving model performance.

### Feature Evaluation and Selection Strategy

Every feature in this dataset will be judged on its theoretical relevance for the prediction of bestseller status. Numeric features regarding book characteristics, like length, ratings, and engagement metrics, will be checked for their predictive power. The categorical ones, like author, genre, and language, will be evaluated based on their potential influence on market success and reader interest.

These will be administrative features in nature, which do not serve a predictive purpose. These kinds of features bring noise and not signal to the model and hence should not form a part of the analysis. The reason behind keeping or discarding each feature will be clearly documented.

Statistical methods will be employed to reinforce subjective evaluation. Correlation analysis will identify those features highly correlated with each other, pointing toward redundancy. Features with very low correlations with the target variable will be considered for removal.

### Feature Engineering and Transformation

Features showing right-skewed distributions will be logarithmically transformed to stabilize the variances and thereby enhance the performance of models. This especially holds for count variables, such as review counts or engagement metrics.

Categorical variables will be appropriately encoded for use within logistic regression. One-hot encoding will first be used on categorical features, which means there would be binary indicators created for each category. For such features with high cardinality, either a category grouping or encoding strategies will be considered to reduce the number of features created.

Interaction terms can be created that describe the combined effect of multiple features. For instance, the interaction between rating and engagement may help to understand the quality engagement synergies. Derived metrics can also be engineered, such as ratios or aggregations, which merge multiple base features.

Numeric features will be standardized or normalized to put them on similar scales. Standardization techniques that are robust with respect to median and interquartile range will be used; these will be fitted only on the training set in order to avoid data leakage. This ensures features with large magnitudes do not dominate the learning process.

Therefore, after engineering and transformation, multicollinearity shall be assessed again. Those features with correlation coefficients above a certain threshold shall be considered for removal in order to reduce model complexity. Backward elimination via statistical significance testing will help further refine the feature set by removing only those features that contribute meaningfully towards predicting the target.

## Train Model

### Objective

Model training aims to build an optimally hyperparameter-tuned logistic regression model. The goal of this stage is to find the best combination of a learning rate, strength of regularisation, and number of training iterations that results in minimal loss and good generalisation on unseen data.

### Model Architecture

A custom logistic regression classifier will be implemented from scratch, with the aim of demonstrating an understanding of the underlying mathematics. The model will use a sigmoid activation function to output probabilities between zero and one. The cost function will be binary cross entropy loss combined with L2 regularisation in order to prevent overfitting and improve model generalisation.

The model will be trained using gradient descent optimisation. In its learning process, it has to compute the gradients of the loss function with respect to model weights and bias and then update these in the direction that reduces the loss. This is an iterative process for a pre-defined number of epochs until convergence.

### Hyperparameter Selection and Tuning

Hyperparameter optimisation will be performed using the grid search methodology. Several candidate values of each hyperparameter will be specified, and all combinations will be systematically evaluated.

Learning rate tuning will be used to determine the step size for weight updates. If the learning rate is too small, convergence will be very slow, while if it is large, the optimisation may diverge. Values to be tested would range from very small to moderate.

The number of training epochs, which defines the number of complete passes through the training data needed for convergence, will be tuned. Too few epochs might lead to underfitting, and too many will cause overfitting or unnecessary computation. It will test a range of epoch values.

The regularisation parameter lambda will be tuned to control the strength of the L2 penalty applied to weights. A larger value of lambda will increase the regularisation penalty, thus forcing weights to take smaller values and thereby reducing model complexity. A smaller value will allow greater flexibility but also increases the likelihood of overfitting. Several values of lambda will be tried in order to find an optimal balance.

The selection criterion will be F1 score, not accuracy. F1 score provides a balanced measure of model performance that accounts for both false positives and false negatives. Using F1 score is even more important in case of class imbalance, since the accuracy can be misleading when one class dominates.

### Class Imbalance Mitigation

Any class imbalance, if present, where one class is much more frequent than the other, will be dealt with using resampling techniques. Random oversampling will be utilized in order to increase the representation of the minority class within the training data. The samples from the minority class will be duplicated to approximately balance out the classes. In this way, the model will see enough examples from each class during training.

### Data Partitioning Strategy

The dataset will be divided into three subsets with distinct roles. The training set, which will comprise the majority of the data, will be used exclusively for fitting model parameters via gradient descent. The validation set will be held separate and used during training for monitoring model performance and overfitting. The test set will be completely held out and used only for the final evaluation of the model after all hyperparameter tuning and training decisions are made.

Stratified random sampling will be used for all data splits to ensure that class proportions are maintained across all three sets. This is particularly important in the case of imbalanced classification problems. Cross validation using k fold stratified sampling may be employed during hyperparameter selection to ensure robust evaluation.

Interpret and Evaluate Model

### Objective

Model evaluation will estimate the effectiveness of the model by using various performance metrics and visualizations. This thorough assessment shall, therefore, determine how well the model discriminates between the two classes and whether the performance is acceptable for the intended application.

### Performance Indicators

Accuracy will be the proportion of correct predictions among all the predictions. While this metric is intuitive, it can be misleading when there are imbalanced datasets, since the baseline model predicting the majority class could give quite high accuracy without actually being useful.

Precision will be calculated as the ratio of positive predictions that turn out to be correct. This metric answers the question: when the model predicts the positive class, how often is this prediction correct? Precision is important when false positives are costly or undesirable.

Recall will be calculated as the share of actually positive cases that the model correctly identifies. This metric addresses the following question: out of all actual positive cases, how many are correctly found by the model? Recall is relevant in cases where the cost of false negatives is high or when comprehensive identification of the positive class is important.

F1 score will be calculated as the harmonic mean of precision and recall. This metric provides a single balanced measure that weighs both types of errors equally. The F1 score is especially useful for imbalanced classification problems and will be used as the main metric when comparing models.

ROC AUC will be calculated, representing the area under the receiver operating characteristic curve. This curve plots the true positive rate against the false positive rate across different classification thresholds. ROC AUC measures the model's ability to discriminate between classes independent of any specific threshold choice.

### Model Diagnostic Visualisations

Confusion matrixes will be created for both training and validation sets to display the counts of true positives, true negatives, false positives, and false negatives. The matrix will clearly indicate where the model makes the right and wrong predictions, and from those, the insight about the type of mistakes that are made will be drawn.

Loss curves for training loss and validation loss across all training epochs will be plotted. The curves will provide insight into whether the model is converging properly, overfitting (diverging curves), or underfitting (both high). Accuracy curves will also show the training and validation accuracy per epoch to monitor the progress of learning.

This will plot a true positive rate against the false positive rate, creating an ROC curve. The area under this curve will be the ROC AUC score. A precision recall curve will show, for various classification thresholds, the trade-off between precision and recall helping identify the optimal threshold for the intended application.

A detailed classification report will be created showing per class metrics including precision, recall, F1 score, and support values. This will provide insight into if the model performs equally well on both classes and exhibits differential performance.

### Success Criteria

The model will be successful when it can give acceptable performances across multiple metrics. The target values will need to be identified for the business context and requirement of the classification task. Generally speaking, the model should achieve strong performance with respect to F1 score and ROC AUC, with recall being particularly important if identifying all positive instances is critical.

The model should generalize well. That means validation performance must be close to the training performance. Any large difference between the training and validation metrics would indicate overfitting. The convergence should be smooth, with stable loss curves showing consistent improvement across epochs.

Write a Report

### Objective

The report aims to document the whole analysis in a professional way, showing all the processes clearly. The report will communicate findings from data cleaning, exploratory analysis, feature engineering, model development, and evaluation in a manner that denotes rigorous methodology with reliable results.

### Report Section Structure

The data cleaning and preparation section will introduce the overview of the raw dataset, including records and features. This includes detailing which records were removed as invalid, why they were removed, and the final dataset size post-cleaning. Strategies for dealing with missing values within each feature will be discussed. Outlier detection results will be presented, showing which features had extreme values and the rationale on how they were handled. Feature engineering steps will be summarized, including encoding methods for categorical variables and scaling methods for numeric features. The approach used for splitting the data will be explained, including the proportions for training, validation, and test sets.

Description of Central Tendency and Spread: The exploratory data analysis section will present descriptive statistics on all numeric features. Distribution visualizations will show original distributions and the effect of transformations. A correlation heatmap will identify issues of multicollinearity. Boxplots will show outlier distributions before and after transformation. Target variable distribution will be presented, showing balance between classes.

The feature selection and engineering section will justify the retention or removal of each original feature and explain the categorical encoding method used. The final number of features after engineering will be reported. Correlation analysis results will demonstrate how redundant features were identified and removed. Results of statistical significance testing will show which features were retained based on their meaningful contribution to predicting the target.

The model training section describes the architecture of the logistic regression, including the activation function, the loss function, and the regularisation approach. It documents the hyperparameter tuning process, including which values of parameters are tested and based on what criteria the best parameters are chosen. Justification is provided for why the chosen optimal values of parameters are best, with convergence and performance criteria. If class imbalance was addressed, the resampling strategy is explained and justified.

The model evaluation and results section will include thorough metrics comparing performance across multiple measures. Confusion matrices will be displayed for visual inspection. The ROC curve will be presented, with the AUC score highlighted. Any trade-offs between metrics will be discussed and justified in the context of the business problem.

What the results mean in practical terms will be explained in the interpretation and insights section. If interpretable, the section includes a discussion on the relative importance of different features. The business implications will be explored in light of the practical utility of the model. Consider limitations of the analysis-both in data, methodological assumptions, or potential biases.

The conclusion section will summarize the key findings and state whether the model reaches its intended objectives. Recommendations for future enhancements may include collecting more features, trying different algorithms, or gathering more data. The conclusion will reiterate the main contribution of the analysis.

### Report Formatting and Presentation

The report will maintain a succinct and professional tone without excessive length or wordiness. Where appropriate, tables will be used to present comparisons and metrics in easily interpretable formats. Visualisations including plots and heatmaps will be included with descriptive captions. Mathematical notation will be used for key formulas. The report will follow Harvard referencing style for citations to academic or technical sources.

# Model Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
|  | Initial Model | Final Model | Improvement |
| Accuracy | 0.9701 | 0.9767 | 0.0066 |
| Precision | 0.9271 | 0.9488 | 0.0216 |
| Recall | 0.9332 | 0.9417 | 0.0084 |
| F1-Score | 0.9302 | 0.9452 | 0.0150 |
| ROC AUC | 0.9947 | 0.9973 | 0.0027 |
| Best Improvement: Precision (+0.0216) | | | |
| Worst Change: ROC AUC (+0.0027) | | | |

Table 1: Detailed Model Comparison

The model evaluation compares the initial versus final model on five key performance metrics. The accuracy has increased from 0.9701 to 0.9767, hence a gain of 0.0066. That means this retrained model predicts correctly more often than the original one and gives roughly 98 percent accuracy on bestsellers and non-bestsellers together.

Precision had the most significant boost among the metrics, increasing from 0.9271 to 0.9488, with a gain of +0.0216. This is the best improvement observed in the retraining process. The higher the precision, the more certain it is that for each prediction the final model makes as to whether or not a book is a bestseller, it will be correct about 95% of the time, as opposed to the 93% rate of the initial model. From a business perspective, this is informative because it avoids wasting resources on false positives.

Recall moved from 0.9332 to 0.9417, an increase of 0.0084. While this gain is smaller than the precision improvement, it remains important for the core objective of identifying bestsellers. The final model now captures approximately 94 percent of actual bestsellers, up from 93 percent in the previous model. That is, the model has become more successful in recognizing those books that eventually turn out to be bestsellers.

This corresponds to an F1-score that increases from 0.9302 to 0.9452, with an improvement of 0.0150. This means that this model improved the overall trade-off between precision and recall since the retraining managed to improve both instead of one at the cost of the other.

The ROC AUC increased from 0.9947 to 0.9973, which had the smallest increase of 0.0027. Since both models already had near-perfect ability to discriminate across any threshold of classification, this change is numerically very small. With the final model's AUC at 0.9973, it remains excellent in ranking books on their likelihood of being bestsellers.

Improvement was realized in the retraining of all five metrics with no trade-off or degradation in any one area. Precision yielded the most significant increase at 0.0216, while ROC AUC had the least improvement at 0.0027. Improvement in these different dimensions of classification performance indicates that the final model is more reliable along all dimensions of classification performance.

# Report

## Data Cleaning and Preparation

### Initial Data Overview

The Goodreads 100k Books dataset, obtained from Kaggle, consisted of 100,000 records of books, each represented by 13 features that included book characteristics, reader engagement metrics, genre classifications, and platform-specific information. Initial inspection showed 7,752 records with zero page counts, indicating missing or corrupted data; these were removed entirely from the dataset. A thorough check for duplicates revealed none. Sparse missing values in the critical columns were treated by removing rows, while median imputation was chosen where necessary to maintain data integrity without bias.

The total valid records in the dataset, after cleaning, were 92,248, which is a 7.5 percent reduction from the original figure of 100,000. The cleaned dataset was then divided into training and testing sets by stratified random sampling, with 73,798 samples (80 percent) for training and 18,450 samples (20 percent) for testing. Stratified sampling maintained representative class proportions in both sets.

### Outlier Detection and Feature Transformation

It finds outliers across all numerical features. However, instead of removing them, logarithmic transformations have been performed on highly right-skewed engagement metrics. Specifically, log(reviews + 1) and log(totalratings + 1) were computed to normalize the distributions while preserving information from extreme values. These transformations converted multiplicative relationships into additive relationships more suitable for linear models and reduced the influence of outliers during gradient-based optimization.

### Feature Engineering and Standardisation

The genre information was one-hot encoded to avoid spurious ordinal relationships. All numeric features were standardized using robust scaling based on the median and interquartile range; these statistics were computed only on the training data to avoid leakage, then identically applied to the test data. After this entire process, the final feature set contained 17 engineered features that included original variables, logarithmic transformations of some of those, interaction terms, and one-hot encoded genres.

## Exploratory Data Analysis

### Numerical Feature Distributions

A graph of a graph of a number of data

AI-generated content may be incorrect.

Figure 1: Distribution of Numerical Features

Examination of numerical feature distributions revealed critical patterns informing subsequent modelling decisions. The rating variable showed left-skewed distribution concentrated between 3.5 and 4.5 stars with peak near 4.0, indicating favourable ratings predominated. The reviews variable exhibited extreme right-skewness with approximately 60,000 books clustered near zero and a long tail extending to thousands. The pages variable displayed approximately normal distribution centred around 250 to 300 pages with multimodal peaks suggesting genre conventions. The totalratings variable mirrored the reviews pattern with approximately 70,000 books at the origin and extreme values extending to 4 million ratings. These distributions directly justified logarithmic transformations of engagement metrics to enable effective linear model learning across the full engagement spectrum.

### Bestseller Class Distribution

A screen shot of a graph

AI-generated content may be incorrect.

Figure 2: Breakdown of Bestseller distribution.

The target variable analysis revealed 78,000 non-bestsellers (78.6 percent) and 21,000 bestsellers (21.4 percent). This moderate class imbalance necessitated careful attention during model training to prevent bias toward the majority class. Although less extreme than typical severely imbalanced datasets, the imbalance remained sufficient to render accuracy inadequate as an evaluation metric, as naive models could achieve 78.6 percent accuracy through systematic non-bestseller prediction.

### Feature Correlations

A screenshot of a grid

AI-generated content may be incorrect.

Figure 3: Correlation Matrix

The correlation heatmap revealed reviews and totalratings showed strong correlation of 0.87, indicating substantial collinearity and overlapping measurement of reader engagement. Rating showed near-zero correlation with all other variables (0.03 with reviews, 0.07 with pages), indicating that average rating operated independently as a distinct book characteristic. Pages showed minimal correlation with all other features (0.03 with reviews and totalratings), confirming that book length operated independently from engagement metrics. These relationships informed feature selection and model interpretation, validating the inclusion of multiple engagement measures while highlighting the distinction between quality (rating) and engagement (reviews, totalratings).

### Genre Distribution

A graph of orange bars with names

AI-generated content may be incorrect.

Figure 4: Genre Breakdown

The top ten genres showed balanced distribution with Romance at approximately 34,000 books, followed by Fantasy at 30,000, Fiction at 29,000, and Nonfiction at 29,000. Remaining genres ranged from approximately 13,000 to 18,000 books. This diversity indicated that any learned patterns needed to generalise across multiple content types rather than reflecting genre-specific phenomena.

### Feature Distributions Stratified by Target Class

A screenshot of a graph

AI-generated content may be incorrect.

Figure 5: Violin Plot of Distribution of Numerical by Bestseller

Violin plot comparisons revealed stark differences between bestsellers and non-bestsellers for engagement metrics. Rating showed substantial overlap between classes with both centring around 4.0 stars, indicating limited discriminatory power. Pages showed similar overlap around 250 pages, confirming minimal length-based differentiation. Conversely, reviews showed non-bestsellers concentrated near zero with minimal spread, while bestsellers demonstrated dramatic extension to approximately 160,000 reviews with pronounced right tail. Totalratings displayed analogous patterns with non-bestsellers at the origin and bestsellers extending to approximately 4 million ratings. These visualisations provided compelling evidence that engagement metrics, not quality or structural characteristics, distinguished bestsellers from non-bestsellers.

### Comparative Box Plot Analysis

A group of graphs with lines

AI-generated content may be incorrect.

Figure 6: Box Plot of Numerical Features by Bestseller

Box plots confirmed and quantified the violin plot findings. Reviews showed bestseller median substantially exceeding non-bestseller median with extended interquartile range. Totalratings demonstrated identical patterns. Conversely, rating and pages showed minimal differences between classes with substantial overlap in distributions. This analysis reinforced that engagement metrics possessed strong discriminatory power while structural and quality metrics lacked predictive value.

## Model Training

### Feature Engineering Strategy

Strategic feature engineering created derived variables to capture non-linear relationships. Logarithmic transformations of reviews and totalratings addressed right-skewness. An interaction term (rating × totalratings) captured synergistic effects of quality and engagement. Review engagement ratio and title length captured additional dimensions. The final 17-feature set combined original variables, transformations, interactions, and encoded genres.

### Initial Model Development and Optimization

Initial model training employed custom logistic regression with gradient descent optimisation using a learning rate of 0.01, 1,000 epochs, and regularisation parameter of 0.1. Despite initial convergence to loss 0.18 by epoch 1,000, test set evaluation revealed critical performance issues. Accuracy reached 98.37 percent but recall for the bestseller class measured only 4.78 percent with precision of 93.75 percent and F1-score of 0.0909. The model had effectively defaulted to non-bestseller prediction, achieving high accuracy through majority class bias while failing the core objective of bestseller identification.

A comprehensive grid search with five-fold cross-validation explored 27 hyperparameter configurations varying learning rates (0.001, 0.01, 0.1), regularisation parameters (0.01, 0.1, 1.0), and epochs (500, 1,000, 2,000). Critically, model selection employed F1-score rather than accuracy, ensuring optimisation toward balanced performance on both classes. The optimal configuration featured learning rate 0.01, regularisation 0.01, and 2,000 epochs.

### Class Imbalance Mitigation and Final Model Training

Random oversampling balanced the training set to 50 to 50 class representation by duplicating bestseller examples. This expanded the training set from 73,798 to 145,084 samples and eliminated majority class bias in gradient calculations.

A graph with a green line

AI-generated content may be incorrect.

Figure 7: Training Loss Curve of Final Model

The final model's training curve demonstrated rapid initial descent from 0.70 at epoch zero to approximately 0.20 by epoch 200, indicating faster convergence than the initial model. Loss continued decreasing to approximately 0.06 by epoch 2,000, representing improved convergence quality and substantially lower final loss than the initial model's 0.18 at 1,000 epochs. The smooth, consistent decrease throughout training confirmed successful model adaptation to balanced data.

## Model Evaluation and Results

### Overall Performance Metrics and Comparative Analysis

A graph of blue and orange bars

AI-generated content may be incorrect.

Figure 8: Comparative Bar graph of Metrics between Initial and Final Models

Grouped bar comparisons quantify the consistent improvements between the initial and final models across all evaluation metrics. The original model achieved an accuracy of 0.9701, which improved to 0.9767 in the final model for a gain of 0.66 percentage points. Precision increased from 0.9271 to 0.9488, increasing 2.16 percentage points and representing the strongest performance gain across all metrics. Recall improved from 0.9332 to 0.9417, contributing 0.84 percentage points of advancement. The F1-score improved from 0.9302 to 0.9452, up 1.50 percentage points and reflecting balanced improvements in both precision and recall. ROC AUC increased from 0.9947 to 0.9973, reflecting a 0.27 percentage point enhancement in the model's discriminative ability across all classification thresholds.

The precision improvement to 0.9488 is the most prominent success point, whereby the model improved in correctly identifying positive instances and being highly confident in the predictions. Meanwhile, its high recall at 0.9417 has been maintained, ensuring the reliable detection of positives across the prediction cycle. Overall, improvements on all metrics indicate consistency in refinement rather than trade-offs, where the model performs better overall without sacrificing any dimension of evaluation.

A graph showing a number of green bars

AI-generated content may be incorrect.

Figure 9: Change in Metrics from Initial to Final Model

The change visualization chart clearly depicts these improvements. Accuracy, precision, recall, F1-score, and ROC AUC all display green bars indicating positive improvement in model performance. The magnitude of precision improvement, measuring 0.0216 in absolute terms, visually dominates the chart, highlighting this metric's critical importance in the model refinement process. Even the smallest improvement in ROC AUC (0.0027) reflects the already high baseline performance, demonstrating that the final model achieves superior classification capability while maintaining the strong discriminative characteristics of the original implementation.

### Classification Performance Analysis

The fundamental difference between models emerged through their classification behaviours: the first model had severe bias toward the non-bestseller class because of class imbalance in the training data and using accuracy as an optimization metric.

A blue squares with white text

AI-generated content may be incorrect.

Figure 10: Confusion Matrix for Initial Model

The confusion matrix from this initial model showed 15,433 true negatives from 15,714 non-best sellers with 281 false positives, but only 3,983 true positives from 4,286 actual bestsellers with 303 false negatives. This pattern demonstrated catastrophic minority class underperformance despite the majority class's accuracy and established a baseline from which improvement was measured. It was imply making positive predictions at extremely low rates, essentially deferring to the majority class.

A blue and white chart

AI-generated content may be incorrect.

Figure 11: Confusion Matrix for Final Model

On the confusion matrix, the final model predicted 15,390 true negatives out of 15,714 non-bestsellers with 324 false positives and 4,174 true positives from approximately 4,286 actual bestsellers with 112 false negatives. This represented 97.94 percent specificity on the majority class and 97.38 percent sensitivity on the minority class-a balanced and practical performance of classification. Compared to the initial model's 303 false negatives, the reduction to 112 represented approximately 63 percent improvement in missed bestsellers. The resampling strategy forced the model to learn patterns from both classes equally, fundamentally changing its decision boundaries.

### Receiver Operating Characteristic Analysis

A graph with a line

AI-generated content may be incorrect.

Figure 12: ROC AUC Curve Comparison

The ROC curve comparison presents both the initial model (blue, AUC 0.995) and the final model (orange, AUC 0.998) that have excellent discrimination capability, with curves hugging the upper-left corner and positioned far above the random baseline diagonal. Whereas the initial model showed an AUC of 0.995, indicating strong probabilistic discrimination, the final model improved to an AUC of 0.998, representing near-perfect discrimination. Both curves positioned dramatically above the random classifier baseline confirm real predictive power across all thresholds. Such improvement from 0.995 to 0.998, while modest in appearance, represented consistent enhancement across the entire operating range. The key difference did not lie in the discrimination of either individual model but in how probabilities are translated into discrete predictions through the combined effects of threshold selection and resampling strategy, allowing the final model to achieve substantially higher recall while sustaining competitive precision.

The ROC curve for the final model revealed an almost ideal performance, with the curve hugging the upper-left corner at most operating points. The curve rose quickly from the origin to a true positive rate of about 1.0 at a false positive rate of about 0.05, remaining close to the upper-left corner across false positive rates 0.0 to about 0.15, before transitioning more gradually. An AUC of 0.998 indicated near-perfect discrimination capability. The curve lay significantly above the random classifier baseline, the diagonal dashed line at AUC 0.50, confirming real predictive power across all thresholds. This AUC of 0.998 represented a 0.3-percentage-point improvement on the initial model's 0.995, showing resampling and hyperparameter optimization improved discrimination uniformly across operating regions.

The ROC curve for the first model showed high discrimination with an AUC of 0.995, hugging the upper left corner and well above the random baseline. Despite excellent AUC, the poor recall performance highlighted the disconnection between probabilistic discrimination and practical classification under class imbalance.

### Precision-Recall Trade-off Analysis

A graph with a line

AI-generated content may be incorrect.

Figure 13: Precision-Recall Curve Comparison

The precision-recall comparison showed the curve for the first model (in blue) maintaining high precision approximately 0.99 to 1.0 across recall 0 to approximately 0.85, then dropping steeply, reflecting precision bias with average precision at 0.982. On the other hand, the final model curve (orange) showed that it maintained high precision, 0.99 to 1.0 across recall 0 to approximately 0.50 and then decreased gradually towards zero, achieving average precision at 0.991. This superior balance of the final model across higher recall values represents genuine performance improvement, and its stakeholders have flexibility toward picking operating points that better match the requirements on precision and recall. The trade-off was clear: whereas the initial model sacrificed recall to achieve precision, the final model balanced both metrics more effectively. Granted, the average precision of the final model, 0.991, was above that of the initial model, 0.982. But more importantly, this was accomplished while driving practical recall levels above 78 percent.

The precision-recall curve for the first model remained high, at about 0.982, as recall ranged from 0 to about 0.85, which is then displayed as a plateau before steeply dropping off at high recall. Precision dominance was reflected in an average precision score of 0.982, but low overall recall rendered the model unsuitable for the business objective.

### Feature Importance and Model Interpretability

A graph with red and green squares

AI-generated content may be incorrect.

Figure 14: Top Features by Coefficient

The coefficients from the learned model showed feature contributions to bestseller prediction. *reviews* was the strongest positive predictor, representing the bestseller probability increasing significantly for every unit increase in log-transformed reviews. *log\_reviews* reflected the importance of transformed engagement metrics. *totalratings* had a substantial positive contribution. The *rating\_x\_totalratings* interaction confirms that engineered interactions captured meaningful non-linear information combining the dimensions of quality and engagement.

On the other hand, *title\_*length insinuated slight negative association with bestseller status. *review\_engagement\_ratio* exhibited inverse relationship. *pages* confirmed negligible predictive power. These coefficients validated that reader engagement metrics dominated bestseller prediction, with quality (rating) and engagement (reviews, totalratings) components combining multiplicatively rather than additively for optimal prediction.

### Decision Threshold Optimization

A graph of a number of different colored lines

AI-generated content may be incorrect.

Figure 15: Threshold Optimization

The threshold analysis showed the F1-score (blue, circles), precision (orange, squares), and recall (green, triangles) across thresholds 0.1 to 0.8. At a threshold of 0.1, the F1-score was about 0.75, the precision 0.60, and recall 1.0; this captured nearly all bestsellers with reduced precision. For a threshold set at 0.5, the F1-score went up to 0.94, precision increased to 0.90, and recall went up to 0.98. For a threshold between 0.6 and 0.7, the peak F1-score, at approximately 0.95, gave a precision of 0.95 and recall of 0.86, which is the best balance that can be used for general applications. At a threshold of 0.85, the F1-score decreased to 0.93, with the precision going up to 0.98 and recall falling to 0.87, indicating over-conservative prediction. Using this chart, stakeholders could select the thresholds that match their specific business needs. The application that prioritizes comprehensive identification of bestsellers would want to use a lower threshold, whereas cost-sensitive applications that need high precision will select higher thresholds.

## Conclusions

A high-performing logistic regression classifier was developed for predicting Goodreads bestselling books by following systematic data cleaning, comprehensive exploratory analysis with visualization, strategic feature engineering, careful model training with hyperparameter optimization, and rigorous multi-metric evaluation. The final model achieved 97.67 percent accuracy, 94.17 percent recall, 94.88 percent precision, 94.52 percent F1-score, and 99.73 percent ROC AUC, representing consistent improvement over the initial model.

This project proved that systematic hyperparameter optimization and careful model refinement could develop a strong baseline model into one delivering superior performance across all evaluation metrics. The improvement of precision by 2.16 percentage points enhanced the model's capability to correctly identify positive instances while maintaining confidence in prediction. The gain in recall by 0.84 percentage points ensured even more reliable detection of positive cases, whereas an increase by 1.50 percentage points in F1-score reflected balanced improvements across both metrics.

The model's insights seemed to confirm that reader engagement metrics, particularly review volume, dominated bestseller prediction, while structural characteristics such as page length and genre classifications were of minimal predictive power. This finding aligns with market dynamics where visibility and community engagement drive commercial success independent of intrinsic book qualities or content type.

The threshold optimisation analysis showed that the model allowed flexibility in deployment across a wide variety of business scenarios and enabled the selection of thresholds that could meet specific precision-recall requirements. The near-perfect ROC AUC of 0.9973 confirmed that the model's probabilistic outputs reflect real discriminative capability, not threshold-dependent artifacts.

# References

Dhamani, M. 2021. *GoodReads 100k books*. Available: https://www.kaggle.com/datasets/mdhamani/goodreads-books-100k/data [2025, October 30].

Figiel, M. n.d. *Data Transformation - Sustainability Methods*. Available: https://sustainabilitymethods.org/index.php/Data\_Transformation [2025, October 30].

Hilbe, J.M. 2009. Logistic Regression Models. *Logistic Regression Models*. (May, 11). DOI: 10.1201/9781420075779.

Lakhloufi, H. 2025. *When Data is Skewed, Log Transformations Save the Day!* Available: https://metricgate.com/blogs/log-transformations-skewed-data/ [2025, October 30].

Little, R.J.A.. & Rubin, D.B.. 2020. Statistical analysis with missing data. 449. Available: https://www.oreilly.com/library/view/statistical-analysis-with/9780470526798/f03.xhtml [2025, October 30].