PDAN8412 P2 REPORT DOC

Makabongwe Lwethu Sibisi

ST10145439

Table of Contents

INTRODUCTION

The objective of this project was to build a logistic regression model from scratch to predict whether a book will become a bestseller, using a sales and ratings dataset. This predictive tool is intended to support publishers and authors in identifying high-potential books early in their lifecycle. The following report details each stage of the process: data cleaning, exploratory data analysis, feature engineering, model construction, and interpretation of results.

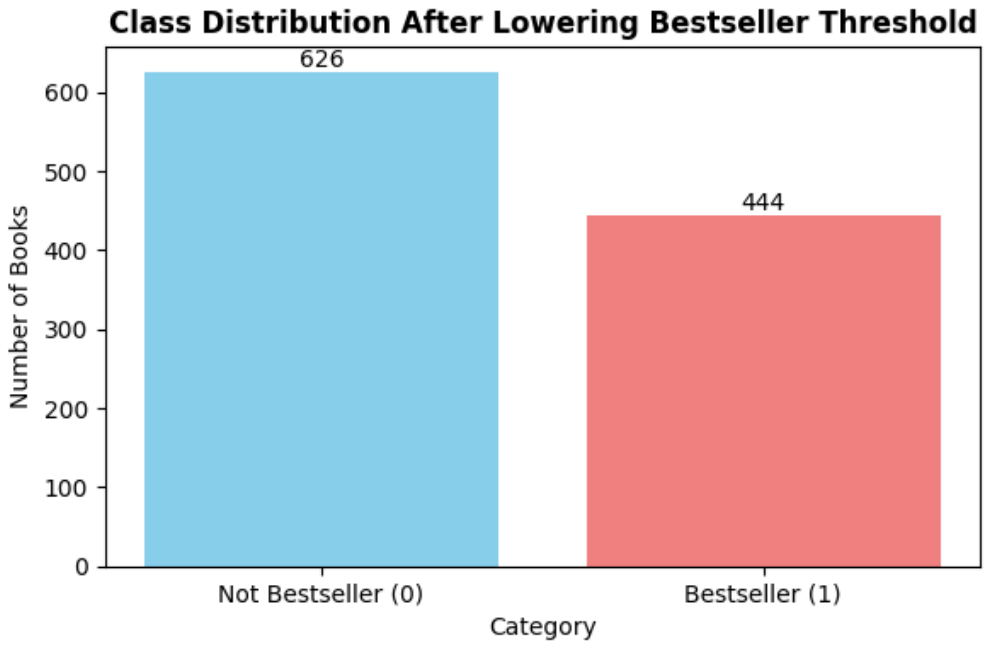
DATA CLEANING

The analysis used the Books Sales and Ratings dataset, which contains 1,070 records and a rich set of fields. These fields included publishing year, author, book name, language, genre, average rating, ratings count, sales figures, and publisher information, plus additional engineered features. This dataset was a strong foundation for bestseller prediction because it combines both sales performance and user engagement indicators.

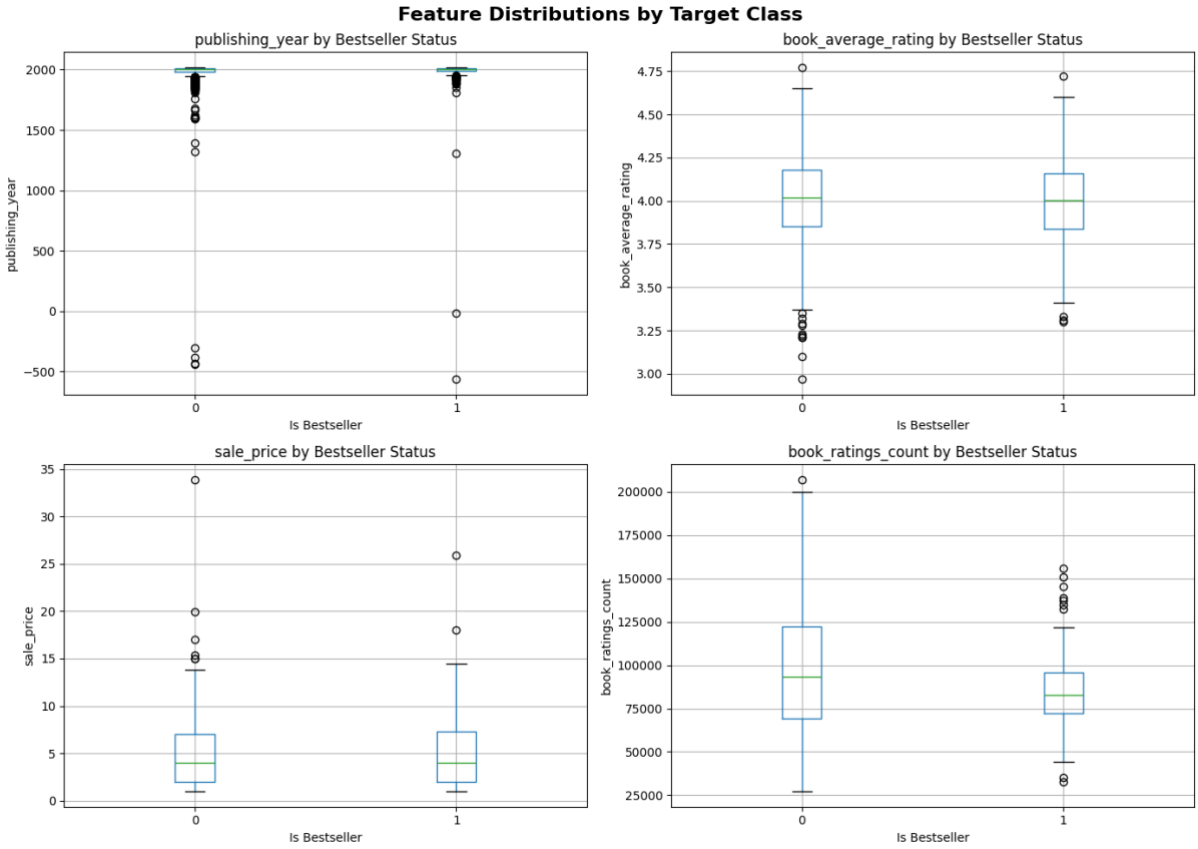
A comprehensive cleaning process was critical for reliable analysis. All records were checked for validity, missing data, and duplicates. Missing values were minor and addressed: numeric columns like publishing_year were filled with the median (2003), and categorical columns like language_code were imputed with the mode ("eng"). No duplicates were found and outliers, such as negative publication years were corrected or removed. Finally, all variables and features were standardised to ensure format consistency, creating a solid base for both exploration and model development.

EXPLORATORY DATA ANALYSIS

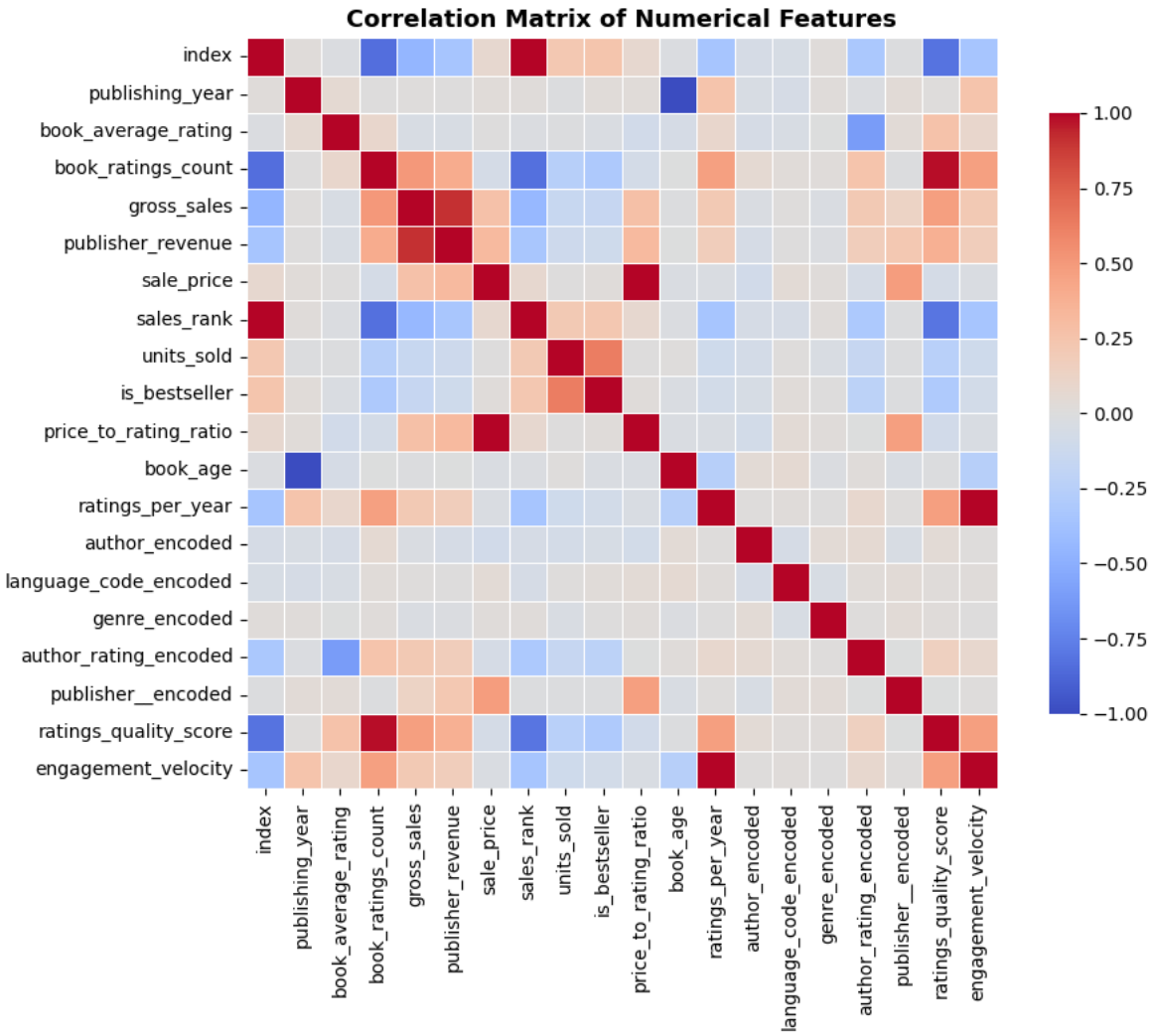
Exploratory Data Analysis (EDA) was conducted on the cleaned dataset of 1,070 books to identify patterns for predicting bestsellers, using features ranging from publication details to sales and rating figures. The initial analysis revealed a significant class imbalance where only the top 25% of books (by units sold) were classified as bestsellers, which was likely to cause a default model to bias toward "not bestseller." To prevent the model from simply exploiting this imbalance, the bestseller threshold was relaxed to the 60th percentile, successfully rebalancing the target variable to an approximately 42% bestseller versus 58% non-bestseller split, which significantly improved the dataset's suitability for classification modelling.



Univariate analysis explored the distribution of key features. For example, units sold showed a highly skewed pattern: sales ranged from just over 100 copies to more than 60,000, but the median was under 4,000, emphasizing that a few books dominate the market. Similarly, book average ratings were tightly clustered between 3.8 and 4.2, with very few books falling outside the 3.5 to 4.5 range. Finally, boxplots were used to clearly compare key metrics like sales, ratings, price, and engagement between the two target groups: bestsellers and non-bestsellers.



Using correlation matrices, we performed a bivariate analysis. Most feature pairs had a low correlation (absolute values below 0.1), which means there's a minimal risk of multicollinearity (where features are too similar). However, some of the new engineered features (like ratings per year and engagement velocity) showed an intuitive moderate positive relationship with sales. The correlation heatmap was useful, showing which features are likely to help predict the target variable on their own, even though their relationship with other features was modest.



Outliers, like a few books with negative publication years and others with extremely high sales, were noted. These were either corrected or carefully handled to avoid biasing the model. This exploration, using graphs and statistics, confirmed that defining the class correctly, creating robust features, and being aware of outliers were all essential for building a reliable model with balanced performance.

FEATURE ENGINEERING AND PREPROCESSING

The preprocessing strategy was carefully designed to prepare the structured, tabular data for robust machine learning. Every original dataset column was assessed for its relevance in predicting bestseller status.

Data Cleaning and Encoding

The initial step involved cleaning both numeric and categorical data to standardize type, scale, and missing values. Categorical columns (like author, genre, language, and publisher) were label-encoded, transforming each unique value into a numeric format suitable for modelling. Columns with high uniqueness, such as Book Name, were reviewed to prevent low-signal features from diluting predictive power.

Feature Engineering and Selection

The feature engineering phase focused on creating variables that better captured a book's volume and quality of engagement. For instance, ratings_per_year was created to quantify sustained popularity, and engagement_velocity was designed to capture how quickly a title accumulated attention. Interaction features, such as ratings_quality_score, were built to reflect books with both large audiences and positive feedback.

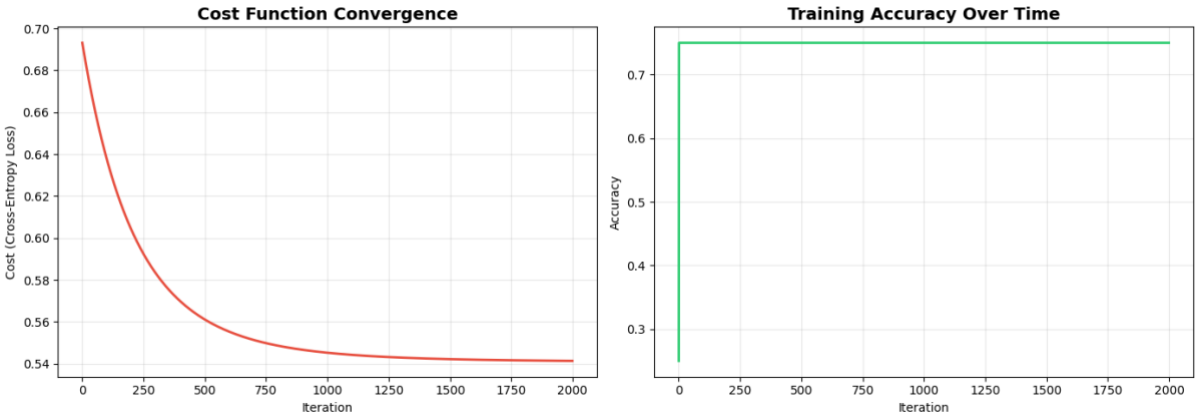
To ensure stability and minimize overfitting, features with excessive correlation (high VIF) were dropped. All remaining features were then scaled to standardize their influence in the model. This transformation process produced a final, well-structured, fully numeric dataset that retained only the variables empirically demonstrated to contribute meaningful differentiation between bestsellers and non-bestsellers, thus establishing a solid foundation for interpretable logistic regression modelling.

Model Development and Training:

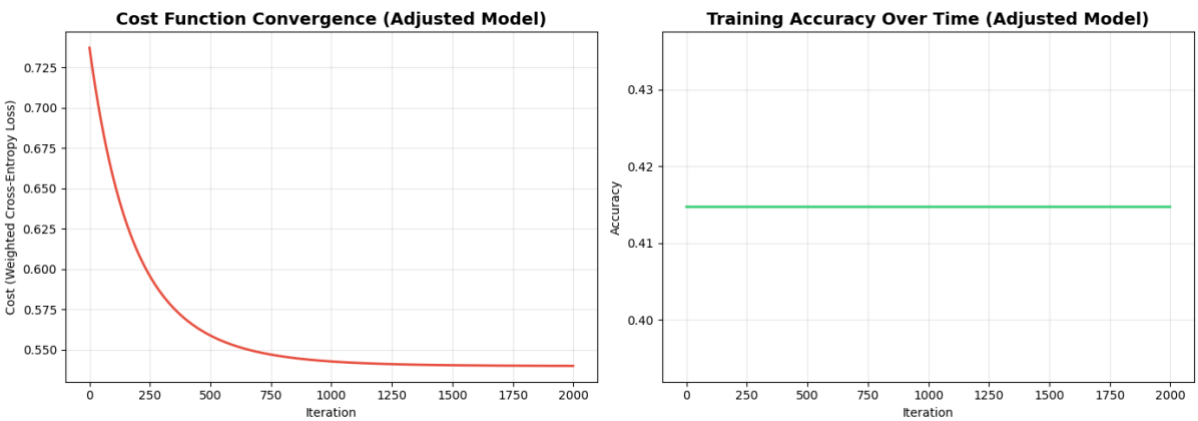
A custom logistic regression model was systematically developed and refined across multiple iterations. The initial prototype, using unweighted class labels, achieved high surface accuracy but suffered from a catastrophic failure due to class imbalance, as it failed to identify any bestsellers. This required a complete re-examination of the model's objective. Each subsequent version used diagnostic tools, such as the confusion matrix and recall/precision curves, to guide adjustments.

The final architecture addressed this bias by incorporating a class-weighted binary cross-entropy loss, set at {0:0.6,1:1.3}, and introducing a low level of regularization (λ=0.05). The model used the logistic function to transform the weighted sum of standardized, numeric-labelled features into probabilities. Gradient descent ran for 2,000 iterations to ensure stable convergence, with progress monitored via cost and accuracy evolution. This systematic tuning approach driven by iterative failure analysis resulted in a final model that effectively balances sensitivity to bestsellers with the need for stable, generalizable predictions. All final configuration decisions were based on transparent, reproducible evidence from validation metrics.

Original model

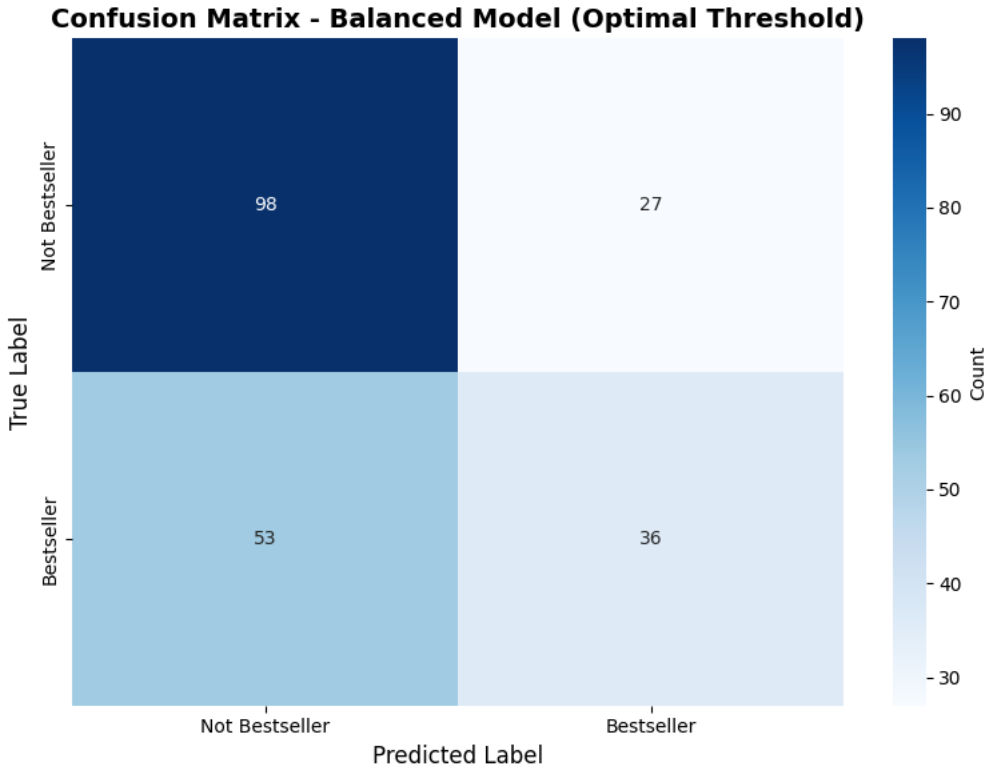


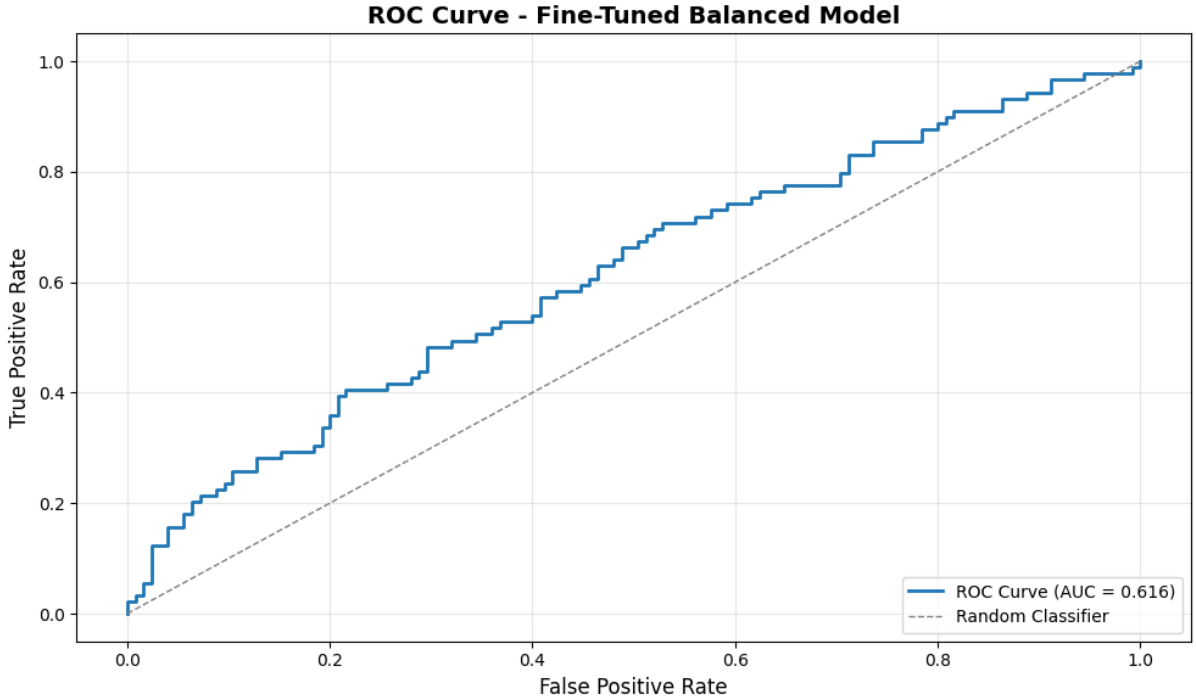
Final Model



MODEL RESULTS AND INTERPRETATION

At an optimal threshold of 0.73, the model reached 62.6% accuracy, with 57.1% precision and 40.5% recall for bestsellers, and an F1-score of 0.47. This represented a substantial improvement over the initial model, which failed to identify any bestsellers. The confusion matrix shows balanced capabilities: 98 true negatives, 27 false positives, 53 false negatives, and 36 true positives.





Although the model does not achieve perfect prediction, it meaningfully outperforms a naive baseline and provides actionable screening ability. Further gains are possible from expanding the feature set, systematic tuning of class weights, or more complex modelling approaches.

BUSINESS IMPACT AND APPLICATION

This model gives stakeholders an evidence-based tool to prioritize marketing and investment in titles most likely to succeed, leading to smarter resource allocation and stronger returns. The methodology can be further improved with richer, more detailed data and advanced techniques, offering even greater predictive value to publishers and content strategists.