Table of Contents

[Evaluation and Justification of dataset 2](#_Toc212214637)

[Planning My Analysis 3](#_Toc212214638)

[EDA 3](#_Toc212214639)

[Feature Selection 3](#_Toc212214640)

[Train Model 3](#_Toc212214641)

[Interpret and Evaluate Model 3](#_Toc212214642)

[Preliminary Report 3](#_Toc212214643)

[Final Report 4](#_Toc212214644)

[Objective 4](#_Toc212214645)

[Data Preparation 4](#_Toc212214646)

[Model Development 4](#_Toc212214647)

[Evaluation 4](#_Toc212214648)

[Insights 5](#_Toc212214649)

[Limitations 5](#_Toc212214650)

[Conclusion 5](#_Toc212214651)

[References 6](#_Toc212214652)

# Evaluation and Justification of dataset

For this project, I selected the [books-prediction/data](https://www.kaggle.com/datasets/shayanshahid997/books-prediction/data) from Kaggle to perform a logistic regression analysis. This dataset is particularly suitable because it offers a variety of relevant and well-structured fields that align with the objectives of my analysis. Additionally, the dataset has received positive usability feedback on Kaggle, indicating its reliability and ease of use for data-driven projects.

The abovementioned dataset has comprehensive sales metrics (columns such as ‘gross sales’, ‘publisher revenue’, ‘units sold’, and ‘sales rank’ offer multiple ways to determine whether a book could sell well or not at all. This allows the model to be trained either for continuous outcomes (e.g. units sold) or classify books into ‘high vs low’ sales categories.

Some challenges and considerations to keep in mind as that some rows have missing values. There is also some inconsistency in the language code column. Multiple authors listed in one field could also skew the data. Author rating should be encoded from ordinal to numeric values.

# Planning My Analysis

## EDA

I will begin by thoroughly exploring the dataset to understand its structure, distributions, and potential anomalies. This includes visualising sales-related variables, examining the distribution of categorical features such as language codes and publishers, and identifying missing values or outliers. Special attention will be paid to the 'author' and 'author rating' columns, as well as the consistency of language codes. I will also use correlation matrices to uncover relationships between numerical variables and the sales outcome (Scikit-learn Developers, 2023).

## Feature Selection

Based on the EDA findings, I will select features that are most predictive of sales performance. Variables such as 'gross sales', 'publisher revenue', 'units sold', and 'sales rank' will be prioritised. Categorical variables like publisher and language will be encoded appropriately, and I will consider aggregating or engineering new features to enhance model performance. Features with a high proportion of missing data or low variance may be excluded (GeeksforGeeks, 2023).

## Train Model

I plan to use logistic regression to classify books into 'high' versus 'low' sales categories. The target variable will be defined based on a sales threshold, such as the median units sold. The data will be split into training and test sets, and cross-validation will be used to tune hyperparameters and prevent overfitting (NumberAnalytics, 2022). I will implement standard preprocessing steps, including assigning missing values and scaling numerical features as needed (Scikit-learn Developers, 2023).

## Interpret and Evaluate Model

Model performance will be assessed using metrics such as accuracy, precision, recall, and the F1-score. I will also generate a confusion matrix to gain insight into the types of errors the model makes. Feature importance or coefficient analysis will help interpret which variables most strongly influence the predicted outcome (Towards Data Science, 2022).

## Preliminary Report

Based on the dataset's richness and structure, I expect to identify key drivers of book sales, such as publisher reputation, author rating, and sales rank. I anticipate that handling missing values and inconsistent entries will be crucial to achieving robust results. Ultimately, I expect the analysis to provide actionable insights into the factors that distinguish high-performing books, and to produce a model capable of predicting sales categories with reasonable accuracy (GeeksforGeeks, 2023; NumberAnalytics, 2022).

# Final Report

## Objective

The goal of this project was to build a predictive model that classifies whether a book is likely to sell well, based on features such as author rating, average book rating, genre, and publisher data. A binary target variable Will\_Sell was created using a threshold of 6000 units sold.

Data Preparation

* Source: A structured dataset of books with 14 features including Book Name, Author, Genre, Publisher, and Units Sold.
* Target Variable: Will\_Sell (1 if units sold ≥ 6000, else 0).
* Preprocessing:
  + Missing values were handled via row removal.
  + Categorical variables were one-hot encoded.
  + Features were scaled using StandardScaler.

Model Development

* Model Used: Logistic Regression
* Hyperparameter Tuning: Performed using GridSearchCV with 5-fold cross-validation (GeeksforGeeks, 2023).
  + Best parameters: C=1, penalty='l2', solver='liblinear'
* Cross-Validation Accuracy:
  + Mean: **98%**
  + Standard Deviation: Low, indicating stable performance across folds (NumberAnalytics, 2022)

Evaluation

* Test Set Accuracy: 98%
* Confusion Matrix: All predictions were correct on the test set.
* Predicted Probabilities: Most confident predictions were clustered near 0 or 1, indicating high model certainty.

## Insights

* Top Influential Features:
  + Book\_average\_rating
  + Author\_Rating
  + sale price
  + Certain publishers and genres
* Overlap Between Predicted and Actual Top Sellers:
  + 4 out of the top 10 predicted bestsellers matched the actual top sellers.
  + Some books were over-predicted due to high ratings or genre bias despite lower actual sales.

## Limitations

* Potential Overfitting: The model achieved perfect predictions on the test set, which may indicate overfitting — especially given the high dimensionality after one-hot encoding (Scikit-learn Developers, 2023).
* Data Imbalance: The dataset may have had more low-selling books, which could skew performance metrics.
* Feature Leakage Risk: Some features (e.g., sales rank) may correlate too directly with the target and should be reviewed for leakage.

## Conclusion

The logistic regression model performed exceptionally well, achieving 98% accuracy and correctly identifying all books in the test set. However, further validation on unseen data is recommended to confirm generalizability. Future work could explore regularization, feature selection, or alternative models such as Random Forest or XGBoost.

Additionally, a custom implementation of logistic regression was built from scratch using NumPy, following the mathematical principles outlined by Towards Data Science (2022). This exercise helped reinforce understanding of the sigmoid function, loss function, and gradient descent.

# References

* GeeksforGeeks, 2023. *How to optimize logistic regression performance*. [online] Available at: <https://www.geeksforgeeks.org/machine-learning/how-to-optimize-logistic-regression-performance/> [Accessed 24 Oct. 2025].
* NumberAnalytics, 2022. *Logistic regression accuracy: 5 factors that affect performance*. [online] Available at: [https://www.numberanalytics.com/blog/logistic-regression-accuracy-5-factors](https://www.numberanalytics.com/blog/logistic-regression-accuracy-5-factors%20) [Accessed 24 Oct. 2025].
* Scikit-learn Developers, 2023. *Logistic Regression — scikit-learn documentation*. [online] Available at: [https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression](https://scikit-learn.org/stable/modules/linear_model.html%23logistic-regression%20) [Accessed 24 Oct. 2025].
* Towards Data Science, 2022. *Understanding logistic regression from scratch*. [online] Available at: <https://towardsdatascience.com/logistic-regression-from-scratch-8d2c5f6c2f3e> [Accessed 24 Oct. 2025].