PDAN8412 POE PART 2

Makabongwe Lwethu Sibisi

ST10145439

Table of Contents

QUESTION 1

Data Needed for Logistic Regression Classification

Logistic Regression is a supervised machine learning algorithm used to predict the probability of a binary outcome (represented as 1 or 0). Unlike linear regression, it models the probability that a given input belongs to a specific class using a sigmoid function, which ensures the output remains between 0 and 1. This approach allows the model to classify outcomes, such as predicting if a book will be a bestseller, by learning from known inputs and class labels. The algorithm computes a weighted sum of input features, applies the logistic transformation, and then optimizes its coefficients using methods like gradient descent to minimize prediction error (GeeksforGeeks, 2025) (Lee, 2025).

Logistic Regression requires data containing a mix of clean, well-labelled, and relevant structured numerical and categorical features that influence the target variable. The Books Sales and Ratings dataset by Josh Murrey from Kaggle is ideal, as it includes core attributes such as author, language, average book rating, genre, and sales metrics like gross sales and units sold. These variables are perfect for the model because they can be standardised or numerically encoded, which allows the algorithm to learn patterns associated with bestselling performance. Furthermore, the dataset's structured nature, moderate size, and meaningful quantitative relationships between predictors and the binary outcome make it a strong foundation for implementation (Murrey, 2023).

Data Quality Considerations & Common Pitfalls:

Common pitfalls like missing values, inconsistent formats, and outliers distort the model's interpretation of relationships between features and the target variable. Multicollinearity, where two or more features are highly correlated, can also bias coefficient estimation and reduce model interpretability. Furthermore, imbalanced datasets, where one class (such as non-bestsellers) dominates, may cause the model to favour the majority class, leading to poor prediction accuracy on the minority. The Books Sales and Ratings dataset addresses these issues with structured data, minimal missing values, and well-defined categories including author, language, ratings, sales, and genre. But with that being said it requires careful preprocessing. This includes scaling numerical features, encoding categorical variables, and handling potential skews in sales data for optimal performance. Paying close attention to these factors ensures the resulting model is robust, interpretable, and accurately identifies patterns distinguishing bestselling books (Ranganathan, et al., 2017).

Question 2

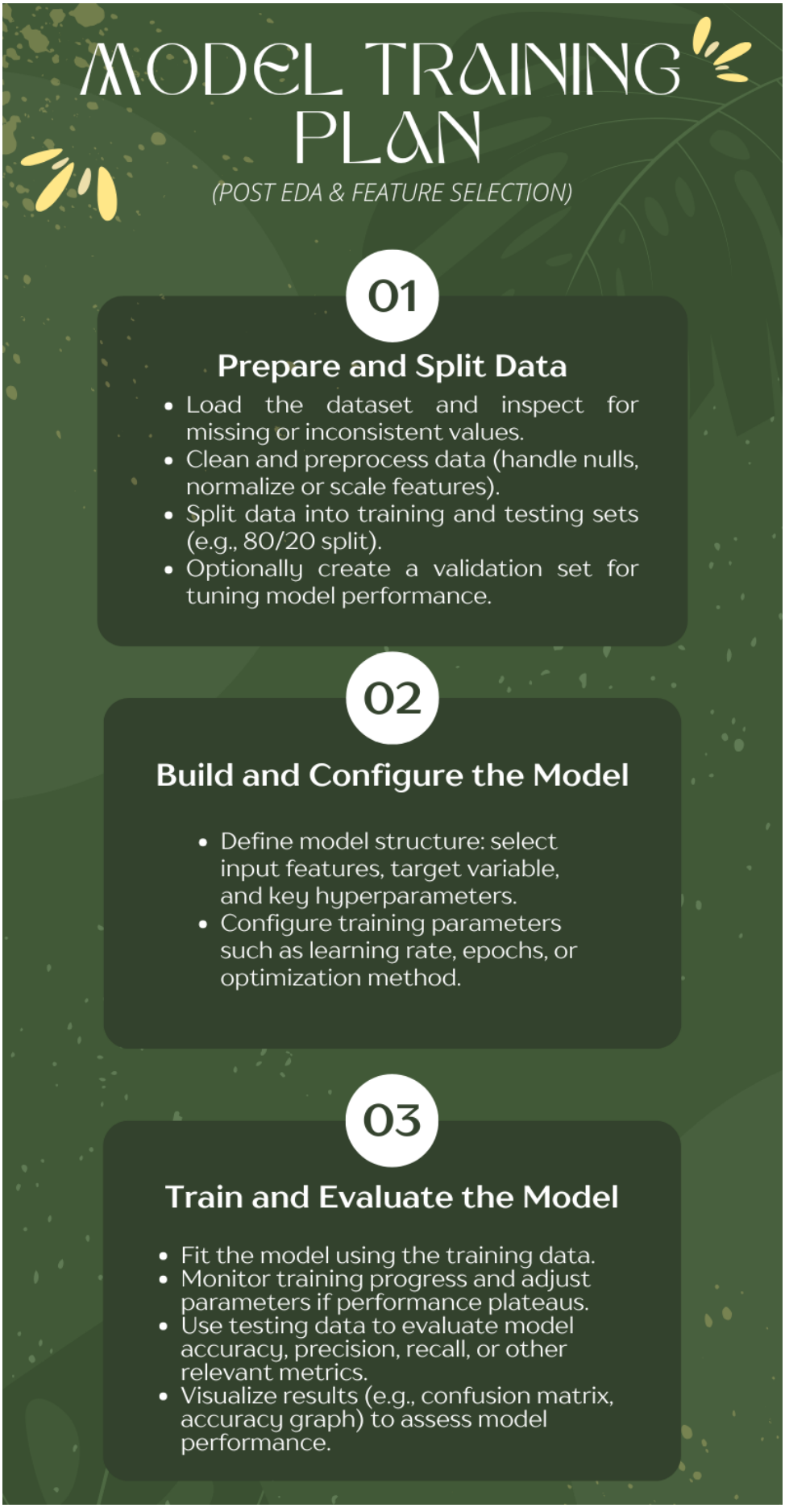
EDA

(Naralasetty, 2023)

Feature Selection

The Books Sales and Ratings dataset contains many features that could help predict whether a book will become a bestseller but using all of them can be computationally expensive, making a clear feature selection plan essential. The process begins with analysing each feature's individual relationship with the target variable, utilizing correlation for numerical data and chi-square or ANOVA tests for categorical features. Next, I will check for multicollinearity by reviewing a correlation matrix or calculating Variance Inflation Factors (VIF) to remove redundant features and enhance model interpretability. To identify the most relevant predictors, I will apply stepwise selection or regularization methods, such as the L1 penalty. Throughout this process, I will consider computational efficiency and aim for a small set of strong features that balances accuracy and performance. Finally, the selected features will be validated through cross-validation by comparing performance metrics like accuracy, precision, and AUC, ensuring the final set maintains logical consistency with predicting bestseller potential (RITHP, 2023).

Train model


MODEL TRAINING PLAN
(POST EDA & FEATURE SELECTION)
01
Prepare and Split Data
• Load the dataset and inspect for missing or inconsistent values.
• Clean and preprocess data (handle nulls, normalize or scale features).
• Split data into training and testing sets (e.g., 80/20 split).
• Optionally create a validation set for tuning model performance.
02
Build and Configure the Model
• Define model structure: select input features, target variable, and key hyperparameters.
• Configure training parameters such as learning rate, epochs, or optimization method.
03
Train and Evaluate the Model
• Fit the model using the training data.
• Monitor training progress and adjust parameters if performance plateaus.
• Use testing data to evaluate model accuracy, precision, recall, or other relevant metrics.
• Visualize results (e.g., confusion matrix, accuracy graph) to assess model performance.

Interpret and evaluate model

I will use the following evaluation metrics to interpret the results of my model:

1. **Accuracy:**
   Measures the proportion of correctly predicted instances compared to the total number of predictions, providing an overall indication of model performance.
2. **Precision, Recall, and F1-score:**
   These metrics will be obtained from a classification report.
   - **Precision** evaluates how many of the predicted positive cases are actually positive.
   - **Recall** assesses how well the model identifies all true positive instances.
   - **F1-score** combines precision and recall into a single measure to provide a balanced evaluation.
3. **Confusion Matrix:**
   A confusion matrix will be generated to visualize model predictions versus actual outcomes. This helps to identify which classes are being correctly predicted and where misclassifications occur.

I will classify the model as successful if it is able to determine whether the book will be a best seller without being susceptible to the unbalanced dataset

Report


Introduction
Explain the goal of the project and outline the predictive model developed to identify whether a book becomes a bestseller. Summarize the dataset used and
Business Impact and Application
Breakdown the key features of the dataset and any idenentifiable characteristics highlighted within the EDA
Data Cleaning
Describe the cleaning process, including handling of missing values, removal of duplicates, and standardization of data types. Explain how these steps ensured the dataset was ready for analysis.
Model Results and Interpretation
Summarize the key performance metrics obtained from testing, including accuracy, precision, recall, and F1-score. Interpret what these results indicate about the model's overall effectiveness.
Report Mind Map
EDA
Break down the key features of the dataset and highlight any identifiable patterns or characteristics found through visualizations and statistical summaries.
Model Development and Training
List the main model parameters and training setup, including learning rate, iterations, and regularization. Summarize how model adjustments and class weighting improved prediction balance.
Feature Engineering and Preprocessing
Explain how new features were created and existing features transformed for modeling. Include encoding, scaling, and any feature selection methods used to enhance model performance.

Question 3

Please find Reponses within the github repo.
https://github.com/just-makab/PDAN8412_POE_PART_2.git

Question 4

https://github.com/just-makab/PDAN8412_POE_PART_2.git

Question 5

Please refer to the Report Document.

Bibliography

GeeksforGeeks, 2025. *Logistic Regression in Machine Learning*. [Online]
Available at: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/
[Accessed 25 October 2025].

Lee, F., 2025. *What is logistic regression?*. [Online]
Available at: https://www.ibm.com/think/topics/logistic-regression
[Accessed 25 October 2025].

Murrey, J., 2023. *Books Sales and Ratings*. [Online]
Available at: https://www.kaggle.com/datasets/thedevastator/books-sales-and-ratings/data
[Accessed 25 October 2025].

Naralasetty, H. K. S. D. P., 2023. *Books Sales | EDA | Sales Prediction(99%)*. [Online]
Available at: https://www.kaggle.com/code/hknaralasetty/books-sales-eda-sales-prediction-99/comments
[Accessed 25 October 2025].

Ranganathan, P., Pramesh, C. & Aggarwal, R., 2017. *Common pitfalls in statistical analysis: Logistic regression*. [Online]
Available at: https://journals.lww.com/picp/fulltext/2017/08030/common_pitfalls_in_statistical_analysis__logistic.9.aspx
[Accessed 25 October 2025].

RITHP, 2023. *Logistic Regression for Feature Selection: Selecting the Right Features for Your Model*. [Online]
Available at: https://medium.com/@rithpansanga/logistic-regression-for-feature-selection-selecting-the-right-features-for-your-model-410ca093c5e0
[Accessed 25 October 2025].