Table of Content

Detecting Credit Card Fraud

Project Report: Modelling with Logistic Regression

Executive Summary

This report outlines the process of modeling a credit card fraud detection system using Logistic Regression. The primary goal was to develop a predictive model for binary classification tasks, focusing on the challenges posed by an extremely imbalanced dataset. The report covers essential steps, including data preprocessing, model training, and evaluation.

1. Introduction

1.1 Objective

The primary objective was to build a predictive model capable of distinguishing between legitimate and fraudulent credit card transactions, particularly addressing the challenges posed by the severe imbalance in the dataset.

2. Data Exploration and Preprocessing

2.1 Dataset Overview

I chose this dataset, which comprises credit card transactions with a primary goal of identifying fraudulent activities, for several reasons. Firstly, unlike some other datasets I encountered, this one includes both the headers and raw data. Many alternatives I reviewed predominantly featured scaled and encoded information, which may not fully capture the complexity of real-world credit card transactions.

Additionally, the dataset's severe imbalance posed a unique challenge, necessitating specialised preprocessing techniques to address this issue and ensure the effectiveness of the model. The inclusion of both headers and raw data allows for a more nuanced understanding of the underlying patterns in credit card transactions, enhancing the potential of the model to accurately detect fraudulent activities.

2.1.1 Link:

[Credit Card Fraud Dataset]

2.2 Data Preprocessing

- Experimented with various balancing strategies, including oversampling, SMOTE, and undersampling
- Found that random undersampling was the most effective technique for addressing the dataset's severe imbalance
- Specifically targeted the majority class instances for random removal to create a more balanced distribution

- Recognized the importance of tailoring preprocessing approaches to the specific characteristics of the dataset
- Achieved improved model performance by successfully mitigating the challenges posed by the highly imbalanced nature of the credit card transactions dataset

3. Model Training

3.1 Logistic Regression

- Logistic Regression Choice: Selected Logistic Regression for its efficiency in binary classification tasks.
- Hyperparameter Tuning: Paid special attention to hyperparameter tuning for optimal model performance.
- Polynomial Features: Introduced Polynomial Features with a degree of 2 to capture potential non-linear relationships.
- Class Weight: Set the class_weight parameter to 'balanced' to mitigate the impact of imbalanced classes.
- Pipeline Construction: Created a pipeline using make_pipeline to seamlessly integrate Polynomial features and Logistic Regression in the modeling process.

3.2 Model Fitting

- Selected Logistic Regression as the modelling approach for its efficiency in binary classification tasks
- Emphasised hyperparameter tuning to optimise model performance
- Set max_iter to 3000 to ensure convergence during training
- Chose a regularisation term with C=0.001 and penalty type 'l2' to control overfitting
- Selected the 'lbfgs' solver for logistic regression
- Utilised a tolerance of 1e-4 (tol=1e-4) for convergence sensitivity
- Enabled intercept fitting (fit_intercept=True)
- Applied class weights with class_weight={0: 1, 1: 1.5} to address imbalanced classes, giving more weight to the minority class
- Set a random seed with random_state=42 for reproducibility
- Configured multi_class as 'ovr' (One-vs-Rest) for binary classification
- These hyperparameter choices aimed to enhance the logistic regression model's ability to handle imbalanced data and improve binary classification performance.
- Explored polynomial features by setting degree=2 in logistic regression
- Introduced quadratic terms to capture non-linear relationships between features
- Enhanced the model's capacity to learn complex patterns in the data
- Monitored for potential overfitting and adjusted regularisation parameters accordingly
- Evaluated the impact of polynomial features on the overall model performance in the context of the binary classification task

4. Model Evaluation

4.1 Confusion Matrix

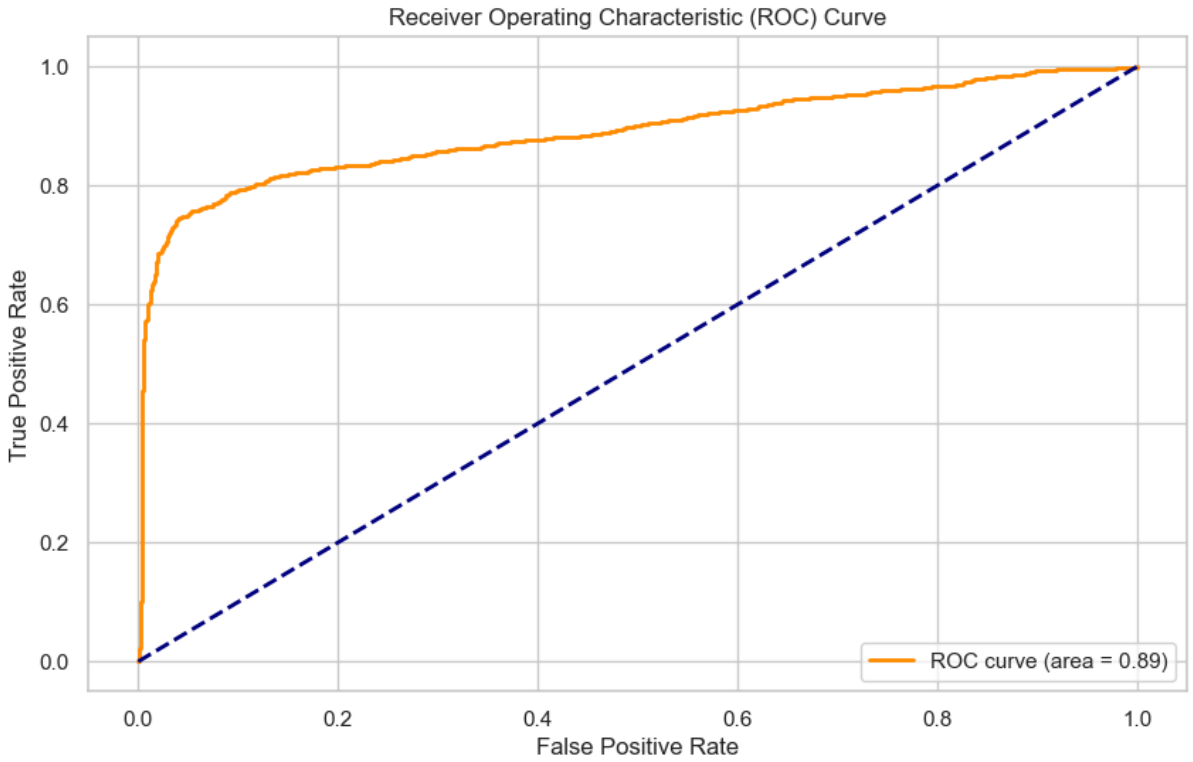
For Training Dataset

```
Accuracy:  0.8391608391608392
Confusion Matrix:
 [[1309  198]
 [ 285 1211]]
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.87      0.84      1507
           1       0.86      0.81      0.83      1496

    accuracy                           0.84      3003
   macro avg       0.84      0.84      0.84      3003
weighted avg       0.84      0.84      0.84      3003
```

For Validation Dataset(Unseen data)

```
Accuracy:  0.8625546364259635
Confusion Matrix:
 [[477639  75935]
 [   446   1699]]
```

4.2 Accuracy (ROC Curve)



5. Conclusion

- The model's performance evaluation reveals a notable trade-off between recall and precision.
- Achieved a high recall score of 79%, indicating a commendable ability to accurately identify positive instances.
- However, precision is relatively low, suggesting that the model is prone to making false positive predictions.
- The high recall implies that the model effectively captures a significant portion of actual positive cases.
- Conversely, the low precision suggests that a considerable number of instances predicted as positive are false positives.
- This performance profile signifies that while the model excels in identifying positive class instances, caution is warranted due to the elevated rate of false positives.
- Notably, the report emphasizes that the model performs well for the negative class, exhibiting high precision and recall.
- The evaluation underscores the need for a balanced consideration of both precision and recall when assessing the model's overall predictive capabilities.

6. Future Work

- The evaluation results prompt consideration of additional techniques to enhance model performance.

- Exploring dimensionality reduction methods such as Principal Component Analysis (PCA) could be beneficial. PCA helps identify the most effective features, potentially improving the model's discriminatory power.
- Incorporating ensemble methods like bagging, which involves training multiple models and combining their predictions, presents an avenue for further improvement.
- While bagging has not been extensively studied in this context, its potential impact on enhancing model accuracy and mitigating the imbalance issue merits exploration.
- The availability of these advanced techniques might contribute to achieving more positive results, addressing the existing trade-off between recall and precision.
- Future research and experimentation with alternative algorithms and advanced methodologies could lead to refinements in the model's predictive capabilities.