Credit Card Fraud Detection Using Logistic Regression

Authors: Ritu kanchi , Geet shukla
Affiliation
Date: 8/11/24

---

1. Abstract

This report presents a machine learning project focused on developing a credit card fraud detection model using logistic regression. Due to the significant financial impact of fraud, identifying fraudulent transactions is crucial for financial institutions. Logistic regression, a well-known binary classification technique, was selected for its interpretability and effectiveness in distinguishing between fraudulent and non-fraudulent transactions. To address the common issue of class imbalance in fraud datasets, we applied Synthetic Minority Over-sampling Technique (SMOTE) and IBM sampling, which enhanced data quality and improved the model's accuracy. Data visualization was conducted using Seaborn and Matplotlib, while model implementation and evaluation were performed using Scikit-Learn, with additional analysis through Keras to explore potential deep learning applications.

---

2. Introduction

Problem Definition:

Credit card fraud poses a serious challenge for financial institutions, leading to losses in the millions of dollars annually and harming consumer trust. Fraudulent activities range from unauthorized purchases to identity theft, making it crucial for institutions to have reliable detection systems.

Objective

The goal of this project is to develop a robust logistic regression model that can accurately detect fraudulent credit card transactions, thereby helping reduce financial losses and improving fraud detection efficiency.

Scope

This project focuses on binary classification using a labelled credit card transaction dataset. The report covers data preprocessing, handling class imbalance, model development, and performance evaluation, with the aim of providing a real-time fraud detection solution.

---

3. Data Collection and Preprocessing

Data Sources:
The dataset was sourced from publicly available credit card transaction datasets, often including anonymized features for transaction details.

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Data Description:
The dataset includes various features (time, amount, anonymized features V1-V28) and a target variable indicating fraud status (fraud or non-fraud).

Data Cleaning and Preprocessing:
Initial steps involved checking for missing values, handling outliers, scaling the features, and normalizing values to optimize model performance.

Data Imbalance:
As fraud transactions are rare, the dataset is highly imbalanced, with fraudulent cases comprising a small percentage of the total records. This imbalance necessitated techniques like SMOTE to improve the minority class representation.

---

4. Exploratory Data Analysis (EDA)

Descriptive Statistics:
Summary statistics were calculated for each feature, providing insights into the range, distribution, and central tendency of transaction data.

Visualizations:
Using Seaborn and Matplotlib, we visualized relationships between features and the target variable, highlighting patterns that distinguish fraudulent from non-fraudulent transactions. Notable techniques included correlation matrices, box plots, and histograms.

Insights:
EDA revealed some distinct patterns in transaction amounts and times for fraudulent vs. non-fraudulent transactions, informing our approach to model development.

---

5. Model Selection and Justification

Choice of Model:
Logistic regression was selected due to its simplicity, interpretability, and suitability for binary classification tasks. It provides clear insights into feature importance and probability outputs that can be used to prioritize high-risk transactions.

Evaluation Metrics;
Given the imbalanced nature of the dataset, metrics such as Precision, Recall, F1 Score, and AUC-ROC were prioritized. These metrics help assess the model's effectiveness, particularly in reducing false negatives.

---

6. Data Balancing Techniques

SMOTE;
The Synthetic Minority Over-sampling Technique (SMOTE) was applied to oversample the minority (fraud) class, helping the model learn from a more balanced distribution of fraud cases.

IBM Sampling
IBM sampling was employed to enhance data representativeness, ensuring that the dataset captures a range of transaction types without introducing bias.

---

7. Model Training and Evaluation

Training Process:
The model was trained using Scikit-Learn's logistic regression implementation. We used cross-validation to validate the model's performance across different subsets, reducing overfitting risk.

Results:
The logistic regression model demonstrated promising results, achieving high recall and precision rates, critical for minimizing missed fraud cases. AUC-ROC was particularly high, indicating strong discriminatory power.

---

8. Model Optimization and Fine-Tuning

Hyperparameter Tuning:

Grid Search was applied to optimize model parameters, including regularization strength (C parameter) to improve model generalization.

Final Model Performance
After tuning, the model achieved improved recall and precision, further reducing false negatives and providing a more reliable tool for fraud detection.

---

9. Results and Analysis

Key Findings:
The logistic regression model proved effective in identifying fraudulent transactions with minimal false negatives, thus reducing the likelihood of undetected fraud. A high AUC-ROC confirmed the model's ability to distinguish between classes accurately.

Confusion Matrix Analysis:
An analysis of the confusion matrix showed an acceptable balance between false positives and false negatives, indicating reliable fraud detection without overburdening financial resources.

Practical Implications:
The model could be implemented in real-time transaction monitoring systems, providing alerts for high-risk transactions and allowing institutions to take preemptive action against potential fraud.

---

10. Challenges and Limitations

Challenges Faced:
Handling data imbalance was a primary challenge, requiring careful sampling techniques to avoid overfitting. Additionally, logistic regression has limitations in capturing complex patterns, which could limit its performance in cases with more intricate fraud patterns.

Limitations:
The model's effectiveness is limited by the feature set and the specific dataset used.

Additional features or data sources could further enhance its performance, especially if applied to larger datasets or extended to deep learning approaches.

---

11. Conclusion and Future Work

Summary of Findings:
The logistic regression model effectively detected fraudulent transactions, providing a reliable tool for fraud mitigation. The model's high recall and precision rates support its applicability in real-time fraud detection.

Future Improvements:
Future work could explore ensemble models or deep learning techniques through Keras to capture more complex relationships. Additionally, real-time data monitoring could be introduced to adapt the model to changing fraud patterns.

---

12. References

- List any datasets, research papers, libraries, or tools referenced, such as Scikit-Learn, Keras, SMOTE, IBM sampling documentation, and any related articles on fraud detection techniques.

---

13. Appendix

Include additional visualizations, charts, tables, and details about hyperparameter tuning if needed.