Baris Yuksel- Personal Project Data Analyst 25

Business Report: Credit Card Fraud Detection

1. Executive Summary

In this project, I have developed machine learning models to detect credit card fraud. The main objective is to identify fraudulent transactions to help businesses mitigate financial losses. I employed Logistic Regression and Random Forest models, addressing the issue of class imbalance with RandomOverSampler. Both models achieved high performance based on evaluation metrics, but there are limitations that need to be considered for real-world applications. I recommend deploying the models for real-time fraud detection with continuous monitoring and retraining to keep up with evolving fraud patterns.

2. Problem Statement

Credit card fraud detection is a critical issue for financial institutions as it helps to prevent unauthorized charges. Fraudulent transactions can lead to financial losses for both customers and businesses. Therefore, detecting such transactions early on is essential for maintaining trust and reducing potential risks.

3. Data Overview

The dataset I used consists of transactions made by European cardholders in September 2013, covering two days of transactions. Out of 284,807 transactions, only 492 are fraudulent, representing 0.172% of all transactions, which presents a significant class imbalance challenge.

- Source of Data: European cardholders' transactions from September 2013.
- Time Period: Two days of transactions.
- Total Transactions: 284,807.
- Features: The dataset contains 30 numerical features, of which 28 are PCA-transformed components (V1, V2, ..., V28), with 'Time' and 'Amount' being the original features.
- Target Variable: The 'Class' feature, where 1 indicates fraud and 0 indicates non-fraud.

Given the class imbalance, I focused on metrics like the Area Under the Precision-Recall Curve (AUPRC) for model evaluation, as accuracy alone would not provide meaningful insights.

4. Methodology

I used two machine learning models for this project: Logistic Regression and Random Forest.

- Model Selection: Logistic Regression was chosen for its simplicity and effectiveness in binary classification, while Random Forest was selected for its ability to handle imbalanced datasets through its ensemble nature.
- Data Preprocessing: To address the confidentiality of the original data, I worked with PCA-transformed features. I handled class imbalance using the RandomOverSampler technique to oversample the minority class (fraudulent transactions).
- Resampling Strategy: The RandomOverSampler was applied to ensure a balanced training set, given the large number of non-fraudulent transactions.
- Model Evaluation: I used cross-validation and evaluated the models with metrics such as precision, recall, F1-score, and the AUPRC.

5. Findings

- Fraudulent Transactions: Fraudulent transactions constituted only 0.172% of the dataset, underscoring the imbalance.
- Transaction Amounts: Fraudulent transactions had a higher average transaction amount compared to non-fraudulent ones.
- Model Performance: Both Logistic Regression and Random Forest models achieved perfect scores in cross-validation, suggesting strong predictive performance.
- However, despite these results, I recommend evaluating the models further in the future on unseen data to ensure their generalization beyond this dataset.

6. Results Interpretation

- Implications of Perfect Scores: While perfect evaluation metrics are promising, they might indicate overfitting, especially given the cross-validation results. Testing the models on entirely new data will be critical to confirm their robustness.
- Logistic Regression vs. Random Forest: The Random Forest model, with its ability to handle class imbalance, may be better suited for this type of problem. However, Logistic Regression also performed exceptionally well.
- Real-World Application: The results indicate that these models can effectively detect fraud, but further validation is necessary before deployment.

7. Limitations

- PCA Transformation: The use of PCA-transformed data limits interpretability, as the original meaning of the features is unknown.
- Overfitting Risk: Despite cross-validation, the models' perfect scores may suggest overfitting. Evaluating the models on more diverse and unseen datasets will help confirm their generalization capability.
- Data Specificity: The models were trained on a specific dataset, and applying them to other datasets may require adjustments and retraining.
- Evolving Fraud Patterns: Fraud patterns constantly change, so the models will need continuous retraining and monitoring to remain effective.

8. Recommendations

- I recommend deploying these models for real-time fraud detection in credit card transaction systems.
- Continuous monitoring and retraining of the models will be necessary as new data becomes available, particularly to address the risk of overfitting.
- A feedback loop should be implemented to adapt to new fraud patterns as they emerge.
- I suggest exploring additional machine learning techniques, such as neural networks, to further enhance detection performance.

- Conducting further analysis on interpretability and feature importance could provide deeper insights into the models' decisions and improve trust in their predictions.

9. Visualizations

I recommend that stakeholders consider implementing the following visualizations to better support the findings and enhance understanding:

- A confusion matrix for both models, which would help visualize the true positive, false positive, and false negative rates.
- A feature importance plot for the Random Forest model to highlight the most significant features contributing to fraud detection.
- A ROC curve or Precision-Recall curve, which would allow for a clearer comparison of model performance under different threshold settings.

10. Future Work

- I plan to test these models on more recent and unseen data to validate their effectiveness.
- Investigating additional machine learning algorithms, such as gradient boosting or neural networks, could further improve detection accuracy.
- I would also suggest examining the economic impact of fraud detection, including how the models can help businesses save costs.

11. Ethical Considerations

While the models have demonstrated high performance, it is important to consider the ethical implications, particularly the consequences of false positives, where legitimate transactions may be flagged as fraud. This could lead to customer frustration or loss of trust. I recommend developing a strategy for minimizing such errors and ensuring that affected customers have an easy way to dispute or clarify any misclassifications.

12. Conclusion

In conclusion, the credit card fraud detection models I developed have shown strong potential for identifying fraudulent transactions. The recommendations for deployment, alongside continuous monitoring, will help businesses enhance their fraud prevention strategies. I recognize the limitations and potential for overfitting, and suggest a cautious approach with ongoing evaluation to ensure these models remain effective and reliable over time.