Find Default

Prediction of Credit Card fraud

Mehul Raj
Data Science Bootcamp, upGrad

Table of Contents

1. Introduction

Credit cards are a popular financial product, facilitating online purchases and payments. However, they are also susceptible to fraud, where unauthorized users exploit someone else's credit card information. This unauthorized use not only leads to financial losses for consumers but also poses significant challenges for credit card companies in identifying and preventing fraudulent transactions.

The primary goal of this project is to build a classification model that predicts whether a given transaction is fraudulent or legitimate. By analyzing a dataset of credit card transactions, we aim to develop a solution that enhances the ability of credit card companies to detect fraud efficiently.

2. Dataset Overview

The dataset used in this project comprises credit card transactions recorded in September 2013 by European cardholders. The dataset contains a total of 284,807 transactions, with 492 fraudulent transactions. This results in a highly imbalanced dataset, where the fraudulent transactions represent only 0.172% of all transactions. Such an imbalance poses a significant challenge in model training and evaluation, necessitating the implementation of specific techniques to handle it effectively.

---

3. Methodology

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand the dataset better. This involved:

- **Descriptive Statistics**: Analyzing summary statistics to identify basic trends and anomalies in the data.
- **Visualizations**: Using graphs and charts to visualize the distribution of transaction types, amounts, and other relevant features**.**

3.2 Data Cleaning

Data cleaning involved several steps**:**

- **Standardization**: Ensured consistent formatting across all data entries.
- **Handling Missing Values**: Addressed any missing data points through imputation or removal.
- **Outlier Detection**: Identified and treated outliers to prevent them from skewing the results.

3.3 Dealing with Imbalanced Data

Given the highly imbalanced nature of the dataset, various techniques were implemented to address this issue, such as:

- **Oversampling**: Increasing the number of instances in the minority class (frauds) using methods like SMOTE (Synthetic Minority Over-sampling Technique).
- **Under sampling**: Reducing the number of instances in the majority class (non-frauds) to balance the classes.

3.4 Feature Engineering

Feature engineering was conducted to improve the model's performance:

- **Creation of New Features**: Derived additional features that may provide more insight into transaction behavior.
- **Transformation of Existing Features**: Modified existing features for better interpretability and modeling.

3.5 Model Selection

Different classification models were considered for this project, including:

- Logistic Regression
- Decision Trees
- Random Forest
- Gradient Boosting
- Support Vector Machines (SVM)

The choice of model was based on their performance on the validation dataset and their ability to handle class imbalance.

3.6 Model Training

The dataset was split into training and testing sets, typically with a ratio of 80/20. The training set was used to fit the model, tuning hyperparameters as necessary to improve performance.

3.7 Model Validation

Model performance was evaluated using various metrics, such as:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**
- **ROC-AUC Score**

This evaluation helped ensure that the model generalizes well to unseen data.

---

4. Results and Discussion

Upon evaluating the model's performance, it was found that the accuracy exceeded 75%, meeting the project's success criteria. Key performance metrics demonstrated that the model was capable of detecting fraudulent transactions effectively, balancing both precision and recall.

Key Findings:

- The Random Forest model outperformed others in terms of accuracy and F1 Score.
- The feature importance analysis indicated that certain features significantly influenced the model's predictions, such as transaction amount and time.

---

5. Future Work

Several avenues for future improvement were identified:

- Exploring Advanced Algorithms: Investigating more sophisticated models like neural networks or ensemble methods could potentially enhance prediction accuracy.
- Real-Time Monitoring: Developing a system for continuous learning, where the model adapts to new data over time.

---

6. Conclusion

In this analysis, we performed a comprehensive exploration and modeling process for credit card fraud detection using an imbalanced dataset. The steps included:

Data Exploration and Preprocessing:

We began by loading and exploring the dataset, visualizing class imbalance, and performing necessary preprocessing steps.

Correlation Analysis:

By investigating the correlations between features, we gained insights into how features interact with each other. This helped in identifying potential redundancies and understanding feature relationships.

Handling Imbalanced Data:

We applied techniques such as undersampling and oversampling to address the class imbalance. This was crucial in ensuring that our models could better learn from the minority class.

Model Training and Evaluation:

We trained and evaluated various classifiers, including Logistic Regression, Decision Tree, and RandomForest. We assessed model performance using metrics such as confusion matrices, classification reports, ROC-AUC scores, and ROC curves.

Model Saving:

The trained models were saved for future use, ensuring that we can easily load and apply them for predictions on new data**.**

Key Findings:

- Feature Correlations: Our correlation analysis revealed important relationships between features. This understanding can guide feature selection and engineering in future analyses.
- Model Performance: The RandomForestClassifier demonstrated high accuracy in detecting fraud, showing that it is a strong candidate for deployment. The ROC-AUC scores and ROC curves provided insights into each model's performance, particularly in distinguishing between fraudulent and non-fraudulent transactions.
- Impact of Imbalance Handling: Techniques for balancing the dataset were essential in improving model performance and ensuring that the minority class (fraudulent transactions) was adequately represented in the training process.

Overall, this analysis has provided a robust framework for credit card fraud detection. The insights gained from feature correlations and model evaluations will be instrumental in refining our approach and improving detection capabilities. Future work could involve fine-tuning models further, experimenting with additional features, and exploring other advanced techniques for handling imbalanced data.