Big Data Security Problem

Find a machine learning method to avoid the fraud transactions and reduce financial loss

Dataset: https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset

CONTENTS

Introduction and Background Analysis

Background Information

Big data Security

Big data security aims to ensure the confidentiality, integrity, and availability of large-scale data. The urgent task for enterprises is to develop security management measures to address security issues.

One Application

Credit Card Fraud Detection

The increasing use of credit card bring much convenience to people, while also resulted in increased credit card fraud, resulting in financial loss.

Problem

Find an optimal machine learning algorithm to avoid credit card fraud, reducing the financial losses

Aims and Objectives

Aim

Select an optimal machine learning model comparing **logistic regression, decision tree and random forest** for fraud detection to improve the security of credit card transactions.

Objectives

1. EDA: Conduct an in-depth exploration of dataset to unveil data distributions, correlations, and potential patterns.
2. Model Selection: Evaluate performance logistic regression, decision trees and random forests to find the best model.
3. Best Model Optimization: Choose the best model and further optimize it by analyzing wrong classification and employing down sampling.

Dataset

A simulated credit card transaction dataset that includes information of 1000 customers from January 1, 2019 to December 31, 2020, as well as their transactions with 800 merchants is used.

Exploratory Data Analysis (EDA)

- Only parts of the EDA plots with some special or usual pattern/trends are shown here, more plots could be discovered in report.

Distribution of predictive variable

- This is a binary classification task, and the labels are highly imbalanced. There are 492,494 non-fraud samples (98.5%) and 7,506 fraud samples (1.5%).

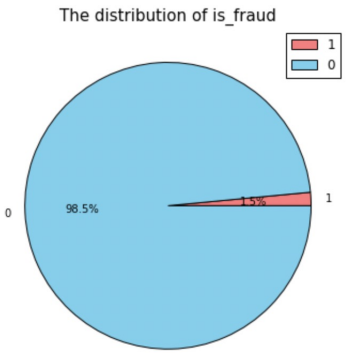


Figure 1. Pie chart of fraud distribution

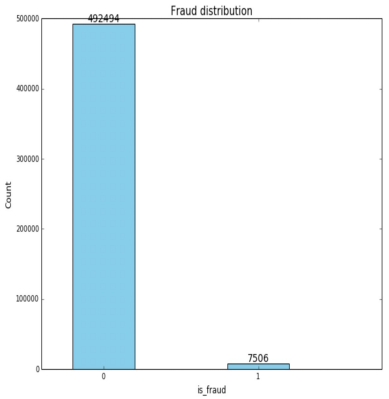


Figure 2. Bar chart of fraud distribution

Feature distribution by predictive variable classification (0 vs. 1)

- The selected feature here is category, which represents in which scenario the transaction happened.
- There are significant differences in the proportion of fraud among different scenarios [Figure 3]. Therefore, a bar chart with fraud rates are shown in Figure 4, where

$$Fraud\ Rate = percentage\ of\ non\_fraud\ - percentage\ of\ fraud$$

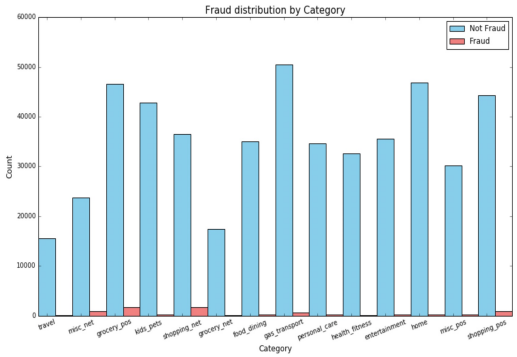


Figure 3. Category distribution by count

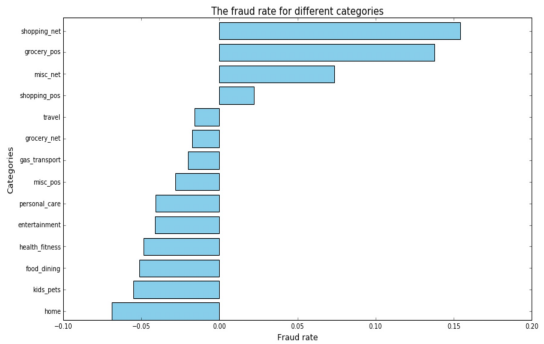


Figure 4. Category distribution by fraud rate

Exploratory Data Analysis (EDA)

- Only parts of the EDA plots with some special or usual pattern/trends are shown here, more plots could be discovered in report.

Time factor distribution related to fraud

- Figure 5 shows the proportion of transaction volume in each time of the day, divided by hours. Fraudulent transactions are exceptionally active between 22:00 and 3:00, far higher than the total number of fraud transactions in other periods.

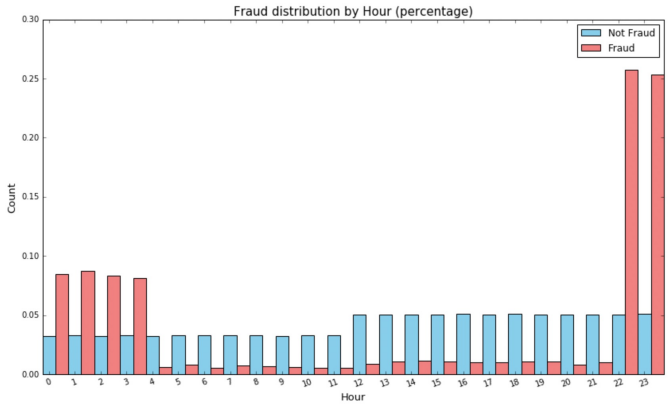


Figure 5. Hour distribution

Correlation heatmap of numeric variables

- Correlation heatmap is shown in Figure 6. It can be concluded that there are several variables with extremely high correlations, such as lat and merge_lat, the correlation of which is as high as 0.9936, long and merge_long reached 0.9991, while zip and long, zip and merge_long is also reached -0.9094 and -0.9086 respectively.

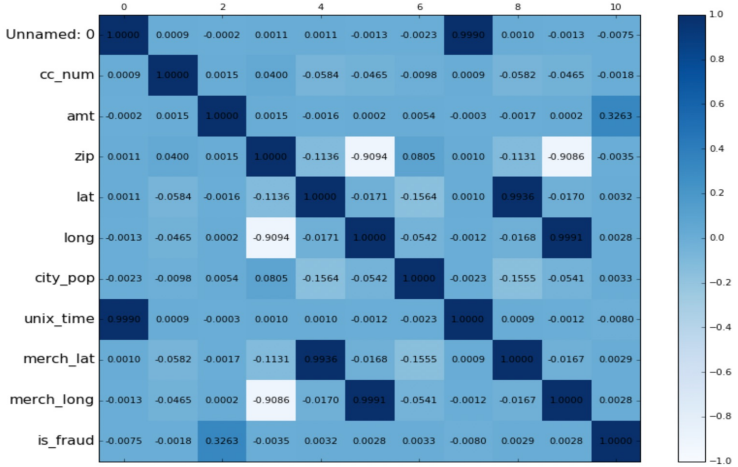


Figure 6. Correlation of numeric variables

Methodology and Experiments

01 ***Preprocessing***

- **Feature engineering:** Dividing some feature into new features such as hour, month, weekday, etc. Combine 2 features to a new feature like age
- **Data Deletion:** In response to the observations of EDA, some features that are weekly correlated with predictive variables were removed.

02 ***Modeling***

Methods of **logistic regression, decision tree and random forest** are chosen because they are all simple, have high interpretability and accuracy, which is important for fraud detection experts in explanation.

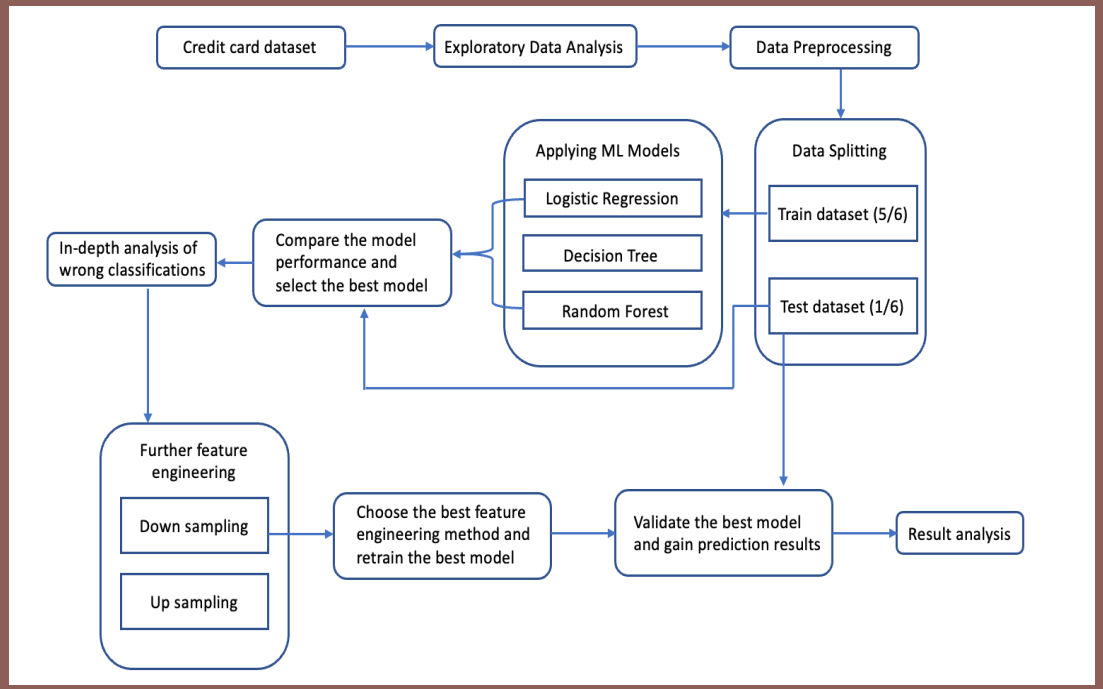
03 ***Evaluation criteria and model selection***

For classification tasks, we generally plot the **confusion matrix**, **accuracy, recall, false positive rate (FPR) and false negative rate (FNR)** are also calculated, which is important for imbalanced classification tasks.

04 ***In-depth analysis and further optimization on selected best model***

Further analyze the abnormal and correct predictions based on the importance of features. Adopt **down sampling** in optimization scheme to balance the distribution

Figure 7. The flow chart of the experiments



Result – Choose the best model

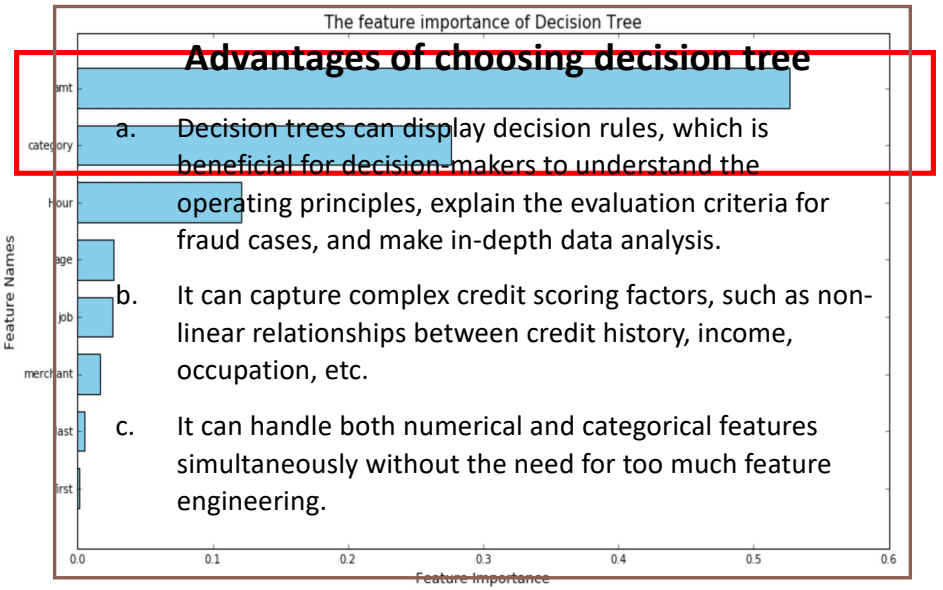
- Accuracy: High, but **not quite meaningful** due to unbalanced dataset.
- FNR: Should pay more attention to, reach **a low FNR** is needed and could help financial institutes reduce financial losses.
- FPR: **Relatively low** is acceptable.

Evaluation indicators (%)	Different Machine Learning Models		
	Logistic Regression	Decision Tree	Random Forest
Accuracy	99.49	99.54	99.68
Recall	0.25	46.10	28.72
False Positive Rate (FPR)	0.13	0.24	0.04
False Negative Rate (FNR)	99.75	53.90	71.28

Table 1. Evaluation indicator results for different models

Decision Tree is Chosen

- Feature importance is plotted.
- **Further analysis of wrong classification** to reduce the important FNR based on important features.



Advantages of choosing decision tree

a. Decision trees can display decision rules, which is beneficial for decision-makers to understand the operating principles, explain the evaluation criteria for fraud cases, and make in-depth data analysis.

b. It can capture complex credit scoring factors, such as non-linear relationships between credit history, income, occupation, etc.

c. It can handle both numerical and categorical features simultaneously without the need for too much feature engineering.

Figure 8. Feature importance of decision tree

Result – Wrong classification analysis

- Important for those should be fraud but predicted as non-fraud, to discover why misclassification happens.

Feature importance of top1 -- atm

Atm has the highest feature importance, which mostly occurs when the value is between 0-50 or 200-400, where the number of correctly classified samples within this interval is very small. Correctly classified samples have atm mostly between 650 and 1200.

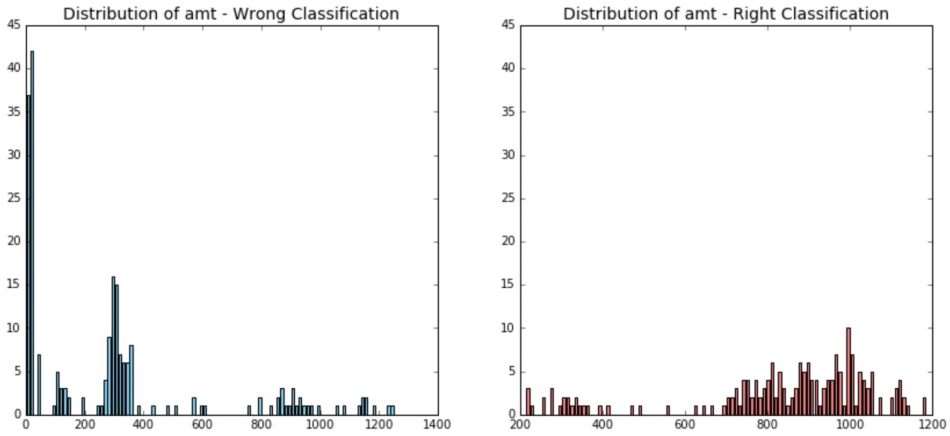


Figure 9. Distribution of atm

Feature importance of top2 -- Category

The model has significantly higher misclassification rates in some categories, especially in the "grocery_pos". Among the samples that were correctly classified, 7 categories even have no correctly classified samples. This suggests that the model's failure to learn the distinctive features of these categories led to incorrect classifications.

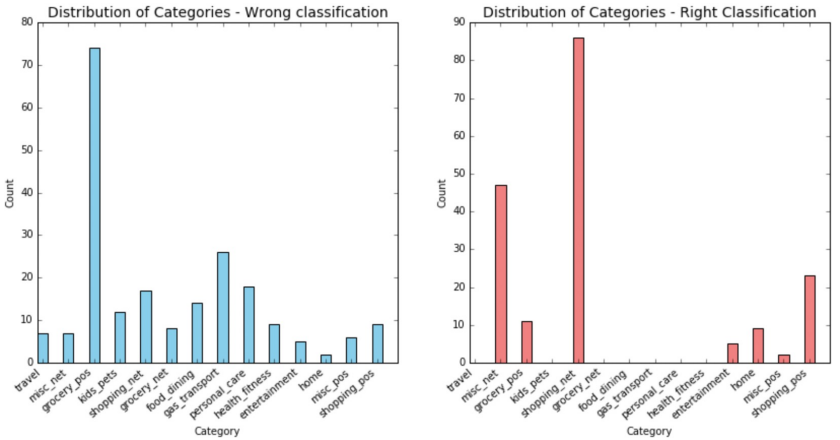
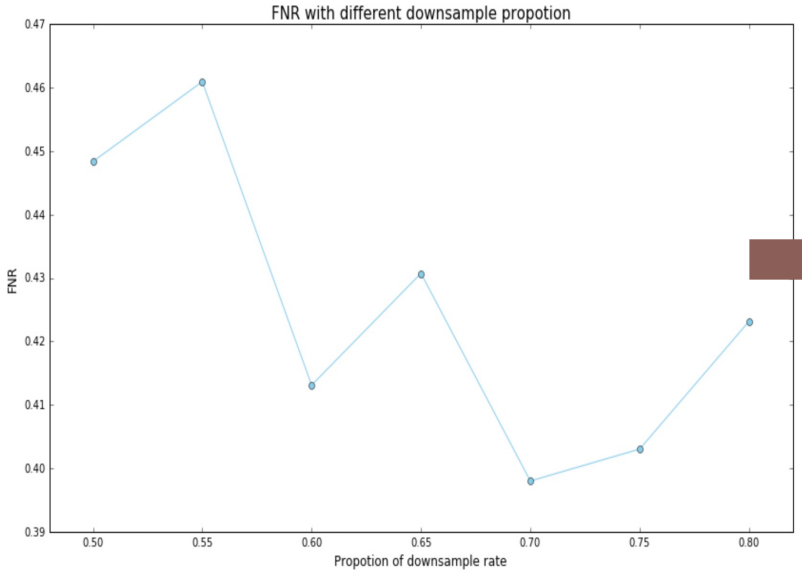


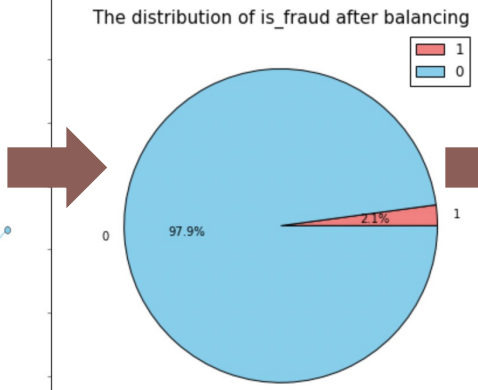
Figure 10. Distribution of Categories

Result – Best model optimization (Reduce FNR)

Down sampling with different ratios to select the best model with lowest FNR, and down sample ratio of 0.70 is chosen.

FNR indeed dramatically decreased. The aim of reducing FNR successfully reached.





Evaluation indicators (%)	Decision Tree Model	
	Before sampling	After sampling
Accuracy	99.54	99.13
Recall	46.10	60.20
False Positive Rate (FPR)	0.24	0.71
False Negative Rate (FNR)	53.90	39.80

Thank you

This study could help financial institutes potential credit card fraud transactions through some machine learning methods, thus reduce potential losses.