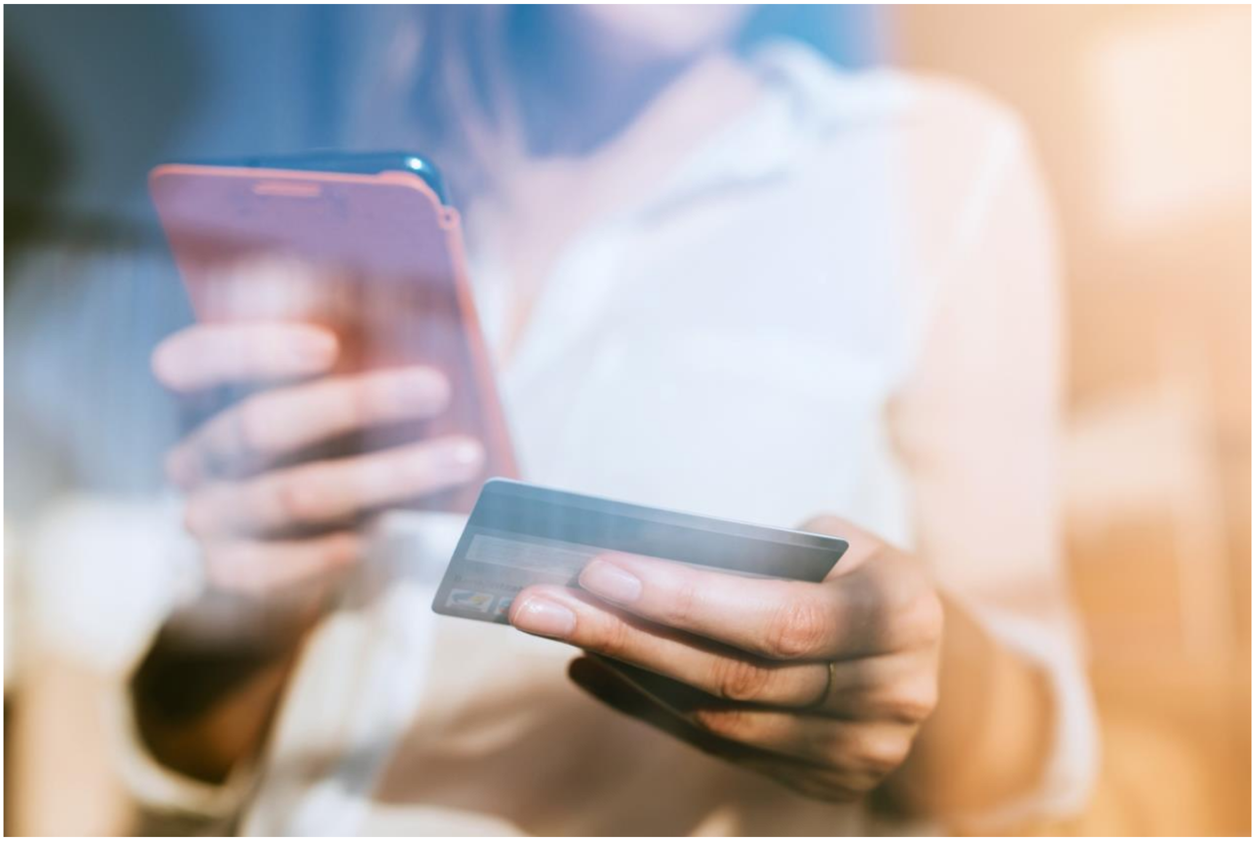
Predictive classification model for identifying fraudulent credit card transactions



Project Proposal by: Himani Kaushik

Abstract

For many banks, retaining high profitable customers is the number one business goal. Banking fraud, however, poses a significant threat to this goal for different banks. In terms of substantial financial losses, trust and credibility, this is a concerning issue to both banks and customers alike. CreditCards, the client wants to proactively monitor and implement fraud prevention mechanisms. It plans to build machine learning models to predict the fraudulent transactions to not only prevent financial losses, but also reduce chargebacks and fees, save time on lengthy manual reviews and decrease denials of legitimate transactions. Four baseline models- Logistic Regression, Decision Tree, random Forest and XGBoost were trained initially and then after the class balancing, Logistic regression model with SMOTE is the best model for its simplicity and less resource requirement.

Design

The purpose of the model is to predict fraudulent credit card transactions using various classification methods. A credit card fraud involves obtaining the credit card information without the proper authorization of the account holder to engage in an illegal financial transaction. This results in financial loss as well as the loss of credibility and trust and affects both the customers and the credit card company. This model can be implemented to detect credit card fraud and prevent the said losses.

Data

The data is obtained from the Kaggle website - https://www.kaggle.com/datasets/mlgulb/creditcardfraud. The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. Out of a total of 2,84,807 transactions, 492 were fraudulent. This data set is highly unbalanced, with the positive class (frauds) accounting for 0.172% of the total transactions. The data set has also been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from 'time' and 'amount', all the other features (V1, V2, V3, up to V28) are the principal components obtained using PCA. The feature 'time' contains the seconds elapsed between the first transaction in the data set and the subsequent transactions. The feature 'amount' is the transaction amount. The feature 'class' represents class labelling, and it takes the value 1 in cases of fraud and 0 in others.

Algorithms

- Exploratory Data Analysis in python
    - Handling missing values and outliers
    - Univariate and bivariate analysis for time and amount.
    - Mitigation of skewness of data
    - Feature scaling of amount.
- Split the data into train and test set

- Train the model with four models – Logistic Regression, Decision Tree, random Forest and XGBoost
- Tune the hyperparameters with GridSearch cross validation and find optimal values of hyperparameters
- Feature importance of the best baseline model
- Handle class imbalance using undersampling, oversampling and SMOTE
- Evaluate all the models using classification metrics to find the best performing model.

Tools

- Pandas and Numpy: For EDA and cleaning data.
- Scikit-learn: For implementing various classification models and performing cross validation, regularization, hyperparameter tuning and feature engineering.
- Matplotlib and Seaborn: For visualizing the data.

Communication

Logistic regression model with SMOTE is the best model for its simplicity and less resource requirement.

Model Metrics:

Train set

- Accuracy:- 0.948
- Precision:- 0.974
- Recall:- 0.922
- ROC :- 0.99

Test set

- Accuracy:- 0.974
- Precision:- 0.056
- Recall:- 0.896
- ROC :- 0.97