Credit Card Fraud Detection Modeling Project

Zaki Alawami, October 21, 2021

Abstract

Credit card providers, Banks, and Merchants can save billions of dollars lost annually to credit card fraud if they can prevent credit card fraud. Also, and more importantly credit card holders will have a higher level of satisfaction, trust, and loyalty to their credit card providers if the latter can efficiently detect and arrest fraudulent transactions before processing the charges, and without annoying their customers with blocked legitimate payments. I will analyze a credit card transactions dataset, derive and highlight insights, and build a predictive classification model to predict if a transaction fraud or legitimate to take appropriate action.

Design:

Can we effectively predict credit card fraud transactions and successfully stop them? I will attempt to answer this question by using a credit card fraud dataset from Kaggle and building a predictive model that can prevent fraud with a good level of performance. Credit card fraud has been on a sharp rise in Saudi Arabia and this is part of the motivation to tackle this problem.

Data:

The Kaggle dataset has 1,842,743 Million records with only 9651 fraud records, which is highly imbalanced as the ratio of fraud records are only 0.52%. The dataset has 21 numerical and categorical features, and the target is a binary class of being Fraud or not. From EDA, the highest correlated features are: the amount of the transaction (amt), the category of the acquired goods (category), and the hour of day when the fraud transaction occur. The latter is included as part of date, so it will have to be created as a new feature. Dataset features:

Algorithms

Feature Engineering

1. Converted categorical features ('category') to one-hot encoding
2. Create new feature ('hour') from the Transaction date as EDA has shown a possible correlation between fraud transactions and time of day.
3. Selected subsets of the total features that included most numerical values (removed ones that caused overfitting such as 'cc_num'), 'category' feature, and the newly created 'hour' feature.

Models

Logistic regression, k-nearest neighbors, and random forest classifiers were used before settling on xgboost as the model with the best performance. PyCaret was used to identify the best model.

Model Evaluation and Selection

The entire training dataset of 1,842,743 records was split into 80/20 train vs. test (holdout). I used PyCaret 10-fold cross validation and was able to replicate the performance results using Scikit-learn directly.

Since this is an unbalanced dataset, I am mainly using Precision and Recall as the performance metrics. I also created a new metric, which I will call **LTPFR** (Legitimate Transactions Predicted as Fraud Rate). LTPFR is important IMHO, because it is important that the Bank (or credit card provider) detects most of the fraud transactions (Precision), without falsely flagging legitimate transactions as fraud, which will cause very low satisfaction and possibly churn for the customers. LTPFR = (fp / (fp +tn))

Tools

- Jupyter, NumPy, Pandas for data processing
- Matplotlib, Plotly, Seaborn for visualization
- Scikit-learn, PyCaret for modeling

Communication: See Notebook for additional charts