Credit Card Fraud Detection - Capstone

Presented by:

Anand Pratap
Sukanya Sarkar
Vijaya Lakshmi Badam

Agenda

- Overview
- Problem Statement
- Objective
- Approach
- Key Insights
- What more we found:
  - Target Problem
  - Model Building
  - Observations
- Cost & Benefit Analysis

Problem Statement

- Modification or falsification of authentic cards
- Making of phony playing cards
- Lost or stolen credit cards
- Dishonest telemarketing

Objective

Identify preventive Action

Getting in place a credit card fraud detection system to save on costs incurred by the bank.

Explore benefits

Reduce time-consuming manual reviews, costly chargebacks and fees, and denial of legitimate transactions.

Approach

An analytics view

A machine learning model has been built to detect frauds early and mitigate losses

Business Impact

A cost benefit analysis has been done for the deployment of the same

Key Insights

# 2,84,807

Credit Card Transactions per month

492

Fraudulent transactions per month

18%

Total costs incurred from fraud transactions

$104

Average loss per fraud transaction

What more we found

- Data plots showed Gaussian distributions, suggesting the effects of transformations on the data.
- 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.

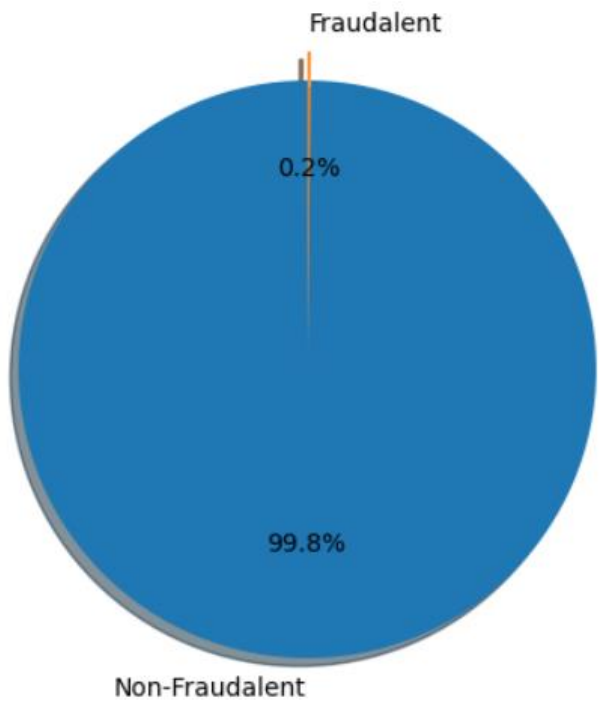
Target Problem

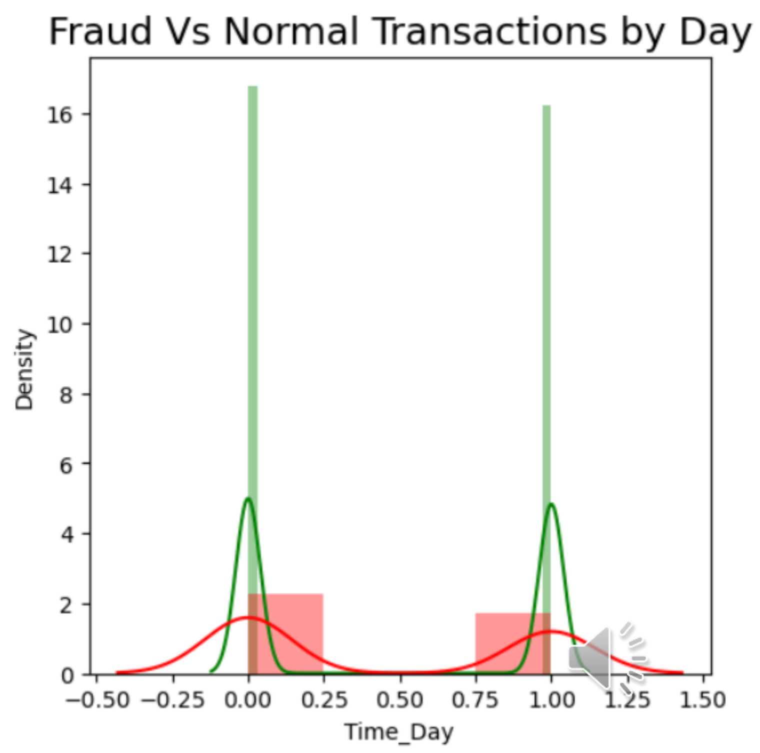
How we analyzed

So we have 492 fraudulent transactions out of 284807 total credit card transactions.

Analysis Approach

Target variable distribution shows that we are dealing with an highly imbalanced problem as there are many more genuine transactions class as compared to the fraudalent transactions. The model would achieve high accuracy as it would mostly predict majority class — transactions which are genuine in our example. To overcome this we used other metrics for model evaluation such as ROC-AUC , precision and recall, Create model functions for Logistic Regress, KNN, SVM, Decision Tree, Random Forest, XGBoost





Model Building

How we analyzed

Build different models on the imbalanced dataset and see the result

Analysis Approach

Performed cross validation with RepeatedKFold

- XGBOost with Repeated KFold cross validation has provided us wih best results with ROC_Value of 0.984352

Performed cross validation with StratifiedKFold

- As the results show Logistic Regression with L2 Regularisation for StratifiedFold cross validation provided best results

Proceed with the model which shows the best result

•Applied the best hyperparameter on the model

•Predicted on the test dataset

Model Building

How we analyzed

looking at the results Logistic Regression with L2 Regularisation with RepeatedKFold Cross Validation has been provided best results without any oversampling.

Analysis Approach

We used Random Oversampling method to handle the class imbalance

1.First we will display class distribution with and without the Random Oversampling.

2.Then We will use the oversampled with StratifiedKFold cross validation method to generate Train And test datasets.

Once we have train and test dataset we will feed the data to below models:

1.Logistic Regression with L2 Regularisation

2.Logistic Regression with L1 Regularisation

3.KNN

4.Decision tree model with Gini criteria

5.Decision tree model with Entropy criteria

6.Random Forest

7.XGBoost

8.We did try SVM (support vector Machine) model , but due to extensive processive power requirement we avoided using the model.

9.Once we get results for above model, we will compare the results and select model which provided best results for the Random oversampling technique

Observations

Overall conclusion after running models on Oversampled data:

We have selected XGBOOST model with Random Oversampling and StratifiedKFold CV

Model Accuracy:

0.9993855444953564

XGboost roc_value:

0.9852138347557161

XGBoost threshold:

0.0050878780893981 46

Cost & Benefit Analysis

Final Savings

$ 186806

$ 213392

Cost incurred per month before the model was deployed (b*c)

$ 8608

Average number of transactions per month detected as fraudulent by the model

$ 1.5M

Cost of providing customer executive support per fraudulent transaction detected by the model

$ 28

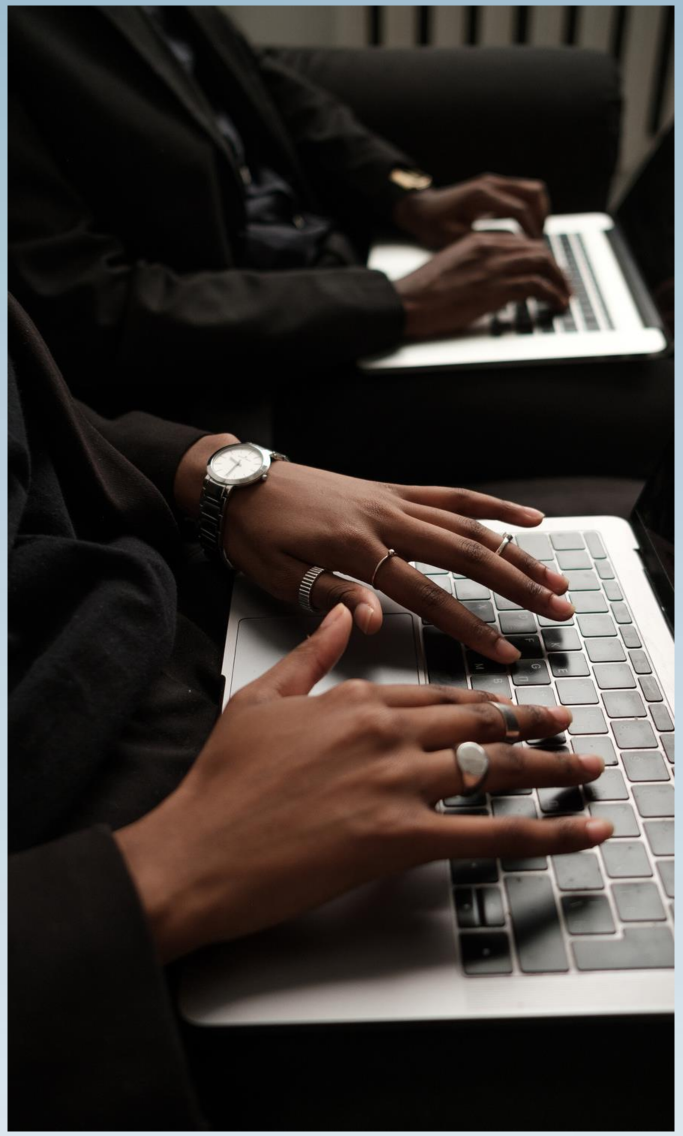
Average number of transactions per month that are fraudulent but not detected by the model

$ 27305

Cost incurred per month

$ 12910

Cost incurred due to fraudulent transactions left undetected by the model



THANK YOU

Reference files attached:

- Cost Benefit Analysis
- Credit_card_fraud detection Model
- Credit Card Raw data - CSV