Credit Card Fraud Prediction

Milestone 3: Preliminary Analysis

Vinay Nagaraj & Vikas Ranjan

DSC630, Spring 2021

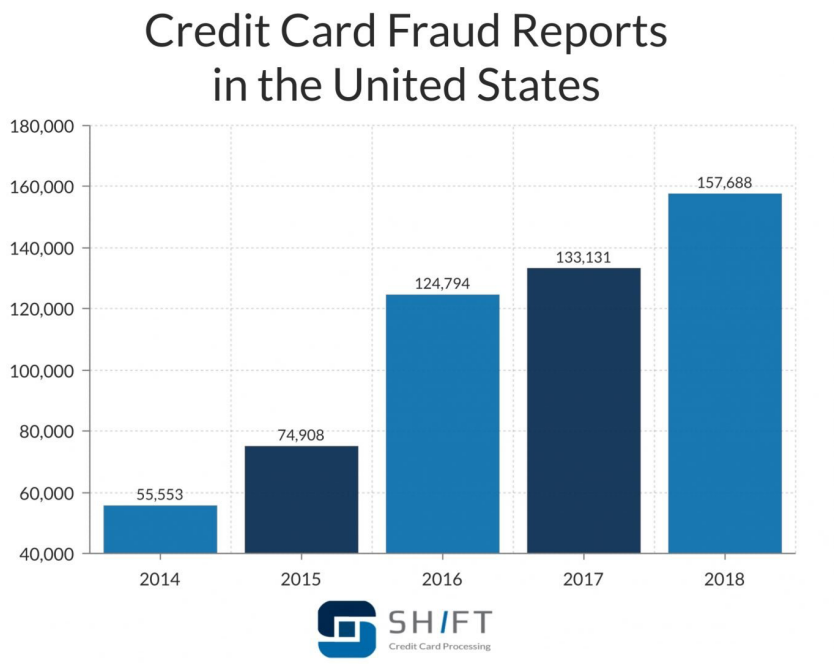
Bellevue University, NE

Abstract

Credit card fraud happens when consumers give their credit card number to unfamiliar individuals, when cards are lost or stolen, when mail is diverted from the intended recipient and taken by criminals, or when employees of a business copy the cards or card numbers of a cardholder. With credit card frauds rampant in the financial sector, we saw it as a good use case for our project. One of the challenges behind fraud detection is that frauds are far less common as compared to legal transactions. After initial struggle to find a good dataset, we came across the dataset on Kaggle which we considered good to be able to build and train our model. As part of this project, we are developing a few models using anonymized credit card transaction data.

Introduction

Credit cards and electronic payments make overall functioning in a global marketplace much easier. Each year financial institutions lost a chunk of money as a result of credit card fraud. In year 2018, a total of $24.26 Billion was lost due to payment card fraud across the globe and United States being the most fraud prone country. Credit card fraud was ranked number one type of identity theft fraud. Credit card fraud increased by 18.4 percent in 2018 and is still climbing. Credit card fraud includes fraudulent transactions on a credit card or debit card. There can be two kinds of card fraud, card-present fraud and card-not-present fraud. Card not present fraud is almost 81 percent more likely than point-of-sale fraud.



It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase. This would not only result in financial loss but also loss of customer confidence in payment industry.

Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling

5. Evaluation
6. Deployment

As part of this paper, first 4 steps (mentioned above) will be covered in this methods section. Evaluation and deployment steps are detailed in the results and conclusion section.

- Business Understanding - Banking sector plays a very important role in our economy and Credit cards are a big-ticket item in this sector. Credit cards are a widely used mode of payment in today's world. Fraudsters are taking advantage of this fact and the number of credit card fraud cases have increased. Therefore, banking industry is coming up with techniques to minimize/eliminate the occurrence of credit card fraud. As part as of this project, we are trying to build a predictive model which determines if a transaction is fraud or not. Test data for such models are difficult to get due to the sensitivity of the data used. After extensive searches, the dataset we have used is one of the clean & reliable dataset which met most of our requirements. The dataset is highly imbalanced, and we will be using methods to overcome that challenge.
- Data Understanding - The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation.

Attribute Information:

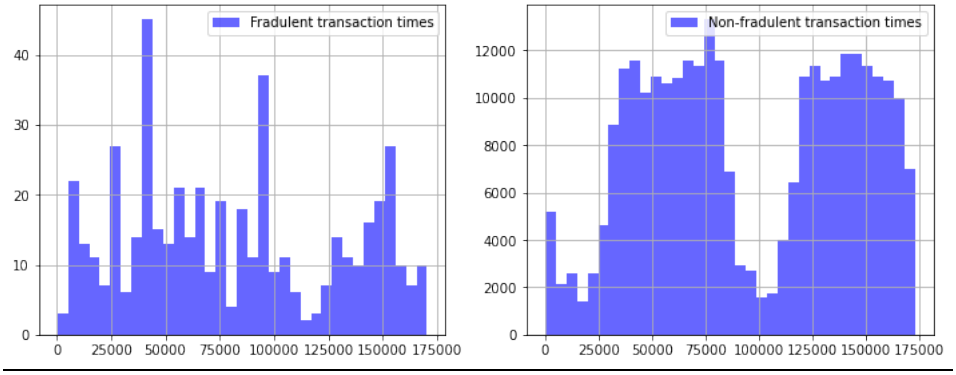
1) Time - Number of seconds elapsed between this transaction and the first transaction in the dataset.
2) V1- V28 – These are the result of a PCA Dimensionality reduction to protect user identities and sensitive features.
3) Amount – Transaction amount
4) Class – This is a response variable and has the values of 1 for fraudulent transactions, and 0 for non-fraudulent transactions.

There are 29 decimal fields and 2 integer fields in the dataset.

As part of this step, we performed graph analysis to see the distribution of the variables.

Fig 1: Below plot shows that the dataset is highly imbalanced as the number of non-fraudulent transactions (284,315) are much higher than fraudulent ones (492).

Fig 2: Below plots shows the distribution of transaction times for fraudulent vs non-fraudulent transactions.



- Data Preparation – The source dataset is clean and contains only numerical input variables which are the result of a PCA transformation. Hence, there wasn't much scope with regards to data cleaning and preparation.
- Modeling – As part of modeling, we did notice that this dataset is severely imbalanced as most of the transactions are non-fraudulent. So, the algorithms are much likely to classify new observations to the majority class and high accuracy won't tell us anything. To address the problem of imbalanced dataset, we choose to use oversampling data approach technique. Oversampling increases the number of minority class members in the training set. In order to make our data set balanced, we are using a type of oversampling called SMOTE (Synthetic Minority Oversampling Technique) and by doing that we are not losing any information from the original training set as all the observations from the minority and majority classes are retained. SMOTE works by utilizing a k-nearest neighbor algorithm to create synthetic data.

  Using the new balanced dataset, we perform feature selection which helps us select the features in our dataset which contributes most to the target variable.

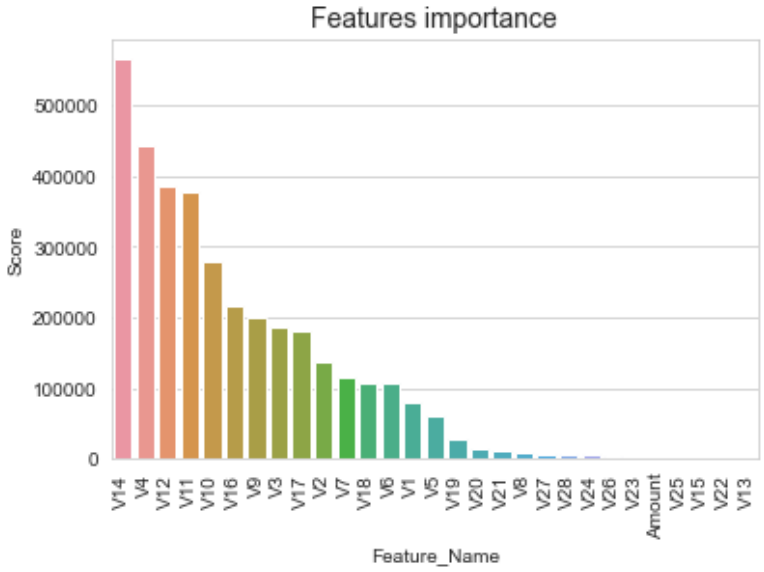
We are using SelectKBest technique to determine 10 best features for the model.

This step helps to improve our model accuracy and reduces training time.

Below is the SelectKBest scores of all the features of our balanced dataset.

	Feature_Name	Score
13	V14	566758.102191
3	V4	444501.743337
11	V12	385289.038158
10	V11	377006.248815
9	V10	278583.089169
15	V16	215311.649804
8	V9	198808.933349
2	V3	185735.366121
16	V17	180622.151619
1	V2	137978.192105
6	V7	115932.706722
17	V18	106806.964661
5	V6	106269.406743
0	V1	78431.271716
4	V5	61168.065828
18	V19	27781.207846
19	V20	14737.694713
20	V21	10680.707515
7	V8	8595.820046
26	V27	6281.318702
27	V28	5123.335450
23	V24	4707.695905
25	V26	3267.452181
22	V23	1228.592812
28	Amount	826.136183
24	V25	738.222047
14	V15	367.094773
21	V22	145.961864
12	V13	122.232423

Below plot shows the features in the order of their importance.



Results

Using the top 10 features determined by SelectKBest technique, we normalized the data and performed model comparison.

	roc_auc
RandomForestClassifier	0.999972
DecisionTreeClassifier	0.997136
SGDClassifier	0.989926
LogisticRegression	0.990047

The roc_auc values for all the 4 models we compared looks good, and in the upcoming weeks we will build the models and evaluate its performance. We will implement ways on how we can improve our model efficiency.

Acknowledgements

We are indebted to the communities behind the multiple open-source software packages on which we depend. We would like to thank our families for their understanding of our time in this endeavor.

References

ULB, M. (2018, March 23). Credit card fraud detection. Retrieved March 27, 2021, from https://www.kaggle.com/mlg-ulb/creditcardfraud

Credit card Fraud Statistics. (2021, January 04). Retrieved March 27, 2021, from https://shiftprocessing.com/credit-card-fraud-statistics/

Siegel, E. (2016). Predictive analytics: The power to predict who will click, buy, lie, or die. Hoboken, NJ: Wiley.