Credit Card Fraud Prediction

DSC630 Course Project: Milestone 5 - Data Selection and Project Proposal

Ganeshkumar Muthusamy, Vasanthakumar Kalaikkovan & Venkata Kanaparthi

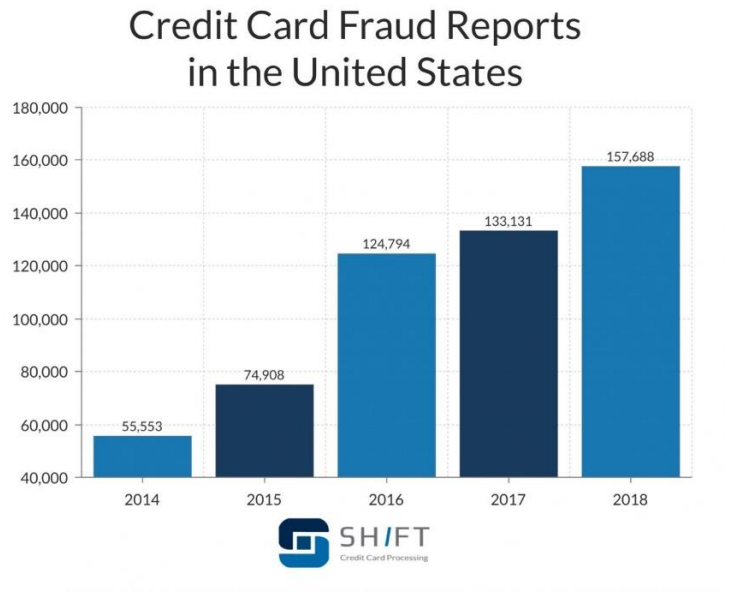
Abstract

According to a study by big organizations, digital payments are expected to reach a record of 726 billion by 2020. Credit card fraud happens in different ways, the new technology on contactless payment on the card allows anyone to read the card details with a contactless card reader. Also, when consumers give their credit card details to unfamiliar individuals when a card is lost or stolen. Many techniques have been introduced to detect fraud in credit card transactions. Fraudsters around the world are always looking for new ways to commit fraud. One of the challenges behind fraud detection is that frauds are far less common as compared to legal transactions. With the increasing number of credit card frauds in the financial sector, we are planning to work on this topic for our project. We found the dataset on Kaggle which is being used to build and train our model. As part of this project, we are developing a few models using anonymized credit card transaction data.

Intro/background of the problem

Credit card fraud is a major problem in financial services and costs billions of dollars every year. Credit card fraud continues to increase due to the rise and acceleration of Phone Order / Mail Order / E-Commerce. There has been tremendous use of credit cards for online shopping which led to a high amount of fraud related to credit cards. Financial institutions like Visa, MasterCard, Amex, Discover, and all debit networks have mandated that banks and merchants introduce EMV (Contact & Contactless) card technology to counter the fraud. In the year 2018, a total of $24.26 Billion was lost due to payment card fraud across the globe, and United States is the most fraud-prone country. Credit card fraud was ranked the 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.



Most credit card fraud occurs when an unauthorized person gains access to our information. The following are the common ways fraudsters get our information:

1. Lost or stolen credit cards.
2. Skimming your credit card, such as at a gas station.
3. Hacking our computer.
4. Calling about fake offers.

Methods

The Cross-Industry Standard Process for Data Mining (CRISP-DM). This process model has 6 phases that naturally describe the data science life cycle for this project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Business Understanding:

Any business or personal use, no matter what size, will have a large surface area for credit card theft and fraud. Devices known as skimmers can illegally obtain credit card details. These machines capture information from the credit card's magnetic stripe, which the criminal can then encode into a counterfeited, faked, or doctored card. It might be hard to detect the difference between a regular card reader or ATM and one with a skimmer attached to it. Rather than stealing existing credit card details, a criminal may instead apply for new credit in someone else's name. They do this by using the victim's personal information, such as their full name, date of birth, address, and Social Security Number. They may even steal supporting documentation to substantiate their application.

The following best practices can be used to detect credit card frauds:

1. Review monthly credit card statements in detail to identify any unauthorized transactions.

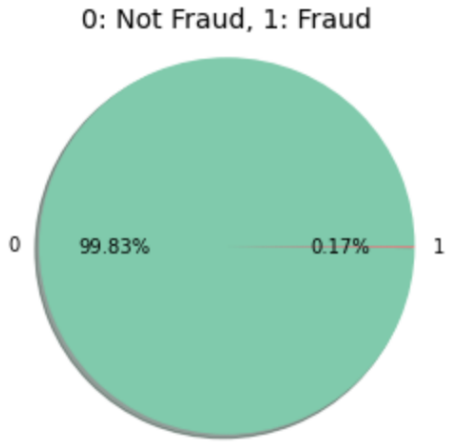
2. Regularly check your credit report to see if anything appears unfamiliar, such as new credit searches and inquiries, the opening of new accounts, or the registration of unknown addresses.
3. Review bills and invoices to ensure you are not receiving correspondence and collection notices for unfamiliar accounts. You can also use your credit report to check if you are on any collection agencies' lists, as most report debts to credit bureaus.

Data Understanding:

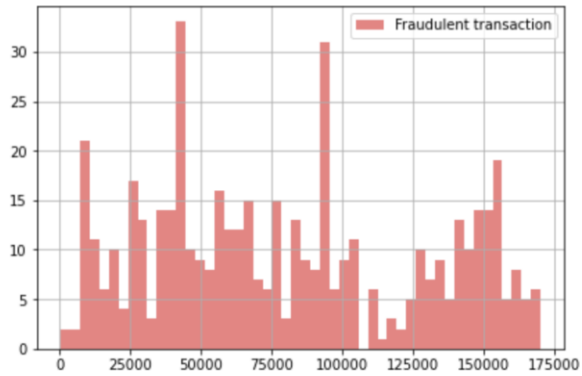
The dataset that is selected has transactions from European cardholders made in 2013. It has 285,000 transactions out of which 492 are fraudulent. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Due to privacy concerns, some principal components are PCA transformed. Time and Amount values are not transformed

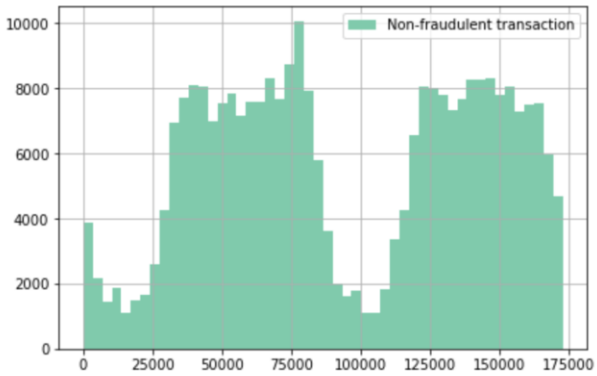
1. Time - Number of seconds elapsed between this transaction and the first transaction.
2. V1- V28 – These are the result of a PCA Dimensionality reduction to protect user identities and sensitive features.
3. Amount – Transaction amount
4. Class – This is a response variable and has the values of 1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

The pie chart shows that the data is highly imbalanced. There is a 0.17% percentage of Fraud transactions in the whole dataset.



The below Histogram shows the Fraudulent vs Non-Fraudulent transaction distribution.





Data Preparation: The dataset contains only numerical input variables which are the result of a PCA transformation and data clean enough.

Modeling: 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. we are choosing this technic because the dataset is highly imbalanced. There are many non-fraudulent transactions compared to fraudulent transactions.

Using the SelectKBest technique to find the best 10 features for the model.

	Feature_Name	Score
13	V14	961.878974
3	V4	768.888297
10	V11	703.241720
11	V12	661.244182
9	V10	534.163705
15	V16	453.489625
2	V3	390.362289
16	V17	384.380278
8	V9	369.572827
1	V2	247.774452

Classification models

With the selected 10 features, we normalized and perform the model comparison. The 4 models shown below have good roc_auc values.

	roc_auc
RandomForestClassifier	0.971959
DecisionTreeClassifier	0.889183
SGDClassifier	0.975939
LogisticRegression	0.978982

Using the training data set, an aggregate measure of performance across the 4 classification models which we created. Based on the accuracy scores, we have finalized to use Logistic Regression and RandomForest model. The hyperparameters are tuned for the model using Gridsearch to evaluate the performance.

Logistic Regression models

Based on a high roc_auc value, the logistic regression model provides us with high true prediction values for our dataset. This model has an AUROC of 0.985 which means that the

model has a very good discriminatory ability. 98.5% of the time, our model will correctly predict if the transaction was fraudulent or not.

```
Confusion Matrix
[[106   3]
 [ 13 124]]

Classification report
              precision    recall  f1-score   support

           0      0.891     0.972     0.930       109
           1      0.976     0.905     0.939       137

    accuracy                          0.935       246
   macro avg      0.934     0.939     0.935       246
weighted avg      0.938     0.935     0.935       246

Scalar Metrics
        AUROC = 0.985
```

106 cases are true negative, meaning they are non-fraudulent transactions and the model predicted them as non-fraudulent. 124 transactions are truly positive, meaning they were predicted as fraudulent and are fraudulent ones. 13 instances are false negative and 3 are false positives.

Random Forest models

Based on a high roc_auc value, Random Forest Model provides us high true prediction values for our dataset. This model has an AUROC of 0.973 which means that the model has a very good discriminatory ability. 97.30% of the time, our model will correctly predict if the transaction was fraudulent or not.

```
Confusion Matrix
[[104   5]
 [ 11 126]]

Classification report
              precision    recall  f1-score   support

           0      0.904     0.954     0.929       109
           1      0.962     0.920     0.940       137

    accuracy                          0.935       246
   macro avg      0.933     0.937     0.934       246
weighted avg      0.936     0.935     0.935       246

Scalar Metrics
        AUROC = 0.973
```

104 cases are true negative, meaning they are non-fraudulent transactions and the model predicted them as non-fraudulent. 126 transactions are truly positive, meaning they were predicted as fraudulent and are fraudulent ones. 11 instances are false negative and 5 are false positives.

Discussion/conclusion

Transaction time does not have any influence on the prediction of fraudulent transactions. This was proven by SelectKBest. SMOTE oversampling technique helped overcome the Imbalanced datasets challenge. Based on a comparison of the confusion matrix of classification models, we recommend that logistic regression classification would be a better performing model to predict whether a transaction is fraudulent or not. The logistic regression model has fewer false positives than random forest.

Acknowledgment

The successful completion of any work would always be incomplete unless we mention the valuable cooperation and assistance of those people who were a source of constant guidance and encouragement. The success and outcome of this assignment required a lot of guidance and assistance from many people and we were extremely fortunate to have got this all along with the completion of our assignment work. Whatever we have done is only due to teamwork. We respect and thank Prof. Fadi Alsaleem for allowing us to do this assignment work and providing us all support and guidance which made us complete the assignment on time. We are extremely grateful to him for providing such nice support and guidance.

References

https://www.kaggle.com/mlg-ulb/creditcardfraud

https://www.self.inc/info/credit-card-fraud-statistics/

https://shiftprocessing.com/credit-card-fraud-statistics/