**Credit Card Fraud Prediction**

**DSC630 Course Project: Milestone 3 - Data Selection and Project Proposal**

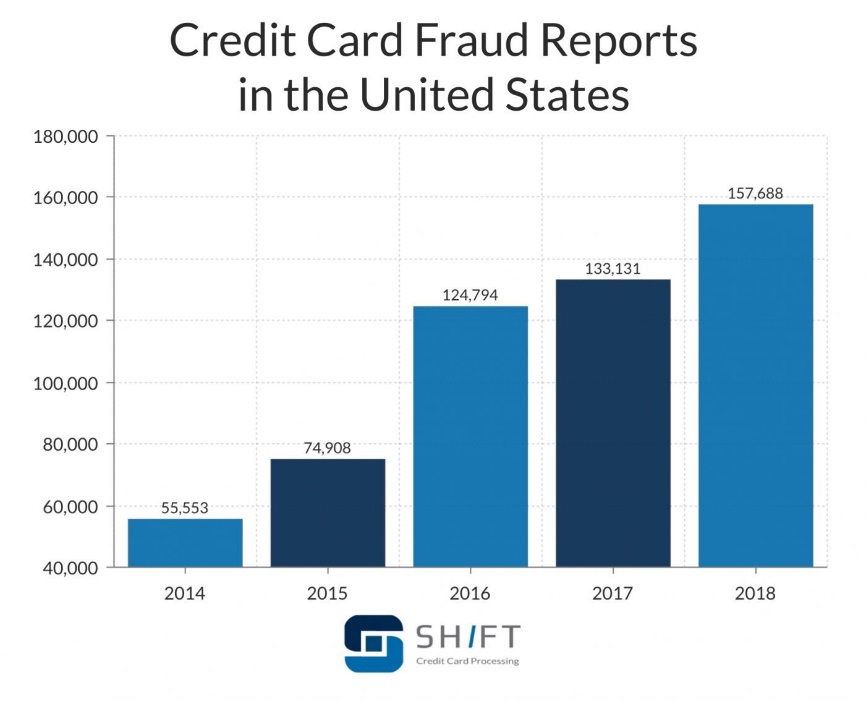
**Ganeshkumar Muthusamy, Vasanthakumar Kalaikkovan & Venkata Kanaparthi**

**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:

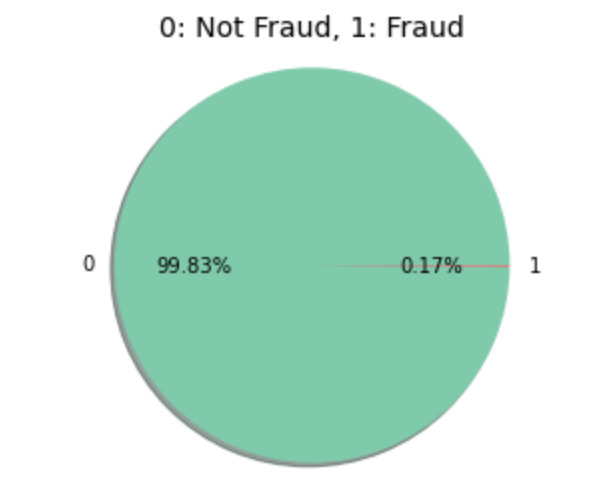
1. Reviewing monthly credit card statements in detail to identify any unauthorized transactions.
2. Regularly checking your credit report to see if anything appears unfamiliar, such as new credit searches and inquiries, the opening of new accounts, or the registration of unknown addresses.
3. Reviewing 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:**

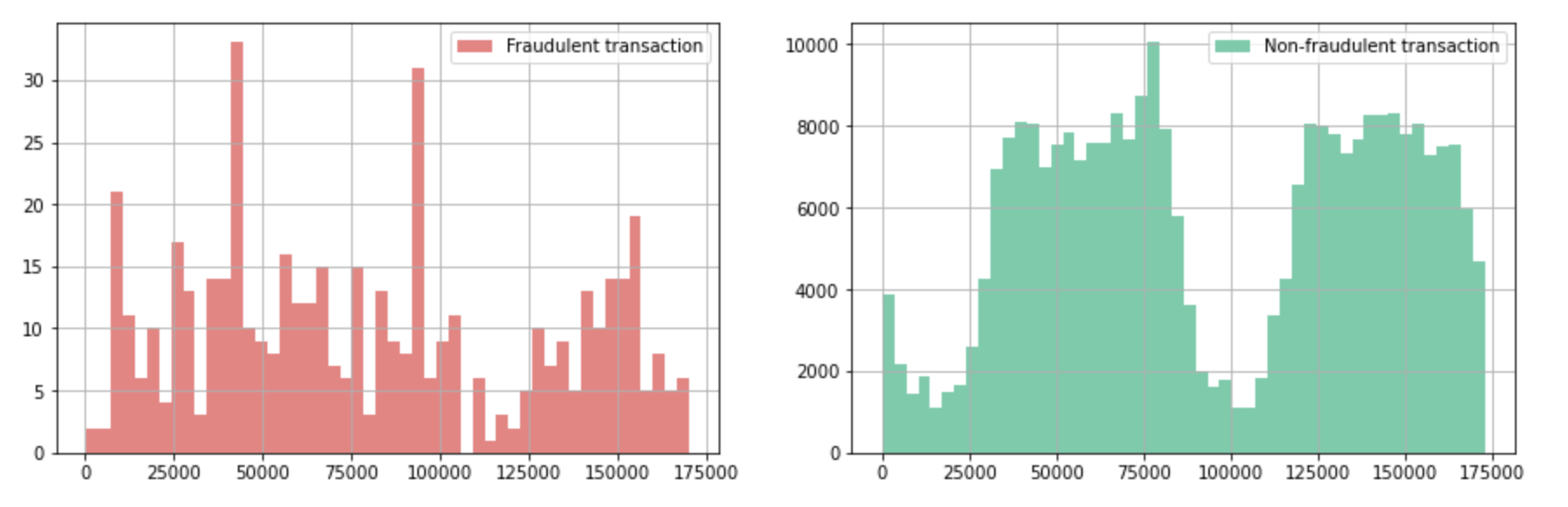
The dataset that is selected has transactions from European cardholders made in 2013. It has 285,000 transactions out of which 492 are fraudulent. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. Due to privacy concerns, some principal components are 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.

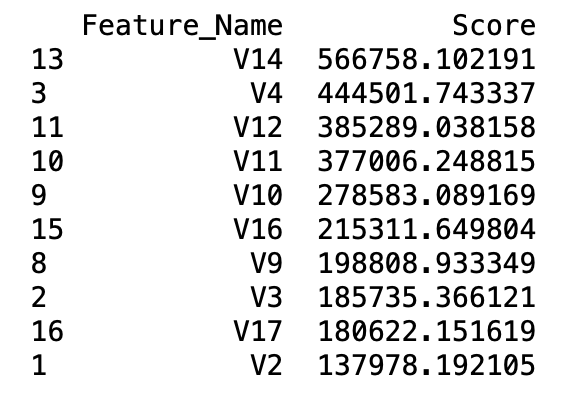


The below Histogram shows the Fraudulent vs Non-Fraudulent transaction distribution.



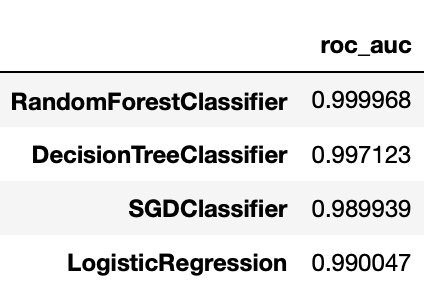
**Data Preparation:** The dataset contains only numerical input variables which are the result of a PCA transformation and data clean enough.

**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.****

**Results**

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

****

**Discussion/conclusion**

Credit card fraud is an act of criminal dishonesty. This article has reviewed recent findings in the credit card field. From an ethical perspective, it can be argued that banks and credit card companies should attempt to detect all fraudulent cases. Yet, the unprofessional fraudster is unlikely to operate on the scale of the professional fraudster and so the costs to the bank of their detection may be uneconomic. The bank would then be faced with an ethical dilemma. Should they try to detect such fraudulent cases or should they act in shareholder interests and avoid uneconomic costs? As the next step in this research program, the focus will be upon the implementation of a models on a real data-set and its evaluation. The main tasks will be to build models to predict fraudulent behavior.

**Acknowledgement**

The successful completion of any work would be 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 final outcome of this assignment required a lot of guidance and assistance from many people and we extremely fortunate to have got this all along the completion of our assignment work. Whatever we have done is only due to team work. We respect and thank Prof. Fadi Alsaleem for giving us an opportunity to do this assignment work and providing us all support and guidance which made us complete the assignment on time. We extremely grateful to him providing such a 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/>