Credit Card Fraud Detection with Machine Learning

Aslı Yılmaz Senih Erdem Çağrı Tapan Yunus Kara

Abstract

One of the fields in finance that machine learning is most commonly used is fraud detection, and credit card fraud detection is a highly prominent area in this field. In this paper, we detected fraudulent credit card transactions using machine learning. We have a sufficiently big data of both fraud and non-fraud transactions that are realized in the time span of two days. In order to prepare the data for training, Principal Component Analysis (PCA) and Random Under-Sampling is used. Then, we employed Logistic Regression classification to create a model and separate the fraudulent transactions from the genuine transactions. Since it is important to not miss any fraud activity, recall is the main measure to evaluate our model. Results indicate that our model is quite successful as we have 0.9444 *recall rate. Additionally, our f1 score is 0.955 which is, again, pretty good.*

1. Introduction

Machine learning has a wide variety of regulatory and supervisory usage in finance. These areas range from money laundering, criminal transactions such as supporting terrorism or other criminal related transactions to risk assessment for investment and detecting fraudulent activities. Credit card fraud detection is one such area that has relatively big importance among other activities. With the increasing share of card payments among financial transaction tools, fraudulent activity and the losses emerging from such activities increases likewise. Especially, using cards in online platforms creates a bigger risk and contributes a great deal to fraud. In 2015, $21.84 billion was fraudulent transaction among all credit card transactions. In US, which accounts for almost two-fifths of the total fraudulent transactions, this amount was $8.45 billion.

With the rising use of data mining, useful and insightful information is discovered from large amounts of data, helping to discover different patterns of card breaching activity. These information are, then, modeled with machine learning in order to predict possible future credit card fraud.

We also aim to create such model to differentiate small and significant amount of fraudulent transactions from large amounts of everyday transactions. Our factors include time of the transaction, amount of the transaction and other confidential variables of credit card owners. As we will classify the instances into two classes (fraud and non-fraud), binary classification techniques will be available for us. Also, we will aim a high recall rate due to the fact that false negatives, that is, labeling fraud as non-fraud, would be much costlier than false positives (labeling non-fraud as fraud).

2. Methods Review and Background

As a beginning, we analyse how our data is distributed. Our data is imbalanced since most of the transactions are non-fraud. The proportion of non-fraud transaction is 99.83% of the time and fraud transaction posses 0.17%. While testing the dataset, oversampled or undersampled datasets might get quite high amount of errors and algorithms will probably overfit. In our case, in order to balance our dataset and understand how features affect the results, correlations are checked.

In our analysis we use logistic regression. Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X. Our output that we test based on 1 or 0 value. Fraud transactions are “fraud=1” and non-fraud transactions are “fraud=0”.

Additionally, in order to evaluate a model performance, several metrics are used. In our model we will check accuracy, precision, recall, and f1 score. In order to calculate these metrics, we need to calculate True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). True positives are the correctly predicted positive values which means that the value of actual class is yes, and the value of predicted class is also yes.True negatives are the correctly predicted negative values which means that the value of actual class is no, and value of predicted class is also no. False positives occurs when actual class is

no and predicted class is yes. Lastly, false negatives occur when actual class is yes but predicted class in no.

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Accuracy = TP+TN/TP+FP+FN+TN. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Precision = TP/TP+FP. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Recall = TP/TP+FN. Last but not least, F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

3. Data Cleaning and Restructuring

The Credit Card Fraud Detection dataset we used is from https://www.kaggle.com/mlg-ulb/creditcardfraud. This dataset contains transactions made by credit cards in two days in September 2013 by European cardholders. There are 284,807 transactions in total and 492 of them are classified as fraud (positive class). Since the positive class account for only 0.172% of all transactions, the dataset can be determined as a highly unbalanced dataset.

The dataset comprises of 30 features, but due to the data confidentiality reasons, features are named as V1, V2, …, V28, except the amount and time. Time feature shows the number of seconds elapsed between a specific transaction and the first transaction in the dataset and amount feature defines the total amount of transaction. Figure 1 shows the distributions to better understand how skewed these features are among all transactions. The time of

transactions is relatively better distributed, however, the amount feature does not show a similar behavior. Most of the transaction amount are relatively small and the mean of them is approximately $88.

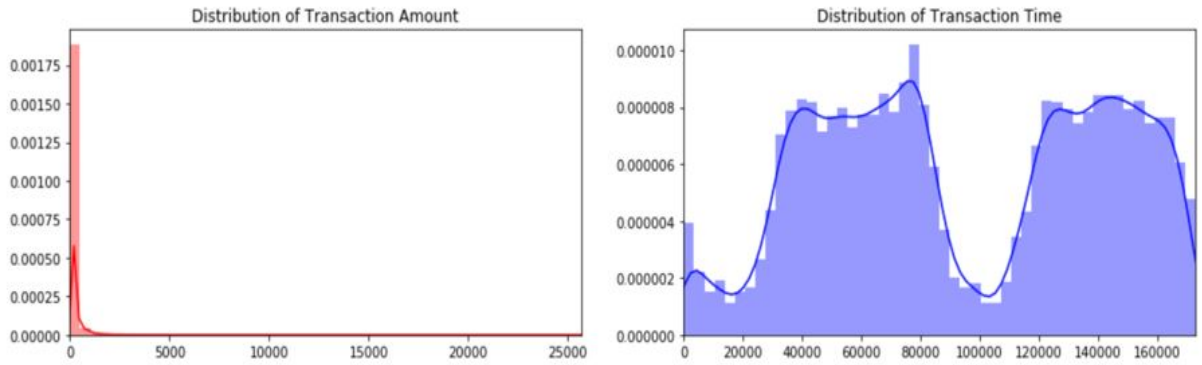

Figure 1

The description of the data says that all the features except time and amount went through a PCA transformation which is a dimensionality reduction technique. Since the other columns have already been scaled, the amount and time columns should be scaled as well before processing the data.

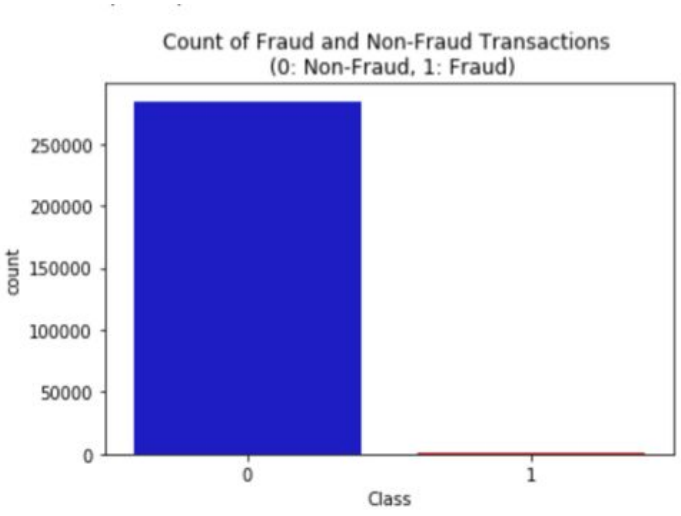

Figure 2

Figure 2 illustrates the histogram of positive (fraud) and negative (non-fraud) classes. It can be clearly seen from the histogram that the number of fraud transactions are almost at a negligible level compared to the non-fraud transactions. If we use this dataset as the base for our predictive models and analysis, we might get a lot of errors and the algorithms we further apply will probably overfit since it will assume that most transactions are not fraud. Besides, it is unknown what "V" features stand for and it will be useful to understand how each of these features influence the result (fraud or non-fraud). By having an unbalanced dataset, it is not possible to see the true correlations between the classes and features. To overcome these overfitting and wrong correlation issues, we use Random Under-Sampling technique to have a dataset with a 50/50 ratio of fraud and non-fraud transactions.

Since the number of fraud transactions is 492, we also randomly choose 492 non-fraud transactions, get them together and finally shuffle them to have our evenly-distributed sub-sample dataset (Figure 3).This dataset will be our clean, base dataset for further processing.

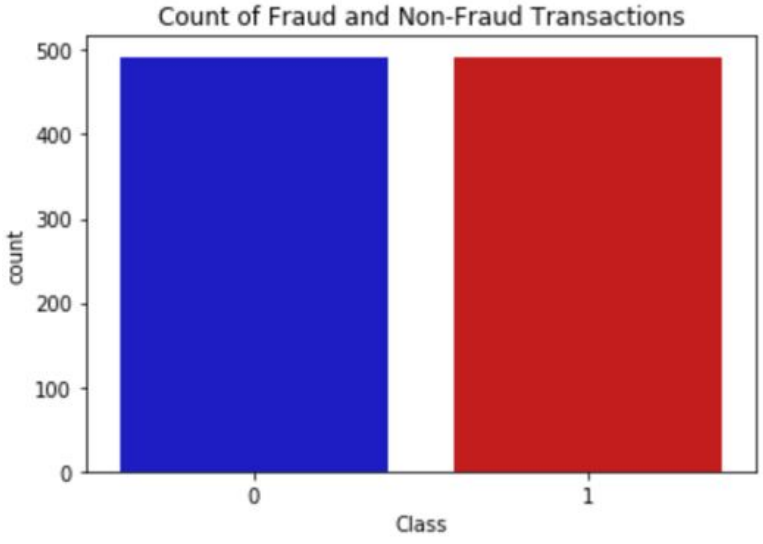

Figure 3

4. Statistical Analysis and Data Exploration

In order to explore our sub-sampled data more deeply, we first construct a correlation matrix between the features (Figure 4). Correlation matrix is one of the most useful techniques to understand data and see which features have more influence on results. In our case, it is important that we use the correct subsample in order for us to see which features have a high positive or negative correlation with regards to fraud transactions. As the second step, we use boxplots for having a better understanding of the distribution of these features in fraudulent and non-fraudulent transactions (Figure 5).

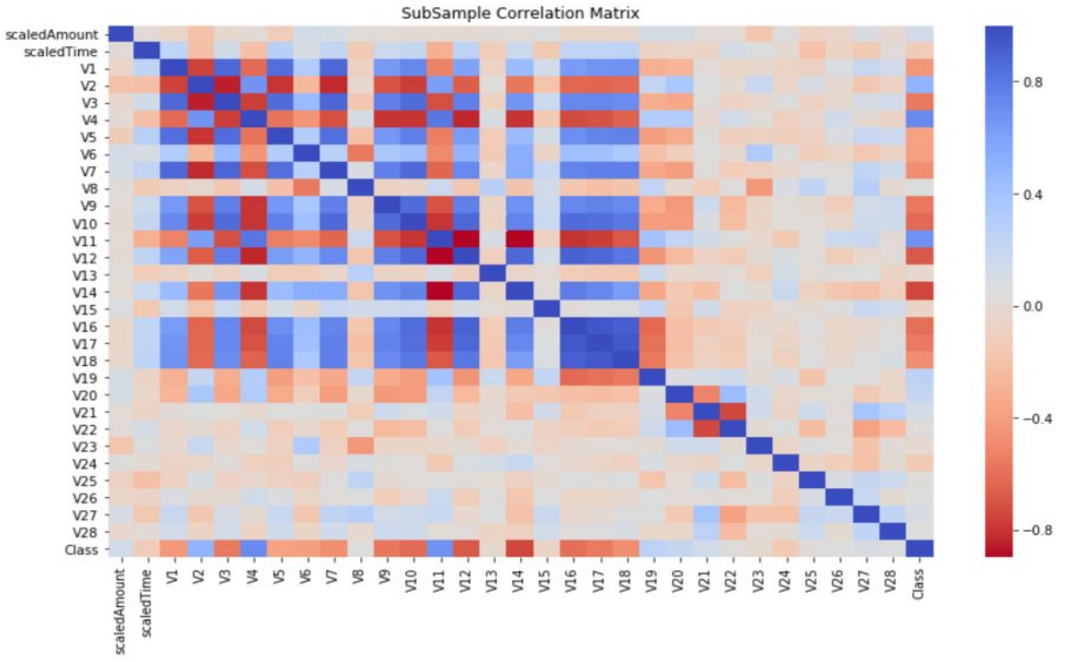

Figure 4

As seen in Figure 4, V17, V14, V12 and V10 are negatively correlated with the Class. The lower the values of these features are, the more likely the end result will be a fraud transaction. Also, V2, V4, V11, and V19 are positively correlated. The higher the values of these features are, the more likely the end result will be a fraud transaction.

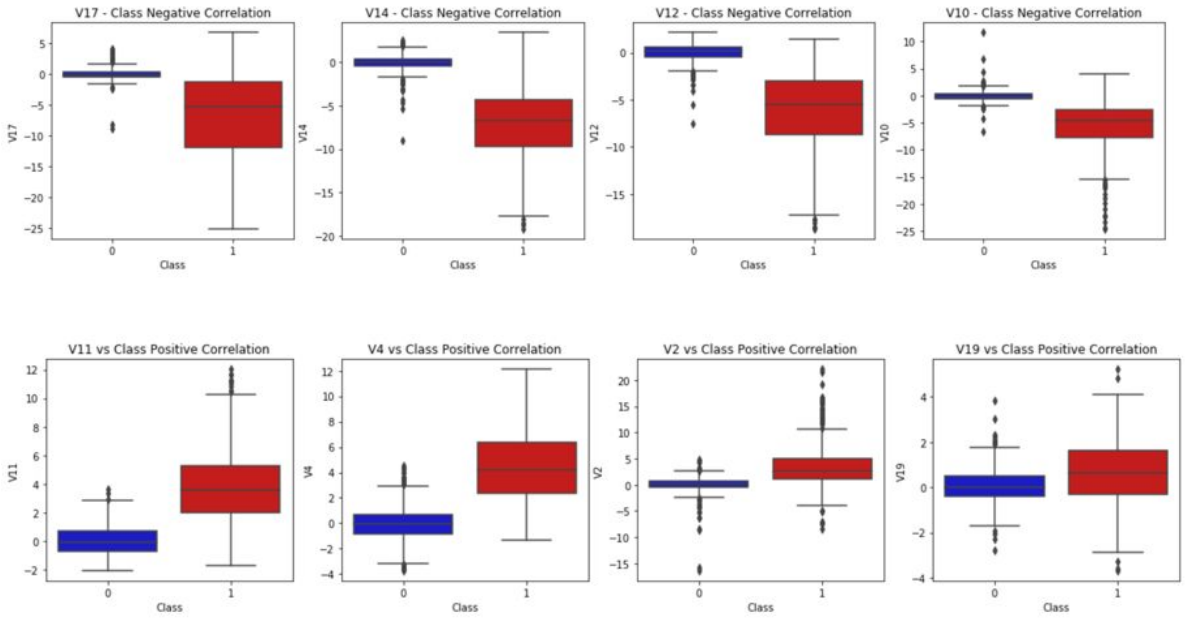

Figure 5

To detect anomalies and have a positive impact on the accuracy of our models, we remove extreme outliers from features that have a high correlation with our classes. First, we apply Interquartile Range (IQR) method which will be calculated by the difference between the 75thpercentile and 25thpercentile so that we can create a threshold beyond the 75thand 25thpercentile and if an instance passes this threshold, it will be deleted.

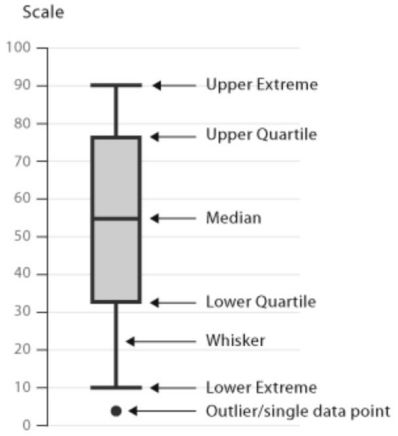

Figure 6: IQR Scale

We use boxplots to see the 25thand 75thpercentiles (both end of the squares in Figure 5) it is also easy to see extreme outliers (points beyond the lower and higher extreme in Figure 5). From the figure, we can see that there are many extreme outliers of V14, V12, and V10. To better understand, we plot the distributions of these features in fraud transactions (Figure 7).

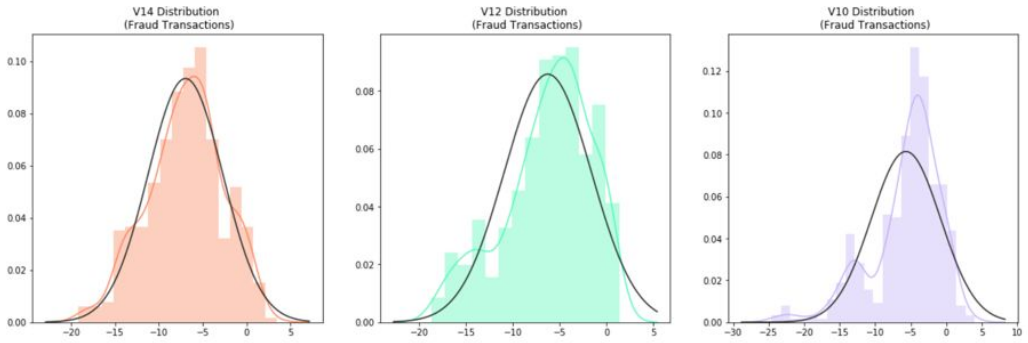

Figure 7

After eliminating the extreme outliers, the boxplots of instances of V14, V12, and V10 become as shown in Figure 8.

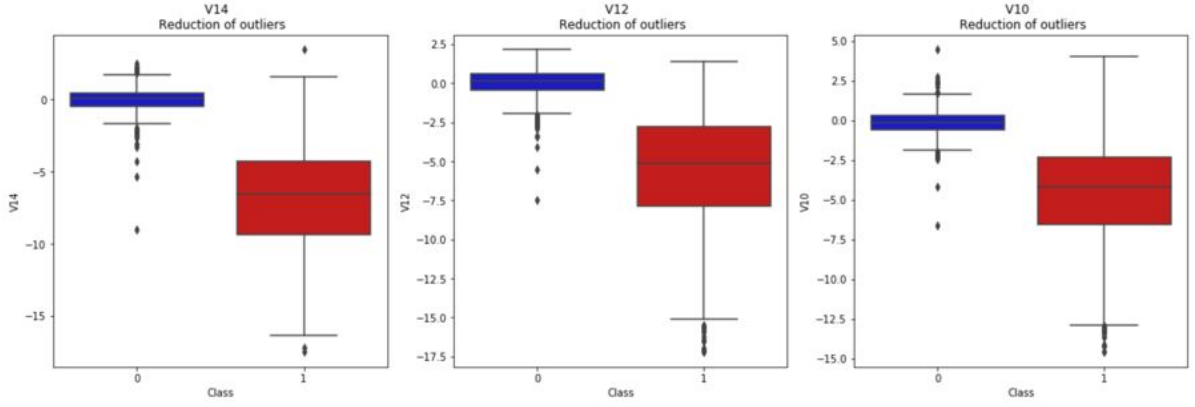

Figure 8

5. Methods and Tools of Use

We use logistic regression classifier because of the fact that, we examine four types of classifiers and logistic regression is the classifier with the highest accuracy on our data frame. Logistic Regression has the best Receiving Operating Characteristic score (ROC), meaning that Logistic Regression quite accurately separates fraud and non-fraud transactions. The wider the gap between the training score and the cross validation score, the more likely our model is overfitting (high variance). If the score is low in both training and cross-validation sets this is an indication that our model is underfitting (high bias). Logistic Regression Classifier shows the best score in both training and cross-validating sets. Our classifiers training and cross validation scores are seen above:

Classifiers: LogisticRegression Has a training score of 93.0 % accuracy score

Classifiers: KNeighborsClassifier Has a training score of 93.0 % accuracy score

Classifiers: SVC Has a training score of 94.0 % accuracy score

Classifiers: DecisionTreeClassifier Has a training score of 90.0 % accuracy score

Logistic Regression Cross Validation Score: 93.92%

Knears Neighbors Cross Validation Score 92.99%

Support Vector Classifier Cross Validation Score 94.05%

DecisionTree Classifier Cross Validation Score 89.55%

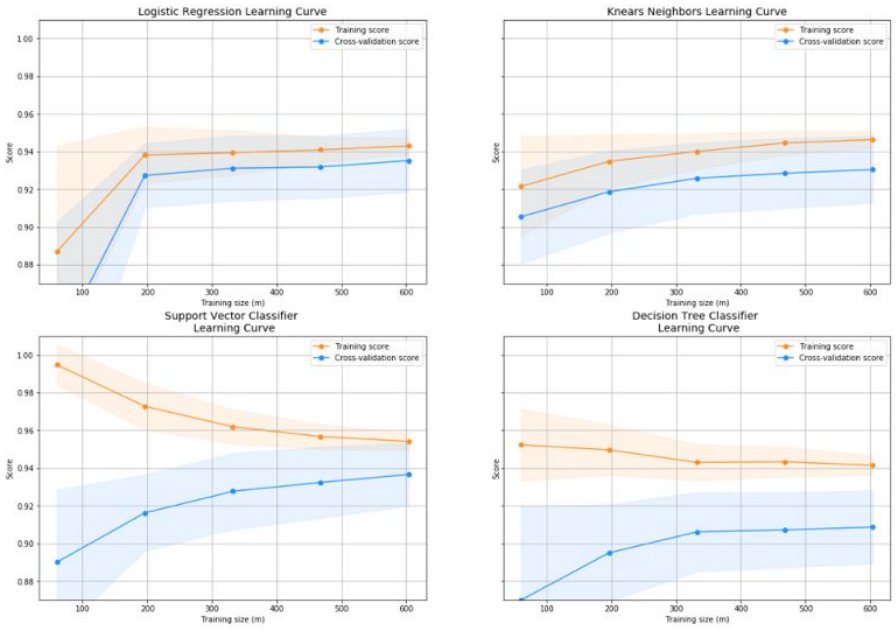

Figure 9

On the figure above, it is seen how our classifiers scored in training and cross- validation.

6. Results

We can see the performance of our logistic regression model, which is trained with random under-sampled and 80-20 split data, with confusion matrix in Figure 10. Our accuracy is pretty high with 0.947 which is significant as our data was balanced with under-sampling. Besides, we care about false negatives and recall rate because although false positives could be neglected as this model would provide possible targets for further interrogation, false negatives would be much costly since fraud would be missed.

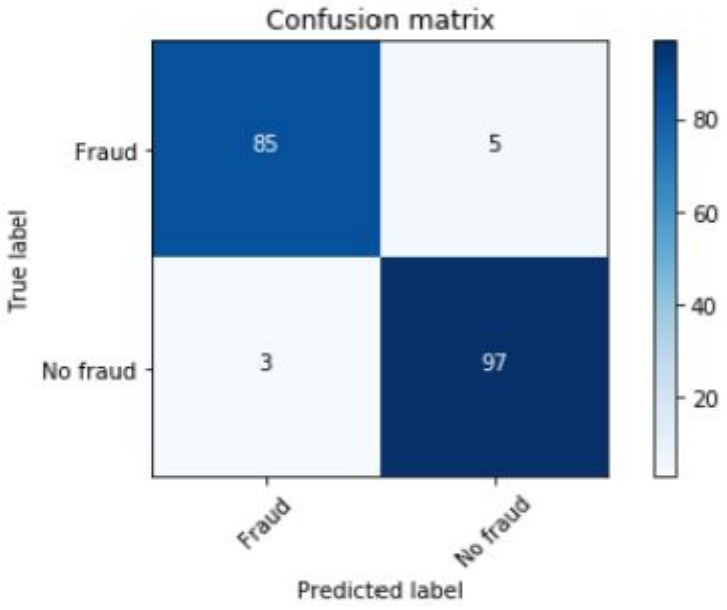

Figure 10

Number of false negatives is not big in our model and our recall rate is sufficiently good with 0.9444. Moreover, our precision is not bad also and our f1 score is 0.955.

These scores could be improved with more data. We had an unbalanced data, so we had to feed our data with relatively small amount of data. But, more data, more informative features that would be discovered with data mining and data analysis and more complex models could boost these results.

References:

https://becominghuman.ai/logistic-regression-in-python-from-scratch-954c0196d258

https://github.com/martinpella/logistic-reg/blob/master/logistic_reg.ipynb

https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets

https://www.forbes.com/sites/rogeraitken/2016/10/26/us-card-fraud-losses-could-exceed-12bn-by-2020/#39f076f6d243