Credit Card Fraud Detection Using Machine Learning

Zheng Li, Yash Shingadiya, Nan Xu

***Abstract*—It is important for credit card companies to identify all the fraudulent transactions to avoid financial losses. Therefore, an effective fraud detection method is important that can identify potential fraud in time. One way to do this is, model past credit card transactions with the knowledge of the ones that turned out to be fraud and, then use it to identify whether a new transaction is fraudulent or not. This paper discusses about the analysis of the dataset that has been taken from Kaggle, data preparation, outlier detection, dimensionality reduction and, then finally the results that are obtained by comparing different classification algorithms like Logistic Regression, Random Forests and Support Vector Machines using different parameters so, as to see which algorithm works better under which circumstances.**

***Keywords*—fraudulent transactions, non-fraudulent transactions, Random under-sampling, Random Forest, Logistic Regression and Support Vector Machine**

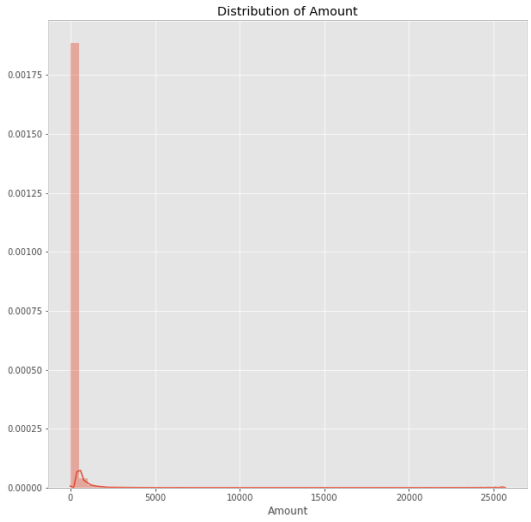
I. Introduction

Every year, billions of dollars are lost due to credit card frauds. It is important that credit card companies are able to recognize fraudulent transactions so that, the customers are not charged for the items that they did not purchase. This paper is about the identification of all such fraudulent transactions while minimizing incorrect fraud classification using machine-learning techniques such as Random Forest, Logistic Regression and Support Vector Machine. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, consisting of 492 frauds out of 284,807 transactions. The dataset contains only numerical input variables which are the result of a PCA transformation. Due to confidentiality issues, the original features and more background information about the data are not provided. Features V1, V2, through V28 are the principal components obtained with PCA, the only features which are not transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. As the dataset is highly unbalanced, first random under-sampling is performed so, as to create a balanced dataset. Then, the two features which are not transformed using PCA are scaled so, as to get better performance for different classification algorithms, especially Logistic Regression. After the data preparation, is done, outliers are removed so that, they do not impact the predictions and, also cannot skew the results. Once, the data is cleaned and, all outliers are removed, the dimension is reduced using T-SNE technique so, as to project higher dimensional data to lower dimensional data because it is not possible to produce a 31 dimensional plot using all the predictors. After all of these data preparation, the dataset is splitted into 80/20 train-test set and, then finally different classification algorithms are used using evaluation techniques like F1-Score and ROC-AUC score to see which algorithm works better and, under which circumstances.

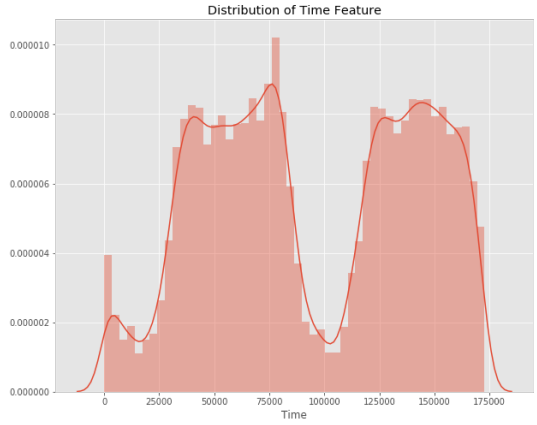
II. Proposed Work

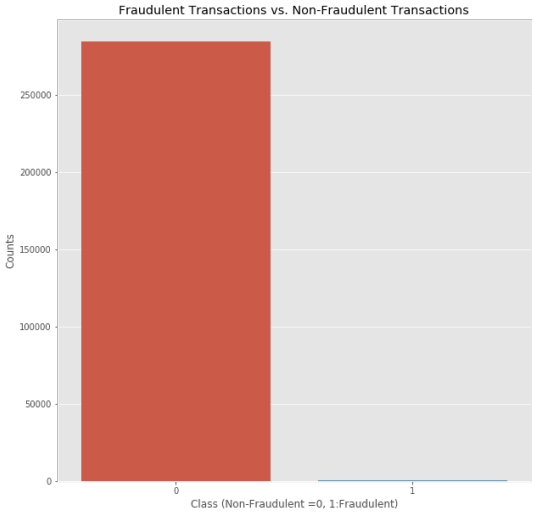
A. Exploratory Data Analysis

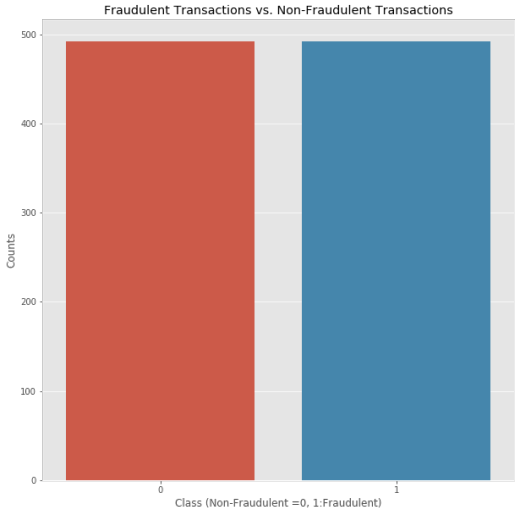
The dataset consists of 284807 rows * 31 columns. The mean value of all the transactions is $88.35 while the largest transaction in this dataset seems to be $25,691.16. The distribution of the amount value of all the transactions is heavily right-skewed. The vast majority of the transactions are relatively small and only a tiny fraction of transactions comes even close to the maximum. As opposed to the distribution of the amount value of the transactions, the distribution of time is bimodal. This indicates that approximately 28 hours after the first transaction there was a significant drop in volume of transactions guessing, that it might be probably because of the night time. As shown in the below visualization, 99.83% of the transactions in this data set are non-fraudulent and only 0.17% are fraudulent which shows that the dataset is highly unbalanced.



Credit Card Fraud Detection Using Machine Learning







B. Data Preparation

Scaling the Data Set:

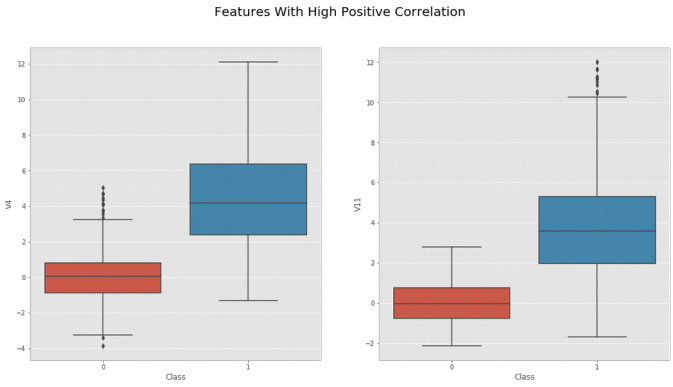
Not scaling the data would result in certain machine-learning algorithms like Logistic Regression to perform worse because it assigns weights to the features. So, the Time and Amount features are scaled first.

Balancing the dataset:

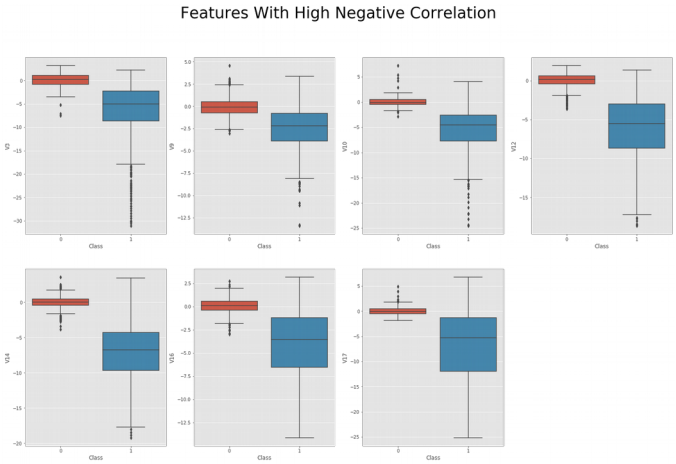
As the dataset has 99.83% non-fraudulent transactions and only 0.17% fraudulent transactions, it is very important to balance out the dataset because if such a highly unbalanced dataset is trained then the model will always predict future transaction as non-fraudulent as it has more number of non-fraudulent transactions. To solve this problem, random under-sampling is used to create a training data set. For this, all the fraudulent transactions are chosen and random samples are taken from non-fraudulent transactions so, as to create 1:1 dataset.

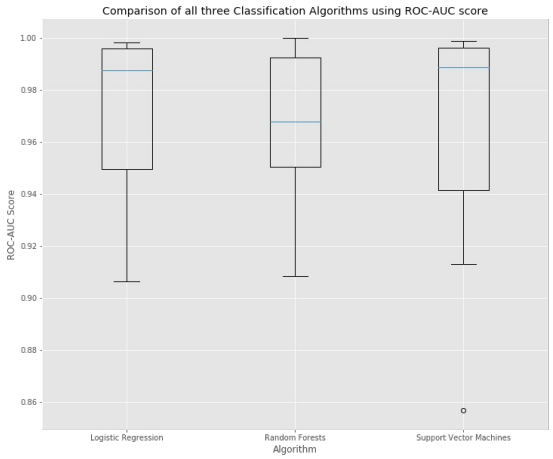
Outlier detection and removal:

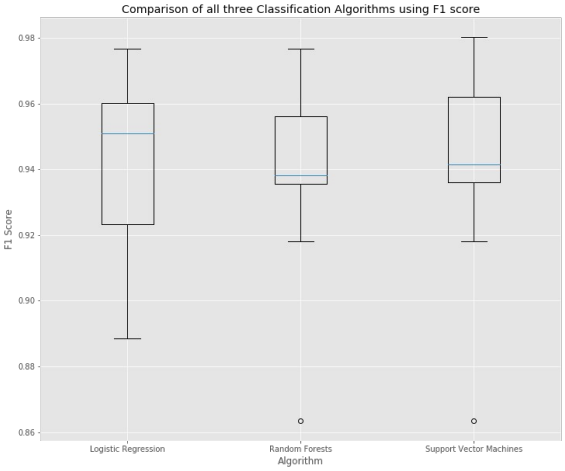
The outliers can harm the predictions so, it is necessary to remove the outliers. For this, main focus is given on the correlation between the features. First, the features that have correlation greater than 0.5 are identified as the features with high positive correlation and, the features that have correlation less than -0.5 are identified as the features with high negative correlation. As an experiment the Inter-Quartile Range (IQR) is varied from 1.5 times the IQR to 2.5 times the IQR and, finally from 2.5 times the IQR to 3.5 times the IQR. From these experiments, it was observed that when the range is 1.5 times the IQR then, instead of removing outliers, the data was getting removed. And, when the range was 3 or more times the IQR then, all of outliers were not getting removed, some of the outliers were still present along with the data. So, finally, the range is set as 2.5 times the IQR and, only the extreme outliers are focused. As shown in the visualization below, all the transactions that lie outside 2.5 times the Inter-Quartile Range (IQR) are removed.



Credit Card Fraud Detection Using Machine Learning







C. Classification Algorithms

To test the performance of the Random Forests, Logistic Regression and Support Vector Machines first, the data is splitted into 80/20 train-test dataset. To avoid overfitting, the resampling technique of k-fold cross validation is done to separate training data into k folds and then the model is fitted on k-1 folds. And, then finally this process is repeated for every single fold so, as to obtain the average of the resulting predictions. This process is repetitive and takes time but, it guarantees convergence. Finally, using F1-Score and ROC-AUC score, all of these algorithms are evaluated and, compared along with parameter tuning which are discussed in detail in the results section.

III. RESULTS

The results of the three classification algorithms and their performance comparisons using different evaluation techniques are shown below in the table.

Evaluation Technoques	Logistic Regression	Random Forests	SVM
F1-score (mean)	0.942	0.937	0.941
ROC-AUC score (mean)	0.969	0.964	0.964

IV. CONCLUSION

This paper demonstrates an overall data preparation process for highly unbalanced dataset, fraud detection using three different classification algorithms and, their performance for the credit card fraud detection problem. From the results, it can be seen that, for F1-score Logistic Regression performs better than the other two. But, for ROC-AUC score both SVM and Logistic Regression gives same performance. Among both the evaluation techniques, Random Forests shows less performance as compared to the other two algorithms. In overall, it can be concluded that for the given dataset, the Logistic Regression shows better performance as compared to the other two algorithms. The performance of the Random Forests can be further improved by better parameter tuning and feature extraction.

REFERENCES

[1] *https://www.kaggle.com/mlg-ulb/creditcardfraud*

[2] *https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8*

[3] *John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis, IEEE, 2017*

[4] *Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with under-sampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015*

[5] *L.J.P. van der Maaten and G.E. Hinton, Visualizing High-Dimensional Data Using t-SNE (2014), Journal of Machine Learning Research*