Streaming Analytics On Credit Card Fraud Detection In Data Analytics

Abhiram Prasad, Nithin Sylesh, Ritika R Prasad, Vyshak Nair, Lakshmi G Pillai
Dept of Computer Science & Engineering
Amrita Vishwa Vidyapeetham
Kollam.

Abstract*: Due to the rapid development of Internet technology, the use of credit cards has increased dramatically, leading to an increase in the number of credit card frauds. Big data analytics can help detect fraud and establish stealing and might facilitate digital forensic analysis. The aim of this paper is to spot the user model that best identifies fraud cases.

1. Introduction

Cybercrime is any type of crime that can be violated or committed on or against computer networks and systems. The banking and financial industries are facing severe challenges in the form of cyber frauds, Credit card fraud being one example. This paper focuses on issues related to fraudulent methods in credit card transactions. Credit card fraud cases have grown exponentially over the past few years. Unfortunately, fraud is one of the main problems that consumers face every day.

There are different categories of credit card transaction fraud, big data enables financial institutions to combat fraud in different ways. Today, the banking industry has become a very important sector of our generation, and everyone must deal with banks online or in person. With this development, online transactions have become one of the most important forms of transactions. A credit card is one of the ways to make transactions. Online Trading Credit cards are nothing more than payment cards issued to users so that cardholders can pay merchants for the goods and services they need. The most important challenge in credit card fraud detection is to improve detection accuracy and computing capacity with the explosive growth of business data. Fraudulent transactions can be detected by a classification method or by detecting external transactions from normal transactions. Various methods and comparisons for detecting fraudulent transactions are presented in this paper.

II. Project Background

Society nowadays is dependent on the internet more than ever. This has led to increased use of credit and debit cards. The number of card transactions seems to have a rising trend over the years leading to a similar trend of stolen card numbers. This paper focuses on issues related to fraudulent methods in credit card transactions. There are different categories of credit card transaction fraud, and big data enables financial institutions to combat fraud in different ways. Credit card fraud is a huge issue that costs banks and card issuing companies a lot of money. Because of this major problem in the transaction system, bank credit card fraud is taken very seriously, and nowadays we have advanced monitoring measures in place to track transactions and identify fraud as soon as possible. As modern technology grows annually, credit card fraud continues to grow. Credit card fraud results in billions of dollars being lost for consumers and financial companies. Hence, banks and financial

institutions need better systems to detect this fraudulent activity and minimize their losses. What makes this project an interesting but challenging one is that the number of fraudulent transactions is significantly fewer than the actual transactions, making them difficult to detect.We first downloaded the credit card fraud records from the Kaggle website, and then tried some exploratory data analysis to understand the relationship between them, and then we will use it to determine the authenticity of the transaction. Anti-card fraud solutions can be divided into two categories: mainly prevention (that is, the prevention of fraud prevention) and detection (that is, measures taken after fraud occurs). The user cannot manually check whether each transaction is valid.

III. Objectives

The main objective is to reduce the impact of class imbalance and to ensure that the models work perfectly even when the number of fraud cases are low. Our focus is to reduce the false positives as well as false negatives while detecting fraud transactions so as to not to affect the customer experience. Different methods are implemented to deal with the class imbalance problem and a series of models are also implemented to date.

IV. Literature Survey

M. Sathyapriya and Dr.V. Thiagarasu proposed a model in which the performance of Apache Spark is better than other methods when implementing a credit card fraud detection system.[2]

Suraj Patil, Varsha Nemade, and PiyushKumar Soni discussed a big data analytical framework for working with a large volume of data and implemented various machine learning algorithms for observing the performance of the dataset on a real-time basis.[3]

Sk. Kamaruddin and Vadlamani Rav used a hybrid architecture including particle swarm optimization (PSO) and auto-associative neural network (AANN) to perform one-class classification (OCC) in the big data paradigm. Inside of which they attain 89% of true classification in credit card fraud detection.[1]

You Dai, Jin Yan, and Xiaoxin Tang implemented the latest massive knowledge technologies like Hadoop, Spark, Storm, HBase, and so forth. A paradigm is enforced and tested with an artificial dataset, that shows nice potentials of achieving the higher than goals.[4]

Rajeshwari U and B Sathish Babu proposed Apache Spark and Hidden Markov model in which the foremost will attain the transaction data in real-time basic and former determines the fraud in the incoming transaction.[5]

V.Methodology

The technique is divided into two sections.

- Stabilizing the data
- Learn by machine

1.The dataset we're using is unbalanced because it contains a total of 2,82,807 documents, but only 492 fraud reports. With such a low number of fraud records, it would be almost difficult to develop a successful model. As a result, we strive to reduce the number of non-fraud reports to a level that is comparable to the fraud records.

2.We convert the data to pandas type so that we can easily experiment with certain parameters while also reducing the data to 50/50. As a result, we will be able to escape the imbalances that exist in our results. After we have reduced the data, we can look at the key parameters that may have an impact on our model. Better visualization, such as heatmap plots, will help us understand this.

3.Taking machine learning as an aspect. We must first train our model with new data. The data can then be divided into training and test sets. Then we use Classifiers, in this case, Random Forest Classifier and Logistic Regression are used. Card transactions are always unknown compared to previous customer transactions. The main aim of our research is to overcome the problem of Credit

Card fraud detection, the approach proposes machine learning algorithms using Pyspark.

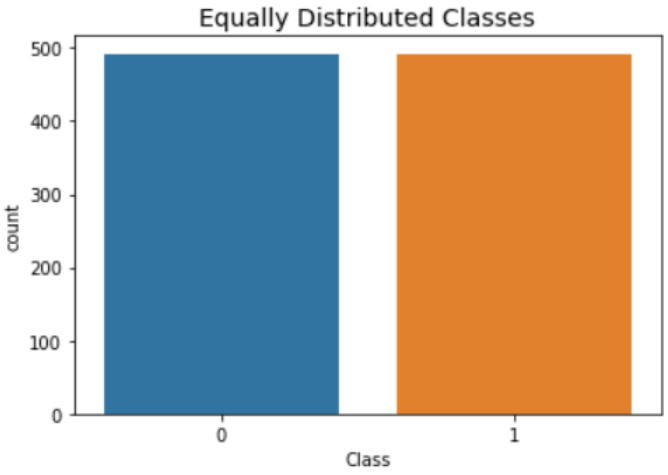
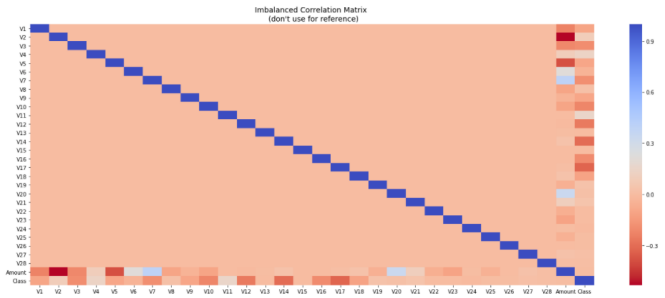


Fig1.Equally Distributing and Correlating

Fig (1) shows the plotted view of the new data balance. For examining how many numbers of exchanges are fraudulent and how many are not we introduced the graph in (Fig 1) which shows the count of fraudulent and non-fraudulent transactions. The dataset used here consists of 99 % of data that are non-fraudulent while only 0.1% seem to be fraudulent.

correlation matrix: is a table showing correlation coefficients between variables. It consists of rows and columns that show the variables. The correlation matrix is used to summarize data, as input for more advanced analysis and as a diagnosis for advanced analysis.

Correlation Matrices are used to understand the data that is present. This is used to check the amount of impact each variable has on the other. The stark difference in the two matrices is proof that sampling the data has positive effects on the correlations.The difference between the two datasets is shown below.



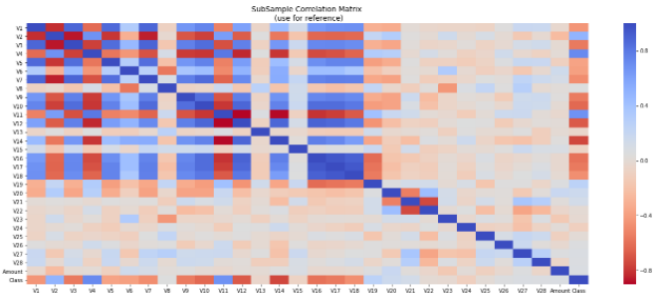


Fig 2: Correlation Matrices before and after Under-sampling

There are two main types of correlations: positive and negative. V17, V14, V12, and V10 are on the negative side, which shows that the variables are inversely correlated to the response variable. The lower these values are, the higher are the chances that the transaction is fraud. The variables with positive correlations are V2, V4, V11, and V19. The higher these values are, the more likely it is that the transaction is fraud.

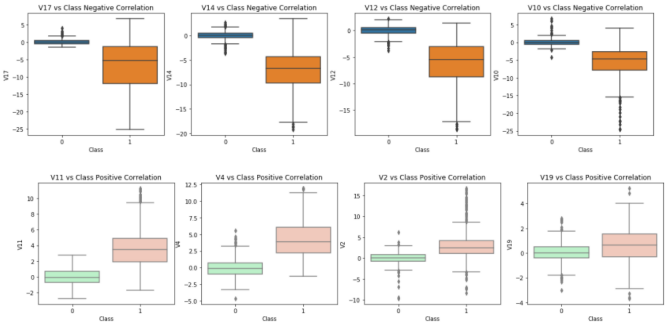


Fig 3. (a) shows Negative Correlation Box Plots (b) shows Positive Correlation Box Plots

The main agenda of this step is that the extreme outliers are removed so that the accuracy of the the model can be improved. This is done with the help of box plots and Interquartile Ranges (IQR). The interquartile range is calculated by taking the difference between the 75 percentile and the 25 percentiles. The objective is to create a threshold that can be used to delete the values which pass this border. This threshold needs to be decided in such a manner that the information loss is not at the cost of the performance of the model. The distribution of

the various variables is plotted to check the match with Gaussian distribution.

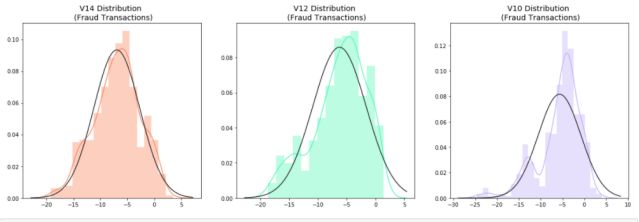


Fig 3.Gaussian Distribution of the variables

As evident from figure 3, only V14 has Gaussian distribution. The upper and the lower threshold are then calculated. Once these borders are calculated for each of the features, then the extreme outliers are removed. We use the outlier detection technique to measure the distance of each data similar to the clustering technique but is used to find specific data and rules that are separated from the total data. The values which are not an inflow of the linear graph are considered as outliers. Here our aim is to reduce the outliers to have a better-trained model. We use the NumPy library in python for this.

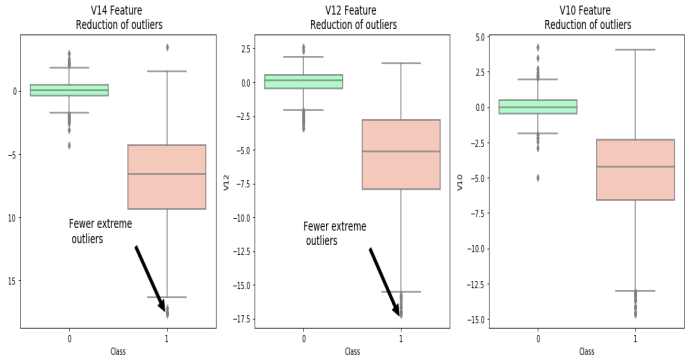


Fig 4:Feature reduction of outliers

A. Dataset

Machine Learning algorithms usually work best when the different classes contained in the dataset are more or less equally present. There is no evidence to learn how to detect fraud if there are few cases. This is referred to as a "class imbalance," and it is one of the most difficult aspects of detecting fraud.The datasets include credit card purchases made by European cardholders in September 2013. In this dataset, we have 492 frauds out of 2,84,807 transactions that occurred in the last two days. The dataset is heavily skewed, with the optimistic class (frauds) accounting for just 0.172 percent of all transactions.It contains 30 parameters in which time, amount, class and other 28 characteristics are the results of the dimensional reduction of the principal component analysis to protect the identity and confidential information associated with these credit card transactions.

B. Machine Learning Algorithm

In this part, we are going to explore two classifiers that are supported by MlLib: Random Forest and Logistic Regression.After obtaining the trained models, we will use them to make predictions on the testing data and evaluate their performance using BinaryClassificationEvaluator.

i. *Random Forest Algorithm*

In the context of supervised learning methods, random forest is a popular machine learning algorithm. It can be used for classification problems and regression problems in machine learning. It is based on the concept of ensemble learning, which combines multiple classifiers to solve complex problems and improve model performance. In the proposed system, we use the random forest algorithm to identify fraudulent transactions and their accuracy.

ii. *Logistic Regression*

Logistic regression is one of the classification algorithms used to predict the binary value of a set of specific explanatory variables. The proposed system uses logistic regression to create a classifier to prevent fraud in credit card transactions. The preprocessing step is used to process the contaminated data and ensure a high level of detection accuracy.

VI. Result

Weighted logistic regression is the best model when compared to Random Forest as Logistic Regression provides the highest metric score of 0.9888.

Random Forest generates 0.9786 as the metric score.

Algorithm	metric score
Random Forest	0.978685
Logistic Regression(weighted)	0.9888
Logistic Regression(Non-weighted)	0.9888684

VII. Conclusion

Fraud detection is a difficult problem that necessitates a great deal of planning before using machine learning algorithms. Nonetheless, it is a good application of data science and machine learning because it means that the customer's money is safe and secure. The method of detecting anomalous transactions using electronic payment log analysis and machine learning was investigated in this report.The use of electronic payment log analysis and machine learning techniques to identify anomalous transactions was explored in this report. The results demonstrate the importance of the algorithms used on the dataset, as well as efficient classification.

VIII. Reference

[1] Sk, Kamaruddin & Vadlamani, Ravi. (2016). Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification. 1-8. 10.1145/2980258.2980319.

[2] M. Sathyapriya , Dr. V. Thiagarasu (2015)Big Data Analytics Techniques for Credit Card Fraud Detection: A ReviewInternational Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

[3] Suraj Patil, Varsha Nemade, Piyush Kumar Soni,Predictive Modelling For Credit Card Fraud Detection Using Data Analytics, Procedia Computer Science, Volume 132, 2018, Pages 385-395, ISSN 1877-0509

[4] Y. Dai, J. Yan, X. Tang, H. Zhao and M. Guo, "Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies," *2016 IEEE Trustcom/BigDataSE/ISPA*, 2016, pp. 1644-1651, doi: 10.1109/TrustCom.2016.0253.

[5 Rajeshwari U and B. S. Babu, "Real-time credit card fraud detection using Streaming Analytics," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 439-444, doi: 10.1109/ICATCCT.2016.7912039.

[6] Amirneni, S. (2019). Anomaly Detection in Highly Imbalanced Dataset.

[7] Credit Card Fraud Detection Dataset, https://www.kaggle.com/mlg-ulb/creditcardfraud/tasks.