Credit Card Fraud Detection using Machine Learning

V.Sellam, P.Tushar, G.Rohit, S.Sanyam

ABSTRACT: *These days online transactions are the most Preferred mode of transactions. It's basically a constant payment method which has become a key part of our lives. But there are some problems associated with this mode of transaction which are fraud transactions that are associated with it and as the count of the online transactions increase, the count of the fraudulent transactions increases along with it. If not being completely put to an end, these. Fraud transactions can definitely be reduced to some extent. There are various methods for that, out of which data analytics and machine learning are one of the methods First a data set is provided from which the maximum, minimum and standard deviation is found. Using this a histogram is plotted which provides a visualisation of the data. Once this is done, 2 groups of graphs are created using the data which are the amount vs class graph and type of transactions vs time graph. Then later 3 machine learning algorithms are used that is light GBM , Adaboost and random forrest classifier to provide the recall , precision and accuracy of the model. A function to find the time taken to run these algorithms is also used. In the end, the value provided by these 3 algorithms are compared to find the one which provides the best result.*

I.INTRODUCTION

Ever since the rapid growth of E-commerce, theusage of credit cards for online purchases has dramatically increased and has caused an explosion in the credit card fraud.Many modern technologies based on artificial lintelligence, datamining, fuzzylogic, machine learning Sequence alignment, genetic programming etc. has alsoevolved in detecting various credit card frauds. The traditional detectionmethod mainly depends on database system and the education of customers, whichare usually delayed inaccurate and not in time. After that methods based on discriminate analysis and regression analysis are widely used which can detect fraud by credit rate for cardholders and credit card transaction. For a large amount of data, it is not efficient.

II.SYSTEM ARCHITECTURE

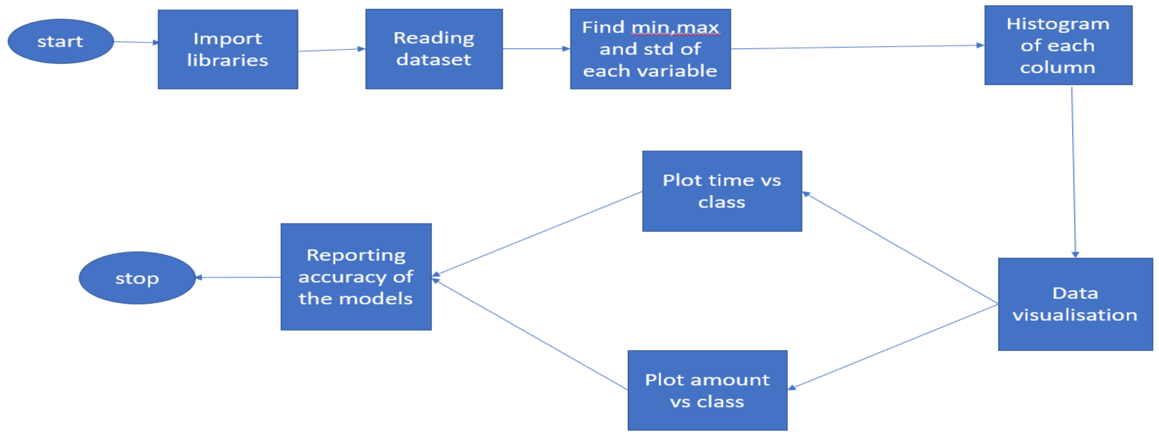

Figure-1

Figure-1 represents the system architecture of the systemwhich begins by importing the libraries are imported such as numpy, pandas, matlab which are an essential part for visualising the data.

After this, the sample dataset contains transactions made by credit cards in September 2013 by European card holders and then the max, min and std of each variable is found after which histogram of each of the predictor columns is produced. The next step is to visualise the data by producing graphs which are the time of transaction vs amount by class graph and amount per transaction by class along with this the total number of fraud cases in test dataset is also produced thus providing a clear picture of the data and number of fraudulent transactions. Once this is done the accuracy , precision and recall of the project is produced using 3 machine learning algorithms which are Light GBM , Adaboost and random forest classifier along with this a time function is added to produce the time taken for these algorithms to run and so a comparison can be done using the 4 outputs.

Manuscript received on 27 January 2021 | Revised Manuscript received on 06 February 2021 | Manuscript Accepted on 15 February 2021 | Manuscript published on 28 February 2021.
***** Correspondence Author

V.Sellam*, Asst Professor, Department of CSE, SRM University, Ramapuram, Chennai, Email: sellamv@srmist.edu.in

P.Tushar, Student, Computer Science Engineer, SRM University, Ramapuram, Chennai, Email: tushar.dbz25@gmail.com

G.Rohit, Student, Computer Science Engineer, SRM University, Ramapuram, Chennai, Email: rohitgurmith88@gmail.com

S.Sanyam, Student, Computer Science Engineer, SRM University, Ramapuram, Chennai, Email: rebelamy12@gmail.com

III.HISTOGRAM OF EACH OF THE PREDICTOR COLUMNS

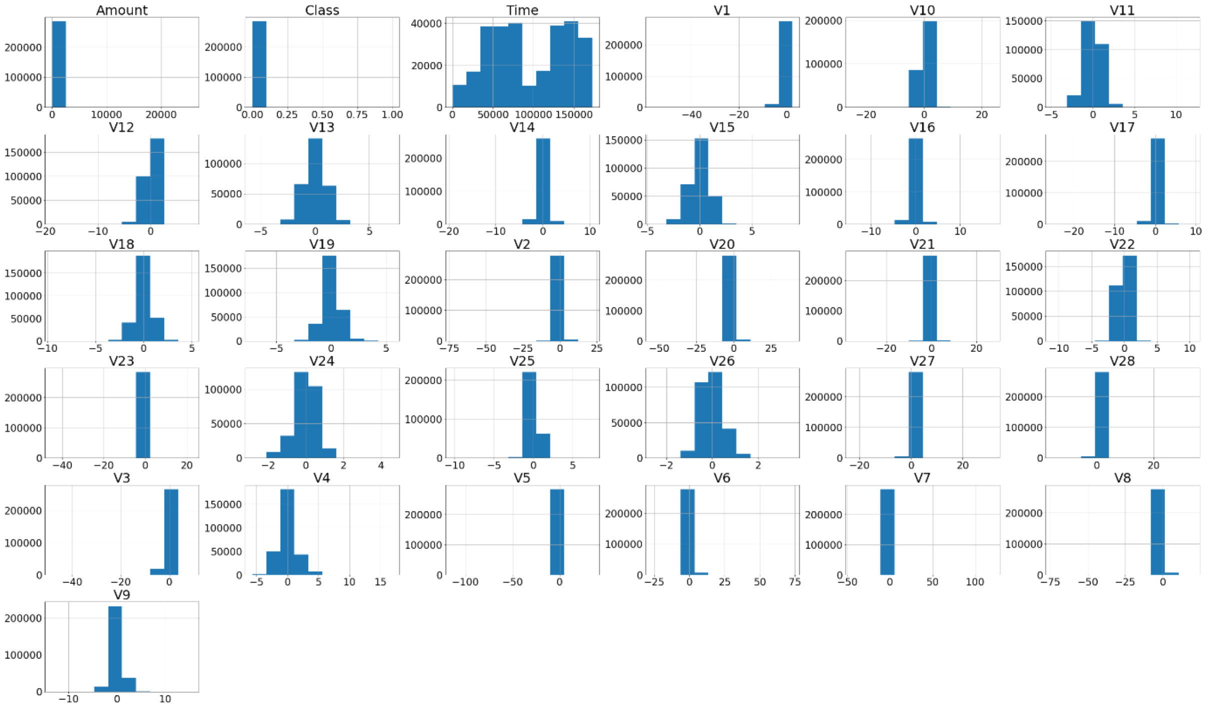

Figure-2

Figure-2 represents the histogram of each column in the dataset. This is obtained by first reading the provided dataset , then the maximum , minimum and std of each variable is found using describe() function .Once this is done, a histogram of each of the predictor columns is generated . This allows for a better visualization of our dataset once the max , min and std has been found.

IV.DATA VISUALIZATION

Data visualization is done using various python libraries which are made available such as Matplotlib , Seaborn etc. Using the various data visualization libraries made available , the data will be visualized into

a couple of graphs illustrating the number of fraud cases in the dataset along with a time of transaction vs amount by class graph and a amount vs class graph.

4.1.NUMBER OF FRAUD CASES

By incorporating sampling and normalization using standard scaler , the number of fraud cases present in test dataset can be generated .In the case of this test dataset, the total fraud cases in test data set is 156.

4.2.TIME VS CLASS

The Figure-3 graph is generated to illustrate the time of transaction vs amount by class graph. The Figure-3of the time vs class graph helps acquire a better understanding of the type of transaction that occurs, that is the number of transactions that are fraudulent and normal. This is visibly apparent from figure-3.

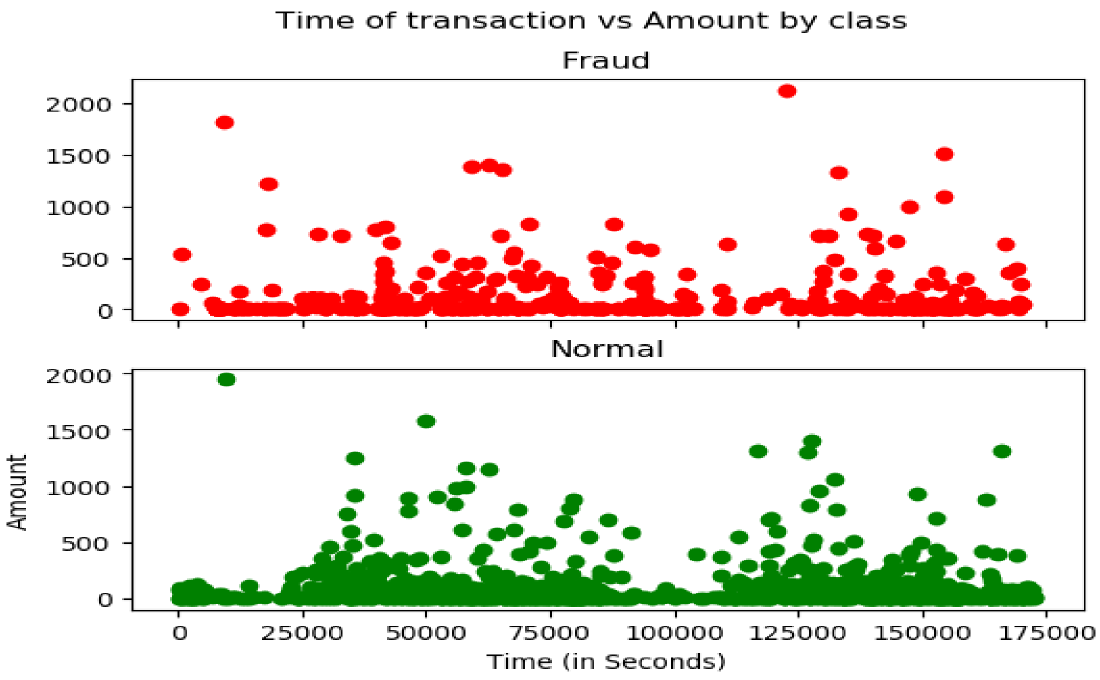

Figure-3

4.3.AMOUNT VS CLASS

The other aspect of data visualization is the amount vs class graph. Figure-4 produces a clear graphical representation of the amount per transaction by class graph. It consists of 2 types of graphs, one which covers amount of transactions vs fraud , and ,another which covers amount of transactions vs normal .This presents a clearer picture of the number of transactions that occur in both cases. Figure-4 paints a clear picture of the dataset thus providing the required data regarding fraudulent transactions.

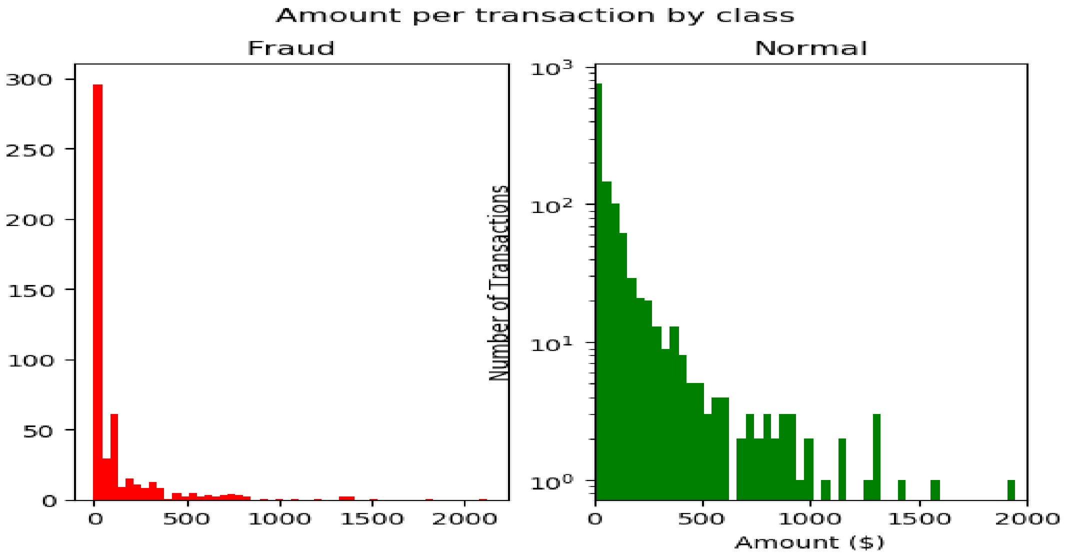

Figure-4

V.SELECTION OF ML MODEL AND REPORTING ITS ACCURACY

5.1.LIGHTGBM

LightGBM is a gradient boosting framework that uses tree based learning algorithm. LightGBM is implemented tree leaf-wise while other algorithms are implemented level-wise. The accuracy , precision and recall of the model is generated using light GBM that is

Confusion Matrix: [[395   2]
 [22 150]]
Accuracy: 0.9578207381370826

Precision: 0.9868421052631579
Recall: 0.872093023255814
Time of Light GBM : 0.8786146640777588

5.2. ADABOOST CLASSIFIER

Adaboost stands for “Adaptive Boosting”. It aims to convert a set of weaker classifier into a strong one. This is a useful algorithm for fraud detection systems. The accuracy , precision and recall of this model is implemented as well thus giving us the following result

```
Accuracy: 0.961335676625659
Confusion Matrix: [[394   3]
 [ 19 153]]
Precision: 0.9807692307692307
Recall: 0.8895348837209303
Time of AdaBoost : 1.091081142425537
```

5.3. RANDOM FOREST CLASSIFIER

It is probably the most popular classification algorithm. This algorithm is used for both classification and regression problems. As the name suggests it creates a forest with trees, the more trees in the forest the more accurate the model is. The accuracy, precision and recall of the model is produced using random forest classifier which is

```
Accuracy: 0.9595782073813708
Confusion Matrix: [[394   3]
 [ 20 152]]
Precision: 0.9806451612903225
Recall: 0.8837209302325582
Time of Random Forest Classifier: 0.4198784828186035
```

VI.CONCLUSION

In conclusion, the number of fraudulent transactions has been identified. In the beginning, the minimum, max and standard deviation of the sample dataset is found. Using this, histogram (Figure-2) and 2 graphs (Figure-3 and Figure-4) have been generated , one is the time vs class graph and the other being the amount vs class graph. As well as calculating the accuracy of the model using 3 algorithms, which are Light GBM, Adaboost and Random forest classifier. Comparison made of the 3 algorithms reveals which of produces the best result. In terms of accuracy, Ada Boost provides the highest result with 0.9613. In terms of precision, Light BGM produces the highest result with 0.986. In terms of recall , Adaboost provides the highest recall with 0.889. In terms of execution time, Random Forest Classifier executes the fastest among the 3 present algorithms.

FUTURE WORKS

With regards to the project, there is still a lot of room for improvement in terms of efficiency. After recognising the impediments in methodology of the program it can concluded that this program can benefit from improvements from other fields as well as adding even more parameters when it comes to detecting credit card transaction fraud. Some of the parameters for detecting credit card fraud transactions that can be included are location , real time credit card information as well improved efficiency to support the stated parameters. Future works will include adding the above stated features into a successful working model which can efficiently detect fraudulent credit card transactions with a real time dataset rather than a sample one.

REFERENCES

1. Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015 [CrossRef]
2. Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon [CrossRef]
3. Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE [CrossRef]
4. Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)
5. Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier [CrossRef]
6. Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019 [CrossRef]
7. Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud DetectionInformation Sciences, 2019