Fraud Detection in Credit Card using Machine Learning Techniques

Mr.Manohar.s[1], Arvind Bedi[2], Shashank kumar[3], Shounak kr Singh[4]

[1](Professor/ CSE, SRM Institute of Science and Technology)
[2](B.Tech - Student/ CSE, SRM Institute of Science and Technology)
[3](B.Tech - Student/ CSE, SRM Institute of Science and Technology)
[4](B.Tech - Student/ CSE, SRM Institute of Science and Technology)

Abstract: Credit card fraud happens frequently and leads to massive financial losses .Online transaction have increased drastically significant no of online transaction are done by online credit cards. Therefore, banks and other financial institutions support the progress of credit card fraud detection applications. Fraudulent transactions can happen in different ways and they can be placed into various categories. Identification of fraud credit card transactions is important to credit card companies for the prevention of being charged for items transaction of items which the customer did not purchase. Data science along with machine leaving helps in tackling these issues. The fraudulent transactions are mixed up with legitimate transactions and the simple recognition techniques which include comparison of both the fraud and the legitimate data are never sufficient to detect the fraud transactions accurately. This project intends to illustrate the modelling of a knowledge set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling of credit card transactions which has happened earlier with the data of fraud transactions. Our model will determine whether a new transaction tends to be fraud or legitimate. We have an objective to detect 100% of the fraud transactions while reducing invalid fraud classifications.

I. INTRODUCTION

Fraud credit card transaction is the unbidden use of someone's account without the owner being aware of it .Prevention measures need to be taken against such fraudulent practices by analysing and studying these fraud transactions to avoid similar situations in the upcoming transactions.

Briefly, Credit Card transaction Frauds can be explained as a scenario in which a fraudster utilizes a credit card of another person for personal means without seeking permission or authorization of the owner of the credit card and the credit card issuing officials or institutions are unknown of the fraud. In order to detect fraud, there is a need to monitor the activities of the users to avoid abnormal behaviour which include intrusion, Fraud and defaulting. Machine learning and data science are the communities that focus on problems like these since the solution has greater possibility and feasibility for being automated .From the perspective of learning this is a very challenging problem as it is distinguished by many factors such as class disparity. Usually the legitimate transactions are more than the fraudulent ones. Furthermore, the statistical properties of transaction arrangement changes frequently over a time period .Real -word execution of credit card fraud detection faces many more obstacles. However In real world instances, automatic tools scan the vast stream of transaction requests that tells which transaction to legitimise. Machine learning algorithm are used to inspect all the ligament transaction and list out transaction which are suspicious.

The reported suspicious transaction one investigated by professionals. They contact the cardholder to identify whether the transaction was authentic or fraudulent. The automated systems are the updated by the investigators as a feedback which helps the system to further train and improve the effectiveness of the fraud detection over time .Fraud detection needs to be constantly updated to defends against against merging fraudulent strategies by the criminals.

II. LITERATURE SURVEY

Fraudulent transactions act as the illegal activities which are meant to generate personal or financial profit. It is an intentional act that is criminal with an intention to make financial profit.

There are a number of literature or research papers available on this domain in the public platform. A detailed survey study conducted by Clifton Phua and his interns suggested methodologies used in this domain are data mining applications, Fraud detection, adversarial detection. In another research paper Suman threw lights that techniques like supervised and unsupervised learning for fraud detection. Indeed these methods were very effective and efficient in some areas of the domain but they failed to give a permanent solution to the fraud detection in credit cards. In a similar research paper by Wen-Fang Yu and Na Wang, in which they used outlier mining, outlier detection mining and Distance sum algorithm to predict fraud in an experiment conducted on the credit card data of some particular commercial banks. Outlier mining is a technique of data mining which is mostly applied in the fields related to finance or internet. It detects the fields which are not genuine. In this technique we take fields of customers behavior and on the basis of that we determine the distance between the observed value of the field and the predetermined value. There is some literature which suggests a completely new perspective to detect fraud. In case of fraudulent transactions, there have been some studies to make the alert feedback interaction more efficient. Whenever a fraudulent transaction is encountered an alert will be generated and sent to the authorised server system which in return will deny the transaction. One of the many other aspects of Artificial Genetic Algorithm is an advancement to deal with the problem in a totally different aspect. This method resulted in more accurate fraud detection and less number of false alerts.

III. SYSTEM FRAMEWORK

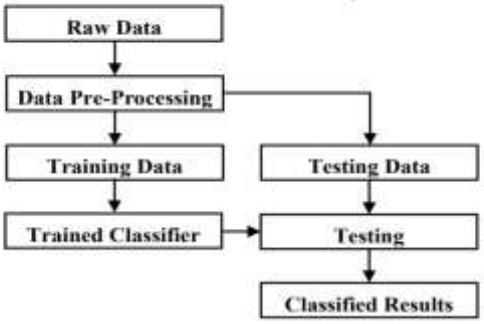


Fig.1 demonstrates the process involved in developing the model. The diagram represents the key steps involved in the development of the proposed model. The sequence of operations like data processing, data cleaning and feature extraction takes place and in the end the classification is performed.

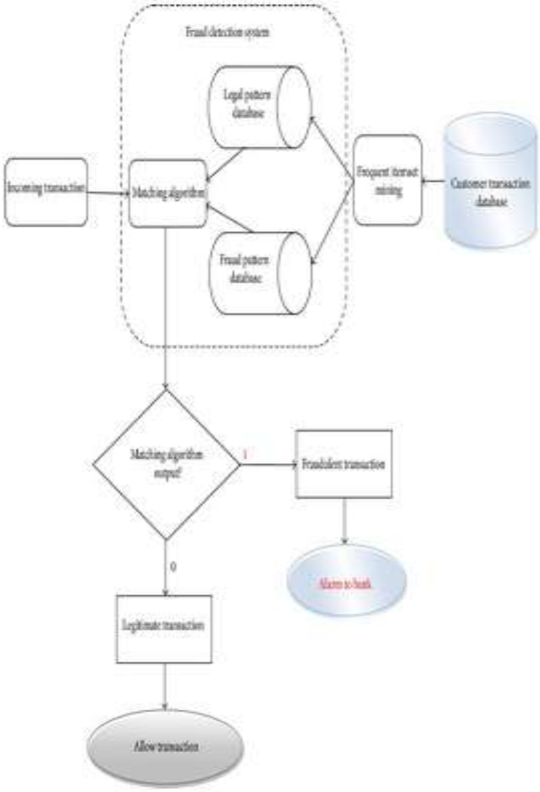


Fig.2

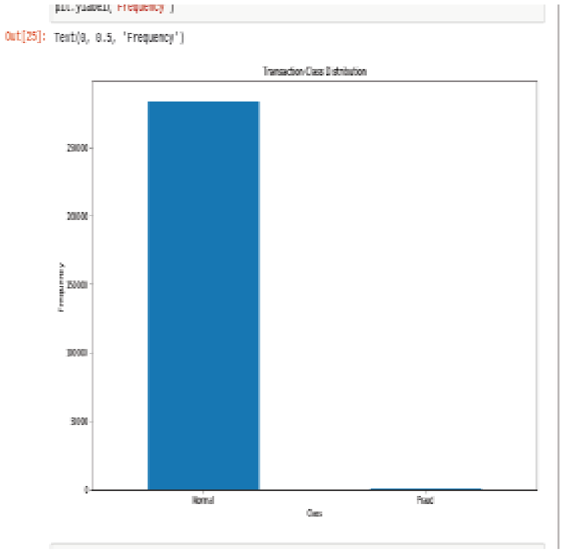
IV. METHODOLOGY

Our paper suggests an aspect of latest machine learning algorithms to detect fraudulent or anomalous events commonly called outliers. Above (Fig.2) is the architecture diagram of our proposed system.

We have used a dataset provided by Kaggle, this dataset comprises the transaction records of the european card holders in the year 2013. Inside the dataset there are 31 columns out of which 30 are used as features and the remaining 1 column is used as class. Our features include Time, Amount and Number of transactions.

We have plotted a graph which shows the inconsistency in the dataset. Kindly refer to Fig.3 for the same.

This graph shows that the number of fraud transactions is very less as compared to authorised transactions.



```
plt.ylabel('Frequency')
```

```
Out[25]: Text(0, 0.5, 'Frequency')
```

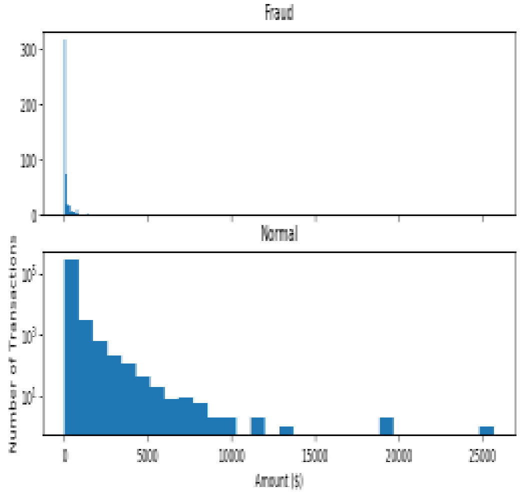
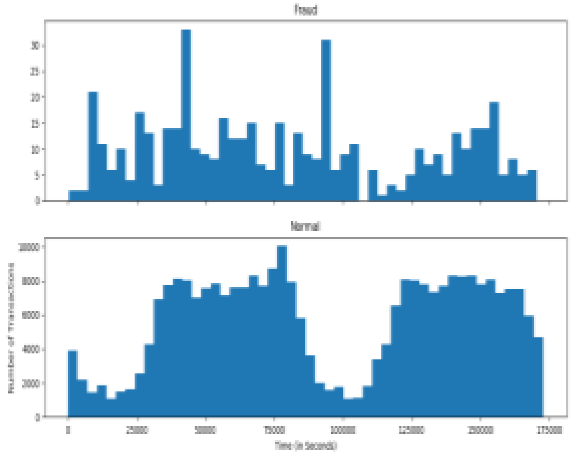


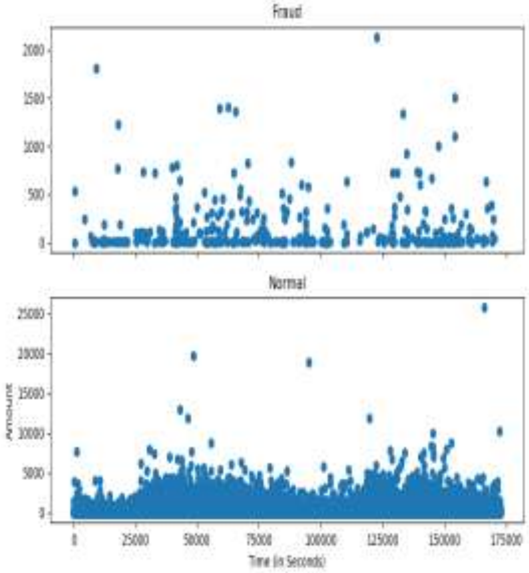
Fig. 3

To determine the pattern of the fraudulent transactions we have plotted three graphs between: Number of transactions vs Time, Number of transactions vs amount, Amount vs Time.

This graph shows the relation between Number of transactions and Amount.

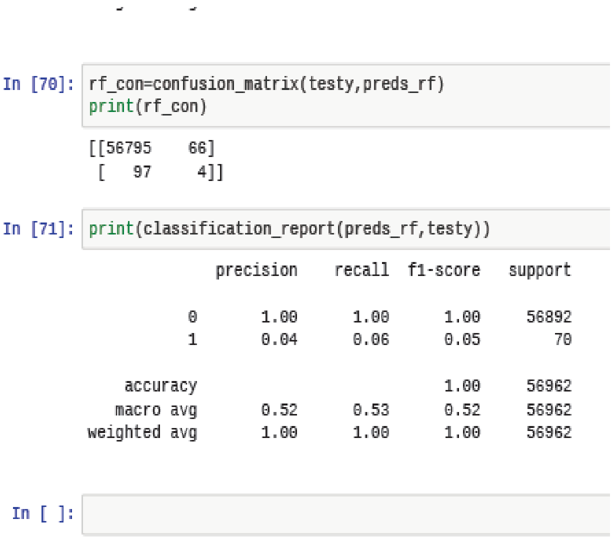


This graph shows the relation between Number of transactions and Time.

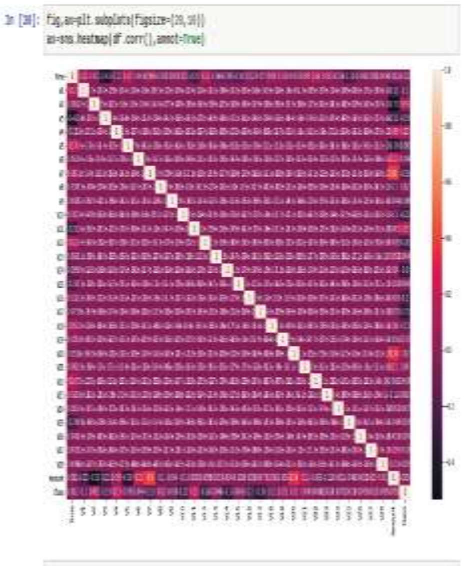


This graph shows the relation between Amount and Time.

These graphs give us a correlation between all the features of the dataset which will help us to detect an anomaly. After this step we plot a Heatmap in order to obtain a correlation between the predicting variable and the class variable.

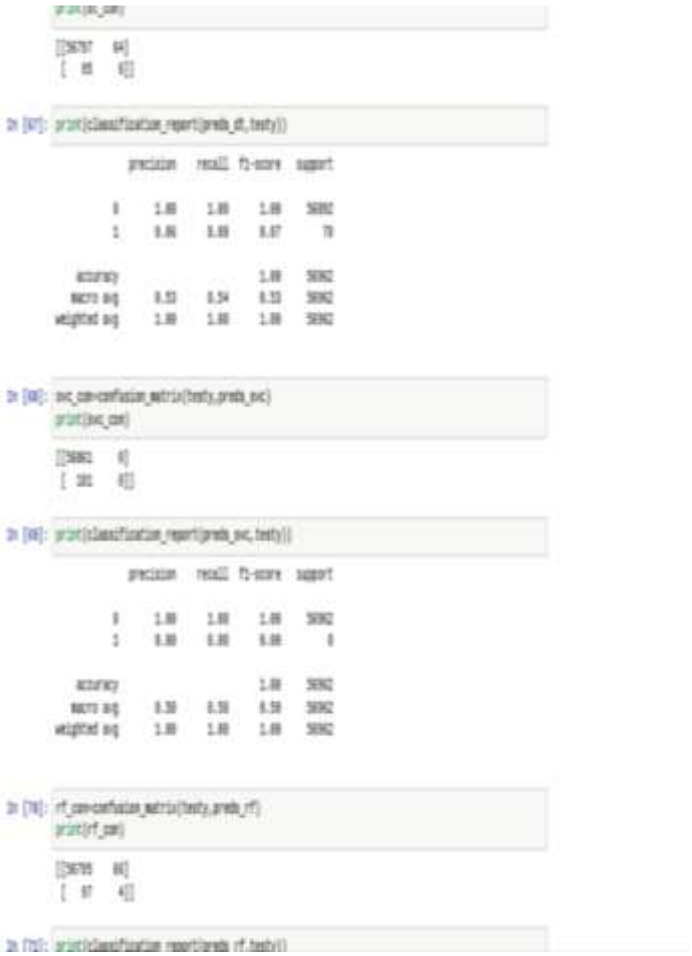


```
In [70]: rf_con=confusion_matrix(testy,preds_rf)
         print(rf_con)

         [[56795    66]
          [   97     4]]

In [71]: print(classification_report(preds_rf,testy))

                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00     56892
                    1       0.04      0.06      0.05        70

             accuracy                           1.00     56962
            macro avg       0.52      0.53      0.52     56962
         weighted avg       1.00      1.00      1.00     56962

In [ ]:
```



```
In [30]: fig,ax=plt.subplots(figsize=(20,16))
         ax=sns.heatmap(df.corr(),annot=True)
```

In succession to generating the Heatmap we will train our system with different Machine Learning algorithms so that we can extract the data and find a pattern between the data i.e. Fraudulent transactions. We will use the pattern determined between the normal transactions and the fraudulent transactions which are already in the dataset to predict a fraudulent transaction in future.



```
         print(dt_con)

         [[56797    64]
          [   95     6]]

In [67]: print(classification_report(preds_dt,testy))

                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00     56892
                    1       0.06      0.09      0.07        70

             accuracy                           1.00     56962
            macro avg       0.53      0.54      0.53     56962
         weighted avg       1.00      1.00      1.00     56962

In [68]: svc_con=confusion_matrix(testy,preds_svc)
         print(svc_con)

         [[56861     0]
          [  101     0]]

In [69]: print(classification_report(preds_svc,testy))

                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00     56962
                    1       0.00      0.00      0.00         0

             accuracy                           1.00     56962
            macro avg       0.50      0.50      0.50     56962
         weighted avg       1.00      1.00      1.00     56962

In [70]: rf_con=confusion_matrix(testy,preds_rf)
         print(rf_con)

         [[56795    66]
          [   97     4]]

In [71]: print(classification_report(preds_rf,testy))
```

The following algorithms are used to build the model and train the model for detection of frauds in credit card transactions:

Random Forest Classifier:

Random Forest can be used to classify the data into one of the many classes. In Spark, the Random Forest model can be achieved using the RandomForest class which is part of pyspark.mllib.classification module. The data needs to be made available in the required format for analysis to be performed on the same. RandomForest takes multiple parameters. The parameters are: data, numClasses, categorical Features Info, num Trees, feature Subset Strategy, impurity, max Depth, max Bins, seed.

Decision Tree:

A decision tree is a tree-like structure in which the root node and each internal node represent a "test" on an attribute of an instance in the dataset, the outcome of each test is represented by the corresponding branches and the node that does not branch further is called a leaf node and represents the class labels. It takes three parameters:

Instances – the set of instances for which class label is already known. Target_Attribute – the class label attribute. Attributes_List – the list of predictor attributes.

Support Vector Machine:

Support Vector Machines (SVM) is another classification algorithm that classifies data into one category or the other by using hyperplanes based on the training data. SVM essentially creates a model such that it finds the widest possible margins and thus the optimal hyperplane. SVM creates a separating hyperplane by transforming the data into higher dimensions where the data is separable using the kernel trick.

VI. CONCLUSION

Detection of credit card fraud is an intent part of testing for the researchers over a long time and will be an interesting part of testing in the coming time. We are introducing a fraud detection system for credit-cards by applying three different algorithms and training our machine using these algorithms with the transaction records we have. The model that we built helps the authorities to get notified of the fraud in credit-cards and take the further necessary steps over the transaction and label the transaction as fraud or legitimate transaction. These algorithms show us that the given transaction tends to be a type of fraud or not, these algorithms were selected using experimentation, discussion and feature importance techniques as shown in methodology. It is a real-time transaction data from European credit-card holders which explains the skewness of data. Therefore, we can infer that there is a requirement of applying feature selection technique. We used a PCA algorithm to select the features from our transaction dataset which uses correlation and variance as parameters to select the features. We have set the summation of variance as 95% for selecting the features using the PCA algorithm. We also applied the PCA feature selection algorithm as there are no features which have high variance and correlation with the class column which determines the transaction as fraud or not. We have applied the three machine learning algorithms as stated in methodology and the models indicate a high accuracy score for each one of them. The scores of each model were 99.7%, 99.8%, 99.7% for the decision tree, support vector machine and random forest classifier algorithms respectively. As these models have high accuracy but the predicted values have a low precision, so with the upcoming time we will be focusing on improvement in our model and get the best results with high precision in determining the fraud detections in credit card transaction records.

VII. FUTURE ENHANCEMENTS

Reaching a Goal of 100% should be the target of our accuracy score for our machine model which detects the fraud detection of credit card transactions. But reaching a 100% accuracy score we can infer that our model is being over fitted with data which is giving us the output which is being already trained for it. So, for future enhancements we can convey that the precision and confusion matrix values can be improved with a high range. We can furthermore implement new algorithms to our European credit-card holders transaction dataset and combine the results for these algorithms for precision and confusion matrix to get more legitimate values. The data set can be improved too with replacing the highly skewed values to normalized values and bringing a pattern to it which helps us in building a more accurate model. The outliers can be minimized in the present data set as they confuse the model while training it. The correlation and variance is also less in the dataset which can be improved to by minimizing the skewness of the data. These will increase the modularity and versatility of our project and make it more accurate to predict the transactions are tending to fraud or not. These improvements require the knowledge and support of experts from different sectors like: machine learning and artificial intelligence experts, data scientists and also bankers who will help us to provide the better form of data.

REFERENCES

1. *Credit Card Fraud Detection Based on TransactionBehaviour-by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 20107*

2. *CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, MonashUniversity, Wellington Road, Clayton, Victoria 3800, Australia*

3. *"Survey Paper on Credit Card FraudDetection bySuman", Research Scholar,GJUS&T Hisar HCE,Sonepatpublished by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014*

4. *"Research on Credit Card Fraud Detection Model Based on Distance Sum –by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence*

5. *"Credit Card Fraud Detection through Parenclitic Network Analysis-By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages*

6. *"Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018*

7. *"Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya,Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering .*