Exploratory Data Analysis for Credit Card Fraud Detection

Dhiraj Kumar, Diptirtha Chatterjee, Prasoon Singh Patel, Shailendra Nath Yadav, Jyoti Singh Kirar
Banaras Hindu University, India

***Abstract* – In this modern era where everything is online, making a transaction is not safe anymore, so it is duty of every bank to check the transaction details and analyse them whether it's a fraud transaction or not which is not an easy task and very time consuming in the era where time is money here comes Data Science and Machine Learning algorithm to rescue us from this vital problem. This project includes to give a complete view how to solve this problem using very sophisticated Machine Learning algorithm with Credit Card Fraud Detection, which includes modelling with the past records of Credit Card Transactions with the ones who have the Fraud Transactions. Here we want to classify all the correct Fraud Transactions to avoid any misclassifications. In this project, first we have focused in data pre-processing and data transformation, then we have performed various Classification algorithm to get a model having maximum accuracy score**

Keywords – Credit fraud detection, Classification, Data Science, Random Forest , ANN, Machine Learning.

I. INTRODUCTION

Fraud is intentionally deception action designed to provide the perpetrator with an unlawful gain or to deny a right to a victim. Type of Fraud include tax fraud, Credit card fraud, wire fraud, securities fraud and bankruptcy fraud.

Fraud involves false representations of facts, whether by intentionally withholding important information or providing false statements to another party for specific purpose for gaining something that may not have been provided without the deception.
In Credit card fraud detection necessary prevention needs to be take care to prevent such kind of unauthorized and unwanted usage of one's account by others.

This is very relevant problem needs to be solved properly using Data Science and Machine Learning Algorithm.
In practice things are not so easy to detect fraud transactions because data is not so cleaned and numerous factors are dependent for this classification. In practice massive payment requests are quickly scanned by an automatic tool to authorize those.

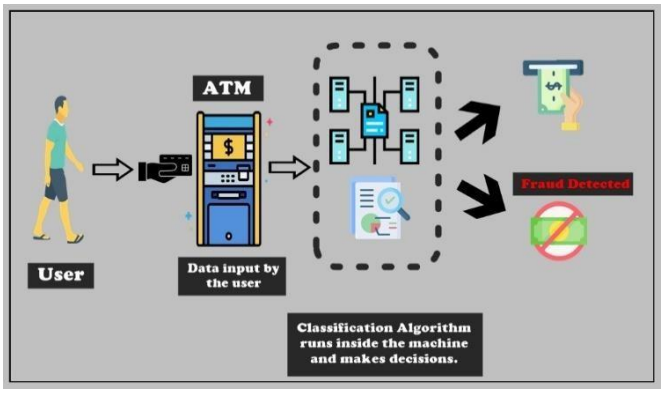

Figure 1 User interface

Various Machine Learning algorithms are deployed to analyze these authorized transactions and to predict the suspicious ones. The reports will be investigated further to confirm the accuracy of the model.
Some sort of instructions will be given on the basis of the feedback from the investigators to improve the performance of the Automated system to avoid misclassification.
This type of Fraud Detection techniques is widely used to detect:

- Credit Card Fraud
- Online Payment apps
- Application Fraud
- Telecommunication Fraud
- Account Bankruptcy etc.

Several Machine Learning Algorithms are used to detect Fraud efficiently. They are :

- Logistics Regression
- K- Nearest Neighbors
- Support Vector Machine
- Decision Tree
- Random Forest
- Artificial Neural Network etc.

II. LITERATURE REVIEW

'Fraud' itself a word that belongs to something unlawful and un authorized. Several laws are there to stop this kind of financial benefit.

Several research papers have been published already and are on the web about 'Credit Card Fraud detection'[1-10]. But many of them are focused on some particular problem like 'Outlier detection', 'model building' etc. Many of the papers have applied several Supervised and Unsupervised Machine learning Algorithm to detect Fraud. Even though there is a good result to detect Fraud in some areas, where in many situations they don't run well.

A research was done by Wen-Fang YU and Na Wang where they have used Outlier Mining, distance measure algorithms to detect Fraudulent transactions efficiently and more accurately. Outlier Mining is a part of Data Mining which is used to perform analysis in Internet and monetary fields. It can deal with those transactions which are often ignored by the automatic system. They have the customer behaviour as an attribute and based on the values they have calculated the distances between the observed and predetermined values in the Fraud detection.

III. METHODOLOGY

In this paper we have used some very sophisticated Machine Learning Algorithms. Our whole analysis is based on this flowchart given below:

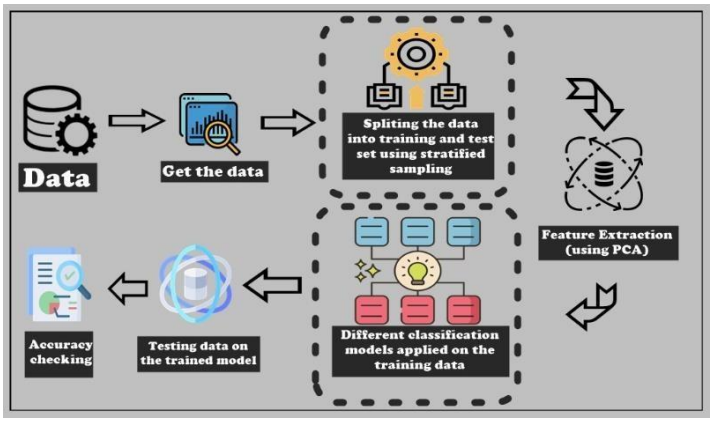

Figure 2 Flowchart of the proposed work

We have found the dataset for our analysis from Kaggle which is one of the most used platforms by the future Data Scientists.
Our dataset has 31 columns out of which 28 are named as v1-v28. Others are 'Time', 'Class', 'Amount'. 'Time' represents the time gap between the transactions on the dataset. 'Amount' is the amount of money that had been withdrawn or transferred. 'Class' has two values '0' and '1' where '0' represents a genuine transaction and '1' represents a fraud one.

First, we have checked for missing values in dataset. There were no missing values. Then, we have performed several Exploratory Analysis to draw some insights about the dataset. These are diagrams we have got by our analysis:

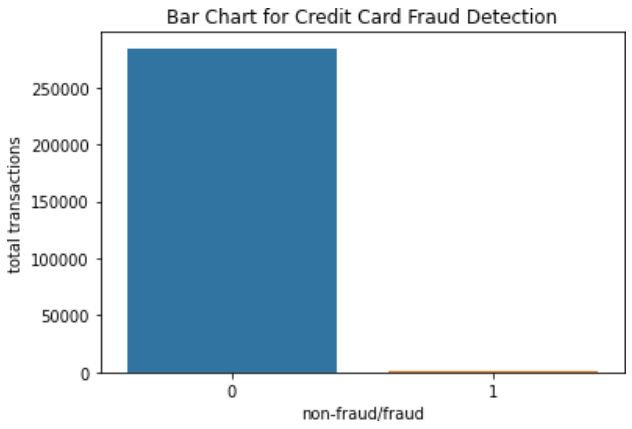

Figure 3. Bar chart of data distribution

From the above Bar chart, it is very clear that in our dataset we have very high amount of '0' means valid transactions and low amount of '1' means fraud ones. To have more numerical idea

about the percentage of valid and fraud transactions, we have also plotted a pie chart.

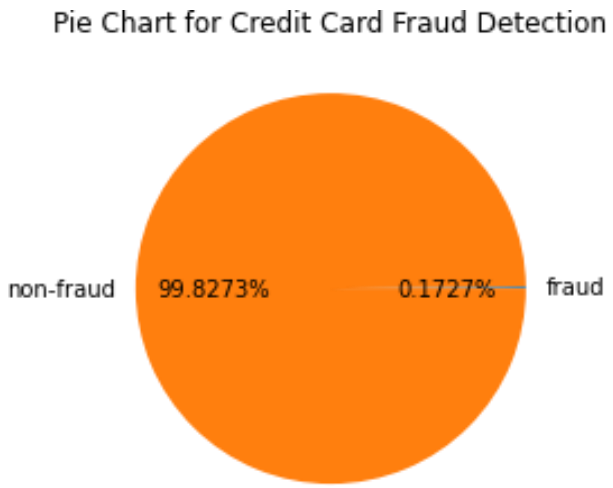

Figure 4 Pie chart

It clearly shows that 99.8273% of the whole transaction list are valid and only 0.1727% are fraud one. It's a very challenging situation. Now we have plotted a heatmap to have an idea about the linear relationship (correlation) between the features of interest. The heatmap is given below:

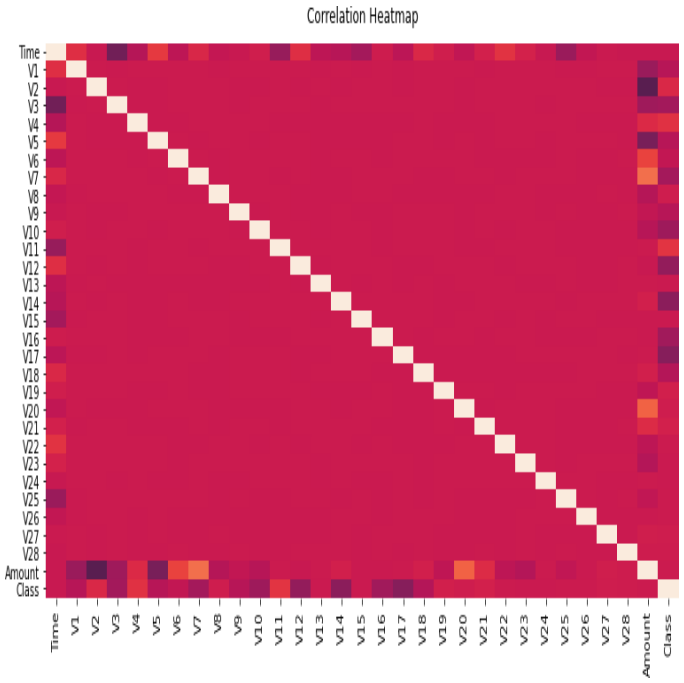

Figure 5 Correlation between features of interest

Exploratory Analysis

Now we are done with the Exploratory analysis process. We have splitted our dataset into training and testing. Here we have not used Random Sampling technique because our dataset is dominated by '0' means valid transactions. So, if we randomly sampled dataset, there is higher possibility that our train set can contain only '0' resulting bad performance of our model. So, we have used Stratified Sampling technique for splitting. It ensures that train and test data set have reasonable amount of '0' and '1'.

Next, we have scaled our features. It's a necessary step to convert all the features values to the same unit. Here we have scaled all the features into Standard Normal variable by subtracting all of them with their mean and divide them with the standard deviation.

We have 30 features which more enough to be processed by several classifiers. So, we have used Data Transformation techniques to reduce the number of features and move forward efficiently. Here we have used Principal Component Analysis (PCA) for data transformation.

We have finished our Data preprocessing phase of our analysis.

Now, it's time to build some classifier model to predict the Fraud transactions. In this analysis, we have used some sophisticated Machine Learning algorithms.
They are : Logistic Regression, K-nearest Neighbors, Naïve Bayes, Support Vector Machine, Random Forest and Artificial Neural Network

- **Logistic Regression:** It's a statistical model which gives a linear boundary to classify the predictor variable very efficiently. As it gives linear boundary, sometimes there is a chance of misclassification.

- **K-nearest Neighbors:** It's a non-parametric Supervised Machine learning algorithm used for classification. It calculates the distances of new data point from the k nearest neighbors and count the number of data point in each category.

Assign the point to that class which have counted the most neighbors.

- **Naïve Bayes:** It's a collection of classification algorithms based on Bayes theorem. It used the probability of dependent variable given that all the independent variables have already occurred. Based on the probability, it classifies the dependent variable.

- **Support Vector Machine:** It's a supervised machine learning algorithm used for Classification and regression both. Here we have used Radial basis function to classify the dependent variable. Radial Basis Kernel is a kernel function that is used in machine learning to find a non-linear classifier or regression line.

- **Random Forest:** It's one of the mostly used Supervised Machine learning algorithm for classification. Random forest is formed through a bunch of decision tree. Decision tree is collection rules through which classification is made. Using more than one decision trees, make the prediction stable and strong.

- **Artificial Neural Network (ANN):** An Artificial Neural Network is an information processing technique. It works like the way human brain processes information. ANN includes a large number of connected processing units that work together to process information. They also generate meaningful results from it.\

We have used K-fold cross validation to predict the misclassifications very precisely. Here we have taken k=5 which is sufficient enough to reduce misclassification.

IV. RESULTS

Here we have got the results about the cumulative explained variation ratio when we have used 27 components.

Table 1 Explained variance from different components

No of Component	Cumulative Explained Variance
1	0.0655522
2	0.12173668
3	0.15678751
4	0.1910612
5	0.22500437
6	0.25882544
7	0.29256692
8	0.32615887
9	0.35971397
10	0.39321304
11	0.42669142
12	0.46014037
13	0.49356824
14	0.52695778
15	0.56030777
16	0.59363685
17	0.62695067
18	0.66022101
19	0.69346699
20	0.72669242
21	0.75990356
22	0.79306064
23	0.82616005
24	0.859239
25	0.89223047
26	0.92515551
27	0.95800657

From the above table it is very clear that 27 components can explain 95.8% of the total variability. We have finished our Data preprocessing phase of our analysis.
All the confusion matrix for all the classifiers used are given below:

1. Logistic Regression:

	Predicted No	**Predicted Yes**
Actual No	56854	10
Actual Yes	37	61

2. K-nearest Neighbors:

	Predicted No	**Predicted Yes**
Actual No	56860	4
Actual Yes	28	70

3. Naïve Bayes:

	Predicted No	**Predicted Yes**
Actual No	55666	1198
Actual Yes	18	80

4. Support Vector Machine:

	Predicted No	**Predicted Yes**
Actual No	56862	2
Actual Yes	36	62

5. Random Forest:

	Predicted No	**Predicted Yes**
Actual No	56862	2
Actual Yes	33	65

6. ANN:

	Predicted No	**Predicted Yes**
Actual No	56844	17
Actual Yes	17	84

We have also calculated the accuracy score of each of the classifiers. Results are given below:

Name of Classifier	**Accuracy Score**	**Standard deviation**
Logistic regression	0.9992	0.01%
K-nearest neighbors	0.9995	0.01%
Naïve Bayes	0.9790	0.10%
Support vector machine	0.9993	0.01%
Random forest	0.9995	0.01%
ANN	0.999403	---

V. CONCLUSION

Credit card fraud detection is a mostly used machine learning problem. We have used some of the common techniques of Machine learning algorithm. We have got 99.95% accuracy by Random Forest classifier which is no doubt a great accuracy for this kind of serious crime detection problem. In Random Forest classifier we have got 34 misclassifications which is negligible to the seriousness of this problem. If we put more data in this algorithm, we expect to get more accurate predictions.

VI. FUTURE ENHANCEMENT

However, we have reached a very good accuracy, it's not the end. We will try our best to improve the model to get 100% accuracy where our model can classify correctly the valid and fraud transactions without having any false prediction. We will also add some more sophisticated algorithms to make the predictions more accurate in future.

More data always increases the precision of any algorithm. We will also try to predict this problem with more data. However, this needs some official supports from banks to face the real-life data problems.

VII. REFERENCES

1. Raj, S. B. E., & Portia, A. A. (2011, March). Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET) (pp. 152-156). IEEE.

2. Shimpi, P. R. (2016b). Survey on Credit Card Fraud Detection Techniques. *International Journal Of Engineering And Computer Science*. Published.

3. Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. IEEE Transactions on dependable and secure computing, 5(1), 37-48.

4. Padvekar, S. A., Kangane, P. M., & Jadhav, K. V. (2016). Credit Card Fraud Detection System. *International Journal Of Engineering And Computer Science*.Published.

5. C., D. V. P. (2020). Analysis of Performance on Classification Algorithms for Credit Card Fraud Detection. *Journal of Advanced Research in Dynamical and Control Systems*, *12*(SP3), 1403–1409.

6. Analysis of Credit Card Fraud Detection Techniques. (2016b). *International Journal of Science and Research (IJSR)*, *5*(3), 1302–1307.

7. Credit Card Fraud Detection System: A Survey. (2020b). *Journal of Xidian University*, *14*(5).

8. Garg, V., Chaudhary, S., & Mishra, A. (2021). ANALYSING AUTO ML MODEL FOR CREDIT CARD FRAUD DETECTION. *International Journal of Innovative Research in Computer Science & Technology*, *9*(3).

9. Porkess, R., & Mason, S. (2011). Looking at debit and credit card fraud. *Teaching Statistics*, *34*(3), 87–91.

10. Shimp, P. R., & Kadroli, P. V. (2016). "Banking Expert System"With credit card fraud detection using HMM algorithm. *International Journal Of Engineering And Computer Science*. Published.