Credit Card Fraud Detection Using Machine Learning

Ayan Neyazee[1], Jayesh Kumar[2], Soumen Sourabh[3], Prof. Avinash Kumar[4]
[1, 2, 3] Dept of Computer Science and Engineering
[4]Assistant Professor, Dept of Computer Science and Engineering
[1, 2, 3, 4] Atria Institute of Technology,Bangalore, India

Abstract- *The detection of frauds in credit card transactions may be a vital question in financial research, of intense economic implications. While this has hitherto been tackled through data analysis techniques, the resemblances between this and other problems, just like the design of advice systems and of diagnostic and prognostic medical tools, suggest that a fancy network approach may yield important benefits.The performance of fraud detection in credit card transactions is greatly affected by the systematic sampling approach on dataset, selection of variables and detection techniques used. This paper investigates the performance of Logistic Regression, Random Forest Classifier, Decision tree, Hidden Markov Model, K-Nearest Neighbour, Local Outlier Factor, Isolation Forest, Histogram,Heat map and Correlation matrix highly skewed credit card fraud data.The Dataset of credit card fraud transactions is sourced from European cardholders containing 284,807 transactions. it's supported a recently proposed network reconstruction algorithm that permits creating representations of the deviation of one instance from a reference group.This article defines common terms in credit card fraud and highlights key statistics and figures in this field.*

Keywords- Credit card fraud, applications of machine learning, isolation forest algorithm, local outlier factor, CNN Model, Heatmap.

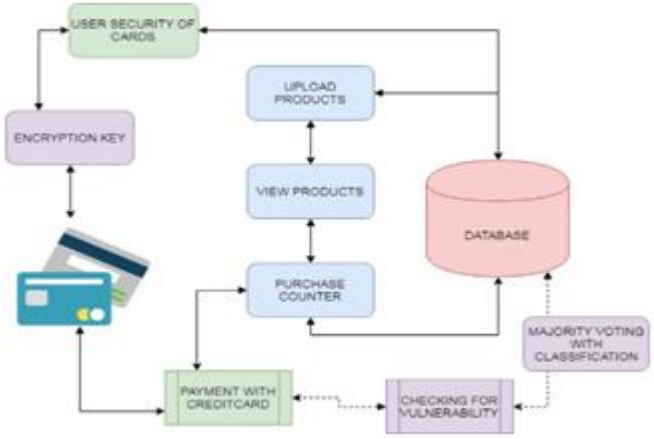
I. INTRODUCTION

In daily routine we use credit cards to buy goods and services using online transaction or physical card for offline transaction In credit card based purchase, the card holder issues his card to merchant to do payment .the person has to sneak the card to make the transaction fraudulent . If the user is not aware of loss of card it leads to financial loss to the user as well as credit card company. When the payment mode is online, attackers require only little information for doing false transaction. Example card number. The only way to detect these kind of fraud is to analyse the spending patterns on every card and irregularities are figured with respect to normal pattern. Fraud which is perceived using existing purchase data of card holder is way to reduce the growing rate of frauds. Every card holder is distinguished by figure or the patterns containing information about particular purchase category the time since the last buying, money spent and other things.

'Fraud' in the credit card transactions is unauthorized and unwanted usage of an account by someone apart from the owner of that account. Necessary prevention measures may be taken to prevent this abuse and therefore the behavior of such fraudulent practices may be studied to reduce it and protect against similar occurrences within the future.

In other words, credit card Fraud may be defined as a case where an individual uses someone else's credit card for private reasons while the owner and therefore the card issuing authorities are unaware of the very fact that the cardboard is getting used. Also the transaction screen patterns or the figure often change their statistical properties over the course of your time. These aren't the sole challenges within the implementation of a real-world fraud detection system, however. In this real world instance, the immense or the vast stream of payment requests or the solicit is quickly scanned by automatic tools that determine which transactions to authorize.The Machine learning algorithms are employed in this paper to analyse all the authorized or unauthorized transactions and report the suspicious or the doubtful ones. These reports are investigated by expert or the professionals who contact the cardholders to substantiate if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is employed to coach and update the algorithm to eventually improve the fraud-detection performance overtime.

Flow Diagram:



Some of the currently used approaches to detection of such fraud are:

- Artificial Neural Network
- Fuzzy Logic
- Genetic Algorithm
- Logistic Regression
- Decision tree
- Bayesian Networks
- Hidden Markov Model
- K-Nearest Neighbour

II. LITERATUREREVIEW

The Credit card fraud detection has received the outstanding or the significant consideration from researchers within the world. Several techniques are developed to detect fraud using credit card which are supported by many algorithms such as Bayesian networks, neural network, data processing, clustering techniques, genetic algorithms, decision tree etc. the thought of similarity tree, a spread of Decision Trees logic was proposed by Kokkinaki in 1997. a choice tree is defined recursively; it contains nodes and edges that are labeled with attribute names and with values of attributes, respectively. All of those satisfy some condition and find an intensity factor which is defined because the ratio of the quantity of transactions that satisfy applied conditions over the whole number of legitimate transaction.

The advantages of the strategy are the convenience to grasp and implement. While its disadvantages include a necessity for continual updates when user habits and fraud patterns change, because the user profiles don't seem to be dynamically adaptive. In the other way Ghosh and Reilly proposed or recommended algorithm which is neural network method to detect credit card fraud transactions. They analyzed this method and built a detection system, which uses a three-layered feed-forward network with only two training passes and is trained on an oversized sample of labeled master card account transactions. These transactions which contains a sample of all the fraud cases which is stolen cards, lost cards,unknown card, application fraud, stolen card details, counterfeit fraud etc. They tested on an information based on all the set of all transactions of credit card account over a subsequent period of your time. The system significantly reduced the investigation workload of fraudanalysts.

III. PROBLEM STATEMENT

Our goal is to implement three different machine learning models or the algorithm so as to classify, to the best possible degree of accuracy and to give exact or the nearest results, Credit card fraud from a dataset gathered in Europe in 2 days in September 2013. After initial data exploration by using this techniques, we knew we might implement other algorithm such as logistic regression model, a k-means clustering model, and a neural network. Enormous Data is processed each day and therefore the model build must be fast enough to retort to the scam in time. Contrary to popular belief, merchants are way more in danger from master card fraud than the cardholders. While consumers may face trouble trying to urge a fraudulent charge reversed, merchants lose the price of the merchandise sold, pay chargeback fees, and fear from the chance of getting their merchant account closed.

Some challenges and the main problem we observed in the dataset from the beginning were the large imbalance within the dataset which gives that frauds only account for 0.172% of the given dataset fraud transactions. during this case, it's much worse to which it possess false negatives than false positives in our predictions because false negatives mean that somebody gets away with credit rcardfraud.False positives, on the opposite hand from the given dataset which also detect the main issue here that, merely or that cause a complication and possible hassle when a cardholder must verify that they did, in fact, which they must or not complete said transaction and also they are not atheif.

IV. PROPOSED SYSTEM

In this paper, we've used the random forest and isolation forest algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. they are evaluated using both benchmark and the real-world European credit card data sets. To further evaluate the robustness, scalibility or the reliability of the models,which is further handled to the uses of the noise is added to the real-world data set. The key contribution of this paper is that the evaluation or the estimation of a spread of machine learning algorithms here predict or analyze the basic models with a real-world European credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the information sets employed in this paper are extracted from actual master card transaction information over three months.

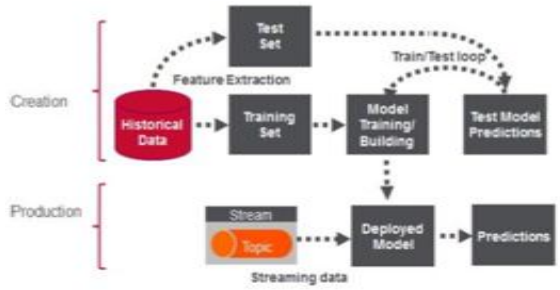


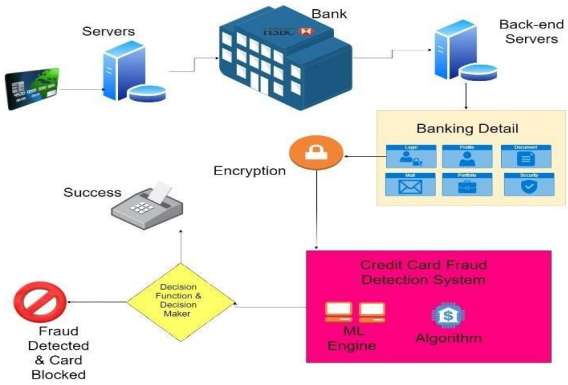
Figure. Project Flow diagram

We have imported different libraries in our project and printed their versions in order to obtain the analysis using different algorithms of machine learning in our codesowe imported necessary packages. we loaded the dataset from the csv file using pandas or numpy. and then we explored the dataset. we have 31 different columns as given in the dataset which is obtained from kaggle.There are v1 to v28 are the results of PCA dimensionality reduction to guard sensitive information in our dataset like we don't want to show identity and placement of a private. class 0 indicates valid transaction and other class 1 indicates fraud transaction.we have 284807 transactions with 31 columns. further while exploring dataset we noticed that mean values are near 0 shown in the dataset it means there are more valid transactions than fraud transactions in our dataset.in order to save lots of time and computational requirements because it could be a large dataset we will take only 10% of the info.so nowwe have for now 28401 transactions left so as to produce desired results. and then now we have observed the different plot of histogram of each parameter to check if there are any unusual parameters obtained from the dataset. and then we have obtained valid or fraud cases after that we have tested the different algorithm and then obtained accurate or the desired results.

V. METHODOLOGY

The approach that this paper proposes, uses the latest machine learning algorithms to detect anomalous activities which is called outliers factors.

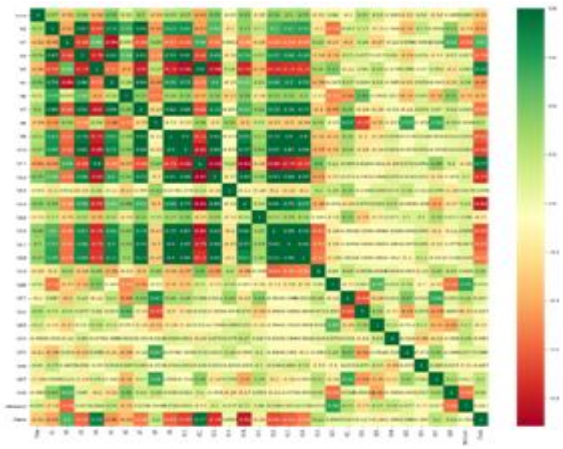
The full architecture diagram can be represented as follows:



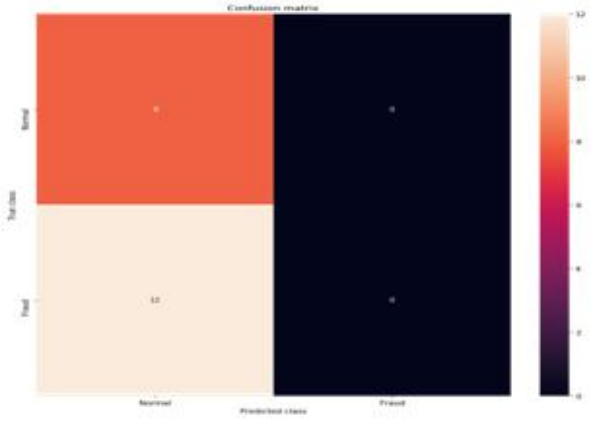
First of all, we obtained our dataset let's separate the Fraudulent cases from the authentic ones and compare their occurrences in the dataset as follows:

After this analysis, we plot a heatmap to get a coloured representation of the data and to study the correlation between out predicting variables and the class variable.

This heatmap is shown below:



The confusion matrix between the normal and fraud transaction as shown below which is obtained from the dataset:

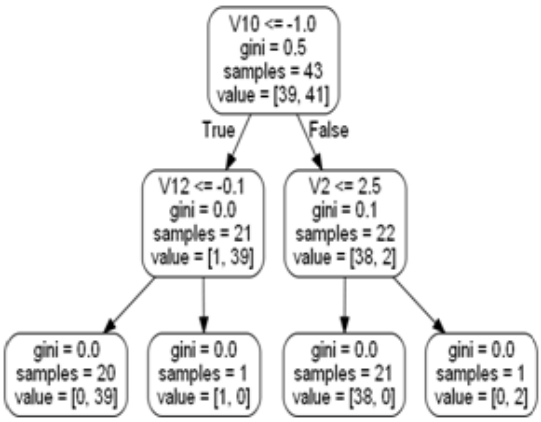


The dataset is now configured and processed further. The time and amount column in the dataset are standardized and also the Class column is removed to reconfirm the decorum of estimation and the given evaluation. the information is now processed by a collection of algorithms from the modules. the succeeding module from the algorithms in the following diagram as shown below explains how these algorithms work together: This data is fit into a model and also the Random Decision forest classifier are applied onthat.

The pseudocode for this algorithm is written as:

```
#visualizing the random tree
feature_list = list(X.columns)
#Import tools needed for visualization
from IPython.display import Image
from sklearn.tree import export_graphviz
import pydot
#pulling out one tree from the forest
tree = rfc.estimators_[5]
export_graphviz(tree, out_file = 'tree.dot', fe
ature_names = feature_list, rounded = True, pre
cision = 1)
# Use dot file to create a graph
(graph, ) = pydot.graph_from_dot_file('tree.do
t')
# Write graph to a png file
display(Image(graph.create_png()))
```

Plotting the results of visualizing the random tree:



VI. RESULTANALYSIS

The algorithm used in this paper is Random forests or Random decision forests so as there needs to be some authentic indicator in our aspect so that models built using those pecuilarity to do better assume it than the random guessing. and to grading the various learning technique for classification, regression and the other details that operate by constructing a heaps of decision trees at the training time and then producing the given output of that the class so this is the mode of the classes which is classification or mean prediction which is regression of the discrete trees.

```
# Building the Random Forest Classifier (RANDOM FOREST)
from sklearn.ensemble import RandomForestClassifier
# random forest model creation
rfc = RandomForestClassifier()
rfc.fit(X_train,Y_train)
# predictions
y_pred = rfc.predict(X_test)
```

Random decision forests correct for decision trees' habit of overfitting to their training set which gives the predictions (and therefore the errors) made by the individual trees need to have low correlations with each other and the results obtained as shownbelow:-.

```
The model used is Random Forest classifier
The accuracy is 0.95
The precision is 1.0
The recall is 0.9166666666666666
The F1-Score is 0.9565217391304348
The Matthews correlation coefficient is0.90267093384844
```

VII. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. this text has listed out the foremost common methods of fraud together with their detection methods and reviewed recent findings during this field. This paper has also explained thoroughly, how machine learning is applied to induce better ends up in fraud detection together with the algorithm, pseudocode, explanation its implementation and experimentation results.

While the algorithm we used in our project does reaches over 99.6% accuracy, althoughits precision remains only at 28% when a tenth percent of the data set is taken into consideration. Although, when the entire dataset is taken into the algorithm, the precision rises to 33%.Then this high percentage of accuracy is to be calculated due to the huge

imbalance between the number of valid and number of genuine transactions.

Since the entire whole dataset consists of only two days' transaction records of the European Bank, its only a fraction of data that can be made available if this project were to be used on a private or merchandise scale. Existence algorithm based on the Machine Learning , the program will only increase or enlarge its efficiency over time as more data is taken into considerations.

VIII. FUTUREENHANCEMENTS

While we couldn't reach out goal of 100% accuracy in credit card fraud detection, we did end up designing a system that can, with sufficient time and data, we were very close to that goal. As with any such project, there is some room for improvement here. The formation and the results obtained in this project will allows for multiple algorithms to be integrated or desegregated together as modules and their results can be combined to increase the accuracy of the final result.This project can be further improved and will be further excellence with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied or fulfilled , then the modules are easy to add as done in the code. This provides a great degree of compatible and versatility to the project.

More room for improvement can be found in the dataset. As demonstrated before, the precision of the algorithms increases when the size of dataset is increased. Consequently, additional data will confidently and surely make this project more accurate in detecting frauds and turn down the number of false cases. However, this requires official support from the banks themselves with the clear understanding of the senario of this frauds.

Further enhancement can be done by making this system more secure with the use of certificates for both merchant and customer and as technology changes new checks can be added to understand the pattern and design of fraudulent transactions and to alert the respective card holders and bankers when fraud activity is identified.

REFERENCES

[1] Wikipedia https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection

[2] Googlebookshttps://books.google.co.in/books/about/Credit_Card_Transactions_Fraud_Detection.html?id=DZqkNwAACAAJ&redir_esc=y

[3] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, ―Detecting credit card fraud using periodic features,‖ inProc.14thInt.Conf.Mach.Learn.Appl., Dec. 2015, pp. 208–213

[4] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301,1982].

[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science,1989.

[7] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., inpress.

[8] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Radoand H. Suhl, Eds. New York: Academic, 1963, pp. 271350.

[9] K. Elissa, "Title of paper if known,"unpublished.