



## **Credit Card Fraud Detection Using Random Forest Algorithm**

**V.Gokula Krishnan<sup>1</sup>, S.DhineshRaj<sup>2</sup>, S.Lokesh<sup>3</sup>, S.Sudharshan<sup>4</sup>**

<sup>1</sup>Associate Professor, Department of CSE, Panimalar Institute of Technology, Tamil Nadu, India

<sup>2,3,4</sup> UG Scholars, Department of CSE, Panimalar Institute of Technology, Tamil Nadu, India

### **ABSTRACT**

Our project mainly focuses on credit card fraud detection in real world. In this proposed project we designed a protocol or a model to detect the fraud activity in credit card transactions. This system is capable of providing most of the essential features required to detect fraudulent and legitimate transactions. As technology changes, it becomes difficult to track the behavior and pattern of fraudulent transactions. With the rise of machine learning, artificial intelligence and other relevant fields of information technology, it becomes feasible to automate this process and to save some of the intensive amount of labor that is put into detecting credit card fraud. Initially we will collect the credit card datasets for trained dataset. Then we will provide the user credit card queries for testing data set. After classification process of dataset random forest algorithm is used for analyzing data set and current dataset provided by the user. After final optimization the results indicates about the optimal accuracy for Random Forest Algorithm which is 98.6% of the accuracy of the result data.

**KEYWORDS** – Machine Learning, Credit Card, Random Forest Algorithm, Data Sets.

### **1. INTRODUCTION**

Billions of dollars of loss are caused every year by the fraudulent credit card transactions. The PwC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime. Therefore, there is definitely a need to solve the problem of credit card fraud detection. Moreover, the development of new technologies provides additional ways in which criminals may commit fraud. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years. Huge Financial losses has been fraudulent affects not only merchants and banks, but also individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that, though difficult to quantify in the short term, may become visible in the long period. For example, if a cardholder is victim of fraud with a certain company, he may no longer trust their business and choose a competitor. Credit card fraud detection is a relevant problem that

draws the attention of machine-learning and computational intelligence communities, where a large number of automatic solutions have been proposed.

In a real-world FDS, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Classifiers are typically employed to analyze all the authorized transactions and alert the most suspicious ones. Alerts are then inspected by professional investigators that contact the cardholders to determine the true nature (either genuine or fraudulent) of each alerted transaction. By doing this, investigators provide a feedback to the system in the form of labeled transactions, which can be used to train or update the classifier, in order to preserve (or eventually improve) the fraud-detection performance over time. The vast majority of transactions cannot be verified by investigators for obvious time and cost constraints. These transactions remain unlabeled until customers discover and report frauds, or until a sufficient amount of time has elapsed such that nondisputed transactions are considered genuine.

## **2. LITERATURE SURVEY**

Jon T. S. Quah and M. Sriganesh[1] identified that the online banking and e-commerce have been experiencing rapid growth over the past few years. Haibo He and Eduardo A. Garcia[2] stated that with the continuous expansion of data availability in many large-scale, complex, and networked systems, such as surveillance, security, Internet, and finance. D. Sanchez, M.A. Vila, L. Cerda, J.M. Serrano[3] stated that association rules are considered to be the best studied models for data mining. M. Krivko[4] proposed a framework for hybrid model for plastic card fraud detection systems. The proposed data-customized approach combines elements of supervised and unsupervised methodologies aiming to compensate for the individual deficiencies of the methods. Siddhartha Bhattacharyya, SanjeevJhab, KurianTharakunnel[5] identified that the Credit card fraud is a serious and growing problem. While predictive models for credit card fraud detection are in active use in practice

Ryan Elwell, and RobiPolikar[6] introduced an ensemble of classifiers-based approach for incremental learning of concept drift, characterized by non-stationary environments. SanjeevJhaa, Montserrat Guillenb, J. Christopher Westland[7] identified that Credit card fraud costs consumers and the financial industry billions of dollars annually. CesareAlippi, GiacomoBoracchi, and Manuel Roveri[8] stated that the Just-in-time (JIT) classifiers operate in evolving environments by classifying instances and reacting to concept drift. Michele Carminati, Roberto Caron, Federico Maggi, IleniaEpifani[9] proposed a semi-supervised online banking fraud analysis and decision support approach. During a training phase, it builds a profile for each customer based on past transactions.

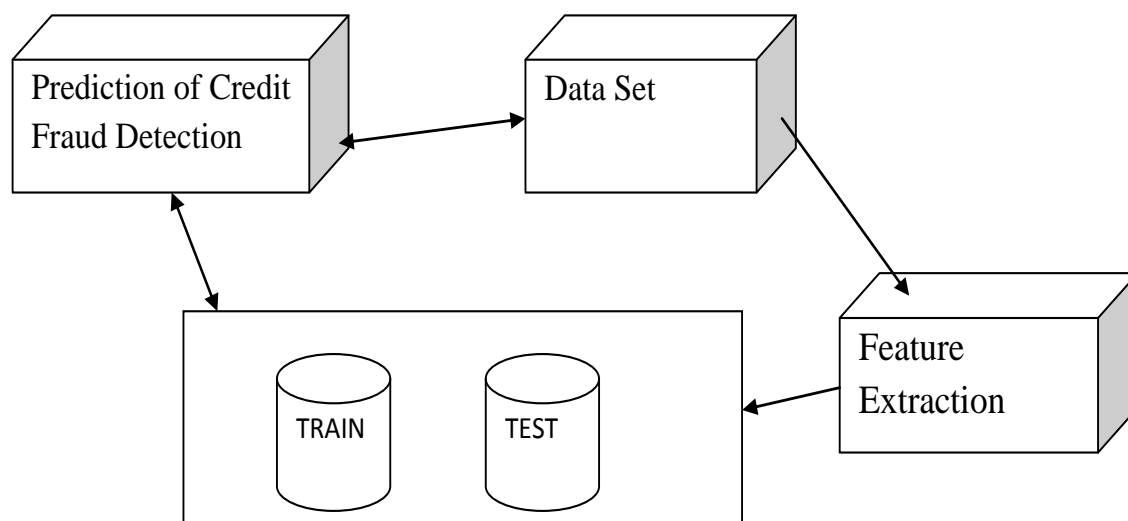
### **3. EXISTING SYSTEM**

In the existing System, a research is done about credit card fraud detection, where data normalization is applied before cluster analysis and with results obtained from the use of cluster analysis and artificial neural networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized and promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential. Accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. Concept drift change their strategies over the time based on customer habit fraudster's. Genuine transaction for outnumber frauds. Small set of transactions are timely checked by the investigators.

### **4. PROPOSED SYSTEM**

Random forest algorithm is used for classifying the credit card dataset. Random Forest is an algorithm for classification and regression. RFA is a collection of decision tree classifiers. Random forest has an advantage over decision tree as it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built. Then each node splits based on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provide a good estimate of the generalization error and to be resistant to over fitting. Random forest algorithm ranks the importance of variables in a regression or classification problem in a natural way. Random forest algorithm uses the 'amount' feature which is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (non fraud).

## 5. BLOCK DIAGRAM



In the above system architecture diagram dataset is being collected and the collected data set is classified, formatted and sampled in the process called data preprocessing and after which the feature extraction process takes place and then the random forest algorithm is used to detect whether the data set is fraudulent or not. By using Random Forest Algorithm we are splitting the data sets as trained and test data sets. The trained data sets will be more in amount than the test data sets. We will compare the trained data sets and we will see whether it matches or not. If it matches it will display the result as 0. It will display the results in 0's and 1's. 0 means (Non Fraud) 1 means (Fraud). The accuracy and speed of the result is high.

## 6. SYSTEM MODULES

Our project consists of the following modules

- 6.1 Data Collection
- 6.2 Data Pre-Processing
- 6.3 Feature Extraction
- 6.4 Evaluation Model

### 6.1 Data Collection

Data used in this project is a set of product reviews collected from creditcard transactions records. The data collection module is concerned with selecting the subset of all available data that you will be working with. ML Problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

## **6.2 Data Pre-Processing**

In this module your selected data is organized by formatting, cleaning and sampling from it. Three data pre-processing steps are:

- **Formatting:** The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.
- **Cleaning:** Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed.
- Additionally, there may be sensitive information in some of the attributes and these attributes may need to be removed from the data entirely.
- **Sampling:** There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

## **6.3 Feature Extraction**

Feature extraction module is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labeled dataset gathered. The rest of our labeled data will be used to evaluate the models. Some machine learning algorithms were used to classify preprocessed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

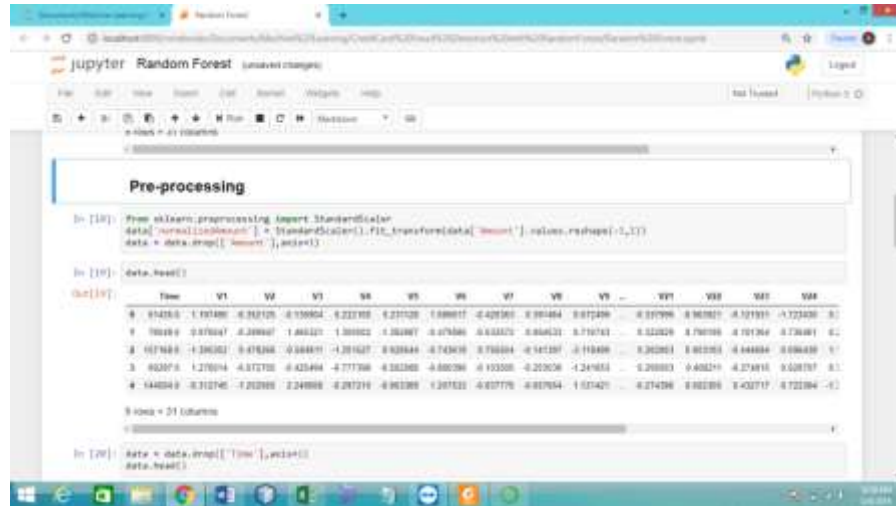
## **6.4 Evaluation Model**

Evaluation model is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) to evaluate model performance. Performance of each classification model is estimated based on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the

percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

## 7. RESULTS AND DISCUSSIONS

The launching of Anaconda navigator and after the launching jupyter is used to run the python code.



```
In [18]: from sklearn.preprocessing import StandardScaler
data['normaliseHeight'] = StandardScaler().fit_transform(data['height'].values.reshape(-1,1))
data = data.drop(['height'],axis=1)

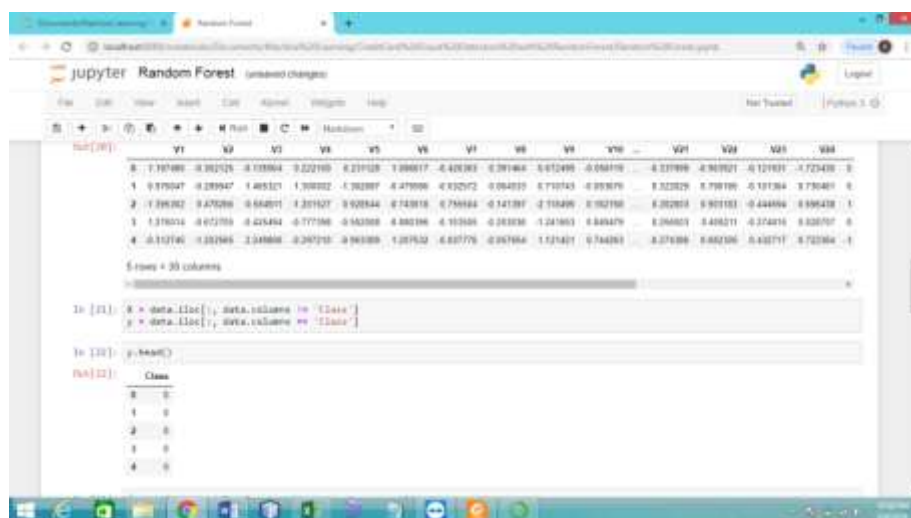
In [19]: data.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
0	97425.0	1.197486	-0.312125	-0.130804	0.233103	0.237128	1.088617	-0.428161	0.391464	0.072496	-0.317996	0.361821	-0.121931	-1.723430	0
1	10029.0	-0.317604	-0.389667	1.065321	1.301022	-0.381807	-0.476486	0.633372	0.066421	0.719763	-0.322026	0.790196	0.701266	0.736081	0
2	612768.0	-1.386303	0.478366	-0.688491	-1.281627	0.628846	0.743636	0.766804	-0.141287	-0.118406	0.382803	0.001903	-0.448884	0.066489	1
3	46307.0	1.276014	-0.172705	-0.425464	-0.777366	0.382360	-0.680396	-0.193505	-0.202636	-1.241653	-0.298803	0.406219	-0.274816	0.028787	0
4	644034.0	-0.312745	-0.212868	0.248668	-0.287216	-0.063389	1.267121	-0.077776	-0.067654	1.037421	-0.274386	0.002386	0.402717	0.722384	-1

```
5 rows x 15 columns

In [20]: data = data.drop(['Time'],axis=1)
data.head()
```

The above figure shows the output of the Data Pre-processing module. In this output screenshot only that data set will be displayed which has undergone sampling, cleaning and formatting.

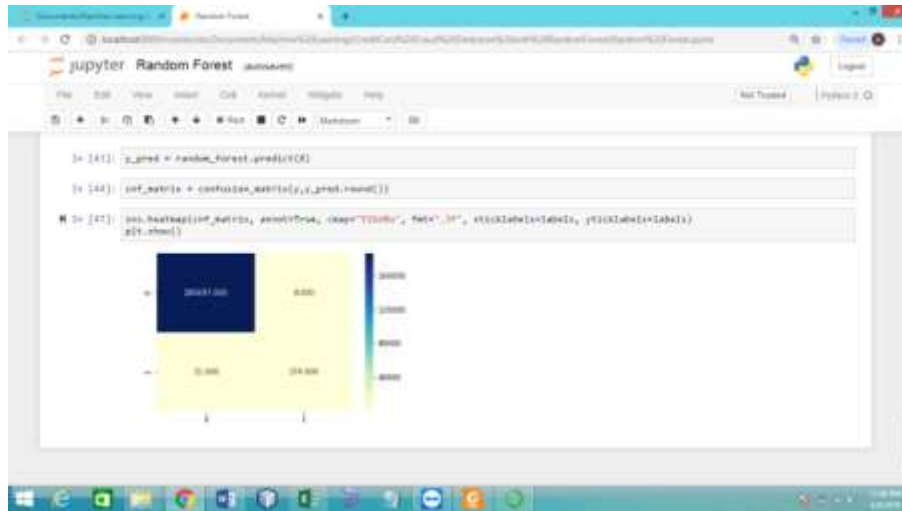


```
In [21]: X = data.iloc[:, data.columns.isin('V1:V14')]
y = data.iloc[:, data.columns == 'Class']

In [22]: y.head()
```

	Class
0	0
1	0
2	1
3	0
4	0

The above figure shows the output after the data cleaning process where the sample data set undergoes data cleaning process in which corrupt or inaccurate data from the data set is being removed.



The above figure shows the output screenshot evaluation module. Here the classified data is being represented in the form of a graph.

## 8. CONCLUSION

Fraud detection is a complex issue that requires a substantial amount of planning before using the algorithm. Here we analyze in detail the real-world working conditions of FDS and provide a formal description of the articulated classification problem involved. The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Our experiments on data sets of transactions show that, in order to get precise alerts, it is mandatory to assign larger importance to feedbacks during the learning problem. The SVM algorithm still suffers from the imbalanced dataset problem and requires more pre-processing to give better results as the results shown by SVM is great but it could have been better if more pre-processing have been done on the data.

## 9. REFERENCES

1. J. T. Quah and M. Sriganesh, "Real-time credit card fraud detection using Computational intelligence," *Expert Syst. Appl.*, vol. 35, no. 4, pp. 1721–1732, 2008.
2. H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.



3. D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "Association rules applied to credit card fraud detection," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3630–3640, 2009.
4. M. Krivko, "A hybrid model for plastic card fraud detection systems," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 6070–6076, 2010.
5. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision Support Syst.*, vol. 50, no. 3, pp. 602–613, 2011.
6. R. Elwell and R. Polikar, "Incremental learning of concept drift in non stationary environments," *Trans. Neural Netw.*, vol. 22, no. 10, pp. 1517–1531, 2011.
7. S. Jha, M. Guillen, and J. C. Westland, "Employing transaction aggregation strategy to detect credit card fraud," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12650–12657, 2012.
8. C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 620–634, Apr. 2013.
9. M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, *BankSealer: A Decision Support System for Online Banking Fraud Analysis and Investigation*, Berlin, Germany: Springer, 2014, pp. 380–394.
10. A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Detecting credit card fraud using periodic features," in *Proc. 14th Int. Conf. Mach. Learn. Appl.*, Dec. 2015, pp. 208–213.
11. A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection and concept-drift adaptation with delayed supervised information," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.
12. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *Proc. IEEE Symp. Ser. Computat. Intell.*, Dec. 2015, pp. 159–166.
13. N. Mahmoudi and E. Duman, "Detecting credit card fraud by modified fisher discriminant analysis," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2510–2516, 2015.
14. C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 246–258, Feb. 2016.
15. Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, Fellow, IEEE, and Gianluca Bontempi, Senior Member, IEEE, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," *IEEE Transactions On Neural Networks And Learning Systems*, vol 10, pp 216-23, 2018.