Credit Card Fraud Detection Using Random Forest Algorithm

M Rajeswari and K M Srinithi

June 12, 2022

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM

Rajeswari.M[1], Srinithi.K.M[2]
Assistant Professor[1], UGScholar[2]
Department of Computer Science and Engineering
IFET College of Engineering, Villupuram

Abstract - **Over few years, Credit card fraud has become one of the most common types of fraud prevalent issues. The major purpose of this project is to detection of fraudulent activity in the actual world. The recent increase in transactions has resulted with such a huge increase in illegal transactions. The purpose is to obtain things without having to pay for them or to remove money from one account without being authorized. All credit card providing institutions must implement robust detecting fraud systems in order to minimize their losses. The fact that neither cards nor card users must be present in the transaction is among the most difficult parts of running a business. As a result, the retailer has no way of determining whether or not the customer is purchasing the actual cardholder. The accuracy of the proposed scheme is achieved by utilizing the random forest. A random forest algorithm involves analyzing the dataset and the user's current dataset. Finally, improve the correctness of the outcome data and then process some of the offered qualities to identify fraud detection. A thorough review of existing and planned fraud detection methods and a comparison of these strategies, have been conducted. As a result, classification models based on the Random forest technique are applied to the data, and the model's performance is measured using graphical representations of precision, classification accuracy and f1-score.**

Keywords: **Fraudulent detection , Random classifier, Credit card fraud, fraudulent transactions.**

I. INTRODUCTION

With emergence of online money transfers in the cashless economy and the migration of businesses to the Internet, efficient detection of fraud has become a major aspect of transaction security. When a fraudster uses a credit card, by that time fraud will occur and uses the card number to buy things without the owner's permission. Because of the increased use of credit cards and an unavailability of solutions for security, identity fraud spends billions of dollars. Due to credit card companies' reluctance to disclose personal information, it's impossible to establish a reliable evaluation of the losses. The public has access to some information concerning the financial losses incurred in the course of fraudsters. When credit cards are used without proper protection, it leads to billion-dollar financial losses. The world's economic losses incurred as a result of transaction fraud totalled 22.8 (US$ by 2017) and will expand to 31 billion (Us$ by 2020). Two types of card frauds are Application and Behavior fraud. Application fraud is in which criminals impersonate a real consumer by applying for a mastercard using stolen or counterfeited documents.While this can be identified by background checks if used it allows thieves to use a valid credit card with a fraudulent written record. An equivalent type of fraud entails impersonating a customer and using a similar false written account to take over a valid Mastercard account. While this can be identified by the background checks, if applied criminals may be able to use a valid credit card with a fraudulent written record as a result of this. An analogous sort of fraud entails having a secure Mastercard account by faking the customer and making a similar counterfeit written account. A similar user can file multiple applications based on same data or by different persons with the same set of information is identity fraud. A card that has been taken or misplaced, cards that have been sold cheaply, or a cardholder who isn't available are the four types of behavioral fraud. In stolen/lost card fraud, cybercriminals acquire credit cards or recover a stolen card. Postal theft occurs when a thief gets a MasterCard or confidential details from a bank through the mail before it reaches the intended recipient. In both fake and cardholders who don't appear to be present and the Mastercard information is taken without their awareness. The proposed method's main objective is to distinguish between legal and illegal MasterCard transactions. The key contributor to recognizing fraud in MasterCard transactions using a random forest classifier is the Supervised ML Approach, which is commonly utilized in Classification and Regression.

II. EXISTING SYSTEM

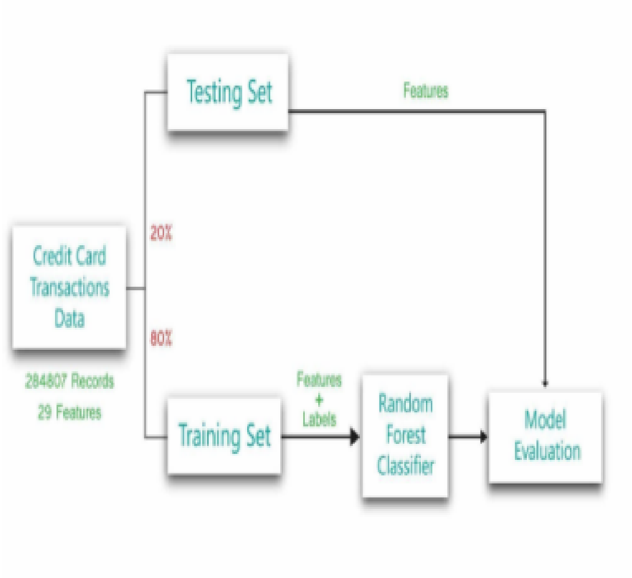
In Existing system, A study involving mastercard fraud detection, in which before Cluster Analysis data normalization is applied. The conclusions achieved from the utilization of Artificial neural networks has shown that The number of neural inputs is reduced to a minimum by clustering attributes. It is also possible to urge results by combining normalized data with MLP-trained data. And it is based on unsupervised machine learning. For classifier training, the foremost prominent algorithms for detecting mastercard fraud are supervised machine learning. The classification of knowledge taken from mastercard transactions is employed to detect fraudulent transactions. The Biological Neural Networks had the best accuracy (90%), while the Decision Trees had a 73.6 percent accuracy. An existing algorithm’s accuracy is within the range of 50% to 90% and it's difficult to classify fraud transactions from non fraud transactions because of the massive number of datasets. Significance of this paper was to improve the precision of the results.

III. PROPOSED SYSTEM

Significance of this paper was to increase the fraud detection accuracy and reduce the cost measure. For this, a Random forest algorithm has been implemented. Because it uses both bagging and randomness to produce an uncorrelated forest of decision trees, the random forest technique is an extension of the bagging method. In relation to decision trees, it has the feature of correcting the overloading of the train set. They have the Target variable values as categorical values(fraud, non-fraud) or integer(0,1). The 'Class' variable in the dataset contains only two labels: 0 (legal transactions) and 1 (illegal transactions). A random forest is made up of Decision Trees, each of which makes its own prediction. The averaged values are Regression or maximum votes are Classification to obtain the result.

IV. MODULE DESCRIPTION

SYSTEM ARCHITECTURE



The transactions dataset is obtained from Kaggle and validated, which involves the removal of duplication and the filling of blank spaces in columns. The data set is first trained using regression analysis, and then tested using the random forest approach. Then the machine learning algorithm of Random forest classifier is implemented and prediction is made based on model evaluation.

DATA COLLECTION

The datasets used for training are Fraud Detection Dataset from Kaggle. This dataset has 31 attributes, 28 of which are classified V1 through V28 and are anonymized. The remaining attributes are the amount and time of transactions also as a label whether that transaction was legal or illegal. The variables that have been anonymized are the transactions of 284,807 European cardholders. Because the average transaction value is 88.35D, the highest transaction amount is 25,691.16 USD. The majority of the transactions are minimal, as one would expect in daily transactions.The transactions take place over the course of two days. During this time, 99.8% of transactions were not normal, whereas only 0.17 percent were legit. There is also a loss of relationship between variables, which could be due to Principal component analysis converted variables.

DATA PRE-PROCESSING

Because of their different origins, the vast amount of actual-world data is susceptible to misplaced, unreliable and unpredictable. Machine learning algorithms will fail to recognize patterns effectively in this noisy data, resulting in inaccurate results. As a result, Data Processing is critical to ensuring the highest possible data quality.

The Kaggle Fraud Detection dataset was used. It has attributes of 31, which has the major features generated by using Principal component analysis. And time feature has been neglected which is useless in the creation of models. The 'Amount' feature, which contains the actual amount of money being transacted, and the 'Class' feature, which indicates whether the transaction is a fraud case or not, are the other variables.

FEATURE EXTRACTION

By reducing the number of recent variables to a smaller number, which is composed of the input parameters and essentially has almost the same data as the input parameters. It produces additional features which are a linear combination of the existing ones. When compared to the initial feature values, different values will be assigned to the new set of features.The main goal is to acquire the same data with limited functionality. We might think that choosing fewer features might cause underfitting, but within the case of the Feature Extraction technique, the additional data is usually noise.

MODEL EVALUATION

A classifier trained and evaluated by using the same data set which usually reports a higher accuracy due to the model of the same data and it has the target label of each one of instance. Model appears to be very effective due to overfitting. When the model is evaluated with training data it becomes ineffective and that results in unknown data. A confusion matrix can be generated for each classifying model prediction, representing the number of test instances that can be classified. considering the target classes as 1 for Positive and 0 for Negative.

(i) Accuracy

Accuracy is the proportion of correctly classified instances made by the models to the total number of instances predicted based on the given data. It works well only if there are an equal number of samples belonging to each class.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

(ii) F1-Score

It's also known as F-Measure. It's a metric for the developed model's Test accuracy. It simplifies our effort by removing the need to evaluate recall and precision individually by understanding the model performance. The average of recall and precision is the F1 Score. The larger the F1 Score, the greater the model's performance. Because the F1 score is accurate and stable, the model performance is obtained without separately calculating Precision and Recall.

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

(iii) Precision

Precision is proportional to the accuracy of minority class predictions, or the percentage of accurate positive predictions to all positive predictions made. When there is an imbalance in classification, the maximum class is taken as negative and there are two positive minimum classes. Precision calculates the percentage of true predictions in both the positive class and negative class.

$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Positive}}$$

(iv) Recall

The recall is a ratio of true positive predictions that models are used.The number of instances which the model correctly identified as relevant out of the total relevant instances.

$$\text{Recall} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Negative}}$$

(v) Support

The amount of real instances of the class in the dataset is known as support. It does not differ between models rather it analyzes the process of performance evaluation.

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

BUILDING MACHINE LEARNING MODEL

Now the machine learning algorithm of random forest classifier is implemented. Classifier is the ML algorithm that gives a classifier to a data point after learning the model from train data. It's an ensemble learning method for supervised learning, in which a set of specifically identified observations are available for training. They have the Target variable values as categorical values(fraud, non-fraud) or integer(0,1). The 'Class' variable in the dataset contains only two labels: 0 (legal transactions) and 1 (illegal transactions).

ALGORITHM UTILIZED

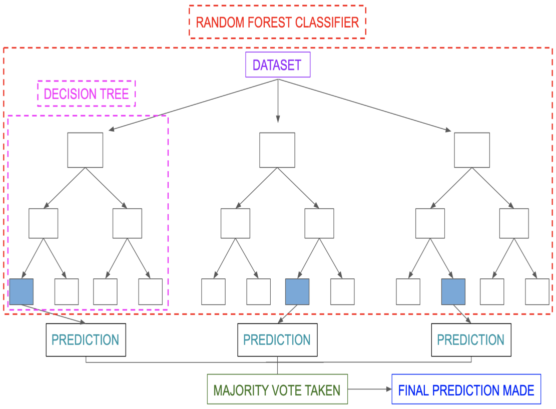
In this project, Random Forest Algorithm is implemented. It is a statistical learning model. It works well with smaller to larger datasets. Because it uses both bagging and randomness to produce an uncorrelated forest of decision trees, the random forest technique is an extension of the bagging method. In relation to decision trees, it has the feature of correcting the overloading of the train set. The prediction will be determined differently depending on the type of difficulty. Individual decision trees will be averaged in a regression task, and a majority vote on the most common categorical variable will produce the projected class in a classification task. Finally, the oob sample is used for cross-validation, bringing the prediction to a conclusion. It exhibits the distinction between the two items, allowing a large number of variables to contribute to prediction at the same time. The essential notion that makes the algorithm stronger than choice trees is aggregating un-correlated trees. To generate a higher random forest, the best idea is to create multiple sample trees and take the mean of those trees.

RANDOM FOREST CLASSIFIER

The random forest algorithm is a bagging modification that uses both bagging and randomness to produce a randomization forest of decision trees. In relation to decision trees, it has the feature of correcting the overloading of the train set. The advanced system of bagging is the random forest, which is essentially a collection of decision trees trained with a bagging process. The random forest concept entails creating many decision trees and aggregating them to obtain an accurate result. This approach is resistant to overfitting and performs well when dealing with unbalanced or missing data.

WORKING OF RANDOM FOREST CLASSIFIER

1. Initially, randomly selected samples are selected from the dataset.
2. Separate decision trees will be created for selected samples by the random forest classifier and it results in predictions from each decision tree.
3. For each predicted result, votes will be performed.
4. Finally, The majority voted predictions will be taken as the final predicted result.



ADVANTAGES OF RANDOM FOREST CLASSIFIER

1. It reduces overfitting, which improves decision tree performance.
2. Its efficiency is particularly notable in large datasets.
3. It can manage both continuous and categorical data.
4. It simplifies the process of replacing missing data values.
5. There is no need for data normalization, thus it uses a rule-based approach.

MODULE IMPLEMENTATION

IMPORTING PACKAGES

For this project, Pandas will be used to work with data, NumPy will be used to work with arrays, scikit-learn will be used to split data, implement and evaluate classification models, and then the random forest package will be used to implement the random forest classifier model algorithm. And then importing all packages into the python environment has been done.

DATA LOADING

The credit card transactions dataset was obtained from Kaggle. It has 31 attributes v(1 to 28) and other major three attributes are generated using Principal component analysis. The attributes are

1.Amount

2.Time

3.Output class

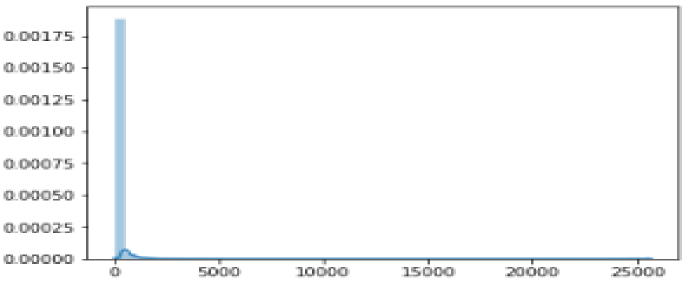
[0 for legal transactions and 1 for illegal transactions].

Except for the Output class, which is an int, the dataset contains float data values for all classes. The data from the dataset was put into a Pandas data frame in CSV format. Package for Python. Pandas provides high-performance, data analysis tools for python and open-source python libraries with a BSD license.

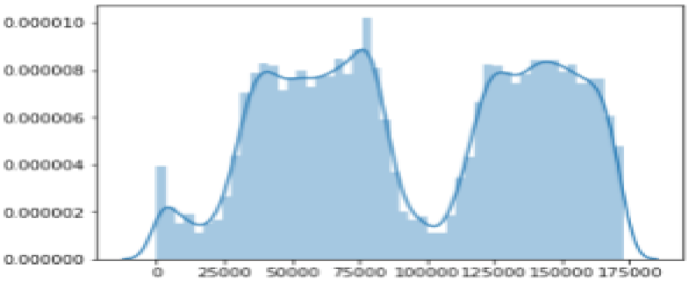
CLASS WISE ANALYSIS

A legal transaction has an Output class of 0 while an illegal transaction has an Output class of 1. There are (2,84,807 rows) and (31 columns) in the dataset. legal transactions accounted for (2,84,315) whereas illegal transactions accounted for (492). The distribution of false points accounts for 0.17 percent of the overall dataset. Using the matplotlib Python tool, the number of legal and illegal transactions were plotted as bar graphs. Graphs are generated with relationships between Time and Amount among illegal and legal transactions because Amount and Time are the only known attributes.

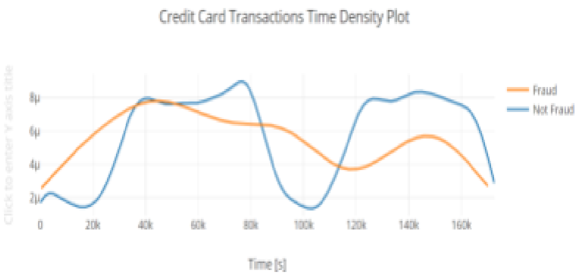
Distribution of Amount



Distribution of Time

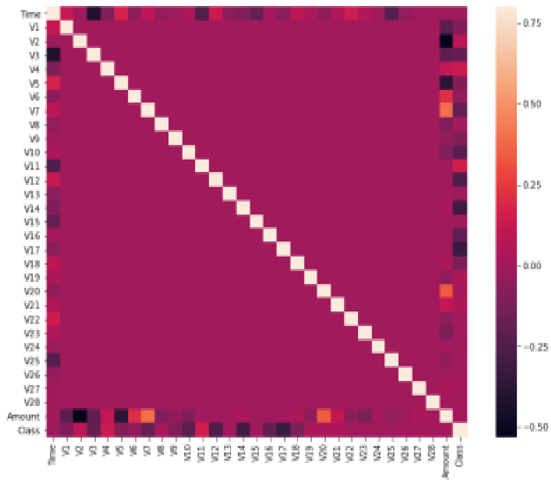


Transactions on Time



CORRELATION MATRIX

By using correlation some of the insight is evaluated if one or more qualities are dependent on another attribute or the source of another attribute. The majority of the attributes have no relation with one another, however some do have a positive or negative correlation with one another. "V2" and "V5", for example, are substantially negatively linked with the attribute "Amount."



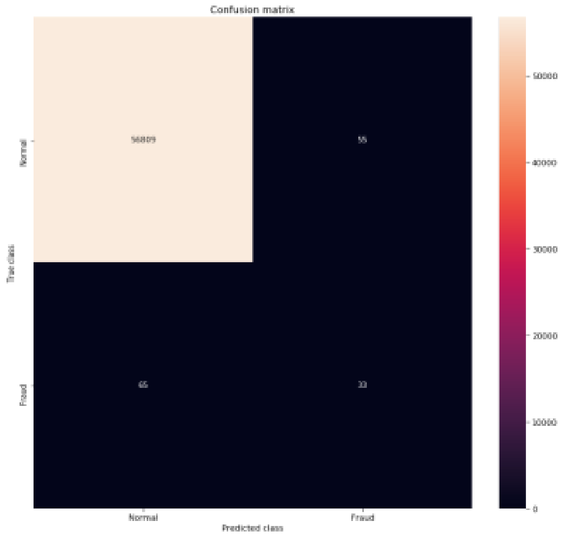
Between characteristics V1 to V28, there's no obvious connection. Because some of the features have a relationship with Amount (opposite correlation with V1 & V5), (direct correlation with V7 & V20), and Time (opposite correlation with V3).

PREDICTIVE MODELING

The Random Forest Classifier is trained using the fit function and train data. The target values will be calculated by using predict function at the same time. The real model is computed using the training set. This dataset accounts for 60% to 80% of the test set from the data (depends on Cross-Validation set). The final dataset is usually 20% of the test data. The performance of the Machine Learning algorithm is measured by how well it predicts the test set.

CONFUSION MATRIX

The confusion matrix is an NxN matrix structure for evaluating a classification model's performance, where N represents the number of predicted classes. It operates on a test dataset with known actuality values. The number of wrong and right assumptions made by a classifier and the model's accuracy is obtained. It measures the performance of a classification model. And the classification evaluation is determined by the predicted result of test data. It displays the total number of errors generated by a classifier and also the types of faults generated by the classifier. It can be used to measure Accuracy, recall, precision, specificity and sensitivity.



MODEL RESULTS

The data is investigated and unbalanced data is checked. The relationship between various characteristics is analyzed and visualized. Then the Random forest classifier is used to split dataset into train and test.

Kaggle distributions used to implement this project. Its goal is to make package management and distribution easier. The performance measurements of precision, classification accuracy, f1 score and recall, are used to compare classification algorithms.

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	1.00	1.00	56864
1	0.97	0.77	0.77	98
ACCURACY	-	-	1.00	56864
MACRO AVG	0.99	0.88	0.93	56864
WEIGHTED AVG	-	1.00	1.00	56864

CONCLUSION

The Random forest will perform efficiently with more training data and pre-processing methods, and it is the most efficient machine learning algorithm. For the varied data sets, it provides a highly accurate classifier. It works well with large databases. It can deal with tens of thousands of input variables without removing any. As a result, the Random Forest Algorithm provides more accurate credit card fraud detection results, can estimate missing data, and can handle massive amounts of data.

11. REFERENCES

[1] Adi Saputra, "Suharjito2L: Fraud Detection using Machine Learning in e-Commerce (IJACSA)", *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, 2019.

[2] Heta Naik and Prashasti Kanikar, "Credit card Fraud Detection based on Machine Learning Algorithms", *International Journal of Computer Applications (0975 - 8887)*, vol. 182, no. 44, March 2019.

[3] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi and Gianluca Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy", *IEEE on Neural Networks and Learning Systems*, 2018.

[4] V Sahayasakila, D. Kavya Monisha, Aishwarya and Sikhakolli VenkatavisalakshiseshsaiYasaswi, "Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm", *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 5, June 2019.

[5] John, Hyder, and Sameena Naaz. "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest.", International Journal of Computer Sciences and Engineering. Vol. 7, no. 4, pp. 1060-1064,Apr (2019).

[6] Varre Perantalu K., Bhargav Kiran, "Credit card Fraud Detection using Predictive Modeling "(2014) Volume 3, Issue 9 IJIRT, ISSN: 2349-6002.

[7] Hafiz K.T., Aghili S., Zavarsky P., "The use of predictive analytics technology to detect credit card fraud in Canada", 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016).

[8] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., "Cost sensitive credit card fraud detection using Bayes minimum risk". 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.

[9] Sonepat H.C.E., Bansal M., Survey Paper on "Credit Card Fraud Detection", International Journal of Advanced Research in Computer Engineering & Technology 3(3) (2014).

[10] Adewumi AO, Akinyelu AA. "A survey of machine - learning and nature- inspired based credit card fraud detection techniques." International Journal of System Assurance Engineering and Management, vol 8,no 2,pp. 937-53, Nov.2017.

[11] Yashvi Jain, Namrata Tiwari and Shripriya Dubey, "Sarika Jain: A Comparative Analysis of Various Credit Card Fraud Detection Techniques", *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 7, no. 5S2, January 2019.

[12] Wei Sun, Chen-Guang Yang, Jian-Xun Qi: Credit Risk Assessment in Commercial Banks Based On Support Vector Machines, vol.6, pp 2430-2433, 2006.

[13] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi Conference of Engineers scientists, vol. I, 2011.

[14] Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar," Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 92-95.

[15] Raj S.B.E., Portia A.A., "Analysis on credit card fraud detection methods", Computer, Communication and Electrical Technology International Conference on (ICCET) (2011), 152-156.