IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

1V. Mounica, 2S. Sandhya Rani, 3V. Ramya Lakshmi, 4 R. Anusha, 5 Mrs .Kiranmayi.

1Student, 2Student, 3Student, 4Student, 5Assistant Professor

1Vignan's Institute of Engineering for Women, 2Vignan's Institute of Engineering for Women, 3Vignan's Institute of Engineering for Women, 4Vignan's Institute of Engineering for Women, 5Vignan's Institute of Engineering for Women.

Abstract

Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm and classification algorithms such as Random Forest on the PCA transformed Credit Card Transaction data.

Keywords: Data Science, Principle Component Analysis (PCA).

INTRODUCTION

Credit card fraud events take place frequently and then result in huge financial losses. Credit card fraud is an illegal act of using credit card information without the knowledge of card holder. Credit card is used physically or online. In case of physical use, cardholders use their cards at merchant end. The fraudster has to acquire the card in its physical form through fraudulent means and use it to commit fraud. In online card transaction, information, such as, CVV code, expiry date, card number and pin code are required to commit fraud. Fraudsters acquire card information through intercepting of mails, phishing and skimming of victim's online transactions. The credit card fraud can be categorized as application fraud, stolen card, account takeover and counterfeit card. In application fraud fraudsters acquire card by submitting the fake personal documents or possess it from postal services or from card issuing company. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways. It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. The objective here is to detect major issues of the fraudulent transactions while minimizing the incorrect fraud classifications. Credit Card Fraud Detection is a typical sample of classification. In this process, we have focused on analysing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm and classification algorithms such as Random Forest algorithms.

ALGORITHMS USED

After doing research on detecting credit card online frauds, we have decided to use machine learning algorithms. Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. Based on the data (input, output) and the tasks machine learning algorithms can be classified into two types:

1. Supervised learning Algorithms
2. Unsupervised learning Algorithms

UNSUPERVISED LEARNING ALGORITHMS:

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labelled, classified or categorized.

Local Outlier Factor:

It is an Unsupervised Outlier Detection algorithm. The local outlier factor is based on a concept of a local density, where locality is given by k nearest neighbours, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbours, we can identify regions of similar density, and points that have a substantially lower density than their neighbours. These are considered to be outliers.

• LOF(k) ~ 1 means Similar density as neighbour's,

• LOF(k) < 1 means Higher density than neighbour's (Inlier),

• LOF(k) > 1 means Lower density than neighbour's (Outlier)

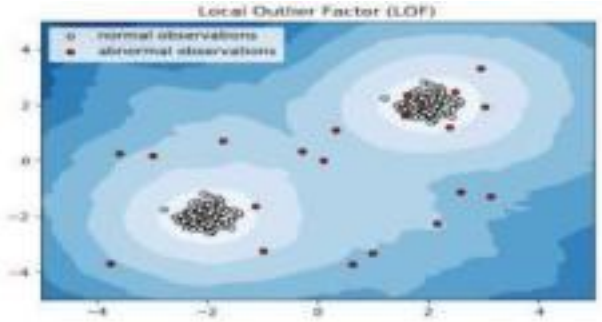

Fig. Local Outlier Factor

Isolation Forest Algorithm:

Isolation Forest explicitly identifies the anomalies by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature. The number of splits required to isolate a sample is equivalent to the path length root node to the terminating node.

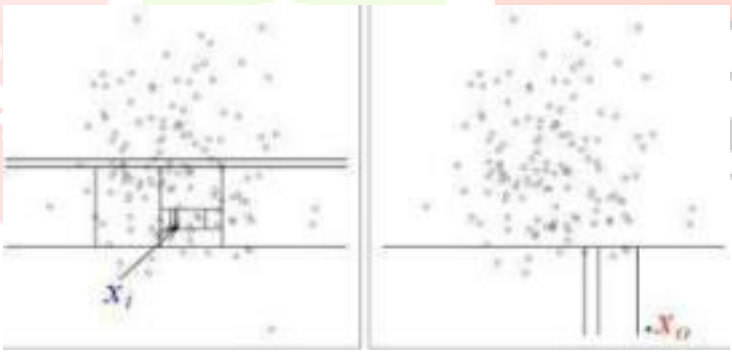

Fig. Identifying Normal vs Abnormal observations.

• if *s* is close to 1 then *x* is very likely to be an anomaly

• if *s* is smaller than 0.5 then *x* is likely to be a normal value

• if for a given sample all instances are assigned an anomaly score of around 0.5, then it is safe to assume that the sample doesn't have any anomaly.

SUPERVISED LEARNING ALGORITHMS:

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal.

Random Forest Algorithm:

Random forest is a supervised learning algorithm, which is used for both classification as well as regression. But however, it is mainly used for classification problems. Random forest is an ensemble classifier, which involves building several trees and combining the output to improve the generalization ability of the model. A subset of the training set is sampled randomly to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set.

The **Random Forest** is a model made up of many decision trees. Rather than just simply averaging the prediction of trees, this model uses two key concepts that gives it the name random:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

Working of Random Forest Algorithm:

1. **Step 1** − First start with the selection of random samples from a given dataset.

2. **Step 2** − Next this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

3. **Step 3** − In this step voting will be performed for every predicted result.

4. **Step 4** − At last select the most voted prediction result as the final prediction result.

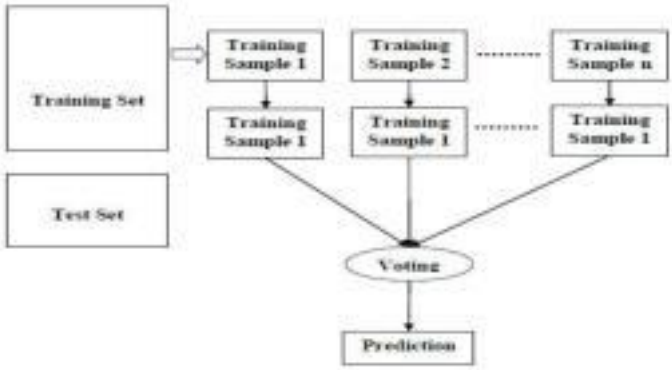

Fig. Working of Random Forest Algorithm

THE METHODOLOGYATTEMPTED

1. **Dataset description:**

   The dataset is obtained from Kaggle website, contains 284,807 transactions made by credit card holders in September 2013 for two days. There are 492 fraudulent transactions and hence the dataset is highly imbalanced. the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables, which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the original features and background information about the data are not given. There are 31 columns out of which 28 columns have been named V1, V2,

   …V28 and the remaining three columns are Time, Amount and Class. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

2. **Data Pre processing:**

After examining the dataset, it's analyzed that all the columns except Amount and Time have been scaled using PCA transformation technique. Hence Time and Amount columns are scaled using the RobustScaler to ensure uniformity.

Splitting of Data:

Separating the Independent variables from the Dependent variable: Independent variables are the features in a dataset, which are used to obtain the dependent variable. The dependent variable is the target variable, which we are trying to find. Separating the independent variables from the dependent variable in machine learning is a very crucial task.

Splitting the Dataset: The dataset is split into training set and test set. The purpose of this is to train our model with the oversampled or undersampled training set, and test it on the original testing set.

3. **Oversampling:**
   oversampling technique used for dealing with imbalanced data. The new minority instances are not just copies of existing minority cases; instead, the algorithm takes samples of the feature space for each target class and its nearest neighbors, and generates new examples that combine features of the target case with features of its neighbors. This approach increases the features available to each class and makes the samples more general.
4. **Performance metrics:**

Metrics are based on a well-known machine learning concept confusion matrix with the following four categories defined:

- TP (True Positive) – correctly predicted positive class
- FP (False Positive) – incorrectly predicted positive class
- TN (True Negative) – correctly predicted negative class
- FN (False Negative) – incorrectly predicted negative class

Confusion Matrix is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

The confusion matrix is represented as:

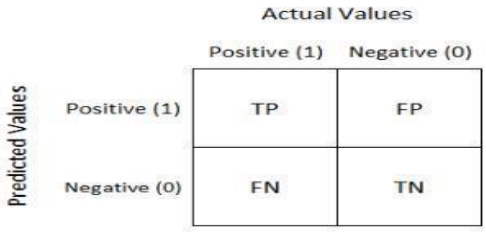

Fig. Confusion Matrix

CONCLUSION

Every application has its own merits and demerits. The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. Changing the existing models or adding new modules can append improvements. Further enhancement can be made to the application, so that the website functions very attractive and useful manner than the present one. One important development that can be added to the project in future is file level backup, which is presently done for folder level.

No.	Algorithm	Accuracy
1	Local Outlier Factor	97 %
2	Isolation Forest	71%
3	Random Forest	99 %

FUTURE WORK

Every application has its own merits and demerits. The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. Changing the existing models or adding new modules can append improvements. Further enhancement can be made to the application, so that the website functions very attractive and useful manner than the present one. One important development that

can be added to the project in future is file level backup, which is presently done for folder level. future work is to decrease the number of ownership share or reduce the size of ownership share.

REFERENCES

1.https://www.bing.com/search?q=kaggle+credit+card+fraud+detection+dataset&pc=COS2&ptag=D042220-N0340A5AF4E3D53C&form=CONBDF&conlogo=CT3335878&SearchUrlPostfix=/search&toWww=1&redig=29C58D6037B74FF6B490CB14129C1EC0

2.https://medium.com/codex/credit-card-fraud-detection-with-machine-learning-in-python-ac7281991d87#:~:text=Credit%20Card%20Fraud%20Detection%20With%20Machine%20Learning%20in,Modeling.%20...%207%20Evaluation.%20...%208%20Final%20Thoughts%21