DETECTION AND PREDICTION OF CREDIT CARD FRAUD TRANSACTIONS USING MACHINE LEARNING

Alexander Turner[1], William Carter[2]
[*1] School of Geophysics, University of Oxford, UK
[2] Department of Environmental Science, University of Sydney, Australia

ABSTRACT

The era of variety in digital convenience to make customers loyal to brands, brings along with it an empirical universe of financial fraud. In banks and other financial institutions have shown a higher leap of financial fraud over the past year. This paper describes probability of fraudulent transactions in prevalence and context of credit card usage. A conscious effort is put to bring in a conceptual distinction between fraud detection and predicting probable fraudulent opportunities on the digital space of financial transactions. This emerges a new dimension of financial fraud as a complex phenomenon that can take very different forms, depending on the market segments and the actors involved.

KEYWORDS: Fraud, machine learning, Fraud detection, artificial intelligence.

1. INTRODUCTION

Online electronic transactions are increasing with credit cards as its easier and time saving for the customers to purchase .But the number of frauds or fraudulent transactions are costing the credit card companies a huge amount annually. Hence using machine learning algorithms ,an attempt has been made to identify predict whether the transactions are fraudulent or genuine.

With the widespread usage of ecommerce and online shopping ,the customers credit card passwords, cvv numbers and other vital information are always vulnerable. The fraud users easily crack into the vital information and hence fraud cases are on the rise. The banking system also are vulnerable to online fraudulent behavior of fraud users. The identity theft in social media is on rise as the number of fraud cases are getting increased. As the fraudsters are finding new avenues or ways for fraud, the fraud prevention is a constant evolving process. Due to this, the traditional way of handling fraudulent behavior is slowly getting replaced with online fraud detection software using various machine learning algorithms.

[1] Many of the online fraud detection systems use transaction rules and also manually viewing of transactions to find discrepancies in transactions. It is time consuming process and also the fraudsters can take advantage of the time consuming process. The main disadvantage of this process is the occurrence of false positives. The false positives make the genuine users apprehensive of the transactions and may lead to cancellation of genuine user accounts.

Machine learning and deep learning plays a pivotal role in fraud detection and prediction. Machine learning algorithms along with high processing or computing power, increased capability of handing large datasets helps in detecting fraudulent transactions efficiently. With datasets involving lakhs of rows of transaction data ,it takes a lot of time to manually review and find patterns in the data for fraudulent transactions. Machine learning algorithms and deep learning also gives fast and efficient solutions to real time problems like fraud detection, medical diagnosis, email spam.

The null hypothesis is the credit card transaction is correct and not fraud. Hence false positive is whether it is a correct and genuine transaction and the system model predicts it as fraud transaction and raises a false alarm. This means completely normal customers looking to make a purchase would deter away from making purchases. False negative is a serious issue as the transaction is fraudulent and the system model predicts it as non -fraudulent. In our case, a false negative is much more serious than false positive as our system model would prove costly if it predicts fraudulent transactions as genuine.

2. LITERATURE SURVEY

A survey of various papers discussing various techniques to detect fraud are analyzed below.[2] The fraud detection in online advertising using association rule mining or APRIORI algorithm is discussed. The frequent item set and the phish tank database are analyzed for fraud patterns. The frequent item sets are analyzed from the credit card transactions and the anomalous patterns or outliers are found out. The dataset used is credit card transactions from the users. The disadvantage of the above method is when there is increase of size of the

dataset the search time of the algorithm also increases.[3]The method proposed is finding fraudulent transactions from unsupervised data using single linked hierarchical clustering .The data was extracted from the bank in Tehran. Machine-learning methods have been divided into two methods of supervised and unsupervised and the supervised method cannot be taken into practical use because it makes use of unlabeled data (the fraudulent and non-fraudulent data have not been separated and distinguished), hence unsupervised (hierarchical clustering)algorithm is used.[4] Dahee Choi and Kyungho Lee discusses a system using machine learning and artificial neural networks approach to detect fraud and process large amounts of financial data. The class imbalance problem was addressed and the usage of Synthetic Minority Oversampling Technique (SMOTE) and Random Under sampling (RUS)was applied.[5] Online mobile advertising plays a vital financial role in supporting free mobile apps, but detecting malicious apps publishers who generate fraudulent actions on the advertisements hosted on their apps is difficult, since fraudulent traffic often mimics behaviors of legitimate users and evolves rapidly. In this paper, a novel bipartite graph-based propagation approach is proposed, iBGP, for mobile apps advertising fraud detection in large advertising system. The fraud detection problem in mobile advertising is analyzed to detect fraudulent apps and introduce the initial score learning model to a large user-app bipartite graph propagation method for fraud detection. With the careful investigation of behavior patterns of mobile app users, two key characteristics are identified: power law distribution and user pertinence.[6] This paper proposes a fraud detection model based on the convolution neural network in the field of online transactions, which constructs an input feature sequencing layer that implements the reorganization of raw transaction features to form different convolutional patterns. Its significance is that different feature combinations entering the convolution kernel will produce different derivative features. The advantage of this model lies in taking low dimensional and non derivative online transaction data as the input.[7],[[10],[11],[12] papers suggest using Neural networks and Bayesian classification algorithms for credit card fraud detection.[8] Web phishing targets at stealing confidential data such as usernames, passwords, and credit card details, impersonating a legitimate entity. It will lead to data leakage and property damage. This paper suggests on applying a deep learning framework to detect phishing websites. The Deep belief Networks algorithm is used in Web phishing Detection Model.[9] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten states the feature engineering strategies for credit card fraud detection.

3. ARCHITECHTURE

The system model accepts input of real time customer credit card transactions. It is necessary to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

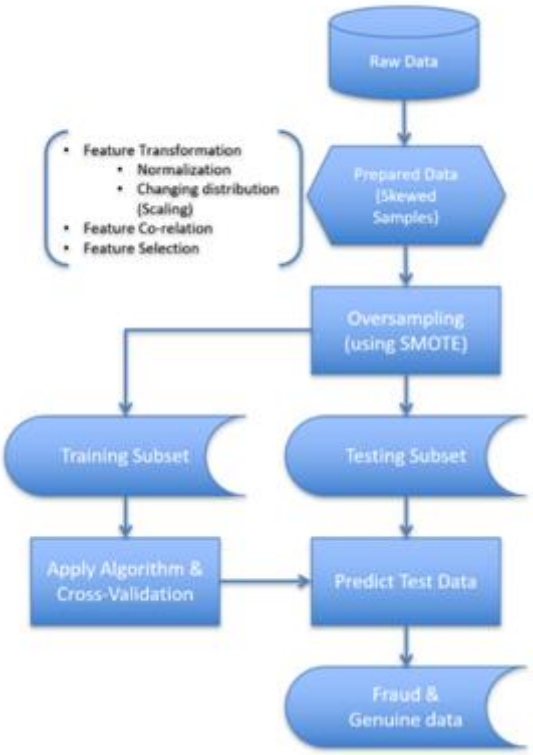

Fig 1 System Architecture Diagram

A. **Raw Data**: The collected input data is in the form of csv files.

B. **Prepared Data**: A process to gather context to the input data. Understanding the data for pre processing and cleaning of datasets. The two columns 'amount ' and 'time' were not normalised. The remaining columns were normalised using Principal Component analysis. The two features 'amount' and 'time' are scaled between -1 and 1 using Standardize data as its used in gaussian distribution.
C. **Oversampling(Using SMOTE)**: The fraud transactions are 492 samples which is unbalanced .Hence oversampling of fraud cases is performed using Synthetic Minority Oversampling Technique.
D. **Training and Testing Subset**: As the dataset is imbalanced, many classifiers show bias for majority classes. The features of minority class are treated as noise and are ignored. Hence it is proposed to select a sample dataset.
E. **Applying algorithm**: Following are the classification algorithms used to test the sub-sample dataset .
a. Logistic Regression
b. Random Forest
F. **Predicting results**: The test subset is applied on the trained model .The metrices used are precision and Recall score. The ROC Curve is plotted and the desirable results are achieved.

Understanding Data
The dataset contains 284,807 transactions made by credit card holders in September 2013 for two days.[13] 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. 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'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection(https://www.kaggle.com/mlg-ulb/creditcardfraud). It can be observed that Amounts in fraudulent transactions were always below 2,500 .
...: df.head(5)
Out[3]:

Time	V1	V2	..	V28	Amt	Class
0.0	-1.359807	-0.072781		-0.021053	149.62	0
0.0	1.191857	0.266151		0.014724	2.69	0
1.0	-1.358354	-1.340163		-0.059752	378.66	0
1.0	-0.966272	-0.185226		0.061458	123.50	0
2.0	-1.158233	0.877737		0.215153	69.99	0

[5 rows x 31 columns]

Fig 2. Credit card dataset containing V1 to V28 columns, Time and Amount.

The Time column is examined to check if any fraudulent transactions are occurring within a specific time pattern. In our dataset, it is observed that the fraudulent transactions occurring are not dependent on a specific Time pattern. Hence Time column is dropped.

After visualizing the features of dataset ,it is observed that all the feature values are uniform and hence after visualizing it is decided that all features are equally important .Hence all the features except Time are selected for further processing .It is observed that the dataset is highly imbalanced. In real time , fraudulent transactions would be very less than genuine transactions. For achieving higher results, the fraudulent and genuine transactions are separated. The fraud transactions which have class 1 are termed fraudulent and are selected.

Fig 3 Class Distributions (0:No Fraud,1:Fraud) of original dataset

No Frauds 99.83 % of the dataset

Frauds 0.17 % of the dataset

As shown above the dataset is imbalanced , the re-sampling technique is applied. After examining the amount column ,its analyzed that there are no patterns present in the amount column.

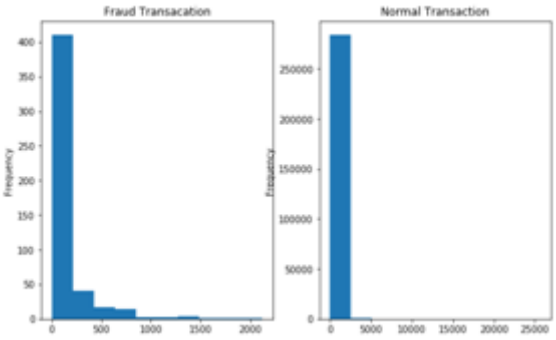

Fig 4: Amount column in Fraud and Non-Fraud Transactions.

Data Pre-processing

Scaling :In this phase ,after examining the dataset, its 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 Dimensionality Reduction Technique to ensure uniformity.

Out[**6**]:

scaled_ amount	scaled_ time	V1	..	V28	Clas s
1.783274	-0.994983	- 1.359807	..	- 0.021053	0
-0.269825	-0.994983	1.191857	..	0.014724	0
4.983721	-0.994972	- 1.358354	..	- 0.059752	0
1.418291	-0.994972	- 0.966272	..	0.061458	0
0.670579	-0.994960	- 1.158233	..	0.215153	0

[5 rows x 31 columns]

Fig 5: Scaled Amount and Time column in credit card dataset.

Feature Co-relation

The credit card dataset is examined and visualized. The feature values are examined .The correlation between Class attribute with other attributes are examined Search method-> Corelation Ranking and Attribute Evaluator-> Correlation Ranking Filter.

=== Attribute Selection on all input data ===

Search Method:
Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 31 Class):
Correlation Ranking Filter
Ranked attributes:
0.154876 12 V11
0.133447 5 V4
0.091289 3 V2
0.040413 22 V21
0.034783 20 V19
0.02009 21 V20

0.019875 9 V8
0.01758 28 V27
0.009536 29 V28
0.005632 30 Amount
0.004455 27 V26
0.003308 26 V25
0.000805 23 V22
-0.002685 24 V23
-0.004223 16 V15
-0.00457 14 V13
-0.007221 25 V24
-0.012323 1 Time
-0.043643 7 V6
-0.094974 6 V5
-0.097733 10 V9
-0.101347 2 V1

-0.111485 19 V18
-0.187257 8 V7
-0.192961 4 V3
-0.196539 17 V16
-0.216883 11 V10
-0.260593 13 V12
-0.302544 15 V14
-0.326481 18 V17

Fig 6:Corelation between Class attribute and other attributes

The attributes like V11,V4,V2,V21 have positive correlation, i.e have higher values when Class=1. The attributes like V17,V14,V12,V10 have negative correlation i.e have lower values when Class=1.

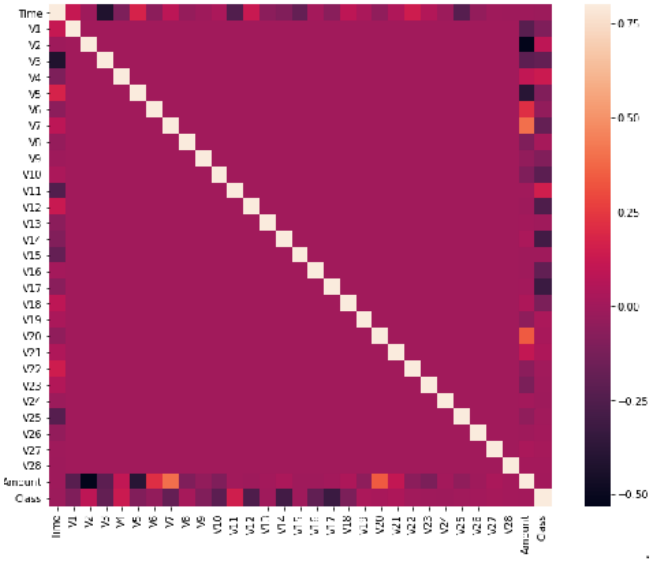

Fig 7 Heat map to visualize correlation matrix

The feature V14 was randomly selected for removing outliers as it was having extremely lower values when there was a fraud transaction.[14]The interquartile range rule is useful in detecting the presence of outliers. Outliers are individual values that fall outside of the overall pattern of the rest of the data. The interquartile range can be used to help detect outliers. The steps in calculating interquartile range are

1. Calculate the interquartile range for our data
2. Multiply the interquartile range (IQR) by the number 1.5
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Quartile 25: -9.69272296475 | Quartile 75: -4.2828208495

iqr: 5.40990211525
Cut Off: 8.114853172875002
V14 Lower: -17.807576137625002
V14 Upper: 3.8320323233750013
Feature V14 Outliers for Fraud Cases: 4
V14 outliers:[-19.21432549, -18.82208674, -18.49377336, -18.04999769]

Fig 8 V14 attribute outliers.

The feature attribute V14 having values greater than V14 Upper and less than V14 Lower are dropped and hence outliers for V14 are removed.

Creating Sub-sample from Dataset
Split the data into train and test sets, for anomaly detection algorithms, it is preferable not to include the outliers; here the outliers are the fraudulent transactions. The major issue with the given data is that it is highly skewed, if the entire data is used to train the model, it would predict majority of fraudulent transactions as genuine. The best method to avoid this issue is to select equal percentage of genuine and fraud sample transactions

As the dataset is imbalanced[18][20] ,the classifiers produce the result in the favor of majority class ,in this case non-fraud transactions. This results in inaccurate results as many a fraud transactions would be labeled inaccurately as non-fraud transactions.
The different ways of sampling data are
Under sampling: The samples of majority class are under sampled. The disadvantage of under sampling is there may be information loss due to under sampling of data. The sub sample of 50-50 ratio of fraud and non fraud transactions are selected.
Oversampling: The samples or data points are created of minority class .SMOTE(Synthetic Minority Oversampling Technique) SMOTE is an oversampling method.

The oversampling technique gave better fraud prediction results as compared to random under sampling technique.

A subsample of equal ratio of fraud and non-fraud transactions are selected using over sampling for a balanced dataset. There are 492 fraud transactions. Hence randomly select 492 non-fraud transactions. The fraud and non fraud transactions are concatenated to form a subsample.
Distribution of the Classes in the subsample dataset
1 0.5
0 0.5
Class =0 and Class=1 transactions are of equal percentage

Fig 8: Sub sample of equally distributed classes (0:No Fraud,1:Fraud)

Metrics Used

The output of the metrices depend on the results obtained by True positive(TP),True Negative(TN),false positive(FP),false negative(FN). The transaction cases which are not fraud and the system model has predicted as not fraud as True Positive(TP). The transaction cases which are fraud and the system model has predicted as fraud as True Negative(TN). The transaction cases which are fraud and the system model has predicted as not fraud as False Positive(FP). The transaction cases which are not fraud and the system model has predicted as fraud as True Neagtive(TN).

The other metrices used are Precision[18] which is defined as In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).The ROC Curve is also known as relative characteristic curve. It is plotted with true positive rate (TPR)against false positive rate(FPR).The True positive rate is calculated as the true positive cases system has identified divided by actual positive cases. The False positive rate is calculated as False positive cases system has identified divided by negative actual cases .

Metrics used	Formula
Precision	TP/(TP+FP)
F1 Score	2*(Precision*Recall)/(Precision + Recall)
Recall	TP/(TP+FN)

Identify fraudulent credit card transactions. Given the class imbalance ratio, it is recommended measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC).[15] It also explains using Recall as a metric in large skewed dataset.

Applying Algorithim

Logistic Regression and Random Forest classifier is applied. Logistic Regression classifier was selected as it is used for predictive analysis. [19] Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Random Forest is an ensemble of random decision trees. Random Forest achieved better results than decision trees. Its default hyperparameter settings produces good prediction results. The main limitation of random forest is due to multiple trees in forest the algorithm is slow.

Performance Tuning of Classifiers:

Logistic Regression:

Using GridSearchCV to search for optimum parameters ,given the list of parameters and Logistic Regression optimum parameters:penalty:l2 C=0.1.Precision of training data was 86 percentage and precision of test data achieved was 91 percentage. Recall metric is 0.53 percentage.

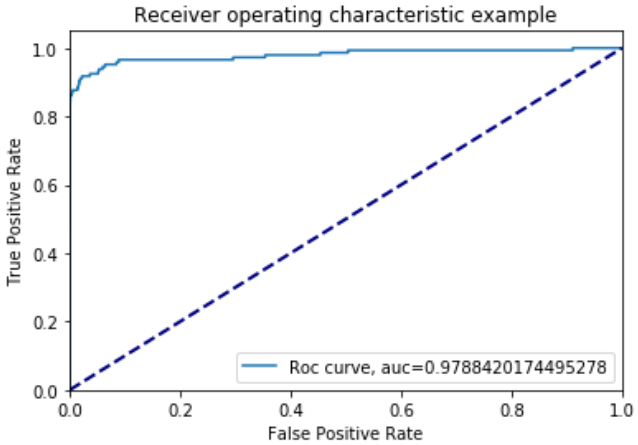

Fig 9. Results of Roc Auc curve in Logistic Regression

Random Forest: The following parameters of Random Forest using Grid search CV bootstrap are max_depth , max_features,min_samples_leaf, min_samples_split, and n_estimators were fine tuned.

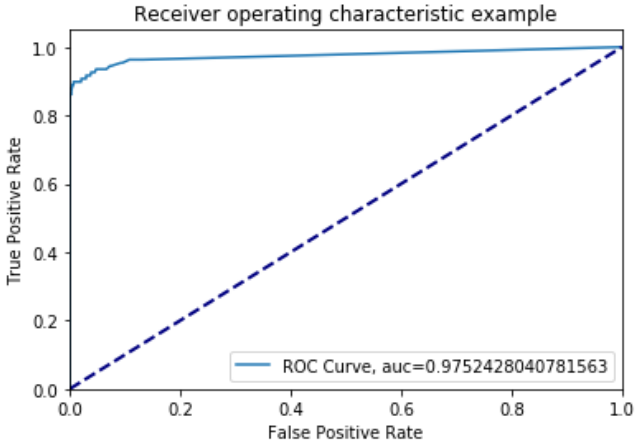

Fig 10. Results of Roc-Auc curve in Random Forest

4. TESTING

The algorithms were tested with the below parameters and the results are given below:

Metrices	Logistic Regression	Random Forest
Precision	0.88	0.93
F1 Score	0.72	0.85
Recall	0.61	0.78
Roc-Auc curve	0.978	0.975

5. FUTURE ENHANCEMENT

One-class SVM is a classifier used for anomaly detection.[22],[23]Neural networks and using combination of Hidden Markov model or K-Nearest neighbor also may help to achieve better anomaly detection in fraud detection .

6. CONCLUSION

The imbalanced datasets for applications like fraud detection, medical applications for cardiac surgery is used.[23] There are still aspects of this approach that are subjected to change. The ratio of fraud and non -fraud transactions in sub sample are 50-50.The other ratios like 80-20,40-60 of fraud and non- fraud transaction subsample need to be analyzed and examined. The oversampling, under sampling of data for accuracy of classifiers is promising. The feature analysis and examining the positive, negative correlation of the features are also important for the accuracy of various classification algorithms. The future work includes analyzing fraud detection using Neural Network and also analyzing the features and subsample ratios for imbalanced datasets.

REFERENCES

[1] Maruti Tech Labs:How Machine Learning facilitates fraud detection https://ww4aw.marutitech.com/machine-learning-fraud-detection/

[2] Diwakar Tripathia,_, Bhawana Nigamb, Damodar Reddy Edlaa (2017)“A Novel Web Fraud Detection Technique using Association Rule Mining “ *,7th International Conference on Advances in Computing & Communications, ICACC-2017.*

[3] Zohreh Darbandian, Alimohammad Latif , Sima Emadi “The discovery of the credit card transactions suspicious of fraud using unsupervised data-mining methods (single-link hierarchical clustering)”(2016) *International Journal Of Humanities And Cultural Studies ISSN 2356-5926*

[4] Dahee Choi and Kyungho Lee,”An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation”(2018)

[5] Jinlong Hu, Junjie Liang, and Shoubin Dong,"iBGP: A Bipartite Graph Propagation Approach for Mobile Advertising Fraud Detection"(2017)
[6] Zhaohui Zhang, Xinxin Zhou,Xiaobo Zhang,Lizhi Wang,and Pengwei Wang,"A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection"(2018)
[7] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit Card Fraud Detection Using Bayesian and Neural Networks," in Proceedings of the 1st International Naiso Congress on Neuro Fuzzy Technologies, pp. 261–270, 2002.
[8] Ping Yi,,Yuxiang Guan,Futai Zou,Yao Yao,Wei Wang, and Ting Zhu "Web Phishing Detection Using a Deep Learning Framework"(2018)
[9] A. Correa Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134–142, 2016. View at Publisher · View at Google Scholar · View at Scopus
[10] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," in Proceedings of the 27th Hawaii International Conference
[11] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural-network," in Proceedings of the 27th Hawaii International Conference on System Sciences, vol. 3, pp. 621–630, Wailea, Hawaii, USA, 1994..
[12] R. Patidar and L. Sharma, "Credit Card Fraud Detection Using Neural Network," in International Journal of Soft Computing and gineering (IJSCE), vol. 1, pp. 32–38, Citeseer Press, 2011. View at Google Scholar
[13] https://www.kaggle.com/mlg-ulb/creditcardfraud/home.
[14] https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244
[15] http://www.chioka.in/differences-between-roc-auc-and-pr-auc/.
[16] https://www.newgenapps.com/blog/precision-vs-recall-accuracy-paradox-machine-learning
[17] JesseDavis,MarkGoadrich, http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf" The Relationship Between Precision-Recall and ROC Curves https:// data-sets using SMOTE and rough sets theory
[18] towardsdatascience.com/accuracy-precision-recall-or-f1
[19] https://www.statisticssolutions.com/what-is-logistic-regression/
[20] https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets
[21] J. Holton Wilson , "An Analytical Approach To Detecting Insurance Fraud Using Logistic Regression "Journal of Finance and Accountancy
[22] Y. Sahin ; E. Duman "Detecting credit card fraud by ANN and logistic regression" 2011 International Symposium on Innovations in Intelligent Systems and Applications
[23] Masoud Khodabakhshi, Mehdi Fartash,"Fraud Detection in Banking Using KNN (KNearest Neighbor) Algorithm",2016.