GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

EARLY PREDICTION OF CREDIT CARD FRAUD DETECTION USING ISOLATION FOREST TREE AND LOCAL OUTLIER FACTOR MACHINE LEARNING ALGORITHMS

A PROJECT REPORT OF CAPSTONE PROJECT – 2

Submitted by

SUBHASH SINGH NEGI

(1613101754)

In partial fulfillment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE ENGINEERING

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Under Supervision Of

MR. ARJUN KUMAR KP

Assistant Professor

APRIL / MAY - 2020

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **"EARLY PREDICTION OF CREDIT CARD FRAUD DETECTION USING ISOLATION FOREST TREE AND LOCAL OUTLIER FACTOR MACHINE LEARNING ALGORITHMS"** is the bonafide work "**SUBHASH SINGH NEGI"** who carried out the project work under my supervision.

SIGNATURE OF HEAD

Dr. MUNISH SHABARWAL,

PhD (Management), PhD (CS)

Professor & Dean,

School of Computing Science & Engineering

SIGNATURE OF SUPERVISIOR

MR. ARJUN KUMAR KP

Assistant Professor

School of Computer Science & Engineering

TABLE OF CONTENTS

Abstract

A credit card is one vital thing nowadays for everyone, even in developing countries. We are using a credit card to pay bills, shop both online as well as offline. With the increase in the use of credit cards, fraud with the credit card also increasing side by side. Credit card fraud is one of the major crimes now as days. In this study, we proposed two machine learning models to predict the fraud transaction outbreak in a credit card accurately. Machine learning models can effectively help bankers and customers accomplish this objective because of their quick and accurate recognition efficiency. Our proposed work, we used Isolation Forest (IF) tree and Local Outlier Factor (LOF) algorithms, which used for anomaly detection for detecting fraud transactions.Isolation Forest (IF) tree algorithm randomly selecting credit card features and make a decision tree from a given dataset and finally score calculated as path length of tree to isolate outlier.Local Outlier Factor (LOF) algorithm calculates outliers by computing the local density of given data concerning its neighbors. Those two models trained and tested with the dataset contain 4092 entries of customer's credit card details made by European cardholders. The data sample is consisting of 80% of fraudulent transactions and the remaining is authenticated transaction done by the customer. We compared our two models with all the existing models that used to identify the fraud transactions, and the prediction accuracy reaches 99%.

LIST OF TABLE

LIST OF FIGURES

INTRODUCTION

Credit card is one of vital thing nowadays for everyone, even in all the developing countries. We can use credit card to pay bills, shop both online as well as offline. With increase in use of credit card, fraud with credit card also increasing side by side. Every year millions of dollars loses caused by credit card fraud [1]. Credit card fraud are define as someone using else credit card for their own use without owner doesn't knowing about it that his card in used. So it's become very necessary that credit card companies are able to correctly identify fraud transaction very efficiently. Necessary major decision should take by credit card companies to avoid and prevent credit card fraud. Credit card companies also make their system more secure so that information are not get leak from their side.

Credit card fraud is one of the major crimes now as days. Credit card fraud can happen through online platforms as well as offline. According to Reserve Bank of India total 972 cases are reported in 2017-18. So, it's become very important that banks are correctly able to identify fraud transactions. Around 25 billion dollar lost in credit card payments all around world in 2018. According to shift credit card processing website United States of America is the global leader in credit card fraud prone country

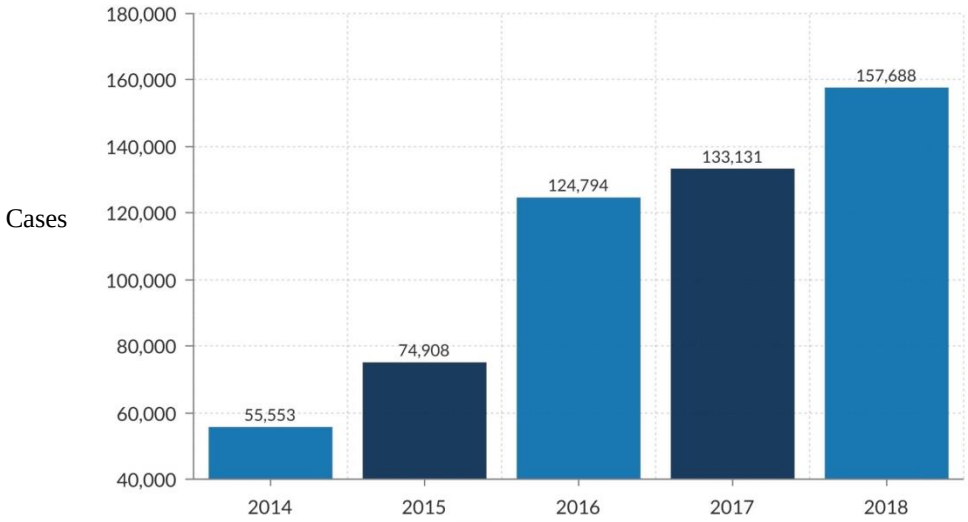

Figure 1 Credit Card Fraud Reported in US

Identity is stolen through via scams, theft and information by scammer and thieves can target and acquire your personal information [2]. Identity theft is third biggest cause of financial fraud. Identity theft occurs when someone uses information such as name, address, birthday, bank statements, etc. to apply for new card or to access your credit card account. In figure 2 represented number of different types of identity thefts and we can clearly understand credit card fraud are the most reported identity theft compared to other type of identity thefts.

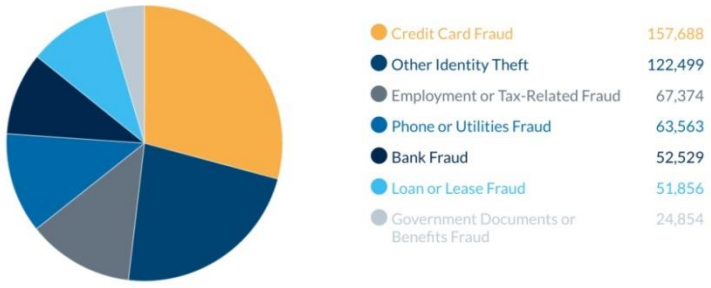

Figure 2 Identity theft reports by SHIFT on 2018

Increase in exposed data breaches are by 54% in 2019. According to shift credit card processing website eight of 3800 data breaches of 2019 exposed more than 3,2 million records i.e., nearly 80% of all record exposed so far in 2019.Figure 3 shows the different sector data breaches are represented in pie graph. The top companies are also data breaches that exposed consumer records like Yahoo, Facebook, Marriott, etc [3].

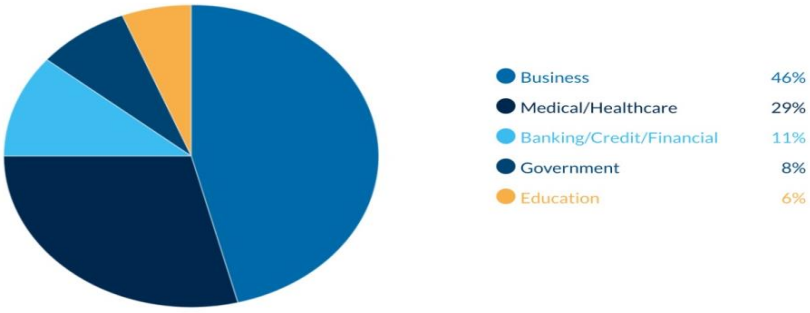

Figure 3 Data Breaches Reports in different Areas

.

Credit card fraud can be prevents through various easy way. Various methods through which we can take precaution for prevent credit fraud are monitoring your credit card account statement frequently. Keep your wallet secure where you kept your cards and close to you all time.When doing online payment make sure doing through secure website and kept your card details not save in public computer [4]. During point of sale payment method make sure no one see your pin while entering. If you are a victim of credit card fraud than make sure report of that to your credit card company and if you misplaced your card or lost contact your bank and closed that card. Final one is changing your card pin regularly over 2 – 3 months to prevent fraud.

Machine Learning is sub- set of Artificial Intelligence (AI). Machine Learning is field of study concerned with the design and development of algorithms and techniques that allows machine learn by it. Machine learning works as similar ways as human learning. In Machine learning, statistical and mathematical methods are used for learning through datasets. Machine learning can helps in many areas where problems are very complex n difficult such as in medical field, banking sector, etc [5]. Machine learning are applied to our every daily bases such as in Netflix, Amazon as well as Facebook, etc.

Machine learning also play an important role in providing cyber security. Machine learning easily safeguard against malware, assessing network security, developing secure online transactions as well as online interaction and developing authentication systems. Credit card fraud can be happen in numerous ways. In case of offline credit card fraud, fraudster required credit card physically while, in case of online credit card fraud, fraudster required only details of credit card user. The major challenges fraud detection are [6-9], first one huge large amount of data processed every day and model built that must fast enough to detect fraud transaction in time. Second is imbalanced dataset i.e., most of the transaction are valid transaction and only few about 0.1% of transactions are fraud transaction. Final the banking companies are not sharing their fraudulent transaction details because of company's credentials.

LITERATURE REVIEW

Fraud detection for credit card can be done through various techniques such as machine learning techniques, neural networking and data mining. Most commonly ways of detecting fraud in credit card through machine learning is using supervised learning i.e., Decision tree, Random Forest, KNN, etc. there are many research focus on use on data mining for data processing and data analysis. Research show development of fraud detection model along with machine learning increases the prediction result.

Netty Setiawan, Suharjito and Diana [10] they have used the BPSOSVM-ERT and BPSOSVM-RF which are compared on the several performance metrics. The dataset is used form the LeadingCub which provides the loan dataset and the data issued from 2007 to 2017. The BPSOSVMERT model which produced the accuracy as 64%.SurajPatil, VarshaNemade and Piyush Kumar Soni [11] they have used the German credit card fraud dataset and the models are used in this approach are logistic regression, decision tree and random forest decision tree with accuracy as 72%, 72% and 76% respectively.

LutaoZheng, Guanjun Liu, Chungang Yan and Changjun Jiang [12] they have used the Markov chain models for the transaction fraud detection based on total order relation and behavior diversity. In this they have also used the behavior profiles (BPs) helps in finding the fraud. They have used the dataset from Kaggle and they got accuracy score about 0.912.SurayaNurainKalid, Keng-Hoong Ng, Gee-Kok Tong and Kok-Chin Khor [13] they have used the Multiple Classifiers System for the anomaly detection in credit card in which they have used the two datasets as credit card fraud dataset in which they have overlapping class samples and unbalanced class distribution, and the another dataset is credit card default payments. In the multiple classifier system, they have used the two models as Naïve Bayes (NB) and C4.5, with

this model they got the accuracy score different for both the datasets as for credit card fraud as 0.99 and for credit card default detection as 0.93.

Sara Makki, ZainabAssaghir, YehiaTaher, RafiqulHaque, Mohand-SaidHacid and Hassan Zeineddine [14] they have used the imbalanced classification approaches for the credit card fraud detection. Their models used in the approach are LR, C5.0 decision tree algorithm, SVM and ANN which performed well and got the same accuracy score as 96% for all the models. They have used the credit card fraud labelled dataset.

AltyebAltaherTaha and SharafJameelMalebary [15] they have used the optimized d light gradient boosting machine (OLightGBM). In their approach they have used a Bayesian-based hyperparameter optimization algorithm is used to tune the parameters of a light gradient boosting machine (LightGBM). They have used two different datasets; first dataset consists credit card transaction of an owner in Europe and the second dataset is from UCSD-FICO Data mining contest in which it has the real dataset of e-commerce transcations. They got the accuracy 98.04%, precision 97.34% and f-1 score 56.95%.

FayazItoo, Meenakshi and Satwinder Singh [16] they have used logistic regression, Naïve Bayes and KNN models for the credit card fraud detection. They have used the dataset from the kaggel which provides the dataset for the credit card fraud detection. They got the best accuracy for the logistic regression model with accuracy 0.959.

PROPOSED MODEL

Dataset Description

Kaggle provided the dataset which contains transaction made by European credit card-holders. The dataset contains transactions that occur in two days, where it has 492 Fraud transaction and 284315 valid transactions. Dataset has features V1, V2, V3 ... V28 that are transform into PCA values due to confidentiality issues. It also contains three more features that are not PCA transform, amount, time and class. Class represents transaction is fraud or valid. If class is 1 then transaction is Fraud while 1 when transaction is valid. Table 1 shows the overall details of credit card fraud problem's dataset provided by Kaggle.

Table 1 Dataset Explanation

Variables	Explanation	Data Type	Scale
V1 - V28	Confidential data of credit card holders which numerical input variable that is PCA transform	Numerical	Numerical
Amount	Numerical input feature that represent amount debit or credit from credit card	Numerical	Money in euros
Time	Time represent second elapsed between each transaction and first transaction in the dataset	Numerical	Time in seconds
Class	Class represent fraud and valid transaction i.e., 0 for normal transaction and 1 for fraud transaction	Numerical	(0,1)

Model Performance Evaluation Matrix

Table 2 shows the performance evaluation matrix that we used to show the performance of our model compared to other existing models.

Table 2 Performance Evaluation Matrix

Performance Matrix Name	**Explanation**
True Positive (TP)	Credit card fraud cases that model predicated as "fraud". $TP = Count\ of\ positive\ \gamma \rightarrow positive\ \hat{\gamma}$
False Positive (FP)	Credit card non- fraud cases that model predicated as "fraud". $FP = Count\ of\ negative\ \gamma \rightarrow positive\ \hat{\gamma}$
False Negative (FN)	Credit card fraud cases that model predicated as "non- fraud". $FN = Count\ of\ positive\ \gamma \rightarrow negative\ \hat{\gamma}$
True Negative (TN)	Credit card non- fraud cases that model predicated as "non –fraud". $FP = Count\ of\ negative\ \gamma \rightarrow negative\ \hat{\gamma}$

Accuracy	Model accuracy which means our model must report positive cases are predicted as positive and negative cases are predicted as negative. $$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$
Precision	Precision is the propositions of predicted true positive cases are relevant among all true positive and false positive cases. $$Precision = \frac{TP}{TP + FP}$$
Recall	Precision is the propositions of predicted relevant true positive cases are among all true positive and false negative cases. $$Recall = \frac{TP}{TP + FN}$$
F1-Score	F1 score describes about balance between the precision and recall. $$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Pre-processing

In figure 4 represents heatmap which means one variable that could be gently connected with another variable. It will be giving more effective outputs for investigations and displays more readily between factors. We can clearly understand the features V1 to V28 are not connected each other so these features wouldn't produce a good model. In feature selection process we skip or remove these features to get a good prediction model.

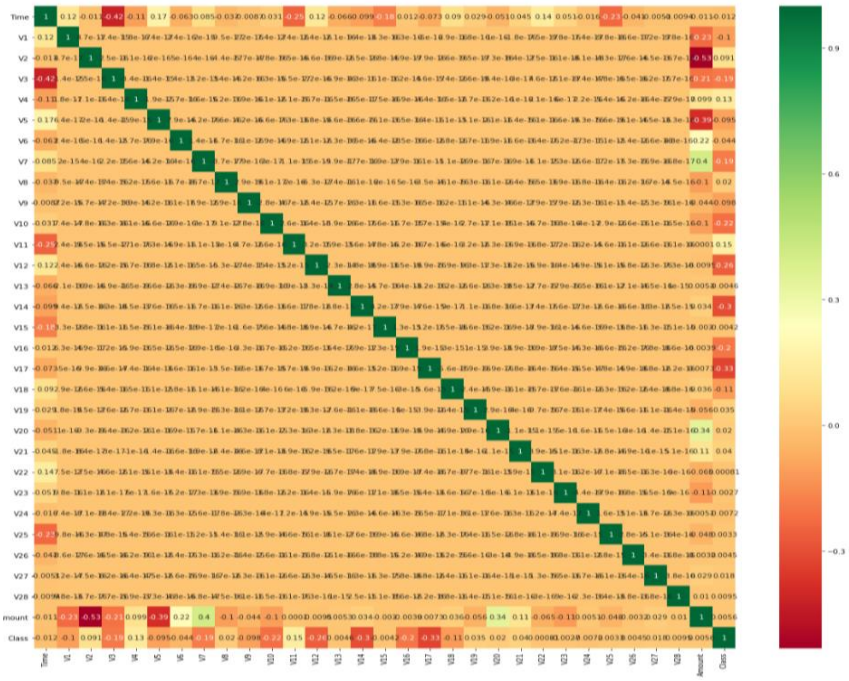

Figure 4 Heat map for correlated features values in the dataset.

Model Description

Here we focus on the fraud transaction recognition using the Isolation Forest Tree (IF) and Local Outlier Factor (LOF)model. Kaggle provides Kaggle notebook, a cloud-based machine learning

platform that gives the advantages of reproducible and collaborative analysis.Kaggle provideddataset feed the ML models was trainedon the two models.

A) Isolation Forest Tree Model

Isolation Forest Tree (IF) machine learning algorithm is an anomaly detection method. IF algorithm can work either supervised neither unsupervised learning method. Isolation forest tree algorithm different from other type of distance or density based method for outlier detection and algorithm tried tree to build extremely randomized decision tree for separating outlier. The Equation for calculating outlier in Isolation forest trees as following:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

- h(x) is path length of observation x
- c(n) is average length of unsuccessful search in Binary Search Tree
- n is number of external nodes

Parameters in IF are as followings:

- n_estimators :int , optional(default=100) The number of base estimators in the ensemble.
- max_sample :int or float, optional (default = “auto”) The number of samples to draw from X train each base estimator.
- Contamination : ‘auto’ or float, optional (default = ‘auto’) The amount of contamination of the dataset. Used when fitting to define the threshold on the score of the samples.
- max_feature :int or float, optional (default = 1.0) The number of feature to draw from X to train each base estimator.

- bootstrap : bool, optional (default = False) If True, then individual trees are fit subsets of training data sampled with replacement. If False, then sampling without replacement is performed.
- n_jobs :int or None, optional (default = None) The number of jobs to run in parallel for both fit and predict. None means 1 unless in a joblib.parallel_backend context. -1 means using all processors.
- random_statee :int, RandomState instance or None, optional (default=None) If int, random_state is the seed used by the random number generator. If RandomState instance, random number generator. If None, the random number generator is the RandomState instance used by np.random.
- verbose :int, optional (default = 0) Control the verbosity of tree building process.

B) Local Outlier Factor for outlier detection

The Local Outlier Factor algorithm is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbors. It is considered as outlier samples which has substantially lower density than their neighbors.

- **n_neighbors :int, optional(default = 20)** Number of neighbors to used by default for **kneighbors**queries. If n_neighbors is larger than the number of samples provided, all samples will be used.
- **algorithm : {'auto', 'ball_tree', 'kd_tree', 'brute'), optional** Algorithm used to compute the nearest neighbors
- **leaf_size :int, optional (default = 30)** Leaf Sizee passed to BallTree or KDTree. This can affect the speed of the construction and query, as well as the memory required to store the tree,

- **metric : string or callable, default 'minkowski'** metric used for the distance computation. Any metric from scikit-learn or scipy.spatial.distance can be used.
- **metric_params :dict, optional (default=None)** Additional keyword argument for the metric function.

contamination : 'auto' or float, optional (default = 'auto') The amount of contamination of the dataset. When fitting is used to define the threshold on the scores of the samples.

RESULT

In result and discussion section, we analyzed our proposed model performance and we also compared our proposed model with existing models showed in table 3.

Table 3 Proposed Model Evaluations

Models	Accuracy	Precision	F1-Score	Recall
Naïve Bayes	75.8	90.5	84.5	84.5
Random Forest	86.10	87.10	92.40	92.40
Support Vector Machine(SVM)	70.09	75.00	61.50	61.50
Local Outlier Factor	99.65	75.50	75.50	75.50
Isolation Forest Tree	99.74	81.50	81.50	81.50

Isolation Forest tree algorithm detected 73 errors and its accuracy is about 99.74% which is greater than local outlier factor algorithm. Local Outline Factor (LOF) detected 97 error and it's accuracy is around 99.65%. Using 10% of dataset for faster execution of learning and training model for predicting results.Figure 5 shows the graphical representation of comparison result, from that graph we can understand our two proposed model is far better than all the existing methods.

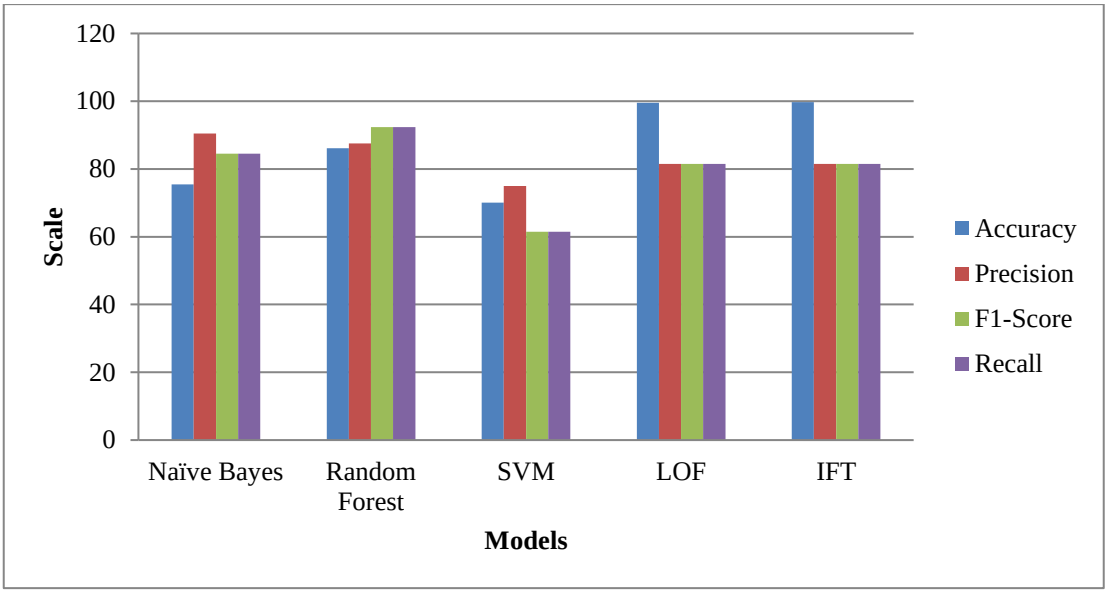

Figure 5 Graphical representation of Proposed Models Comparison

When we compare our model with existing model we can easily say that our models prediction accuracy is far better than all the existing methods. When comparing the error precision and recall for Isolation Forest (IF) and Local Outlier Factor (LOF), the other models like Navie Bayes and Random Forest performed good because its dataset distribution of fraud and non-fraud transaction almost equal. Overall performance of Isolation Forest (IF) is much better that Local Outlier Factor (LOF). Performance of model can be improved by using Deep learning or Neural Network however at the cost of computational expensive.

CONCLUSION

Credit card fraud detection involves the deep analysis of large group of credit card users in order to identify properly. According to federal law and issuer card network terms and policies,credit card owner don’t have to pay cost of unauthorized purchases made with his cards. Financial institutions and merchants assumed responsible for the most of the money spent as product of fraud.Credit card fraud is happen when someone other than owner of credit card uses credit card or credit account to make transactions. In our proposed, we implemented two machine learning algorithms are Isolation Forest (IF) tree and Local Outline Factor (LOF).Isolation Forest tree algorithm accuracy is about 99.74% which is greater than local outlier factor algorithm. Local Outline Factor is around 99.65%. In both of these cases results are approximately same. When comparing the error precision and recall for Isolation Forest and Local Outlier Factor, the Isolation Forest Tree algorithm preformed much better than the Local Outlier Factor . Overall performance of Isolation Forest (IF) is much better that Local Outlier Factor (LOF). Performance of model can be improved by using Deep learning or Neural Network however at the cost of computational expensive.

REFERENCES

[1] VenkatKrishnapur, “Thieves only need your credit card data, not your card to defraud you”, Economic Times

[2] Barry Wong,“The 8 Different Types of Card Fraud”, Mastercard

[3] Chanellebessette,“How Serious a Crime Is Credit Card Theft and Fraud?”, Nerdwallet.

[4] Setiawan, N., Suharjito, & Diana. (2019). *A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Machine Learning.Procedia Computer Science, 157, 38–45.* doi:10.1016/j.procs.2019.08.139

[5] Zheng, L., Liu, G., Yan, C., & Jiang, C. (2018). *Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity. IEEE Transactions on Computational Social Systems, 1–11.* doi:10.1109/tcss.2018.2856910

[6] Kalid, S. N., Ng, K.-H., Tong, G.-K., &Khor, K.-C. (2020). *A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes. IEEE Access, 8, 28210–28221.* doi:10.1109/access.2020.2972009

[7] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., &Anderla, A. (2019). *Credit Card Fraud Detection - Machine Learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH).* doi:10.1109/infoteh.2019.8717766

[8] Awoyemi, J. O., Adetunmbi, A. O., &Oluwadare, S. A. (2017). *Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI).* doi:10.1109/iccni.2017.8123782

[9] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., &Kuruwitaarachchi, N. (2019). *Real-time Credit Card Fraud Detection Using Machine Learning.2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence).* doi:10.1109/confluence.2019.8776942

[10] Setiawan, N., Suharjito, & Diana. (2019). *A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Machine Learning.Procedia Computer Science, 157, 38–45.* doi:10.1016/j.procs.2019.08.139

[11] Patil, S., Nemade, V., &Soni, P. K. (2018). Predictive Modelling For Credit Card Fraud Detection Using Data Analytics.Procedia Computer Science, 132, 385–395. doi:10.1016/j.procs.2018.05.199

[12] Zheng, L., Liu, G., Yan, C., & Jiang, C. (2018). *Transaction Fraud Detection Based on Total Order Relation and Behavior Diversity. IEEE Transactions on Computational Social Systems, 1–11.* doi:10.1109/tcss.2018.2856910

[13] Kalid, S. N., Ng, K.-H., Tong, G.-K., &Khor, K.-C. (2020). *A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes. IEEE Access, 8, 28210–28221.* doi:10.1109/access.2020.2972009

[14] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.-S., &Zeineddine, H. (2019). *An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access, 1–1.* doi:10.1109/access.2019.2927266

[15] Altyeb, A. A., &Malebary, S. J. (2020). *An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine. IEEE Access, 1–1.* doi:10.1109/access.2020.2971354

[16] Itoo, F., Meenakshi, & Singh, S. (2020). *Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology*. doi:10.1007/s41870-020-00430-y