Turkish Journal of Computer and Mathematics Education *Vol.12 No. 11 (2021), 3189-3195*

Research Article

A Novel Framework for Credit Card Fraud Detection

Heena Kochhar[a], Dr. Yogesh Chhabra[b]

[a,b] Department of Computer Science & Engineering, CT University Ludhiana, Punjab, India

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Credit card fraud is the challenge of predicting fraudulent transactions based on specific rules. In this paper, Various classification algorithms are implemented on an imbalanced dataset concerning the performance analysis to detect fraud in the credit card. In this study, the dataset is sourced from Kaggle. There are 284,807 transactions, out of which 17% of transactions are fraudulent. Various classifiers that are logistic regression, naive Bayes, AdaBoost, and voting classifiers that are combinations of all mentioned above algorithms are refined. The AI model needs significant historical data to prepare the model. For this, a huge amount of information is given to the model as Training data, While dealing with a lot of information, the model's execution time is expanded, which influences execution. In this study, the voting classifier is applied for the expectation analysis which is a mix of different AI calculations.This voting based classifier increases the complexity of the prediction analysis and also increases execution time. In the future, a hybrid classification model will be designed to detect credit card scams.

Keywords: Naive Bayes, Credit card, Logistic regression, random forest, majority voting

1. Introduction

Credit cards are commonly utilized to purchase several goods and access several services in our daily lives.[1] Credit card fraud is defined as a fraudulent transaction fake exchange by an unapproved individual for his advantage The approved cardholder and the card supplier are uninformed of the exchange right now of its realization[23]. From providing a successful payment method, the cards are provided by the cardholder to the merchant during the purchasing method's execution based on a physical card. By stealing a credit card, a fraudulent attack has been conducted through the attacker. The credit card company can also face a massive loss if the cardholder is not aware of losing it. For performing any fraudulent In online transactions, the attacker requires fewer data. The internet and telephones are used for purchasing products and services online. For reducing the rate of successful credit card fraud, it is essential to introduce fraud detection methods. This method is proposed based on the purchase information of the particular cardholder. It is possible to know that the card is stolen with recognizing patterns [1].

fraud has been expanding definitely with the movement progression of state-of-art technology and worldwide communication.[5] Fraud can be checked two ways: prevention and detection. Anticipation of information is the place where a layer of assurance is shaped to evade any assaults from untouchables. It attempts to prevent misrepresentation from happening in any case. Conversely, extortion discovery helps in distinguishing and cautioning when it has been executed. In this way, discovery comes into the scene once the avoidance has effectively fizzled. Thus, location should constantly be running as nobody can anticipate when a penetrate may happen to the security given by extortion counteraction strategies.[9]

2. Machine Learning Algorithms For Fraud Detection

- Logistic Regression (LR) is a kind of generalized linear model. It is not apposite to implement a simple linear regression if the variable is binary, which will be predicted because of normality assumptions. [2].
- Decision Trees (DT) has a tree structure in which test is denoted through every node on a feature, and every branch is employed to reveal a result obtained in testing. Hence, the observations are split into mutually exclusive subgroups by the tree [3]
- Neural Networks (NN) is an experienced innovation that has a set up hypothesis and perceived application regions. Various neurons are formed in these organizations. The weight is a mathematical worth that is identified with each association[3].

- Support Vector Machines (SVM) has utilized a straight model for applying nonlinear class limits. Information vectors are planned in a nonlinear manner into a high-dimensional component space. The improvement of an ideal isolating hyperplane is completed inside a novel space[4].
- Bayesian belief network (BBN) facilitates a demonstration of the dependencies among subsets of features. Every node is used to show an attribute in this graph, and probabilistic dependence is demonstrated by every arrow [5].
- K-nearest neighbor (KNN) is carried in the systems which are executed for detection. The KNN is proved efficient in CCFD systems with the utilization of supervised learning schemes. [6].
- Hidden Markov Model (HMM) is different from the standard statistical Markov model as it contains invisible states; however, a visible form is produced through every state at random. A hidden Markov model is represented as the most straightforward dynamic Bayesian network [7].
- Artificial neural networks (ANN) are initially constructed to impersonate the nature of the human brain. A NN is the relationship of rudimentary items perceived as the straightforward neuron[8].

3. Literature Review

KuldeepRandhawa et al. [15] suggested a method in which ML was utilized to detect credit card fraud. At first, the execution of standard models was done, which was followed by hybrid models. The AdaBoost and majority voting techniques were carried out in these models. The model's efficacy was evaluated using a publicly available dataset, and the other dataset was carried out from the financial institution to analyze fraud. For quantifying the robustness of the algorithms, the insertion of noise was done to the data sample. The outcomes obtained after the experiments revealed that the higher accuracy rates were provided through various voting methods to detect the CCF. There was about10% and 30% noise added in sample data so as the hybrid models were evaluated. When 30% noise was added, a good score of 0.94 was obtained from various voting techniques. Therefore, it demonstrated that much stable performance had been acquired using the voting technique when the noise was available.

Abhimanyu Roy et al. [16] suggested DL topologies detect fraud from money's online transaction. The ANN was employed to carry out this approach and the in-built time and memory elements in long-term and short-term memory and various other parameters. There were 80 million online exchanges utilizing Mastercards and pre-marked as false and legitimate contingent on the viability of these components, which had been used while identifying extortion.The distributed cloud computing environment of superior performance has been implemented for these elements. According to their account, the researchers suggested a study in which an efficient guide to suggested parameters' sensitivity analysis for detecting the fraud had been offered. A framework was also proposed by researchers for the parameter alteration of DL topologies to detect fraud. Thus, the financial institution for reducing the losses was facilitated in it when fraudulent activities were avoided.

ShiyangXuan et al. [17] presented two types of Random Forest classifiers that were utilized for training the behavior attributes included transactions- normal and abnormal. Comparing these two RFs, which had distinguished their classifiers, performance while detecting fraud in credit cards was carried out. The data was extracted from an e-commerce company, and the analysis of the presented model's efficiency was done. The author employed the B2C dataset to detect and recognize fraud in credit cards. Thus, the outcomes proved that superior results were obtained from suggested RF on a small dataset; however, some issues were present, such as imbalanced data, which became less efficient than any other dataset.

Deshen Wang et al. [18] recommended a novel technique in which any of the two previously existing methods can improve detecting frauds. No prevention and applying ML detection models to perform all transactions are these two strategies. It is seen that the performance of no method is up to the mark when there is a high fraud rate as well as the secondary verification compensation rate. None of these strategies individually provides benefits to the merchants. It is possible to adopt a different approach when there is very little compensation needed so that the incentivized consumers can accept the secondary verification for all transactions. As there is a removal of the advantage through which the secondary guarantee is being applied when the merchant can consider less fraud rate, the two strategies mentioned above. As all the false positive drops are removed when all the transactions are accepted, the first approach can be used. Otherwise, the second approach is preferred.

Johannes Jurgovsky et al. [19] studied that the fraud detection issue can also be called the sequence classification task. The Long Short-Term Memory (LSTM) networks are deployed here. The traditional feature aggregation techniques are integrated, and the conventional retrieval metrics are used for reporting the results

achieved. In the offline transactions in which the cardholder is available physically at the merchant, the accuracy of detection is better in the LSTM approach than the traditional RF classifier. For the manual feature aggregation techniques, the performance of sequential and non-sequential processes is better. Integration of these two techniques is suggested since various frauds are detected by analyzing both approaches' true positives.

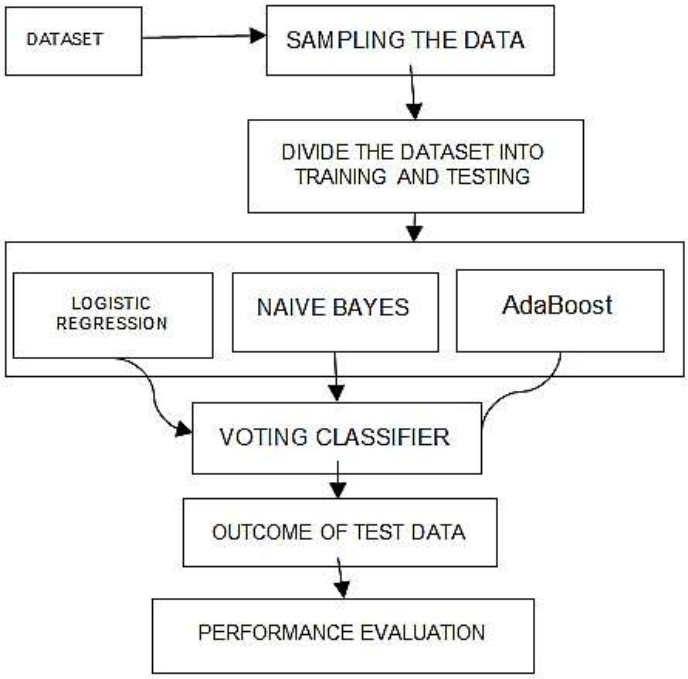
Alex G.C. de Sá. et al. [20] proposed an enhanced BNC algorithm called Fraud- Bayesian Network Classifier. The taxonomy of the knowledge related to BNC algorithms is created, and the most pleasing element combination within the dataset is identified by HHEA, which further generated the Fraud-BNC. PagSeguro is the well-known online payment service of Brazil. This online service was used to create a dataset through which this proposed technique was further generated. To handle the cost-sensitive classification, the proposed approach was tested along with the two strategies. Comparisons were made amongst the results achieved by seven different algorithms and the proposed algorithm. The presented technique had performed superior, and the company's economic efficiency was improved to around 72% as per the current state.

Akila S. et al. [21] proposed Risk Induced Bayesian Inference Bagging (RIBIB) model in which a bag creation technique was utilized to detect frauds. The base learner and cost-sensitive weighted voting integrated approach were used in this technique. Simulations were performed by implementing the proposed method on the bank data of Brazil. It was seen that the cost was minimized to 1.04–1.5 times when the proposed technique was applied. Without using the domain-specific parameter fine-tuning, the model proved highly robust when implemented on UCSD-FICO data.

FabrizioCarcillo et al. [22] proposed a novel technique by integrating the big data tools with machine learning techniques for handling specific complex scenarios. This technique was named Scalable Real-time Fraud Finder (SCARFF). It was initially challenging to maintain the enormous streaming data by the fraud detection techniques. Thus, improvements were suggested so that the excessive data could be handled efficiently. It is seen that on a significant stream of transactions, the scalability, efficiency, and exactness of the presented technique have improved more than the outcomes achieved by previously existing approaches.

4. Methodology

In the process flow (Fig.1), we first import the dataset. In this paper, the problem of class imbalance is addressed by using sampling techniques. In oversampling, we have used the SMOTE method. We used three classifiers- logistic regression, Naive Bayes, and AdaBoost, to classify the data. Then apply a voting classifier.Then, to analyze each of the models' performance, four types of performance metrics: accuracy, precision, f1 score, and recall, are used.[9]



(FIGURE 1)

5. Performance Evaluation

A.DATASET:

The dataset is available in source form Kaggle. The payments of credit cards done through European cardholders in September 2013 are composed in this dataset. The transactions made in 2 days and further include 284,807 transactions have been described in this dataset. The dataset is unbalanced and skewed in the positive class. The input variables are available in only numerical form as these values have resulted from the PCA. Therefore, thirty input attributes have been carried out in it. The details and background information regarding the features are not defined as there are some personal issues. The seconds elapsed amid every transaction, and the first transaction in the dataset is included in the time attribute. The amount attribute is the amount of the payment. The feature class is a destination class employed to perform the binary classification and provides the value 1 in the positive case situation and 0 for the non-fraud case.

B. DATA PRE-PREPROCESSING

The model accuracy depends on the amount of data on which it is trained. The more data, the better will be the performance of the model. In this first step, the data is cleaned and preprocessed as follow:

Cleaning: Fixing missing data or removal of duplicate data from a dataset is called cleaning. The dataset may contain records that may be duplicated, incomplete, or may have null values. Such documents need to be removed by cleaning.[25]

C. RESAMPLING TECHNIQUE

As the number of frauds in the dataset is less than an overall transaction, class distribution is unbalanced in credit card transactions. the sampling method is used to solve the issue. [25].The over-sampling is done on a too unbalanced dataset for obtaining two sets of distribution to perform the analysis. The stepwise expansion and deduction of an information point are inserted among existing information focuses until the over-fitting limit is reached.

D.METRICS

There are four basic metrics to compute the experiments based on TPR, TNR, FPR, and FNR rates metrics.

True Positive(TP): The actual positive rate represents the fraudulent transactions' portion correctly classified as fraudulent transactions.

True positive=Tp/TP+FN (1)

TrueNegative (TN): The actual negative rate represents the regular transactions' portion correctly classified as routine transactions.

True negative=TN/TN+FP (2)

False Positive (FP): The false-positive rate indicates the bit of the non-fake exchanges wrongly delegated as fraudulent transactions.

False positive=FP/FP+TN (3)

False Negative (FN): The false-negative rate demonstrates the part of the non-deceitful exchanges wrongly being named normal exchanges.

False negative=FN/FN+TP (4)

E. CLASSIFICATION

1 Logistic Regression Classifier: Logistic regression is a supervised classification algorithm.In this the objective variable(or yield), y, can take just discrete qualities for a given arrangement of features(or inputs), X. The calculated relapse model outlined the association between indicators that might be persistent, twofold, and absolute. Strategic relapse turns into a grouping procedure where just a call edge is brought into the picture. The edge worth setting is a fundamental feature of strategic relapse and relies upon the arrangement drawback itself. It predicts the opportunity that a given information passage has a place with the class numbered as "1"

2 Naïve Bayes Classifier: Naive Bayes (NB) technique classifiers depend on Bayesian hypothesis that chooses the choice dependent on contingent likelihood [24].

This algorithm performs decision-making according to the maximal probability. Bayesian probability makes use of given values for estimating indefinite probabilities. This algorithm enables past knowledge and logic to be implemented to unclear descriptions. This algorithm assumes the conditional independence of features within the data. This classification model follows the idea of conditional probabilities of the two classes called fraudulent and non-fraudulent.

Here, n corresponds to the maximal number of attributes"" signifies the probability of feature value in the class. Refers to the likelihood that generates the feature value of a known type.) is the probability of class incidence. Also,) depicts the likelihood of the incidence of feature value.

The result is C1 when

The result is C2 when

Ci denotes the target class for classification. Moreover, C1 and C2 represent non-fraudulent and fraudulent cases, respectively [15]

3. AdaBoost: Adaptive Boosting or AdaBoost is utilized related to various calculations to improve execution. The outputs square measure combined by employing a weighted add that represents the combined output of the boosted classifier, which represents the combined output of the boosted classifier, i.e.,

$$F\mathrm{t}(x) = f\mathrm{t}(x) \quad (5)$$

Where each ft is a classifier (weak learner) that profits the anticipated class concerning input x. Each feeble student gives a yield forecast, h(xi), for each preparation test. In each cycle t, the frail student is picked and is designated a coefficient, αt, so the preparation blunder whole, Et, of the subsequent t-stage helped classifier is limited,

$$E\mathrm{t} = \sum [F\mathrm{KO0}(x'') + \alpha \mathrm{K}h(x'')] \quad (6)$$

In whichFt–1(x) is a boosted classifier that is constructed in the preceding phase, E(F) represents the error function, and ft(x) = αth(x) is a weak learner who is considered for the final classifier. The vulnerable learners are squeezed concerning misclassified data samples by AdaBoost. However, it has susceptibility against noise and outliers.

AdaBoost has the potential for enhancing the individual outcomes from diverse algorithm till the classifier performs randomly [14]

4. Majority voting: Majority voting is frequently used in data classification, again utilized in information grouping, which includes a joined model with at any rate two calculations. Every calculation makes its forecast for each test. The last yield is for the one that gets most of the votes, as follows.[15]

Algorithm steps for finding Results

Step1: Import the dataset.

Step2: Data preprocessing.

Step3: Do oversample.

Step4: Give 70% data for training and 30% for testing.

Step5: Assign train dataset into models.

Step6: Apply the algorithm.

Step7: Predict the test dataset for each algorithm.

Step8: Perform various metrics.

6. Results

There are 2,84,807 observations in our test dataset, out of which 17% of transactions are fraudulent. In this study, four classifiers: logistic regression, naive Bayes, AdaBoost, and voting classifier, are refined. The performance evaluation of machine learning models is based on F1 score, precision, recall (sensitivity), and accuracy. Moreover, from the confusion matrix the value of True Positive (tp) ,True Negative(TN), False Positive(FP), False Negative(fn) are concluded. the whole dataset is divided into a 70:30 ratio. 70% of the dataset is used for training, and the other 30% used for testing. The results show accuracy for logistic regression, Naive Bayes, Adaboost, voting classifiers, 94.51%, 91.41%, 95.67%, and 94.69%.

Table 1 represents the result of logistic regression, which give 94.5% of accuracy,

```
Total number of observations: 284807
Percentage of fradulent transactions in data.: 0.1727485630620034
Accuracy on the Logistic regression model is: 94.51259805593452
[[91344  7948]
 [ 2349 86007]]
              precision    recall  f1-score   support

           0       0.97      0.92      0.95     99292
           1       0.92      0.97      0.94     88356

    accuracy                           0.95    187648
   macro avg       0.95      0.95      0.95    187648
weighted avg       0.95      0.95      0.95    187648
```

Performance evaluation of logistic regression machine learning algorithm

Table 2 represents the result of Naive Bayes, which give 91.41% of accuracy,

```
Total number of observations: 284807
Percentage of fradulent transactions in data.: 0.1727485630620034
Accuracy on the naive_bayes is: 91.4144522800415
[[82974 12448]
 [ 2198 72969]]
              precision    recall  f1-score   support

           0       0.97      0.87      0.92     95422
           1       0.85      0.97      0.91     75167

    accuracy                           0.91    170589
   macro avg       0.91      0.92      0.91    170589
weighted avg       0.92      0.91      0.91    170589
```

Performance evaluation of Naive Bayes machine learning algorithm

Table 3 represents the result of Adaboost, which give 95.67% of accuracy,

```
Total number of observations: 284807
Percentage of fradulent transactions in data.: 0.1727485630620034
accuracy on the adaboost is: 95.67674351804631
[[82657  4860]
 [ 2515 80557]]
              precision    recall  f1-score   support

           0       0.97      0.94      0.96     87517
           1       0.94      0.97      0.96     83072

    accuracy                           0.96    170589
   macro avg       0.96      0.96      0.96    170589
weighted avg       0.96      0.96      0.96    170589
```

Performance evaluation of Adaboost machine learning algorithm

Table 4 represents the result of the Voting classifier, which give 94.69% of accuracy,

```
Total number of observations: 284807
Percentage of fradulent transactions in data.: 0.1727485630620034
accuracy on the testing set: 94.69719610498263
[[83693  7567]
 [ 1479 77850]]
              precision    recall  f1-score   support

           0       0.98      0.92      0.95     91260
           1       0.91      0.98      0.95     79329

    accuracy                           0.95    170589
   macro avg       0.95      0.95      0.95    170589
weighted avg       0.95      0.95      0.95    170589
```

Performance evaluation of voting classifier

7. Conclusion

The models designed for credit card fraud detection perform well on the little datasets, however the exactness gets decreased when the dataset size increments.The models intended for credit card fraud detection perform well in terms of some parameters but do not perform well in terms of all. The credit card fraud detection model needs large quantities of historical information to train models. While handling such a large amount of data, the model's execution time is increased, affecting performance. In this study, the voting classifier is applied for the prediction analysis. This increases the prediction analysis complexity and increases execution time; the efficient prediction analysis technique must establish a relationship between attribute and target set. So far, the proposed plans are not so efficient to drive relationships between target sets and attributes of the dataset. In the future, a hybrid classification method will be designed to detect credit card scams.

References

Jain, R., Gour, B., & Dubey, S. (2016). a hybrid approach for credit card fraud detection using a rough set and decision tree technique. International Journal of Computer Applications, 139(10), 1-6.

Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., Dedenne, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research, 176(1), 565-583.

Kirkos, E., Spathis, C., Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert systems with applications, 32(4), 995-1003.

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques.Decision Support Systems, 50(2), 491-500.

Li, C., Poskitt, D. S., & Zhao, X. (2019). The bivariate probit model, maximum likelihood estimation, accurate pseudo parameters, and partial identification. Journal of Econometrics, 209(1), 94-113.

Hoogs, B., Kiehl, T., Lacomb, C., &Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement

Dai, Y., Yan, J., Tang, X., Zhao, H., & Guo, M. (2016, August). Online credit card fraud detection: A hybrid framework with big data technologies. In 2016IEEE Trustcom/BigDataSE/ISPA (pp. 1644-1651). IEEE.

Mubarek, A. M., &Adalı, E. (2017, October). Multilayer perceptron neural network technique for fraud detection. In the 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 383-387). IEEE.

Sisodia, D., &S.Bhandari, N. r. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. IEEE International conference on power,control,signals and instrumentation ,(Chennai), 2747-2752.(2017)

Sadi Gali, I., Sael, N., Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia computer science, 148, 45-54.

Bhanusri, A. (n.d.). Credit card fraud detection using machine learning algorithms. Journal of Research in Humanities and Social Science, Volume 8 ~ Issue 2 (2020), 04-11.

awoyemi, j. o., adetunmbi, a. o. (2017). Credit card fraud detection using machine learning techniques.IEEE, 978-15090-3/17.

zheng, Lutao, and Guanjun liu. "ieee." A New Credit Card Fraud Detecting Method Based on Behavior Certificate, vol. 978–1–5386–505–0/18/$31.00, no. 2018.

Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," IEEE Access, vol. 6, pp. 14277-14284, 2018.

A. Roy and J. Sun and R. Mahoney and L. Alonzi and S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," in Systems and Information Engineering Design Symposium (SIEDS), pp. 129-134, 2018.

Guanjun Liu, Zhenchuan Li, Lutao Zhang, Shuo Wang, and Changjun Jiang Shiyang Xuan, "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference On Networking, Sensing and Control (ICNSC), pp.1-6, 2018.

Wang, Deshen, Bintong Chen, and Jing Chen, "Credit card fraud detection strategies with consumer incentives." Omega (The International Journal of Management Science), 2018.

Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-EdouardPortier, Liyun He-Guelton, and Olivier Caelen. "Sequence classification for credit-card fraud detection." Expert Systems with Applications vol.100, pp: 234-245, 2018.

Adriano C.M. Pereira, Gisele L. Pappa Alex G.C. de Sá, "A customized classification algorithm for credit card fraud detection," vol. 72, pp. 21-29, 2018.

Akila S, Srinivasulu Reddy U, "Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection," Journal of Computational Science, vol. 27, pp. 247–254, 2018.

FabrizioCarcillo, Andrea Dal Pozzolo, Yann-A¨el Le Borgne, Olivier Caelen, YannisMazzer, GianlucaBontempi, "SCARFF: a Scalable Framework for Streaming Credit Card Fraud Detection with Spark", Information Fusion, vol. 41, pp.182-194, 2018.

SADGALI, I., & Sael, N. (2019). Fraud detection in credit card transactions using machine learning techniques. IEEE, *978-1-7281-4368-2/19*.

Mittal, Sangeeta. n.d. "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection." 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 978-1-5386-5933-5/19 (2019).

More, R. S., Hawaii, c. j., & shirgave, D. s.k. (n.d.). credit card fraud detection using a supervised learning approach. International journal of scientific & technology research, volume 9(10 October 2020).

Suma, Dr. V., and Shavige Malleshwara Hills. (2020). "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) Vol.02/ No.03:Pages: 153-159.

Thennakoon, Anuruddha, and chee bhagyani.(2019). "Real-time Credit Card Fraud Detection Using Machine Learning." International Conference on Cloud Computing, Data Science & Engineering (Confluence) 978-1-5386-5933-5/19.