Credit Card Fraud Detection Using Machine Learning

Zainab Firdous[1], Sushma V[2], Aftab Pasha S[3], M Shahista Banu[4], Najmusher H[5]

[1,2,3,4]Dept. of CSE, HKBK College of Engineering, Bangalore
[5]Professor, Dept. of CSE, HKBK College of Engineering Bangalore

---***---

Abstract – Fraud is the act of depriving a person/organization of money through willingness, deception or other unfair means. The unforeseen event of Covid-19 has led to many people embracing digital transactions and online shopping. This combined with other benefits provided by credit card issuers such as rewards has increased the usage of credit card, and in turn increased credit card frauds. Credit card default can have serious implications on credit card holder and can affect financial stability of credit card issuers. There is a need for a system that can predict defaults ahead of time so that appropriate measures can be taken by credit card issuers. In this paper, we have provided a comparative analysis of various machine-learning algorithms often used in fraud detection such as logistic regression, decision tree classifier, random forest classifier and support vector machine classifier. The models were compared on the basis of precision, recall and accuracy to find the best algorithm for predicting probable defaulters.

Key Words: **Credit Card fraud, Credit Card default, Machine Learning, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest**

1.INTRODUCTION

The banks earn money by various means such as lending loans to other customers using the depositor's money. The interest gained is the profit earned by the bank, but when the borrower defaults, the loan becomes a NPA and is a huge blow to the bank's statements. It was estimated that the total amount of NPAs in India increased from 2.39 lakh crore in 2014 to 10.36 lakh crore in 2018. Therefore, an effective system must be developed to curb these defaults even before they occur. Different types of defaults in finance include:

- Loan default: This is the most common type of default. In this case the loan borrower fails to repay the loan.
- Credit card default: This occurs when the credit card holder uses his/her credit card to buy items that they cannot afford but doesn't repay the money spent.
- Bond default: when the organization/government fails to repay the loan/principal amount, it is considered as bond default.

We have focused on developing a system to predict credit card defaulters. A comparative analysis was done on the performance of various algorithms such as logistic regression, decision tree, random forest and support vector machine based on precision, recall and accuracy. The traditional systems available such as CIBIL score uses demographic data, credit history and so on to calculate a score. This score is employed by many money lending organizations to judge whether to issue the loan to this applicant or not and to set the credit limit in case of credit cards. However, this system only gives an idea to banks of the level of risk involved in granting the person loan. Hence, we have developed a model that predict with precision the probability of a user defaulting.

2. LITERATURE REVIEW

The paper by author Yue Yu, **[1]** concentrates on the importance of credit card default prediction by highlighting the need of timely and precise prediction to prevent financial losses for both banks and the users, then proceeding to make an outline of various machine-learning based algorithms by explaining their principles, advantages and comparing their performances. Yue Yu then introduces the dataset used in the paper, which includes thirty thousand entries of credit card users from a Taiwanese bank containing their previous credit card transaction information, card-user demographics, and payment behavior to train and evaluate these algorithms. The paper then presents an analysis of applying the selected machine learning algorithms to the dataset and evaluating their performance using parameters such as accuracy, precision, recall, score of harmonic mean and Receiver Operating Characteristic graph. The results show that all the algorithms considered for the test produce good results but artificial neural networks and support vector machine giving better accuracy and performance rates. The article ends by admitting some drawbacks and potential scope for future study.

Authors, Yashna Sayjadah, Khairl Azhar Kasmiran, Ibrahim Abaker Targio Hashem, and Faiz Alotaibi, **[2]** in their paper, have recognized challenges associated with credit card default prediction, such as imbalanced datasets and the requirement for precise models. To analyze their study, the authors obtained a dataset containing credit card information and default status which is pre-processed by handling missing values, encoding certain variables, and conducting feature ascending. Following which the dataset is categorized into training and testing purposes. Accuracy of different algorithms was noted for the possible occurrence of default credit card. Algorithms used here are logistic regression, decision tree and random forest, the accuracy of these were 75%, 64% and 77% respectively. Based upon this the authors have concluded that banks can adapt the random forest algorithm to detect the credit card defaulters so that they can analyze the risk before giving the credit card to the clients. In general, the paper captures a thorough examination of credit card default prediction using machine learning methods and a seemingly bright platform for future hypothesis.

Alžbeta Bačová and František Babič's paper, **[3]** draws attention towards creating a model that can predict the possibility of a customer bound to miss his payment. For the purpose of their investigation, the authors employed a dataset that contained generic data on credit card users. From the thesis, it was deduced that predictive analytics may effectively identify clients who are more likely to forget to pay using a credit card. The researchers found that specific variable feature, definitely influence how a prediction can be made. The trained model was accurate, which hints that there is a way for them to find purpose in the credit card market. The paper establishes how predictive analytics may be used to identify credit card defaulters and reduce financial risks for credit card issuers. To sum it up, the paper serves as a valuable resource for credit card companies and industry professionals keen in implementing predictive analytics to predict defaults of credit card holders.

The writers Talha Mahboob Alam, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi, **[4]** investigates the problem of credit card default prediction using imbalanced datasets. The authors start with citing the need of a proper detection and prediction system to identify precise and accurate default predicting project, as it helps establishments identify dicey borrowers and make calculated decisions regarding credit approvals. There is always an imbalance present while going through different data sets. This can lead to biased models that perform poorly in predicting defaults. the impact of datasets on credit card default prediction is investigated along with different techniques to address this issue. The authors find that regular classifiers trained on imbalanced datasets tend to benefit the majority class, resulting in smaller recall score for the defaults. However, by applying sampling techniques, the performance of the classifiers improves drastically. Adaptive Synthetic Sampling consistently beats other techniques, providing better prediction accuracy and recall precision for default instances. The authors highlight the need for more comprehensive and diverse credit card default datasets to validate the findings of this study and to facilitate the development of better models.

The paper prepared by authors, Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Saïd Hacid, and Hassan Zeineddine, **[5]** introduces different approaches for detecting credit card frauds occurring through imbalanced datasets. The experimental arrangement involves using a practical credit card transaction. The authors note that while hybrid sampling and certain algorithms provide the required results, choosing the proper algorithm may depend invariably on the problem at hand. The paper shows different overview of researches of existing approaches and techniques to check the ones who have defaulted. They arrive at a solution that no algorithm alone is suffice to address the issue of imbalance in datasets. The research work briefs about confusion matrix and related results are formulated in tables. The confusion matrix, is also given by the term error matrix, it is a table that is frequently used to assess how well a categorization model is working. It summarises the model's predictions and the degree to which they agree with the actual labels or data classes. The genuine class labels and anticipated class labels are represented, respectively, by rows and columns in a grid that represents the confusion matrix. For the purpose of detecting credit card fraud, the paper offers an experimental investigation of unbalanced classification algorithms. To increase the accuracy of fraud detection, it evaluates the body of research, discusses various methodologies, shows experimental findings, and stresses on balancing classes.

D. Tanouz, G V Parameswara Reddy, R Raja Subramanian, A. Ranjith Kumar, D. Eswar, CH V N M Praneeth, **[6]** have made a report about the growing threat of credit card fraud and the necessity for effective detection systems. The approach depicts promising results and contributes to the development of strong fraud detection systems by recognizing the rising threat of credit card fraud, the authors dwell into a broader spectrum of the research of various algorithms suitable for fraud detection. To validate their proposed approach, the researchers conduct experiments using a practical credit card dataset. They as well assess and compare the parameters of the different algorithms employed. The efficiency of machine learning algorithms in detecting credit card fraud by leveraging patterns is marked and regular techniques process the results and upon tabulating the outcomes various features were considered which were put into tables evident in the paper. The authors examine approaches such as primary component study, feature ranking and engineering, providing visions into their impact on the accuracy and efficiency of the systems. The authors also discuss the importance of dataset pre-processing and handling techniques to address the inherent class imbalance problem in credit card fraud datasets. The authors conclude by making a reference to a number of studies, research papers, and publications in the area of credit card fraud detection.

3. METHODOLOGY

3.1 DATA

The dataset used for building the models was obtained from UCI Machine Learning Repository. It consists of 30,000 observations and 24 features. This information was collected from Taiwan as part of a research in the year 2005. It consists of demographic details of the user, payment history, bill amount and the amount paid by the user.

Feature	Description
ID	It is a number used to identify the user
LIMIT_BAL	It is the limit of the credit card
SEX	Indicates the gender of the user (1=male, 2=female)
EDUCATION	It indicates the level of education of the user (1 = graduate school, 2 = university, 3 = high school, 4,5,6,0 = others)
MARRIAGE	It indicates the marital status of the user (1=married, 2= single, 3,0=other)
PAY_0 to Pay_6	It indicates the payment status of the credit card from April to September 2005(-2 = no consumption, -1 = pay duly, 1 = one month delay, 2 = two-month delay....9 = nine-month delay)
BILL_AMT1 to BILL_AMT6	It indicated the bill amounted on the credit card from April to September 2005
PAY_AMT1 to PAY_AMT6	It indicates the amount paid by the user
DEFAULT	It indicates if the user defaulted next month or not (0 = non-defaulter, 1 = defaulter)

Table 1: Description of the dataset

3.2 EXPLORATORY DATA ANALYSIS

We have used matplotlib and seaborn libraries to visualize and get better understanding of the data. It is clear from the bar graph in figure-1 that the data is highly imbalanced i.e., the number of non-defaulters is far more than defaulters, this can lead to a bias towards class 0(non-default) while modelling.

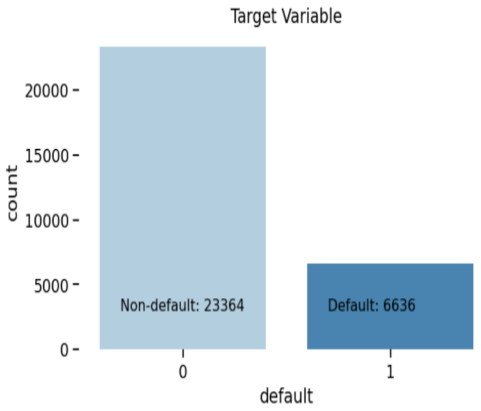

Figure 1: Target variable visualization

The contribution of various categorical values that is gender, education and marital status was analyzed in terms of default.

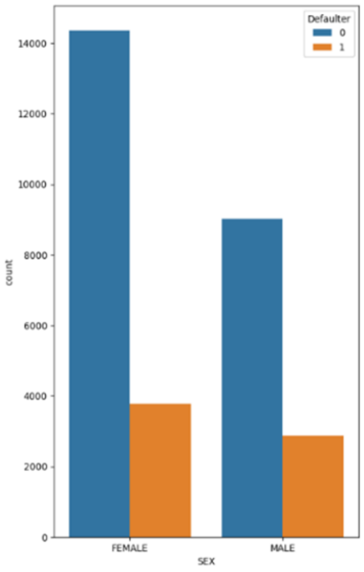

Figure 2: Visualizing gender in terms of default

It can be observed from figure-2 that there was generally more female data in the dataset compared to male data, and as a consequence more females have defaulted as compared to males.

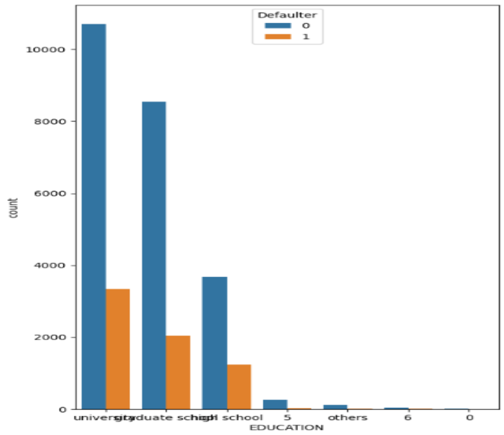

Figure 3: Visualizing educational level in terms of default

Majority of the users in the dataset have educational level of university, the users with educational level have defaulted more, followed by educational level of graduate and finally high school.

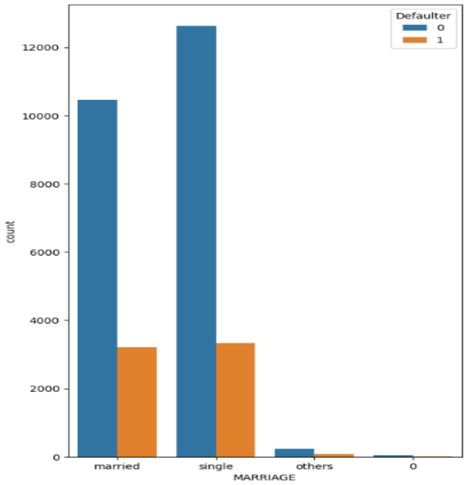

Figure 4: Visualizing marital status in terms of default

It can be noticed from figure-4 that the dataset mostly has users who are single. There is not much difference in number of defaulters who are single and defaulter who are married. A slight percent of 'other' category has also defaulted.

Finally, a correlation heatmap was plotted to understand the relationship of various attributed with each other and the target variable.

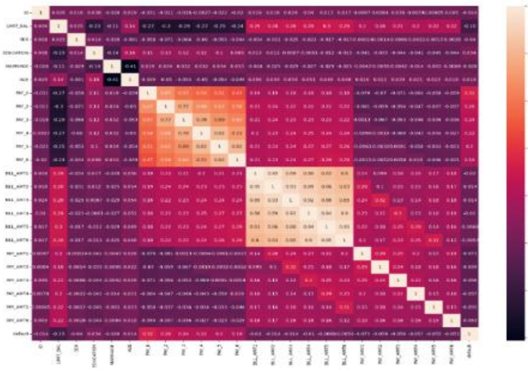

Figure 5: Correlation heatmap

It can be noticed that default is only positively correlated with payment status features. The limit_bal is positively correlated with bill_amt and pay_amt, which is obvious since the credit card issuers tend to give higher credit limit to applicants with good credit history. Also bill_Amt and payment status features are also positively correlated to each other and bill_amt and pay_Amt are also positively correlated.

3.3 DATA PRE-PROCESSING

From the figures-3 and figure-4 it can be noticed that there is inconsistency in the labels of these categorical features, that is multiple labels seem to be indicating the same category. In 'Education' categorical feature, the label 4,5,6 and 0 all indicate the 'other' category. In 'marriage' categorical feature, the label 3 and 0 indicate users in 'other' category. This can cause confusion for the model in prediction of default, so we have removed these inconsistencies by combining these labels into one.

Some algorithms can be sensitive to categorical features. Thus, we encoded these categorical features using the Label Encoder class of scikit-learn library.

We dropped the feature 'ID' since it is only used to indicate the user and has no contribution towards the prediction of default. Then the dataset was split into the train set and the test set (80:20) using the train_test_split class of scikit-learn library.

Finally, we standardized the dataset by scaling the data in terms of median absolute deviation using sklearn library.

3.4 MACHINE LEARNING ALGORITHMS

The main objective of the project is to predict probable defaulters for which we have used four most popular machine learning algorithms often used for fraud detection. These are:

Logistic regression:

This is popular classification algorithm that can takes any input value and gives the probability of the dependent variable between 0 and 1. It then fits a sigmoid function and when the new input data point comes it calculates its probability, if the probability is greater than the threshold it is predicted as 1 or it is predicted as 0.

Decision tree:

Decision tree builds tree that predicts the outcome of the dependent variable. The model asks a question and based on the answer splits the dataset into subset, here one node corresponds to the answer yes and another node corresponds to answer no. The model keeps splitting until max_depth value has reached or the node contains values from same class.

Random forest:

Random forest is based on ensemble learning which combines multiple decision tree and predicts dependent variable by taking average of all the decision tree. This gives better accuracy and reduces the problem of overfitting.

Support vector machine:

Support vector machine algorithm creates a decision boundary (hyperplane) that separates default and non-default cases. This hyperplane is chosen such that it maximizes the margin between the two classes. The margin is created by calculating the distance between the hyperplane and support vectors.

K-Nearest Neighbor:

KNN algorithm stores all the data during the training phase and when the new data point comes in, it calculates the distance between the new point and the neighbors. Then selects the stated number of nearest neighbors. Finally, the number of default and non-default outcomes are counted, whichever has the majority is assigned as the outcome of the new data point's dependent variable.

4. RESULTS

The various algorithms were evaluated on the basis of accuracy, precision and recall. Accuracy is the measure of predictions that are correctly predicted out of the total predictions. Precision is measure of true positive that were actually correct. Recall is measure of true positives that were correctly identified. High recall will lead to fewer defaulters being precited as non-defaulters where as High precision will reduce the number of defaulters predicted as non-default. However, there is a trade-off between precision and recall. For our problem statement it is important to have high precision, since we don't want our system to be predicting non-defaulters as defaulter.

Algorithm	Accuracy
KNN	80
Logistic regression	81
Decision tree	73
Random forest	81
Support vector machine	82

Table 2: Accuracy results

Decision tree has the least accuracy and the best accuracy is given by SVM.

Algorithm	Precision for class 0	Precision for class 1
KNN	0.92	0.36
Logistic regression	0.97	0.24
Decision tree	0.81	0.41
Random forest	0.94	0.36
SVM	0.96	0.36
SVM (changed threshold)	0.84	0.64

Table 3: Precision results

It can be noticed that generally the precision is more for class 0 (non-defaulter), this is due to the imbalance in dataset. As a result, even though the accuracy is good for some of the algorithms, they hardly can predict defaulters correctly. But on changing the decision threshold of support vector machine, the precision for class 1(defaulter) drastically increased, which means it can correctly predict 64% defaulters and 84% non-defaulters.

Algorithm	Recall for class 0	Recall for class 1
KNN	0.84	0.55
Logistic regression	0.82	0.69
Decision tree	0.83	0.38
Random forest	0.84	0.63
SVM	0.83	0.70
SVM (changed threshold)	0.94	0.37

Table 4: Recall results

In case of recall too, generally the recall for class 0 is more than class 1. The best recall for class 1 was from support vector machine. The confusion matrix for the various algorithms were as follows:

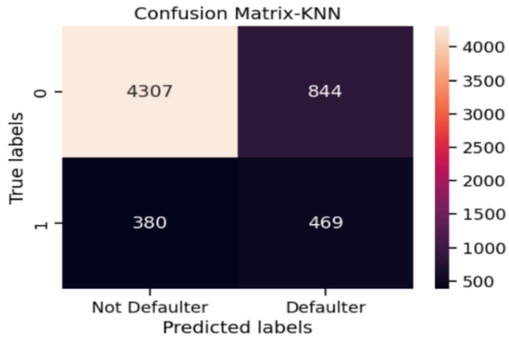

Figure 6: KNN confusion matrix

Although KNN has high true positive and true negatives, still the number of false positive and false negative is quite high.

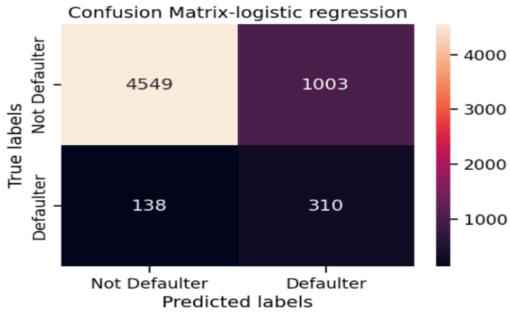

Figure 7: Logistic regression confusion matrix

Logistic regression has high true positives, it performs quite bad at predicting defaulters. Thus, logistic regression doesn't serve the purpose of this proposed system.

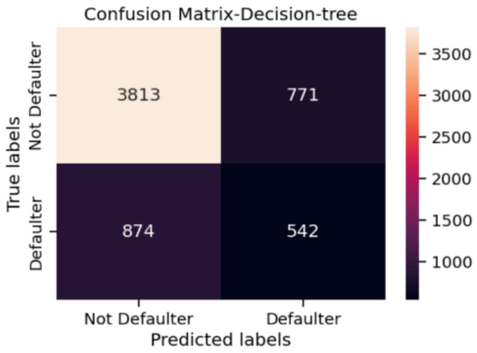

Figure 8: Decision tree confusion matrix

Even though decision tree has predicted the greatest number of defaults by far, the true positives have reduced and false negatives have increased, which means quite a lot of defaulters were let off the radar.

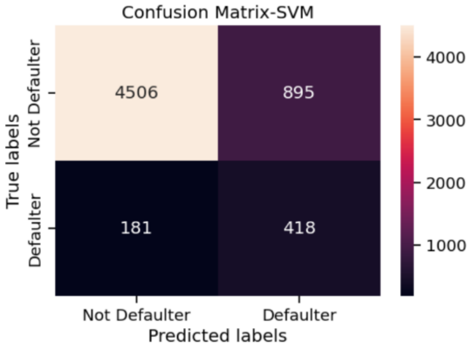

Figure 9: SVM confusion matrix

SVM seems to be doing well with good number of true positives and true negatives and low false negatives.

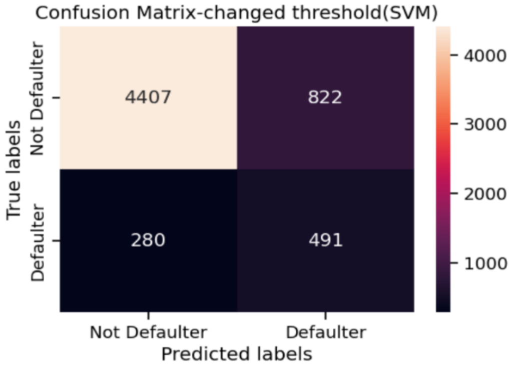

Figure 10: SVM changed threshold

However, changing the threshold to increase the precision for class 1, provides far better results. The number of true negatives has increased and the number of false positives has decreased.

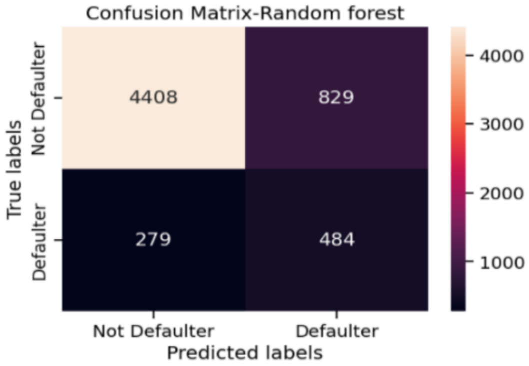

Figure 11: Random Forest confusion matrix

Random forest too has performed good with high true positives, true negatives.

However, SVM with changed threshold seems to be performing the best with both true positive and true negatives being high and reducing false positives.

5 CONCLUSIONS

The main objective of this proposed system is to correctly predict defaulters. We employed KNN, logistic regression, decision tree, random forest and support vector machine to find the best algorithm for the problem statement. We evaluated these algorithms based of accuracy, precision, recall and comparing their confusion matrix. The major challenge we faced was the imbalanced dataset, however under sampling or oversampling seems to be misrepresenting the relationship between X and Y leading to poor performance. Thus, we have used the dataset without any sampling for modelling. It can be noticed that accuracy alone cannot be used to judge the performance of a model. Even though logistic regression has high accuracy, it performed poorly on test set. In conclusion, both support vector machine and random forest have good accuracy. SVM with slightly changed threshold performs the best with reduced number of false negatives and better recall and precision. However, it is hard to narrow down on one of the models, because of the limited user data available due to confidentiality.

REFERENCES

[1] Yue Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction", International Conference on Computing and Data Science, 978-1-7281-7106-7/20 © 2020 IEEE DOI:10.1109/CDS 49703.2020.0050.

[2] Yashna Sayjadah, Khairl Azhar Kasmiran, Ibrahim Abaker Targio Hashem, Faiz Alotaibi, "Credit Card Default Prediction Using Machine Learning Techniques", 978-1-5386-7167-2/18 © 2018 IEEE.

[3] Alžbeta Bačová, František Babič, "Predictive Analysis for Default of Credit Card Clients", SAMI 2021, IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, January 21–23, Herl'any, Slovakia, 978-1-7281-8053-3/21/$31.00 ©2021 IEEE.

[4] Talha Mahboob Alam, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, Matloob Khushi, "An Investigation Of Credit Card Default Prediction In The Imbalanced Datasets", DOI: 10.1109/ACCESS.2020.3033784.

[5] Sara Makki, Zainab Assaghir, Yehia Taher, Rafiqul Haque, Mohand-Saïd Hacid, Hassan Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection", DOI: 10.1109/ACCESS.2019.2927226.

[6] D. Tanouz, G V Parameswara Reddy, R Raja Subramanian, A. Ranjith Kumar, D. Eswar, CH V N M Praneeth, "Credit Card Fraud Detection Using Machine Learning", Fifth International Conference on Intelligent Computing and Control Systems ICICCS 2021, IEEE Xplore Part Number CFP21K74-ART; ISBN: 978-0-7381-1327-2.