Identifying Credit Card Defaulters and Predicting Fraudulent Transactions Using Various Machine Learning Techniques

Shreeyash Bhaskar Mane Deshmukh[1], Dr. Savita Sangam[2]

[1]PG Scholar, Department of Computer Engineering, School of Engineering and Applied Sciences, Kalyan(w)
[2]Professor, Department of Computer Engineering, School of Engineering and Applied Sciences, Kalyan(w)

Abstract

In recent years as increase in number of credit card transactions, large amount of data generated via Razor pay, Billing desk payment gateways. Indian government encourages to each and every individual to do digital transactions via online payment gateway under the scheme of digital India. For going cashless economy under the banking domain credit card transactions for the booking flight tickets, online shopping, railway ticket booking has increased rapidly after the covid pandemic. Credit card banks giving offers like some amount of percentage of cashback and some credit points for the refilling the fuel on the selected petrol pumps like BPCL and HP so a greater number of users to use credit card on petrol pumps for filling up the fuel tank. This is the positive side of the digital India concept. Due to increasing number of online transactions via credit card it also increases the defaulters and fraud by using credit cards transactions. During online payment via credit card a large amount of data is generated including customer information, credit card number, what type of product customer like to purchase, credit history, transactions history and has the customer pay the credit card bill before due date. As the customers CIBIL score is generated on the basis of due date payment of credit card which is used for future reference. Credit card holder have serious implication if credit card holder is a defaulter. So, we are going to propose a solution of generating a model by using different machine learning algorithm techniques to avoid such situation. We are doing comparison of different performance models to identify defaulters using evaluation matrices such as accuracy, precision, recall and F1-score. Including decision making on historical credit card transaction using machine learning algorithms Support vector Machine, Logistic regression and Random Forest to identify credit card defaulters to preventing the financial loss of the credit card lending banks.

Keywords: Logistic Regression, Random Forest, Support Vector Machine, Machine Learning.

1. Introduction

Credit card default occurs when the credit card holder uses credit card to buy items that they cannot afford but doesn’t repay the money spent. Nowadays, credit card offers various offers to users such as 10% or 20% of cashback on particular product on E-commerce platform so a greater number of users attracted to the offers and do the online transactions without verifying the genuine website which leads to online frauds. To avoid such situation, we are monitoring the behaviour of the user and analysing the online credit card transaction history. By using users credit card transaction history and payment history credit card lending banks making decision on approving new credit card to the users to minimise the loss. The main goal is to develop a model which is predict credit card defaulters by using large amount of data. We have to understand the various machine learning algorithm techniques to build model for identifying the credit card defaulters and predicting the credit card fraud. Sometimes saved data like cvv, credit card numbers, first name, last name on google chrome used by hackers to do the fraudulent transactions without using the physical credit card. So, this is financial risk for bank as well as to the customers. To avoid such risk, we are building a model to detect such transaction using machine learning algorithm techniques. By using the dataset provided by the bank we can find the defaulters. As we are monitoring and analyzing the credit card defaulter's dataset, we are making it sorted on the basis of payment on regular basis or not. On the analysis of data, we are finding customers area of interest on which customer spent most amount of money. If the minimum due amount is more than the paid amount means you are default so it will affect your CIBIL score which calculates credit score. Depending upon your credit score financial institute decides to lend credit card to the particular customer or not based on customers purchasing history behavior. The apprehensive transactions are found by analyzing all the transactions using machine learning algorithm. As we monitoring customers purchasing history ended up purchasing more than the credit card limit, more on online shopping. All these purchasing history and user behavior easily understand by the machine learning algorithms techniques which helps banks to make decision on offering credit cards to the customers. Hence, in this paper we are using different machine learning algorithms such as Random Forest, Support Vector machine and Logistic Regression. By implementing Regression and classification algorithms we can identifying credit card defaulters and credit card fraud. As per our exploration we found out that Random Forest algorithms gives the adequate result than the other machine learning algorithm techniques.

2.1 Problem Statement

To build a classification methodology to predict whether a person defaults the credit card payment for the next month or not. Folks uses credit card more than the credit limit by using overlimit facility for their personal use or sometimes they miss the deadline of due date because of fund shortage which affects his credit score as well as financial loss of credit card lending banks. To avoid such vulnerability, it is essential to recognize folks who are failing to pay the minimum amount. Various types of fraudulent transactions commenced by the cardholder or the third person without consent. Card not present fraud is denoting typically situations where physical credit card is not present during transactions. This type of fraud occurs online or remote transactions where information like name, credit card number, CVV are used without physical card. The big challenge is to protect customers private data to avoid such financial loss. One solution for this to build a filter-based machine learning model to identifying credit card defaulters and credit card fraud done by third party without physical card.

2.2 Objectives

1. Propose a system to detect credit card fraudulent transactions.
2. Minimize misclassification of credit card fraud and defaulters.
3. To secure financial sector by Predicting fraudulent transactions.
4. To predict credit card defaulters.

3. Literature review

Machine Learning methods in paper [1] For analysis credit card data utilized for fraud detection, found that data-set was imbalanced and utilized in augmentation. The machine learning algorithms were used Naïve bayes, Random Forest and Logistic Regression

In paper [2] Goyal and Kaur used logistic regression to build model tree that consists of a data set of 13 attributes and the result accuracy is between 69 % to 80% in five runs. For this test they have used R programming language.

Authors, Yashna Sayjadah, Khairl, Azhar Kasmiran, Ibrahim, Abaker, Targio Hashem, and Faiz Alotaibi, in their paper [3], have recognized challenges associated with credit card default prediction, such as imbalanced datasets and the requirement for precise models. To analyse 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.

This research in paper [4] oversampling support banks to overcome with the challenge of reducing risk of borrowers going bankrupt. To address this challenge, they suggest fresh way of evaluating whether someone is payback the due amount on time or not. Using hierarchical clustering and k-means method they propose model of borrowers with similar risks. The models they created are used to make decision about how much money to lend to each borrower what interest rates to charge them and what term to offer them for repayment.

In paper [5] Dataset categorise a big collection of credit card information into different groups based on certain criteria. Their research shows that batch normalisation techniques can used to make model faster while also reducing the risk of overfitting.

Author of paper [6] Zhuhai, studied two variations of random forests to tech the model about both regular and unusual transactions. They looked at two types of random forests that use two different basic classifiers and studied how well they work for spotting credit card fraud.

This paper [7]A dataset of real-world transactions is used to find out fraudulent credit card transaction using supervised machine algorithm. They tested different machine learning algorithms that learned from exampled to see how well they detect fraudulent transactions compared to their own advanced classifier.

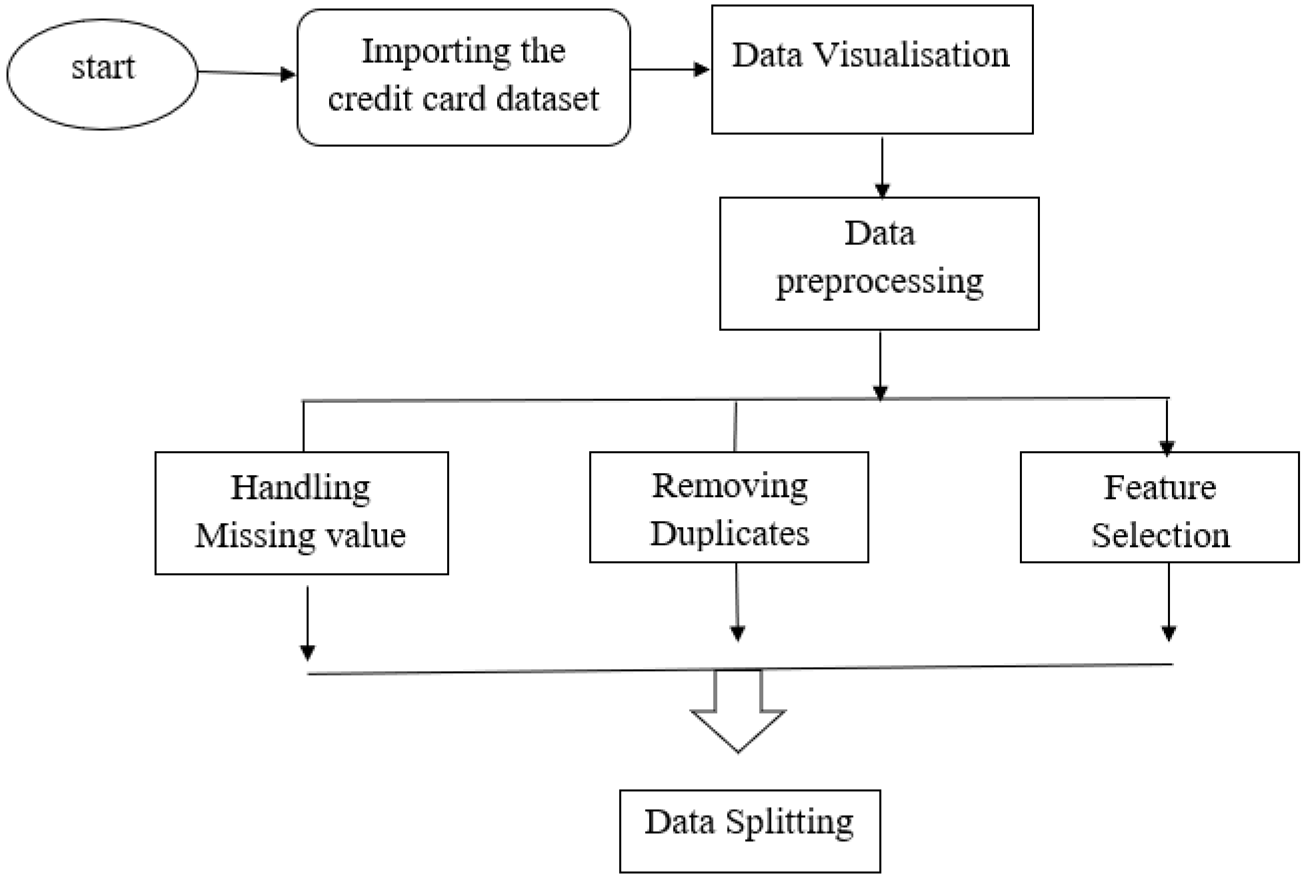
In paper [8] have made a report about the growing threat of credit card fraud and the necessity for effective detection systems. To validate their proposed approach, the researchers conduct experiments using a practical credit card dataset. They as well assess and compare the parameters of the different algorithms employed. The 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.

4. Proposed Method.

- This study primarily aims to find and predict people who might not pay back their credit card bills on time using machine learning algorithm that learn from data set. The suggested solution to build model uses a supervised machine learning algorithm to identify potential credit card users who might not repay their credit card bill on time. Credit card activity that differs significantly from customer's usual spending patterns are considered unusual or unfair.
- The dataset includes information about American Express credit card users and their transactions sourced from Kaggle.com, the dataset consists of total 346,982 transactions of which 761 are

considered fraudulent transaction. This imbalanced dataset with only small fraction of transactions is fraudulent. As Given sensitivity of providing transaction details maintaining confidentiality is crucial.

- In this dataset many columns have missing values indicates by Nan. The next step involves removing any columns from dataset that contain missing values.
- Initially, the original train dataset contains 113 columns. However, after removing columns with missing values only 87 column remains in the train dataset. It is crucial to address the missing values in dataset.
- After combining duplicate customer ID's and removing certain columns from the train dataset, there was a slight impact on distribution. However, this effect was deemed relatively minor and datasets remained largely unaffected.
- Feature selection is a process of choosing a subset of input variables(attributes) by removing features. This process believes to improve accuracy of the resulting classifier and offer leads to a model that generalizes better to new data points which includes thirty-six thousand observations and twenty-nine features with fifty-two thousand rows after grouping customer ID's it now has over thirty-eight thousand rows likewise test dataset has over thirteen million rows now it reduced to eighty-nine thousand rows.
- We are going to split the credit card dataset into two sections train dataset and test dataset. Typically, ratio of 80:20 is commonly used for split the dataset into train dataset and test dataset. The classification model learns from the training dataset to understand the pattern then it's tested on the remaining data in the test set to see how well it predicts outcomes using various methods



\

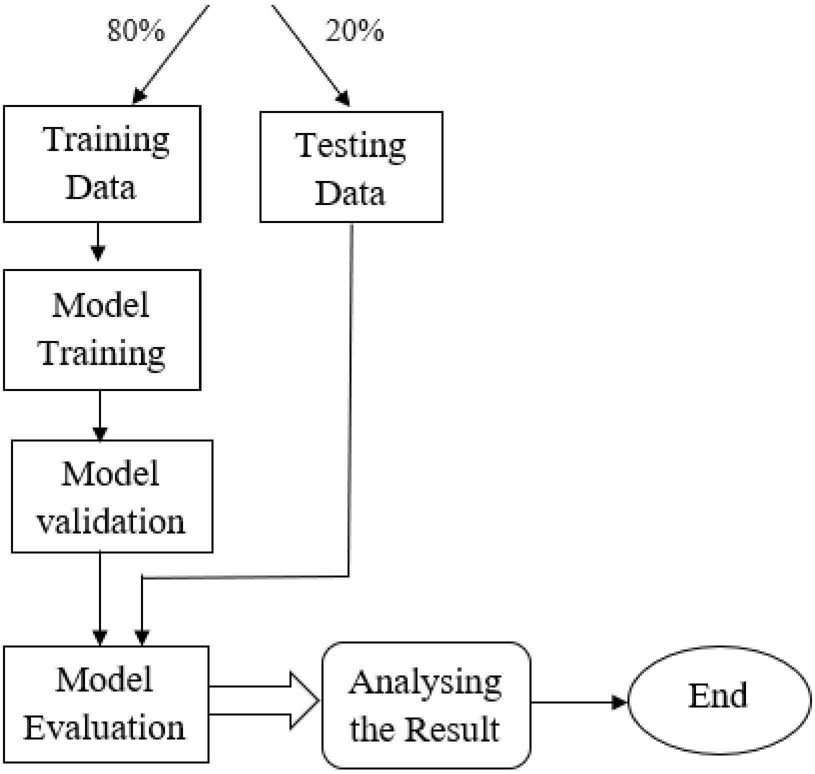


Fig:1 System Flow Diagram

5. Mahine Learning Algorithms:

After splitting the data, it is crucial to determine the most effective method for predicting the credit card defaulters. To predict probable defaulters for which we have used most popular machine learning algorithms often used for credit card defaulters' detection.

5.1. Logistic Regression: The Logistic regression is statistical method for predicting the probability of binary outcome such as whether a credit card payment will default or not based on one or more predictor variable. It fits sigmoid function and calculate probability of new input data point. If the probability higher than the predefined threshold it predicts the outcome as one otherwise it predicts outcome as zero.

Sigmoid function is a mathematical function that maps any real valued number to a value between zero and one. In logistic regression the sigmoid function converts the output of the linear combination of input features into probabilities.

S(X) = 1/(1+e^(-x))

Where **S(X)** is sigmoid function

e is Euler's number

5.2 Random Forest: Random Forest is ensemble learning method that consist of multiple decision trees. For classification task algorithm predicts the class that is mode of the individual tree prediction. For regression task it predicts the average of the individual tree prediction. This approach helps to improve the accuracy and robustness of the model by reducing overfitting and capturing more complex pattern in the data. For classification decision tree are built on different subset of the data and each tree makes its prediction.

Whether the dataset contain numerical value or categorical value Random Forest can effectively learn from the data and make accurate prediction.

5.3 Support Vector Machine: The main idea behind support vector machine is to transform data points into a higher dimensional space where hyperplane can be drawn to separate different classes with the maximum possible margin. Two hyperplanes are constructed on either side of the planes that differentiate data. These hyperplanes are positioned to maximise the distance between them effectively wide margin between the classes which helps to improve the model's ability to generalise new data. The assumption is that the larger the gap between the parallel hyperplanes created by support vector machine, the lower the overall error of the classification model. This is because a larger margin indicated better separation between different classes of data, leading more accurate data.

6. Result of prediction Models

The main aim of this study is to help credit card lending banks to lower their financial risks by correctly identifying people who might not pay back their credit card bill on time. Three different machine learning techniques are used for this analysis those are support vector machine, Logistic regression and Random Forest. The dataset is used for this analysis is sourced from Kaggle.com contained information about credit card transaction from bank. Using those algorithms train the classifier and build the classification model. 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.

The confusion matrix is a tool used to assess the performance of machine learning model. It provides summary of the model's predictions compared to actual outcomes. The confusion matrix organises the data into a matrix based on two factors the actual category of each record and classification prediction made by the model. The performance of different algorithm is based on three key matrices are precision, recall accuracy. Precision is a measure of the accuracy of the positive prediction made by a classification model. It calculates the proportion of true positive prediction among all instances predicted as positive by model. Precision simply answers the question: of all instances predicted as positive how many were actually positive? Recall also known as true positive rate is a measure of the completeness of the positive prediction made by a classification model. It calculates of all actual positive instances how many were correctly identified by the model. Accuracy is a measure of the overall correctness of the prediction made by the classification model. In simple terms of all instances in dataset how many were correctly classified by the model.

6.1 Classifiers according to Confusion Matrix

The results are presented in table highlighting significant difference in accuracy precision and recall. The high dimensional data handled by the random Forest which means there is no need to reduce dimensionality of the feature. We can identify the most influential variables in predicting customer default behavior by using Random Forest.These key variables helps to determine whether a customer is likely to default or not.

Classifier	Accuracy (%)	Precision	Recall
Logistic Regression	91.27	0.875	0.84
Support Vector Machine	89.54	0.869	0.81
Random Forest	93.47	0.891	0.86

Table 1. Result of confusion matrix

As the dataset is biased all classifiers achieved an accuracy over 89%. Banks gives credit cards to people they think they can payback what they borrow if they don't believe someone can do that, they won't give them card. It's all about making sure people can handle the responsibility of using credit card. Our experiment shows that Random Forest (RF) has outperformed than logistic regression and support vector machine (SVM) algorithm. This method is being used is good is not only for predicting risks but also dealing with the dataset that keep getting bigger over time. from this study as seen it indicates that how the training data was managed didn't cause significant changes to original data or affect how well machine learning algorithm performed as it was evaluated at 0.89, it suggest that the method is effective, accurate and adaptable machine learning model.

7. Conclusion

This study tries to figure it out how many people in a specific dataset might not payback their credit card bill on time. It does this by testing different machine learning algorithm to see which one is best at predicting who might default on their credit card payment. This study found that Random Forest method was more effective than other common machine learning techniques logistic regression and support machine algorithm. Using machine leaning to predict credit card defaulters is important because it helps to identify customers who might not pay back their credit card bill on time. This helps banks manage their credit risk more effectively. By accurately predicting credit card defaulters' financial institutions can make smarter decisions when planning their prospects schemes. Among all classifiers tested Random Forest showed the best performance. When there are numerous examples to classify the performance of classify is not as good as when there are fewer examples. It is important to recognize that human response is always evolving from one day to the next. So, this study decided to use newer data to find out if the Random Forest technique is the still best for predicting credit card defaults.

Refrences:

1. Dejan Varmedja; Mirjana Karanovic; Srdjan Sladojevic; Marko Arsenovic; Andras Anderla; 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) 20-22 March 2019
2. A. Goyal, R. Kaur, 2016. Accuracy Prediction for Loan Risk using Machine Learning Models. International Journal of Computer Science Trends and Technology, 4(1), pp. 52-57.
3. 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.
4. Orlova, Ekaterina V. "Methodology and Models for Individuals' Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods." *Mathematics*, vol. 9, no. 15, Aug. 2021, p. 1820. *DOI.org (Crossref)*, https://doi.org/10.3390/math9151820.
5. S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, "Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach," International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 5, p. 4392, Oct. 2021, doi: 10.11591/ijece. v11i5.pp 4392-4402.
6. M. IEEE Systems and Institute of Electrical and Electronics Engineers, "ICNSC 2018," Zhuhai, China, Mar. 2018.
7. S. Dhankhad, E. Mohammed, and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," in 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, Jul. 2018, pp. 122–125. doi:10.1109/IRI.2018.00025
8. 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.

9. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International* Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016.
10. H. Mohapatra and A. Rath, Advancing generation Z employability through new forms of learning: quality assurance and recognition of alternative credentials, ResearchGate, 2020.
11. L. Flores, R. M. Hernandez, L. C. Tolentino, C. A. Mendez, and M. G. Z. Fernando, "A Classification Approach in the Probability of Credit Card Approval using Relief-Based Feature Selection," in *2022* 2nd Asian Conference on Innovation in Technology *(*ASIANCON*)*, Aug. 2022, pp. 1–7.
12. L. Bhavya, S. Reddy, A. Mohan, and S. Karishma, "Credit card fraud detection using classification, unsupervised, neural network models," International journal of engineering research & technology, vol. 9, no. 4, pp. 806–810, 2020.
13. K. SAINANI, (2015), Dealing with Missing Data, PM& R 7.9: 990-994, ISSN 1934 1482, https://doi.org/10.1016/j.pmrj.2015.07.011.
14. T. Chou, M. Lo, and M. Lo, "Predicting credit card defaults with deep learning and other machine learning models," *International Journal of Computer theory and Engineering*, vol. 10, no. 4, pp. 105–110, 2018.
15. Y. SAYJADAH, I. A. T. HASHEM, F. ALOTAIBI, & K. A. KASMIRAN, (2018), Credit Card Default Prediction using Machine Learning Techniques, Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 1-4.
16. S. Wang, B. Yuan, and D. Wu, "A hybrid classifier for handwriting recognition on multi-domain financial bills based on DCNN and SVM," *Treatment du Signal*, vol. 37, no. 6, pp. 1103–1110, 2020. [17] J. M. T. Wu, M. E. Wu, P. J. Hung, M. M. Hassan, and G. Fortino, "Convert index trading to option strategies via LSTM architecture," Neural Computing and Applications, pp. 1–18, 2020.
17. J. Dong and X. Li, "An image classification algorithm of financial instruments based on convolutional neural network," Treatment du Signal, vol. 37, no. 6, pp. 1055–1060, 2020.
18. S. L. STOLBA, (2020), Married consumers have higher credit scores and debt than single adults, In Experian.
19. Y. Alghofaili, A. Albattah, and M. A. Rassam, "A financial fraud detection model based on LSTM deep learning technique," Journal of Applied Security Research, vol. 15, no. 4, pp. 498–516, 2020.
20. M. TSUNADA, S. AMASAKI, & A. MONDEN, (2012), HANDLING CATEGORICAL VARIABLES IN EFFORT ESTIMATION, In Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement (ESEM '12). Association for Computing Machinery, New York, NY, USA, 99–102. https://doi.org/10.1145/2372251.2372267s