International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Default of Credit Card Clients Prediction Using ML Algorithms

Rupali Dasarwar[1], Deepali C. Gajbhiye[2]

[1]Assistant Professor, Wainganga College of Engineering & Management, Nagpur, Maharashtra, India
Email: *rupalidasarwar20[at]gmail.com*

[2]M. Tech. Student (AIDS), Wainganga College of Engineering & Management, Nagpur, Maharashtra, India
Email: *deepaligajbhiye06[at]gmail.com*

Abstract: ***This study aims to use a decision tree machine learning model to predict credit card defaults using imbalanced datasets. Imbalanced datasets occur when the number of observations in one class is significantly lower than the number of observations in the other class. In the context of credit card defaults, this means that the number of non - defaulting cases is much higher than the number of defaulting cases. This poses a challenge for machine learning models, as they may be biased towards the majority class. The data used for this analysis includes demographic information and credit card usage patterns of individuals. The decision tree algorithm will be used to train and test the model using this data. The study will first perform an exploratory data analysis, then data will be pre - processed and cleaned before modeling. The model will be trained and tested using different techniques to handle imbalanced data, such as oversampling and under sampling, and the results will be evaluated using metrics such as accuracy, precision, and recall. The goal of this study is to provide insight into the factors that contribute to credit card defaults and to develop a model that can assist in identifying individuals at risk of default, even when the data is imbalanced. By identifying these individuals, lenders can take steps to mitigate the risk of default, such as offering credit counseling or adjusting credit limits. Additionally, the findings of this study could also be used to inform public policy decisions related to consumer credit. In conclusion, this study aims to use a decision tree machine learning model to predict credit card defaults using imbalanced datasets. The results of this study could be used to assist in identifying individuals at risk of default, which would be of great value to lenders and could help to mitigate the risk of credit card defaults. The study will explore the different techniques to handle imbalanced data to improve the model performance.***

Keywords: Accuracy; precision; recall; metrics; Credit card defaults; Imbalanced datasets; Exploratory data analysis; Risk identification

1. Introduction

The problem of credit card defaults is a significant concern for both consumers and lenders. Credit card defaults can have a negative impact on a consumer's credit score and can result in financial hardship. For lenders, credit card defaults can lead to significant financial losses. Identifying individuals at risk of default is crucial for managing credit risk, as it allows lenders to take steps to mitigate the risk of default, such as offering credit counseling or adjusting credit limits. Additionally, understanding the factors that contribute to credit card defaults can inform public policy decisions related to consumer credit.

Machine learning models have been widely used in credit risk analysis, and decision tree is a powerful tool for classification tasks. It is well - suited for this application because it can handle both categorical and numerical variables and provides clear interpretability, allowing for the identification of important factors that contribute to credit card defaults. However, when the data is imbalanced, which means that the number of non - defaulting cases is much higher than the number of defaulting cases, it poses a challenge for machine learning models, as they may be biased towards the majority class. This issue is commonly seen in credit card default prediction, as defaulting cases are often rare events.

To address this problem, this study aims to use a decision tree machine learning model to predict credit card defaults using imbalanced datasets. The data used for this analysis includes demographic information and credit card usage patterns of individuals. The study will explore different techniques to handle imbalanced data, such as oversampling and under sampling, and the results will be evaluated using metrics such as accuracy, precision, and recall.

The goal of this study is to provide insight into the factors that contribute to credit card defaults and to develop a model that can assist in identifying individuals at risk of default, even when the data is imbalanced. By identifying these individuals, lenders can take steps to mitigate the risk of default, such as offering credit counselling or adjusting credit limits. Additionally, the findings of this study could also be used to inform public policy decisions related to consumer credit. By developing a model that can accurately predict credit card defaults using imbalanced data, this study aims to contribute to the field of credit risk analysis and help lenders and policy makers to better manage credit risk.

2. Objective

The main objective of this study is to develop a decision tree machine learning model to predict credit card defaults using imbalanced datasets. The aims of this study can be summarized in following points:

- To investigate the factors that contribute to credit card defaults by analysing demographic information and credit card usage patterns of individuals.
- To use decision tree machine learning model to predict credit card defaults and evaluate the performance of the model using different techniques to handle imbalanced data, such as oversampling and under sampling.

- To compare the performance of the decision tree model with other machine learning models and evaluate the performance of the model using metrics such as accuracy, precision, and recall.
- To provide insight into the factors that contribute to credit card defaults and to develop a model that can assist in identifying individuals at risk of default, even when the data is imbalanced.
- To contribute to the field of credit risk analysis and help lenders and policy makers to better manage credit risk by providing a useful tool for identifying individuals at risk of default.

3. Problem Statement

Credit card default is a major concern for financial institutions, as it leads to significant losses. Default occurs when a credit card holder is unable to make the minimum payment due on their credit card balance. It is important for credit card companies to identify high - risk customers in advance and take appropriate measures to mitigate the risk of default.

The objective of this project is to build a predictive model that accurately predicts the likelihood of credit card default based on the available information about the credit card holder. The model will be trained on a dataset that includes information about the credit card holder's demographic, financial, and credit history. The model will use this information to predict whether a credit card holder is likely to default in the near future.

The predictive model will be evaluated based on its accuracy in correctly classifying the credit card holders into two categories: those who will default and those who will not. The model should also provide a clear and concise explanation of how the predictions are being made, to ensure that the results are interpretable and can be used to make informed decisions.

4. Research and Methodology

Every time a consumer transacts with a bank, vast amounts of data are collected, including everything from demographic information to online history. This information is utilised to assess customer risk and chance of default using data mining techniques including artificial neural networks, linear regression, Naive Bayes, and random forest regression.

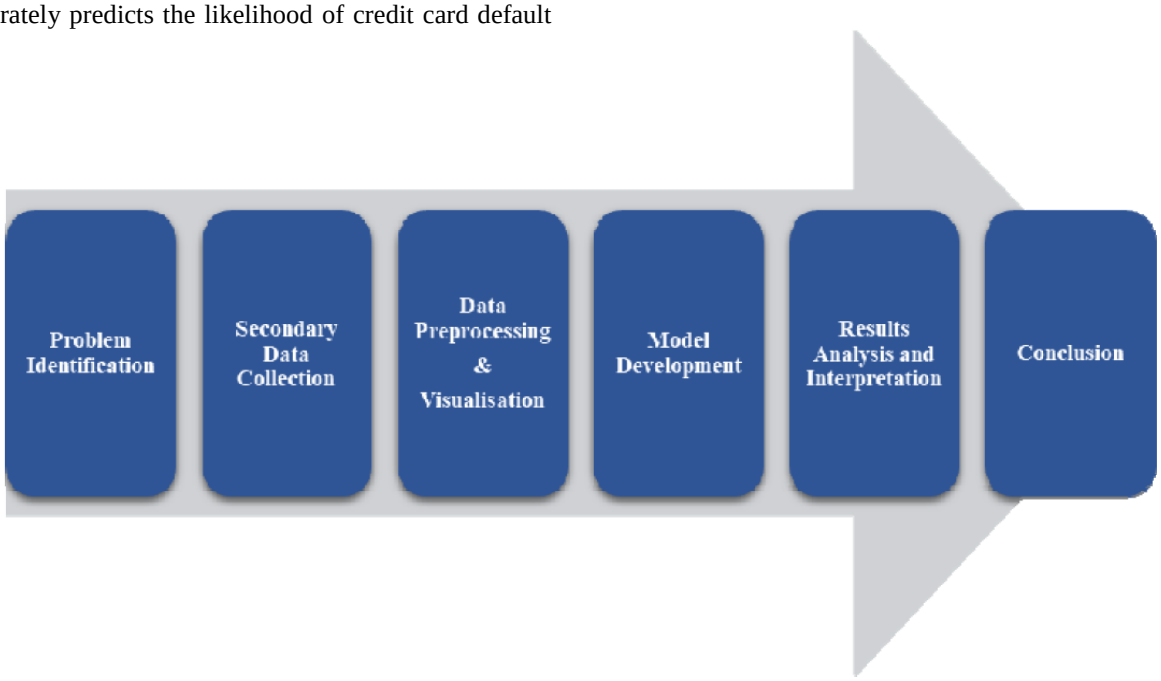

Figure 1: Research Process

a) ***Machine Learning Techniques***

Machine learning techniques are a set of methodologies used for automating analytical model building in data analysis. Within the field of artificial intelligence, machine learning enables systems to learn from data, make predictions, and make decisions without explicit programming. The core objective of machine learning is to develop computer programs capable of accessing data, learning from it, and making predictions or decisions without requiring explicit instructions.

Machine learning encompasses various techniques that can be categorized into two broad categories: supervised learning and unsupervised learning.

1) Supervised learning

Supervised learning is a machine learning approach that involves providing the system with labelled training data to make predictions on unseen data. The objective is to train the system to learn patterns and relationships from the labelled data, enabling it to accurately predict outcomes for new, unseen data. Supervised learning encompasses a range of algorithms, some of the commonly used ones include:

- *Linear Regression:* Based on one or more predictor variables, it is used to forecast a continuous outcome variable.
- *Logistic Regression:* Based on one or more predictor variables, it is used to forecast a binary outcome variable.

- *Decision Tree:* Based on one or more predictor factors, it is used to forecast a categorical outcome variable.
- *Random Forest:* It is an extension of decision trees that uses multiple decision trees to improve the overall performance of the model.

2) Unsupervised learning

Unsupervised learning is a type of machine learning where the system is not provided with labelled data, and the task is to find some structure or pattern in the data. The system is trained to learn from the data, and it will then use this learning to make predictions on new, unseen data. The most common algorithms of unsupervised learning are:

- *Clustering:* The purpose of using this technique is to cluster or group similar objects together based on their shared attributes.
- *Principal Component Analysis (PCA):* It is used to reduce the dimensionality of the data, by finding the principal components that explain most of the variance in the data.
- *Autoencoder:* The autoencoder algorithm is employed to acquire a reduced - dimensional representation of the data. This is achieved by training a neural network to reconstruct the input data, thereby learning the underlying patterns and structure.

b) *Performance Matrix*

There are several metrics used to evaluate the performance of a machine learning model. Some commonly used metrics are:

1) *Accuracy:* This is the most straightforward metric and is calculated as the number of correct predictions made by the model divided by the total number of predictions. The formula for accuracy is:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

2) *Precision:* Precision is a measure of the accuracy of positive predictions, calculated as the number of true positive predictions divided by the sum of true positive and false positive predictions. The formula for precision is:

$$Precision = TP / (TP + FP)$$

3) *Recall (Sensitivity or True Positive Rate):* Recall is a measure of the ability of the model to find all positive instances, calculated as the number of true positive predictions divided by the sum of false negative and true positive predictions. The formula for recall is:

$$Recall = TP / (TP + FN)$$

4) *F1 score:* The F1 score is the harmonic mean of precision and recall, and it ranges between 0 and 1, where 1 indicates perfect precision and recall, and 0 indicates that the model has a low precision and recall. Below is the formula for F1 score:

$$F1\ score = 2 * (Precision * Recall) / (Precision + Recall)$$

5) *ROC - AUC (Receiver Operating Characteristic - Area Under the Curve):* ROC - AUC is a metric used to evaluate the performance of a binary classification model, and it is calculated based on the False Positive Rate (FPR) and True Positive Rate (TPR). The ROC - AUC metric provides a graphical representation of the model's performance by plotting the TPR against the FPR. The formula for the FPR is:

$$FPR = FP / (FP + TN)$$

c) *Prefered Technique*

Decision trees are widely utilized machine learning algorithms capable of addressing both regression and classification problems. The main reasons why decision trees are widely used include:

1) *Interpretability:* Decision trees are easy to interpret and understand, as they provide a clear and concise decision - making process. The structure of the tree is a graphical representation of the relationships between features and the target variable, making it easy to understand the factors that influence the predictions.
2) *Handling Missing Data:* Decision trees can handle missing data by using surrogate splits, which help to estimate the missing data and make predictions.
3) *Handling Non - Linear Relationships:* Decision trees can handle non - linear relationships between features and the target variable, which is important for many real - world problems.
4) *Scalability:* Decision trees are well - suited for handling large datasets and can be efficiently parallelized, making them a viable option for addressing big data challenges.
5) *Handling Imbalanced Data:* Decision trees can handle imbalanced data by adjusting the splitting criteria to favor the minority class, making them a good choice for problems where the data is imbalanced.
6) *Handling Categorical Data:* Decision trees can handle categorical data directly, without the need for feature encoding, making them a good choice for problems where the data is categorical.
7) *Flexibility:* Decision trees can be used for both regression and classification problems, making them a flexible choice for a wide range of problems.

A decision tree is a supervised machine learning algorithm predominantly employed for tackling classification tasks. A decision tree is structured in a manner resembling a flowchart, with each internal node representing a feature or attribute, the branches indicating decision rules, and the leaf nodes representing the outcomes. The initial node at the top of the decision tree is referred to as the root node. It learns to partition the data into subsets based on the values of the input features. The goal of a decision tree is to accurately predict the class label of unseen instances by making a series of decisions based on the values of the input features.

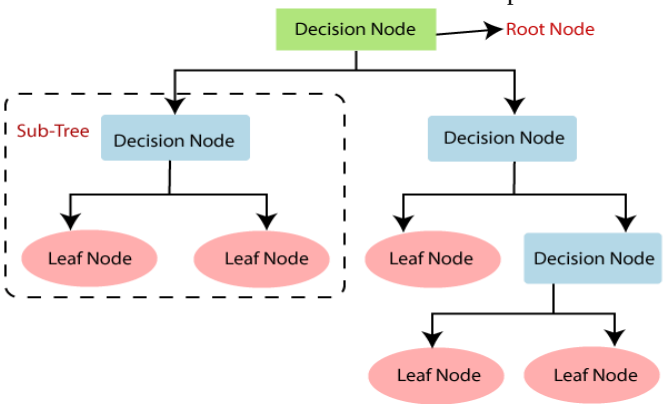

Figure 2: Decision Tree

The basic building block of a decision tree is a decision node and a leaf node. A decision node (also called an internal node) has two or more branches and represents a test on the value of a feature, whereas a leaf node (also called a terminal node) represents a class label. The decision tree algorithm recursively splits the data into subsets and assigns a class label to each leaf node.

The most commonly used algorithm for building decision trees is ID3 (Iterative Dichotomiser 3) which is based on entropy and information gain. The ID3 algorithm uses entropy, which is a measure of impurity in a set of instances, to determine which feature to split the data on at each decision node. The feature with the highest information gain (i. e., the feature that results in the most homogeneous subsets of instances) is chosen as the splitting feature.

Entropy is defined as:
*Entropy (S) = - ∑ (p (i) * log2 (p (i)))*
Where S is the set of instances, p (i) is the proportion of instances of class i in set S, and the summation is over all classes.

Information gain is defined as:
*Information Gain (A) = Entropy (S) - ∑ (|Sv| / |S|) * Entropy (Sv)*

Where A is the feature to split on, S is the set of instances, Sv is the subset of instances for a particular value of feature A, and |S| and |Sv| are the number of instances in set S and Sv respectively.

The ID3 algorithm starts with the root node and recursively splits the data into subsets, selecting the feature with the highest information gain at each decision node, until all instances in a subset belong to the same class, or until a stopping criterion is met.

5. Applications

The prediction of default of credit card clients is a important problem in the field of finance and credit scoring. The applications of this model include:

1) *Credit Risk Management:* The model can be used by financial institutions to predict the likelihood of default for credit card clients and help manage credit risk.
2) *Customer Segmentation:* The model can be used to segment customers based on their likelihood of default, allowing financial institutions to target their marketing efforts more effectively and improve customer engagement.
3) *Loan Approval:* The model can be used by financial institutions to approve or deny loan applications based on the likelihood of default.
4) *Credit Scoring:* The model can be used to generate a credit score for credit card clients, which is an important factor in determining the interest rate and credit limit offered by financial institutions.
5) *Fraud Detection:* The model can be used to detect fraudulent behavior by analyzing the patterns of credit card usage and flagging any anomalies.
6) *Customer Retention:* The model can be used to identify customers who are at risk of default, allowing financial institutions to take proactive steps to retain them and improve customer satisfaction.
7) *Credit Limit Management:* The model can be used to determine the appropriate credit limit for credit card clients, based on their likelihood of default and other factors.

In summary, the prediction of default of credit card clients has a wide range of applications in finance and credit scoring, including credit risk management, customer segmentation, loan approval, credit scoring, fraud detection, customer retention, and credit limit management.

References

[1] Tsungnan Chou and Mingmin Lo, ” Predicting Credit Card Defaults with Deep Learning and Other Machine Learning Models” International Journal of Computer Theory and Engineering, Vol.10, No.4, August 2018
[2] Yang, S. H. and Zhang, H. M. (2018) Comparison of Several Data Mining Methods in Credit Card Default Prediction. Intelligent Information Management, 10, 115 - 122.
[3] Ying Chen and Ruirui Zhang, “Research on Credit Card Default Prediction Based on k - Means SMOTE and BP Neural Network”, Hindawi Complexity Volume 2021, Article ID 6618841
[4] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques, " 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, pp.1 - 4, doi: 10.1109/ICACCAF.2018.8776802.
[5] T. M. Alam et al., "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets, " in IEEE Access, vol.8, pp.201173 - 201198, 2020, doi: 10.1109/ACCESS.2020.3033784.