Predicting default credit card clients

CIND820 Big Data Analytics Project

Student: Chandramathi Vitta (501144398)

Professor: Ceni Babagolu

July 17, 2023

Table of Contents

[Abstract 3](#_Toc140541296)

[Introduction: 3](#_Toc140541297)

[Literature Review: 5](#_Toc140541298)

[Proposed System: 10](#_Toc140541299)

[Dataset: 15](#_Toc140541300)

[Data Dictionary: 16](#_Toc140541301)

[Exploratory Data Analysis 17](#_Toc140541302)

[Model prediction: 24](#_Toc140541303)

[Comparative analysis: 24](#_Toc140541304)

[Conclusion: 25](#_Toc140541305)

# Abstract

Predicting potential default clients is one of the most challenging problems in the financial industry with recent interest rates rising, as it can help banks identify clients that are most likely to default on their credit card products. The failure to make at least minimum payment on your credit card over an extended period results in default. Your credit score is badly affected and the longer you postpone the repayment the worse your credit score will become.

This research is conducted using traditional machine learning classification models, Decision Tree Classifier, Logistic Regression, Random Forest, XG Boost, Support Vector Mechanism, K Nearest Neighbors and compares the classification performance of these 6 classification models. For comparison, model evaluation and selection techniques such as k-fold and cross validation are applied to identify the model that predicts with high accuracy, precision, recall and f1 score. The experimental results show that XG Boost has high AUC value 0.72 when compared with other models. Since the results are not satisfactory and to increase the models performance in prediction, balancing the dataset was necessary, so further experimented applying K-Means smote algorithm to balance the dataset and test the model which resulted in high AUC value 0.91 in Random Forest and XG Boost classifier models. For exploratory data analysis (EDA), data visualization libraries such as Matplotlib and Seaborn are used.

# Introduction:

The federal reserve’s latest interest rate hike will have an impact on credit card debt to become expensive to pay off if the balance is carried month to month. The Fed has continuously raised rates since march 2022 in an effort to combat inflation, since raising rates makes it more expensive for consumers to borrow money. Banks use federal fund rate as a starting point when determining the prime rate, which is the interest rate that’s passed onto consumers. It is usually 3% higher than the federal funds rate. Typically, credit card interest rates are much higher to account for the costs incurred by the card issuer and the risk of some cardholders not paying back their debt. Therefore it is crucial to develop accurate prediction models that can detect clients that could likely go default. Many different fields benefit from the use of machine learning. Different algorithms and, in some cases, statistical models are used in machine learning to enable computers to perform tasks automatically by learning the characteristics of the data. Machine learning methods have since played an important role in automatic fraud detection. However, the performance of machine learning techniques greatly depends on the quality of the training data and the imbalance in the data is not a trivial issue, especially when credit card defaults are considered. In general, only a small percentage of fraudulent transactions are presented in the data. This significantly affects how a trained machine learning algorithm can correctly detect fraud cases. Machine learning techniques are framed for well-balanced training data, thus imbalanced data pose a unique problem to classifier frameworks. However, it is only possible if we remove noise information, lessen the intensity of the imbalance degree, make sure to reduce information loss, and keep sample points which are helpful for the learning of the classifier.

This research paper focuses primarily on three main ideas:

* Apply exploratory data analysis techniques to gain insights into the data to better understand the patterns within the data, detect outliers or find interesting relations among the variables.
* Identifying the key factors that contribute to credit card defaults and train the classification model and achieve better prediction results.
* Explore augmentation techniques to balance the dataset and observe the models performance.

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset contains 25 variables and 30000 records.

Kaggle Datasets -<https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset> (2005).

# Literature Review:

Prediction of Credit card default requires the use of various machine learning techniques. Several publications were reviewed with emphasis being place on determining potential factors that may have significant effects on credit card defaults and analyzing various machine learning techniques to predict accurate results in identifying default clients.

**Research Paper 1:**

Title: "An Investigation of Credit Card Default Prediction in Imbalanced Datasets"

Link: <https://ieeexplore.ieee.org/abstract/document/9239944>

The research explores the challenges and techniques associated with predicting credit card defaults due to imbalanced datasets. Three datasets from different geographical locations related to credit default have been utilized in this study. Credit card default dataset from Taiwan, South German and Belgium have been used and all these datasets are highly imbalanced where the number of non-default instances significantly outweighs the number of default instance.

The authors address the issue of imbalanced data by exploring different data preprocessing techniques, recommends using random undersampling and cluster centroid methods to perform under sampling the majority class. Explains the Issues in oversampling the minority class that results in overfitting instead recommends synthetic data generation using techniques like SMOTE (Synthetic Minority Over-sampling Technique). The authors also evaluate the impact of using different evaluation metrics, such as precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC), to assess model performance.

The findings and recommendations presented in this paper can aid researchers and practitioners in improving credit risk assessment and prediction models in the context of imbalanced credit card datasets.

**Research Paper 2:**

Title: Comparison of Several Data Mining Methods in Credit Card Default Prediction

Link: <https://www.scirp.org/journal/paperinformation.aspx?paperid=87507>

The research discussed in the article focuses on credit card default prediction and explores the effectiveness of various data mining methods in predicting customer default behavior. The study aims to extract useful information from user data to control risks, reduce default rates, and manage non-performing rates.  
 The research utilizes credit card customer data from April to September 2005 in Taiwan. The dataset includes 1 response variable (Default) and 23 explanatory variables (X1 - X23) with a total of 30,000 case data. The authors applied several data mining methods that includes logistic regression, decision trees, artificial neural networks, support vector machine (SVM), Xgboost, and LightGBM and compared the predictions of these algorithms,. They finally found that the correct rate of artificial neural network is slightly higher than the other five methods. The results showed that the accuracy of random forest prediction is higher than that of Lasso-Logistic.

To evaluate the performance of the models, 10-fold cross-validation is conducted. The accuracy rates of the models are above 79%, with LightGBM achieving the highest accuracy of 82.29%. The area under the ROC curve (AUC) is used to assess the classification performance, and LightGBM outperforms the other methods with an AUC of 0.7904.  
 In conclusion, the research demonstrates that data mining methods, particularly LightGBM, can effectively predict credit card defaults indicating a good classification effect. The findings suggest that these methods can assist banks in controlling risks and reducing default rates.

**Research Paper 3:**

Title: Comparison of Data Mining Classification Algorithms Determining the Default Risk

Link: <https://www.hindawi.com/journals/sp/2019/8706505/>

The research focuses on the application of data mining classification algorithms in the banking and insurance industries to predict default risks and improve credit extension processes. The study compares six classification algorithms (Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression) using a dataset obtained from a Turkish survey. The algorithms are evaluated based on statistical criteria such as root mean error squares, receiver operating characteristic area, accuracy, precision, F-measure, and recall. Logistic regression is found to be the best algorithm for predicting default risks.  
 The research aims to identify the demographic and socioeconomic characteristics of individuals that contribute to default risk by using the logistic regression algorithm. The dataset used in the study contains information on household demographics, income, and debt payment history. By applying logistic regression to the dataset, the study determines the attributes that are most likely to increase the default risk of individuals.  
  
 The study emphasizes the importance of selecting the most suitable classification algorithm based on various criteria rather than relying on a single statistical criterion. The findings of this research have implications for the financial industry in terms of improving credit risk prediction and decision-making processes. By understanding the factors that contribute to default risks, financial institutions can take proactive measures to mitigate potential risks and ensure the extension of credit to customers with lower default probabilities.

**Research Paper 4:**

Title: Consumer credit-risk models via machine-learning algorithm

Link: <https://www.sciencedirect.com/science/article/abs/pii/S0378426610002372>

The article titled "Consumer credit-risk models via machine-learning algorithm" published in the Journal of Banking & Finance in 2010 presents a study on the application of machine learning algorithms in developing credit-risk models for consumer lending.  
The author proposes a cardinal measure of consumer credit risk that combines traditional credit factors such as debt-to-income ratios with consumer banking transactions, which greatly enhances the predictive power of the model.   
 The study finds that machine learning algorithms, such as decision trees, neural networks, and support vector machines, outperform traditional statistical models in credit risk prediction. These algorithms demonstrate superior accuracy, flexibility, and robustness in capturing complex patterns and non-linear relationships within the data.  
Furthermore, the authors emphasize the importance of feature selection and variable reduction techniques to enhance the performance of machine learning models. They highlight the significance of incorporating relevant borrower characteristics and loan-specific attributes in credit-risk assessment.  
  
 The research concludes that machine learning algorithms offer a promising approach for credit-risk modeling in consumer lending. By leveraging advanced computational techniques, these models can effectively identify high-risk borrowers and contribute to improved credit decisions. However, the authors acknowledge that model interpretability and regulatory compliance remain important considerations in the implementation of machine learning-based credit-risk models.

Based on the information from the above research papers, I will be applying some of the techniques in my research study ex: handling imbalance dataset, applying classification algorithms and plan to apply some of the advanced algorithms like XGBOOST in my study and evaluate the performance of the dataset in predicting default of credit card clients.

# Proposed System:

In proposed system, the credit card dataset is first pre-processed before applying exploratory data analysis techniques to study the data behavior and patterns. Further dataset is divided into training set and validation set and six different classification algorithms are applied to find the best accuracy in predicting the default candidate for the next month. The classification model is trained using the training set and then on test set to evaluate the model’s performance in accuracy, precision, recall, f1 score and AUC. Further apply augmentation techniques to the proposed model and evaluate the results.

A brief introduction of these techniques as follows:

**Decision Tree Classifier**: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression problems. The goal is to create a tree like model of decisions and their possible consequences. The DT classifier predicts the value of a target variable by learning simple decision rules inferred from the data features. At each node of the tree, the decision tree algorithm selects the best feature to split the data. It evaluates different splitting criteria (eg., Gini index, information gain) to determine the feature that provides the most useful information for classification. Each branch of the tree represents a different outcome of the feature test, leading to a new node or a leaf node. The process continues selecting the best feature splitting the data and creating new nodes recursively until a stopping criterion is met. This criterion could be reaching a maximum depth, having a minimum number of samples in a node, or other predefined conditions. The final nodes of the tree are called leaf nodes. Each leaf node represents a class label, and the samples that reach that node during classification are assigned to that class. The class label assigned to the leaf node is then used as the predicted class label for the input instance.

Decision trees handle both categorical and numerical features and resistance to outliers, however they can be prone to overfitting if the tree becomes too complex and may not generalize well to unseen data.

**Logistic Regression:**

Logistic regression is a widely used statistical technique for binary classification problems, where the goal is to predict the probability of an instance belonging to one of two classes.

The key idea behind logistic regression is to model the relationship between the independent variables (features) and the dependent variable (the binary outcome) using the logistic function (also known as the sigmoid function). The logistic function maps any real-valued number to a value between 0 and 1, which can be interpreted as the probability of belonging to one of the classes.

The logistic regression model estimates the probability of an instance belonging to the positive class (usually labeled as 1) using the logistic function:

P(y=1|x) = 1 / (1 + exp(-(b0 + b1\*x1 + b2\*x2 + ... + bn\*xn)))

Here, b0, b1, b2, ..., bn are the model coefficients learned during training, and x1, x2, ..., xn represent the values of the corresponding features for a given instance.

Once the model is trained, a decision threshold is chosen to classify instances as belonging to one class or the other. By default, a threshold of 0.5 is commonly used. If the predicted probability is above the threshold, the instance is classified as belonging to the positive class; otherwise, it's classified as belonging to the negative class.

**Random Forest:**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random Forest can be particularly effective due to its ability to handle high-dimensional data, capture complex relationships, and mitigate overfitting. Random Forest is an ensemble learning method that combines multiple decision trees. Each decision tree is built by randomly selecting a subset of the training data (bootstrap sampling) and a subset of features at each node of the tree. At each node of a decision tree, the algorithm searches for the best feature to split the data based on a chosen criterion (e.g., Gini impurity or information gain). The split is determined by finding the feature that maximizes the separation of the classes. This process is repeated recursively until a stopping criterion is met (e.g., a maximum depth is reached or a minimum number of samples per leaf node). The number of trees to build is specified beforehand and can be set by the user. The predicted class labels from all the decision trees are combined through voting (e.g., majority voting) to determine the final predicted class label.

**XGBoost:**

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. XGBoost uses decision trees as base learners. A decision tree is a flowchart-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents the outcome or class label. XGBoost employs boosting, which means it trains a series of weak models sequentially. Each model is trained to correct the mistakes made by the previous model. The final prediction is a weighted sum of the predictions from all the weak models. XGBoost has a wide range of hyperparameters that can be tuned to optimize performance, such as the learning rate, maximum depth of trees, number of trees, regularization parameters, and others. Hyperparameter tuning is crucial to achieve the best results.

**Support Vector Mechanism:**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. The goal is to find a hyperplane that separates the data points of different classes. The data points that lie on the margin or violate the margin are known as support vectors. These are the critical data points that influence the position and orientation of the hyperplane. SVM only considers these support vectors during the training process, making it memory efficient. Once the hyperplane is determined, new data points can be classified based on which side of the hyperplane they fall.

SVM has several advantages, including its ability to handle high-dimensional data, its effectiveness in dealing with small datasets, and its resistance to overfitting. However, SVM can be computationally expensive, especially when dealing with large datasets, and the selection of an appropriate kernel function can be a challenging task.

**K Nearest Neighbors:**

KNN classifier is a non-parametric and instance based algorithm used for classification and regression problems. It works by finding the K nearest points in the training dataset and uses their class to predict the class or value of a new data point. K is a user-defined parameter that represents the number of nearest neighbors to consider when making predictions. Selecting the appropriate value for K is important, as it affects the model's performance. A small K may lead to overfitting, whereas a large K may lead to underfitting. It's common to experiment with different values of K and choose the one that yields the best results.

k-Nearest Neighbors (kNN) stores all training samples (including their features and labels) in a space according to its metrics without processing or calculation. When the model receives an object to be predicted, it puts the new object into that space (also according to the metrics). The model then makes prediction by looking at k nearest neighbors to the new object. Usually, the prediction is the label that occurs the most among those k samples. The predicted class label for the new instance is returned as the output of the algorithm.

# Dataset:

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The data set can be found at Kaggle -<https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset> (2005).

Original source - [UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients).

Number of Instances: *30000*

# Data Dictionary:

There are 25 variables and 30000 records:

1. ID: ID of each client
2. LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3. SEX: Gender (1=male, 2=female)
4. EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others)
5. MARRIAGE: Marital status (1=married, 2=single, 3=others)
6. AGE: Age in years
7. PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
8. PAY\_2: Repayment status in August, 2005 (scale same as above)
9. PAY\_3: Repayment status in July, 2005 (scale same as above)
10. PAY\_4: Repayment status in June, 2005 (scale same as above)
11. PAY\_5: Repayment status in May, 2005 (scale same as above)
12. PAY\_6: Repayment status in April, 2005 (scale same as above)
13. BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
14. BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
15. BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
16. BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
17. BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
18. BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
19. PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
20. PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
21. PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
22. PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
23. PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
24. PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
25. default.payment.next.month: Default payment in June, 2005 (1=yes, 0=no)

# Exploratory Data Analysis

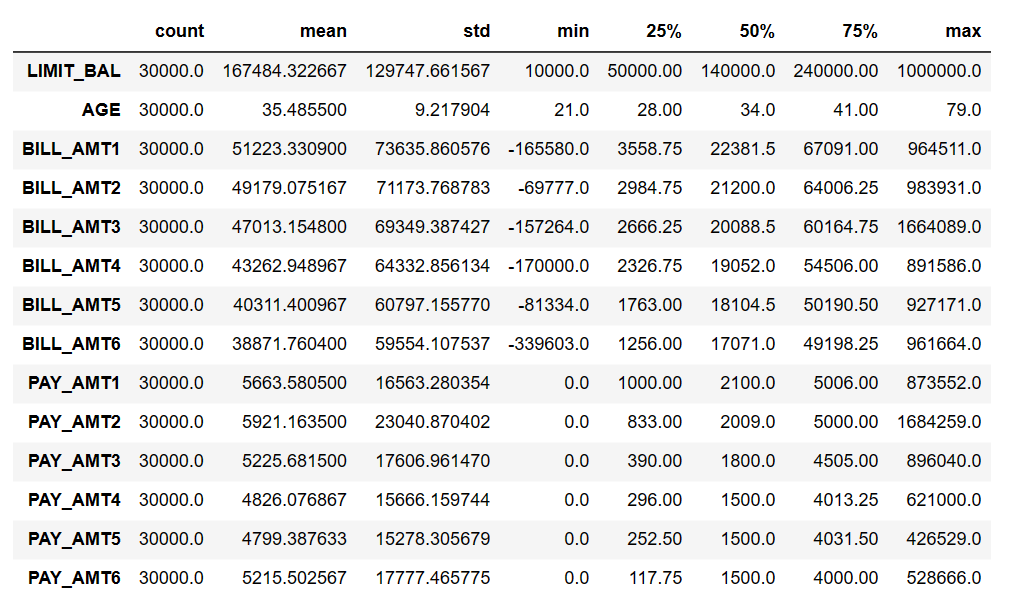
Download the dataset from Kaggle and read the data using python function read.csv. There are no missing values in the dataset. "default.payment.next.month" is a feature and is the target variable.

#### Data Preprocessing:

* Convert categorical data (Sex, Education, Marriage, Pay\_n attributes) to categorical data type.
* Drop attributes of no value – drop first attribute "ID".
* Grouping of undefined categories into unknown category.
  + - Education attribute - grouping 0, 5 and 6 categories to one group ie., unknown (4)
    - Marriage attribute - grouping 0 category to unknown (3)
  + Rename "default.payment.next.month" attribute as ‘default, and "PAY\_0" attribute as “PAY\_1”
  + **Repayment Status (PAY\_X) :** According to the description, this PAY\_X is a set of categorical variables with the levels: -1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above.  Since PAY\_X can take as values only -1,1,2,3,4,5,6,7,8,9 as per the data dictionary - convert -2,-1 values to -1.

**Numerical Data Analysis:**

Conducted 5 number summary on the numerical attributes:



**Data Imbalance:**

Dataset is imbalanced, there are only 6636 defaults in the dataset of total 30000 records.

A black text on a white background

Description automatically generated

**Categorical Data Analysis:**

##### **Social Status Predictors (SEX, EDUCATION, MARRIAGE):**

A graph with blue rectangles

Description automatically generated

1:'Female',2:'Male'

A graph with blue bars

Description automatically generated

1:'GraduateSchool',2:'University',3:'HighSchool',4:'Other'’

A graph with blue bars

Description automatically generated

1: 'Married', 2:'Single', 3:'Others'

**Bivariate analysis:**

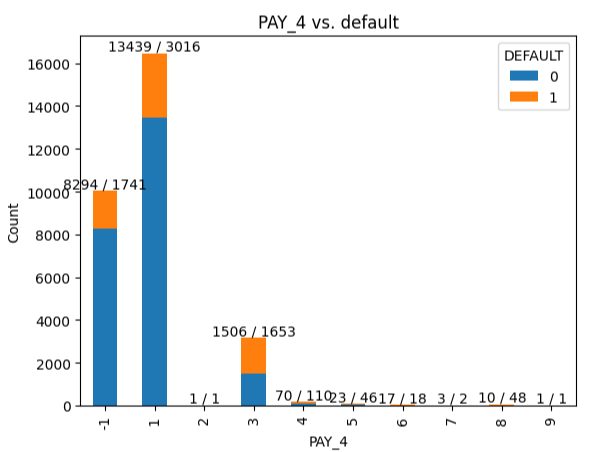
**A graph with numbers and a bar chart

Description automatically generated** **A graph of a graph with numbers and a bar

Description automatically generatedA graph with numbers and a bar chart

Description automatically generatedA graph with blue and orange bars

Description automatically generatedA graph with numbers and a bar chart

Description automatically generatedA graph with numbers and a bar chart

Description automatically generatedA graph with numbers and a bar chart

Description automatically generatedA graph with blue and orange bars

Description automatically generated**

**A graph with numbers and a bar chart

Description automatically generatedA graph of a number of bars

Description automatically generatedA graph of age distribution

Description automatically generated**

**A graph of a graph showing different colored lines

Description automatically generatedA graph of a number of bills

Description automatically generatedA graph of a bill distribution

Description automatically generated**

**A graph of a number of bills

Description automatically generatedA graph of a number of numbers

Description automatically generatedA graph with numbers and a number of numbers

Description automatically generated**

**Outliers:**

A graph with lines and dots

Description automatically generated

**Correlation Matrix:**A table of numbers and a number of numbers

Description automatically generated

We see a high level of linear correlations between the amounts of bill statements in different months. Age, LIMIT\_BAL and PAY\_AMT are less correlated.

**Summary of Bivariate Analysis:**

After exploratory analysis we have a picture about predictors’ impact to response variable:

**LIMIT\_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit)

Impact to default: The lower the amount of given credit limit of the balance owing, the bigger the chances to default. (In general)

**SEX**: Gender (1=male, 2=female)

Impact to default: Male persons have more chances to default. (In general)

**EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others)

Impact to default: The better education the lower chances to default. (In general)

**MARRIAGE**: Marital status (1=married, 2=single, 3=others)

Impact to default: Married persons have more chances to default. (In general)

**AGE**: Age in years

Impact to default: The biggest chance of default is in the age group under 25 and the smallest for 25 - 34 age group.

**PAY\_0..6**: Repayment status in September .. April, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)

Impact to default: Having a delay, even for 1 month in any of the previous months, increases the chance of default.

**BILL\_AMT1..6**: Amount of bill statement in September .. April, 2005 (NT dollar)

Impact to default: The smaller the difference between the amount owed on the bill in September and April, the bigger the chances to default. (In general)

**PAY\_AMT1..6**: Amount of previous payment in September .. April, 2005 (NT dollar)

Impact to default: The smaller the payment amount, the bigger the chance of default. (In general)

# Model prediction:

After performing one-hot encoding on the categorical attributes, create a data frame with the one-hot encoded categorical attributes, and concatenate the categorical, numerical, and target variables into a final pre-processed Data Frame. Build a classification task and compute the feature importance and select the top 5 best features. Split the dataset into training set and validation set on the selected best features by applying k-fold cross validation. Initialize the classification models, Decision Tree Classifier, Logistic Regression, Random Forest, XG Boost, Support Vector Mechanism, K Nearest Neighbors and train the model for each fold, make predictions on the test set and evaluate the performance of the model by calculating accuracy, precision, recall, f1score, ROC-AUC score on each model. Re-train the model by experimenting balancing the dataset using k-means SMOTE and evaluate the performance.

# Comparative analysis:

The results show that the highest AUC value is for XGBClassifier ie., 0.72.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | | **Precision** | | **Recall** | | **F1 Score** | | **AUC** |
|  | **Train Set** | **Test Set** | **Train Set** | **Test Set** | **Train Set** | **Test Set** | **Train Set** | **Test Set** |  |
| **Decision Tree** | 0.99 | 0.71 | 0.99 | 0.35 | 0.97 | 0.36 | 0.98 | 0.36 | 0.6 |
| **Logistic Regression** | 0.81 | 0.8 | 0.69 | 0.66 | 0.27 | 0.27 | 0.39 | 0.38 | 0.69 |
| **Random Forest** | 0.99 | 0.79 | 0.98 | 0.56 | 0.97 | 0.3 | 0.98 | 0.39 | 0.69 |
| **XGBClassifer** | 0.83 | 0.8 | 0.81 | 0.65 | 0.34 | 0.27 | 0.48 | 0.38 | 0.72 |
| **SVCClassifier** | 0.81 | 0.8 | 0.69 | 0.66 | 0.27 | 0.27 | 0.39 | 0.38 | 0.62 |
| **KNNClassifier** | 0.83 | 0.77 | 0.71 | 0.49 | 0.41 | 0.3 | 0.52 | 0.37 | 0.65 |

To further improve the AUC value, and prediction results, model is improved after balancing the dataset using k-means SMOTE algorithm. AUC value increased from 0.72 to 0.91 for XGBClassifier, and 0.90 for Logistic regression and random forest.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | | **Precision** | | **Recall** | | **F1 Score** | | **AUC** |
|  | **Train Set** | **Test Set** | **Train Set** | **Test Set** | **Train Set** | **Test Set** | **Train Set** | **Test Set** |  |
| **Decision Tree** | 0.99 | 0.81 | 0.99 | 0.8 | 0.99 | 0.81 | 0.99 | 0.8 | 0.81 |
| **Logistic Regression** | 0.88 | 0.87 | 0.95 | 0.95 | 0.79 | 0.78 | 0.87 | 0.86 | 0.9 |
| **Random Forest** | 0.99 | 0.87 | 0.99 | 0.92 | 0.99 | 0.8 | 0.99 | 0.85 | 0.91 |
| **XGBClassifer** | 0.89 | 0.87 | 0.96 | 0.95 | 0.81 | 0.78 | 0.88 | 0.86 | 0.91 |
| **SVCClassifier** | 0.88 | 0.87 | 0.95 | 0.95 | 0.79 | 0.78 | 0.87 | 0.86 | 0.88 |
| **KNNClassifier** | 0.89 | 0.85 | 0.93 | 0.89 | 0.84 | 0.8 | 0.88 | 0.84 | 0.9 |

# Conclusion:

In this research paper, various classification techniques were proposed for predicting default credit card clients, and their results were evaluated using k-fold cross-validation methods. However, the initial model predictions did not achieve high accuracy and AUC-ROC scores. To address this, the K-means SMOTE algorithm was applied, resulting in a significant improvement in prediction performance with an accuracy of 0.92. Among the classifiers tested, the XGBoost classifier emerged as a high-performance model for accurate predictions.

Moving forward there are several avenues for further exploration and improvement. Firstly, additional methods for balancing the dataset and applying augmentation techniques can be investigated to further enhance performance. Moreover, conducting further research on hyperparameter tuning is crucial. XGBoost offers a wide range of hyperparameters that can be fine-tuned to optimize model performance, including learning rate, maximum depth of trees, number of trees, regularization parameters, and more. By carefully tuning these hyperparameters, it is possible to achieve even better results in terms of prediction accuracy and overall model performance.

Some of the limitations that can be considered are sample bias, the dataset may not represent the entire population of credit card clients. It is limited by demographic and geographic regions. Dataset might lack certain features that are relevant for credit card default prediction. Additional data or alternative dataset may be necessary to capture a more comprehensive picture of the factors that influence the default behavior. Class imbalance where the number of non-default instances significantly higher and may lead to biased predictions and scores.

**GITHUB REPO**: [chvitta/Project-DataScience (github.com)](https://github.com/chvitta/Project-DataScience)

**References:**

1. Talha Mahboob Alam; Kamran Shaukat; Ibrahim A. Hameed; Suhai Luo (2020) : An Investigation of Credit Card Default Prediction in the Imbalanced Datasets
2. Yang, S. and Zhang, H. (2018) Comparison of Several Data Mining Methods in Credit Card Default Prediction, China
3. Begüm Çığşar1and Deniz Ünal (2018): Comparison of Data Mining Classification Algorithms Determining the Default Risk
4. Amir E. Khandani; Adlar J. Kim; Andrew W. Lo (2010): Consumer credit-risk models via machine-learning algorithms