 DEFAULT OF CREDIT CARD CLIENTS

**By**: Chandramathi Vitta (501144398)

**Supervisor**: Ceni Babaoglu

**Submission** **Date**: 05/05/2023

Table of Contents

[Heading 1 3](#_Toc452457337)

[Heading 2 3](#_Toc452457338)

[Heading 3 3](#_Toc452457339)

[Sample Chart 3](#_Toc452457340)

[Sample Table 3](#_Toc452457341)

# Introduction

The default of credit card clients is one of the major threats faced by the financial industry and risk prediction of credit card clients is growing. Extensive research in this area can help banks and credit card issuers mitigate risks and develop effective strategies to prevent defaults. This literature review aims to summarize by analyzing research papers related to the default of credit card clients.

**Literature Review:**

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 undersampling 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.

**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:

* ID: ID of each client
* LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
* SEX: Gender (1=male, 2=female)
* EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others)
* MARRIAGE: Marital status (1=married, 2=single, 3=others)
* AGE: Age in years
* 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)
* PAY\_2: Repayment status in August, 2005 (scale same as above)
* PAY\_3: Repayment status in July, 2005 (scale same as above)
* PAY\_4: Repayment status in June, 2005 (scale same as above)
* PAY\_5: Repayment status in May, 2005 (scale same as above)
* PAY\_6: Repayment status in April, 2005 (scale same as above)
* BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
* BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
* BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
* BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
* BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
* BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
* PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
* PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
* PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
* PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
* PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
* PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
* default.payment.next.month: Default payment in June, 2005 (1=yes, 0=no)

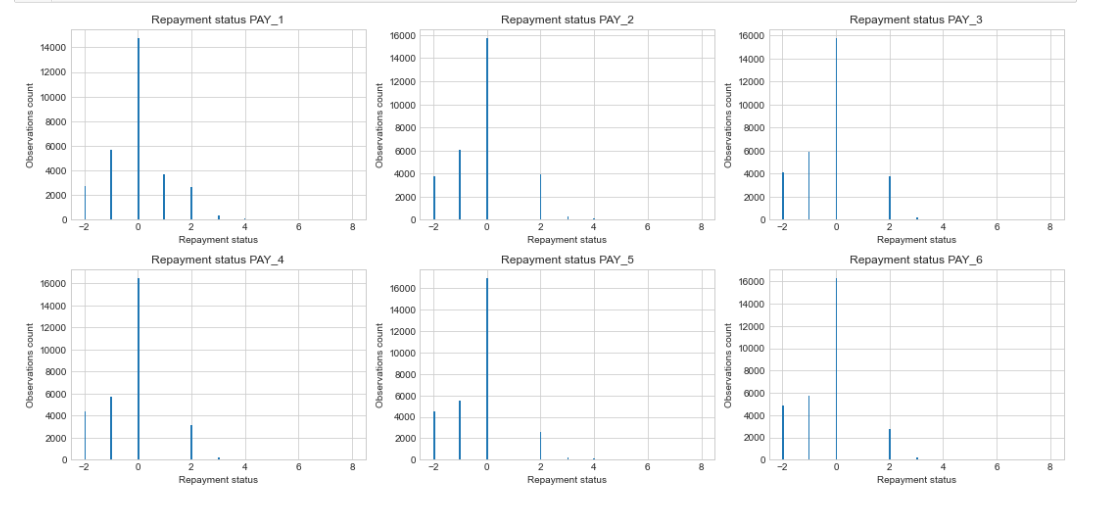
## Importing Data

Download the dataset from Kaggle and read the data using python function read.csv

#### Categorical data:

##### **Repayment Status (PAY\_X)**

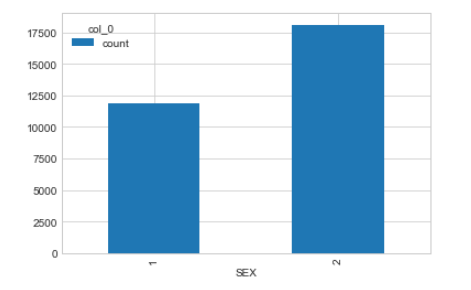
According to the description, this PAY\_X is a set of categorical variables with the levels: -1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above.

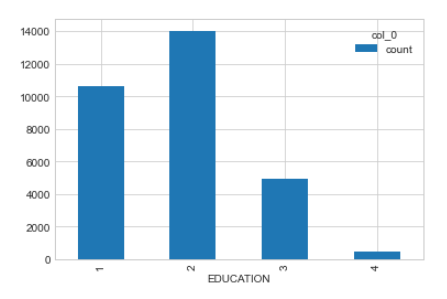


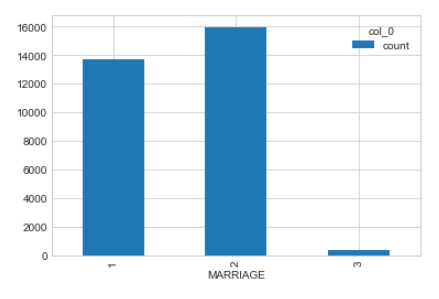
We observe undocumented values for repayment status variables: -2 and 0. Moreover, fraction of it is 86.5%. These could be “**NA**s”. Due to this, there are several standard strategies to deal with:

* Remove observations with NAs. After that we will have 4061 from 30000 observations. Obviously, the loss is too great.
* Replace NAs by major class. But our NA is a major class already. Additionally, it does not look like NAs are random. So, this is a wrong approach in our case.
* I think this is the best decision at this stage of the research and we should keep this data “as is”.

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

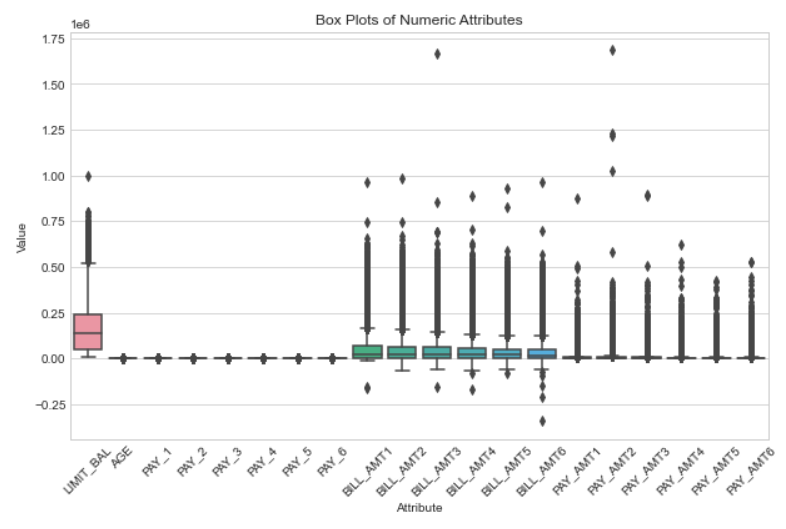






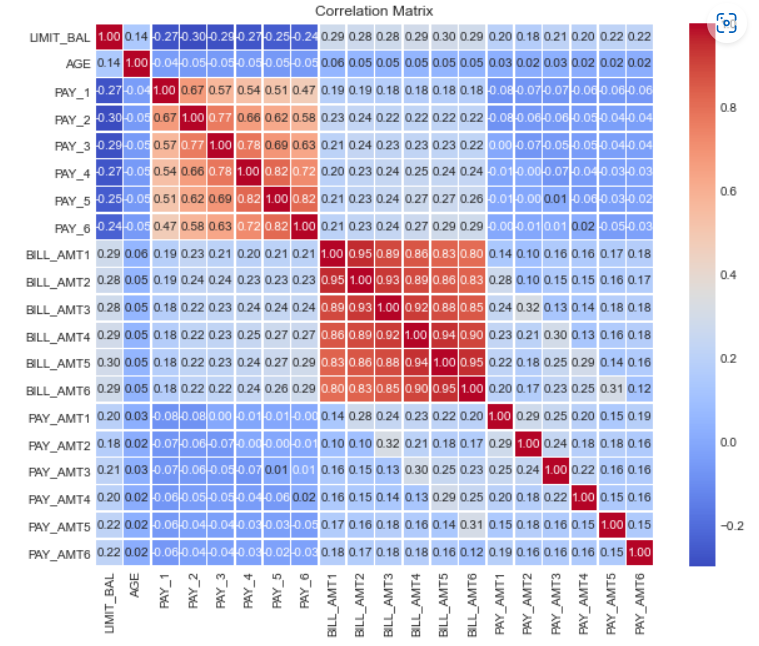
We have some undocumented value 0 for EDUCATION and MARRIGE. I suggest this as NA. Also, let’s rename levels SEX to Male and Female

**Outliers:**



Correlation Matrix:

Numeric attributes:



We see a high level of linear correlations between the amounts of bill statements in different months.

#### Summary

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)

**Flow chart**

**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