

# DATA-2206-01 Capstone

Team Proposal Report

**Credit card payment default prediction**

Professor – Thomas Francescutti

Kanishka Saxena – 100815602

Mayurkumar Chavda - 100839938

Darshanbhai Avaiya – 100848457

Sainadh Chittajallu - 100843160

Date – 4th October, 2022

**Abstract**

This document is our team proposal report for capstone project titled “Credit card payment default prediction”. Most of banks are providing credit cards and line of credit to its customers. Apart from banks, many financial institutions and some retails chains also building tie up with banks and providing credit cards to its customers. Credit limit or line of credit is based upon credit score and usage of any customer but sometimes customer spends too much from their line of credit and at the time of bill payment, they do not have enough balance and they default the payment which is a loss for credit card provider organization and it become non performing assets of the company. Moreover, it can be responsible for raising debt on the company. In this project, we will try to solve this problem and provide a solution using which company will be able to identify potential default payments by any customer in upcoming month.

We will be using the dataset “default of credit card client’s dataset” from UCI machine learning repository. This data was gathered as part of research aimed at the case of customer’s default payments in Taiwan. Data contains 24 variables and 30,000 records. We will use this data in our project to help solving a business problem in financial institution.

It can be useful to the company for risk management. Top management/promotors of any bank or finance company will be the stakeholders and we will be providing this solution to them. After implementing this solution, the company will be able to predict the potential clients who may default their upcoming payments. So, company can be prepared for further actions. This solution can be implemented by other financial institutions as well which provide line of credit to its customers. If company will be able use full potential of solution provided after this project, then there will be significant decrease in company’s default payments of customer and non performing assets of the company.

This solution and recommendations can be useful for any consultancy firm or Information and technology service providers companies as well. Using that they can provide variety of services to their clients in finance sector. It will be helpful to approach the clients in financial needs and provide them appropriate solution to avoid default of credit payment.

**Introduction to Data**

Default of credit card client’s dataset has 24 attributes. One out of 24 is binary variable “default payment”, we can use it as dependent variable. And other 23 variables are amount of the given credit, gender, education, marital status, age, history of past payments for different months, amount of bill statement (NT dollars), and amount of previous payments in different months. We will use these variables as independent variables.

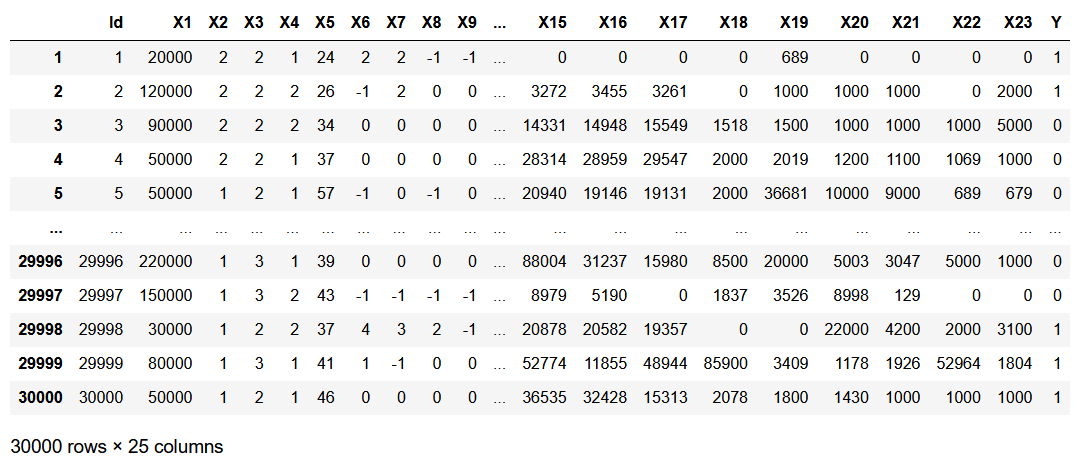
****

Image: Sample of the dataset with 23 independent X variables and 1 dependent variable Y

**Variables:**

**X1:** Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.   
**X2:** Gender (1 = male; 2 = female).   
**X3:** Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).   
**X4:** Marital status (1 = married; 2 = single; 3 = others).   
**X5:** Age (year)

**X6 - X11:** History of past payment (from April to September, 2005)

The measurement scale for the repayment status is: -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.

**X12-X17:** Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

**X18-X23:** Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005.

**Y:** default payment (Yes = 1, No = 0). This will be outcome variable.

**Analytics Approach**

Our analytics approach for this project will be predictive. We will train a binary classification model which will predict whether a customer is going to default next month’s payment or not. Some of variables in dataset are categorical but we can use encoding techniques like one hot encoding to convert it into numerical columns and use it to train a model. We will use Python programming language and libraries like Pandas for data cleaning and manipulation, Scikit-learn to train machine learning algorithm and Seaborn to produce visualizations. If there is any missing data then we will use imputation techniques to fill that data. Using visualizations like density curve and distribution plots, we will know that whether data is skewed or not. If data is skewed then we can use normalization or transformation methods to normalize the dataset.

Feature Engineering will play a vital role in this project because there are 23 features available to use in our dataset. So, team has to identify the relevant features which can be helpful to train a binary classification model. Using feature engineering methods, we can create new features and drop irrelevant features which will be helpful in improving accuracy of algorithm. For binary classification, we can use Logistic regression, Naïve bayes, decision tree, or neural networks. We can try to train all of these algorithms and use whichever fits the dataset best.

This project will be a forward looking one because using the results and implementation of solution, company will be able to improve efficiency and strategy in risk management.

**Implementation**

The solution of this project and algorithm can be embedded into the system of any financial organization. Financial adviser or line of credit authorization officers at bank may access this system and use it to know the probability of client’s payment default for upcoming month. We need more dataset to improve accuracy of prediction using the trained algorithms. So, over the time period team can utilize more dataset to train or reinforce the machine learning model. After gaining satisfactory accuracy of algorithm, employees can use it for real time prediction in company’s system. If the company want to improve the accuracy of prediction, then they must keep collecting new data and use it to train more robust and efficient algorithm. In our team’s opinion they must collect the new data and, in this case, analytics will keep going on.

**References**

1. Data source: <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#>
2. <https://towardsdatascience.com/top-10-binary-classification-algorithms-a-beginners-guide-feeacbd7a3e2>
3. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.