

# DATA-2206-01 Capstone

Group #6 Data Analysis Report

**Credit card payment default prediction**

Professor – Thomas Francescutti

Kanishka Saxena – 100815602

Mayurkumar Chavda - 100839938

Darshanbhai Avaiya – 100848457

Sainadh Chittajallu - 100843160

Date – 27th November, 2022

**Content**

1. **Summary3**
2. **Business Goals4**
3. **Introduction to Data4**
4. **Data cleaning and preparation6**
5. **Predictive model and solution7**
6. **Conclusion9**
7. **Reference10**

**Summary**

Credit card payment default prediction is a data analytics solution to predict potential payment defaults of credit cards payment by customers. We can use dataset of credit cards usage to build a machine learning algorithm which can predict the probability of upcoming month’s payment defaults.

Customers can use the line of credit for a specific time period and then they have to pay bill of credit card. If customer fails to pay credit card bill on time and defaults the payment then bank loses the money. If somehow, we can predict the potential default payments, then organization can take necessary precaution. This solution is completely based on data of credit cards. So, data must be reliable, clean and accurate.

In this credit card payment default prediction project, we will use credit card usage data of Taiwan to train a predictive model and predict payment default. Primary analysis was performed on dataset with some visualizations using Python programming language which is most efficient for data analytics and machine learning. Entire dataset was loaded in Python for analysis and predictive modelling. We divided dataset into training dataset and test dataset. Training dataset is used to train the machine learning algorithm and test dataset is used to validate the performance of algorithm. Different classification algorithms such as Logistic regression, Gradient boosting classifier, Neural networks, Naïve Bayes, and random forest classifier. We finalized an algorithm which gave best prediction accuracy.

This analytics solution 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.

**Business Goals**

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.

Top management/promotors, Analytics team, Financial advisors 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 to 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 data set” was gathered by I-Cheng Yeh from Taiwan. Dataset has 24 attributes. One out of 24 is binary variable “default”, 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 (New Taiwan dollars), and amount of previous payments in different months. We will use these variables as independent variables.

**Independent variables (Input):**

**LIMIT\_BAL:** Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.   
**SEX:** Gender of client (1 = male; 2 = female).   
**EDUCATION:** Education of client (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 – PAY\_6:** 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.

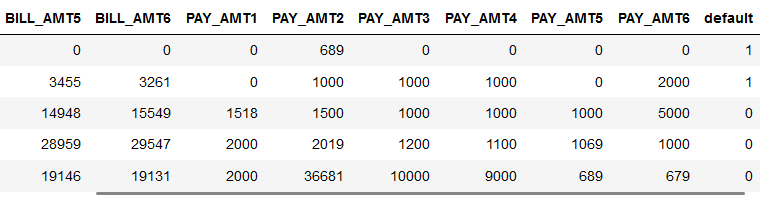
**BILL\_AMT1 - BILL\_AMT6:** Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

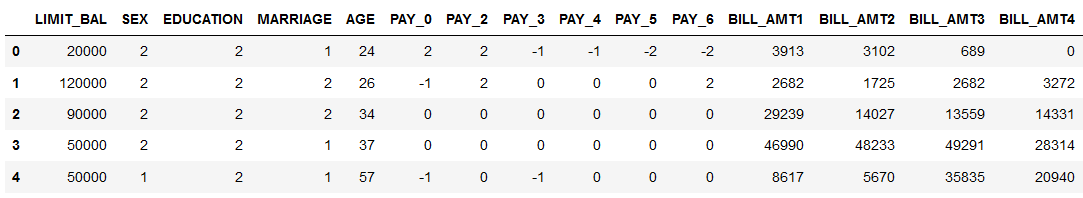
**PAY\_AMT1 - PAY\_AMT6:** Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005.

**Dependent Variable (Output):**

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

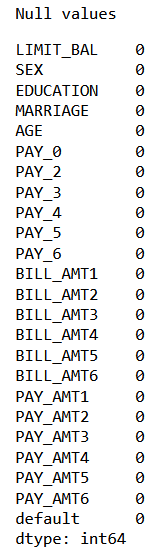
Dataset has 30,000 rows and each rows represents credit card payments, bills and personal data of individual credit card holder.



**Image - 1: Sample of credit card default dataset.**

**Data cleaning and preparation**

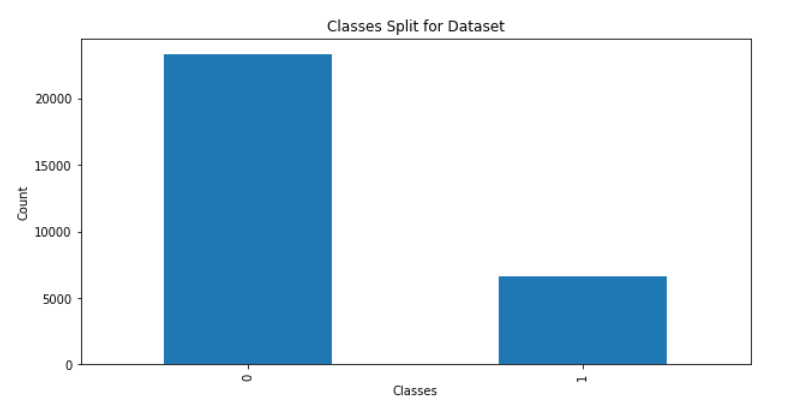
Dataset was downloaded from UCI machine learning repository. Python programming in Jupyter notebook was used to prepare the dataset to train machine learning model.

* Using Pandas library of Python, it was verified that there are no null values in our dataset.

**Image 2: Table showing null values in each column.**

* From dataset attribute information, we found out that education dataset has only four values 1,2,3 and 4 but there were 0,5 and 6 values also present in data which were anomalies.
* So, we dropped those rows with anomalies.

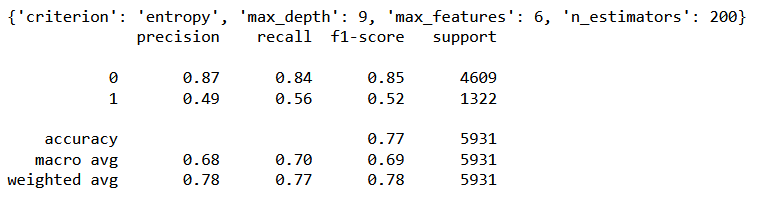
According to primary analysis, data was unbalanced. As there were 23,364 rows with 0 values of “default” variable and 6636 rows with value 1. This could lead to lower accuracy in algorithm. So, we used SMOTE oversampling method to balance the dataset. SMOTE is **an oversampling technique where the synthetic samples are generated for the minority class (“1”)**. This algorithm helps to overcome the overfitting problem posed by random oversampling.

**Image 3: Number of rows with classes “0” and “1” of dependent variable.**

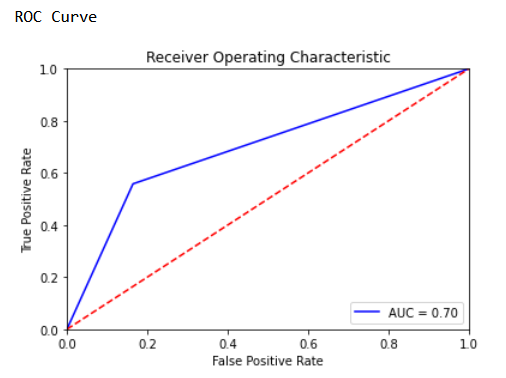
**Predictive Model and Solution**

Random forest classifier model has been trained on the dataset to predict the potential payment defaults of next months. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction.

We will evaluate the model using classification report and AUC/ROC curve.

**Image 4: Classification report of random forest classifier**

* For our model, precision is 78%. It shows, when a model makes prediction, how often it is accurate.
* Precision value is 78% which means 22 out of every 100 default (1) labeled people are not going to default (0) and 76 are going to default the payment next month.
* In our case, recall is 77% which shows our model can predict positive cases correctly 77% of the time.
* Recall value is 77% means that 23 of every 100 potential default payments in reality are missed by our classification model and 77 labeled as default (1).
* Here F1-score is 78% which means model’s predictions exactly matched with classification 78% of the time.
* Accuracy value of 77% means that 23 of every 100 labels is incorrect, and 77 are correct. Which shows our model is performing moderately good and can make good predictions.



**Image 5: ROC curve of random forest classifier**

* **Receiver Operator characteristics (ROC)** curve illustrates the performance of a classification model using two parameters: True positive rate and false positive rate.
* **AUC** is area under the curve which is area between the blue line and axis. The bigger the area covered, the better the machine learning algorithm performs for classification.
* AUC ranges between 0 to 1, predictions are 100% correct if AUC is equal to 1.
* Area under red line is equal to 0.5. and blue line is ROC curve which shows AUC for our model is greater than 0.5
* For our Random Forest classification model AUC = 0.70 which means there is a high chance that the classification model will be able to classify the positive class values from the negative class values.

**Conclusion**

The solution of this project and algorithm can be deployed into the system of any financial organization. Financial advisers 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. The solution can help institutions to reduce its NPA and payment defaults. Nowadays, some retail chains are also providing credits to customers with cashback on shopping offers. To monitor potential payment defaulters, this analytics solution can be helpful to them as well.

**Limitations:**

* Dataset used to train the algorithm was gathers in year 2005 and published in year 2016.
* A lot of systems and services has been changed in recent years in credit card and banking services.
* To keep prediction model efficient and scalable, we need more data to train a robust machine learning mode to make more accurate predictions.
* It is recommended to solution front users that even after the model is deployed into financial organization’s system, enterprises need to keep reinforcing the algorithm with new datasets to improve the accuracy of predictions.

**Reference**

1. Data source: [https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)