Credit card default prediction

Nakshith D N
Data science trainee,
AlmaBetter, Bangalore

Abstract

Despite machine learning and big data having been adopted by the banking industry, the current applications are mainly focused on credit score prediction. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history. According to a literature review report on analysing credit risk using machine and deep learning models, “credit risk management problems researched have been around credit scoring; it would go a long way to research how machine learning can be applied to quantitative areas for better computations of credit risk exposure by predicting probabilities of default.” The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

Problem Statement

The objective of the project is to perform exploratory data analysis, data pre-processing, data cleaning & imputation, and in the end, apply different Data Visualisation techniques to get meaningful insights and apply different models to predict from the given data. This project is aimed at predicting the case of customers' default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Introduction

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can be proven by industry business reports and statistical data. So, assessing,detecting and managing default risk is the key factor in generating revenue and reducing loss for the banking and credit card industry.

Prediction and model building

Here the main purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

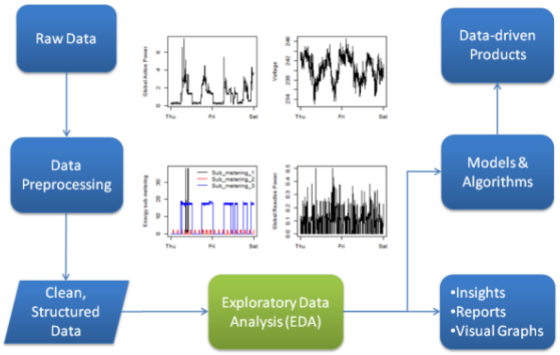
Dataset

Understanding the Dataset can refer to a number of things but not limited to extracting important "variables", identifying

"outliers", "missing values", or "human error". Ultimately, maximising our insights of a dataset and minimising potential "errors" that may occur later in the process.

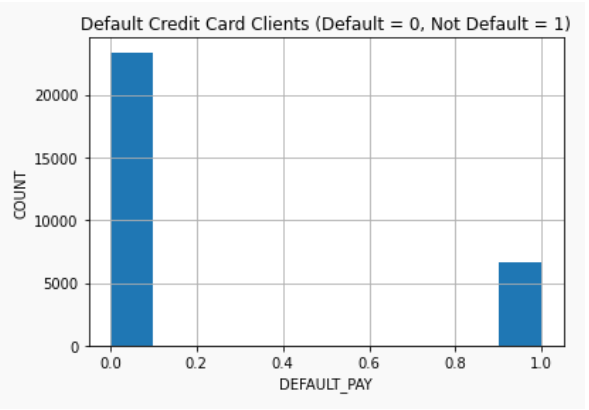
This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information. Data dictionary is available in Appendix A. Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns. More details about the data cleaning can be found in the colab notebook.

Architecture



Data Exploration and visualisation

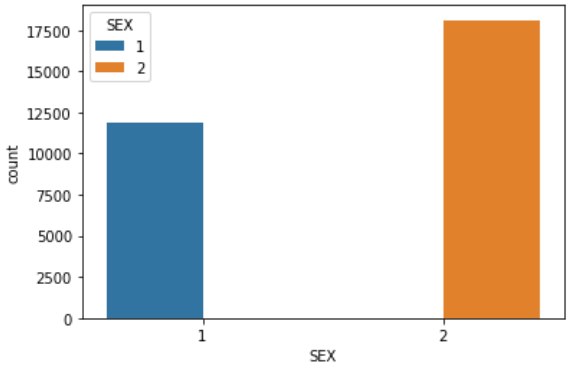
The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualisation and is followed by a statistical test to verify the findings.



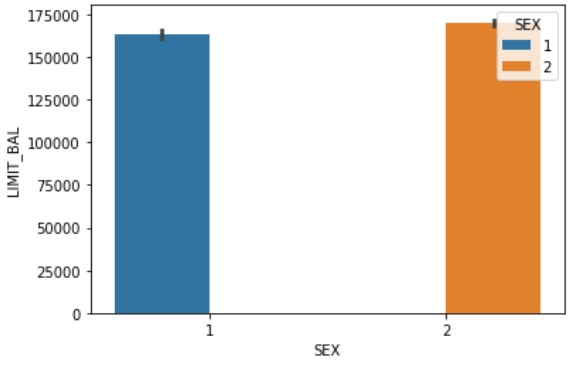
Here we can clearly observe that the percentage of Defaulters are smaller than the Non Defaulters in the given dataset.

Considering gender feature

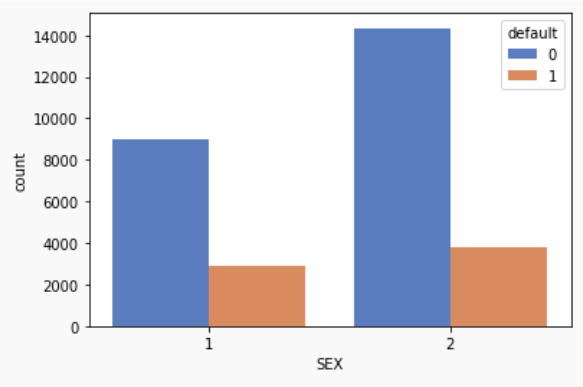
(Here 1- male , 2- female)



In the above figure we can clearly observe that the total female count is higher compared to men, from this we can conclude that females had higher count of credit compared to men.



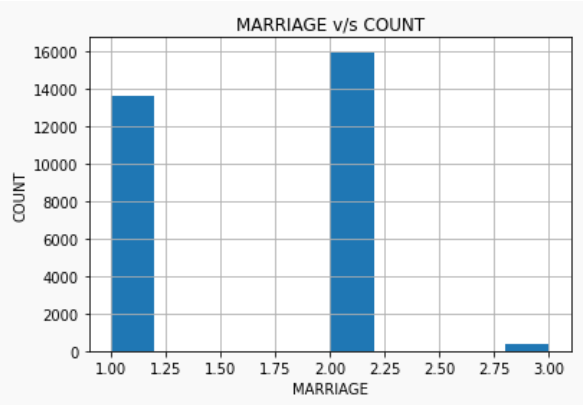
As explained above females have a higher limit balance compared to men and because of their count.



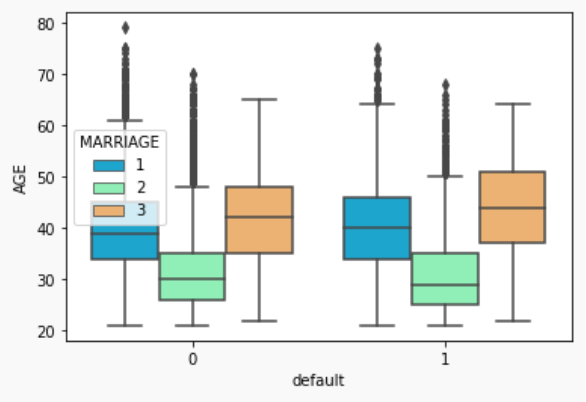
From this above figure we can observe that the total number of counts of defaulters and non-defaulters is sexwise and we can observe that women have overall less default payments compared to men and we can also observe that non defaulters have a higher proportion of females.

Considering marriage feature

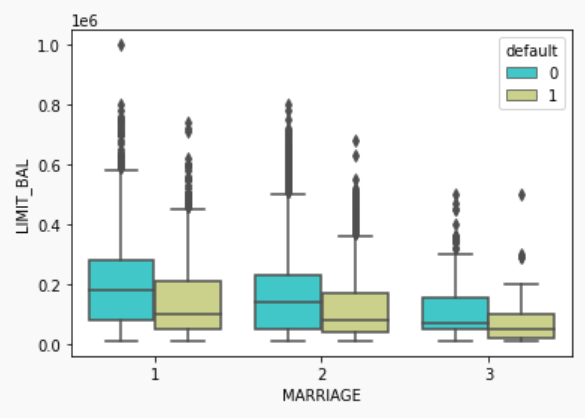
Marital status (1=married, 2=single, 3=others)



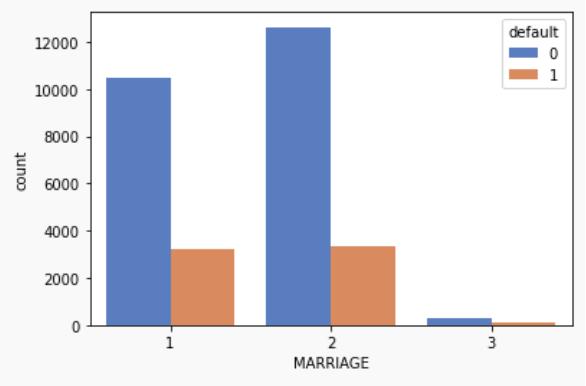
From this figure we can clearly observe that the total number of married people who have taken credit is lower compared to singles.



This plot shows the age group of marriage feature with default and non default and here we can clearly observe that the married people of age group of around 40 years have higher default and singles of age group around 30 years have defaults.

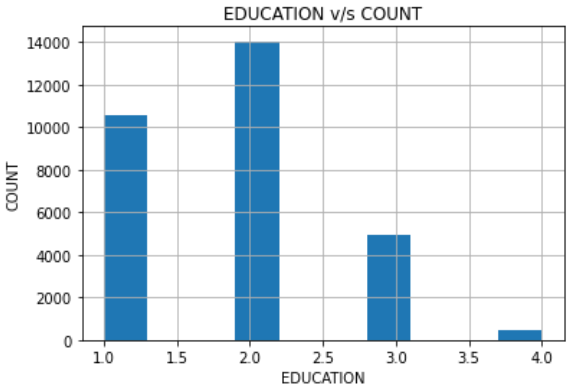


This plot shows the total credit balance by the specific marital status and we can clearly observe that when the credit limit is very low the default is higher but when there is higher credit limit then the default is low, from this graph we can observe that married and single people when they have low limit balance tend to default.

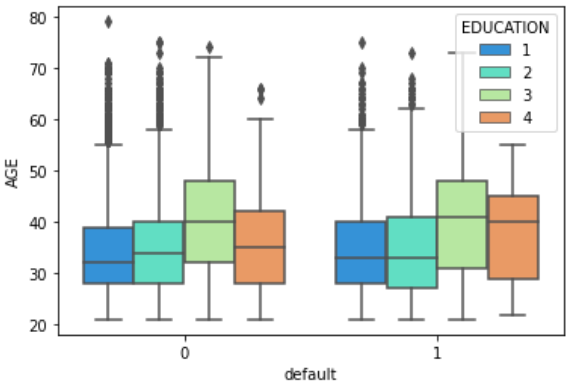


From the above plot it is clear that those people who have marital status single have less default payment wrt married status people.

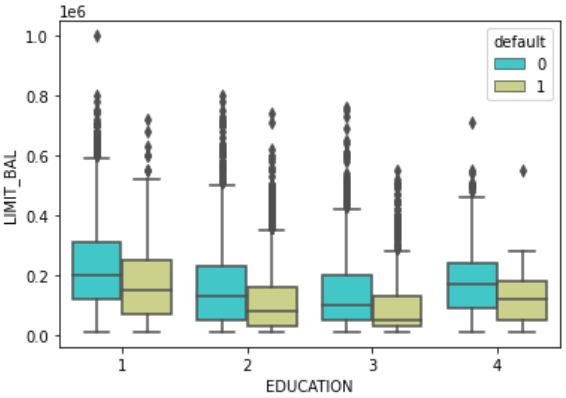
<u>Considering education feature</u> Education (1 = graduate school; 2 = university; 3 = high school; 4 = others)



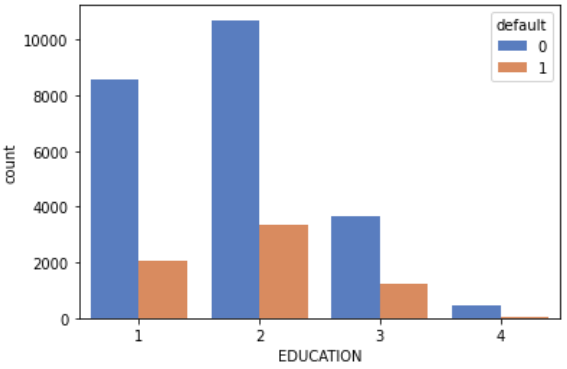
From this above visualisation we can see that more credit holders are university students followed by Graduates and then High school students.



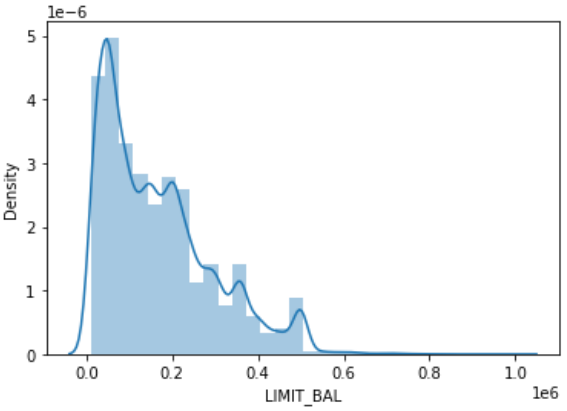
Here we can clearly observe that people with an age group of around 40 years with high school education have higher default compared to graduate and university level people.



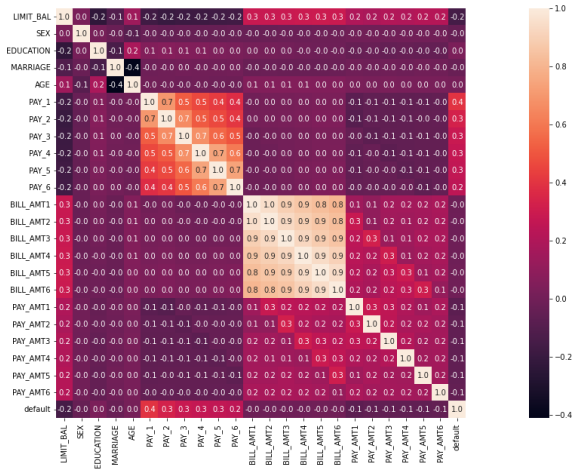
It can be clearly seen that with higher credit balance there is low default but when there is low credit with respect to high school level people the default is present.



From the above graph we can clearly observe that the proportion of non default is higher with graduates and university level people compared to high school level people.



From this plot it is evident that the low amount of credit is provided more whereas the higher credit amount is provided very less.



So from this correlation matrix it looks like the PAY_0, PAY_X variables are the strongest predictors of default, followed by the LIMIT_BAL and PAY_AMT_X variables.

The main findings from exploratory analysis are as following:

- Males have more delayed payment than females in this dataset. Keep in mind that this finding only applies to this dataset, it does not imply this is true for other datasets.
- Customers with higher education have less default payments and higher credit limits.
- Customers aged between 30-50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. However, the delayed rate drops slightly again in customers older than 70.
- There appears to be no correlation between default payment and marital status.
- Customers being inactive doesn't mean they have no default risk. We found 317 out of 870 inactive customers who had no consumption in 6 months then defaulted next month.

Feature Scaling of Numerical Attributes

In this for easier operation we have converted the columns to lower cases and more importantly we have normalised all the numerical values using the equation

(x-np.mean(x))/np.std(x))

After normalisation we have splitted the data into training and test data parts where around 80% of data is training data and around 20% is test data.

Applying Machine Learning Algorithm for Classification Problem

Model building

The major steps involved in model building are,

- Data preprocessing - Feature selection
- Feature engineering
- Train test data split
- Training data rescaling
- ❖ Start with default model parameters
- ❖ Measure roc_auc on training data
- ❖ Model testing and prediction
- ❖ Precision - recall scores

First we applied logistic regression model

Logistic Regression

The Logistic Regression model has been applied to both training and test datasets with default parameters and we have also observed that accuracy obtained was around 0.814167, whereas the precision for the model was obtained around 0.678457, the recall was around 0.315632. The F1 Score that was achieved was around 0.430832 and the ROC was obtained around 0.636371.

To increase the model accuracy we tried a random forest model.

Random forest model

The Random forest model has been applied to both training and test datasets with default parameters and we have also observed that accuracy obtained was around 0.815167, whereas the precision for the model was obtained around 0.661473, the recall was around 0.349289. The F1 Score that was achieved was around 0.457171 and the ROC was obtained around 0.649017.

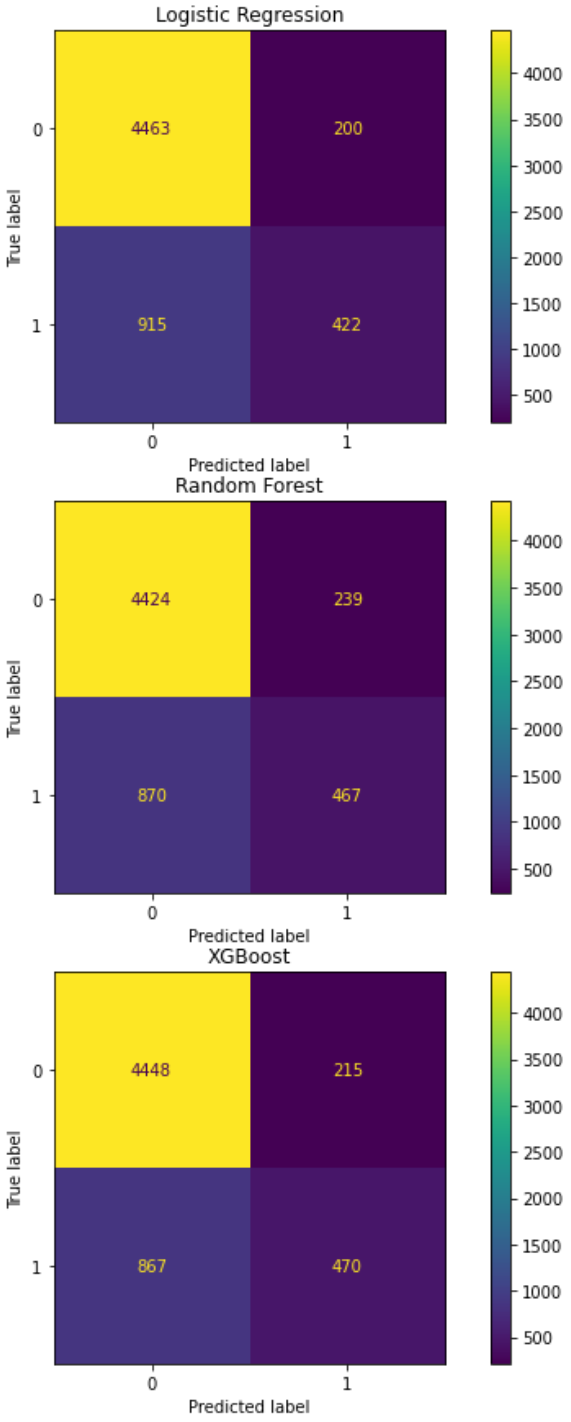
To further increase the model accuracy we have applied the XG boost model and have obtained the following results.

XG boost model

The XGboost model has been applied to both training and test datasets with default parameters and we have also observed that accuracy obtained was around 0.819667, whereas the precision for the model was obtained around 0.686131, the recall was around 0.351533. The F1 Score that was achieved was around 0.464866 and the ROC was obtained around 0.652713.

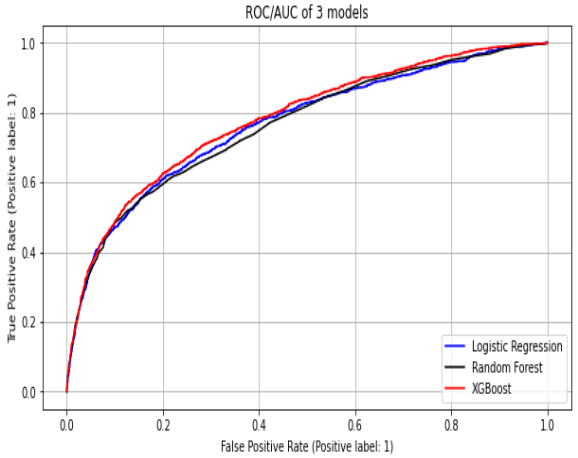
Model comparison

Using confusion matrix



Here we have comprehensively observed the differences between all the models using a confusion matrix.

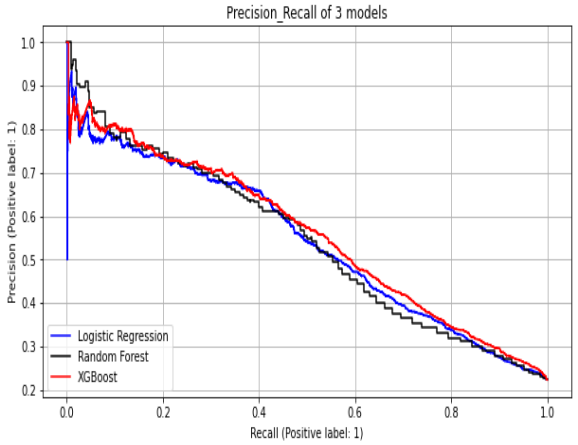
Using ROC_AUC curve



Here we can clearly observe the roc_auc curve of the models and we can clearly observe the model differences with the obtained curves.

Using precision recall curve

Since the classes are imbalanced, the precision_recall curve is more appropriate.



Here we have compared the three models using the precision and recall curves.

Conclusion

The exploratory data analysis and modelling for Credit card default prediction dataset has been successfully done and the following inferences have been made from the obtained visualisations and the predictions from the dataset,

1)Using a Logistic Regression classifier, we can predict with 81.4% accuracy, whether a customer is likely to default next month.

2)Using a random forest classifier, we can predict with 81.5% accuracy, whether a customer is likely to default next month.

3)Using a XG boost classifier, we can predict with 81.9% accuracy, whether a customer is likely to default next month.

The best predictions are made and the results are obtained.