Credit Card Default Prediction

- by Rahul Deshmukh

Points To Discussed:

-Data Summary
-Data Cleaning
-Defining problem statement
-EDA
-Feature Selection
- Feature Engineering
-Preparing dataset for modeling
-Applying Model
-Model validation and selection
-Conclusion

Problem Statement:

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. We have to evaluate which customers will default on their credit card payments.

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faced by commercial banks is the risk prediction of credit clients.

To analyse and predict the above given database, the current project is developed. This project is an attempt to identify credit card customers who are more likely to default in the coming month

Steps involved in supervised ML model:

1. Defining the problem statement
2. Pre-processing the data
3. Splitting the data into train and test data
4. Training the model
5. Evaluating the model
6. Improve the model
7. Deploy the model and monitor real-time

Data Summary/Data Cleaning

- Data processing-In this first part we've removed unnecessary features.

Shape of data
Rows-30000,
columns-25

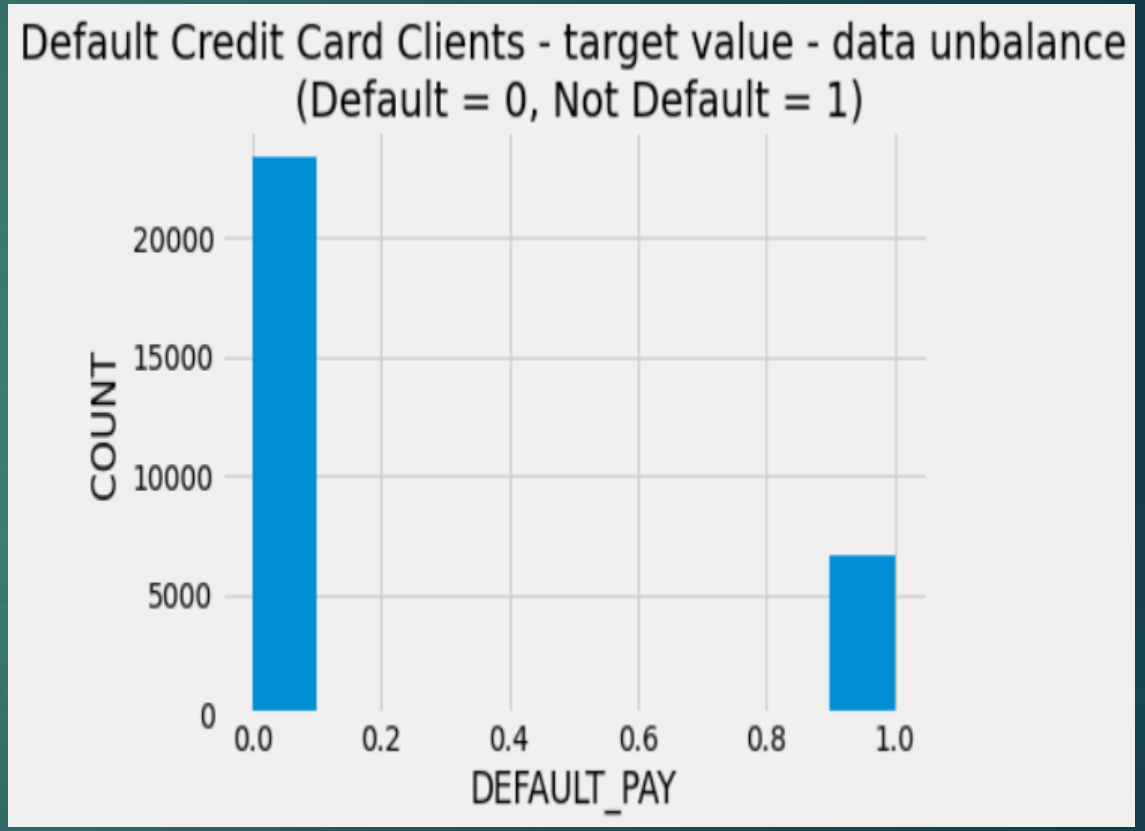
Drop column
dropping the unwanted columns or the column containing constant value
Renaming columns.

Dataset
Fully cleaned data

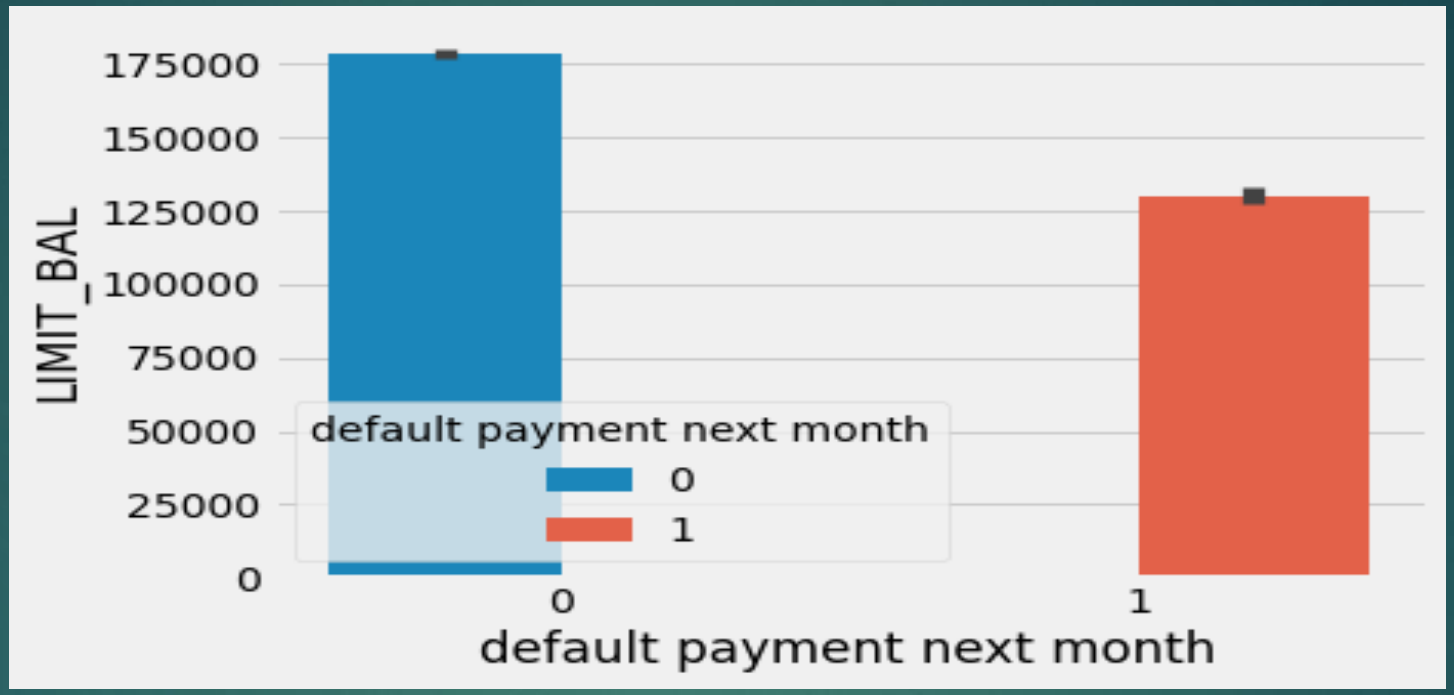
EDA(EXPLORATORY DATA ANALYSIS):

Analysis of different variables:. Analyzing the payment method

From the above graph on x-axis 0 indicates as not a default payment and 1 indicates the default payment. From this we can say that for more customers there are no default payments for next month



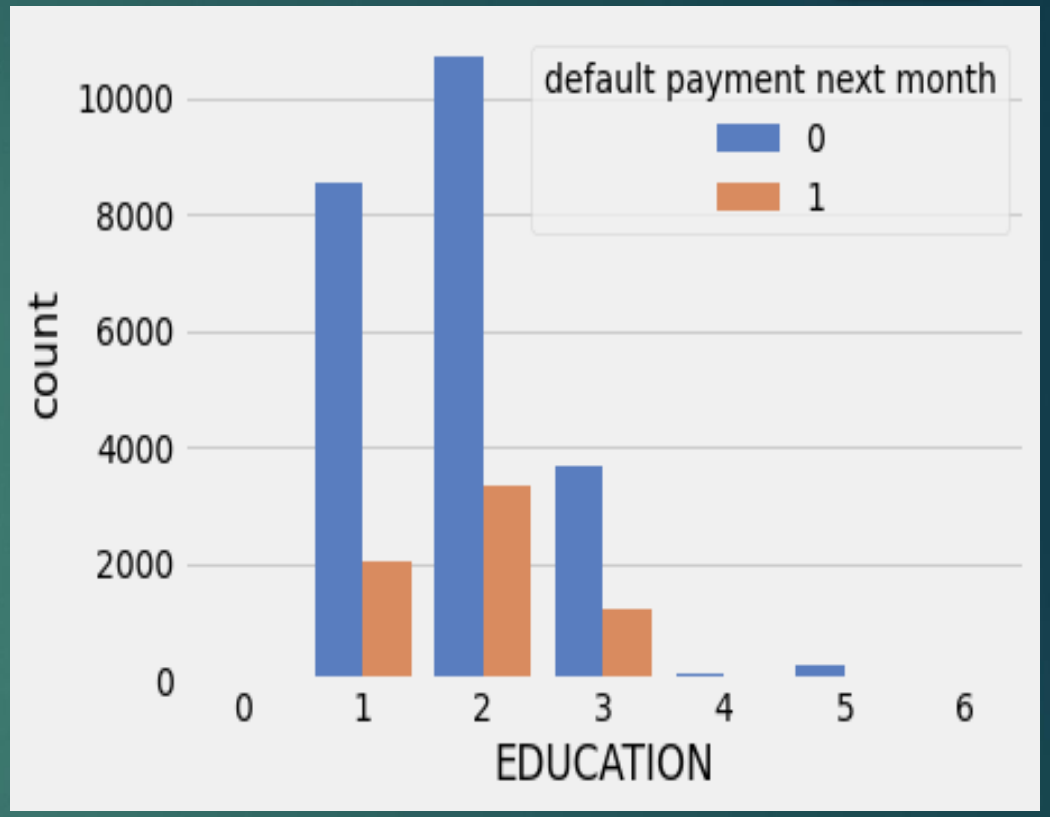
Analysing the default payment next month v/s limit balance



LIMIT_BAL Amount of the given credit (USD): it includes both the individual consumer credit and his/her family (supplementary) credit. Including this variable in the study is important as the credit line of a customer is a good indicator of the financial credit score of the customer. Using this variable will help the model predict defaults more effectively.

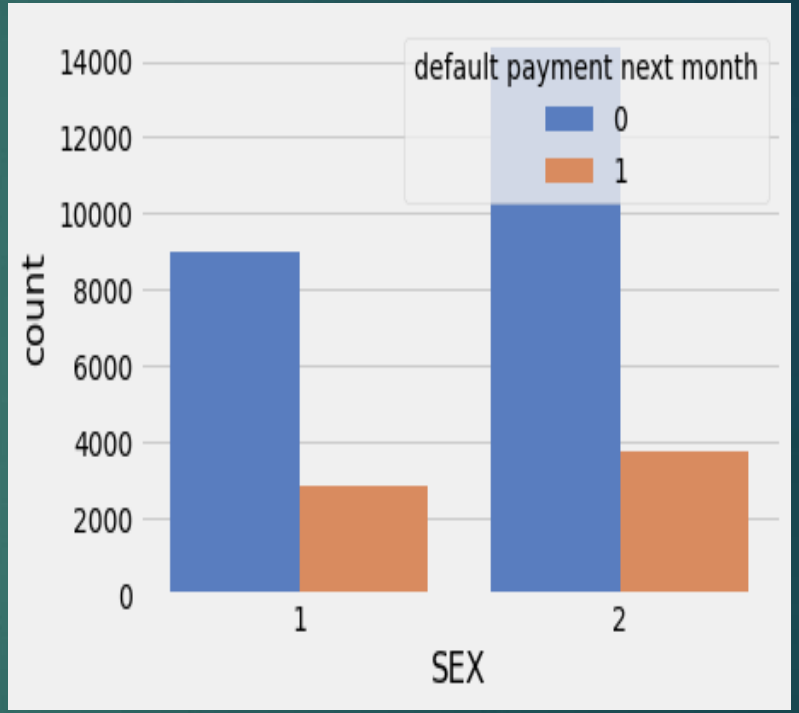
Analysing the customers based on their Education

Education (1 = graduate school; 2 = university; 3 = high school; 4 = others) It might be useful to see whether the education level of the customer is in any way related to his/her probability of default. The distribution of defaults based on education level will be an interesting chart to look at. From the above we can say that most of the people are university educated followed by graduated school More number of credit holders are university students followed by Graduates and then High school students



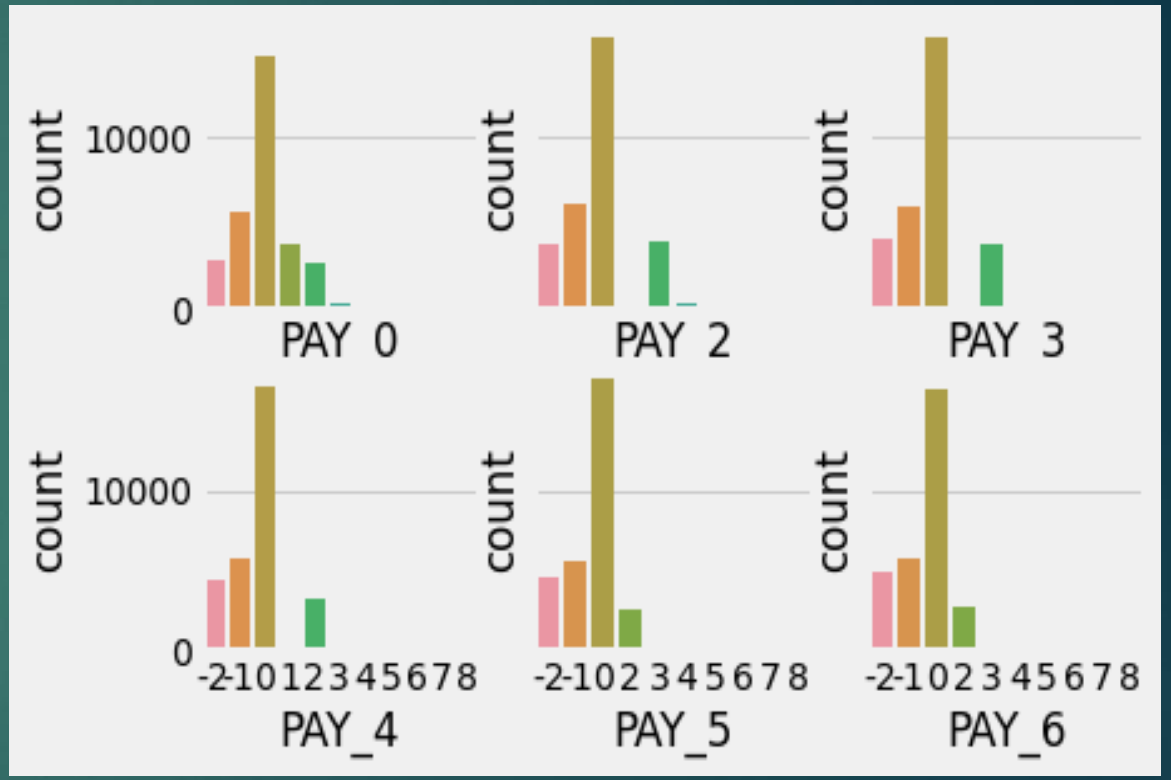
Analysing the LIMIT_BAL, SEX and the default payment for next month variables

Gender (1 = male; 2 = female) It might be useful to see whether the gender of the customer is in any way related to his/her probability of default



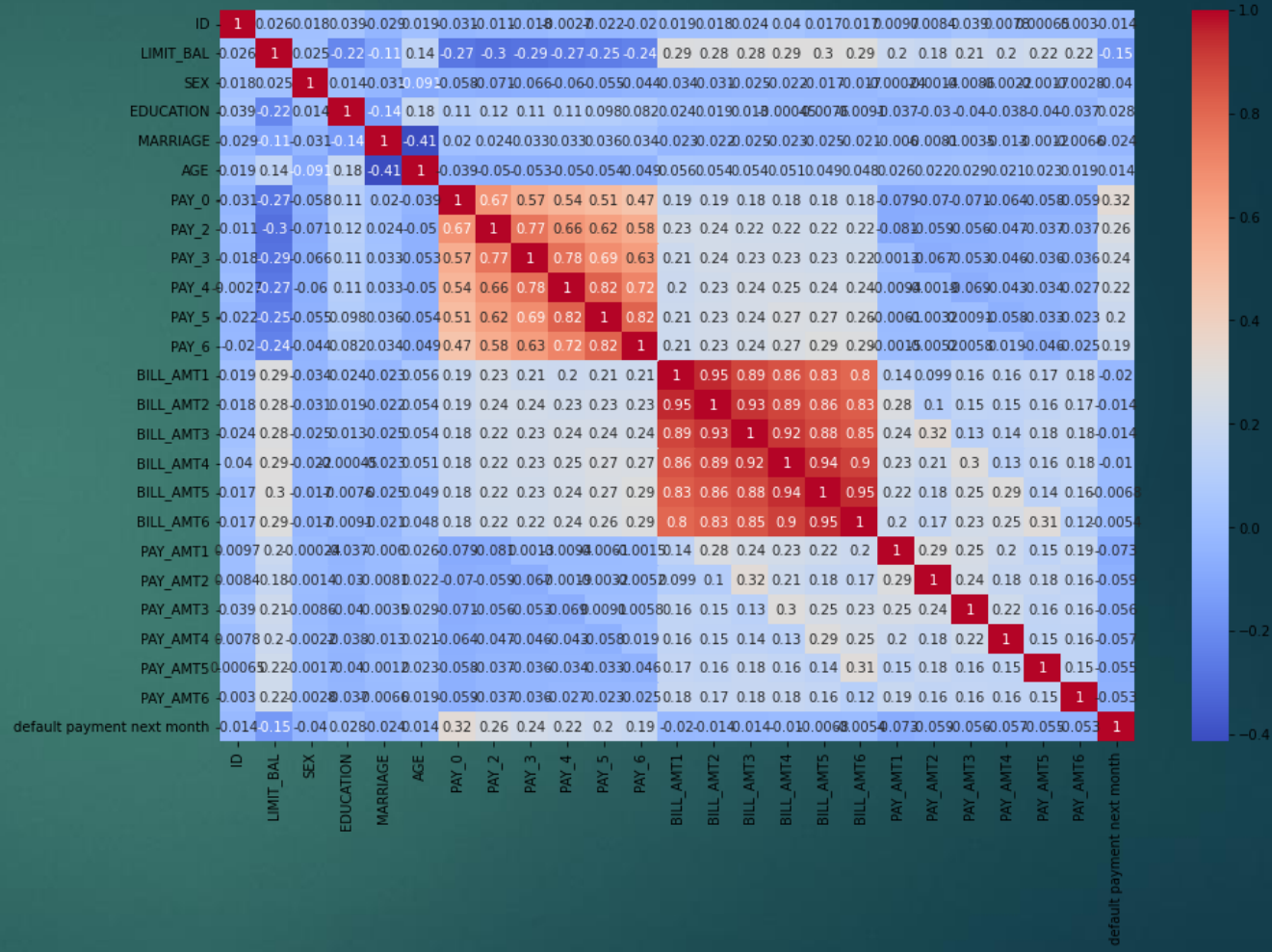
Analyzing the history of payment

History of past payment. Customers' past monthly payment records (from October 2015 to March, 2016) were tracked and used in the dataset as follows: The measurement scale for the repayment status is: -2 = Minimum due payment scheduled for 60 days -1 = Minimum due payment scheduled for 30 days 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months and above; This information is very crucial as it directly provides the payment status of the customer for the past 6 months. Using these variables will train the model efficiently to predict defaults.



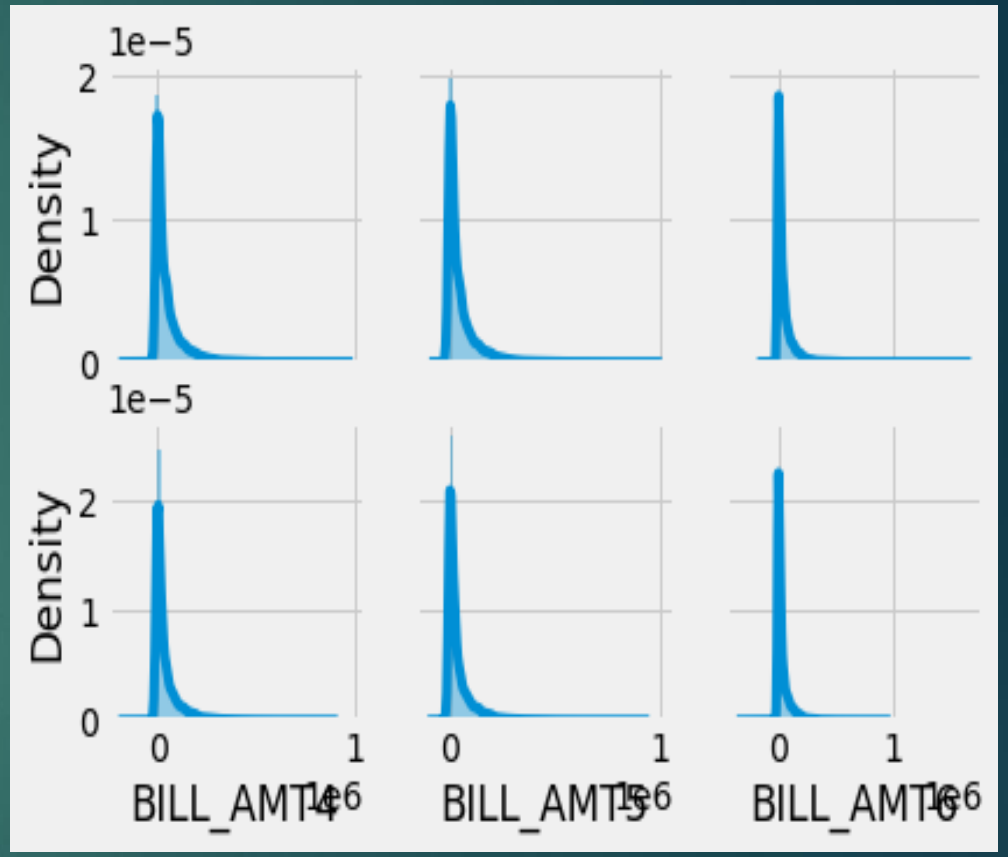
Correlation

The above figure is an indication about the changes between two variables. We can plot correlation matrices to know which variable is having a high or low correlation in respect to another variable.



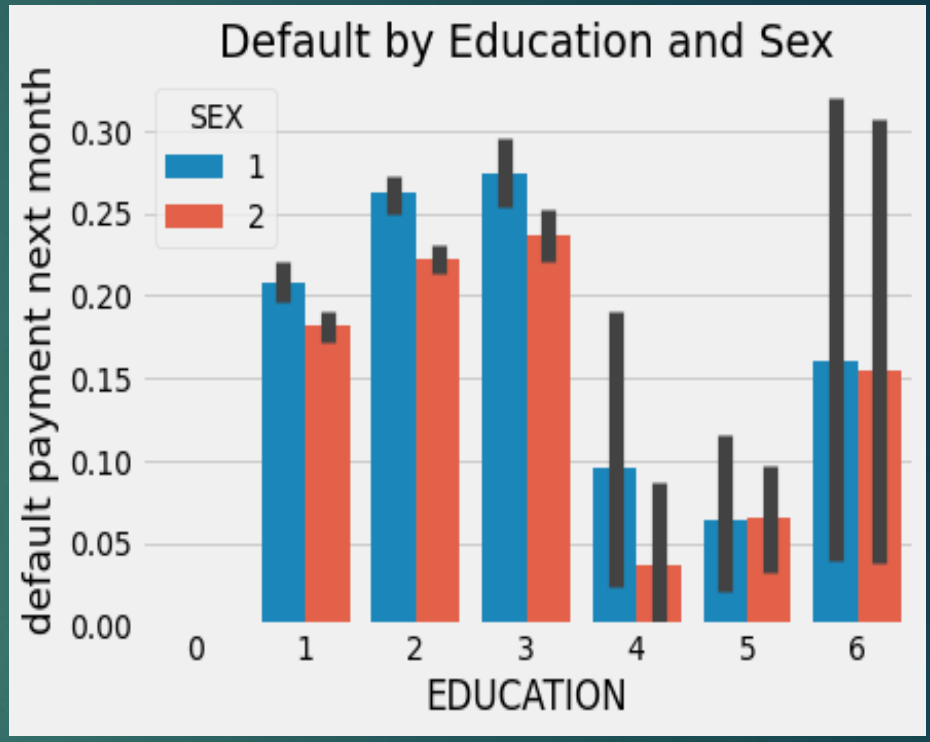
Analyzing the bill statements of the customers

Actual bill statements of the customers for the past 6 months would give a quantitative estimate for the amount spent by the customer using the credit card. Amount of USD paid by the customers in the past 6 months would give the repayment ability of the customer and the pattern for payment could be used to train the model efficiently. We found the 'Bill_AMT' variables contain negative values in the case a customer overpays their bill. This could be caused by an automatic payment set if the bill for that month is not as high as the automatic payment is set for.

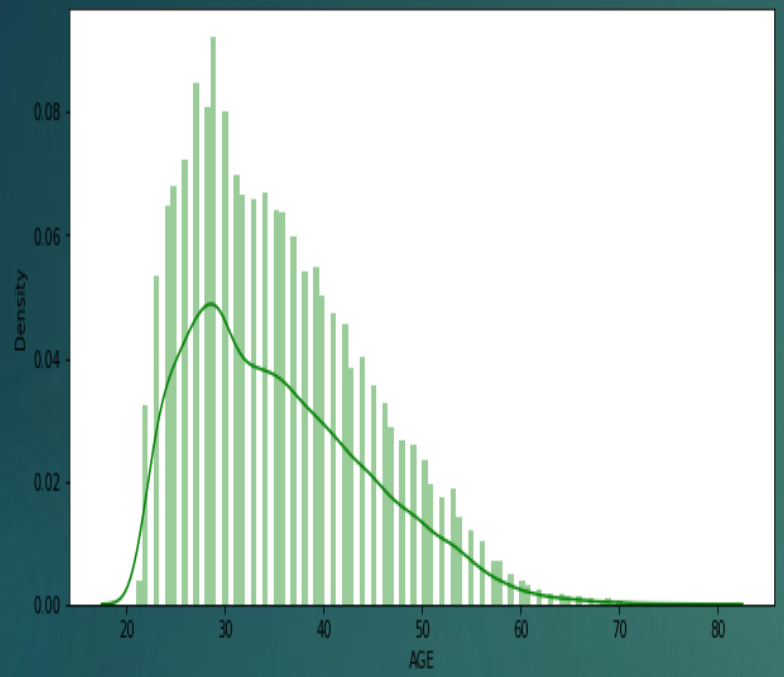


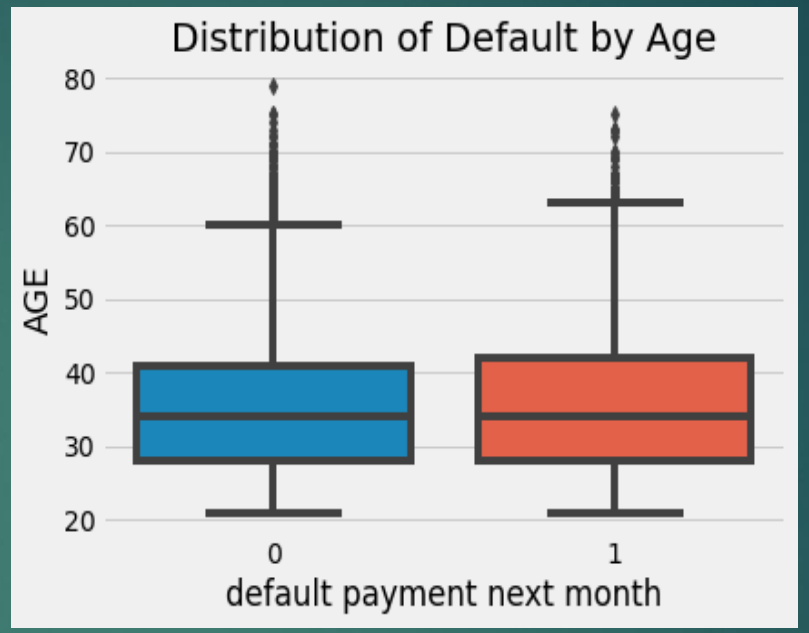
Analyzing the Sex and education variable

As far as education goes there seems to be pretty equal likeliness for each class. Male's tend to have a high default rate for each, which is interesting considering this set contains almost double the amount of females.



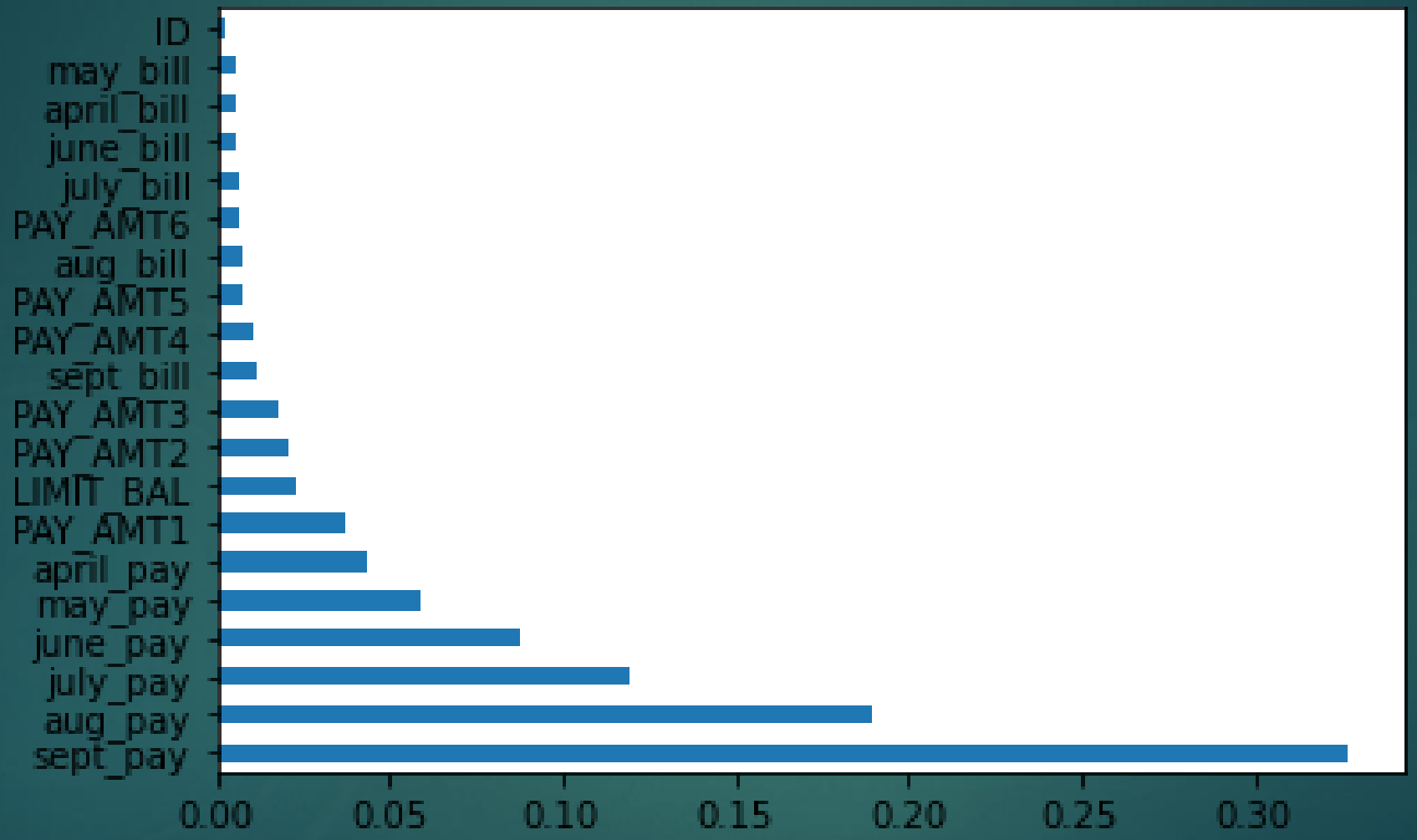
Analyzing the age of the customers





Age (year) It might be useful to see whether the gender of the customer is in any way related to his/her probability of default. From the above graphs we can say that most of the customers are from the age group of 25-30 years.

Futuristic Features



Training the model

Fitting different models

For modelling we tried various classification algorithms like:

- Logistic Regression

  Logistic Regression using LLE

- SVC

- SVC with LLE

Ensemble Learning

- Bagging Classifier

- Voting Classifier

Evaluating the model

After the model is built, if we see that the difference in the values of the predicted and actual data is not much, it is considered to be a good model and can be used to make future predictions.

Few metric tools we can use to calculate error in the model

1) Confusion Matrix: It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting.

2) ROC Curve: Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade-offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume $p > 0.5$ since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of $p > 0.5$. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under the curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0.

ML Models and Metrics

Accuracy of Logistic Regression Model: 0.8206

Accuracy of SVC Model: 0.8198

Accuracy of Bagging Classifier: 0.8206

Accuracy of Voting Classifier: 0.8207

Conclusion

The objective of this project is to train various supervised learning algorithms to predict the client's behaviour in paying off the credit card balance. In classification problems, an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization. We started with the logistic regression model, then compared the results with traditional machine learning-based models. Then K-means SMOTE resampling method on Taiwan client's credit dataset.

In the end, the proposed method has also been deployed on the web to assist the different stakeholders. Therefore, when the financial institution considers issuing the client a credit card, the institution needs to check the payment history of that person because the decision on whether to pay on duly or owe the bill on a specific month usually relates to the previous payment history. For instance, if a person owes numerous bills already, he or she is likely to delay the payment of the current month unless this person gets a windfall so that the total arrears can be paid off. Besides the payment history, it is also imperative to look at the applicants' credit limit of their current credit cards. This is a result of a virtuous circle: people who pay on duly tend to have better credit scores, so the banks prefer to increase these people's credit lines by taking less risk.

Conclusion

As a result, if a potential client already has a credit card with a high credit limit line, this person is unlikely to fail to pay the full amount owed in the future. Although the financial institution often collects clients' personal information such as age, educational level, and marital status when people apply for credit cards, this information also affects the default behaviour. In other words, the financial institution should equally consider their potential clients who are men or women, obtain bachelor degrees or master degrees, single or married when deciding whether to approve their credit card/loan applications. We tried our best to make a thorough analysis, and there are still a few possible improvements that may require longer-term action. For the boosting models, only the GBDT method was trained, but various variants of boosting techniques may also be utilized in the future. The financial market changes rapidly every day, and people's economic status and performance are affected by the market all the time. So, if more economic indicators are added to the dataset, this will lead to a more generic model. After exploring, manipulating and experimenting with different models on the credit card default data set we have obtained a maximum accuracy of 82% to determine whether a person defaults on their credit card or not. Ideally we would have been able to increase this accuracy by trying out various ways of pre-processing the data, utilizing dimensionality reduction, fine-tuning the models' hyperparameters, and applying ensemble learning. First we derived new features from the data set. Since this resulted in about 100 features we explored dimensionality reduction. Using the second form of reduction, LLE, we ran our initial classification models, Logistic Regression and Support Vector Classifier with the original training set and the reduced set.

Conclusion

Seeing no improvement with the set produced in LLE we continued using our original training set.

Next we fine-tuned the better performing model, Logistic Regression, with GridSearchCV as another attempt to improve the model. After using GridSearchCV no significant improvements were seen.

This led us to trying ensemble learning to see if our overall accuracy could be improved by combining various models' predictions. Once again there did not seem to be a significant improvement in both the Bagging Classifier and Voting Classifier when compared to our Logistic Regression model.

In the end, looking at the accuracy scores of each model was not enough information to choose which model performed best when trying to predict whether or not a person would default on their credit card. Above one can clearly see how close all these accuracy scores fall. In fact, by just seeing the accuracy scores the Voting Classifier seems to perform the best. But if looked at performance via ROC curves the Logistic Regression model seems to perform the best and the Voting Classifier does significantly worse.

For future work, we think it would be interesting to develop more complex models, such as implementing a neural network and seeing if there could be a better performance of an 82% accuracy score since all our previous techniques did not seem to affect the accuracy.


Thank you