Default Payments of Credit Card

Business Context

Financial threats are displaying a trend in the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats commercial banks face is the risk prediction of credit clients. Therefore, banks must have a risk prediction model and be able to classify the most relative characteristics that are indicative of people who have a higher probability of defaulting on credit.

A slight increase in the accuracy of identifying high-risk loans could prevent losses. Credit default prediction aims to help financial institutions decide whether to lend to a client or not.

Problem

The default rate on credit loans across all commercial banks is at an all-time high. The climbing delinquencies will result in a significant amount of money lost from the lending institutions, such as commercial banks. Because of the risks inherent in such a large portion of the economy, building models for consumer spending behaviors to limit risk exposures in this sector is becoming more critical. We can improve its ability to anticipate the chance of default for its customers and identify the significant factors that influence this likelihood. It would also assist the issuer in gaining a better understanding of their current and potential clients, which would drive their future strategy, including plans to offer tailored credit. This project aims to let the issuer decide who can get a credit card and the credit limit they can give to that person, and this will be possible by assessing the accuracy of default estimates by comparing different classification methods and selecting the optimal classifier.

Data Source:

UCI Machine Learning Repository
[https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients]

Data Description:

The dataset contains information on 24 variables, including demographic factors, credit data, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not. The number of instances in this dataset are 30000. The data is also clean as there are no null values present in it. The column name and description of the dataset are as follows:

Column Name	Description
ID	ID of each element
Limit_Bal	Amount of given credit in NT dollars
Sex	Gender of the customer
Education	The level of education the customer has studied
Marriage	Marital status of the customer
Age	Age of customer
Pay_0	Repayment status in September 2005
Pay_2	Repayment status in August 2005
Pay_3	Repayment status in July 2005
Pay_4	Repayment status in June 2005
Pay_5	Repayment status in May 2005
Pay_6	Repayment status in April 2005
Bill_Amt1	Amount of bill payment in September 2005
Bill_Amt2	Amount of bill payment in August 2005
Bill_Amt3	Amount of bill payment in July 2005
Bill_Amt4	Amount of bill payment in June 2005
Bill_Amt5	Amount of bill payment in May 2005
Bill_Amt6	Amount of bill payment in April 2005
Pay_Amt1	Amount of previous payment in September 2005
Pay_Amt2	Amount of previous payment in August 2005
Pay_Amt3	Amount of previous payment in July 2005
Pay_Amt4	Amount of previous payment in June 2005
Pay_Amt5	Amount of previous payment in May 2005
Pay_Amt6	Amount of previous payment in April 2005
Default.payment.next.month	Default payment whether it should be paid by the customer or not

Insights From Exploratory Data Analysis:

From Education:

- We can infer that the number of people who have University degree are more compared to others.
- It seen that the limit balance is higher for Graduates, but the default rate is high for people with university degree.

From Marriage:

- We can infer that people who are single are more when compared to Married and Other category.
- We can observe that the default rate is similar for people who are single and married.

From Gender:

- We can observe there are a greater number of Females in the group, but the pattern in the limit balance across them is similar and people between both segments having balance below 50000 are more than 3000.
- Though the number of females is more we can infer that the default rate is greater for men.

From Age:

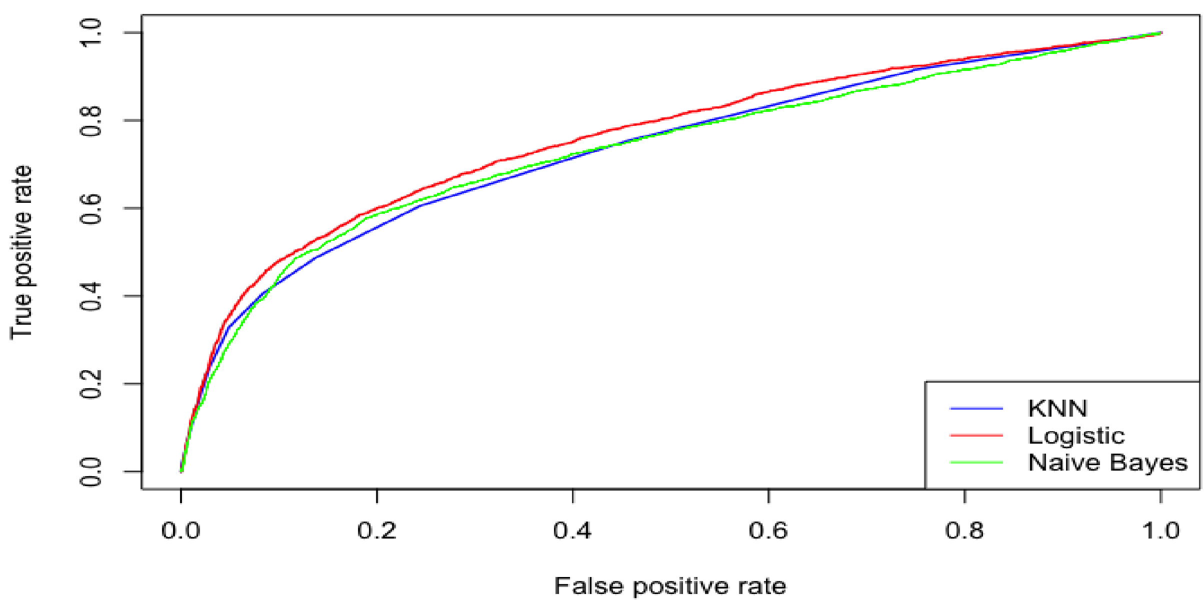
- We see that as the age of a person increases the limit balance goes up gradually.
- It can also be inferred that people having limit balance below the generated trend line are more likely to default.
- The non-default category and the default category are more populated by people in the 20-30 age range.

Prediction of Default using Multiple Classifiers:

We have divided the dataset into training and test data set in the ratio of 60:40 where 60% of the data was used to train the model. We have utilized three classification algorithms to build a prediction model and the classifiers which were used are Logistic Regression, K-NN Classifier and Naïve Bayes Classifier. Upon evaluating the accuracy of the 3 models used it is observed that optimal classifier turned out to be logistic regression with an accuracy of 82.2%. K-NN classifier had an accuracy of 81.1% which is slightly lower than that of logistic regression. But Naïve Bayes Classifier had the least accuracy amongst the 3 with the value of 66.1%.

```
> Accuracy_df_t
                       Accuracy_df_t
KNN Accuracy               0.8115833
Logistic Accuracy          0.8224167
Naive Bayes Accuracy       0.6610000
```

We have also visualized ROC curves for all the 3 classifiers, and it is observed that Logistic model has the highest AUC of 75.92% which can be understood from the picture placed below.



Improvement for the Future

- To collect more demographic variables such as Salary, Credit Score etc. as it can help the model predict better.
- To conduct PCA and clustering as it segments the customers based on various variables and help the financial institution to take risk accordingly.
- Add more data to the training dataset as the model runs against it and this improves the it and helps it identify more cases of default correctly.