GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI
Global Mindset - Indian Roots

GREAT LEARNING

PGP in Data Science and Engineering

Interim Report for:

FINANCIAL RISK ANALYSIS

CREDIT CARD DEFAULT

Group Members :
Enavamshi Gadikota
Mohit Sardar
Sruthi Basani
Radhakrishna Penugonda
Adarsh Raj

Mentor :
Naveen Koneti

Contents

1 Introduction to Domain

Credit cards are typically issued by banks, and it enables the cardholder to borrow funds from that bank and issuers customarily pre-set borrowing limits, based on an individual's credit rating and the banks also impose the condition that cardholders pay back the borrowed money, plus interest, as well as any additional agreed-upon charges.

What is in it for banks?
Credit cards feature higher Annual Percentage rates (APRs) than other forms of consumer loans. Interest charges on the unpaid balance charged to the card are typically imposed one month after a purchase is made.

What makes this product attractive to customers?
Apart from borrowing money when in need, credit card users can avail discounts, travel points, cash-backs and many other perks unavailable to debit card holders. Consumers who pay off their cards in full and on time every month can profit substantially by running their monthly purchases and bills through reward points.

1.1 Abstract

Credit card business is one of the key areas in the banking industry where there exists a huge money flow. Credit card business has been a huge success in the banking market over the years. Most of the banks are also interested in this area expecting a good revenue. Many of the customers use their credit cards beyond their repayment capabilities leading to high debt accumulation and finally defaulting the credits. Classifying the risky and non-risky customers has become a very big challenge for banks. So, the problem we are trying to analyse is to identify the risky and non-risky customers based on the available customer data, helping the bank to decide if a customer can repay his credits to the bank in the coming months.

2 Dataset Description

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card customers of a bank in Taiwan from April 2005 to September 2005.

There are 25 variables:

1. **ID:** ID of each customer
2. **LIMIT_BAL:** Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3. **SEX:** Gender (1=male, 2=female)
4. **EDUCATION:** (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
5. **MARRIAGE:** Marital status (1=married, 2=single, 3=others)
6. **AGE:** Age in years
7. **Repay_Sept:** Repayment status in September, 2005 (-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)
8. **Repay_Aug:** Repayment status in August, 2005 (scale same as above)
9. **Repay_July:** Repayment status in July, 2005 (scale same as above)
10. **Repay_June:** Repayment status in June, 2005 (scale same as above)
11. **Repay_May:** Repayment status in May, 2005 (scale same as above)
12. **Repay_April:** Repayment status in April, 2005 (scale same as above)
13. **Bill_Amt_Sept:** Amount of bill statement in September, 2005 (NT dollar)
14. **Bill_Amt_Aug:** Amount of bill statement in August, 2005 (NT dollar)
15. **Bill_Amt_July:** Amount of bill statement in July, 2005 (NT dollar)
16. **Bill_Amt_June:** Amount of bill statement in June, 2005 (NT dollar)
17. **Bill_Amt_May:** Amount of bill statement in May, 2005 (NT dollar)
18. **Bill_Amt_April:** Amount of bill statement in April, 2005 (NT dollar)
19. **Pre_Pay_Sept:** Amount of previous payment in September, 2005 (NT dollar)
20. **Pre_Pay_Aug :** Amount of previous payment in August, 2005 (NT dollar)
21. **Pre_Pay_July:** Amount of previous payment in July, 2005 (NT dollar)
22. **Pre_Pay_June:** Amount of previous payment in June, 2005 (NT dollar)
23. **Pre_Pay_May:** Amount of previous payment in May, 2005 (NT dollar)

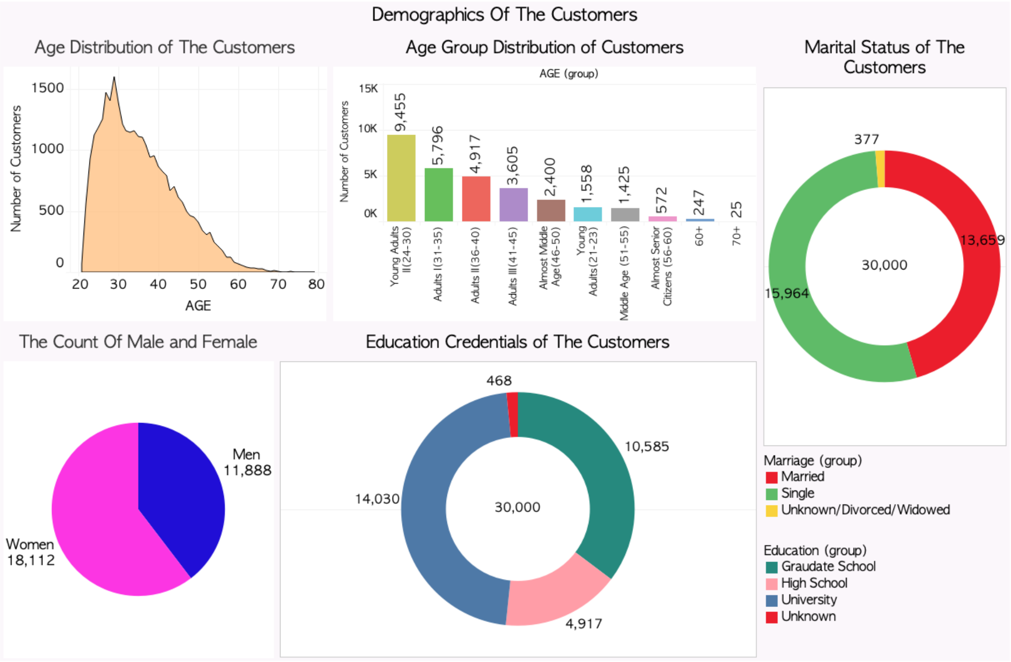
24. **Pre_Pay_April:** Amount of previous payment in April, 2005 (NT dollar)
25. **DEFAULT:** Default payment (1=yes, 0=no)

3 Exploratory Data Analysis

3.1 General Observations

- There are no missing values
- Every data point is important in the credit card business, so we did not consider any outliers in the bill amount, limit balance amount and the payment amount columns. Instead we just considered them as extreme values.
- The *Pre_Pay* amounts for each month correspond to the *Bill_Amt* of the previous month but the six months of *Pre_Pay* do not align with the six months of *Bill_Amt* given i.e *Pre_Pay_April* is given but Bill Amount for March is not given. Similarly Bill_Amt_Sept is given but payment in October is not.

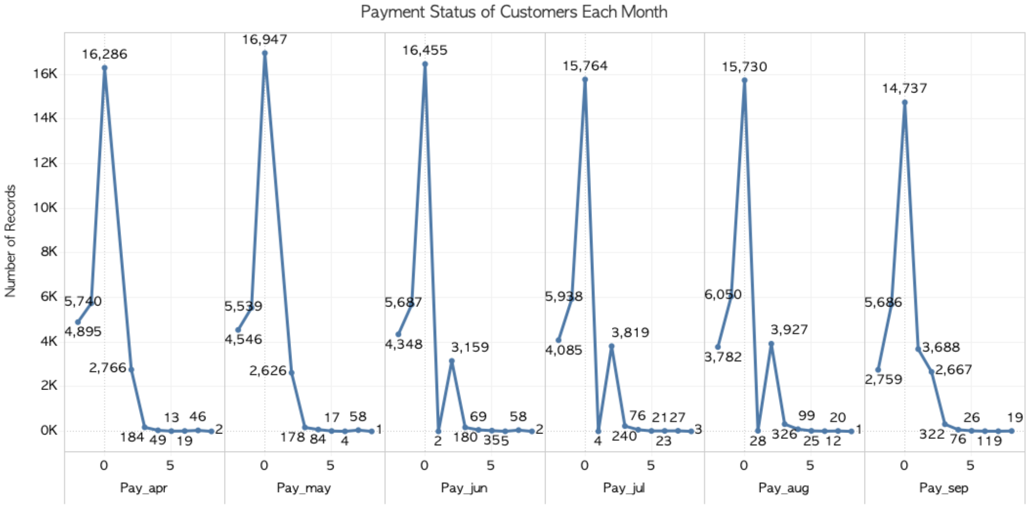
3.2 Univariate Analysis



In the education column, as per the data description 1=graduate school, 2=university, 3=high school, 4=Others, 5=unknown, 6=unknown so we have replaced 5 and 6 with 4 as Others.

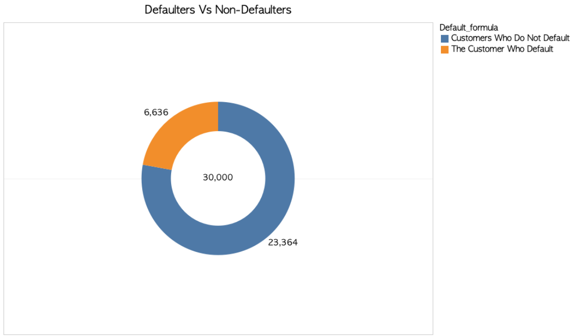
In the marriage column, as per the data description 1=married, 2=single, 3=others, but we had some values in 0=unknown so replaced those values with 3 as others.

The payment status columns for all the months had values like -2,-1,0, 1 and as per the dataset description -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 but we did not have any explanation for what 0, -1 mean and a very important point to be noted is every months there are atleast 50% of rows which had their payment status as 0.So we have taken values less than and equal to zero as the group which made payments on time and 1=payment delay for one month, 2=payment delay for two months and so on.

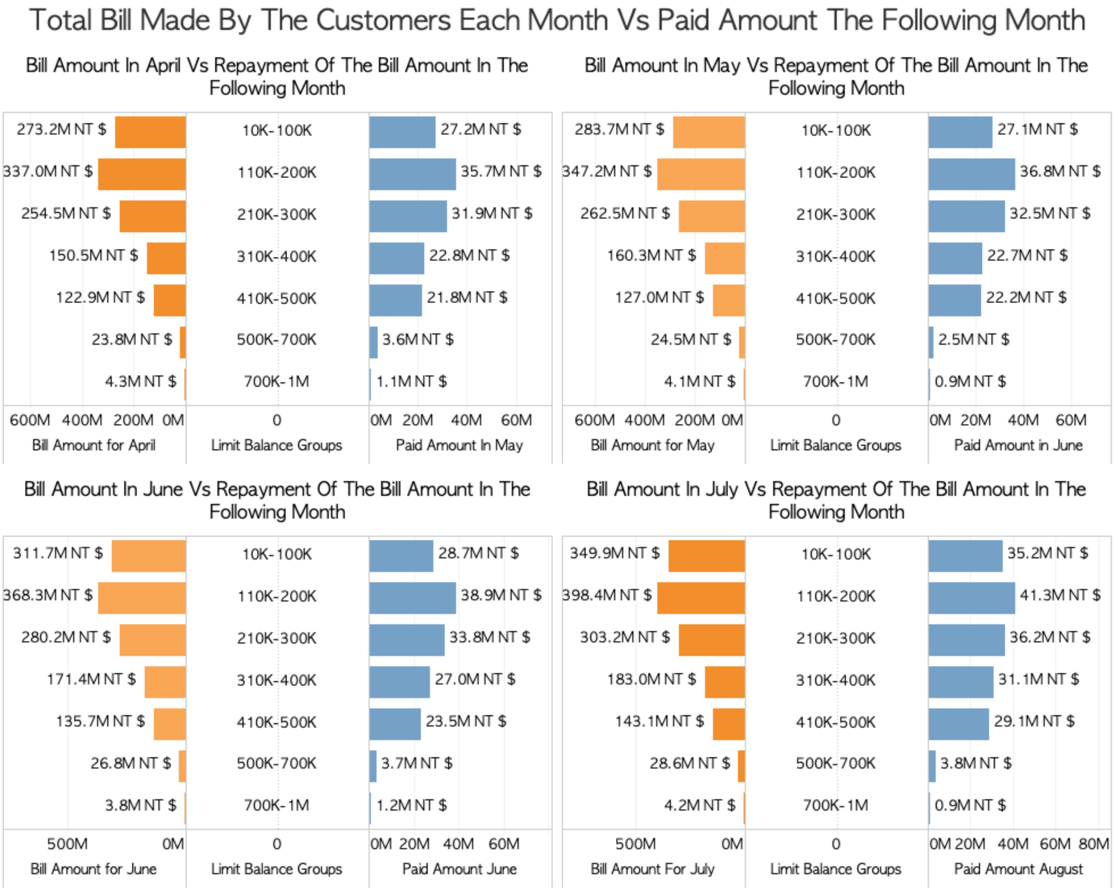


We had no explanaion for what 0 means in the payment status and it is very clear that every month 50% of the customers have their payment status as 0.

The data for our target variable, *DEFAULT* is imbalanced.

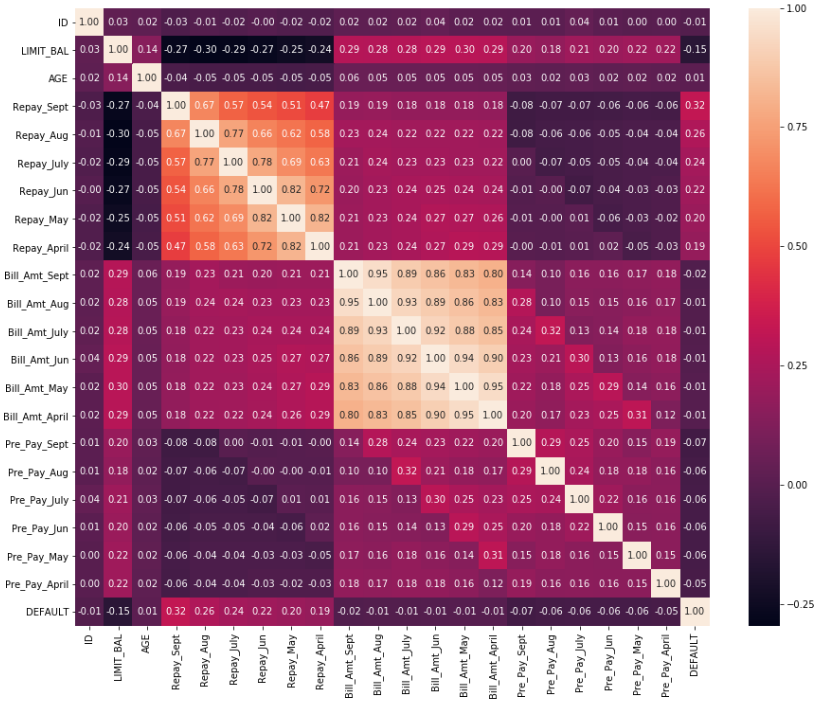


3.3 Bill Amount vs Amount Paid



For all the Limit Balances, The Amount paid is around 10-20% of the bill amount for the previous month.

3.4 Correlation



	ID	LIMIT_BAL	AGE	Repay_Sept	Repay_Aug	Repay_July	Repay_Jun	Repay_May	Repay_April	Bill_Amt_Sept	Bill_Amt_Aug	Bill_Amt_July	Bill_Amt_Jun	Bill_Amt_May	Bill_Amt_April	Pre_Pay_Sept	Pre_Pay_Aug	Pre_Pay_July	Pre_Pay_Jun	Pre_Pay_May	Pre_Pay_April	DEFAULT
ID	1.00	0.03	0.02	-0.03	-0.01	-0.02	-0.00	-0.02	-0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.01	0.01	0.04	0.01	0.00	0.00	-0.01
LIMIT_BAL	0.03	1.00	0.14	-0.27	-0.30	-0.29	-0.27	-0.25	-0.24	0.29	0.28	0.28	0.29	0.30	0.29	0.20	0.18	0.21	0.20	0.22	0.22	-0.15
AGE	0.02	0.14	1.00	-0.04	-0.05	-0.05	-0.05	-0.05	-0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.03	0.02	0.02	0.02	0.01
Repay_Sept	-0.03	-0.27	-0.04	1.00	0.67	0.57	0.54	0.51	0.47	0.19	0.19	0.18	0.18	0.18	0.18	-0.08	-0.07	-0.07	-0.06	-0.06	-0.06	0.32
Repay_Aug	-0.01	-0.30	-0.05	0.67	1.00	0.77	0.66	0.62	0.58	0.23	0.24	0.22	0.22	0.22	0.22	-0.08	-0.06	-0.06	-0.05	-0.04	-0.04	0.26
Repay_July	-0.02	-0.29	-0.05	0.57	0.77	1.00	0.78	0.69	0.63	0.21	0.24	0.23	0.23	0.23	0.22	0.00	-0.07	-0.05	-0.05	-0.04	-0.04	0.24
Repay_Jun	-0.00	-0.27	-0.05	0.54	0.66	0.78	1.00	0.82	0.72	0.20	0.23	0.24	0.25	0.24	0.24	-0.01	-0.00	-0.07	-0.04	-0.03	-0.03	0.22
Repay_May	-0.02	-0.25	-0.05	0.51	0.62	0.69	0.82	1.00	0.82	0.21	0.23	0.24	0.27	0.27	0.26	-0.01	-0.00	0.01	-0.06	-0.03	-0.02	0.20
Repay_April	-0.02	-0.24	-0.05	0.47	0.58	0.63	0.72	0.82	1.00	0.21	0.23	0.24	0.27	0.29	0.29	-0.00	-0.01	0.01	0.02	-0.05	-0.03	0.19
Bill_Amt_Sept	0.02	0.29	0.06	0.19	0.23	0.21	0.20	0.21	0.21	1.00	0.95	0.89	0.86	0.83	0.80	0.14	0.10	0.16	0.16	0.17	0.18	-0.02
Bill_Amt_Aug	0.02	0.28	0.05	0.19	0.24	0.24	0.23	0.23	0.23	0.95	1.00	0.93	0.89	0.86	0.83	0.28	0.10	0.15	0.15	0.16	0.17	-0.01
Bill_Amt_July	0.02	0.28	0.05	0.18	0.22	0.23	0.24	0.24	0.24	0.89	0.93	1.00	0.92	0.88	0.85	0.24	0.32	0.13	0.14	0.18	0.18	-0.01
Bill_Amt_Jun	0.04	0.29	0.05	0.18	0.22	0.23	0.25	0.27	0.27	0.86	0.89	0.92	1.00	0.94	0.90	0.23	0.21	0.30	0.13	0.16	0.18	-0.01
Bill_Amt_May	0.02	0.30	0.05	0.18	0.22	0.23	0.24	0.27	0.29	0.83	0.86	0.88	0.94	1.00	0.95	0.22	0.18	0.25	0.29	0.14	0.16	-0.01
Bill_Amt_April	0.02	0.29	0.05	0.18	0.22	0.22	0.24	0.26	0.29	0.80	0.83	0.85	0.90	0.95	1.00	0.20	0.17	0.23	0.25	0.31	0.12	-0.01
Pre_Pay_Sept	0.01	0.20	0.03	-0.08	-0.08	0.00	-0.01	-0.01	-0.00	0.14	0.28	0.24	0.23	0.22	0.20	1.00	0.29	0.25	0.20	0.15	0.19	-0.07
Pre_Pay_Aug	0.01	0.18	0.02	-0.07	-0.06	-0.07	-0.00	-0.00	-0.01	0.10	0.10	0.32	0.21	0.18	0.17	0.29	1.00	0.24	0.18	0.18	0.16	-0.06
Pre_Pay_July	0.04	0.21	0.03	-0.07	-0.06	-0.05	-0.07	0.01	0.01	0.16	0.15	0.13	0.30	0.25	0.23	0.25	0.24	1.00	0.22	0.16	0.16	-0.06
Pre_Pay_Jun	0.01	0.20	0.02	-0.06	-0.05	-0.05	-0.04	-0.06	0.02	0.16	0.15	0.14	0.13	0.29	0.25	0.20	0.18	0.22	1.00	0.15	0.16	-0.06
Pre_Pay_May	0.00	0.22	0.02	-0.06	-0.04	-0.04	-0.03	-0.03	-0.05	0.17	0.16	0.18	0.16	0.14	0.31	0.15	0.18	0.16	0.15	1.00	0.15	-0.06
Pre_Pay_April	0.00	0.22	0.02	-0.06	-0.04	-0.04	-0.03	-0.02	-0.03	0.18	0.17	0.18	0.18	0.16	0.12	0.19	0.16	0.16	0.16	0.15	1.00	-0.05
DEFAULT	-0.01	-0.15	0.01	0.32	0.26	0.24	0.22	0.20	0.19	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.07	-0.06	-0.06	-0.06	-0.06	-0.05	1.00

There is high correlation between the *Bill Amt* variables but this correlation gradually decreases with difference in months.

Similarly, there is moderate correlation between the *Repay* variables but this correlation also gradually decreases with difference in months.

4 Hypothesis Testing

Statistical tests were performed to see the whether each of the 24 independent variables have a significant relationship with the dependent variable, *DEFAULT*

4.1 Chi-square Test

For the Categorical Columns, a Chi-square Test of independence was performed with the target variable, *DEFAULT* which is also a categorical column.

Here Null Hypothesis H_0: There is NO association between the two variables
And Alternate Hypothesis H_a: There is an association between the two variables

Variable	**p-value**	**Decision**
SEX	$4.94e^{-12}$	Reject H_0
EDUCATION	$1.23e^{-32}$	Reject H_0
MARRIAGE	$8.82e^{-8}$	Reject H_0

For all three variables above The null hypothesis have been rejected which means that they all have a significant relationship with the target variable.

For the *Repay* variables there are 12 categories so instead we will use the simpler version of the columns where ≤ 0 (Duly paid) or > 0 (Payment Delay)

Variable	**p-value**	**Decision**
Repay_Sept	0.0	Reject H_0
Repay_Aug	0.0	Reject H_0
Repay_July	0.0	Reject H_0
Repay_Jun	0.0	Reject H_0
Repay_May	0.0	Reject H_0
Repay_April	0.0	Reject H_0

For all six variables above The null hypothesis have been rejected which means that they all have a significant relationship with the target variable.

4.2 Two-sample t test

For all the numeric variables, A two-sample unpaired t tests was performed between values of the variable for two classes of target variables to compare their means.

Here Null Hypothesis H_0: The means of the two samples are EQUAL
And Alternate Hypothesis H_a: The means of the two samples are NOT EQUAL.

If the means of the two samples are significantly different form each other, then we can conclude that the variable does have a significant relationship with the target variable.

The preliminary Normality Tests (*Shapiro*) and Equality of Variance test (*Levene* and *Bartlett*) were done to determine whether to do a parametric test (*ttest_ind*) or a non-parametric test (*Mannwhitneyu*).

Variable	**Parametric/Non-parametric**	**p-value**	**Decision**
LIMIT_BAL	Non-parametric	$6.12e^{-190}$	Reject H_0
AGE	Non-parametric	0.186	Failed to Reject H_0
Bill_Amt_Sept	Non-parametric	$5.75e^{-6}$	Reject H_0
Bill_Amt_Aug	Non-parametric	0.003	Reject H_0
Bill_Amt_July	Non-parametric	0.014	Reject H_0
Bill_Amt_Jun	Non-parametric	0.074	Failed to Reject H_0
Bill_Amt_May	Non-parametric	0.118	Failed to Reject H_0
Bill_Amt_April	Non-parametric	0.494	Failed to Reject H_0
Pre_Pay_Sept	Non-parametric	$2.31e^{-170}$	Reject H_0
Pre_Pay_Aug	Non-parametric	$4.98e^{-151}$	Reject H_0
Pre_Pay_July	Non-parametric	$4.49e^{-129}$	Reject H_0
Pre_Pay_June	Non-parametric	$3.64e^{-109}$	Reject H_0
Pre_Pay_May	Non-parametric	$5.62e^{-91}$	Reject H_0
Pre_Pay_April	Non-parametric	$1.59e^{-98}$	Reject H_0

Only *AGE* and the Bill amounts for the last three months failed to reject the null hypothesis which means that they do not have a significant relationship with the target variable. While the rest of the numerical variables, *LIMIT_BAL*, Bill amounts for the first three months and all the six Pre Pay amounts rejected the null hypothesis and hence have a significant relationship with the target variable.

5 Evaluation Metrics

The Evaluation Metrics that can be used for a Binary Classification problem are:

1. **Accuracy** - Proportion of correctly identified instances
2. **Precision** - proportion of positive predictions that are correct
3. **Recall** - Proportion of Actual positives predicted correctly
4. **F1 Score** - Harmonic mean of Precision and Recall
5. **ROC AUC** - Area Under Receiver's Operating Characteristics Curve (trade-off between sensitivity and specificity for different thresholds)
6. **Log Loss** - measures the uncertainty of the probabilities of the model by comparing with the true labels
7. **Cohen's Kappa** - measure of inter rater reliability between two raters (actual and predicted labels)

Due to an imbalance in the dataset, Accuracy is an unreliable measure to evaluate a model. For example, say the test dataset has a 70-30 split of the two classes of target variable. A model which simply predicts the majority class (in our case 0) will get 70% accuracy despite the model not predicting even one defaulter.

Since our main goal is to identify possible defaulters we need to focus on the True Positives. Recall is an excellent measure here since it will give us a measure of proportion of actual defaulters identified correctly.

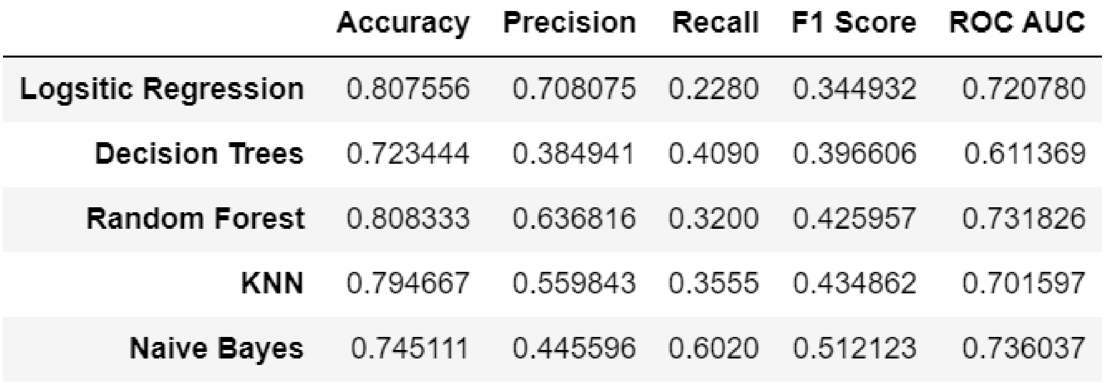
Precision is still an important measure since it identifies how many of the ones predicted as defaulter are correct. From the credit card company's perspective, it is not desirable to predict too many non-defaulters as defaulters since this could result in losing good customers. In comparison Recall is still more important than precision but we should not compromise too much on precision.

F1 Score is a harmonic mean of precision and recall. High F1 Score is achieved only if both precision and recall are high. This is also an excellent measure to evaluate our models.

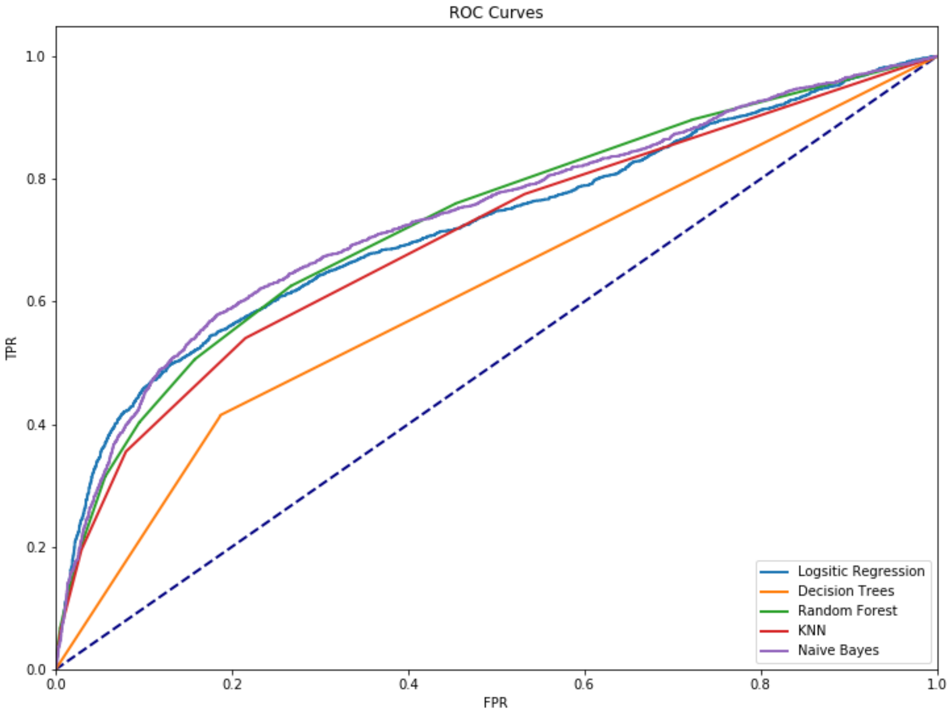
ROC Curve plots True Positive Rate (sensitivity / recall) vs False Positive Rate (1 - specificity) for range of threshold values between 0 and 1. The Area Under the Curve can range from 0 to 1 and is a good measure to evaluate how well model is capable of distinguishing between the two classes. This is also a good evaluation metric for our problem statement.

6 Base Models

Before we launch into creating predictive models for the problem, we start with some baselines to give some sense of what we are trying to achieve. In the due process we have created Logistic Regression, Decision Trees, Random Forest, K Nearest Neighbors, Naïve Bayes.

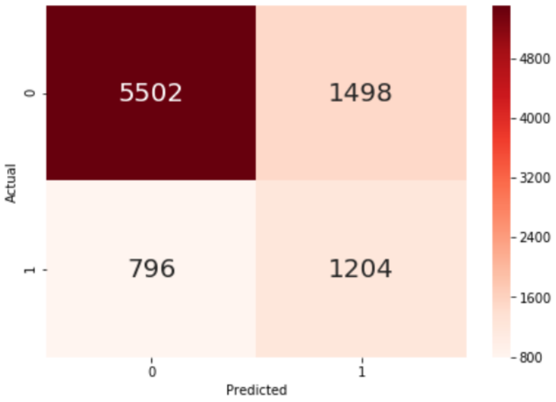


	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logsitic Regression	0.807556	0.708075	0.2280	0.344932	0.720780
Decision Trees	0.723444	0.384941	0.4090	0.396606	0.611369
Random Forest	0.808333	0.636816	0.3200	0.425957	0.731826
KNN	0.794667	0.559843	0.3555	0.434862	0.701597
Naive Bayes	0.745111	0.445596	0.6020	0.512123	0.736037



Even though logistic regression has the highest accuracy and precision scores, its recall is too low comparatively. Since our main aim is to find defaulters logistic regression can't be the best suitable model for the data. Where as the Decision Trees has decent accuracy score but it failed in giving better precision and recall scores and so on. Amongst all the models below, Naive Bayes has the best recall score with 60.2% and with the largest AUC. From the baseline classifiers that we have, Naive Bayes seems to be the best suitable for the data.

Confusion Matrix for the Naive Bayes model:



7 Conclusions

We have started the project by performing some Exploratory Data Analysis and building some base models using different classification algorithms to compare the results using various evaluation techniques.

Based on the above we are able to conclude that whether the person is going to "Default" the Credit card Bill for the immediate next month, but we also see that the prediction is not quite satisfactory due to the low recall and F1 scores.

7.1 Future Scope

1. **Undersampling/Oversampling :** The dataset is slightly imbalanced hence we cna perform some undersampling or oversampling techniques like SMOTE to see if this helps models learn patterns for the minority class (Defaulters) better.

2. **Feature Engineering :** We can create new features from the existing ones and test whether we can build better models with them.
3. **Feature Selection :** We can use some feature selection techniques to check which of the columns are affecting or contributing more towards the target variable.
4. **Parameter Tuning :** For each model, different combinations of respective hyper-parameters can be tried to see which one helps get the best recall/F1 scores.
5. **Ensemble Techniques:** More complex models using Bagging and Boosting techniques can be built to get better recall/F1 scores.
6. **Dimensionality Reduction:** Unsupervised techinques like Principal Component Analysis could be applied to uncover some hidden patterns in the data.