Hello!

I am Abhay Kulkarni

Taiwan Bank Customer Default

acknowledgement

I heartily thank my mentor, **Ms Karuna Kumari** for her guidance and suggestions during this project work.

Problem Statement

- Beginning in 1990, Taiwan Bank with the goal of aggressively expanding their businesses and increasing profits, lavished money encouraging people to apply for credit cards, apparently without consequences.
- Taiwan Bank lowered the requirements for credit card approvals to get more customers
- In 2006, debt from credit cards reached **$268 billion USD**. More than half a million people were not able to repay their loans.

Business Objective

- Taiwan bank wants to identify potential defaulters
- Identify the patterns and factors leading to default.
- Build a model that predicts defaulters. This will help bank to reduce loss and lend money to good borrowers and maintain healthy status of the bank.

Scope Of Project

- Effectively Reach out to potential defaulting customers and prevent Revenue losses.
- Understand the factors contributing to default and mitigate risk accordingly.
- Adjust Interest Rate Per Applicant. Higher the Credit Risk of Borrower the higher interest they pay.
- Completely avoid giving loans to Applicants with Weak Credit History(Score)

What are NPAs?

A non-performing asset (NPA) is a banking industry term for a 'bad loan' – i.e. one that has not been repaid within the stipulated time, or where the scheduled payments are in arrears.

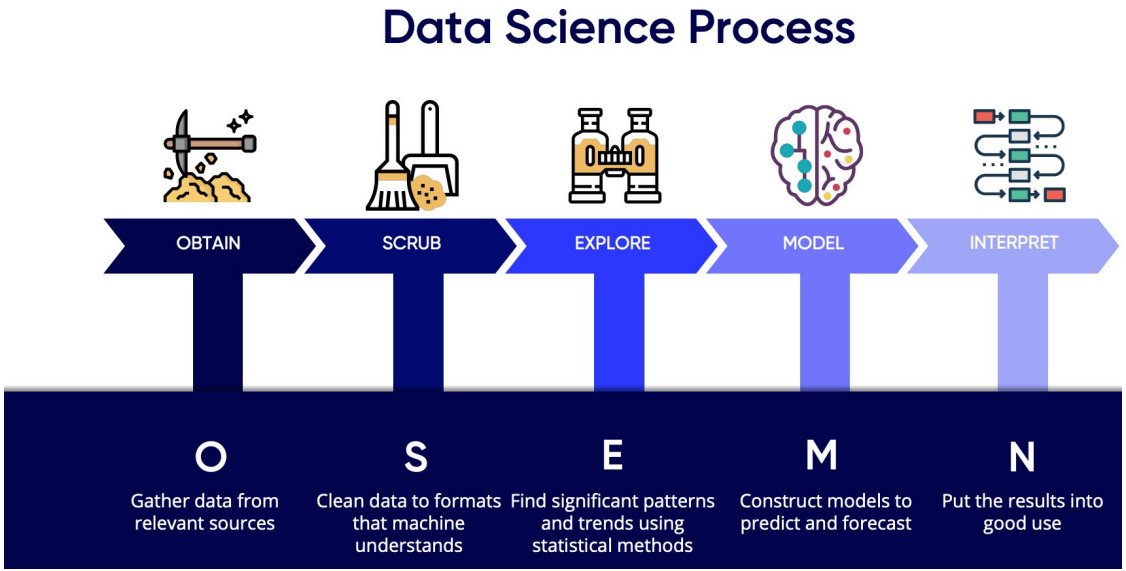
Impact of NPAs

- **Credibility:** The credibility of the banking system gets hit, which can affect the economy as a whole

- **Credit contraction:** Burgeoning NPAs reduces recycling of funds, and by extension, also that of the bank's ability to lend more. This, in turn, results in interest income decline. **On a macro level,** it contracts money circulation that can lead to an economic slowdown

"In financial services, if you want to be the best in the industry, you first have to be the best in risk management and credit quality. It's the foundation for every other measure of success. There's almost no room for error." — John G. Stumpf

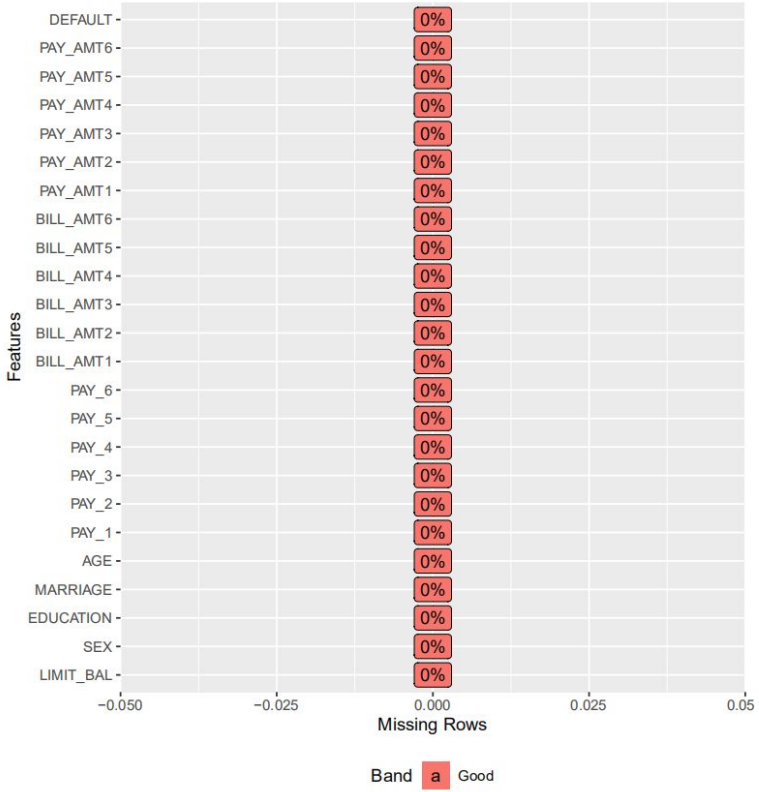
Process Flow using OSEMN framework



Exploratory Data Analysis

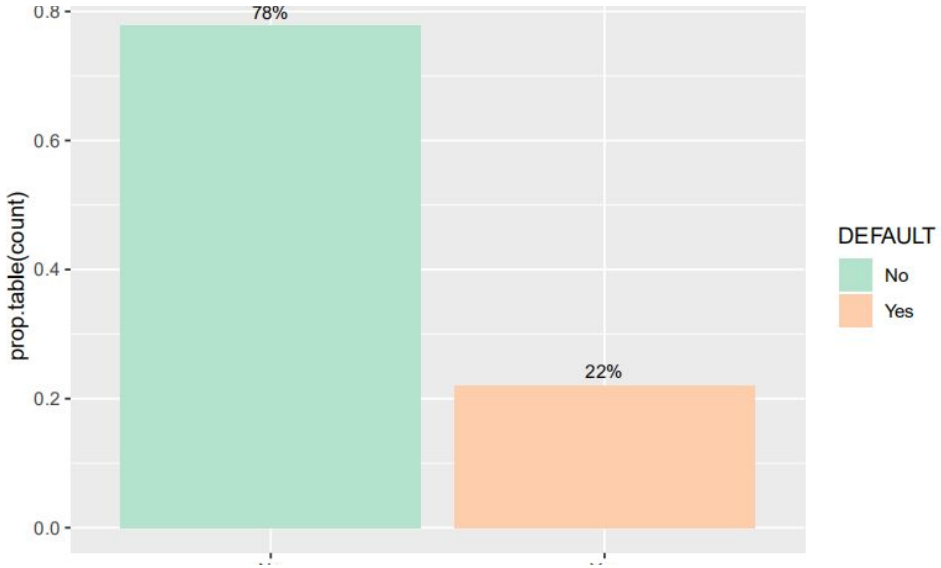
Understanding the data

Missing Values & Data Summary



- There are No Missing Values.
- There are 30000 observations and 25 Features. Few Features have to be converted to factors.

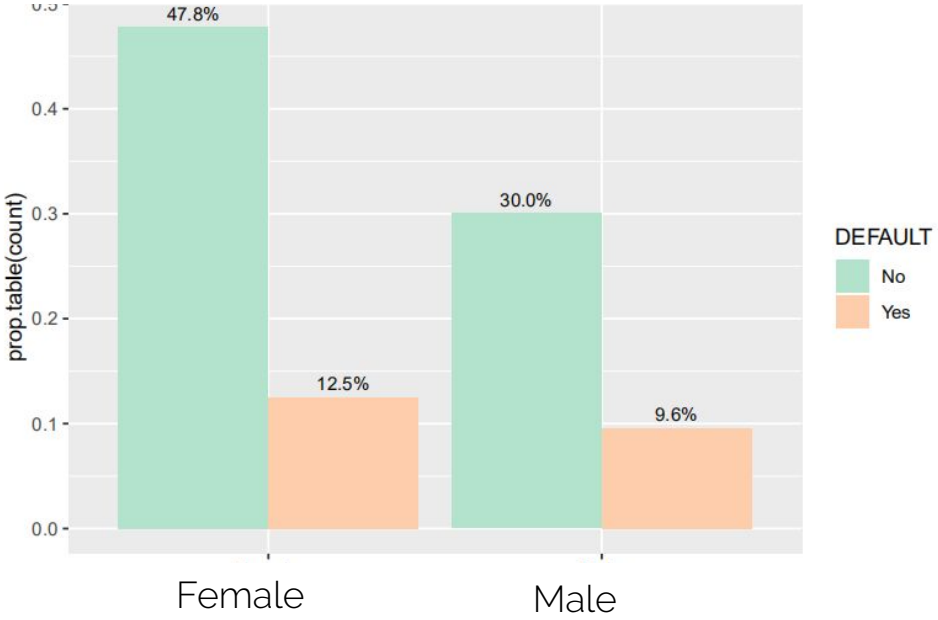
Default (Dependent Variable Split)



Data Constraint :

- Imbalanced Dataset
- “NO 78% and 22%”YES"
- Will use SMOTE to counter the imbalanced dataset.

Gender VS Default



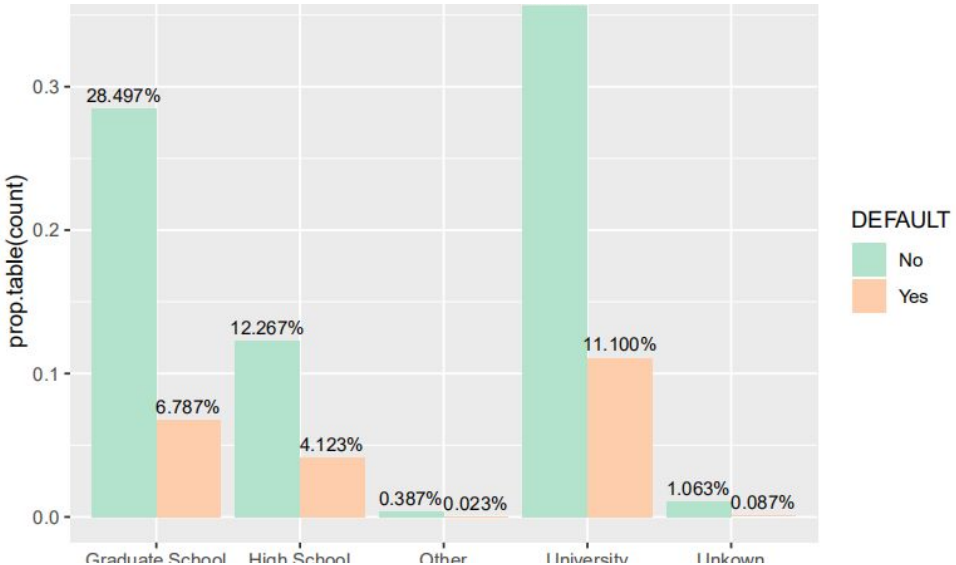
Null Hypothesis(HO) : Default is Independent of Gender

Alternative Hypothesis(H1) : Default is dependent on Gender

Pearson's Chi-squared test with Yates' continuity correction
data: GenderTable
X-squared = 47.709, df = 1, p-value = 4.945e-12

p- value is way lesser than 0.05. We reject Null Hypothesis and go with Alternative hypothesis that Default" depends on “Gender”.

Education VS Default



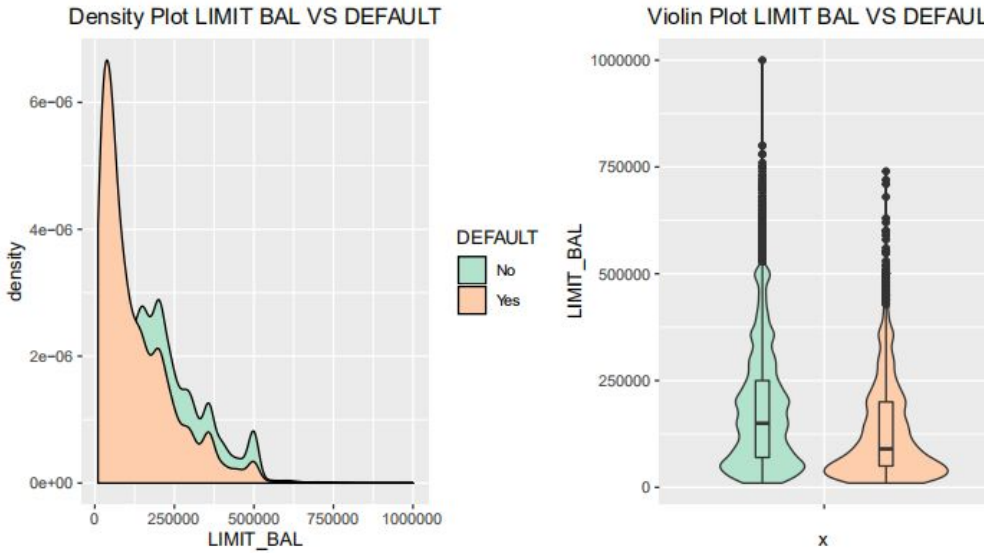
Null Hypothesis(HO) : Default is Independent of EDUCATION

Alternative Hypothesis(H1) : Default is dependent on EDUCATION

Pearson's Chi-squared test with Yates' continuity correction
data: Education Table
X-squared = 160.59, df = 4, p-value < 2.2e-16

p- value is way lesser than 0.05. We reject Null Hypothesis and go with Alternative hypothesis that Default" depends on “Education”.

Credit Limit VS Default

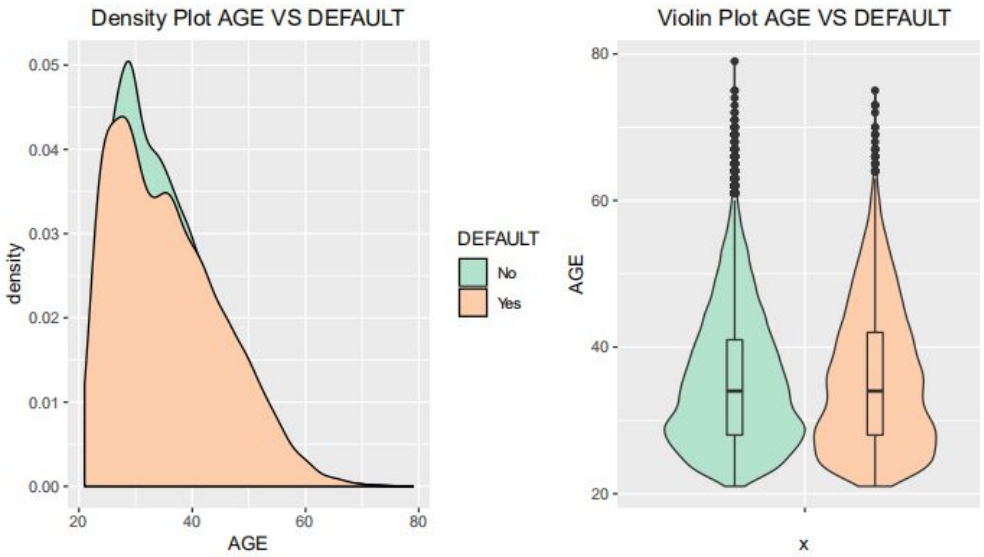


Approx 17% of Customers with Credit Limit **above 100,000 Defaulted**

Approx 29% of Customers with Credit Limit **below 100,000 Defaulted**

Outliers present. Will be treated

Age VS Default



There are **more defaulters between Age 20 and Age 35**

Outliers present. Will be treated

Modelling Approach

This is a **Classification Problem**

- **Predict Binary Output** (Default vs Not Default). Building a model to predict if borrower will default or not default which is a binary problem. So, classification Algorithms are used. Logistic Regression, KNN, Naive Bayes, CART, Random Forest.

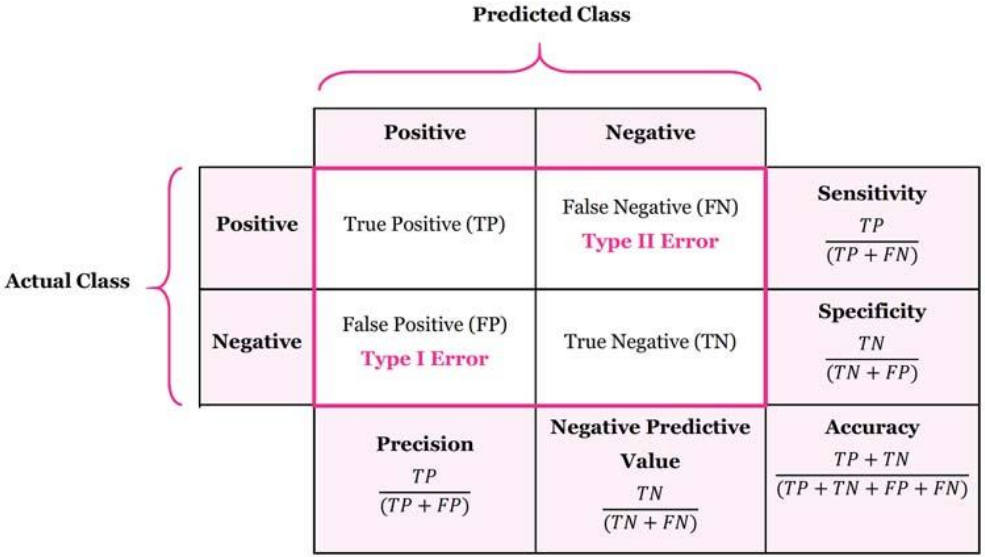
Models	Advantages	Disadvantages
Logistic Regression	1) Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative) 2) Logistic regression is easier to implement, interpret and very efficient to train.	1) Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
KNN	1) No Training Period: KNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training	1) Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

Models Used + Advantages vs Disadvantages

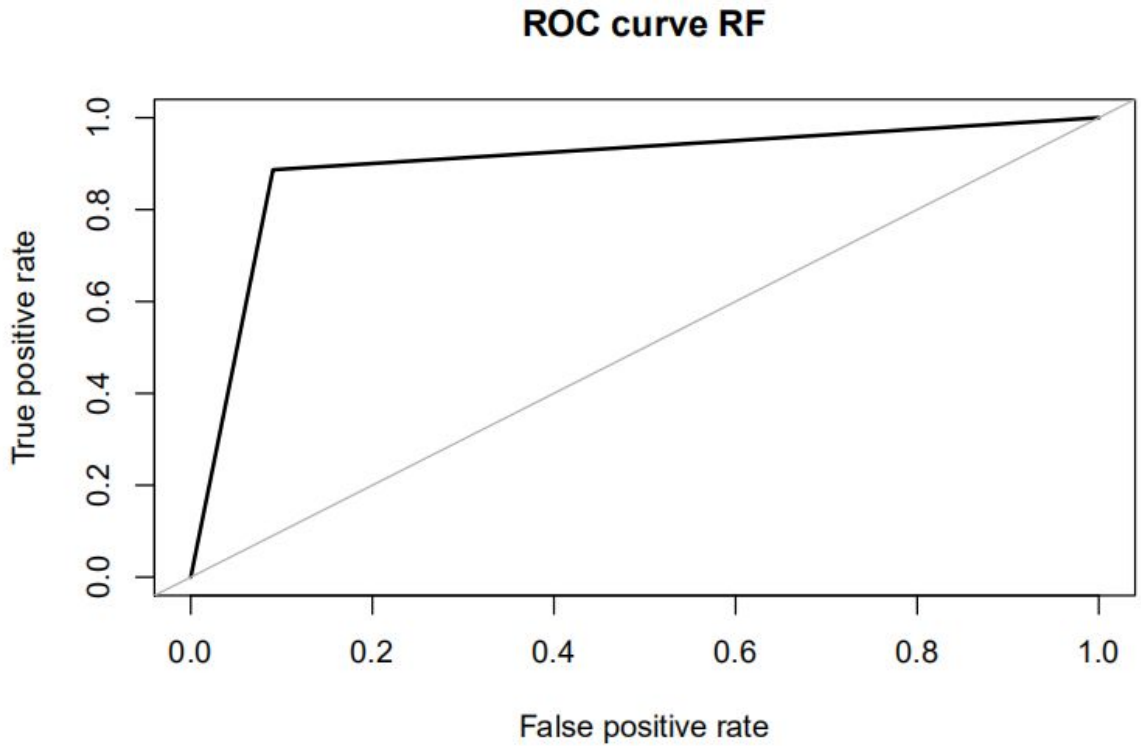
Models	Advantages	Disadvantages
Naive Bayes	1) Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.	1) Main limitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent
CART	1) A Decision trees model is very intuitive and easy to explain to technical teams as well as stakeholders.	1) A small change in the data can cause a large change in the structure of the decision tree causing instability.
Random Forest	1) Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.	1) Longer Training Period: Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes

Models Used + Advantages vs Disadvantages

Model Evaluation Parameters

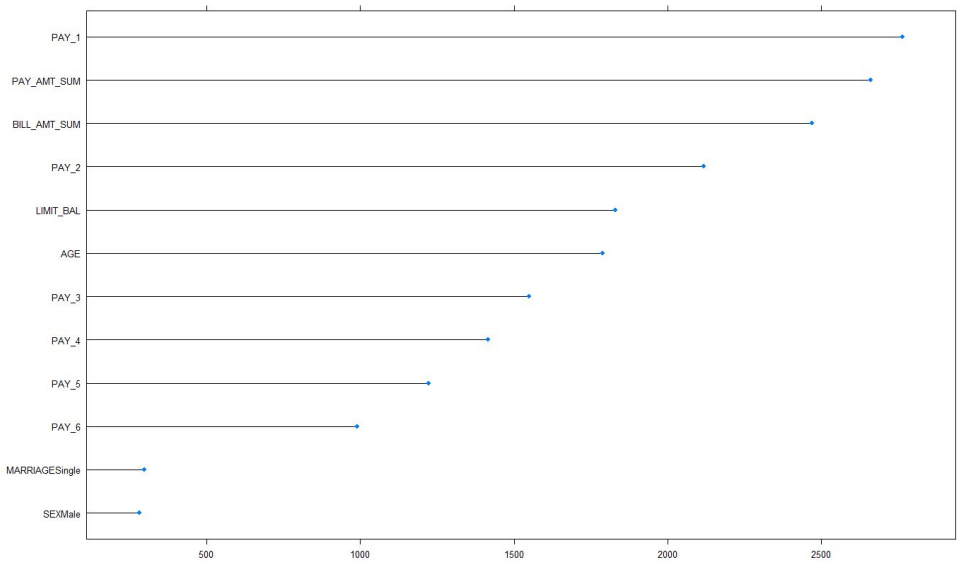


- **Accuracy will NOT be used.** As this is an imbalanced dataset**.** Model cannot be evaluated using accuracy. It is easy to get HIGH Accuracy score in imbalanced dataset.
- **Sensitivity** will be used to evaluate how good or bad a model is . Sensitivity helps us identify **proportion of TRUE POSITIVES that are correctly identified.**
- **Misclassification of defaulters** will **cost heavily** to the bank compared to misclassification of non defaulters. So, the focus is to have the best Sensitivity score.



ROC of Random FOrest

Important Variables



Top 5 Important Variables

- Pay_1
- Pay Amount Sum
- Bill Amount Sum
- Pay_2
- Limit Balance

	Logistic Regression	KNN	Naive Bayes	CART	Random Forest
Accuracy	69.20%	81.70%	76.40%	78.20%	90.00%
Sensitivity	70.80%	87.90%	82.30%	79.40%	88.60%
Specificity	66.90%	75.90%	70.60%	76.70%	91.60%
Precision	75.90%	77.70%	73.20%	81.10%	93.30%
F1	73.30%	82.50%	77.50%	80.20%	90.70%

Model Comparison. Random Forest has performed the best in predicting defaults.

Insights from Analysis

- Recent Month(Sep 05) Payments are the most important
- Credit behaviour, which shows their delay status, is a good indicator for Default.
- When payment is delayed more than 2 months, the chances of default goes higher

- When payment is delayed more than 2 months, the chances of default goes higher.
- Those with High School level have higher chance of Default.
- Those age range from 25 to 35, have lower Default rate.
- Those with higher Pay Amount Sum are less likely to default

Recommendations

- To lower the risk of default, must be very cautious on clients payment behaviour.
- More cautious of High School level clients.
- Marketing campaign should be aimed at clients' age from 25 to 35.

- As the model can predict defaulters. Bank rep can keep an eye on customers who are likely to default and contact them immediately as they default 1st payment. Ensure it doesn't become NPA

- Spread awareness regarding good credit history/score. Advantages of good credit score and disadvantages of bad credit score should be communicated with borrower on regular bases.

- Those age range from 25 to 35, have lower Default rate.

- Those with higher Pay Amount Sum are less likely to default

Thanks!

Any questions?

You can find me at https://in.linkedin.com/in/kulkarni-abhay