Capstone Project - 3

Credit Card Default Prediction

Presented By-
YASHWATI PATEL

Content:-

- Introduction
- Defining Problem Statement
- Data Summary
- Exploratory Data Analysis
- Feature Engineering
- Handling Imbalance data
- Model Creation
- Model Evaluation
- Challenge
- Conclusion

Introduction:-

Credit cards allow you to borrow money from the card issuer up to a certain limit to purchase items or withdraw cash. You probably have at least one credit card and one debit card in your wallet. When you use a debit card, the funds for the amount of your purchase are taken from your checking account in almost real time. When you use a credit card, the amount will be charged to your line of credit, meaning you will pay the bill at a later date, which also gives you more time to pay.

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 can use the K-S chart to evaluate which customers will default on their credit card payments.

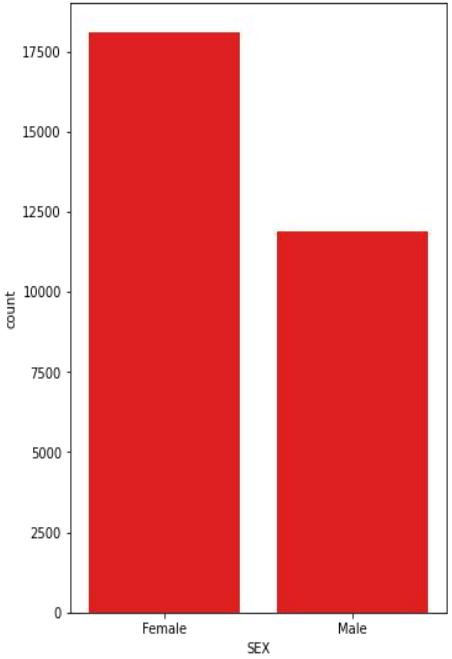
Data Summary:-

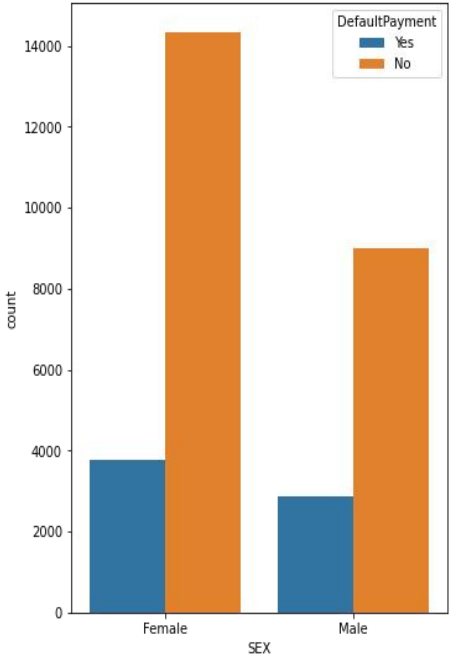
Attributes of Dataset:-

- X1 - Amount of credit(includes individual as well as family credit)
- X2 - Gender
- X3 - Education
- X4 - Marital Status
- X5 - Age
- X6 to X11 - History of past payments from April to September
- X12 to X17 - Amount of bill statement from April to September
- X18 to X23 - Amount of previous payment from April to September
- Y - Default payment

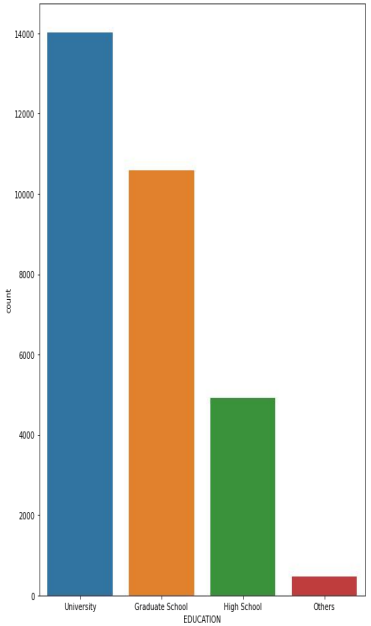
Exploratory Data Analysis:-

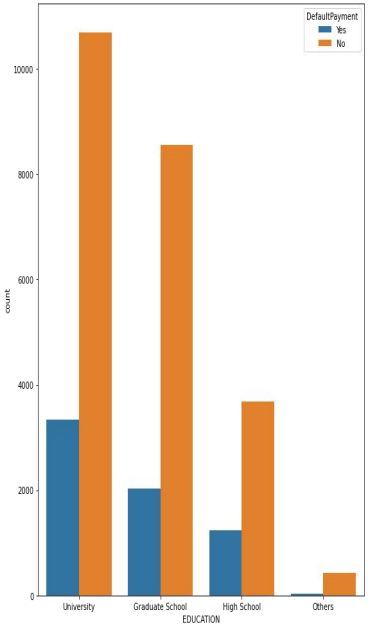
SEX:-



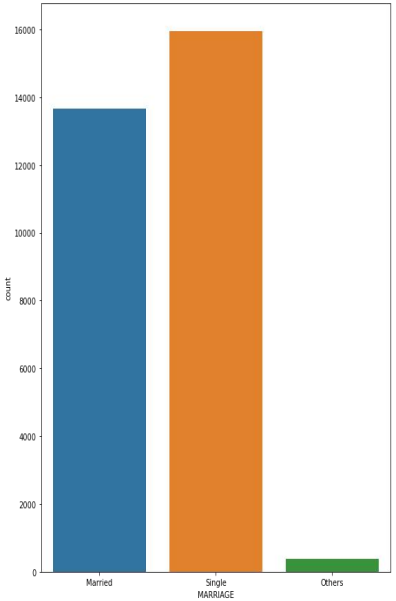


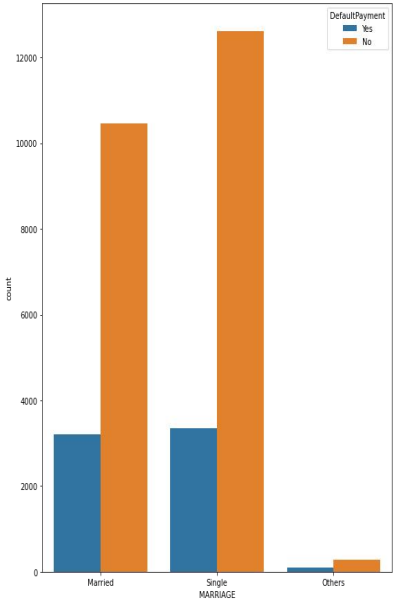
EDUCATION:-



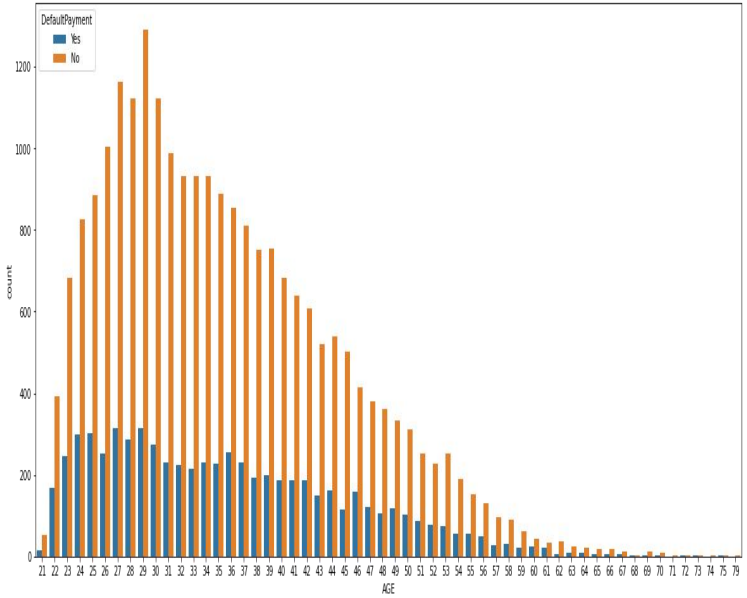


MARRIAGE:-

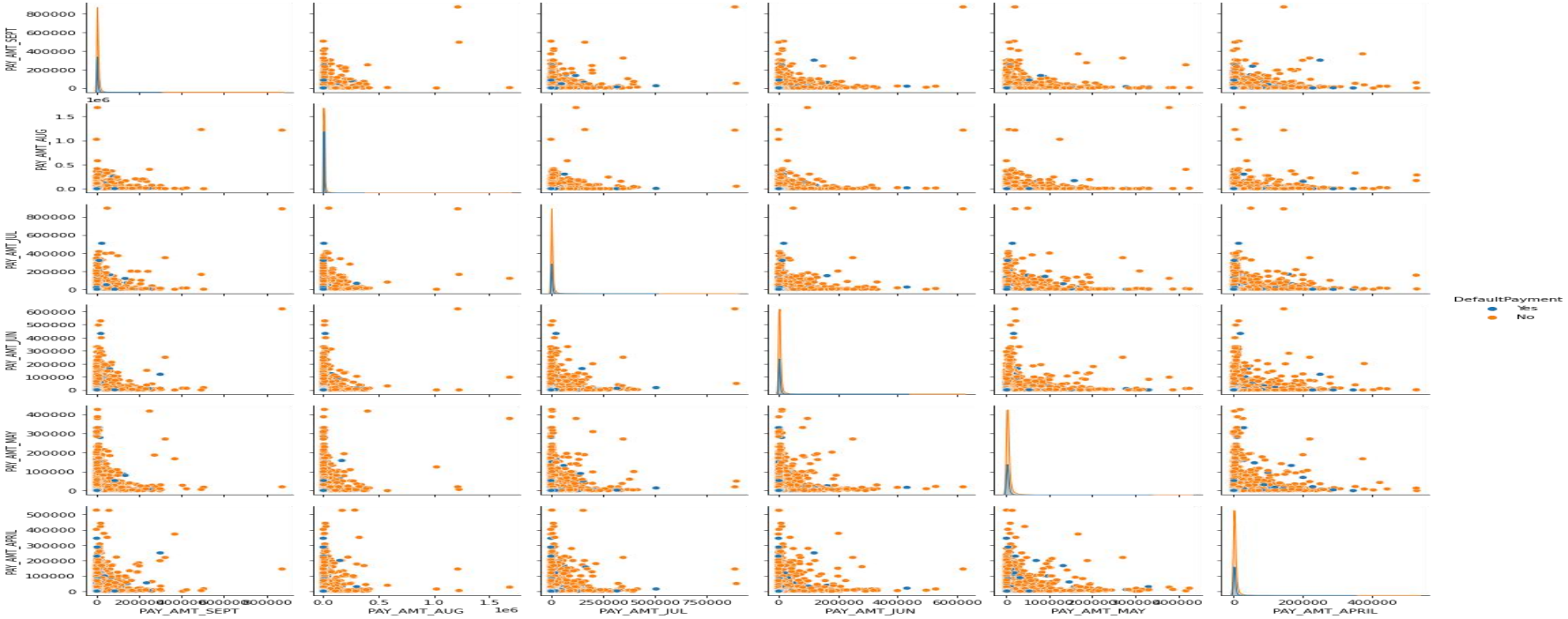




AGE:-



Paid Amount:-

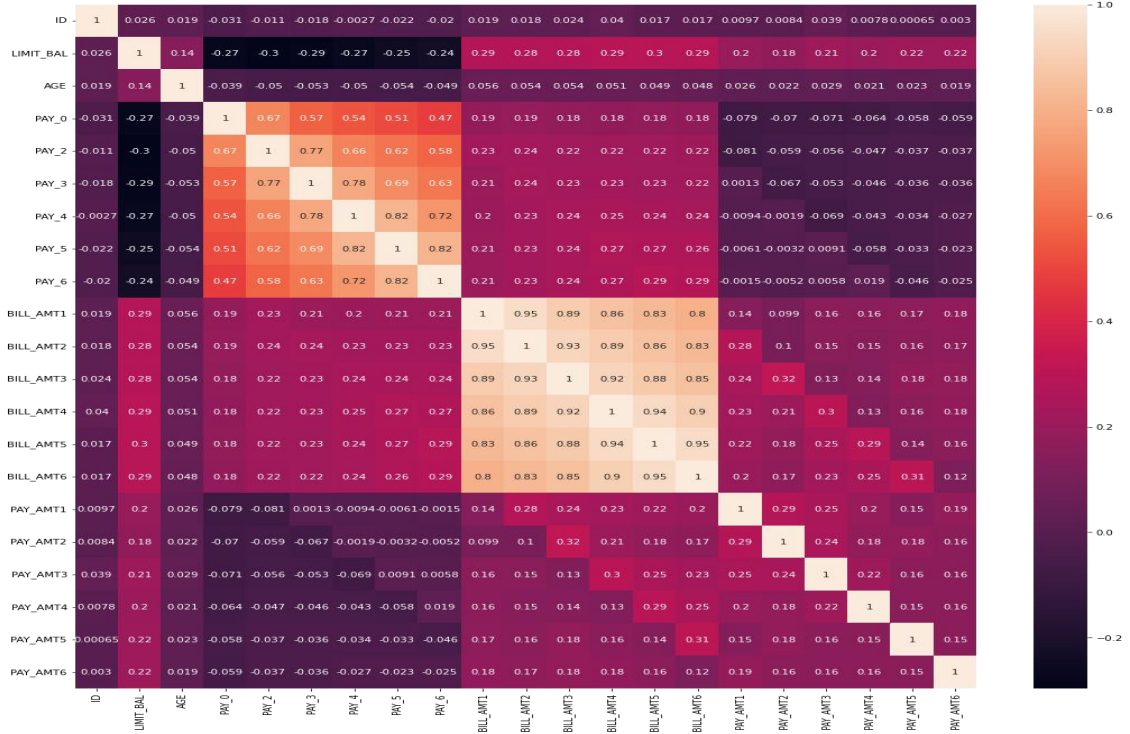

PAY_AMT_SEPT
PAY_AMT_AUG
PAY_AMT_JUL
PAY_AMT_JUN
PAY_AMT_MAY
PAY_AMT_APRIL
1e6
DefaultPayment
Yes
No

From above analysis we can see that:-

- Number of Male credit holder is less than Female.
- More number of credit card holders are university students followed by
- Graduates and then High school students.
- More number of credit cards holder are Single.
- We can see more number of credit cards holder age are between 26-30 years old.

Feature Engineering:-

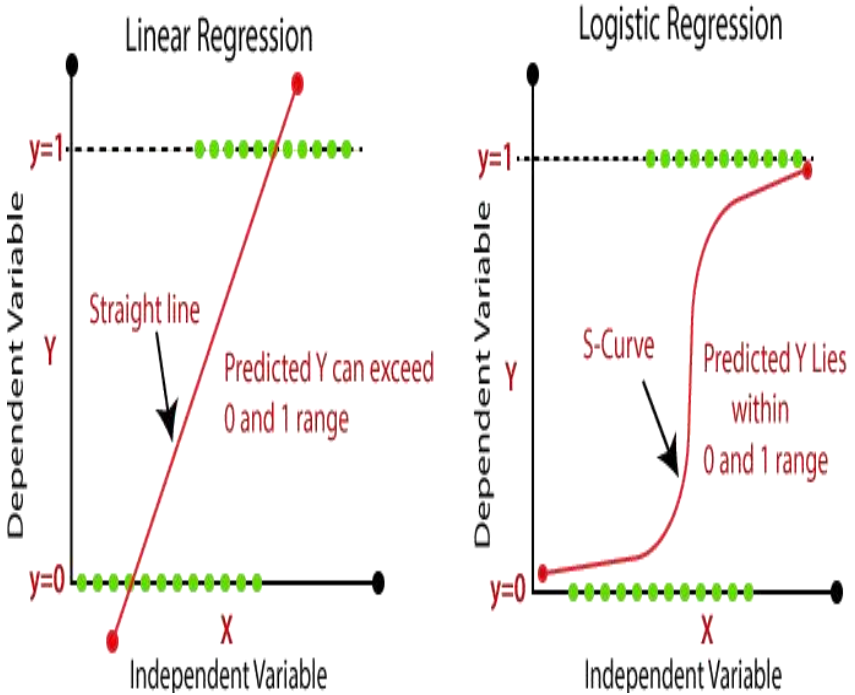
- Here We draw heatmap to find correlation between different independent variable and dependent variable.
- If correlation between independent variable are high and has very less relation with dependent variable ,remove them.



Model Creation:-

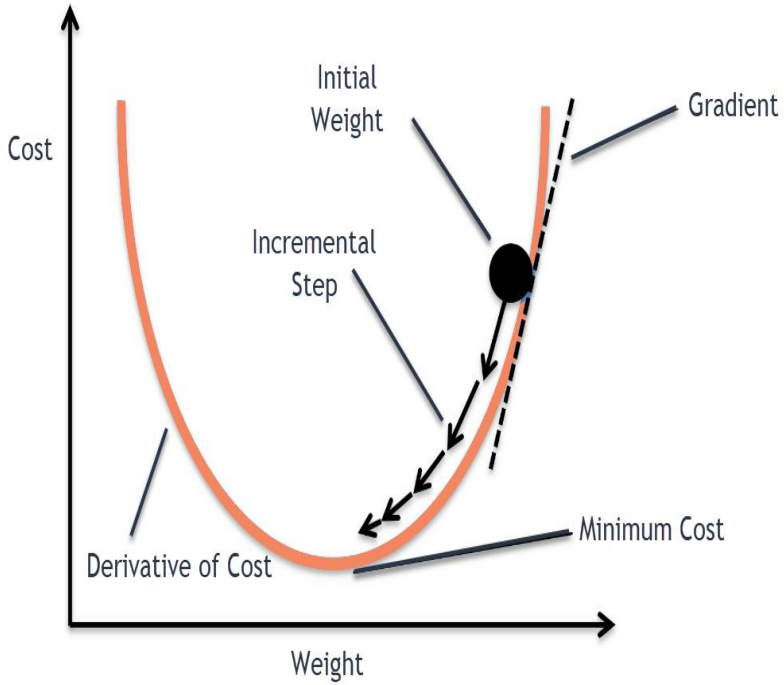
Logistic Regression:-

Logistic Regression used to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.



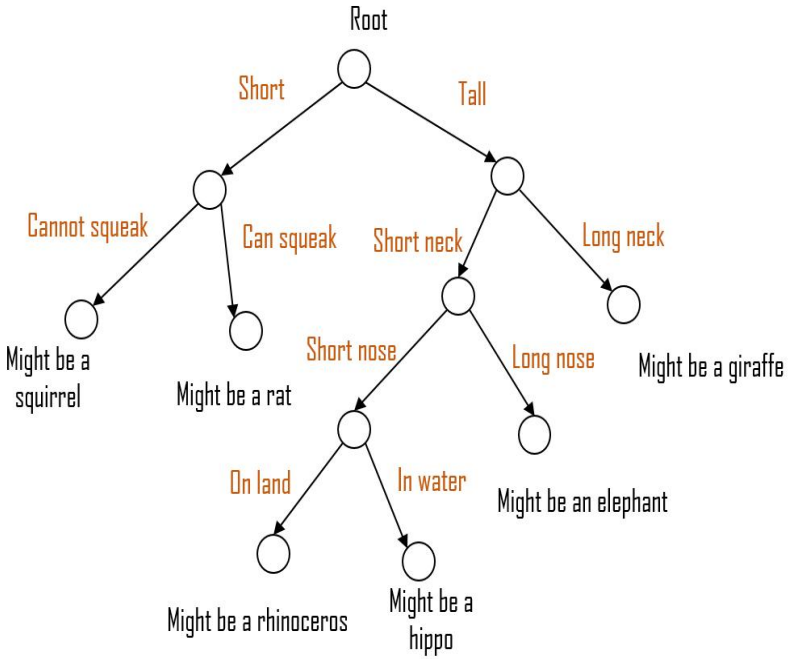
Stochastic Gradient Descent:-

Stochastic gradient descent is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs. It's an inexact but powerful technique. Stochastic gradient descent is widely used in machine learning applications



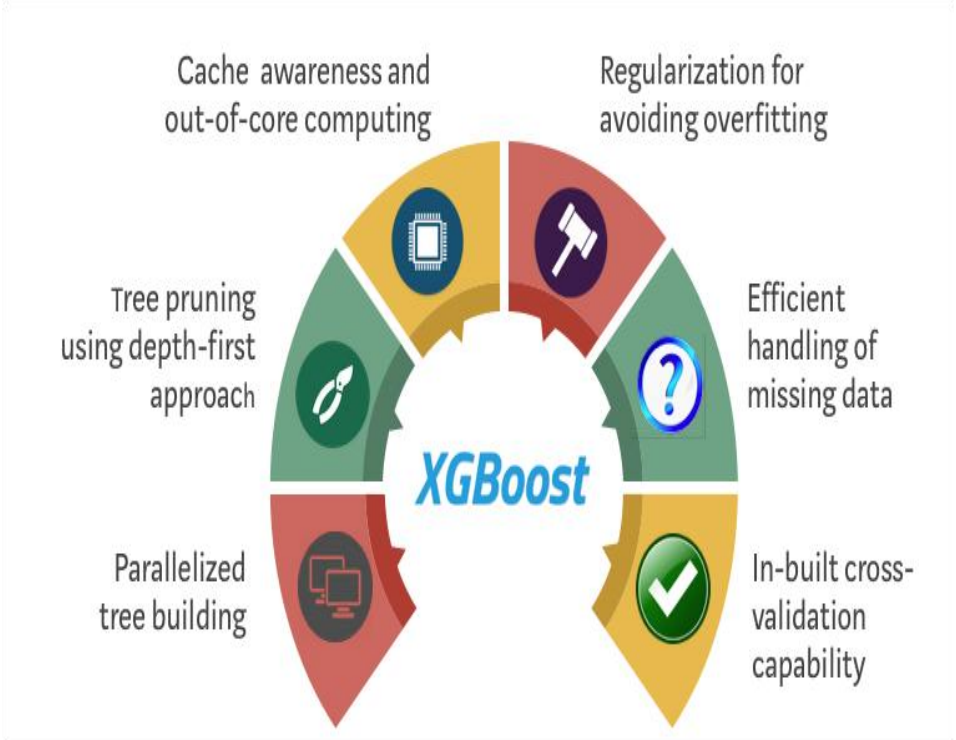
Decision Tree:-

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from training data. In Decision Trees, for predicting a class label for a record we start from the root of the tree.



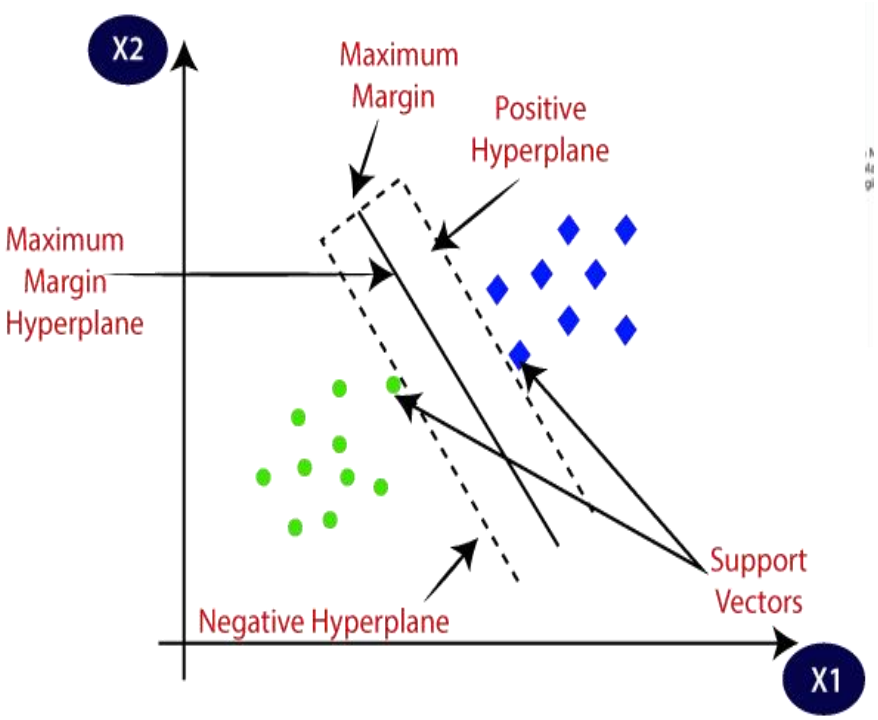
XGBOOST:-

It is generally used for very large dataset. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework.



Support Vector Machine:-

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

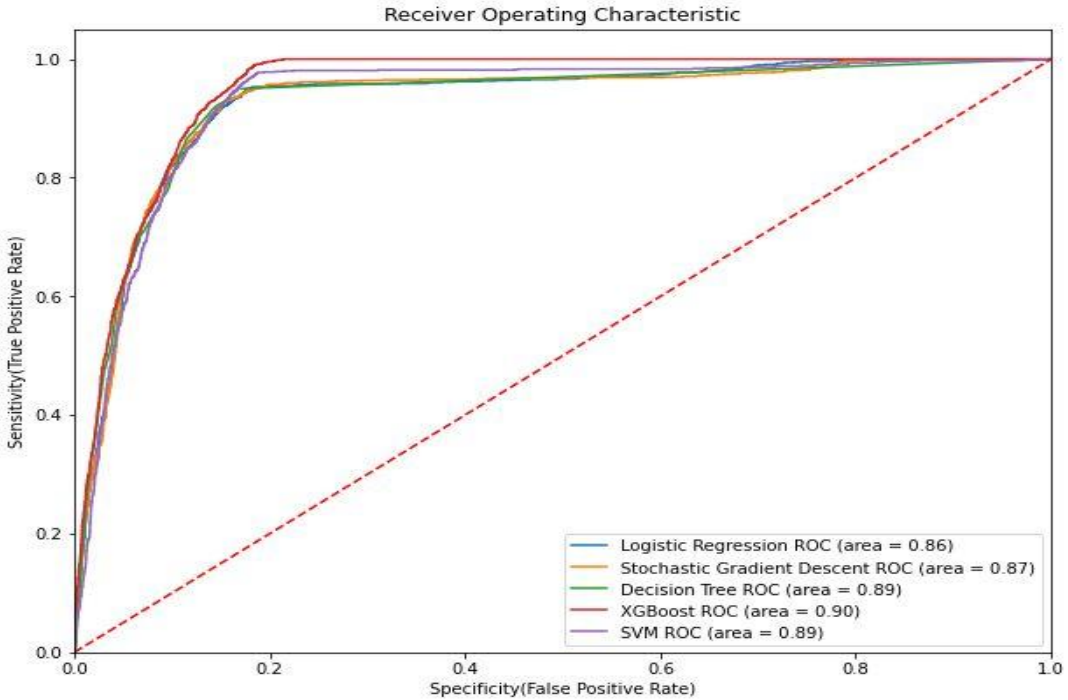


Model Evaluation:-

	Model	Accuracy	Precision	Recall	F1 Score	ROC Score
0	Logistic Regression	0.880648	0.559933	0.843645	0.673115	0.865301
1	Stochastic Gradient Descent	0.879674	0.556955	0.850334	0.673064	0.867505
2	Decision Tree Classifier	0.849105	0.490901	0.969900	0.651869	0.899205
3	Default XGBoost Classifier	0.852028	0.495907	0.962375	0.654535	0.897795
4	GridSearch XGBoost Classifier	0.872610	0.540409	0.838629	0.657274	0.858516

ROC Curve:-

Receiver Operating Characteristic summarizes the model's performance by evaluating the trade offs between true positive rate 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.



AI

Challenges:-

- Understanding the columns.
- Feature engineering.
- Getting a higher accuracy on the models.

Conclusion:-

- Here ,After that we build the Four models Logistic Regression, Decision Tree, Default XGBoost Classifier & Support Vector Machine. The best accuracy is obtained from the Default XGBoost Classifier.
- Consider the age of the applicant younger people are at higher risk of defaulting.Our best prediction accuracy was around 82-83%,Our best prediction accuracy was around 82-83%, our lowest measured prediction accuracy was about 79%.
- Consider the applicants marital status. Married people seem to default more often.
- We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis.

Thank You!