Capstone Project-3

Credit Card Default Prediction
By

Manas Ranjan Behera

Table Of Content :-

AI

1.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.
- So we will Predict whether a customer will default on his/her credit card.

2.Introduction:-

- Credit cards are usually small plastic cards with a unique number attached to an account.
- It impose the condition that cardholders pay back the borrowed money, plus any applicable interest, as well as any additional agreed-upon charges, either in full by the billing date or over time.

Introduction (Continued…)

- We have the dataset of credit card clients contains 30,000 rows and 25 features.
- And the features are
    - **ID**: Unique ID of each client
    - **LIMIT_BAL**: Amount of the given credit (NT dollar).
    - **Gender**: Gender of customer. (1 = male; 2 = female)
    - **Education**: Education qualification of customers. (1 = graduate school; 2 = university; 3 = high school; 4 = others)
    - **Marital Status**: Marital status of customer. (1 = married; 2 = single; 3 = others)

Introduction (Continued …)

- **Age**: Age of customer in years.
- **History of Past Payment**: (PAY) Repayment status in September, August, July, June, May and April 2005.
- **Amount of Bill Statement**: (BILL_AMT) Amount of bill statement in September, August, July, June, May and April 2005.
- **Amount of Previous Payment**:(PAY_AMT) Amount of previous payment in September, August, July, June, May and April 2005
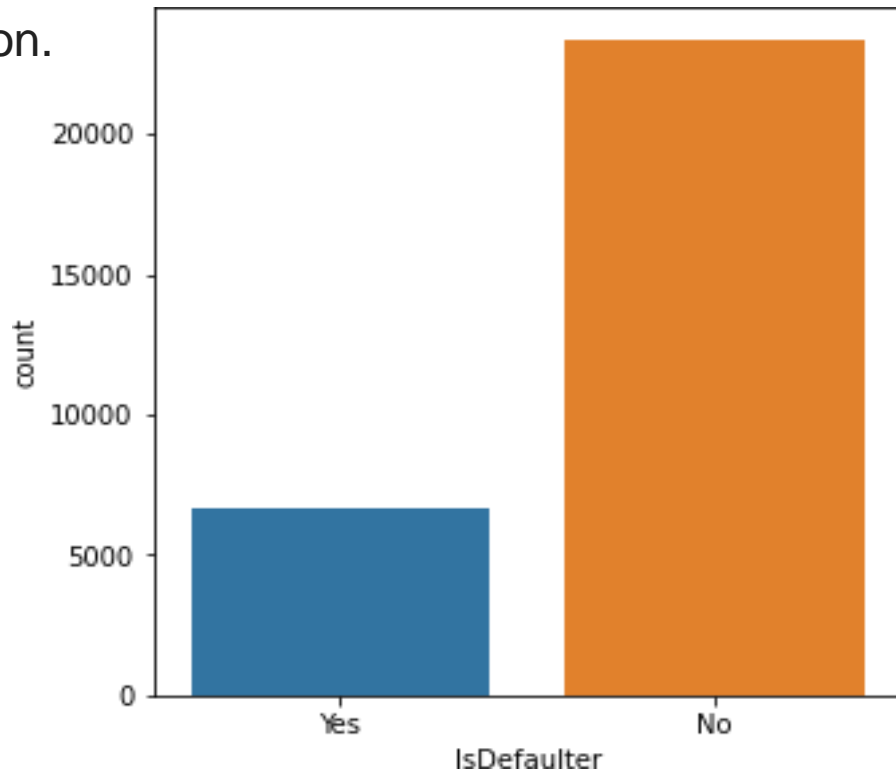
3.Data Cleaning :-

- Dataset contains no null values.
- No duplicates found
- Renamed the features
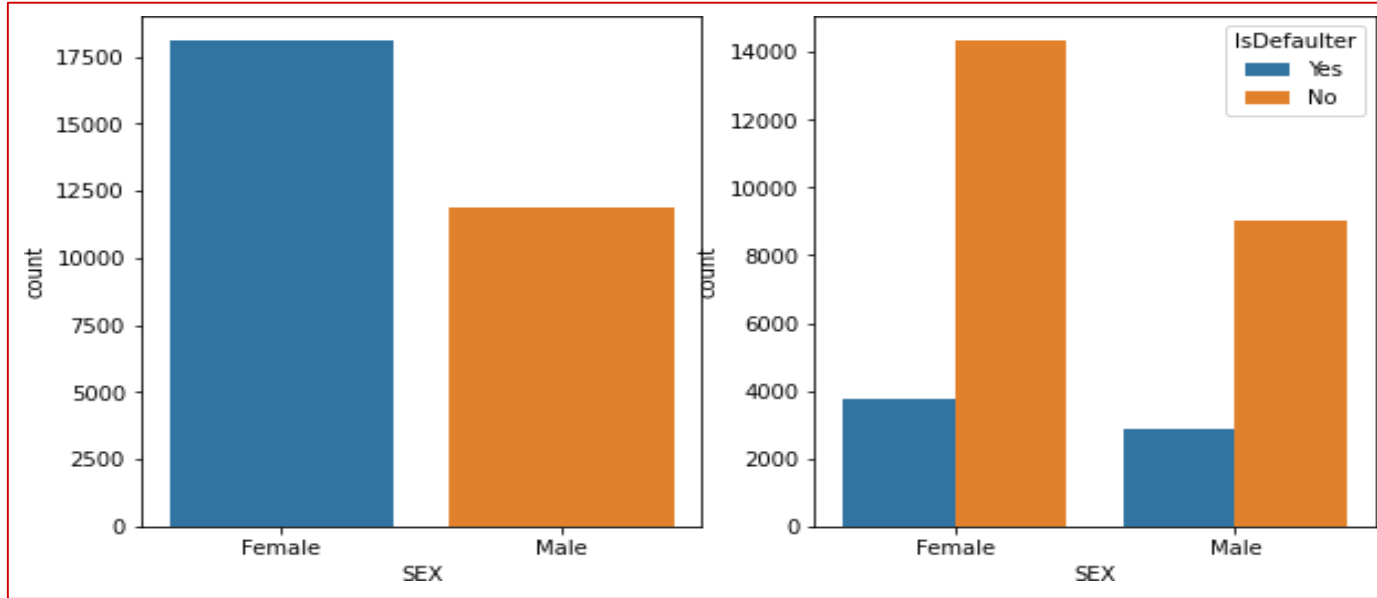- Replaced some columns with categorical values
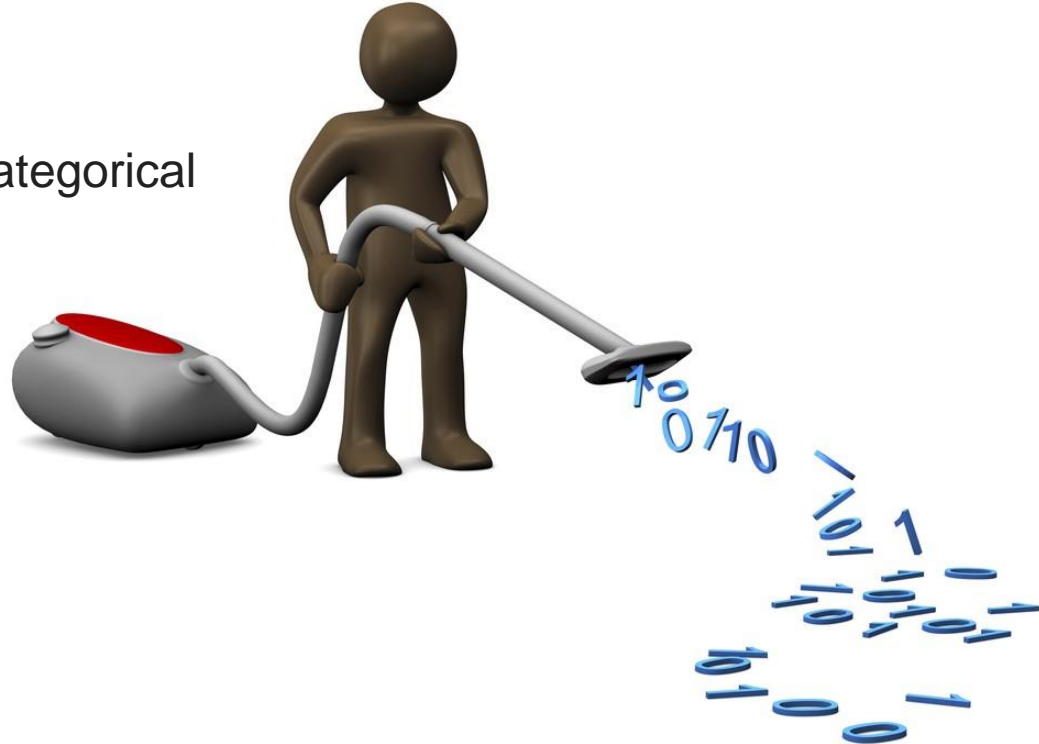
4. Exploratory Data Analysis (EDA)

- Both the classes are not in proportion.
- Which means that the dataset is imbalanced.
- Data balancing is required.

EDA (Continued)

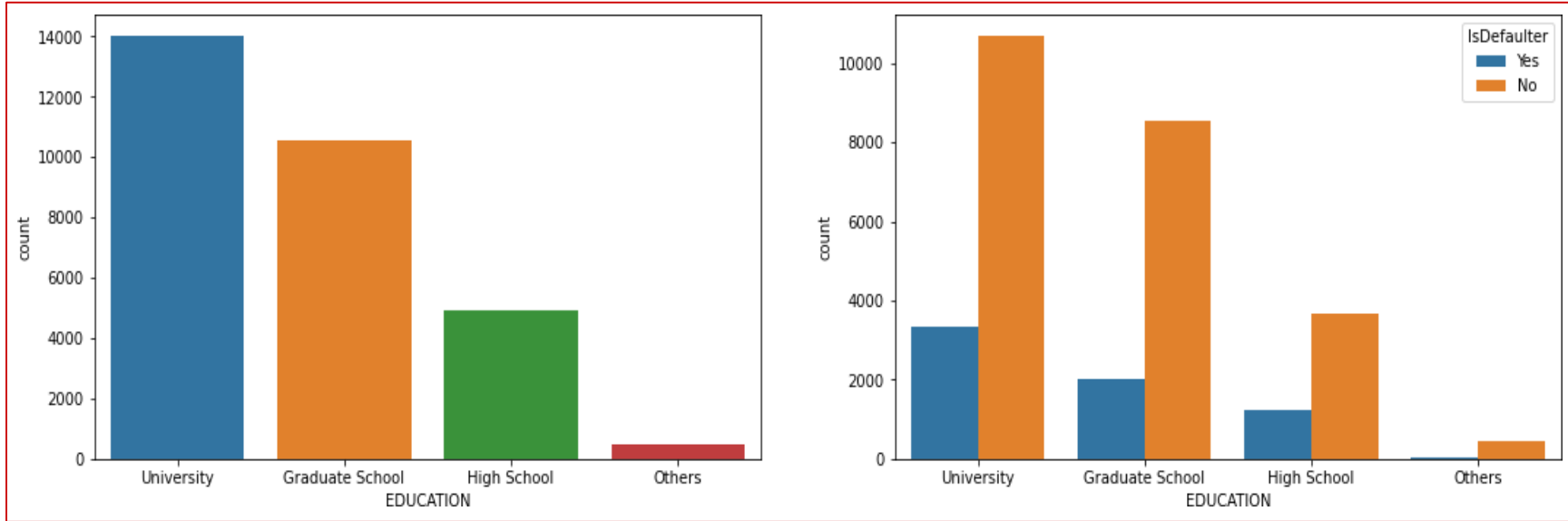- Female credit card holders are larger than the male credit card holders.
- As the number of female credit card holders is larger than the male, their credit card defaults are also higher than male

EDA (Continued)

- University and graduate school has maximum credit card holder.
- As the number of university and graduate school credit card holders is higher, their credit card default are also higher.

EDA (Continued) :-
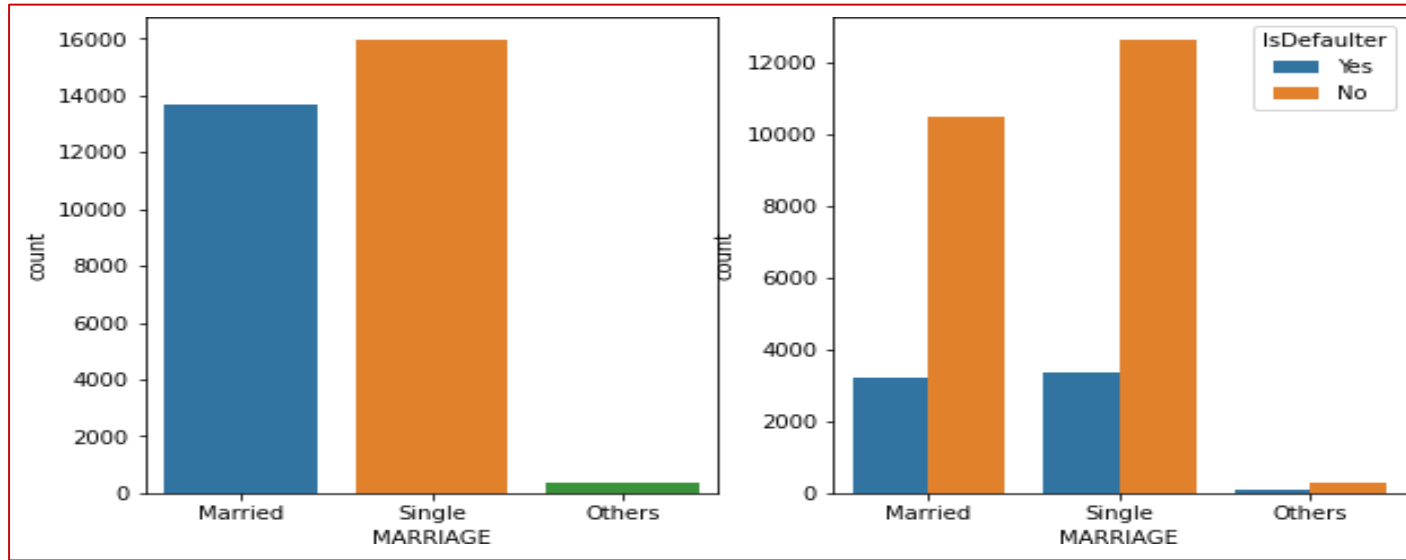
- Number of credit card holder is maximum in singles.
- But credit card defaults are almost same in case of single and married people

5.Handling Class Imbalance :-

To balance the proportion, we have used a technique called SMOTE.

SMOTE :-
- Also called Synthetic Minority Oversampling Technique
- It is the technique to make data class balanced.
- SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

Handling Class Imbalance (Continued)

- This the plotting of the dataset before and after the application of SMOTE
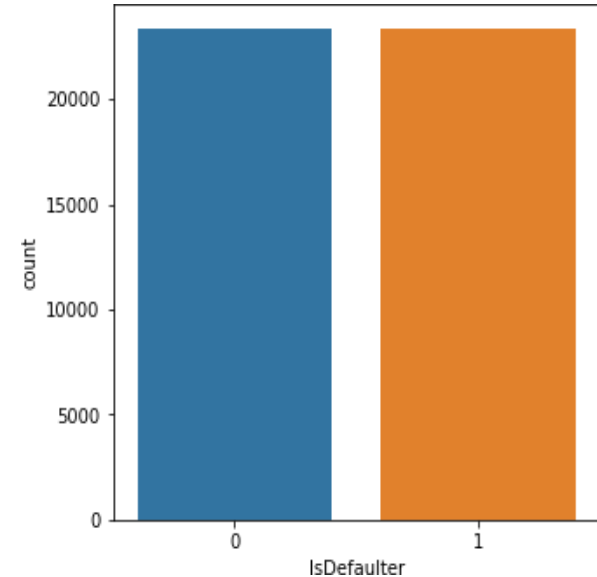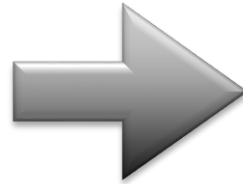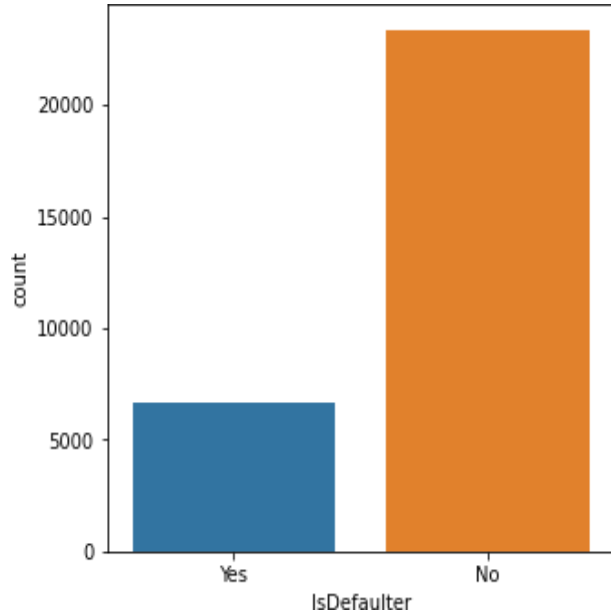
6.Transformation of Data :-

- To scale data into a uniform format that would allow us to utilize the data in a better way.
- For performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to dataset.

EXTRACT   **TRANSFORM**   **LOAD**

7.Splitting Data :-

- Splited the data into training dataset and testing dataset.

- Training dataset is for making algorithm learn and train model.

- Test dataset is for testing the performance of train model.

- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.

Dataset

Training Set	Test Set
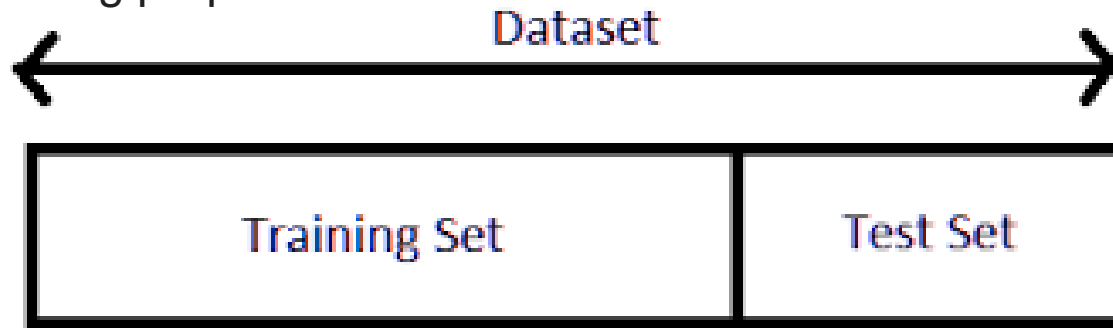
8.Fitting Different Model :-

Following classifier used for prediction credit card default:

- Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier

- Support Vector Machine

- Gradient Boosting

- XG Boosting

9.Cross Validation & Hyperparameter Tuning :-

● It is a resampling procedure used to evaluate machine learning models on a limited data sample.

● Basically, Cross Validation is a technique mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

● Tuning the hyper parameters of respective algorithms is necessary for getting better accuracy and to avoid over fitting.

Logistic Regression :-

- Logistic regression is a machine learning algorithm for classification problem.

- In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.
- It is most useful for understanding the influence of several independent variables on a single outcome variable.

LOGISTIC REGRESSION						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.828	0.832	0.795	0.857	0.825	0.833
Tuned Model	0.827	0.832	0.799	0.855	0.826	0.833

Decision Tree Classifier :-

- Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.
- Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data.

Decision Tree Classifier						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	1	0.791	0.810	0.781	0.795	0.792
Tuned Model	0.837	0.824	0.779	0.857	0.816	0.827

Random Forest Classifier :-

- Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting.
- The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Random Forest Classifier						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	1	0.863	0.821	0.897	0.857	0.866
Tuned Model	0.844	0.833	0.794	0.860	0.826	0.835

Support Vector Machine :-

- Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible.
- New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Random Forest Classifier						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.848	0.840	0.765	0.900	0.827	0.848
Tuned Model	0.846	0.841	0.768	0.900	0.829	0.849

Gradient Boosting :-

- It is a technique of producing an additive predictive model by combining various weak predictors, typically Decision Trees.
- Due to this sequential connection, boosting algorithms are usually slow to learn, but also highly accurate.
- The final model aggregates the result of each step and thus a strong learner is achieved.

Gradient Boosting						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.847	0.843	0.801	0.875	0.836	0.846
Tuned Model	0.951	0.866	0.824	0.899	0.860	0.868

XG Boosting :-

- XG Boost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
- It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time.

Gradient Boosting						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.847	0.843	0.799	0.877	0.836	0.846
Tuned Model	0.995	0.871	0.831	0.904	0.866	0.874

10.Comparison of Model :-

AI

	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
11	Optimal XG Boosting	0.995	0.871	0.831	0.904	0.866	0.874
2	Random Forest	0.999	0.867	0.832	0.895	0.862	0.869
10	Optimal Gradient Boosting	0.951	0.866	0.824	0.899	0.860	0.868
3	SVM	0.846	0.841	0.768	0.900	0.829	0.849
9	Optimal SVM	0.846	0.841	0.768	0.900	0.829	0.849
4	Gradient Boosting	0.845	0.845	0.801	0.878	0.838	0.848
5	XG Boosting	0.847	0.844	0.801	0.877	0.837	0.847
8	Optimal Random Forest	0.844	0.833	0.794	0.860	0.826	0.835
0	Logistic Regression	0.827	0.832	0.796	0.857	0.826	0.834
6	Optimal Logistic Regression	0.826	0.832	0.797	0.857	0.826	0.834
7	Optimal Decision Tree	0.841	0.825	0.779	0.858	0.817	0.828
1	Decision Tree	1.000	0.802	0.814	0.795	0.804	0.802

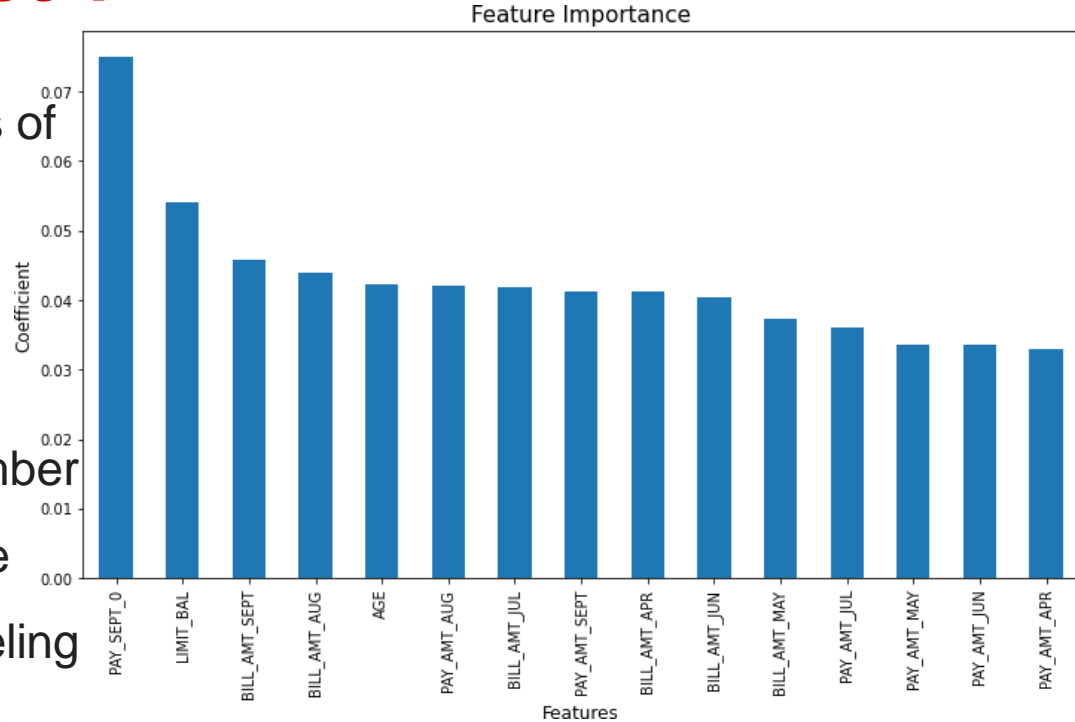- XG Boost shows highest test accuracy score of 87% and AUC score is 0.874.

11.Combined ROC Curve :-

- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.

- An ROC curve plots TPR vs. FPR at different classification thresholds.

- Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

Combined ROC Curve
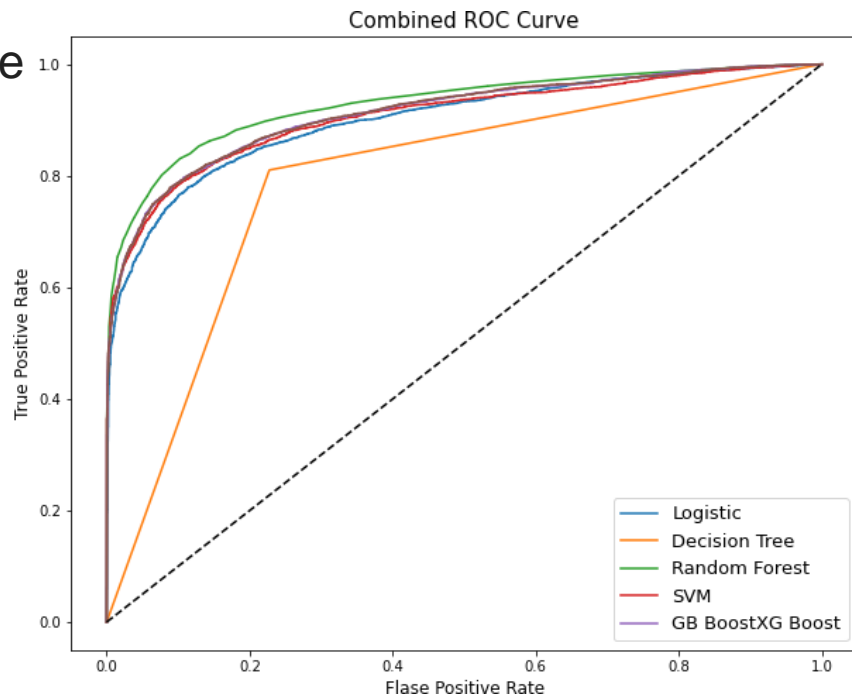
12.Feature Importance :-

- Feature selection is the process of reducing the number of input variables when developing a predictive model.

- It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

Feature Importance

13.Conclusion :-

- From all baseline model, Random Forest classifier shows highest test accuracy and F1 score and AUC.

- Baseline model of Random Forest and decision tree shows huge difference in train and test accuracy which shows over fitting.

- After cross validation and hyper parameter tuning, XG Boost shows highest test accuracy score of 87% and AUC is 0.874.

- Cross validation and hyper parameter tuning certainly reduces chances of over fitting and also increases performance of model.

Thank You