Capstone Project – 3

Supervised Machine Learning
-Classification

Credit Card Default Prediction

Kathiravan.R

POINTS OF DISCUSSION:

1. Problem Statement
2. Introduction
3. Sample Data
4. Data Cleaning
5. Heatmap
6. Exploratory Data Analysis
7. Handling Class Imbalance
8. Transformation of Data
9. Splitting Data
10. Fitting Different Model
11. Cross Validation & Hyperparameter Tunning
10. Comparison of Model
11. Combined ROC Curve
12. Feature Importance
13. Conclusion

1. PROBLEM STATEMENT:

Predicting whether a customer will default on his/her credit card

Problem Statements

2. INTRODUCTION:

- **ID:** Unique ID of each client
- **LIMIT_BAL:** Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- **SEX:** Gender. (1 = male; 2 = female)
- **EDUCATION:** Education qualification of customers. (1 = graduate school; 2 = university; 3 = high school; 0,4,5,6 = others)
- **MARRIAGE:** Marital status. (0 = others, 1 = married, 2 = single, 3 = others)
- **AGE:** Age in years.

2. INTRODUCTION:

- **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.

Approach Overview

Understanding and Cleaning

- Find information on documented columns values
- Clean data to get it ready for Analysis

Graphical

- Examining the data with visualization

Machine Learning

- Logistic
- SVM
- Random Forest
- XGBoost

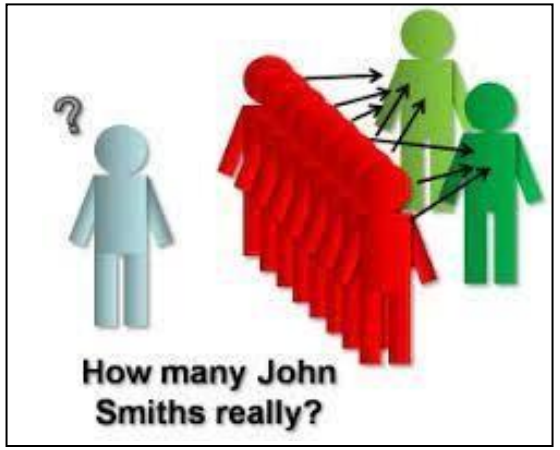
3.SAMPLE DATA

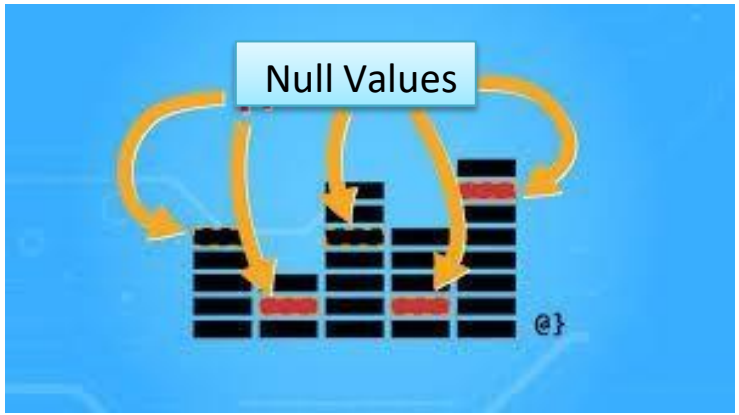
- The given dataset consist of 3000 rows and 25 columns

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY
0	1	20000	2	2	1	24	2	2	-1	-1	...	0	0	0	0	689	0	0	0	
1	2	120000	2	2	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000	1000	0	
2	3	90000	2	2	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000	1000	
3	4	50000	2	2	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100	1069	
4	5	50000	1	2	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000	9000	689	

4. DATA CLEANING

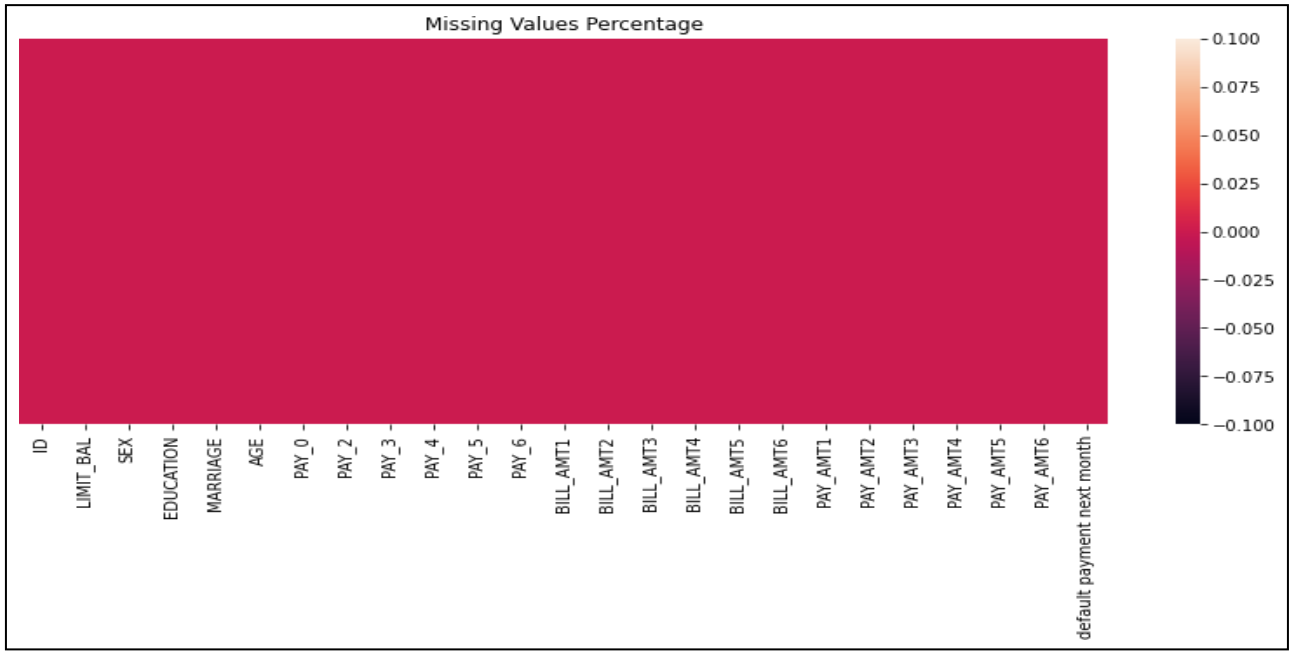
- Null Values Treatment
- Duplicate Values Treatment





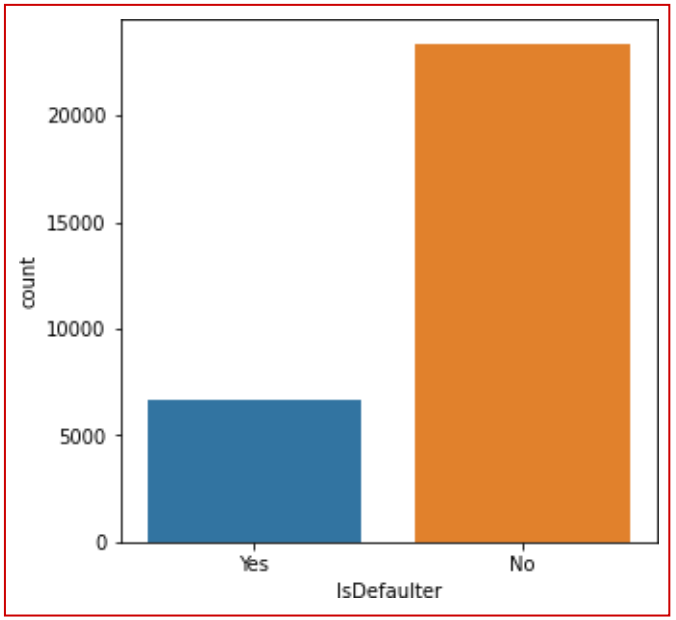
5. HEATMAP

- We can see no NA values, null values and duplicate values are present in data.



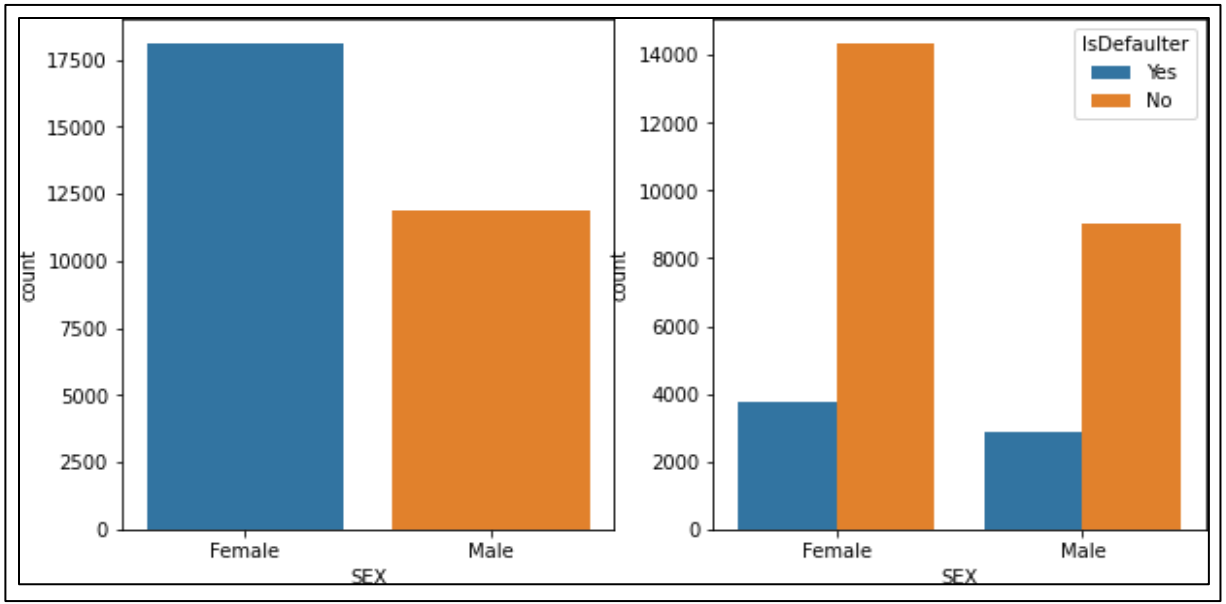
6. Exploratory Data Analysis (EDA)

- EDA is process on data to discover patterns, to spot anomalies and to check assumptions with help of statistical summary and graphical representation.
- Here we can see defaulters are less as compare to Non defaulter in the given dataset
- Which means that dataset imbalanced.
- Data balancing is required.



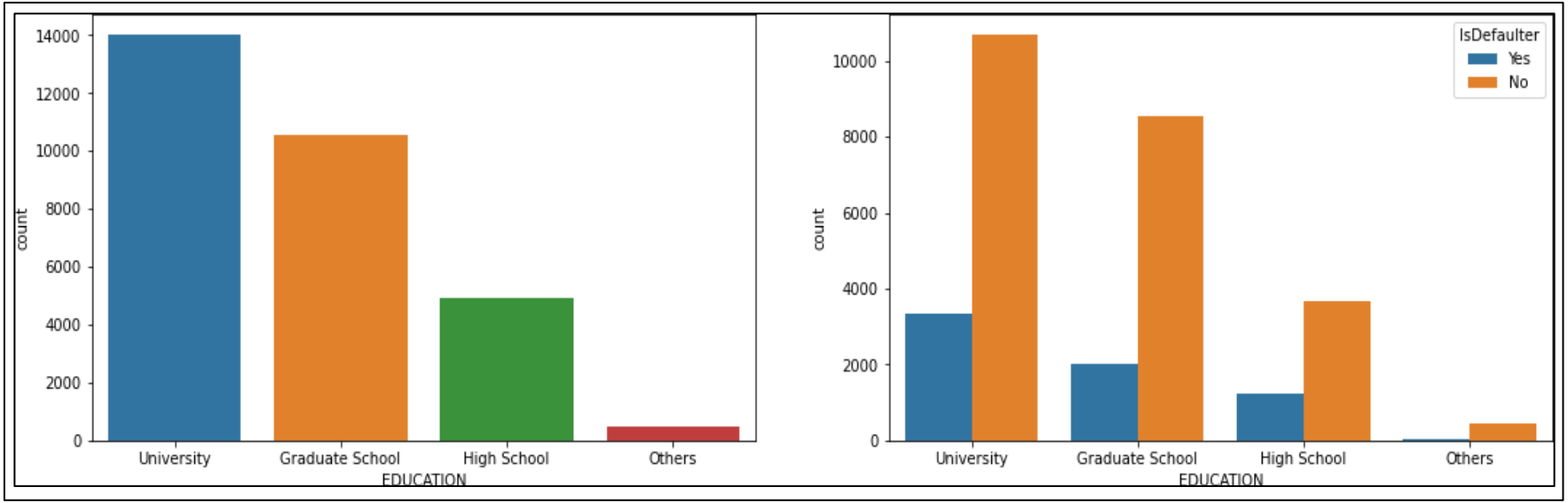
EDA

- Data analysis of category: SEX
- Female credit card holders are larger than male credit cards holders.
- As the number female credit card holder is larger than male, their credit card defaults are also higher than male.



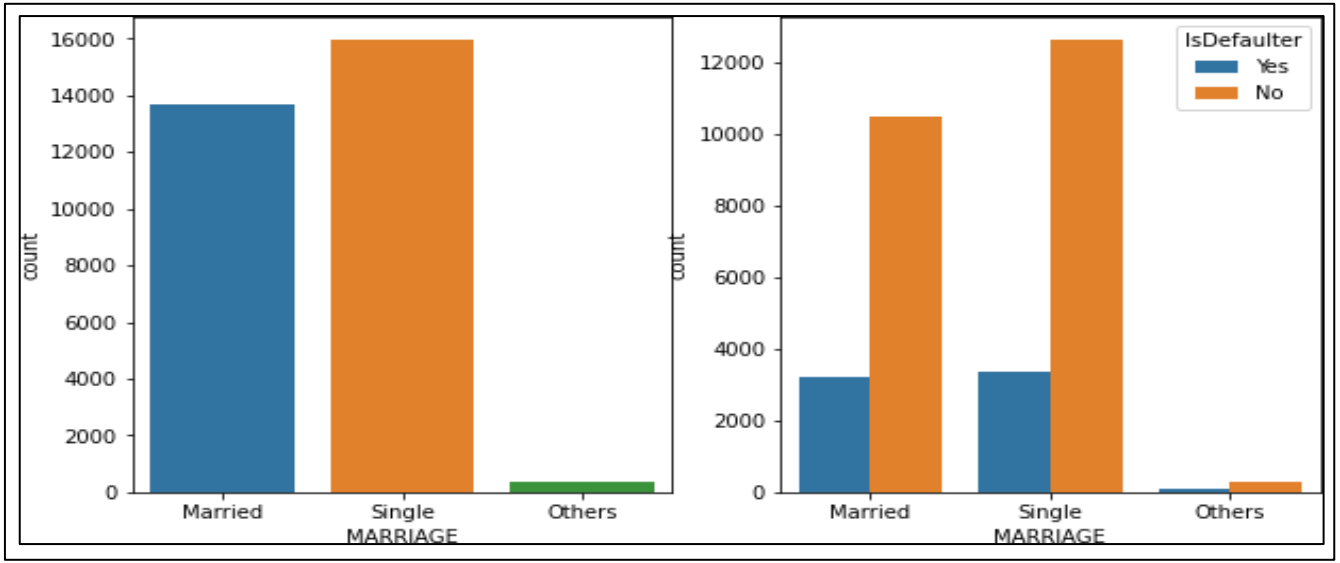
EDA

- Data analysis of category: EDUCATION
- University and graduate school has maximum credit card holder.
- As the number university and graduate school credit card holder is higher their credit card default are also higher.



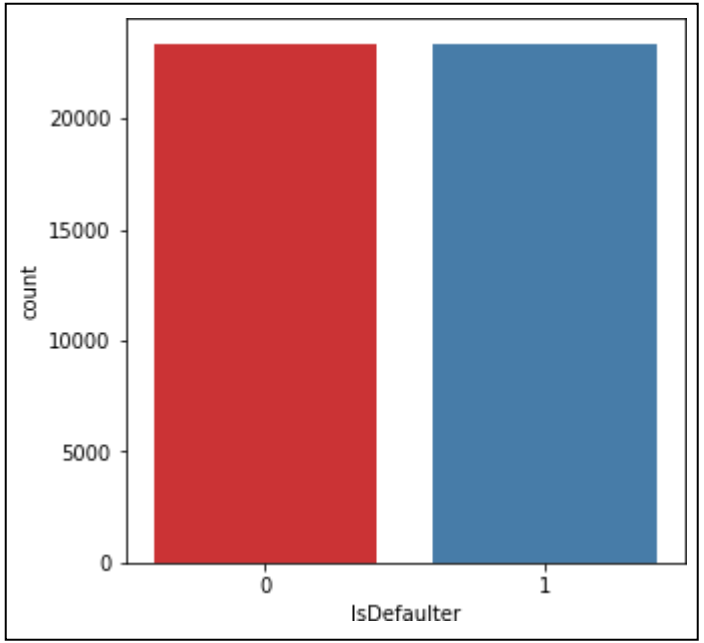
EDA

- Data analysis of category: MARRIAGE
- Number of credit card holder is maximum in singles.
- But credit card defaults are almost same in case of single and married people.



7. Handling Class Imbalance

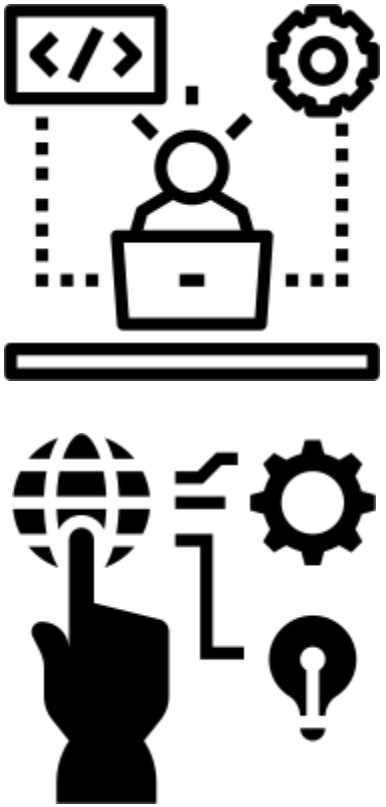
- Both the classes are not in proportion.
- SMOTE (Synthetic Minority Oversampling Technique) is a widely used resampling technique to handle class imbalance problem.
- SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



- After applying SMOTE.
- Data class is balanced now.

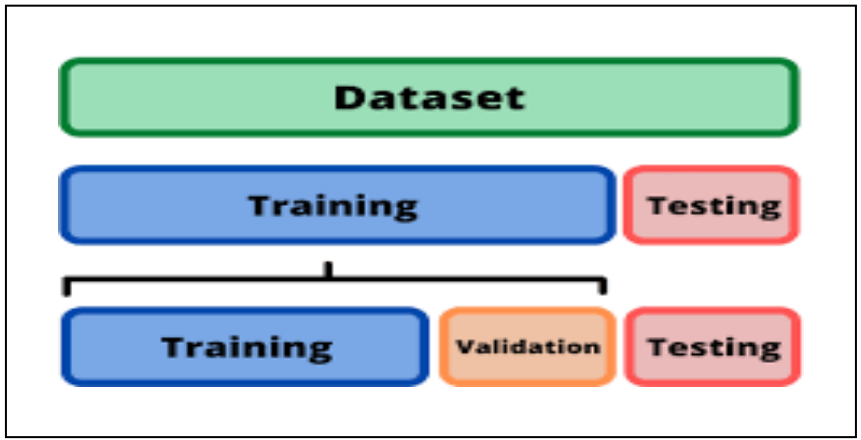
8. Transformation of Data

- 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.



9. Splitting Data

- Data splits 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.



10. 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

11 Cross Validation &

- It is a resampling procedure used to evaluate machine learning models on a limited data sample.
- Basically, Cross Validation is a technique using which Model is evaluated on the dataset on which it is not trained that is it can be a test data or can be another set as per availability or feasibility.
- Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.

10.1 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 REGRESSIO				
	Accuracy		Precision	Reca
	Train	Test		

10.2 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	0.794	0.797	0.792	0.794	0.794
Tunned Model	0.84	0.824	0.779	0.856	0.816	0.827

10.3 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	0.999	0.868	0.829	0.9	0.863	0.871
Tunned Model	0.844	0.833	0.794	0.860	0.826	0.835

10.4 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.

Support Vector Machine						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.845	0.840	0.767	0.898	0.827	0.847
Tunned Model	0.846	0.841	0.768	0.900	0.829	0.849

10.5 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	Rec
	Train	Test		
Baseline Model	0.847	0.843	0.801	0.8

10.6 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.

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

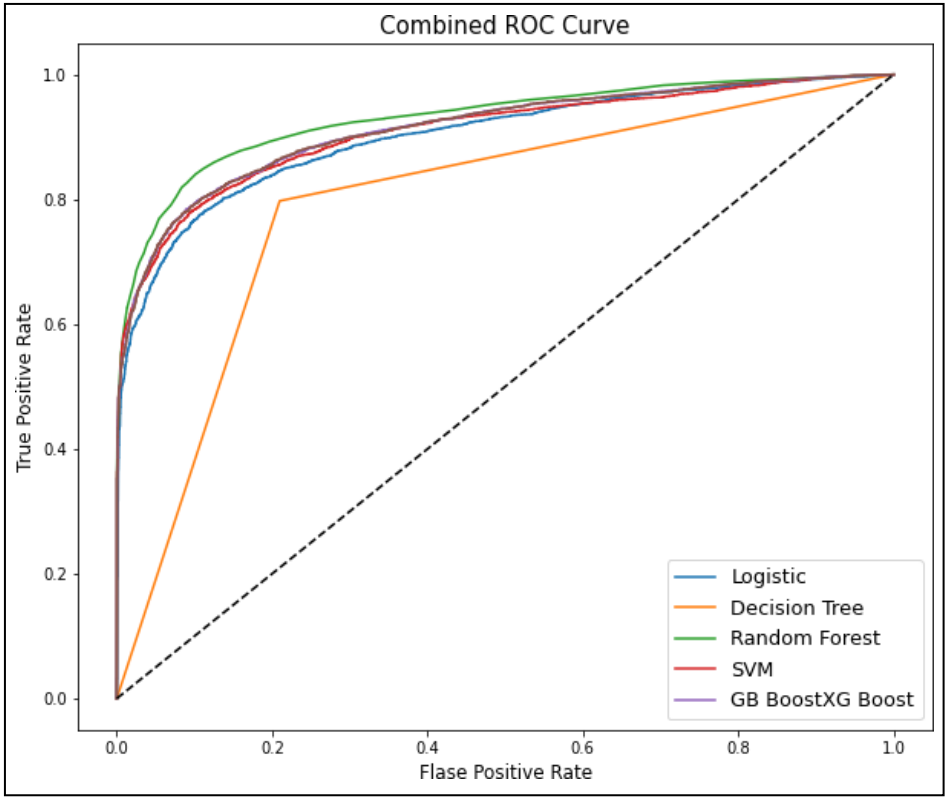
12. Comparison of Model

	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
11	Optimal XG Boosting	0.998	0.870	0.831	0.901	0.864	0.872
2	Random Forest	0.999	0.868	0.829	0.900	0.863	0.871
10	Optimal Gradient Boosting	0.983	0.868	0.830	0.899	0.863	0.871
4	Gradient Boosting	0.846	0.846	0.804	0.878	0.839	0.848
5	XG Boosting	0.846	0.845	0.804	0.877	0.839	0.848
3	SVM	0.845	0.840	0.767	0.898	0.827	0.847
9	Optimal SVM	0.847	0.839	0.765	0.899	0.826	0.847
8	Optimal Random Forest	0.843	0.833	0.796	0.860	0.827	0.835
0	Logistic Regression	0.826	0.832	0.796	0.858	0.826	0.834
6	Optimal Logistic Regression	0.826	0.832	0.796	0.858	0.825	0.834
7	Optimal Decision Tree	0.840	0.824	0.779	0.856	0.816	0.827
1	Decision Tree	1.000	0.794	0.797	0.792	0.794	0.794

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

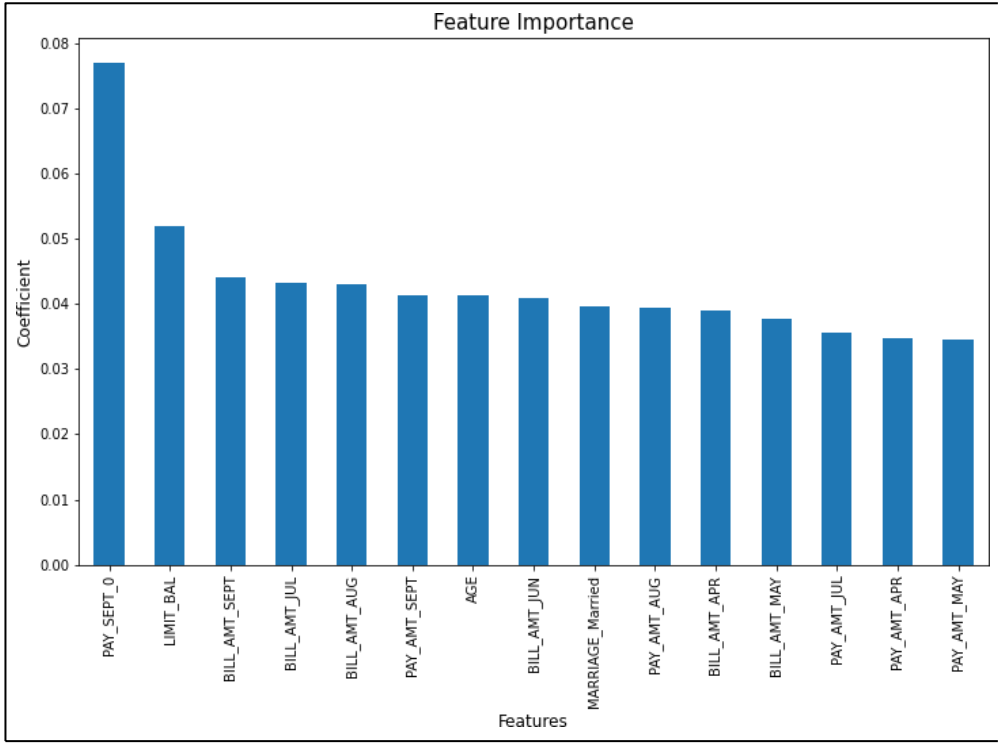
13. Combined ROC Curve

- 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.



14. 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.



15. Conclusion

1. From all baseline model, Random Forest classifier shows highest test accuracy and F1 score and AUC.
2. Baseline model of Random Forest and decision tree shows huge difference in train and test accuracy which shows overfitting.
3. After cross validation and hyperparameter tunning, XG Boost shows highest test accuracy score of 87.10% and AUC is 0.874.
4. Cross validation and hyperparameter tunning certainly reduces chances of overfitting and also increases performance of model.