Capstone Project-3

Credit Card Default Prediction

(Supervised Machine Learning –Classification)

BY

Nehal S Jambhulkar

Content

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

Problem Statement:

•Predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification- credible or not credible clients.

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

Buisness Understanding

As more and more consumers rely on the credit card to pay their everyday purchases in online and physical retail store, the amount of issued credit cards and the overwhelming amount of credit card debt by the cardholders have rapidly increased. Therefore, most financial institutions have to deal with the issues of credit card default in addition to the credit card fraud .

In this project, we have built a machine learning model that predicts whether a certain individual will be a defaulter. This prediction is influenced by a variety of aspects which we will be exploring in subsequent sections. Credit risk of an individual can be calculated based on a number of factors such as - age, gender, education, employment status and marital status, etc.

Features Summary

- **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)

- **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.

Data Cleaning

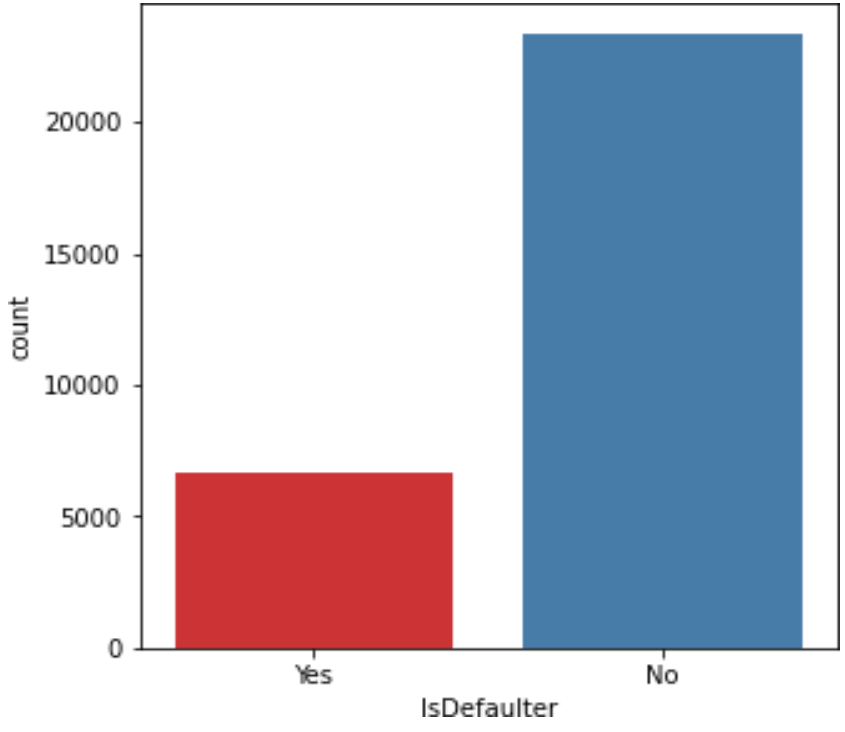
- Check for null value and missing value- No null value is present

- No Duplicate Values are present

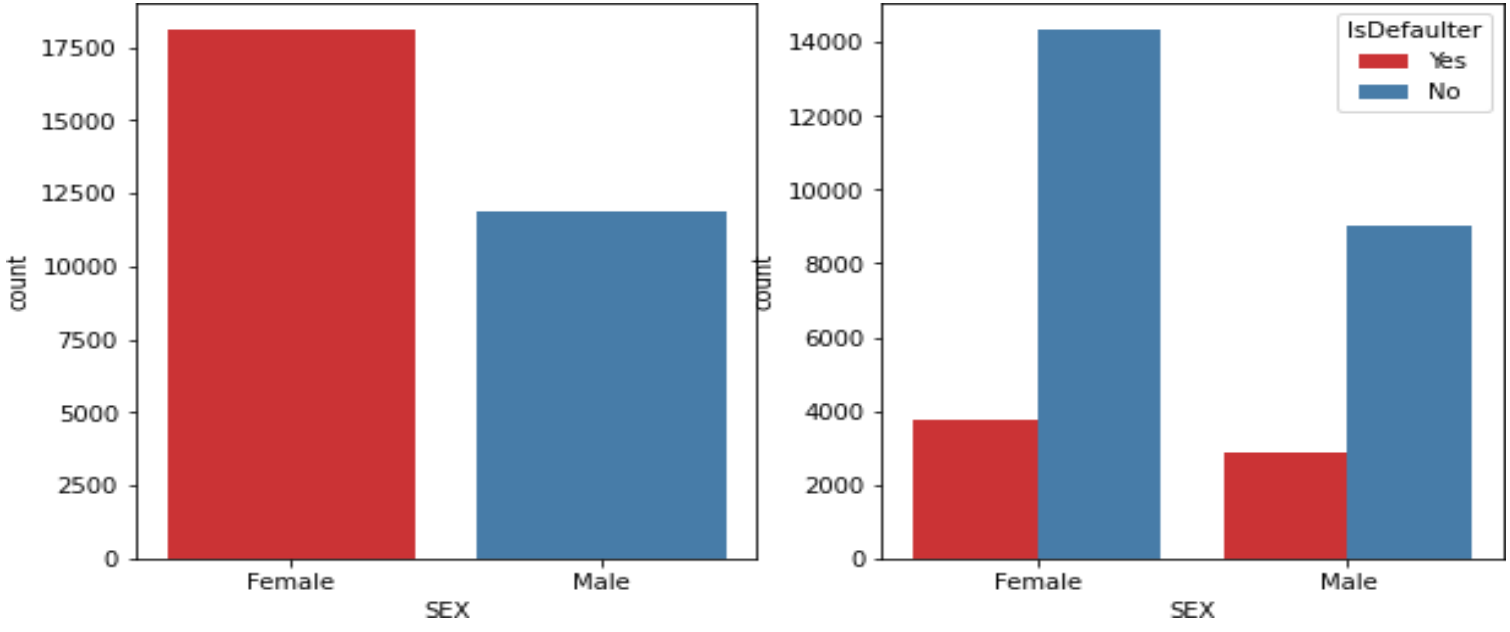
Exploratory Data Analysis

- Defaulters are less as compare to the Non-Defaulters in the given dataset

- Both the classes are not in proportion which means we have an imbalanced dataset.



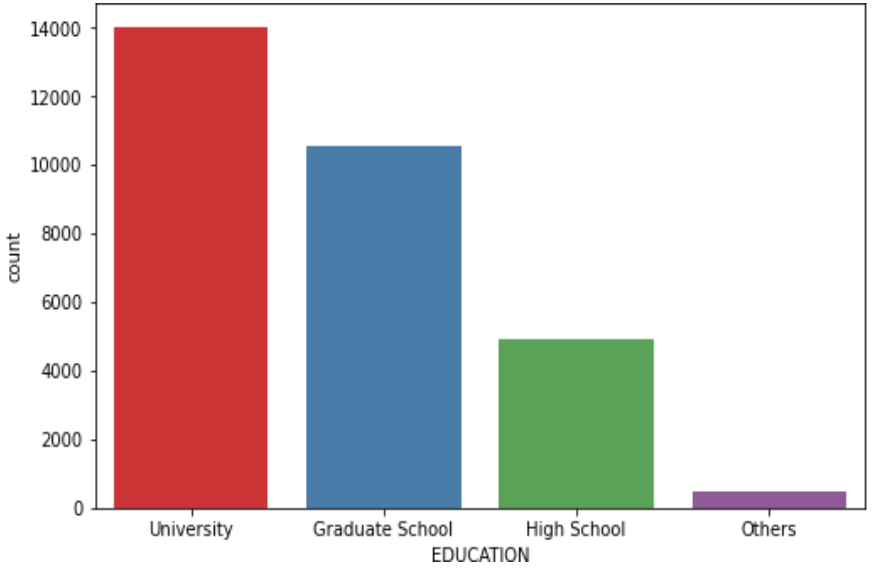
EDA (Continued)

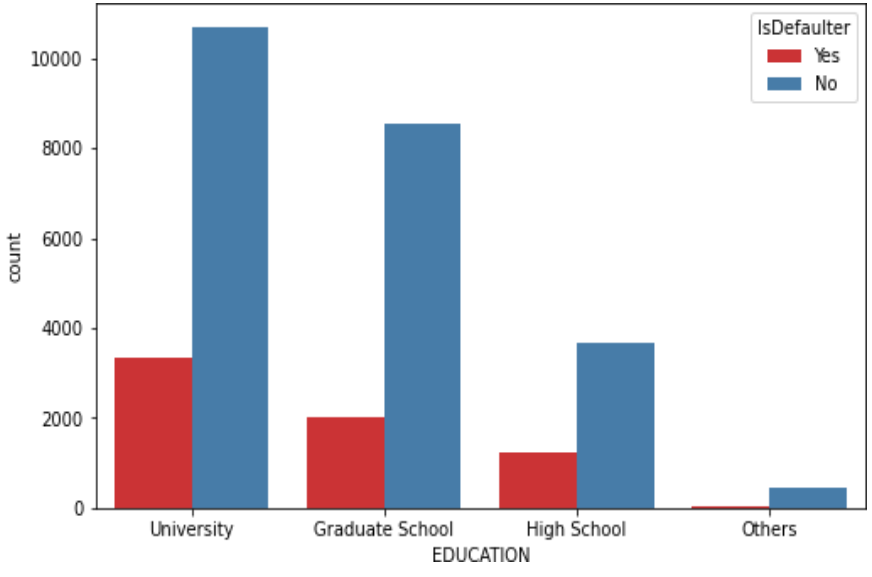


- Female credit card holders are larger than male credit cards holders.

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

EDA (Continued)

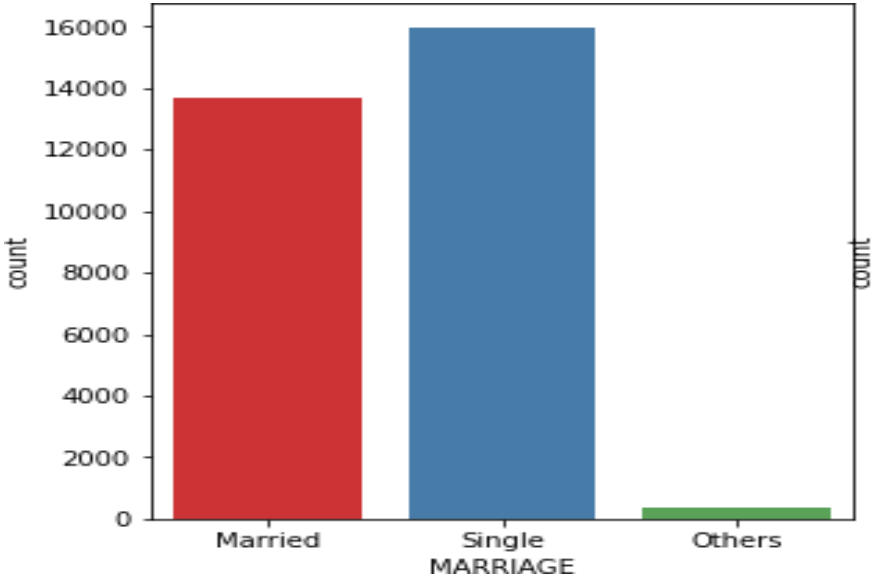


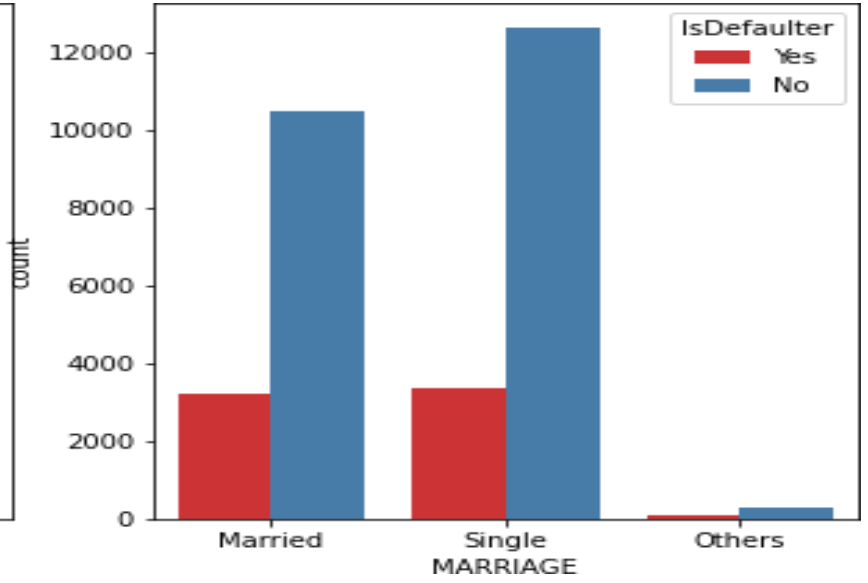


• University and graduate school has maximum credit card holder.

•As the number of university and graduate school credit card holder is higher their credit card default are also higher.

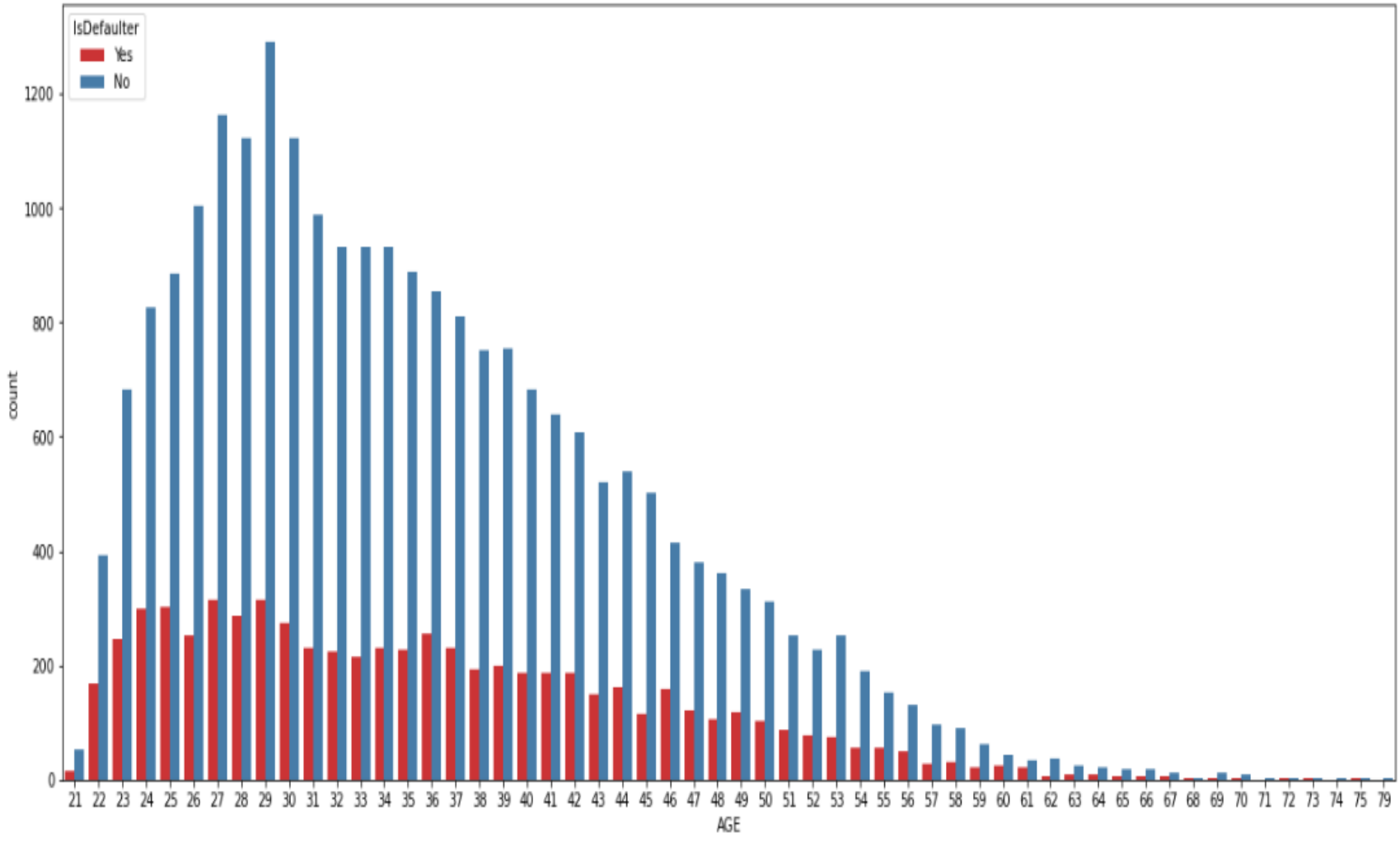
EDA (Continued)





- Number of credit card holder is maximum in singles.

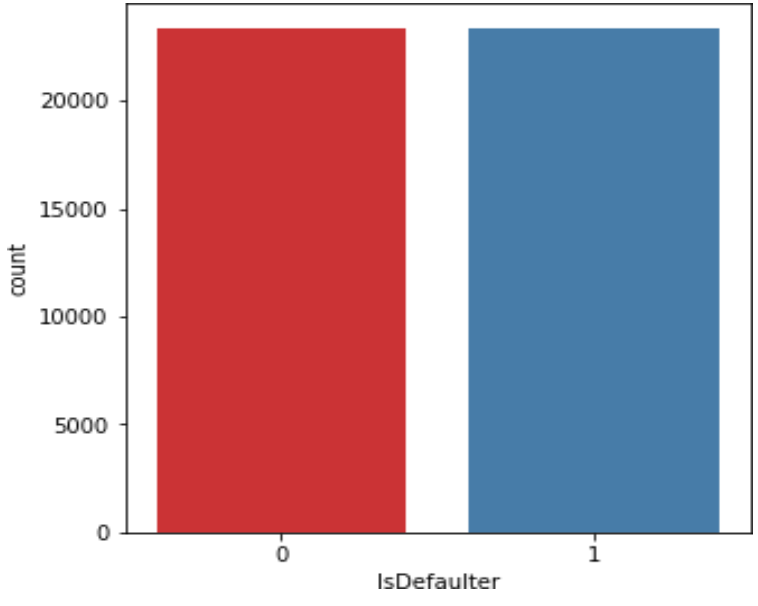
- But credit card defaults are almost same in case of single and married people.



Most of 29th age people used huge credit card and second place was 27th age people

Handling Class Imbalance

- Both the classes are not in proportion.
- SMOTE (Synthetic Minority Oversampling Technique) 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.



- After applying SMOTE Data class is balanced now.

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.

Splitting Data

- Data splits into training dataset and testing dataset.
- Training dataset is used to fit the machine learning model.
- Test dataset is used to evaluate the fit machine learning model.
- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.

Fitting Different Model

Following classifiers are used for predicting credit card default:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine
- Gradient Boosting
- XG Boosting

AI

Cross Validation & Hyperparameter Tunning

•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.

Performance metrics of an algorithm are accuracy, precision, recall, and F1 score.

- **Precision:** is a good metric to use when the costs of false positive(FP) is high.

  Precision = TP / (TP + FP)

- **Recall** is a good metric to use when the cost associated with false negative(FN) is high.

  Recall = TP / (TP + FN)

- **F1score** is a weighted average of precision and recall. Thus, it considers FP and FN. This metric is ve ry useful when we have uneven class distribution, as it seeks a balance between precision and recall.

  F1-score = 2 (precision recall) / (precision + recall)

Different Models

1. Logistic Regression

• Logistic regression is a machine learning algorithm for classification problem.
• It is a predictive analysis algorithm and based on the concept of probability.
• 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	0.827	0.833	0.798	0.858	0.827	0.834
Tunned Model	0.827	0.832	0.798	0.857	0.826	0.834

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, and can handle both numerical and categorical data.

Decision Tree Classifier						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	1	0.794	0.813	0.783	0.798	0.794
Tunned Model	0.844	0.830	0.778	0.868	0.821	0.834

3. Random Forest Classifier

•Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

• 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	0.866	0.829	0.896	0.861	0.869
Tunned Model	0.842	0.832	0.799	0.855	0.826	0.833

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.
•The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

Support Vector Machine						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.846	0.841	0.768	0.900	0.829	0.849
Tunned Model	0.846	0.841	0.769	0.900	0.829	0.849

5. Gradient Boosting

• It is a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.
• It is a sequential ensemble learning technique where the performance of the model improves over iterations.

Gradient Boosting						
	Accuracy		Precision	Recall	F1	AUC
	Train	Test				
Baseline Model	0.846	0.844	0.801	0.877	0.837	0.847
Tunned Model	0.977	0.866	0.823	0.900	0.860	0.869

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	0.845	0.802	0.878	0.838	0.848
Tunned Model	0.999	0.871	0.832	0.903	0.866	0.873

Comparison of Model

	Classifier	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC
11	Optimal XG Boosting	0.999	0.870	0.831	0.900	0.864	0.872
2	Random Forest	1.000	0.868	0.830	0.899	0.863	0.870
10	Optimal Gradient Boosting	0.939	0.862	0.819	0.897	0.856	0.865
3	SVM	0.846	0.840	0.766	0.899	0.827	0.848
9	Optimal SVM	0.846	0.840	0.766	0.899	0.827	0.848
5	XG Boosting	0.846	0.844	0.799	0.878	0.836	0.847
4	Gradient Boosting	0.846	0.843	0.800	0.875	0.836	0.845
8	Optimal Random Forest	0.843	0.834	0.795	0.862	0.827	0.836
0	Logistic Regression	0.827	0.830	0.793	0.856	0.823	0.832
6	Optimal Logistic Regression	0.827	0.830	0.794	0.856	0.823	0.832
7	Optimal Decision Tree	0.847	0.827	0.793	0.852	0.821	0.829
1	Decision Tree	1.000	0.803	0.820	0.793	0.806	0.803

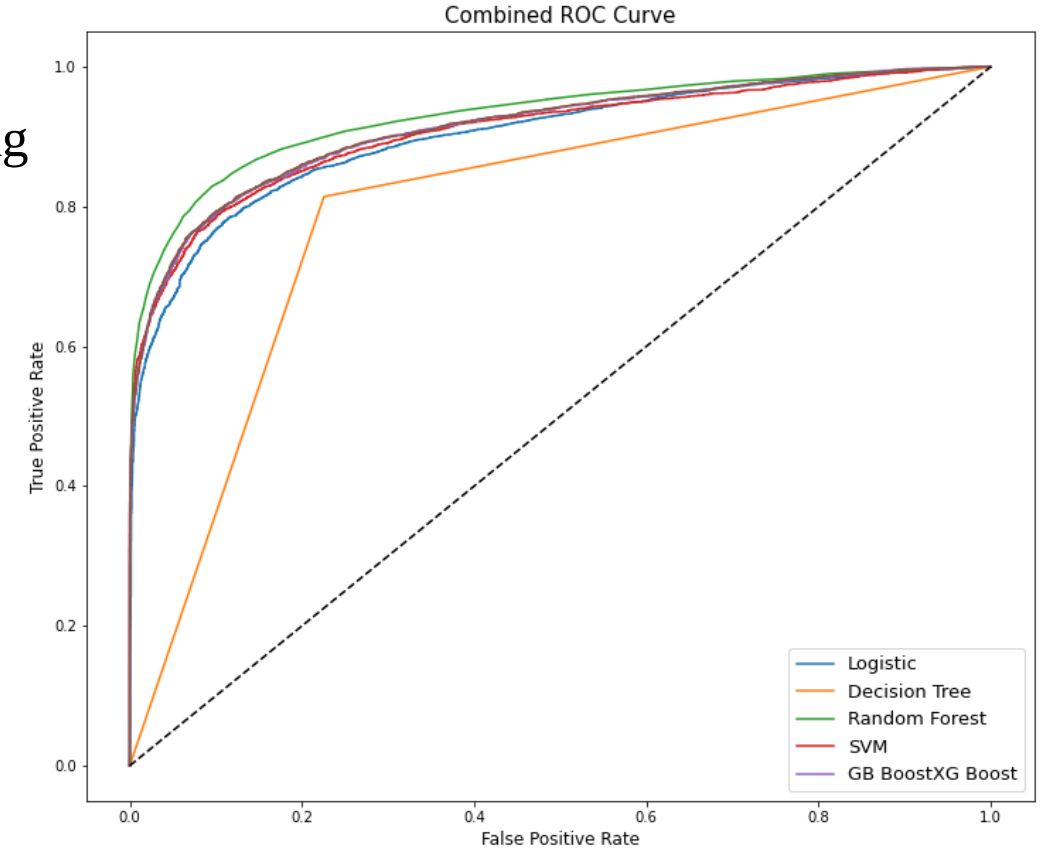
- XG Boost shows highest test accuracy score of 87% and AUC is 0.8732.

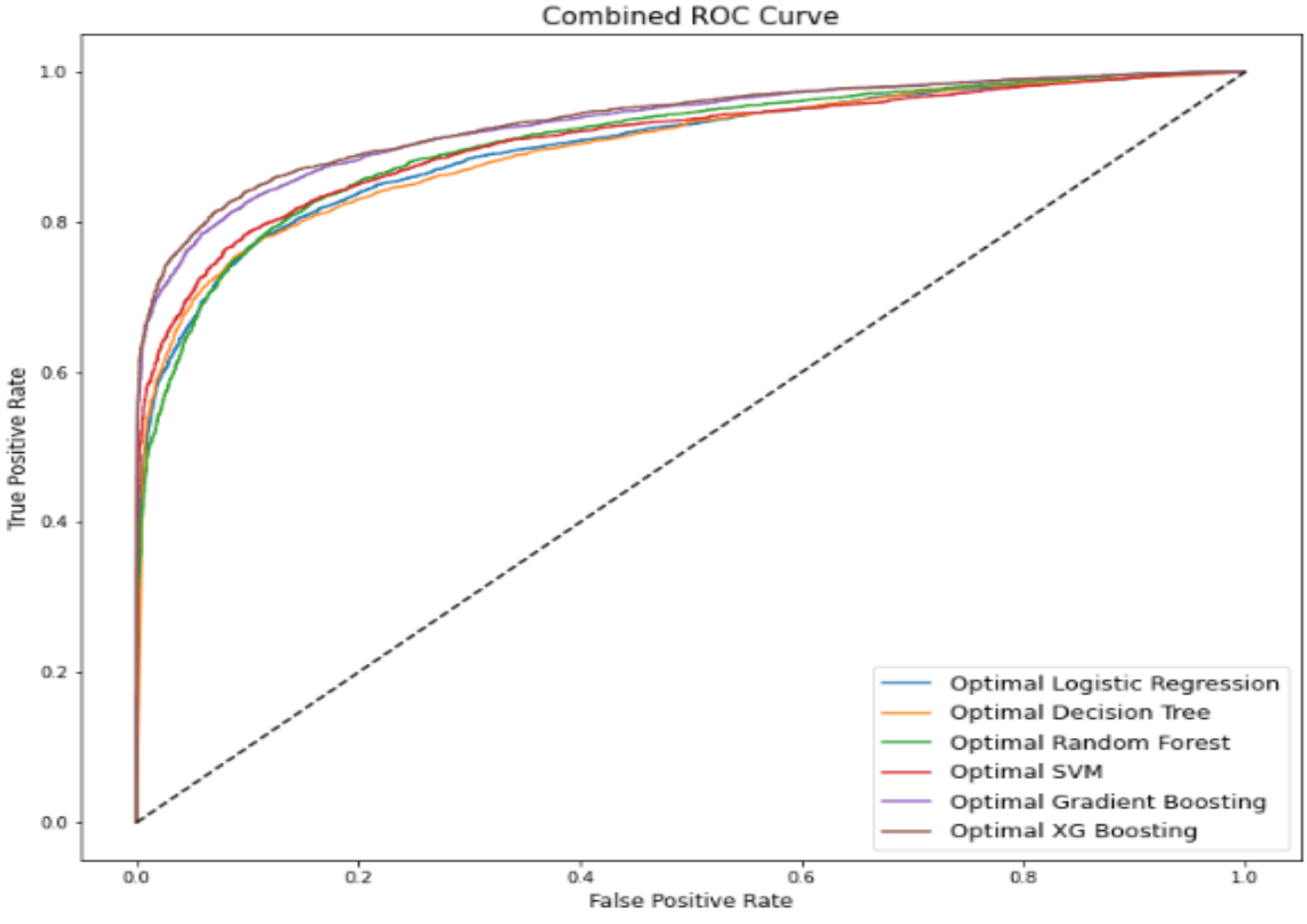
Combined ROC Curve

•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.

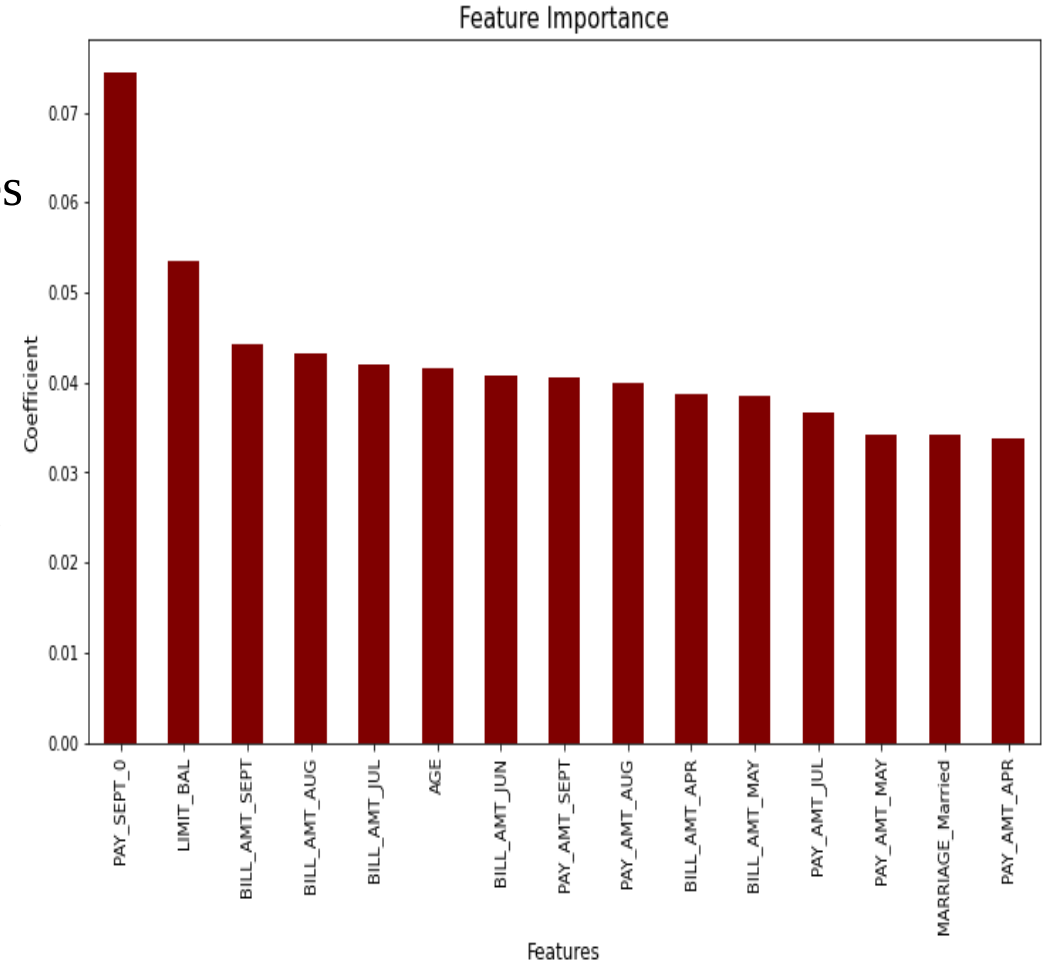



Combined ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Optimal Logistic Regression
Optimal Decision Tree
Optimal Random Forest
Optimal SVM
Optimal Gradient Boosting
Optimal XG Boosting

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 modelling and, in some cases, to improve the performance of the model.



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% and AUC is 0.873.

4. Cross validation and hyperparameter tunning certainly reduces chances of overfitting and also increases performance of model.