



## Predictive Modeling Results

Report Date: April 14, 2020

### **A. Data Input Summary**

1. Total Data Sample: 300
2. Donation Columns:
  - a. Last Gift Amount
  - b. FY 17 Giving
  - c. FY 16 Giving
  - d. FY 15 Giving
  - e. FY 14 Giving
  - f. FY 13 Giving
3. Categorical Columns:
  - a. Gender
  - b. Manager
  - c. Address Type
  - d. Inclination
  - e. Alumni Engagement
4. 80 % of Data used for Training the model: 240
5. 20 % of Data used for Testing the model: 60

### **B. Steps Taken to Run the Predictive Models**

1. Read the input data file provided.
2. Data cleaning: remove null rows and columns and impute missing values.
3. Identify columns containing categorical and textual data and convert it to numerical vectors.
4. Assign target value: Target values are the dependent variable.
5. Splitting the dataset for training and testing to train total of 10 different classifiers.
6. Calculate Feature Importance for each classifier. Feature importance gives a score for each

feature of your data.

7. Plot Confusion Matrix and Classification report. A confusion matrix is a table that is used to describe the performance of a model.
8. Identify and select top 5 classifiers using the F1-Score. The F1-score is a measure of a test's (model's) accuracy.
9. Receiver Operating Characteristic (ROC) Curve. ROC is a probability curve. It tells how much model is capable of distinguishing between classes.
10. Identify optimal threshold (accuracy of the model) and predict.

### **C. Model Summary**

Following terms are used while executing the models.

1. F1-score: It is a harmonic mean of precision and recall.
2. Precision: It is a fraction of correctly classified instances among all predicted instances.
3. Recall: It is fraction of correctly classified instances among all actual/true instances.
4. Support: Number of samples used for the experiment.
5. Confusion Matrix Plot: It is a plot of the true count (x-axis) versus predicted count (y-axis) for both the classes. Top left box represents count of true negatives, top right box represents count of false negatives, bottom left box represents count of false positive and bottom right box represents count of true positives.
6. Feature Importance Plot: Y-axis: variable present in input file and X-axis: relative % of feature importance.

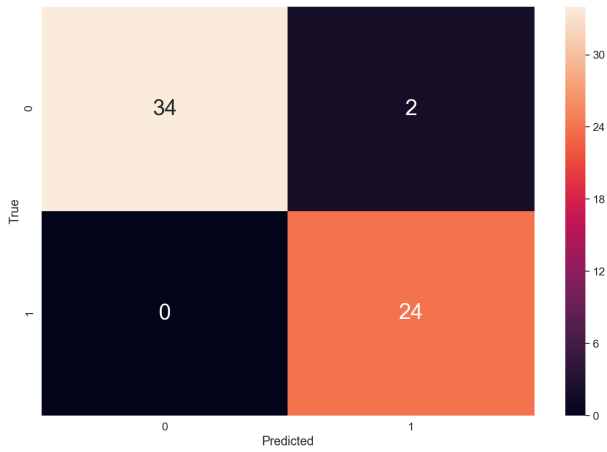
### **D. Top 5 models used to predict**

Top 5 classifiers are selected out of 10 classifiers based on F1-score and used for prediction. We identified optimal threshold to separate donor and non-donor classes. Following are f1-score, threshold and count of donor samples

#### **Model 1. DecisionTreeClassifier**

- a. F1-score: 0.97
- b. Threshold used: 0.8
- c. Donor predicted: 35.0% (105 out of 300)

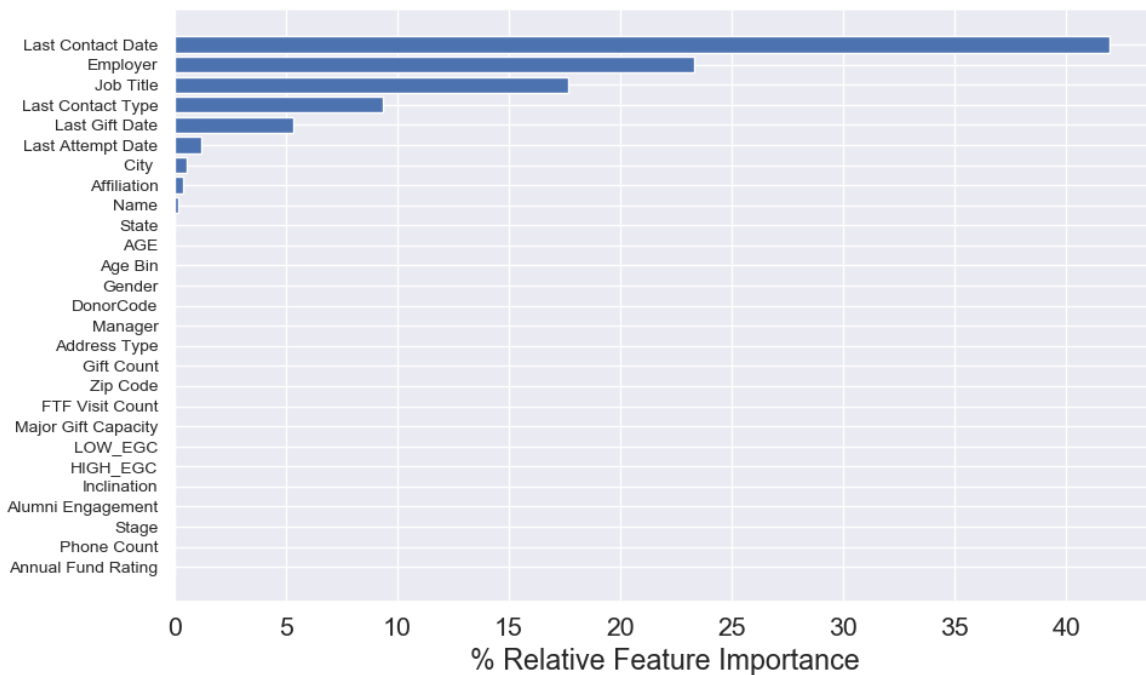
### **# Confusion Matrix Plot**



## # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.97	1.0	0.94	36.0
Donor class	0.96	0.92	1.0	24.0
Macro avg	0.97	0.96	0.97	60.0
Weighted avg	0.97	0.97	0.97	60.0

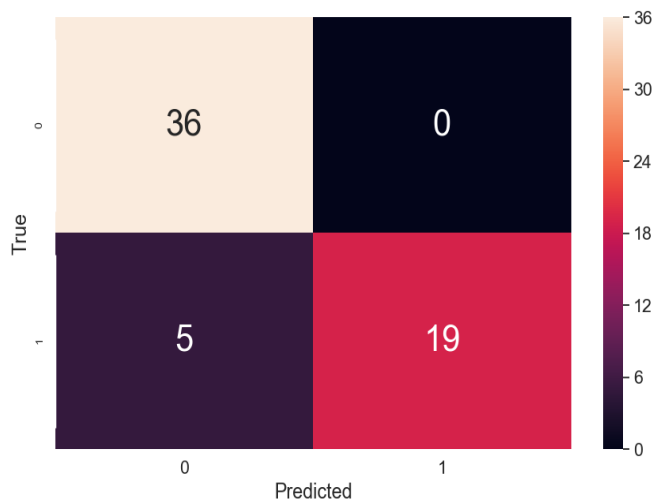
## # Feature Importance Plot



## Model 2. SGDClassifier

- F1-score: 0.91
- Threshold used: 0.7
- Donor predicted: 34.0% (102 out of 300)

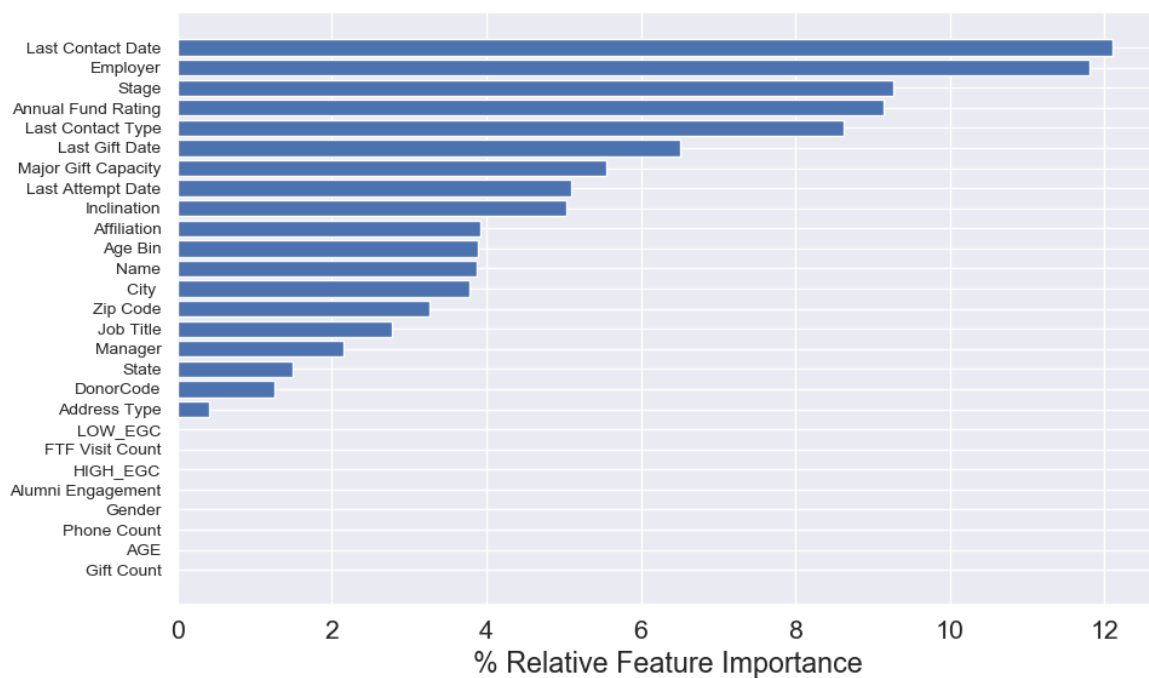
## # Confusion Matrix Plot



## # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.94	0.88	1.0	36.0
Donor class	0.88	1.0	0.79	24.0
Macro avg	0.91	0.94	0.9	60.0
Weighted avg	0.91	0.93	0.92	60.0

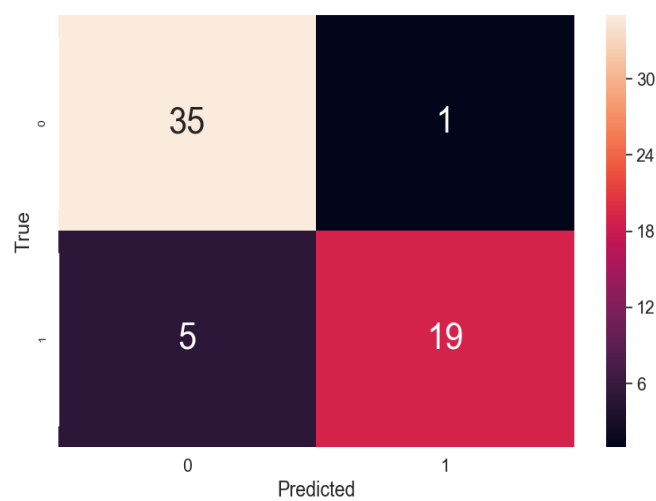
## # Feature Importance Plot



## Model 3. ComplementNB

- F1-score: 0.9
- Threshold used: 0.55
- Donor predicted: 37.0% (111 out of 300)

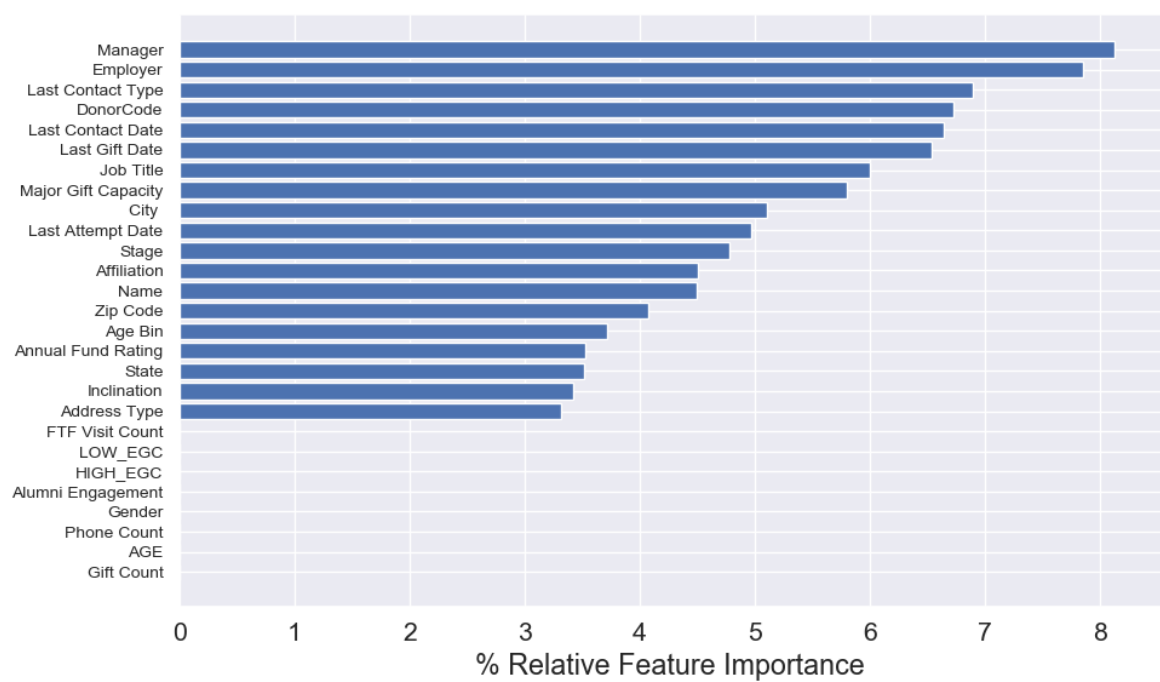
# Confusion Matrix Plot



# Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.92	0.88	0.97	36.0
Donor class	0.86	0.95	0.79	24.0
Macro avg	0.89	0.91	0.88	60.0
Weighted avg	0.9	0.9	0.9	60.0

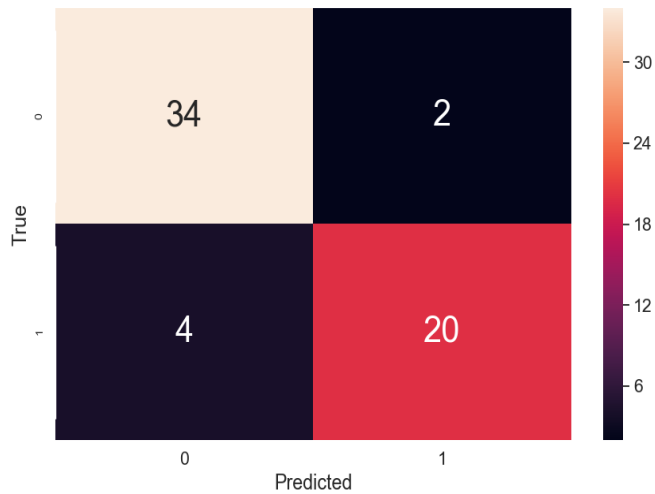
# Feature Importance Plot



Model 4. BernoulliNB

- a. F1-score: 0.9
- b. Threshold used: 0.5
- c. Donor predicted: 34.0% (102 out of 300)

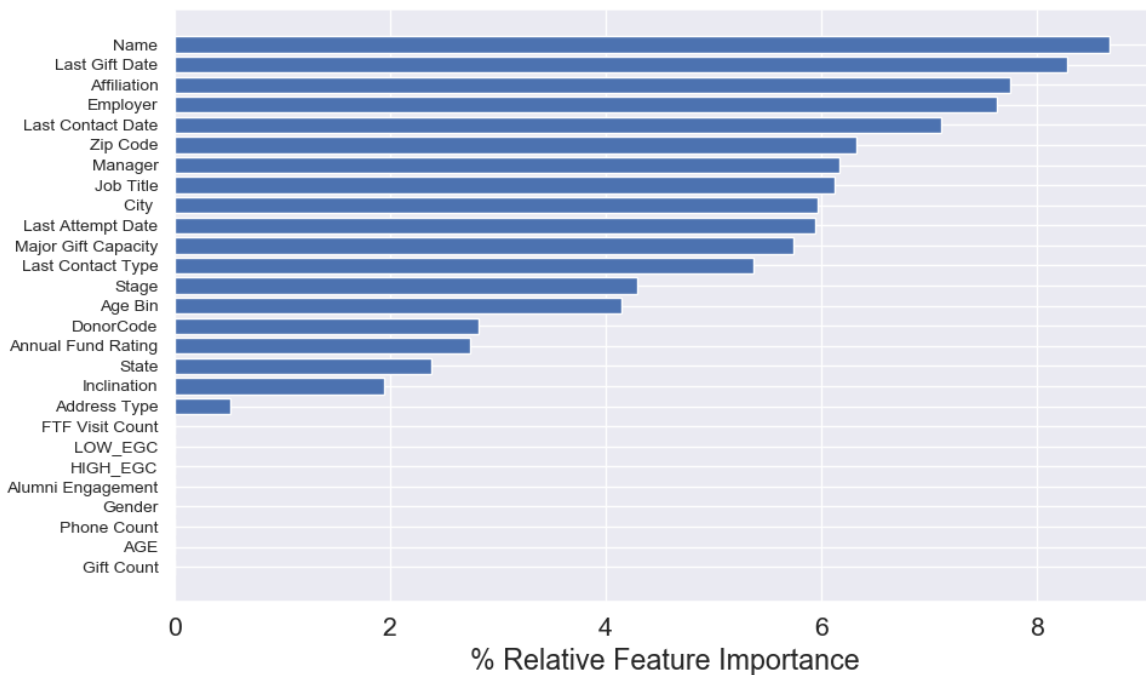
## # Confusion Matrix Plot



## # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.92	0.89	0.94	36.0
Donor class	0.87	0.91	0.83	24.0
Macro avg	0.89	0.9	0.89	60.0
Weighted avg	0.9	0.9	0.9	60.0

## # Feature Importance Plot



## Model 5. PassiveAggressiveClassifier

- F1-score: 0.88
- Threshold used: 0.6
- Donor predicted: 34.0% (102 out of 300)

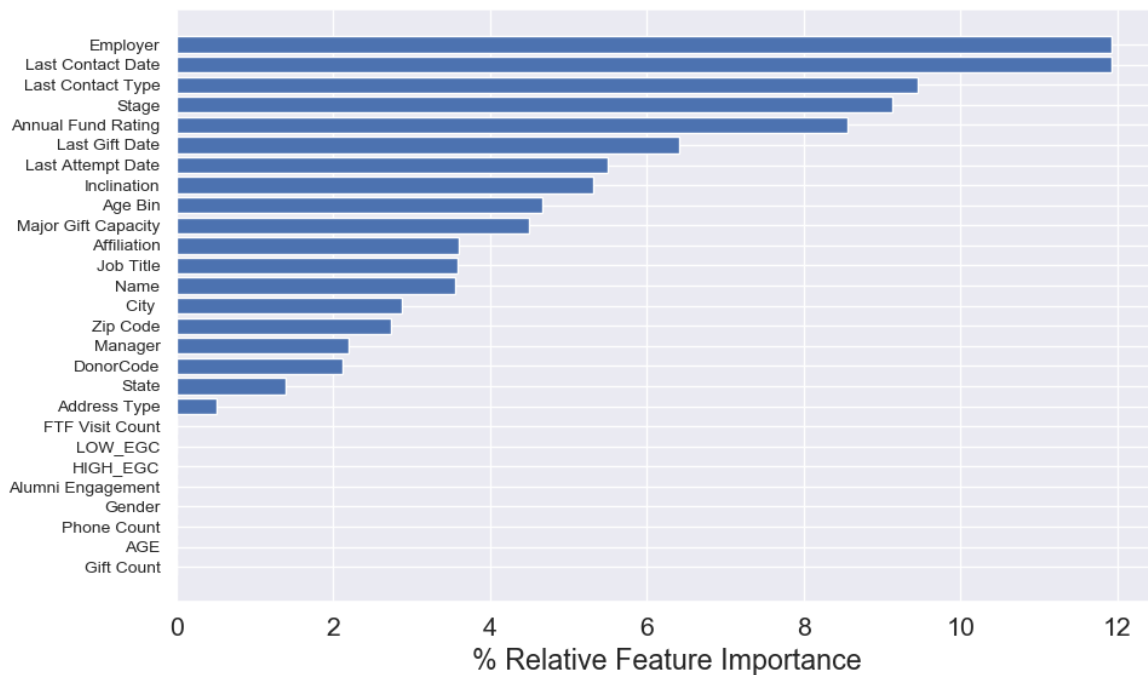
## # Confusion Matrix Plot



## # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.91	0.85	0.97	36.0
Donor class	0.84	0.95	0.75	24.0
Macro avg	0.87	0.9	0.86	60.0
Weighted avg	0.88	0.89	0.88	60.0

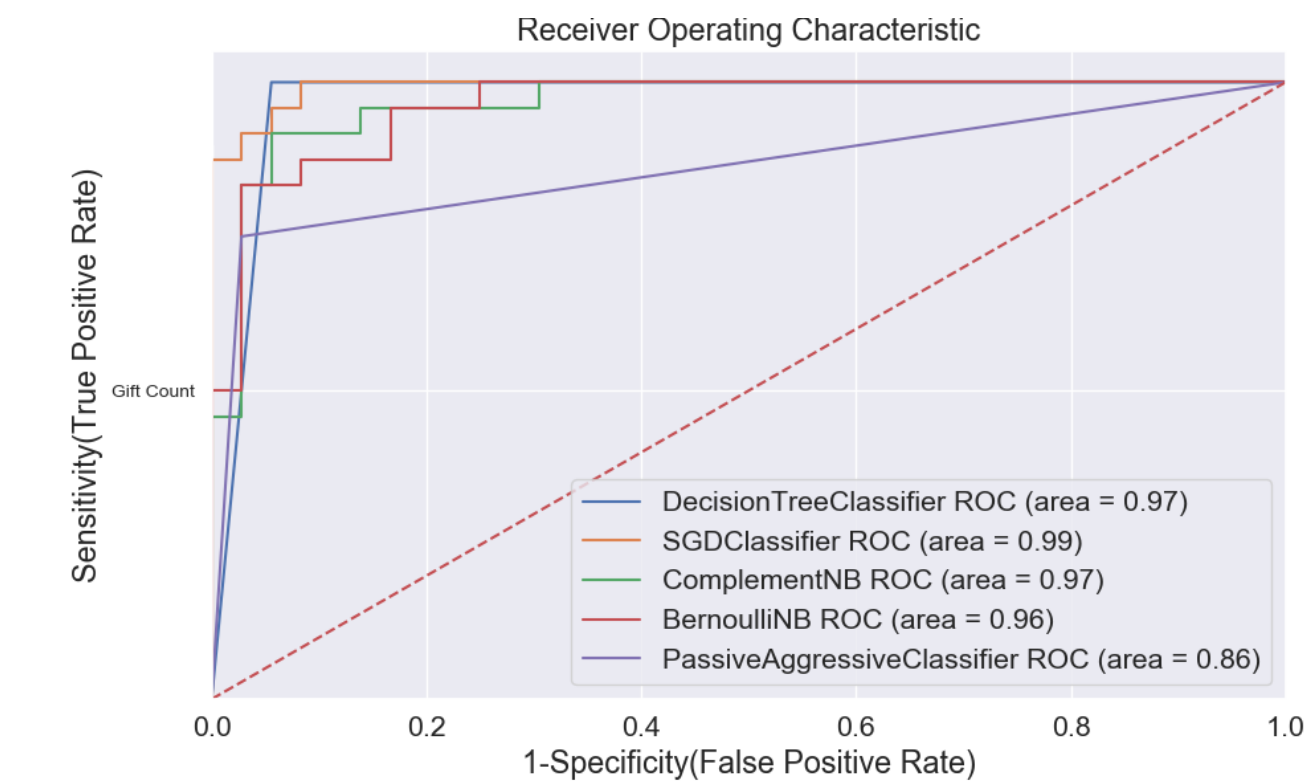
## # Feature Importance Plot



## Receiver Operating Characteristic (ROC) Curve

It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis). True positive rate or sensitivity describes how good the model is at predicting the positive class when the actual outcome is positive. False positive rate describes how often a positive class is predicted when the actual outcome is negative. A model with high accuracy is represented by a line that travels from the

bottom left of the plot to the top left and then across the top to the top right and has Area Under Curve (AUC) as 1. A model with less accuracy is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. We can compare multiple models using AUC value, Best model will have AUC close to 1.



**Correlation Plot**

Correlation explains how one or more variables are related to each other.

