PREDICT ME

Predictive Modeling Results

Data Input Summary

1. Total training samples: 300

2. Donation columns: ['Last Gift Amount', 'FY 17 Giving', 'FY 16 Giving', 'FY 15 Giving', 'FY 14 Giving', 'FY 13 Giving']

3. Positive sample count: 103 and Negative sample count: 197

4. Categorical columns: ['Gender', 'Manager', 'Address Type', 'Inclination', 'Alumni Engagement']

5. Training data (80%) used: 240

6. Test data (20%) used: 60

Steps taken to run the model

1. Read input data file.

2. Data cleaning: Remove null rows and columns.

3. Identify columns containing categorical and textual data.

4. Assign target value to each sample.

5. Train 9 different classifiers by spliting dataset for train and test.

6. Calculate feature importance for each classifier.

7. Plot confusion matrix and classification report.

8. Select top 5 classifier using f1-score.

9. Receiver Operating Characteristic (ROC) Curve.

10. Identify optimal threshold and predict.

Model Summary

Following terms used below are defined as follows

Precision: It is fraction of correctly classified instances among all predicted instances.

Recall: It is fraction of correctly classified instanes among all actual/true instances.

F1-score: It is a harmonic mean of precision and recall.

Support: Number of samples used for the experiment.

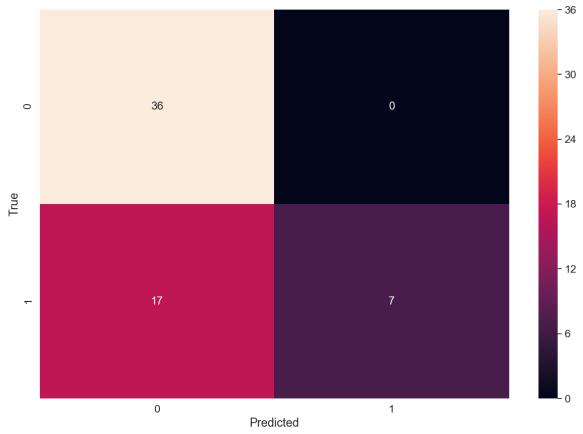
Confusion matrix plot: It is a plot of the true count (x-axis) versus predicted count (y-axis) for both the classes. Top left box represents count of true negatives, top right box represents count of false negatives, bottom left box represents count of false positive and bottom right box represents count of true positives.

Feature importance plot: Y-axis: variable present in input file and X-axis: relative % of feature importance.

1. Classifier: LogisticRegression

f1-score 0.67 and training time(seconds): 0.838

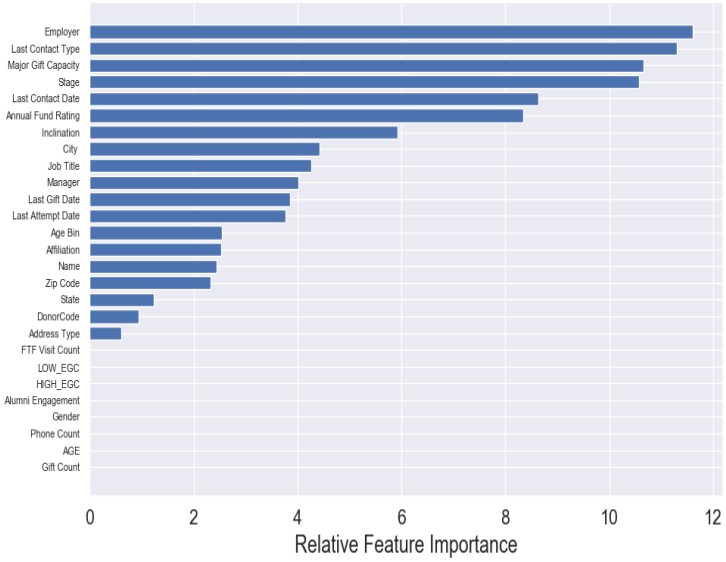
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.81	0.68	1.0	36.0
1	0.45	1.0	0.29	24.0
macro avg	0.63	0.84	0.65	60.0
weighted avg	0.67	0.81	0.72	60.0

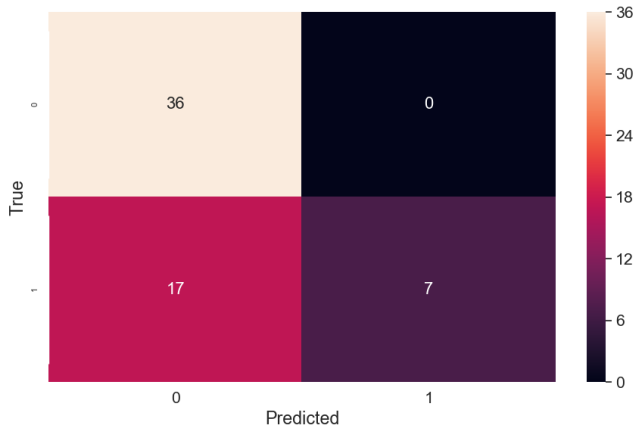
Feature Importance Plot



2. Classifier: MultinomialNB

f1-score 0.67 and training time(seconds): 0.003

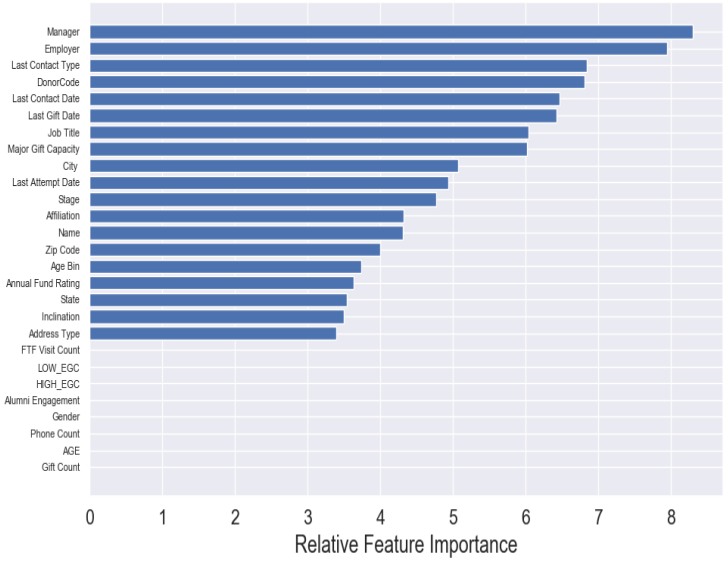
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.81	0.68	1.0	36.0
1	0.45	1.0	0.29	24.0
macro avg	0.63	0.84	0.65	60.0
weighted avg	0.67	0.81	0.72	60.0

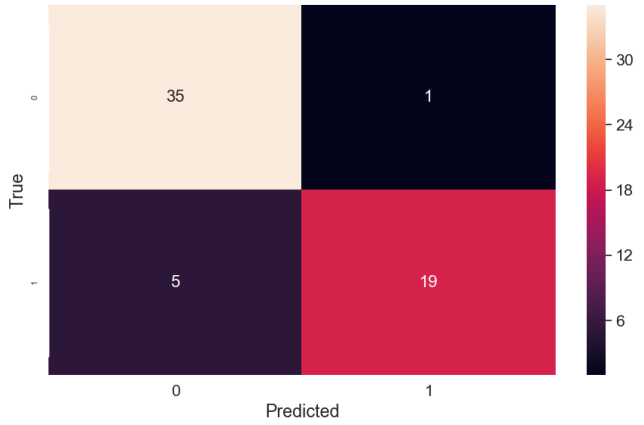
Feature Importance Plot



3. Classifier: ComplementNB

f1-score 0.9 and training time(seconds): 0.003

Confusion Matrix Plot



Classification Report

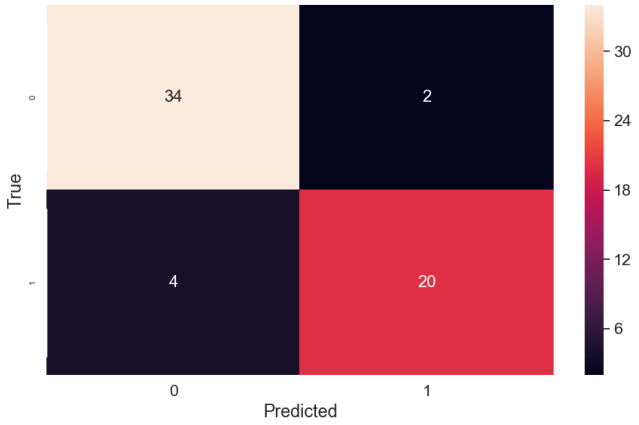
index	f1-score	precision	recall	support
0	0.92	0.88	0.97	36.0
1	0.86	0.95	0.79	24.0
macro avg	0.89	0.91	0.88	60.0
weighted avg	0.9	0.9	0.9	60.0

Feature Importance Plot

4. Classifier: BernoulliNB

f1-score 0.9 and training time(seconds): 0.005

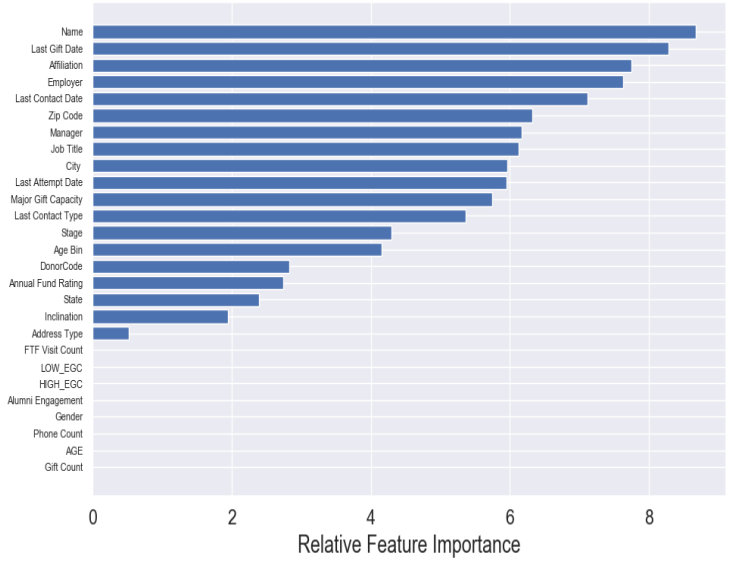
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.92	0.89	0.94	36.0
1	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

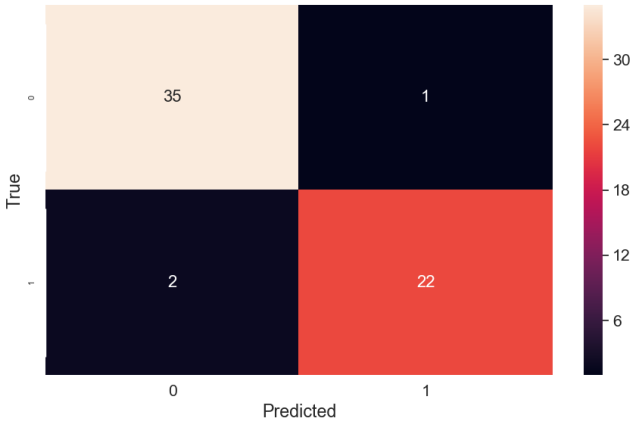
Feature Importance Plot



5. Classifier: DecisionTreeClassifier

f1-score 0.95 and training time(seconds): 0.015

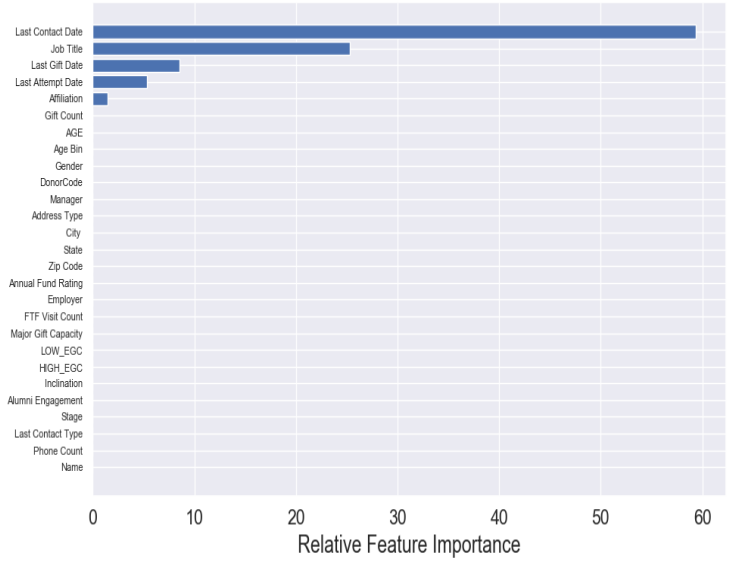
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.96	0.95	0.97	36.0
1	0.94	0.96	0.92	24.0
macro avg	0.95	0.95	0.94	60.0
weighted avg	0.95	0.95	0.95	60.0

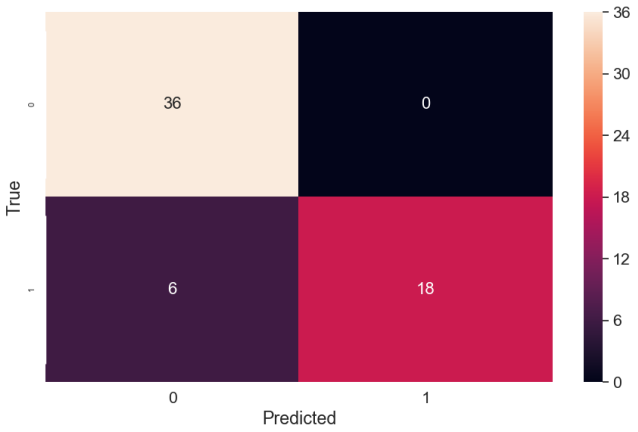
Feature Importance Plot



6. Classifier: SGDClassifier

f1-score 0.9 and training time(seconds): 0.007

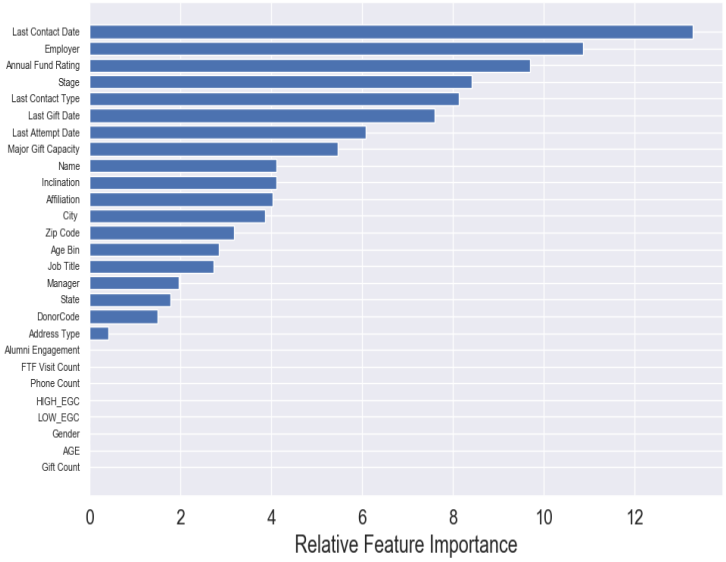
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.92	0.86	1.0	36.0
1	0.86	1.0	0.75	24.0
macro avg	0.89	0.93	0.88	60.0
weighted avg	0.9	0.91	0.9	60.0

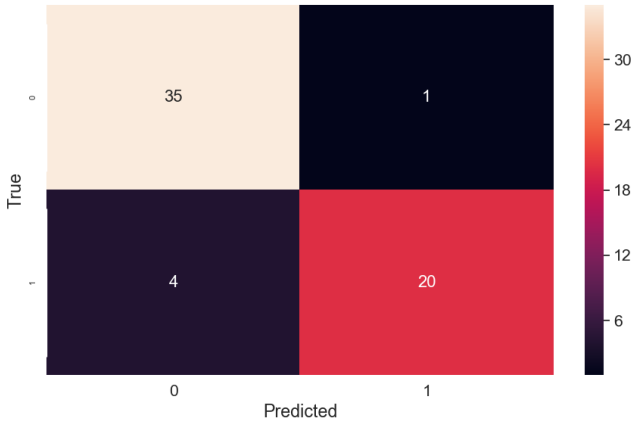
Feature Importance Plot



7. Classifier: PassiveAggressiveClassifier

f1-score 0.92 and training time(seconds): 0.023

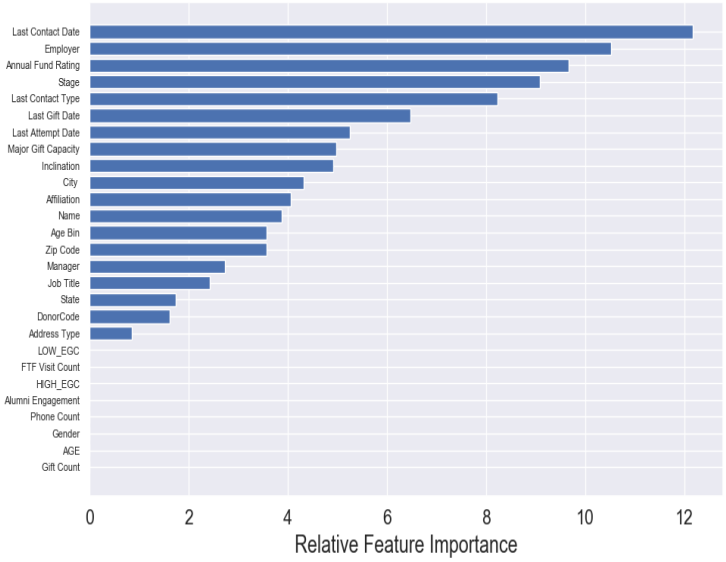
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.93	0.9	0.97	36.0
1	0.89	0.95	0.83	24.0
macro avg	0.91	0.92	0.9	60.0
weighted avg	0.92	0.92	0.92	60.0

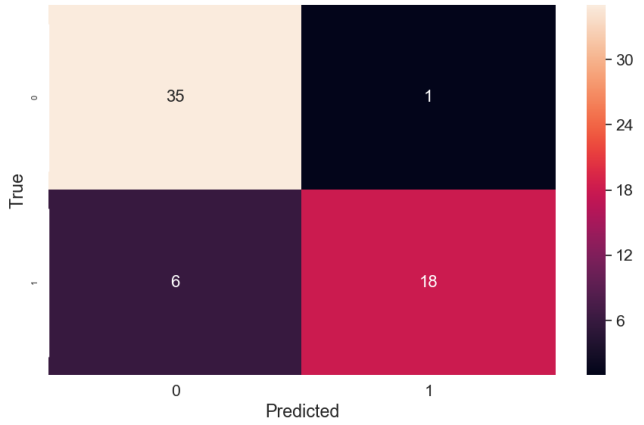
Feature Importance Plot



8. Classifier: LinearSVC

f1-score 0.88 and training time(seconds): 0.014

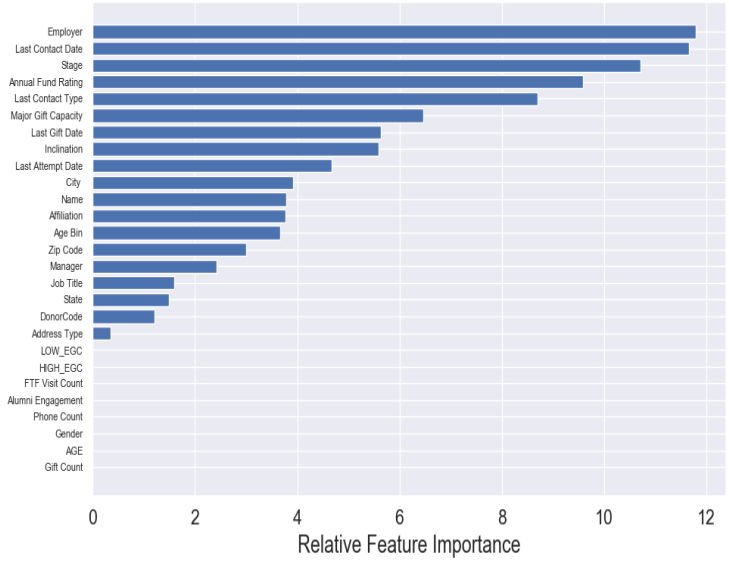
Confusion Matrix Plot



Classification Report

index	f1-score	precision	recall	support
0	0.91	0.85	0.97	36.0
1	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

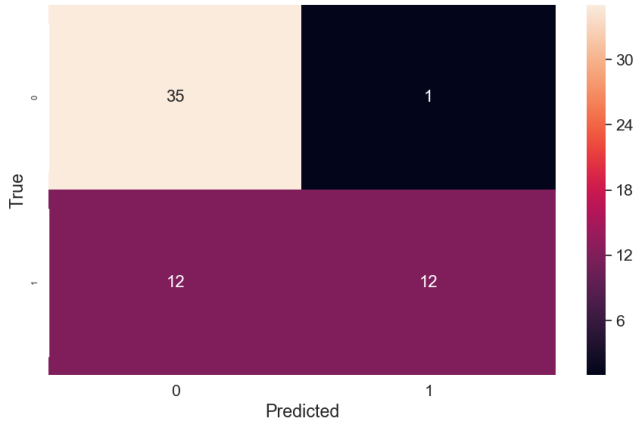
Feature Importance Plot



9. Classifier: RandomForestClassifier

f1-score 0.77 and training time(seconds): 0.019

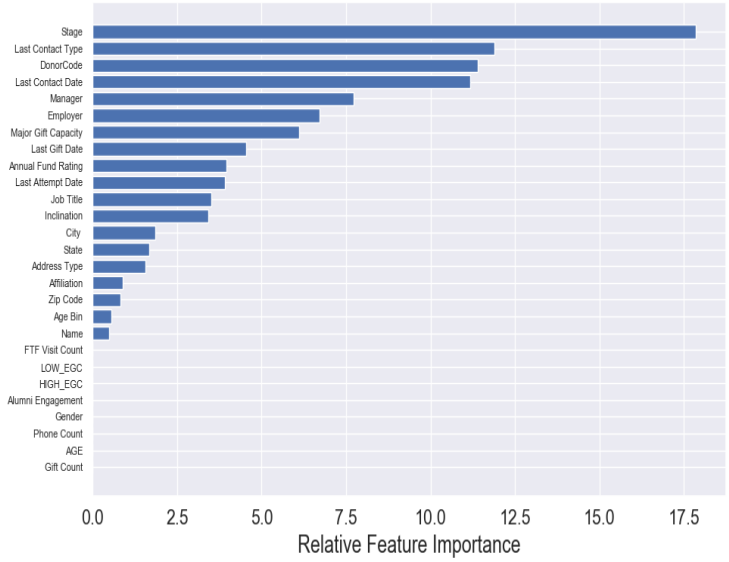
Confusion Matrix Plot



Classification Report

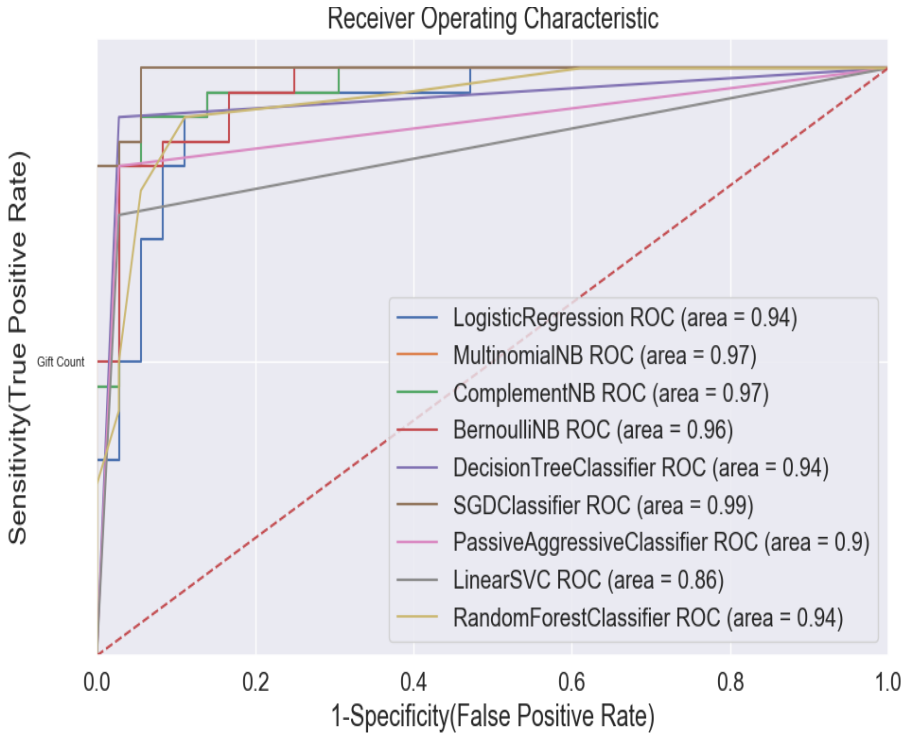
index	f1-score	precision	recall	support
0	0.84	0.74	0.97	36.0
1	0.65	0.92	0.5	24.0
macro avg	0.75	0.83	0.74	60.0
weighted avg	0.77	0.82	0.78	60.0

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 decribes how often a positive class is predicted when the actual outcome is negative. A model with perfect skill 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 no skill 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.



Top 5 models used to predict

Top 5 classifiers are selected out of 9 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 and non-donor samples.

1. DecisionTreeClassifier

F1-score: 0.95

Threshold used: 0.8

Donor predicted: 102

Non-Donor predicted: 198

2. PassiveAggressiveClassifier

F1-score: 0.92

Threshold used: 0.6

Donor predicted: 103

Non-Donor predicted: 197

3. ComplementNB

F1-score: 0.9

Threshold used: 0.55

Donor predicted: 111

Non-Donor predicted: 189

4. BernoulliNB

F1-score: 0.9

Threshold used: 0.5

Donor predicted: 102

Non-Donor predicted: 198

5. SGDClassifier

F1-score: 0.9

Threshold used: 0.8

Donor predicted: 104

Non-Donor predicted: 196