Predictive Modeling Results

Report Date: April 18, 2020

A. Data Input Summary

1. Total Data Sample: 3,800
2. Donation Columns:
    a. funding1
    b. fund2
    c. donations3
    d. gifts4
    e. donat5
    f. transac6
    g. grants7
    h. money8
    i. gift9
    j. gift10
    k. funding11
    l. donations12
    m. donations13
3. Categorical Columns:
    a. Cand_Office
4. 80% of Data used for Training the model: 3,040
5. 20% of Data used for Testing the model: 760

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 (for e.g. Logistic Regression, Naive Bayes and Random Forest etc).

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 instanes 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

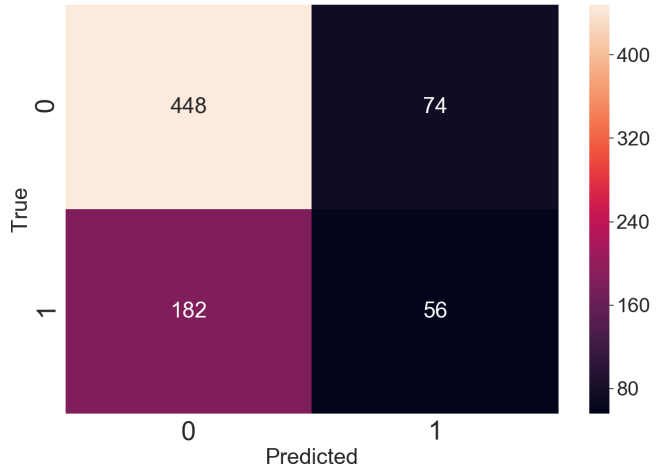
Model 1. BernoulliNB

a. F1-score: 0.63

b. Threshold used: 0.8

c. Donor predicted: 60.92% (2,315 out of 3,800)

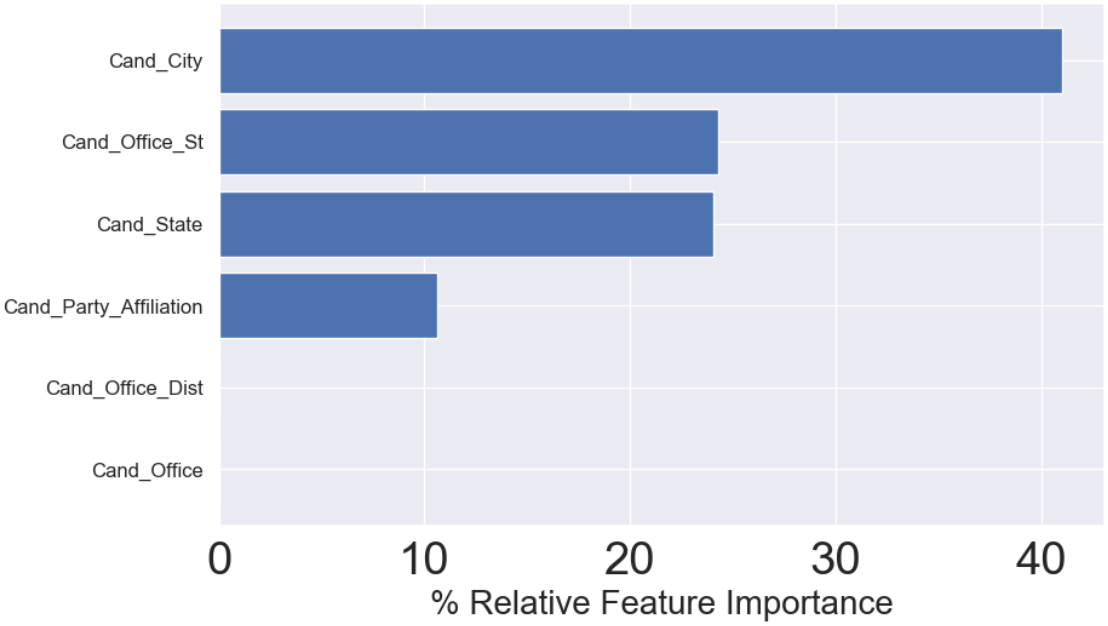
 # Confusion Matrix Plot



 # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.78	0.71	0.86	522
Donor class	0.3	0.43	0.24	238
Macro avg	0.54	0.57	0.55	760
Weighted avg	0.63	0.62	0.66	760

 # Feature Importance Plot



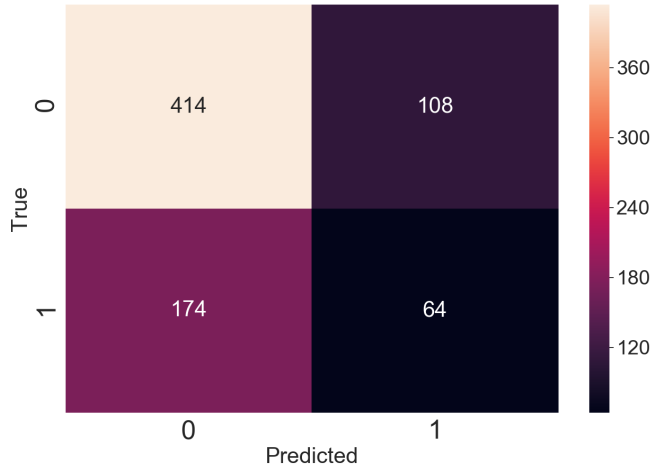
<u>Model 2. DecisionTreeClassifier</u>

a. F1-score: 0.61

b. Threshold used: 0.65

c. Donor predicted: 35.03% (1,331 out of 3,800)

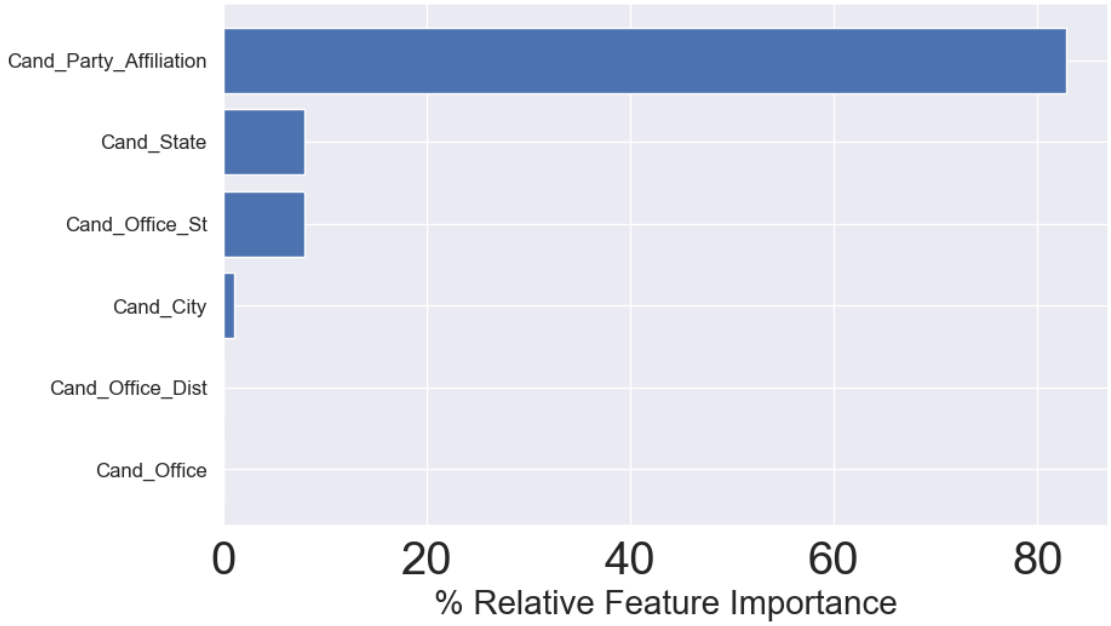
 # Confusion Matrix Plot



 # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.75	0.7	0.79	522
Donor class	0.31	0.37	0.27	238
Macro avg	0.53	0.54	0.53	760
Weighted avg	0.61	0.6	0.63	760

 # Feature Importance Plot



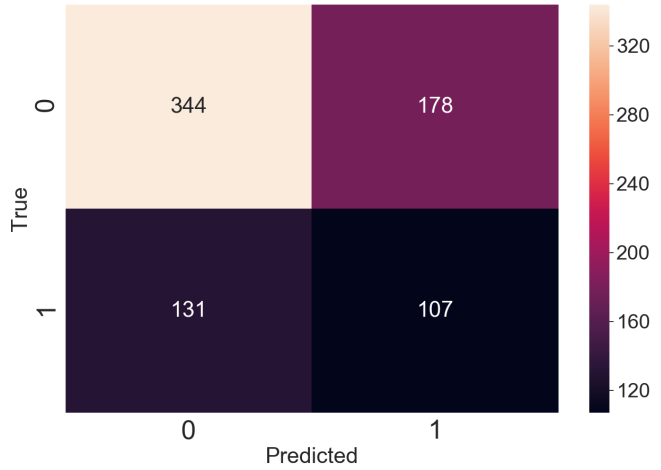
Model 3. ComplementNB

a. F1-score: 0.6

b. Threshold used: 0.55

c. Donor predicted: 53.0% (2,014 out of 3,800)

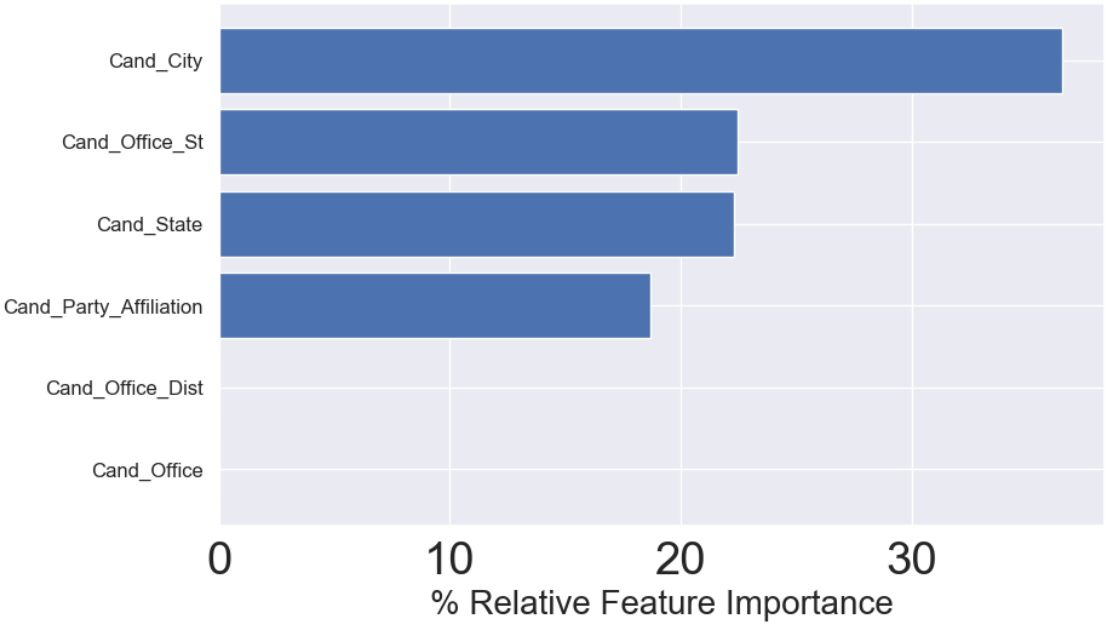
 # Confusion Matrix Plot



 # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.69	0.72	0.66	522
Donor class	0.41	0.38	0.45	238
Macro avg	0.55	0.55	0.55	760
Weighted avg	0.6	0.61	0.59	760

 # Feature Importance Plot



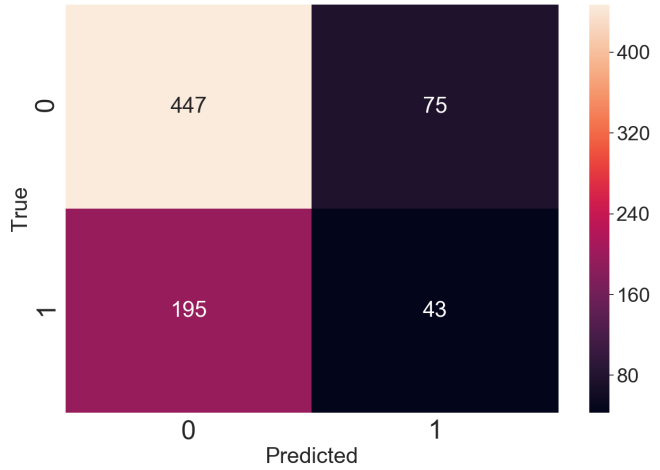
Model 4. RandomForestClassifier

a. F1-score: 0.6

b. Threshold used: 0.65

c. Donor predicted: 35.16% (1,336 out of 3,800)

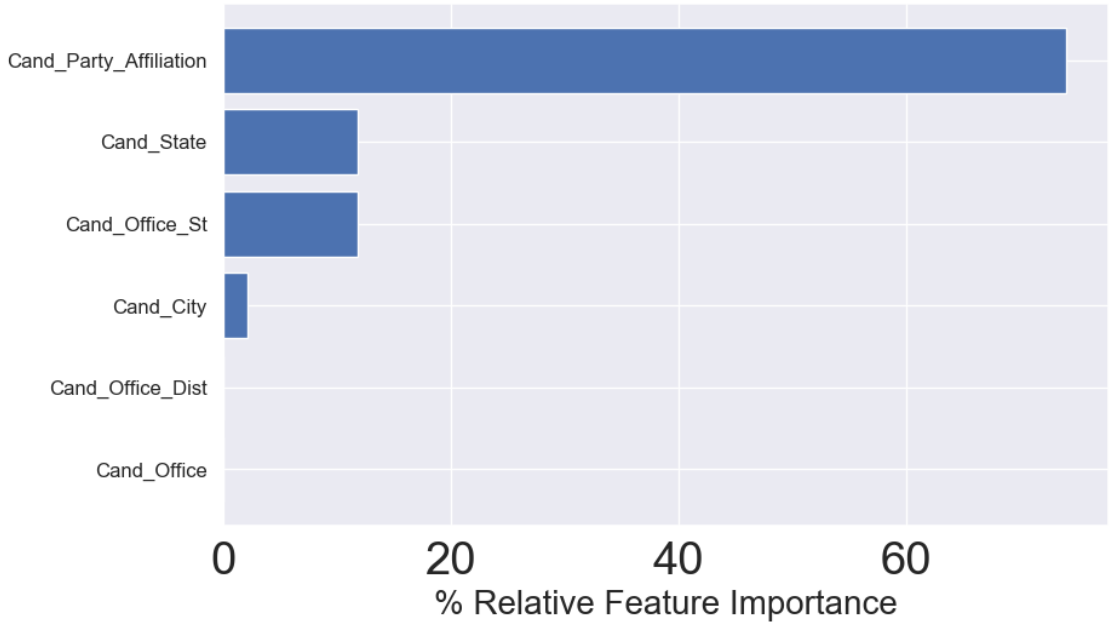
 # Confusion Matrix Plot



 # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.77	0.7	0.86	522
Donor class	0.24	0.36	0.18	238
Macro avg	0.5	0.53	0.52	760
Weighted avg	0.6	0.59	0.64	760

 # Feature Importance Plot



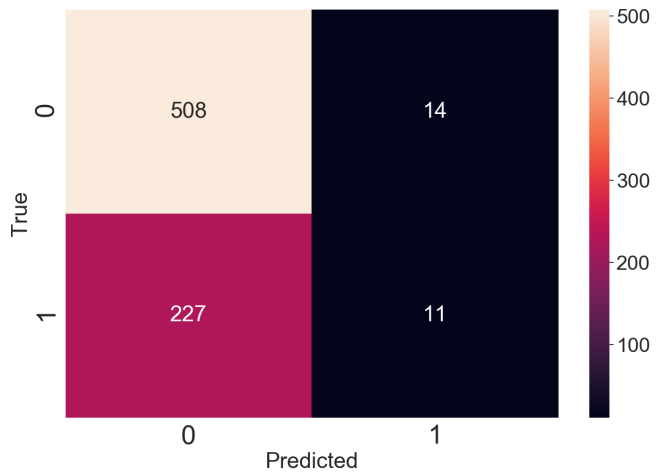
<u>Model 5. SGDClassifier</u>

a. F1-score: 0.58

b. Threshold used: 0.75

c. Donor predicted: 45.66% (1,735 out of 3,800)

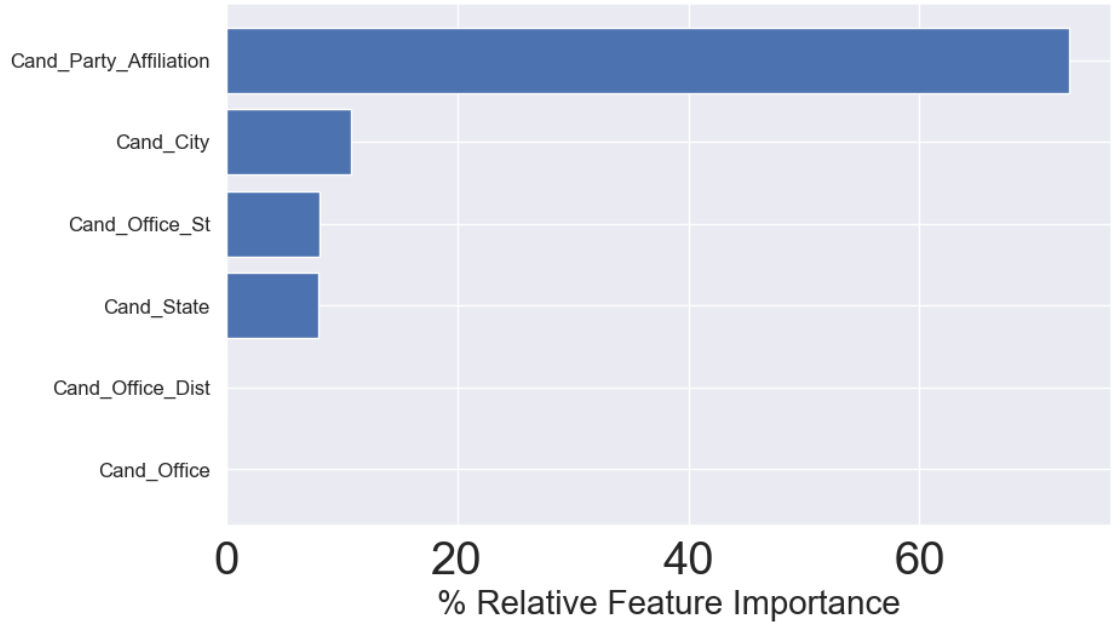
 # Confusion Matrix Plot



 # Classification Report Table

Index	F1-score	Precision	Recall	Support
Non-donor class	0.81	0.69	0.97	522
Donor class	0.08	0.44	0.05	238
Macro avg	0.45	0.57	0.51	760
Weighted avg	0.58	0.61	0.68	760

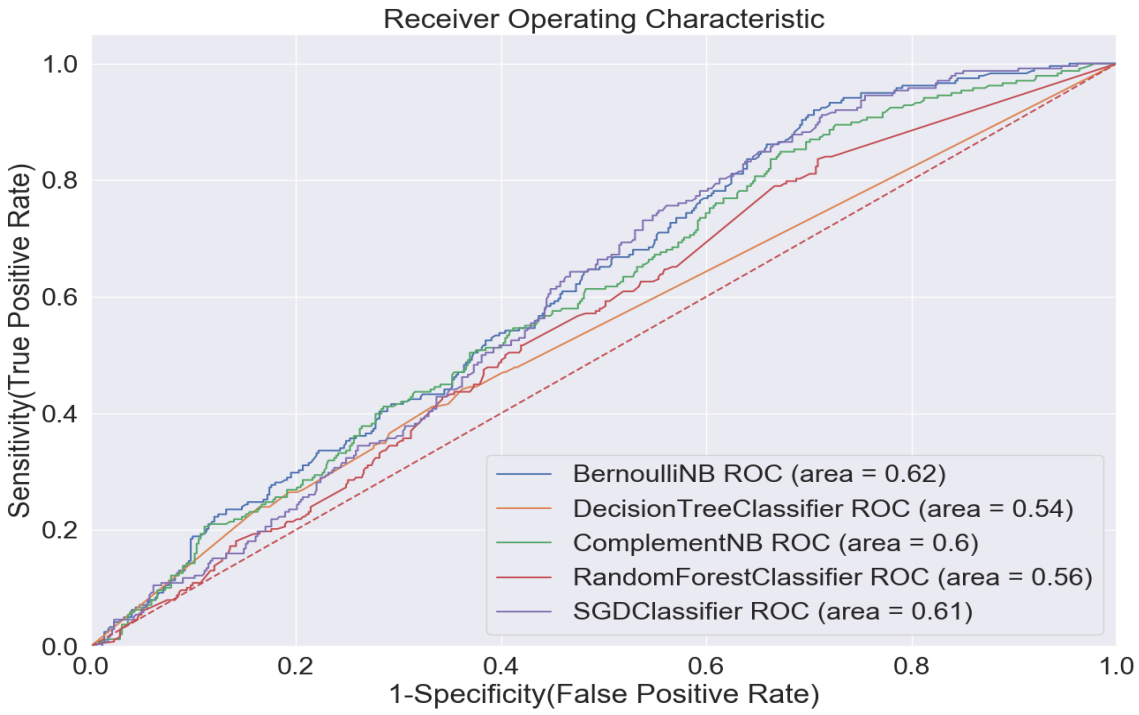
 # Feature Importance Plot



<u>Receiver Operating Characteristic (ROC) Curve</u>

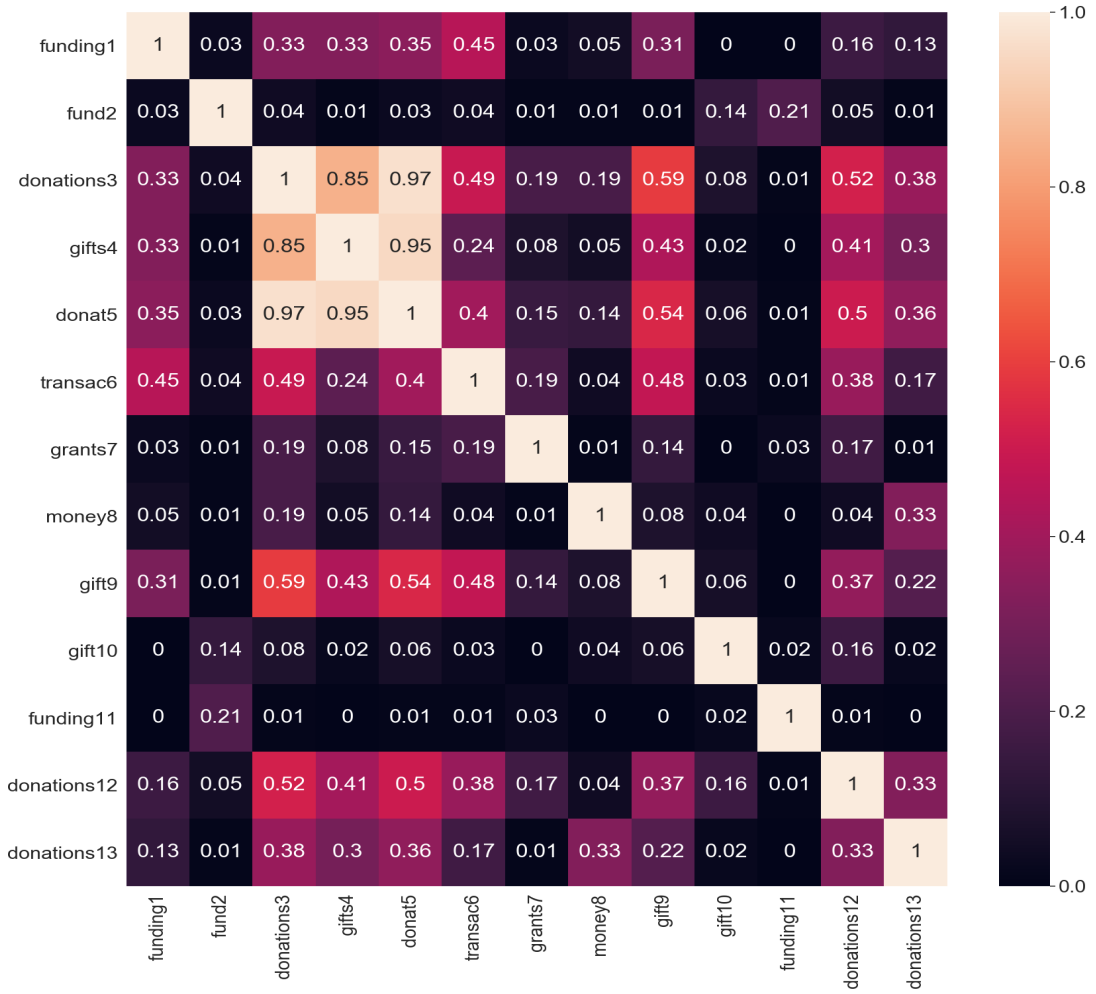
It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis). True positive rate or sensitivity describes how good the model is at predicting the positive class when the actual outcome is positive. False positive rate decribes how often a positive class is predicted when the actual outcome is negative.

A model with 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.


funding1 1 0.03 0.33 0.33 0.35 0.45 0.03 0.05 0.31 0 0 0.16 0.13
fund2 0.03 1 0.04 0.01 0.03 0.04 0.01 0.01 0.01 0.14 0.21 0.05 0.01
donations3 0.33 0.04 1 0.85 0.97 0.49 0.19 0.19 0.59 0.08 0.01 0.52 0.38
gifts4 0.33 0.01 0.85 1 0.95 0.24 0.08 0.05 0.43 0.02 0 0.41 0.3
donat5 0.35 0.03 0.97 0.95 1 0.4 0.15 0.14 0.54 0.06 0.01 0.5 0.36
transac6 0.45 0.04 0.49 0.24 0.4 1 0.19 0.04 0.48 0.03 0.01 0.38 0.17
grants7 0.03 0.01 0.19 0.08 0.15 0.19 1 0.01 0.14 0 0.03 0.17 0.01
money8 0.05 0.01 0.19 0.05 0.14 0.04 0.01 1 0.08 0.04 0 0.04 0.33
gift9 0.31 0.01 0.59 0.43 0.54 0.48 0.14 0.08 1 0.06 0 0.37 0.22
gift10 0 0.14 0.08 0.02 0.06 0.03 0 0.04 0.06 1 0.02 0.16 0.02
funding11 0 0.21 0.01 0 0.01 0.01 0.03 0 0 0.02 1 0.01 0
donations12 0.16 0.05 0.52 0.41 0.5 0.38 0.17 0.04 0.37 0.16 0.01 1 0.33
donations13 0.13 0.01 0.38 0.3 0.36 0.17 0.01 0.33 0.22 0.02 0 0.33 1
funding1
fund2
donations3
gifts4
donat5
transac6
grants7
money8
gift9
gift10
funding11
donations12
donations13
1.0
0.8
0.6
0.4
0.2
0.0