Predictive Modeling Results

Report Date: May 17, 2020

A. Data Input Summary

1. Total Data Sample: 3,800
    a. 80% (3,040) of data used for training the model
    b. 20% (760) of data used for testing the model
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

B. Running the Predictive Model: A Step by Step Guide

1. Read the input data file provided.
2. Cleaning up of data: remove null rows and columns and impute missing values.
3. Identifying columns containing categorical and textual data and converting it to numerical values. If a column has less than or equal to five unique values, then it is identified as categorical value.
4. Assigning Target Value: Target values are the dependent (predicted) variable. Total donation columns are calculated to assign target value. For example, if 50% of the total donation columns have a donation amount (> 0.00 value), the model assigns that row (record) as 1 otherwise 0.
5. Splitting the dataset for training and testing to train a total of 10 different classifiers (for example Naive Bayes, Logistic Regression and Random Forest).
6. Calculating 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. Identifying and selecting the best fit classifier (model) 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 a model is capable of distinguishing between classes (donor and non-donor).
10. Identifying the optimal threshold (accuracy of the model) and predict.

C. Important Terms Used in Predictive Modeling

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 a fraction of correctly classified instances among all actual/valid 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 (donor and non-donor). The top left box represents the count of true negatives, the top right box represents the count of false negatives, bottom left box represents the count of false positives and bottom right box represents the count of true positives.
6. Feature Importance Plot: Y-axis: feature present in input file and X-axis: relative % of feature importance.
7. Correlation Plot: Correlation explains how one or more variables are related to each other.
8. Probability Score: It is a probabilty (likelihood) of an individual to donate.
9. Threshold Value: It is the threshold (cut-off) value used on a probability score to seperate a donor from a non-donor.
10. Predicted Classification (0 and 1): Classification value 1 indicates an individual likely to donate and classification value 0 indicates an individual less likely to donate. They follow the threshold (cut-off) value logic.

D. Best Fit Model Used in Predictive Modeling

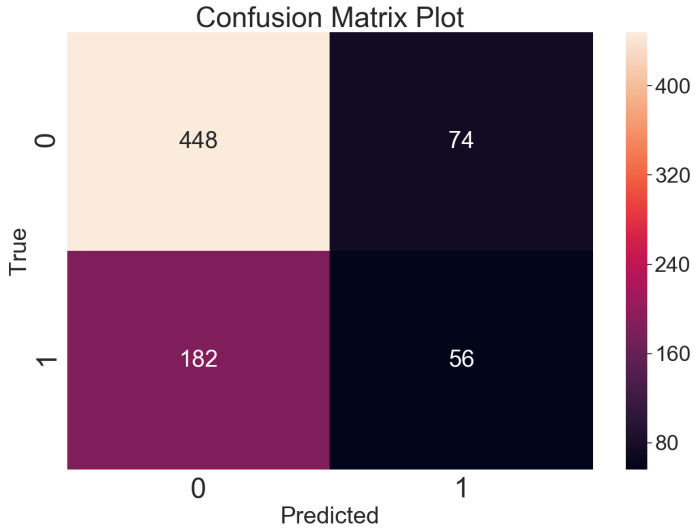
Best fit classifier (model) is selected (out of 10 classifiers) based on F1-score and used for prediction. Model identified the optimal threshold to separate classes (donor and non-donor). Following are F1-score, threshold and count of donor samples.

Best fit model name: BernoulliNB

a. F1-score (accuracy score): 0.63

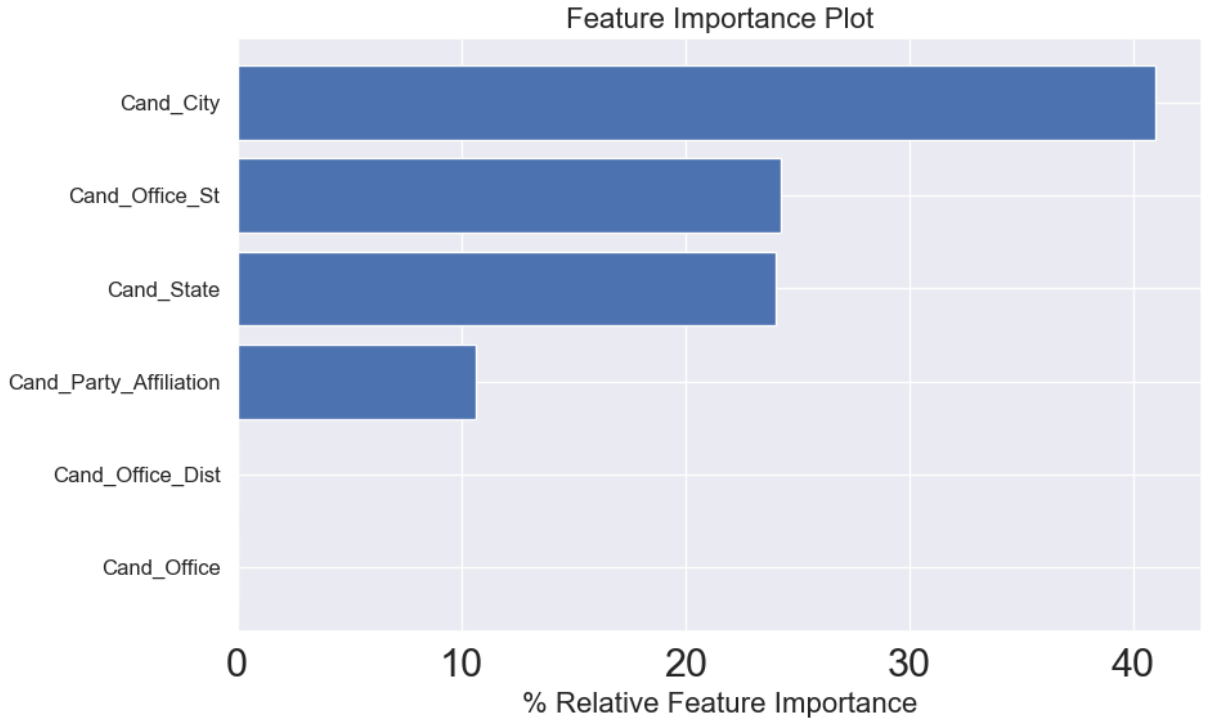
b. Threshold used: 0.2

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



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

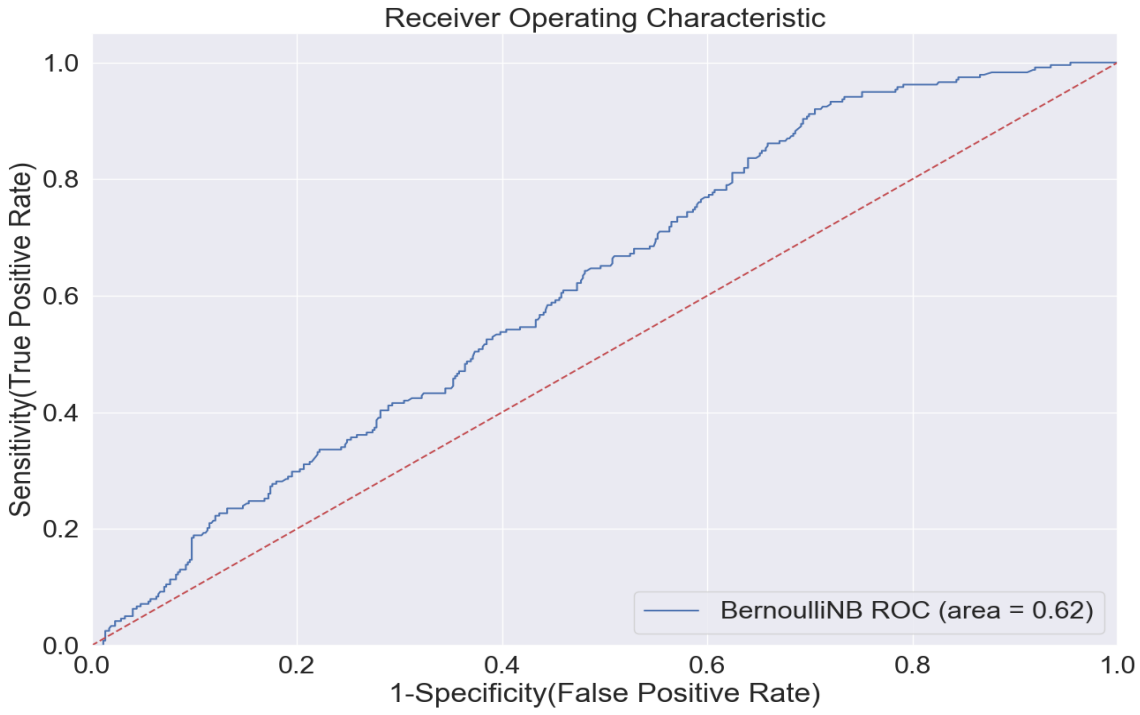


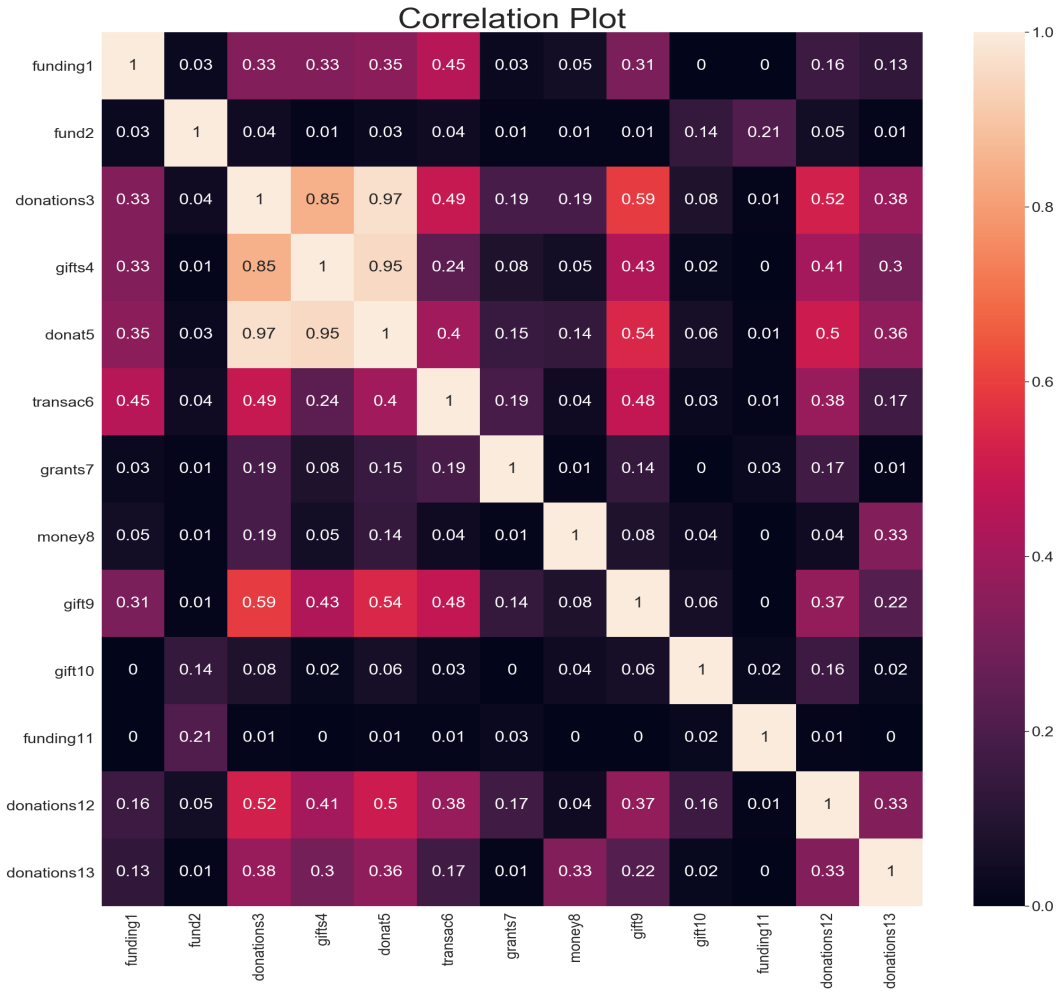
 # Receiver Operating Characteristic (ROC) Curve

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 explains how often a positive class is predicted when the actual result 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; the best model will have AUC close to 1.




Correlation Plot
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