Predictive Modeling Results

Report Date: December 01, 2020

A. Data Input Summary

1. Model Name: Ensemble Method (Top 3 best fit classifiers)
    a. SGDClassifier
    b. DecisionTreeClassifier
    c. RandomForestClassifier
2. Total Data Sample: 392
    a. 80% (313) of data used for training the model
    b. 20% (79) of data used for testing the model
3. Donation Columns:
    a. 2019 Gift
    b. 2018 Gift
    c. 2017 Gift
    d. 2016 Gift
    e. 2014 Gift
4. Categorical Data Features:
    a. Home Address
    b. City
    c. State
    d. Zip
    e. Volunteered in the past
    f. 2011 Gift
    g. CompTotal

B. Important Metrics Definition

1. Text Data Conversion: Process of converting text data into a form that a model can understand.
2. Target Value: Total donation columns from the input file are calculated to assign target values. For example, in each row, if 50% of the total donation columns have a donation amount >=1, the model assigns that row as 1 otherwise 0.
3. Assign Target Value/Class (1 and 0): 1 = class donor or 1. 0 = class non-donor or 0.
4. Data Imbalance: Skewness of the dataset.
5. Training Set: Subset of data to train a model. Test set: Subset of data to test the trained model.
6. Data Feature Importance: Process of selecting data features that contributes the most in prediction.
7. Classifier/Model: Classifier or classification model is an algorithm that predicts classes.
8. Ensemble Method: Technique of combining several models to generate one best model.
9. Soft Voting: Process of generating best result by averaging all the predicted probabilities calculated by distinct models.
10. Performance Metrics: Metrics to explain the performance of a model.
11. Precision: Fraction of correct predictions for a certain class. It refers to the percentage of results that are relevant.
12. Recall: Fraction of correct predictions of all actual classes. It refers to the percentage of total relevant results correctly classified.
13. F1- Score: Measure of a model's accuracy. A perfect model has an F1-score of 1.
14. Confusion Matrix Plot: Visualized table describing the performance of a model. Each row in a confusion matrix represents an actual class, while each column represents a predicted class (or vice versa). Classification report table is the performance metrics of a model.
15. ROC Curve: Graph representing predicted performance of a model.
16. Threshold: Cut-off value on a probability score to separate a donor from a non-donor.

17. Probability Score: Predicted probability (likelihood) score of an individual to donate.
18. Predicted Classification (1 and 0): 1 indicates an individual likely to donate. 0 indicates an individual less likely to donate. They follow the threshold (cut-off) value logic.

C. Steps on Building and Executing Predictive Model

1. Read the input data file.
2. Data cleanse. Impute missing values and remove null rows and columns.
3. Convert text data to numbers.
4. Assign Target Value. Target values are the predicted variable.
5. Evaluate data imbalance.
6. Split data for training and testing.
7. Calculate Data Feature Importance.
8. Evaluate performance metrics of 10 classifiers (models) and select top 3 best fit classifiers.
9. Combine top 3 best fit classifiers predictions using a soft voting ensemble method.
10. Create Confusion Matrix, Classification Report and Receiver Operating Characteristic (ROC) Curve.
11. Identify the threshold and predict.
13. Generate model summary report (PDF) and CSV file with the Assigned Target Value, Donor Probability Score and Donor Predicted Classification columns appended back to the processed file.

D. Ensemble Method Output Metrics

Soft voting ensemble method used to combine the predictions of the top 3 best fit classifiers (models). Following are the Assigned Target Value, F1-score, Threshold and Total Donors Predicted metrics.

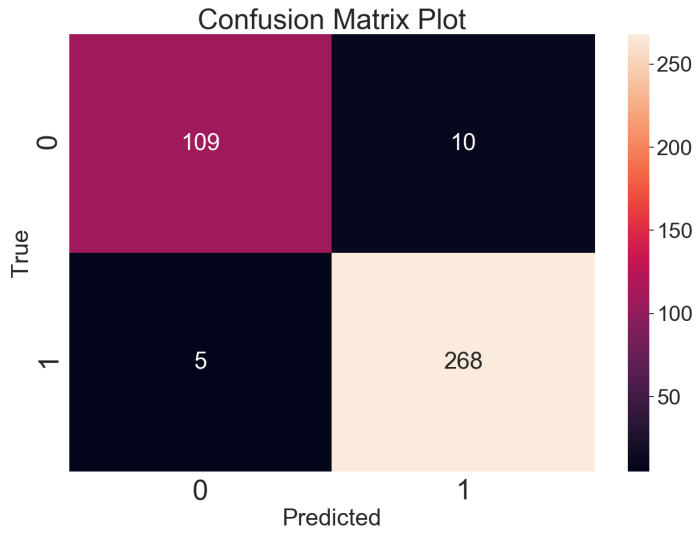
a. Assigned Target Value (class donor): 273

b. F1-Score: 0.76

c. Threshold: 0.6

d. Donors Predicted: 70.92% (278 out of 392)

D (a). Confusion Matrix Plot



Based on the confusion matrix, a total of 273 (5 + 268) samples belong to class donor and 119 (109 + 10) samples belong to class non-donor. The model correctly predicted 268 samples as class donor and 109 samples as class non-donor. The model misclassified 5 class donor as class non-donor and 10 class non-donor as class donor.

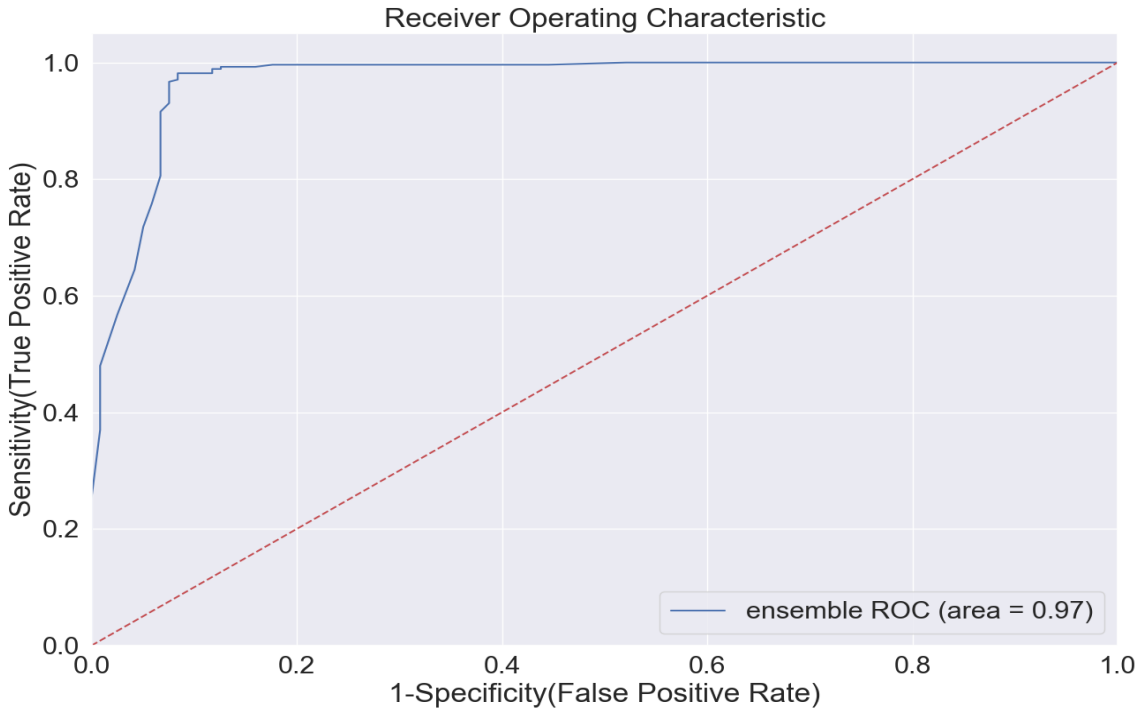
D (b). Classification Report Table

Class	Precision	Recall	F1-score
Non-donor class	0.96	0.92	0.94

Donor class	0.96	0.98	0.97

D (c). ROC Curve

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. The best model has AUC close to 1.



D (d). Feature Importance Plot