Predicting Potential Donors and Dollars

Using Classification and Predictive Models to Increase Profitability of a Random Marketing Campaign

Jorge Leon Frausto, Kyla Whittington
Michael Pichler and Thi Phan

12/8/20

ADTA 5230

Executive Summary

With the emergence of big data and data analytics, more companies are exploring the benefits of making decisions based on data, instead of past experiences or collective knowledge. The days of doing what you have always done and hoping it works is no longer the only option. Using the JMP application for statistical analysis, we built various predictive models to evaluate the classification model's ability to determine binary outcomes andthe prediction model's ability to predict an expected amount. Using statistical measures, such as misclassification rate, we are able to select the best classification and predictive model to not only create a targeted marketing strategy towards expected donors, but estimate the donated amount in dollars from each donor identified. This will significantly change a direct marketing campaign that is not cost-effective into a profitable campaign by simply implementing big data analytics.

Introduction and Business Objective

Based on recent records, the overall response rate is approximately 10 percent of potential donors. Of the mail recipients who respond and donate, the average donation is $14.50. When comparing the average donation to the cost of each mailing, which is $2.00, it would be more cost-effective to send the mailing to a targeted mailing list as opposed to sending to the existing mailing list. The business objective is to develop predictive models for a non-profit organization to improve cost-effectiveness and achieve maximum expected net profit for direct marketing campaigns to previous donors. The first goal is to develop a classification model to effectively capture likely donors using a mailing marketing campaign, so that net profits can be maximized. The second goal is to develop a prediction model to forecast expected gift amounts from potential donors. We will use statistical measures to assess how well each model is able to classify or predict, including the misclassification rate, the confusion matrix (sensitivity and

specificity), the ROC curve (AUC) and the Lift curve. This project will follow the well-known CRISP-DM methodology.

Data Collection and Description

The full dataset consists of 8,009 observations, which is split 50/25/25 between training (3,984), validation (2018) and test (2007) datasets. Weighted sampling has been used to over-represent the responders to equally divide the training and validation sets between donors and non-donors, so that the dataset is balanced. Additionally, the response rate for the test sample is equal to the overall response rate of 10 percent based on historical data. The initial dataset consists of 21 continuous predictor variables, one nominal classification response variable, one continuous prediction response variables, and one nominal partition variable, which shows the 50/25/25 split between the three data sets.

1. ID – ID number was not used as a predictor variable in any models
2. TER1 – Region 1, where 1 = potential donor belongs to this region
3. TER2 – Region 2, where 1 = potential donor belongs to this region
4. TER3 – Region 3, where 1 = potential donor belongs to this region
5. TER4 – Region 4, where 1 = potential donor belongs to this region
6. OWND – Homeowner, where 1 = homeowner, 0 = not a homeowner
7. KIDS – Number of children
8. INC – Household income (seven categories)
9. SEX – Gender, where 0 = Male, 1= Female
10. WLTH – Wealth Rating uses median family income and population statistics from each area to index the relative wealth within each state. Segments are ranked from 0 to 9, where 9 is highest wealth group
11. HV – Average Home Value in potential donor's neighborhood in $ thousands
12. INCMED - Median Family Income in potential donor's neighborhood in $ thousands
13. INCAVG - Average Family Income in potential donor's neighborhood in $ thousands
14. LOW - Percent categorized as "low income" in potential donor's neighborhood

15. NPRO – Lifetime number of promotions received to date
16. GIFDOL – Dollar amount of lifetime gifts to date
17. GIFL – Dollar amount of largest gift to date
18. GIFR – Dollar amount of most recent gift
19. MDON – Number of months since last donation
20. LAG – Number of months between first and second gift
21. GIFA – Average dollar amount of gifts to date
22. DONR – Classification Response Variable, where 1 = donor, 0 = Non-donor
23. DAMT – Prediction Response Variable, donation amount in dollars.

Exploratory Data Analysis: Correlation

Based on the multivariate correlation matrix, the following groups of variables are highly correlated with an r-square greater than ±0.5000:

- Average Home Value (HV), Median Family Income (INCMED), Average Family Income (INCAVG) and Percent categorized as Low Income (LOW)

Variables	HV	INCMED	INCAVG	LOW
HV	1	0.72291	0.839952	-0.63468
INCMED	0.72291	1	0.868454	-0.66001
INCAVG	0.839952	0.868454	1	-0.64141
LOW	-0.63468	-0.66001	-0.64141	1

- Lifetime number of promotions received to date (NPRO) and Dollar amount of lifetime gifts to date (GIFDOL)

Variables	NPRO	GIFDOL
NPRO	1	0.70897
GIFDOL	0.70897	1

- Dollar amount of largest gift to date (GIFL), Dollar amount of most recent gift (GIFR) and Average dollar amount of gifts to date (GIFA)

Variables	GIFL	GIFR	GIFA
GIFL	1	0.709875	0.63215
GIFR	0.709875	1	0.709315
GIFA	0.63215	0.709315	1

Exploratory Data Analysis: Data Quality and Distribution

The size of the data set is significant for data modeling and there are no missing values in the training or validation partitions. The test partition of the provided dataset can only be used for model cross-validation due to the removed or missing outcome variable values (see *DONR* and *DAMT* below), therefore it will be treated as new data for model evaluation by the non-profit organization.

VARIABLE	MIN	MAX	MEDIAN	MEAN	STDEV
Kids	0	5	2	1.6898	1.4018
House Value	51	710	170	182.2263	72.5629
Median Income	3	287	38	43.6175	24.6150
Average Income	12	287	52	56.5626	24.7221
Low Income Percentage	0	87	10	14.2276	13.4115
Lifetime Promotions	2	164	58	60.2633	30.3241
Lifetime Gifts to Date	23	1974	89	113.4346	83.6920
Largest Gift to Date	3	681	16	22.9188	30.0191
Most Recent Gift Amount	1	173	12	15.6637	12.3519
Number of Months Since Last Gift	5	40	18	18.8181	5.6587
Number of Months Between First and Second Gift ("Lag")	1	34	6	6.4680	3.7781
Average Gift Amount	1.29	72.27	10.22	11.6618	6.5222
Donation Amount $	0	27	0	7.2091	7.3612

Figure 1 Descriptive Statistics for Continuous Variables

Numerical Variables Distributions

The distributions graphs for House Value, Median Income and Lag (Months Between First Two Donations) show a positive-skewed distribution. The distributions of Number of Promotions and Number of Months Since Last Donation reflect graphs closer to a normal distribution. Distribution graphs for continuous variables are shown below.

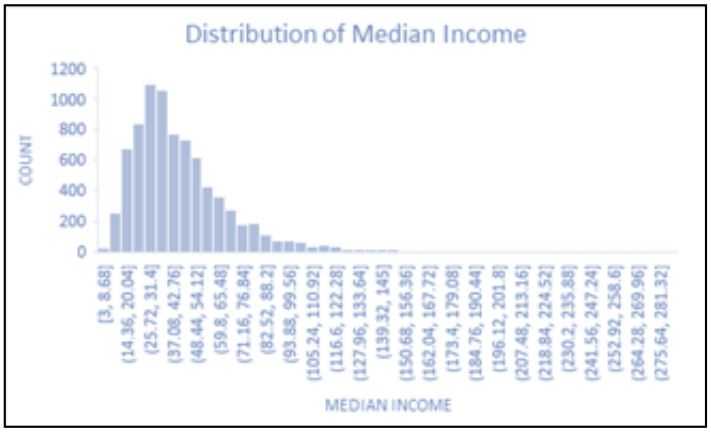


Figure 2. Distribution Graph for Median Income - Positive - Skewed Distribution

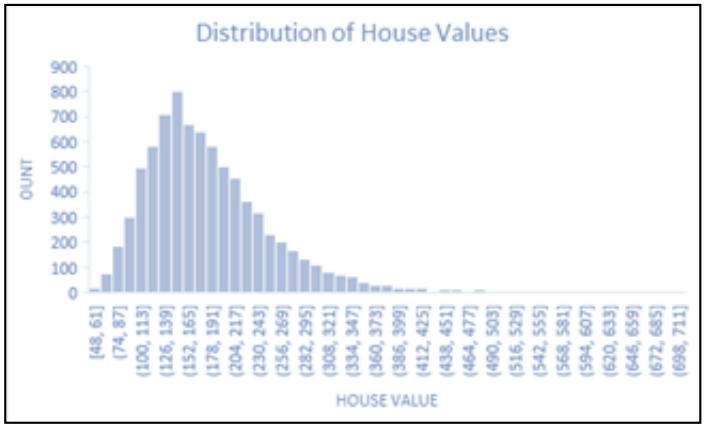


Figure 3. Distribution Graph for House Values- Positive-Skewed Distribution

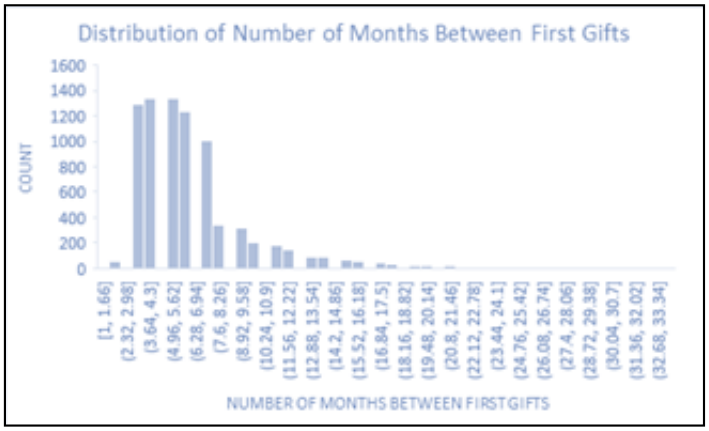


Figure 4. Distribution Graph for Lag(Months Between First 2 Donations) - Positive-Skewed Distribution

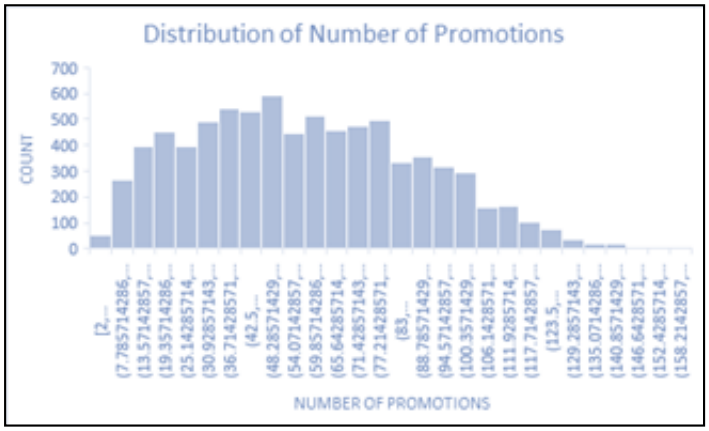


Figure 5. Distribution Graph for Number of Promotions - Normal Distribution

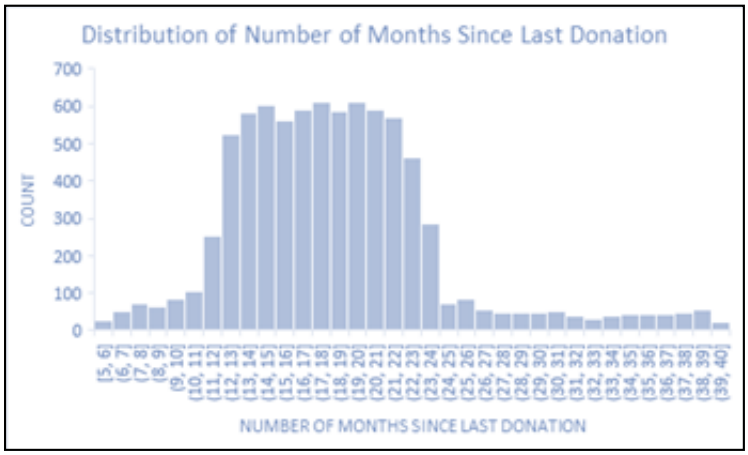


Figure 6. Distribution Graph for Number of Months Since Last Donation - Normal Distribution

The distribution of Months Since Last Donation is of particular interest because it shows a drastic increase in observations between 12 and 24 months. When compared to the average donation, we found a significant influence for this range indicating that people who have not donated in over a year, but in less than two years, are likely to give a larger gift. From this observation, the data was broken into three groups: Those who gave last donations less than a year ago (<=12); those whose last donation is more than one year but less than two; and a third group for those whose last donation was more than two years ago (> 24).

Nominal Variables Distributions

Bar charts were created to show frequency of categorical variables and number of Kids, and indicate the following:

- 43 percent of observations belongs to region 2.
- The most common number of kids are 0 and 2.
- Most of the data belonged to Wealth Rating 8 and 9.
- More than 86% of observations were homeowners.
- More than 82% of observation fall under Month Since Last Donation group 2.
- 61 percent of observations are female.

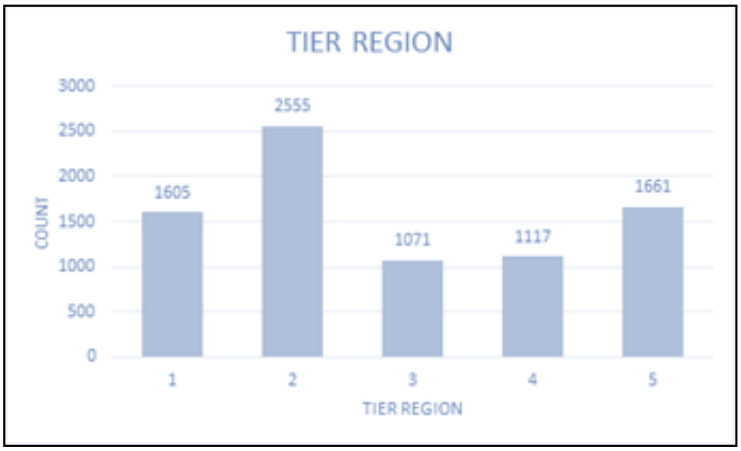


Figure 7. Nominal Variables - Distribution of Tier Regions

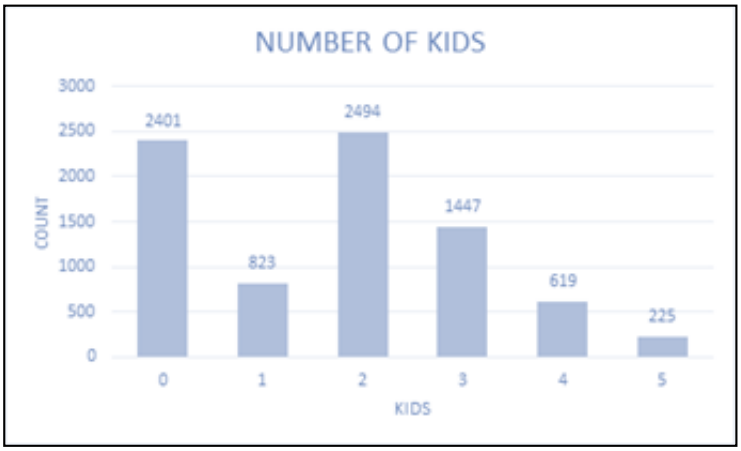


Figure 8. Nominal Variables - Distribution of Number of Kids

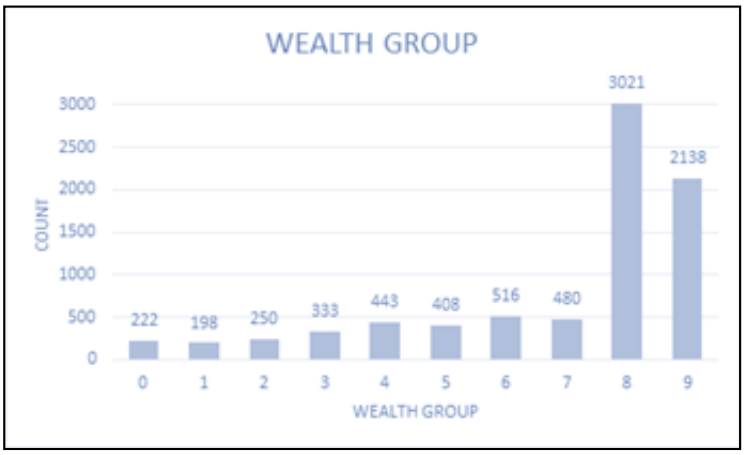


Figure 9. Nominal Variables - Distribution of Wealth Group

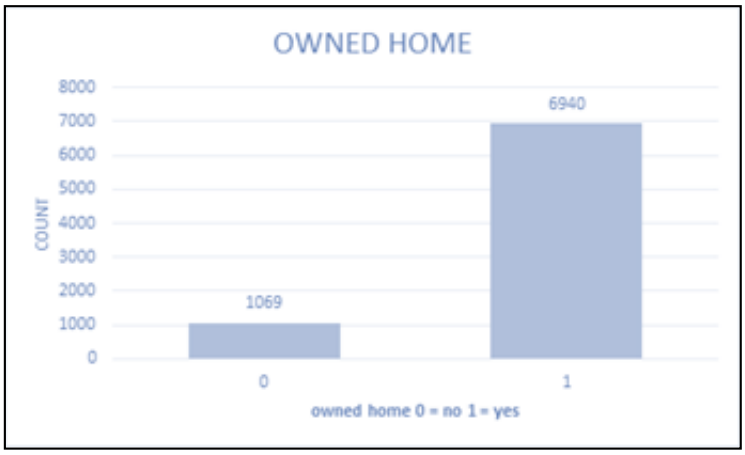


Figure 10. Nominal Variables - Distribution of Owned Home

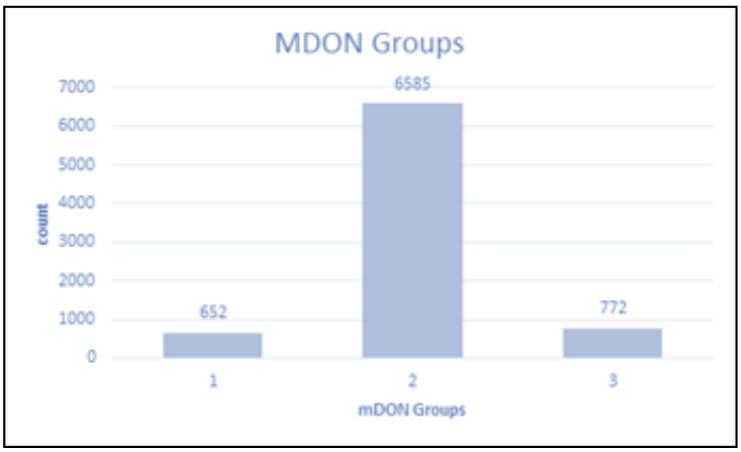


Figure 11. Nominal Variables - Distribution of Months Since Last Donation

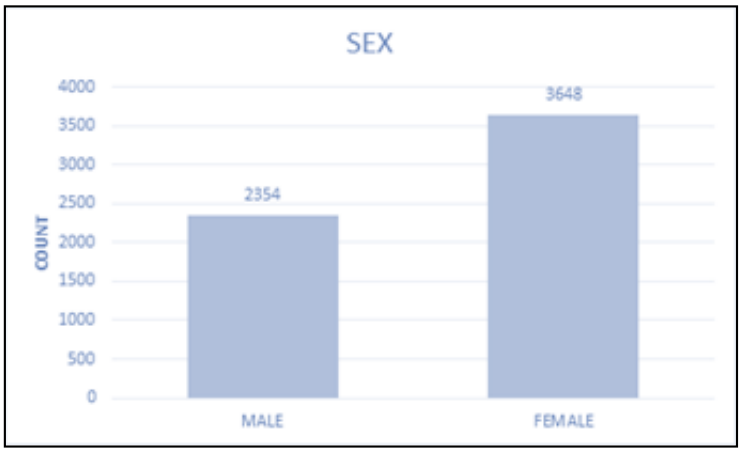


Figure 12. Nominal Variables - Distribution of Sex

Level of Measurement and Principal Component Analysis

Prior to modeling the data, it was necessary to update the levels of measurement for several variables. With exception to the outcome variables Donor (DONR), Donation Amount (DAMT), and the partition flag (PART), all variables were listed as continuous. Upon careful review of the data dictionary, we discovered the following variables should be listed as nominal due their categorical nature: TER1, TER2, TER3, TER4, OWND, INC, and SEX. Additionally, WLTH was updated to ordinal to reflect its rank nature.

In an effort to reduce dimensionality for the classification models, Principal Component Analysis (PCA) was conducted for the predictor clusters surrounding the home (HV, INCMED, INCAVG, LOW), and marketing & response (NPRO, GIFL, GIFR, GIFA, GIFDOL).

The results of the Principal Component Analysis variables are displayed below:

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.1922	79.804		79.804
2	0.4334	10.836		90.640
3	0.2781	6.952		97.592
4	0.0963	2.408		100.000

Scatterplot 3D

Variable Clustering

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
hv	0.50269	0.26262	-0.73475	0.37212
incmed	0.51180	0.18054	0.67073	0.50556
incavg	0.52827	0.33692	0.08975	-0.77419
low	-0.45420	0.88596	0.04699	0.08108

Figure 13. Eigenvalues - PCA Home includes INCAVG, INCMED, HV, and LOW

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	2.3906	47.812		47.812
2	1.6949	33.897		81.709
3	0.3900	7.800		89.509
4	0.2787	5.574		95.083
5	0.2459	4.917		100.000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
npro	0.07142	0.70564	0.33218	0.42033	0.45821
gifdol	0.14571	0.69059	-0.28365	-0.42581	-0.48998
gifl	0.56720	-0.05446	-0.64390	-0.05003	0.50814
gifr	0.58202	-0.10214	0.05615	0.60577	-0.52982
gifa	0.55964	-0.10843	0.62565	-0.52207	0.10509

Figure 14. Eigenvalues - PCA Marketing & Response includes NPRO, GIFDOL, GIFL, GIFR, and GIFA

Based on the Analyses, dimension reduction was possible by leveraging the principal component(s) in place of the included individual predictors. The 'Home' principal component 1 captures nearly 80% of the correlation and allows for an 75% reduction of variables. The 'Marketing & Response" principal components 1 and 2 also capture just over 81% of the correlation as an 60% reduction in variables needed. These principal components were saved as new predictors to be used in the modeling process.

Another effort to reduce dimensionality was to create a predictor to act as a binary classifier of kids, reducing the complexity from 6 categories (0-5) to just 2 (0- no children, 1 – at least 1 child). The predictor WLTH also presented an opportunity for reduction, however it is unclear from the provided data dictionary if the categories are hierarchical in nature, therefore binning would not be appropriate. Additionally, while the partition flag (PART) identified the training, validation, and test data sets, it would not be recognized by the statistical analysis software, so a new "Validation" column was created to be used in the various modeling.

Because the provided data was defined as over-representing the class of interest (DONR=1) in the training and validation partitions, as evidenced in our early data analysis (representing 49.5% of cases within the validation split), we can expect improved performance of modeled classifiers. However, to assess the model performance in the population, where the response rate is much lower, a weight variable was created to be used in such classification models with the relationship as follows:

	DONR = 1 (DONOR)	**DONR = 0 (NON-DONOR)**
% OF POPULATION	10%	90%
% OF VALIDATION SET	49.5%	50.5%
CALCULATED WEIGHT	1/4.950	1/0.561

As models were completed, the prediction formulas, classification probabilities, classifications, profit assessments, and value predictions would also be saved as variables to assess.

Modeling

To address the business requests, separate modeling techniques were used to classify donors and predict donation amounts. Thanks to modern computing power and statistical software, many different classification and prediction modeling techniques could be quickly employed and reviewed. For both class membership prediction and numerical value prediction for donation, models were created using the training partition and measured for goodness of fit and accuracy on the validation data with weight applied.

Evaluation of Classification Models

To select a classification model to best achieve the highest expected net profit from direct marketing to donors, the model which resulted in the highest lift curve incorporating costs, in combination with healthy prediction accuracy measures indicating strong predicted class

membership would be selected. The critical prediction accuracy measures considered for classification models included the misclassification rate, sensitivity, specificity, a receiver operating character (ROC) curve's area under the curve (AUC) metric, and lift. A brief description of these metrics are as follows:

- **Misclassification rate**: Also known as the overall error rate, represents the % of classes that are classified incorrectly. A measurement of 0 would indicate perfect classification for all classes.
- **Sensitivity**: Accuracy measure for the class of interest. A measurement close to 1 indicates a strong ability to identify the class of interest.
- **Specificity**: Accuracy measure for the class not of interest. A measurement close to 1 indicates a strong ability to rule out unimportant class members.
- **AUC:** Measurement of discrimination between classes. A measurement of 1 indicates perfect discrimination.
- **Lift:** Measures the model accuracy when sorted by propensity. A desirable lift curve contains the highest lift for the strongest propensities and falls to a lift of 1 as the portion of data reaches 100%.

Classification Models & Individual Assessment

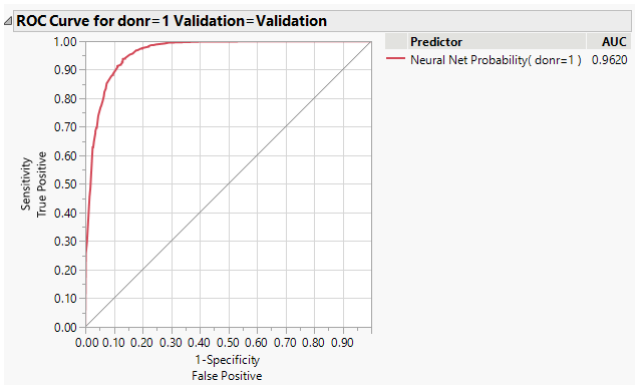
Neural Network

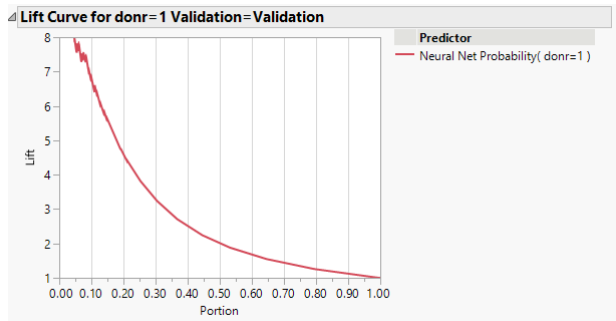
The neural network model was selected due to their high predictive performance, which captures complex relationships between predictors and outcome. Because the correlation analysis revealed weak correlations among most predictors and outcome, the neural network analysis is a good candidate to "learn" what relationships contribute to an outcome. The most successful neural network model was configured to learn using all variables with all principal component replacement to reduce dimensionality. The model included a modest 3 hidden layers using the hyperbolic tangent (TanH) transformation. More complex models with additional hidden layers, boosting and additional tours did not result in improved model performance.

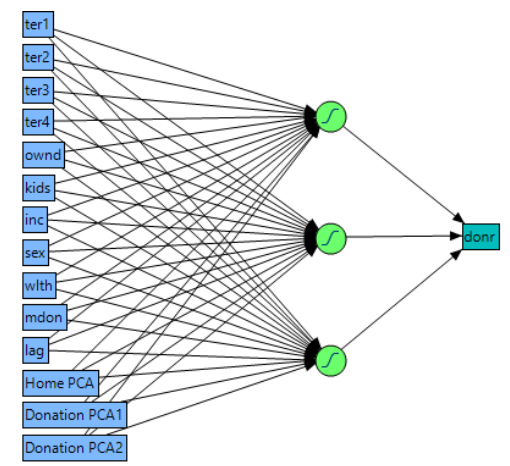
The model performance is as follows:

Measures of Fit for donr

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Neural		0.3514	0.4283	0.2138	0.2618	0.1227	0.0977	3948	0.9750
Validation	Neural		0.1146	0.1502	0.2878	0.2982	0.1453	0.1251	2018	0.9620







Confusion Matrix

Target donr

Predictor Neural

Training

Actual donr	Predicted Count 0	1	NA
0	3187	358	0
1	27	376	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.899	0.101	0.000
1	0.068	0.932	0.000
NA	.	.	.

Validation

Actual donr	Predicted Count 0	1	NA
0	1578	239	0
1	14	188	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.868	0.132	0.000
1	0.067	0.933	0.000
NA	.	.	.

It can be seen from the model performance that this neural network model offers strong predictive accuracy via:

- Low misclassification rate =0.1251
- Strong sensitivity rate = 0.933
- Strong specificity rate = 0.868
- Holding a lift gain (max ~8.0) above 2.0 beyond 50% of the data
- AUC rate = 0.962

Boosted Tree

Boosted Tree Model Advantages and Limitations:

The Boosted Tree was selected due to being highly efficient at building classification and regression models. It is often considered more accurate than the Bootstrap or Random Forest as it can handle mixed measurement types of features and no-preprocessing is required. The disadvantages of using a Boosted Tree model is it requires tuning of the parameters, which include Number of Layers, Splits per Tree, Learning Rate, Minimum Size Split, Row Sampling Rate and Column Sampling Rate. The Boosted Tree model may be at risk for overfitting if too many trees are used in the model, and lastly, it is vulnerable to outliers. The model is based on the following parameters:

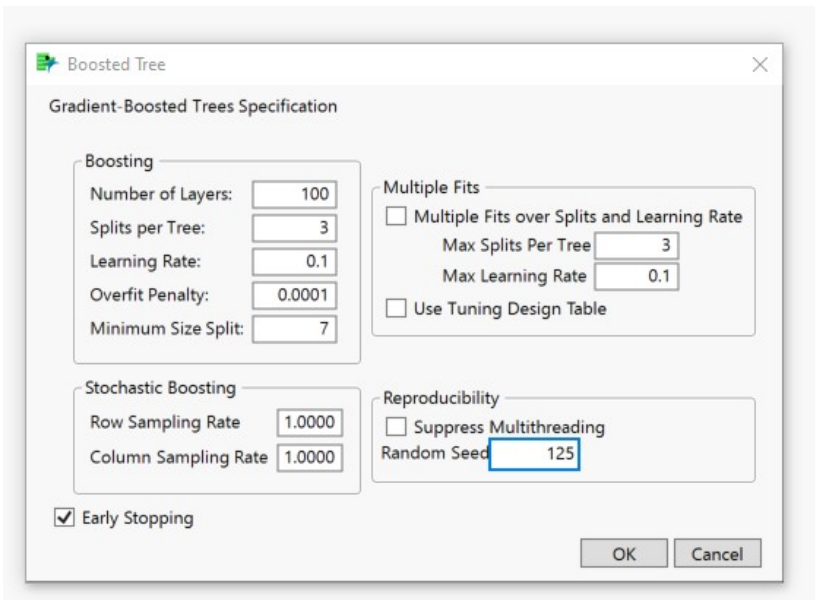
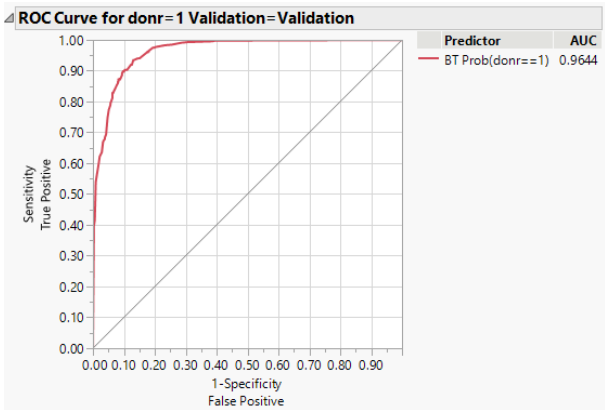


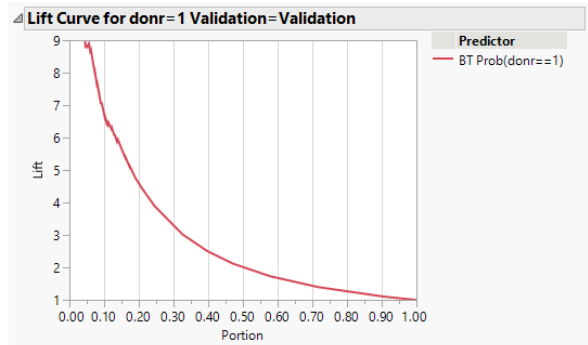
Figure 13. Parameter for Boosted Tree Model

Results and Measure of Fit for DONR using Boosted Tree Model

Measures of Fit for donr

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Boosted Tree		0.2203	0.2801	0.257	0.2821	0.1808	0.1122	3948	0.9709
Validation	Boosted Tree		0.1134	0.1486	0.2882	0.3044	0.1958	0.1314	2018	0.9644





Confusion Matrix

Target donr

Predictor Boosted Tree

Training

Actual donr	Predicted Count 0	1	NA
0	3128	417	0
1	26	377	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.882	0.118	0.000
1	0.064	0.936	0.000
NA	.	.	.

Validation

Actual donr	Predicted Count 0	1	NA
0	1563	253	0
1	12	190	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.861	0.139	0.000
1	0.060	0.940	0.000
NA	.	.	.

It can be seen from the model performance that this boosted tree model offers strong predictive accuracy based on the following statistical metrics:

- Low misclassification rate =0.1314
- Strong sensitivity rate = 0.940
- Strong specificity rate = 0.861
- Holding a lift gain (max ~9.0) above 2.0 beyond 50% of the data
- AUC rate = 0.9644

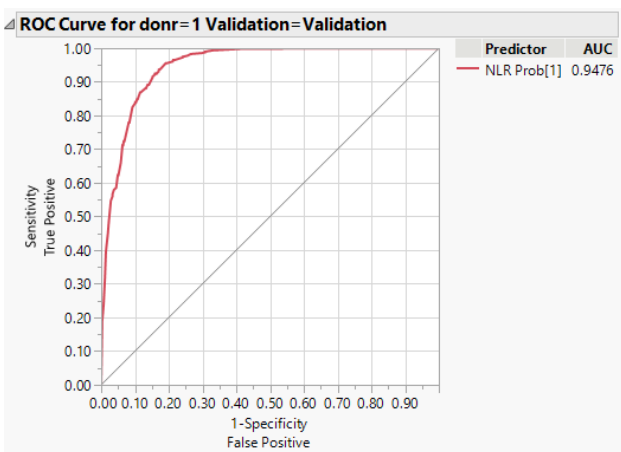
Nominal Logistic Regression

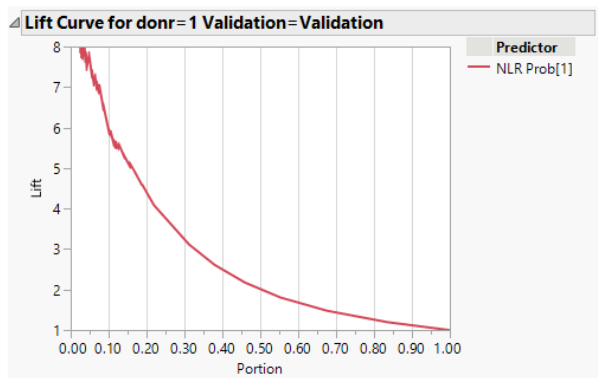
As the nature of our business objective is to identify a portion of an audience to target for marketing, the logistic regression model, with its ability to profile in addition to measuring probability, could serve as an excellent classifier. All predictors with principal component replacement were used. Using backward elimination to remove those with the least statistically significant contribution did not improve the model. Using all predictors in place of the principal components improved the misclassification rate, but the additional variables did not contribute with a contribution of significance, therefore the less complex model was selected to avoid overfitting.

Results and Measure of Fit for DONR using Nominal Logistic Regression

Measures of Fit for donr

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Nominal Logistic		0.1916	0.2458	0.2665	0.2893	0.1537	0.1117	3948	0.9605
Validation	Fit Nominal Logistic		-0.013	-0.018	0.3294	0.3202	0.1740	0.1452	2018	0.9476





Effect Summary

Source	LogWorth	PValue
kids	359.414	0.00000
inc	149.970	0.00000
ownd	91.925	0.00000
ter2	83.780	0.00000
wlth	67.301	0.00000
Home PCA	29.383	0.00000
Donation PCA2	22.329	0.00000
lag	20.741	0.00000
ter1	18.054	0.00000
mdon	6.567	0.00000
Donation PCA1	1.192	0.06424
sex	0.909	0.12339
ter3	0.322	0.47599
ter4	0.162	0.68930

Confusion Matrix

Target donr

Predictor Fit Nominal Logistic

Training

Actual donr	Predicted Count 0	1	NA
0	3141	405	0
1	37	366	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.886	0.114	0.000
1	0.091	0.909	0.000
NA	.	.	.

Validation

Actual donr	Predicted Count 0	1	NA
0	1542	275	0
1	19	183	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.849	0.151	0.000
1	0.092	0.908	0.000
NA	.	.	.

It can be seen from the model performance that this nominal logistic regression model offers strong predictive accuracy via:

- Low misclassification rate =0.1452
- Strong sensitivity rate = 0.908
- Moderate specificity rate =0.849
- Holding a lift gain (max ~8.0) above 2.0 to ~50% of the data.
- AUC rate = 0.9476

Bootstrap Forest

Screening Predictor Variables

Using the predictor screening analysis tool, JMP ranks the variables in order of its contribution to predicting the DONR response variables using the bootstrap forest model. For the DONR variable, the predictor variable KIDS is ranked first for contributing 37 percent using all variables, including PCA variables. All variables were included as the model searches for the best feature among the random subset of features.

	donr		
Predictor	Contribution	Portion	Rank
kids	943.807	0.3726	1
inc	426.279	0.1683	2
ownd	247.018	0.0975	3
ter2	232.104	0.0916	4
wlth	175.981	0.0695	5
mdon	122.433	0.0483	6
Donation PCA2	89.330	0.0353	7
lag	87.851	0.0347	8
Home PCA	82.331	0.0325	9
ter1	36.801	0.0145	10
Donation PCA1	34.187	0.0135	11
ter4	25.192	0.0099	12
ter3	21.171	0.0084	13
sex	8.483	0.0033	14

Bootstrap Forest Advantages and Limitations:

Bootstrap Forest takes random samples with replacement of size n and repeats for each layer (B) and refits the model over each dataset or sample. This model is a relatively stable algorithm, and it adds the following advantages:

- Does not require normalizing or standardizing the data prior to use
- Reduces the risk of overfitting
- Reduces the variance and improves the model accuracy
- Robust to handling outliers, missing data, and non-linear parameters

As with all models, there are limitations or disadvantages to the Bootstrap Forest model. This type of model, compared to Decision Tree models, tends to be complex as the random forest creates a large number of trees. Additionally, the Bootstrap Forest model requires a longer training period due to the tuning of the six hyperparameters. The following hyperparameters were set to run the model:

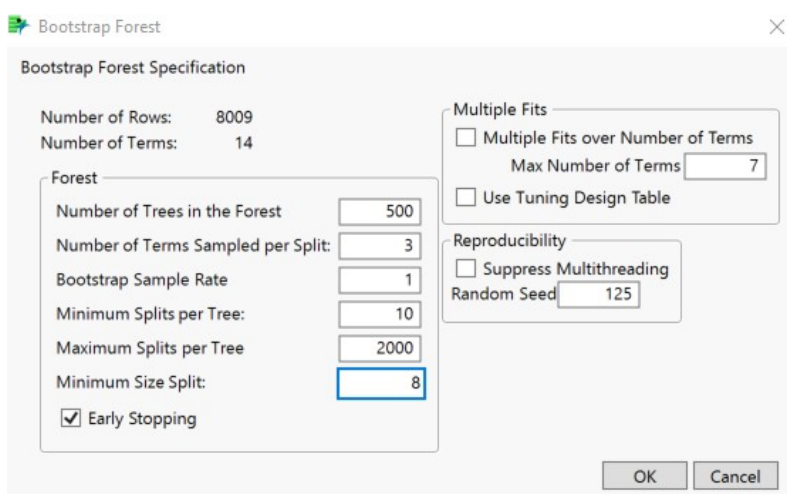
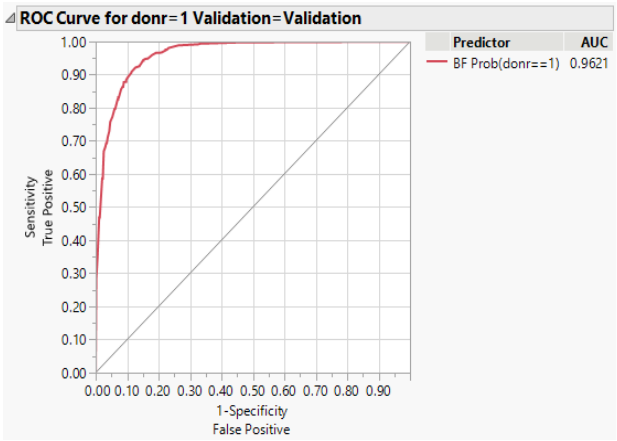


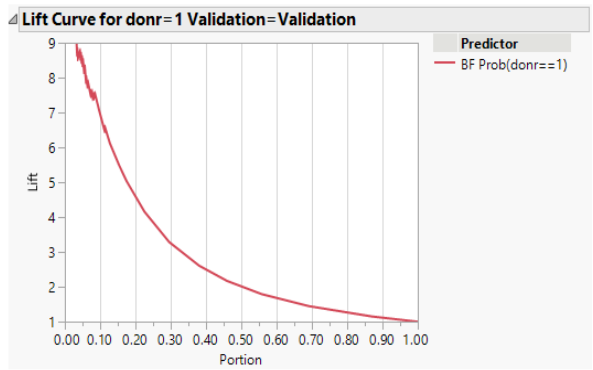
Figure 14. Parameter for Bootstrap Forest

Results and Measure of Fit for DONR using Bootstrap Forest Model

Measures of Fit for donr

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Bootstrap Forest		0.0603	0.0807	0.3098	0.2936	0.2439	0.0915	3948	0.9831
Validation	Bootstrap Forest		-0.126	-0.178	0.366	0.3322	0.2748	0.1400	2018	0.9621





Confusion Matrix

Target donr

Predictor Bootstrap Forest

Training

Actual donr	Predicted Count 0	1	NA
0	3201	344	0
1	17	386	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.903	0.097	0.000
1	0.043	0.957	0.000
NA	.	.	.

Validation

Actual donr	Predicted Count 0	1	NA
0	1545	271	0
1	12	190	0
NA	0	0	0

Actual donr	Predicted Rate 0	1	NA
0	0.851	0.149	0.000
1	0.057	0.943	0.000
NA	.	.	.

It can be seen from the model performance that this bootstrap forest model offers strong predictive accuracy via:

- Low misclassification rate =0.14
- Strong sensitivity rate = 0.943
- Strong specificity rate = 0.851
- Holding a lift gain (max ~9.0) above 2.0 beyond 50% of the data
- AUC rate = 0.9621

Evaluation

Class Prediction Model Result Comparison

Measures of Fit for donr

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Training	Neural		0.3514	0.4283	0.2138	0.2618	0.1227	0.0977	3948
Training	Fit Nominal Logistic		0.1916	0.2458	0.2665	0.2893	0.1537	0.1117	3948
Training	Partition		-0.071	-0.100	0.3531	0.3404	0.2053	0.1703	3948
Training	Bootstrap Forest		0.0603	0.0807	0.3098	0.2936	0.2439	0.0915	3948
Training	Boosted Tree		0.2203	0.2801	0.257	0.2821	0.1808	0.1122	3948
Validation	Neural		0.1146	0.1502	0.2878	0.2982	0.1453	0.1251	2018
Validation	Fit Nominal Logistic		-0.013	-0.018	0.3294	0.3202	0.1740	0.1452	2018
Validation	Partition		-0.211	-0.308	0.3938	0.3621	0.2221	0.1948	2018
Validation	Bootstrap Forest		-0.126	-0.178	0.366	0.3322	0.2748	0.1400	2018
Validation	Boosted Tree		0.1134	0.1486	0.2882	0.3044	0.1958	0.1314	2018

While many of the models presented impressive prediction accuracy measures, our business goal is to maximize expected net profit from direct marketing sent to targeted donors. Capturing the impact each model has on net profit within the validation partition makes use of each of the prediction accuracy measures in a simple way:

1. Assign a net profit value of -$2.00 to false positive donor cases, and a net profit value of $12.50 to true positive donor classifications. Applying the weighting counterbalance.
2. Sort the cases by descendingpropensities.
3. Calculate the cumulative profit for all cases.
4. Select the model with the largest maximum cumulative profit.

This model selection method should work well because the model which has the best accuracy and ability to discriminate accurately should provide the greatest cumulative net profit. By reviewing the descriptive statistics of the models' cumulative profit, we can see that the strongest performing model for this business goal is the boosted tree model.

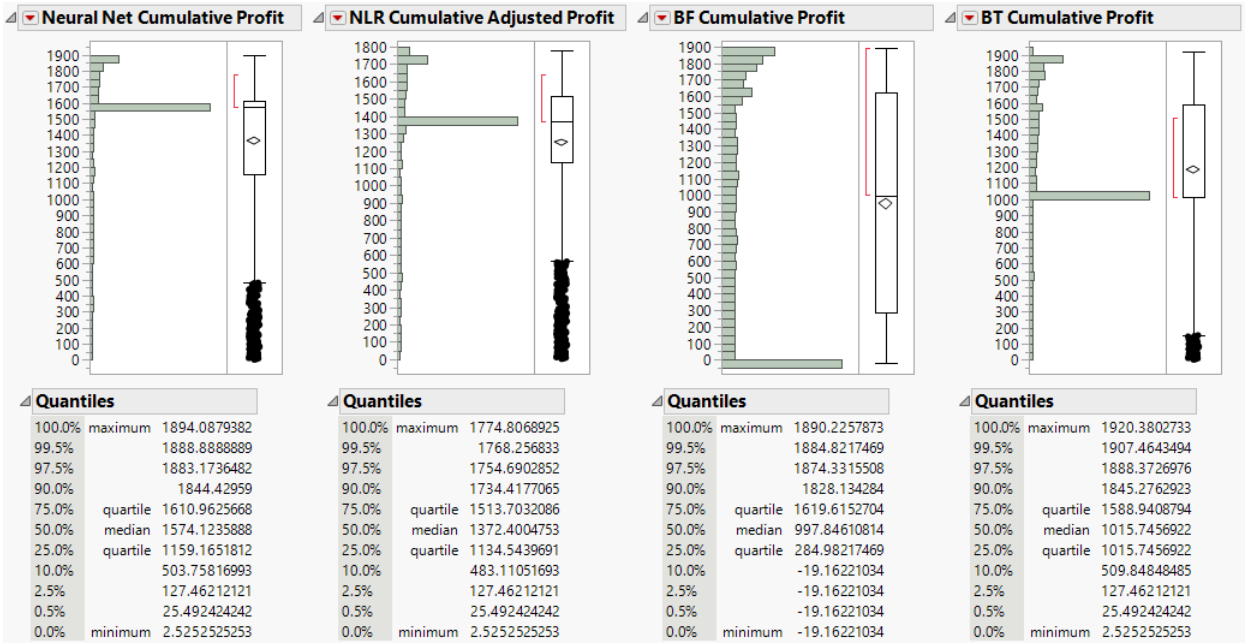
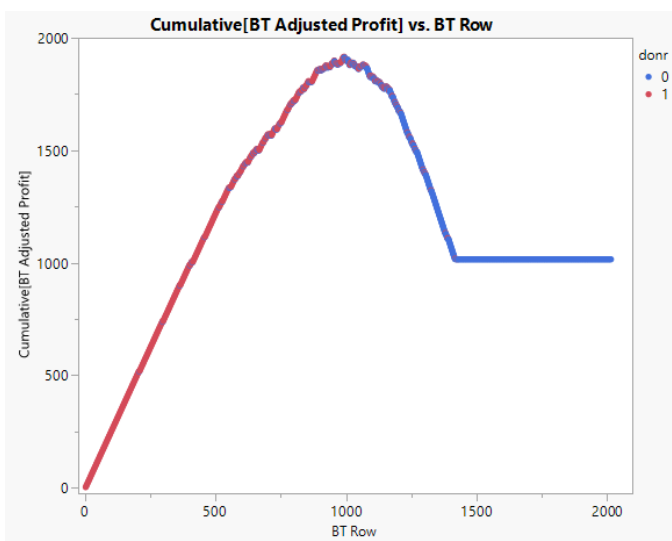


Figure 15: Cumulative Profits for Neural Net, Nominal Linear Regression, Bootstrap Forest, and Boosted Tree

A lift curve incorporating costs for the boosted tree model illustrates how the cumulative profit increase as customers are mailed in order of their propensity for being classified as a donor. The true DONR status is shown to demonstrate the model's accuracy, as increasing profit should result from correctly classified donors where the cost of mailing begins to outweigh the profit once non-donors become the primary recipients.



The maximum of the cumulative profit curve, where cumulative profit = $1,920.38 (for the weighted validation set), serves as the optimal location to set a cutoff for our classification model.

- Classification = 1 when Prob[DONR == 1] >= 0.58144
- Classification = 0 when Prob[DONR ==1] < 0.58144

Prediction Models

Based on the analysis of the non-profit data, four prediction models were created to predict the actual donation amount were created. We present a linear regression model, a decision tree model, a boosted tree model, and a bootstrap forest model. All models were trained with 67% of the data and evaluated with the remaining 33%.

Variable Selection in Prediction Models

All models indicated that the most significant variable contributing to them were Kids, Home Ownership, and Wealth. For model development, these variables were the starting point when selecting variables. Then, forward selection was applied to add variables.

The final selection of variable involved substitutions for variables using those which would contribute the same or similar information to a model. Each model was evaluated to see which were the specific variables which most contributed to its performance.

Evaluation of Prediction Models

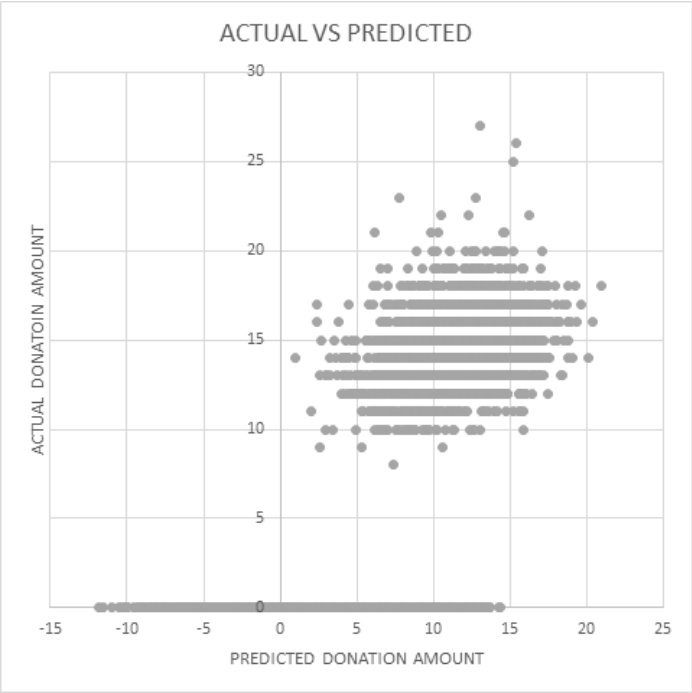
The models were evaluated on two criteria: 1) overall prediction power (based off the R-Squared value) and error (RASE/RMSE values); and 2) Expected cumulative donation amounts based off business model. Where the R-Squared value represents the proportion of explained variability, higher values represent a better fit. RMSE value represent the error and lower values are indicate better performance.

For overall prediction power and error, we review R-Squared and RMSE values. We look for models with the highest R-Squared and the lowest RMSE values possible. This information is directly calculated by the software.

Linear Model

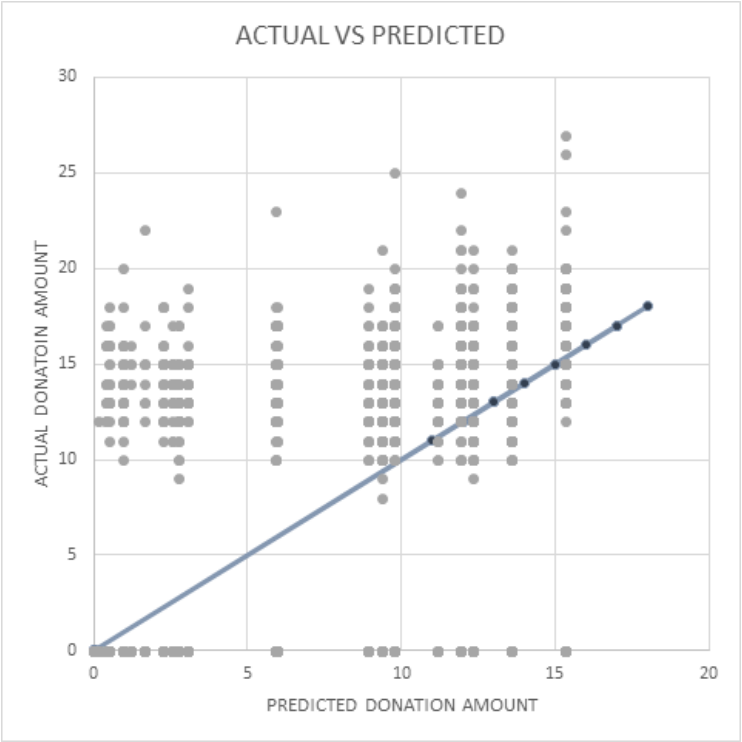
Based on the apparent relationships identified when analyzing the data, a linear regression model was created to predict donation amount. The model uses the following selection of nine variables: Number of Kids; Household Income; Homeowner Status; Tier Region; Wealth Rating; Lag; House Value; Average Dollar Amount of Gifts to Date; and, Months Since Last Gift.

The model's accuracy has an R-Squared value of 0.5459. A scatter plot showing the predicted log prices and the actual log prices is shown below:



Prediction Decision Tree Model

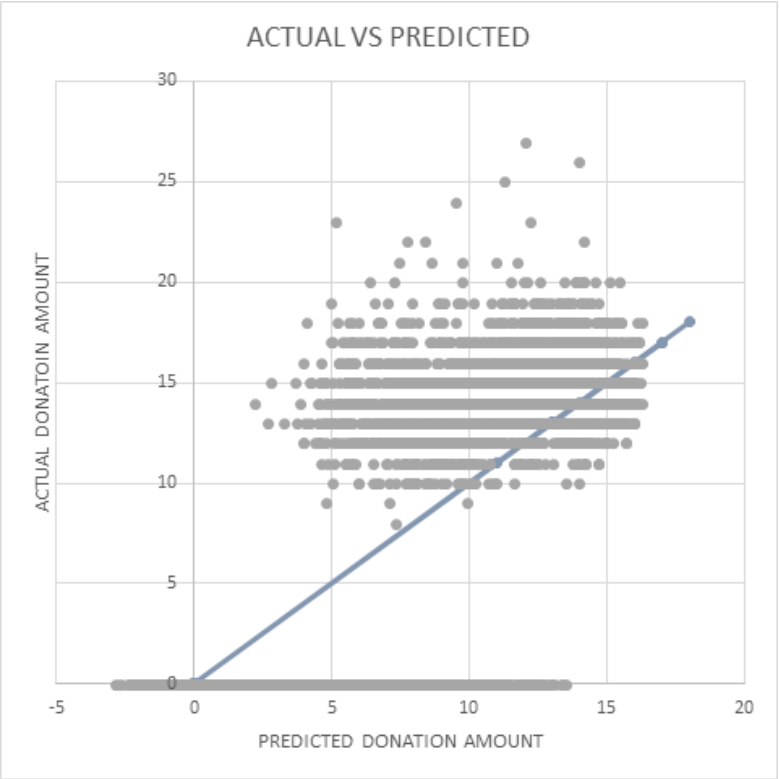
A Decision Tree model was crated with the same target – donation amount. The Decision Tree model uses eight variables: Number of Kids; Household Income; Tier Region; Home Ownership; Wealth Rating; Months Since Last Donation; Average Gift Amount; and Largest Gift Amount. The scatterplot below shows predicted donation amounts and the actual donation amounts below. The decision tree modal had an R-Squared value of 0.589 and a Root Average Square Error of 4.6954. The plot showing predicted values and actual values is shown next:



Prediction Boosted Tree Model

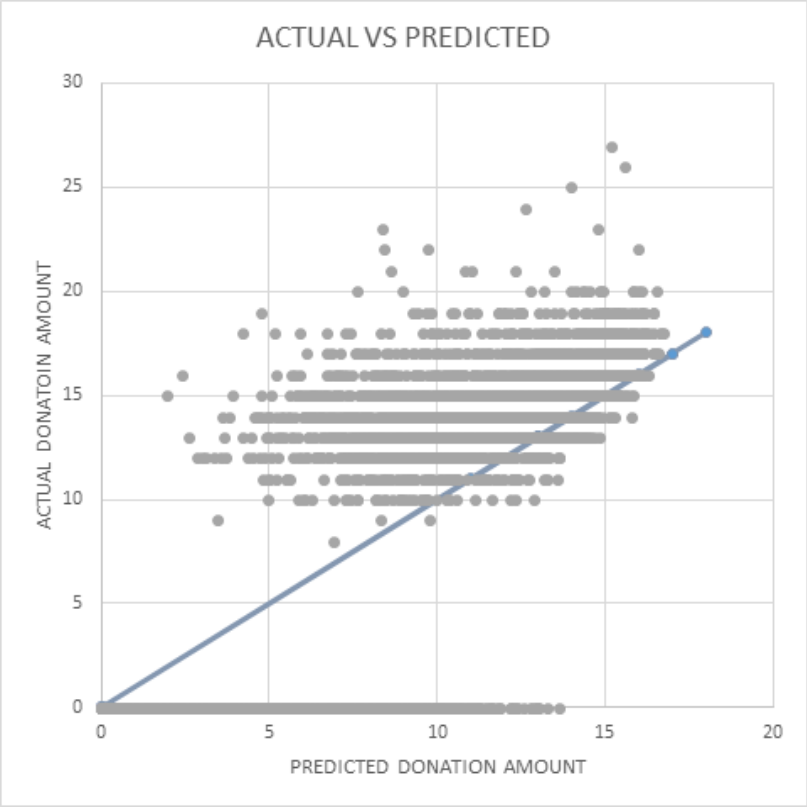
Model was built with ten variables: Number of Kids; Household Income; Tier Region; House Owner Status; Wealth Rating; Median Income; Months Since Last Gift; Lag; Number of Lifetime Promotions; and Largest Gift to Date.

The model has an R-Square of 0.618 and a RASE of 4.5299. The plot below shows the actual and predicted donation amounts.



Prediction Bootstrap Forest Model

The bootstrap forest model was created using 12 variables. The selected variables are: Number of Kids; Tier Region; House Owner Status; Household Income; Wealth Rating; Months Since Last Gift; Median Income; Number of Lifetime Promotions; Average Gift Amount; Lag; House Value; and Sex. The model has an R-Square value of 0.680, and a RASE value of 4.1462.



Evaluation and Model Comparison for Predictive Models

The table below shows the models side-by-side comparison in terms of performance and business application. The Bootstrap Forest model outperforms the other three in all three metrics reviewed. The best performing of the four models is the Bootstrap Forest model.

Prediction Model	RSquare	RASE	AAE
Decision Tree	**0.5893**	**4.6954**	**3.0387**
Boosted Tree	**0.6177**	**4.5299**	**3.5638**
Bootstrap Forest	**0.6797**	**4.1462**	**2.9873**
Linear Regression	**0.5459**	**4.9369**	**4.0443**

Enhanced Prediction Bootstrap Forest Model

Based on the performance statistics and the model application applying costs for reaching out to donors, it is our recommendation that the Bootstrap Forest Model be used. Accordingly, the model was further developed on all available data. The performance on the enhanced Bootstrap Forest Model is shown in the table below.

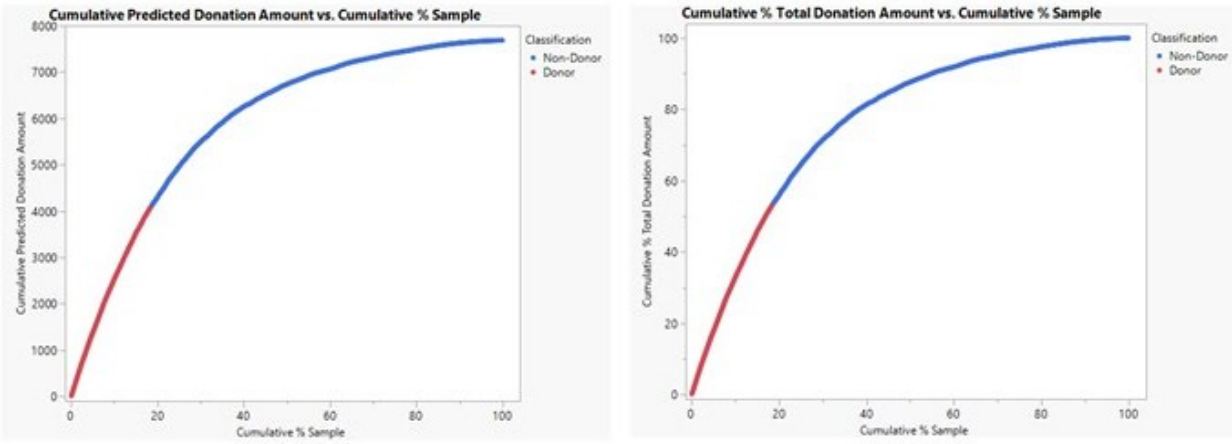
RSquare	RASE	N
0.807	**3.250817**	**6002**

The R-Square value increases to 0.807 (explaining 80.7% of the variation in the data) and the error decreases to 3.2508. It is our recommendation that this model be used to predict donation amounts.

Deployment to Test Partition

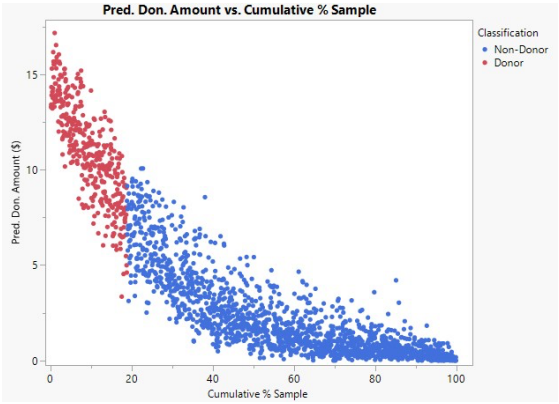
Based on the validation classification lift performance (with weight counterbalance), at a target promotion of ~50% we can identify 2 times as many potential donors. Additionally, at the top ~5%, 9 times as many potential donors are identified (*see boosted tree lift performance*). Combined with the donation value predictive model, if the test partition is representative of the population, we can demonstrate the models' performance on the population.

To capture the maximum expected net profit from direct mail as requested, the propensity for being classified as a donor was increased from 0.50 to 0.58144. In the test sample, this classifies 18.73% of cases as donors, who accounts for 53.67% of potential donations ($).



Donors contribute 53.67% of the predicted donations in only 18.73% of the sample.

We can also see strong alignment between the classification and value prediction by reviewing the donation contributions ($) across the sample. The plotted metrics also illustrate the difference in contribution amounts between segments. If the test partition is to be treated as a population sample, we can expect that individual classified donors will contribute 5 times more on average than classified non-donors.



			Pred. Don. Amount	Pred. Don. Amount
Classification	**N**	**% of Total**	**Sum**	**Mean**
Non-Donor	1631	81.27%	3560.1201233	2.1827836439
Donor	376	18.73%	4125.1166346	10.971054879

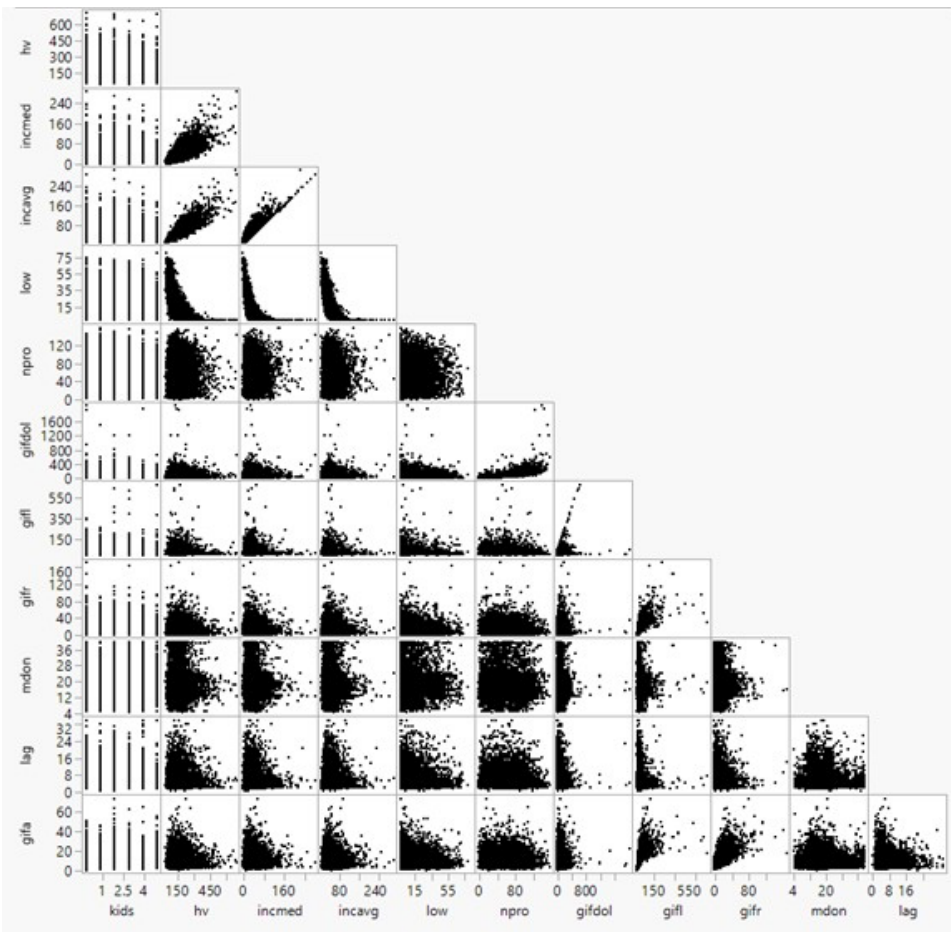
The average classified donor amount is 5 times the average classified non-donor amount.

Thank you for considering our modeling to support your organization. In conclusion, our recommended models not only demonstrate goodness of fit and prediction accuracy among validation data, but when applied to the test partition (described by the organization as more closely aligned with the population), shows strong signs of usefulness. Based on the validation data, by following the classification model we have presented you will accomplish the business goal of maximizing expected net profit on direct mail marketing.

APPENDIX

MULTIVARIATE RELATIONSHIPS AND CORRELATIONS

In exploring the associations between the variable, we're looking for indication of strong correlations via the correlation matrix (shown below). High correlations would indicate an explanation of the same phenomenon in the data. The use of these variables in the same model would cause redundancy at best; and diminish performance and application to future data at worst.



The table below shows the actual values of correlation between the numerical variables in the data set. The more significant values are highlighted via the background color in the cells. For

our models, the options are to either not use highly correlated variables together in the same model, or, if appliable, use a calculated Principal Components in their place.

kids	1.0000											
hv	0.0001	1.0000										
incmed	-0.0133	0.7262	1.0000									
incavg	-0.0075	0.8401	0.8701	1.0000								
low	0.0158	-0.6309	-0.6552	-0.6379	1.0000							
npro	-0.0377	-0.0033	`	0.0154	-0.0206	1.0000						
gifdol	-0.0246	0.0097	0.0315	0.0236	-0.0227	0.7059	1.0000					
gifl	0.0039	0.0051	0.0123	0.0085	-0.0073	0.0024	0.1683	1.0000				
gifr	-0.0024	0.0072	0.0033	0.0075	-0.0108	0.0002	0.0786	0.6973	1.0000			
mdon	0.0365	-0.0195	-0.0226	-0.0278	0.0231	-0.0059	-0.0158	-0.0132	-0.0088	1.0000		
lag	0.0232	0.0134	-0.0041	0.0033	0.0039	0.0088	0.0116	0.0049	0.0047	-0.0062	1.0000	
gifa	0.0138	0.0072	0.0134	0.0102	-0.0116	-0.0056	0.0565	0.6247	0.7107	-0.0109	0.0050	1.0000
	kids	hv	incmed	incavg	low	npro	gifdol	gifl	gifr	mdon	lag	gifa

GENERAL MODEL VARIABLE SELECTION

Initially, models were created through the built-in algorithm in JMP starting with most variables. Then, a backward elimination was applied judging model performance based on the R-Squared value. All models indicated that the most significant variable contributing to them were Kids, Home Ownership, and Wealth. In subsequent evaluations, these variables were the starting point when selecting variables. Then, forward selection was applied to add variables.

The final selection of variable involved substitutions for variables using those which would contribute the same or similar information to a model. Each model was evaluated to see which were the specific variables which most contributed to its performance.

Three types of variable substitution were considered:

- Highly Correlated Variables. Specially in the linear regression model. i.e. the linear regression model could use median income, average income or Low Income percentage as these variables were highly correlated.
- Transformation and Calculations of Variables. Examples include using House Value or the log transformation of House Value given its skewed distribution curve.

- Categorical and Numerical Exchangeable Variables. These were variables which in reality were the same, simply structured around observations. e.g. When considering Tier Region we had the option to select the TIER_REGION variable, or the dummy variables TER1, TER2, TER3, and TER4.

LINEAR REGRESSION MODEL

The linear regression model presented was built using the process above. Below is the model's effect summary describing how much each variable contributed to its performance.

Source	LogWorth	PValue
kids	379.599	0.00000
inc	132.488	0.00000
ownd	77.898	0.00000
ter2	66.619	0.00000
wlth	60.610	0.00000
npro	23.192	0.00000
lag	19.083	0.00000
Log[hv]	17.603	0.00000
ter1	14.640	0.00000
gifa	14.375	0.00000
mdon	5.709	0.00000
ter4	1.986	0.01034
ter3	0.265	0.54276

Of interest is the contribution of Log(HV) rather than the house value itself. Additionally, we can see that Tier Region contributes most to the model with regions 1 and 2. If we're interested in a reduction of complexity at the cost of a slight drop in performance, we could eliminate Ter3 and Ter4 variables. This would mean that for future models the variable would represent Region 1, Region 2 and "all others."

Model Performance on training and validation

RSquare	0.577615
RSquare Adj	0.574839

Root Mean Square Error	4.810981
Mean of Response	7.26004
Observations (or Sum Wgts)	3984

Source	RSquare	RASE	Freq
Training Set	0.5776	4.7947	3984
Validation Set	0.5459	4.9369	2018
Test Set	.	.	0

As we can see, the performance on the training set of 0.5776 and the validation set of 0.54459 suggest the data is not overfit.

DECISION TREE MODEL

The decision tree model we present is built with 8 variables. The contributions of the variables are presented below:

Term	Number of Splits	SS	Portion
kids	3	219782.527	0.4936
ownd	3	65661.0784	0.1475
inc	5	57298.4799	0.1287
tier_region	3	47072.215	0.1057
wlth	3	25816.6771	0.0580
mdon	2	14588.8975	0.0328
lag	1	8591.07682	0.0193
npro	1	4226.74527	0.0095
gifa	1	2201.23717	0.0049

For this model, there was no need for transformations. Additionally, the tree benefits from simplicity by using only the variable TIER_REGION rather than the data as various individual variable (TER1, TER2, etc.).

The model has a R-Square performance of 0.614

	RSquare	RASE	N	Number of Splits	AICc
Training	0.639	4.4185184	3984	22	27868.7
Validation	0.614	4.533393	2018		

BOOSTED TREE MODEL

Model was built with ten variables. The contributions of each variable are presented in the table below.

Term	Number of Splits	SS	Portion
kids	30	370317.848	0.5232
inc	33	94188.0824	0.1331
tier_region	18	73852.2153	0.1044
ownd	19	65553.1054	0.0926
wlth	18	53518.1426	0.0756
incmed	10	16129.6049	0.0228
mdon	7	12458.8931	0.0176
lag	7	11726.5991	0.0166
npro	6	7784.00003	0.0110
gifa	2	2200.72082	0.0031

The model has an R-Square of 0.602 and a RASE of 4.6210.

	RSquare	RASE	N
Training	0.620	4.5468348	3984
Validation	0.602	4.6201978	2018

BOOTSTRAP FOREST

The bootstrap Forest model was created using 12 variables. The contributions of these variables are shown in the table below. As with others Number of Kids; Tier Region; Houseowner Status; and Wealth Rating appear at the top of the list.

Term	Number of Splits	SS	Portion
kids	388	39497.8404	0.4135
inc	517	9516.46985	0.0996
tier_region	459	8570.80848	0.0897
ownd	111	7935.66138	0.0831
wlth	450	7296.25393	0.0764
mdon	503	4505.13043	0.0472
Log[incmed]	538	4047.60861	0.0424
npro	560	4037.21507	0.0423

Term	Number of Splits	SS	Portion
gifa	586	3261.29287	0.0341
hv	568	3209.80683	0.0336
lag	465	3153.04158	0.0330
sex	302	494.095819	0.0052

	RSquare	RASE	N
Training	0.771	3.5319939	3984
Validation	0.661	4.2663669	2018