**Machine Learning for a Charitable Organization**

**Introduction**

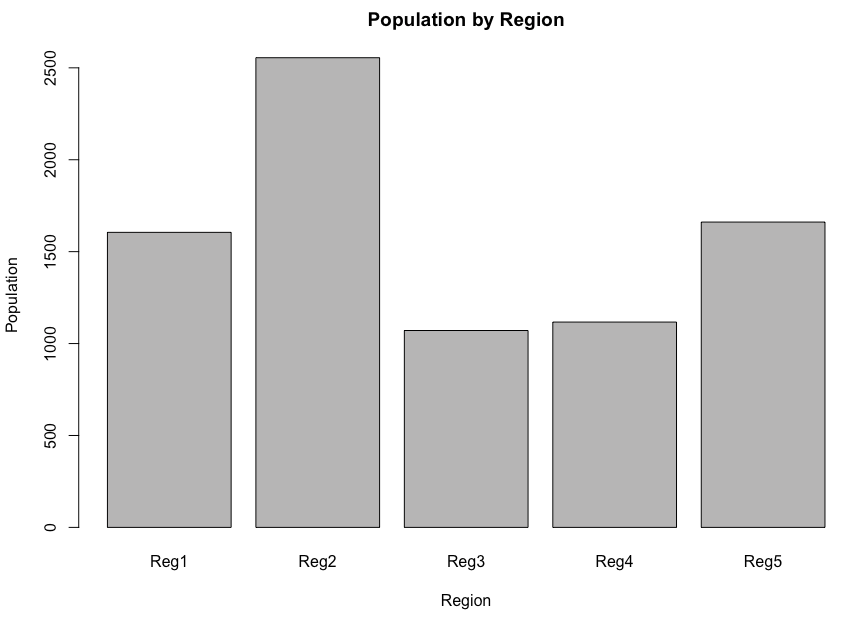
A charitable organization wishes to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to donors. Their current mailing response rate is 10% with an average donation rate of $14.50 per response. There are two different objectives that are to be accomplished. The first objective is to develop a classification model for DONR. The “maximum” profit will be used as the evaluation criteria and the selected model will be used to classify responses. The second objective is to develop a prediction model for the DAMT variable. The “mean prediction error” will be used as the evaluation criteria. The final prediction model will be used to predict DAMT responses in the charity dataset. Models that will be used include logistic regression, GAM, LDA, KNN, ordinary least squares regression, and many more. The analysis will give better insight in the productivity in the direct marketing campaign of the charitable organization and will yield a model to be used in future analysis as well.

**Analysis**

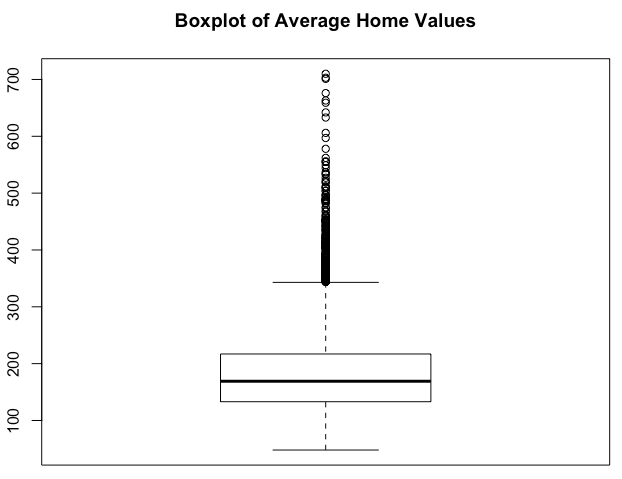
**Exploratory Data Analysis:**

In order to build a prediction model to predict the expected gift amount from donors, exploratory data analysis (EDA) will be performed on the charity data set. This initial analysis will cover the basic description of the data that is in the data set. The data set contains 3,984 records with 24 variables. A breakdown of the population by region was created. Keep in mind that region five is the remainder after regions 1-4 were allocated a population.

|  |  |  |
| --- | --- | --- |
| **Region** | **Population** | **Percentage of Total** |
| Region 1 | 1,605 | 20.03% |
| Region 2 | 2,555 | 31.90% |
| Region 3 | 1,071 | 13.37% |
| Region 4 | 1,117 | 13.94% |
| Region 5 | 1,661 | 20.73% |



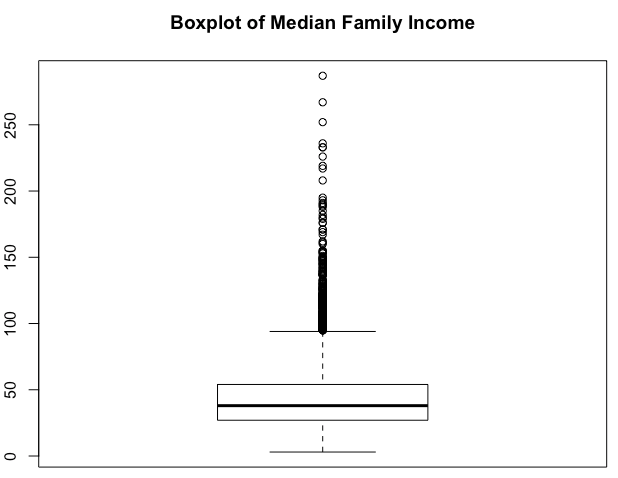
Further analysis found that there are 6,940 homeowners and 1,069 non-homeowners in the data set. For gender, there are 4,848 female donors and 3,161 male donors. Analysis was also performed on the various other variables to find if there were significant outliers or not. All of the variables below were truncated at the maximum usual value.



*AVHV* (Variable Name)

The first variable was average home value. Examining the boxplot of average home values, we can see that there are a lot of outliers at the top part of the box plot. There are 265 records that were higher than the maximum usual value of 343. To accommodate this trend, the variable was truncated at the maximum usual value. The table below shows the summary statistics after truncation.

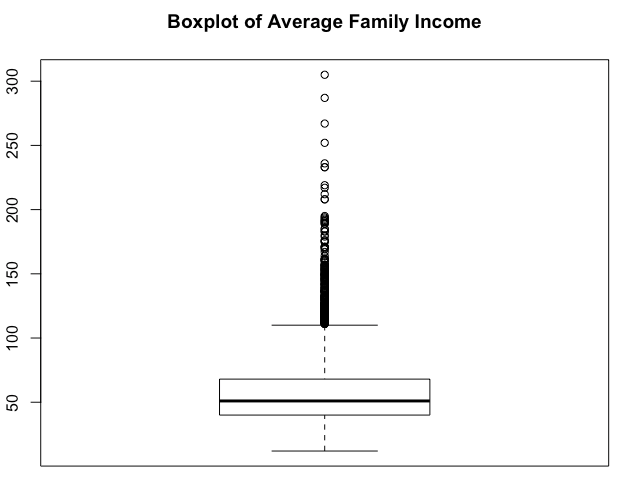
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 48 | 133 | 169 | 182.6 | 217 | 710 |



*INCM*

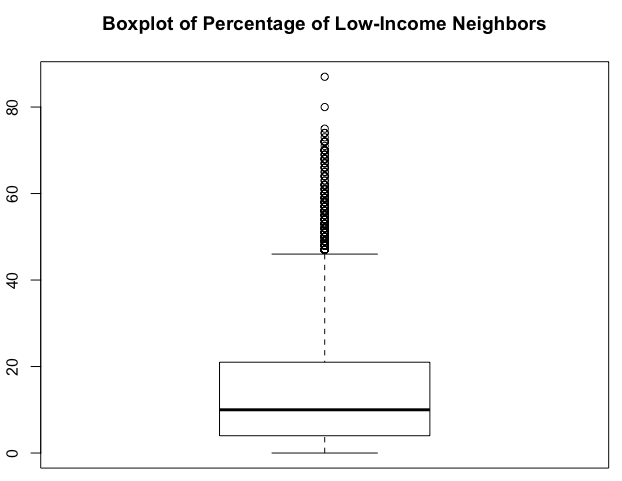
Examining median family income revealed a large number of outliers at the top of the boxplot chart. This value had a maximum usual value at 94.5 and a minimum usual value at -13.5. It was truncated at the maximum usual value.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 3 | 27 | 38 | 42.41 | 54 | 94.50 |

*INCA*

Average family income had outliers only at the top end of the boxplot chart. Average family income was truncated at the maximum usual value of 110. The minimum usual value was calculated to be -2.

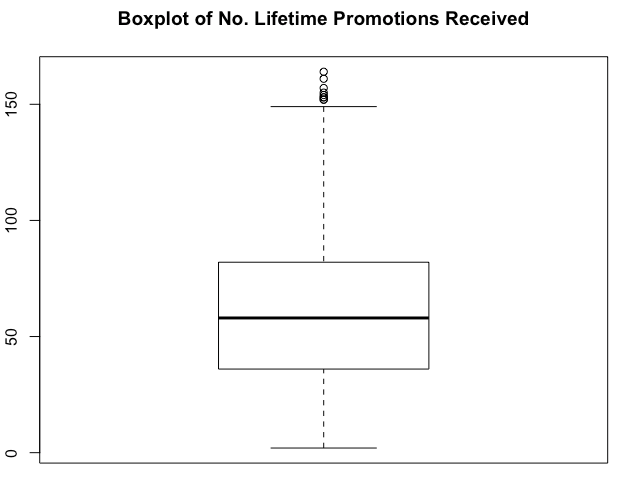
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 3 | 27 | 38 | 42.41 | 54 | 94.50 |



*PLOW*

The percentage of low income neighbors had a lot of variables at the top end of the box plot. This variable was also truncated at its max usual value of 46.50. It has a minimum usual value of -21.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 0 | 4 | 10 | 13.94 | 21 | 46.50 |



*NPRO*

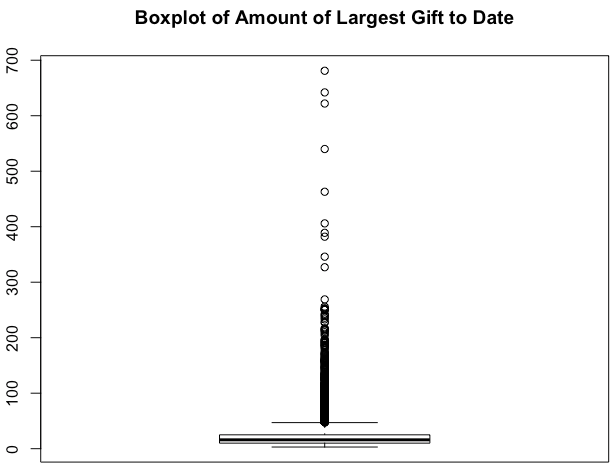
The number of lifetime promotions received had outliers on the top end of the boxplot and was truncated at its maximum usual value of 151.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 2 | 36 | 58 | 60.03 | 82.00 | 151 |

*TGIF*

The number of lifetime gifts to date is extremely skewed in the plot. All outliers fall on the top part of the box plot. The variable was truncated at the top end with the max usual value of 248.

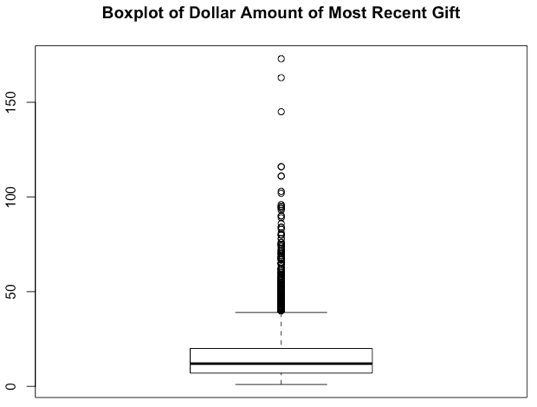
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 23 | 63 | 89 | 107.7 | 137 | 248 |



*LGIF*

The dollar amount of the largest gift to date was also skewed as the number of lifetime gifts to date. The variable was truncated at the maximum usual value of 47.5.

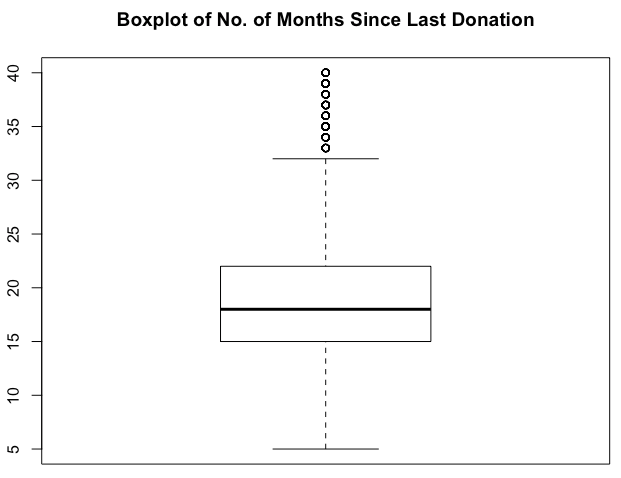
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 23 | 63 | 89 | 107.7 | 137 | 248 |



*RGIF*

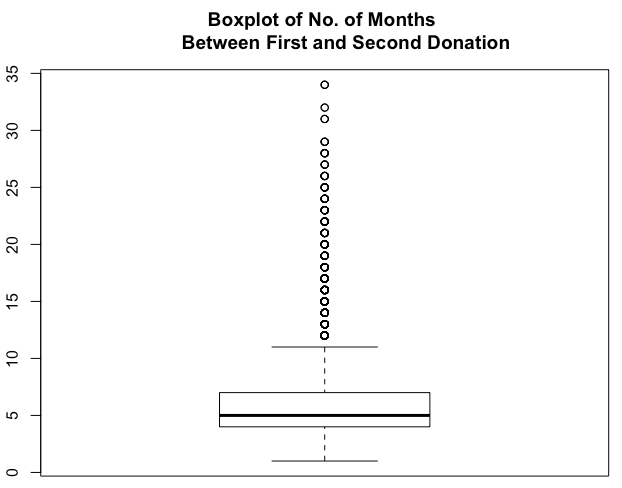
Similar to the lifetime gifts to date variable, the dollar amount of most recent gift box plot shows a large skew. The majority of the outliers fall at the top of the box plot. The variable was truncated at the max usual value of 39.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 1 | 7 | 12 | 14.99 | 20 | 39.50 |

*TDON*

The number of months since last donation was also truncated at the max usual value of 32.5 because of the high number of outliers at the top of the plot.

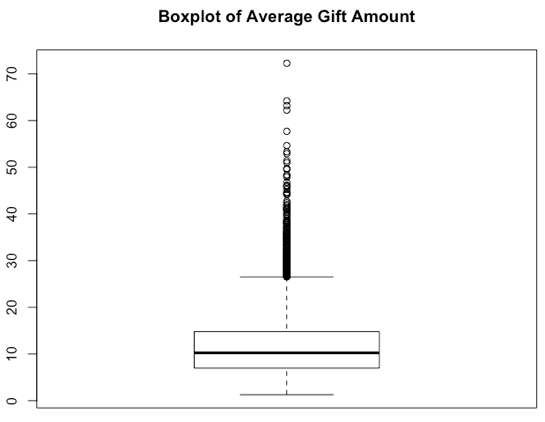
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 5 | 15 | 18 | 18.7 | 22 | 32.50 |



*TLAG*

A boxplot of the number of months between the first and second donation showed outliers at the top of the plot. Subsequently, the variable was truncated at the max usual value of 11.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 1 | 4 | 5 | 5.992 | 7 | 11.5 |

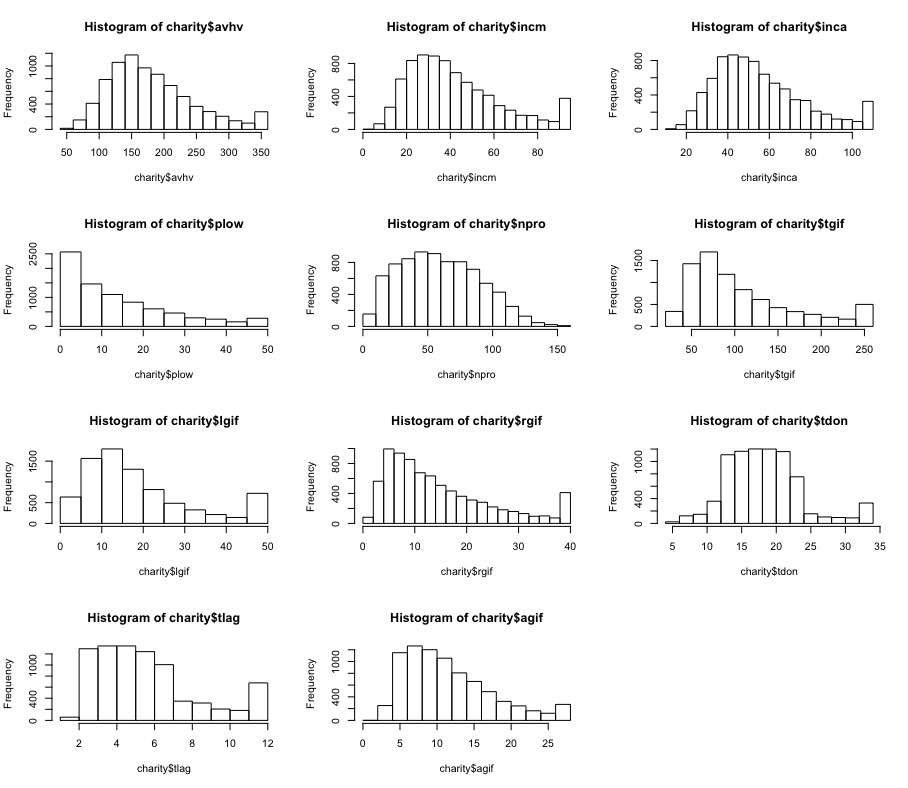


*AGIF*

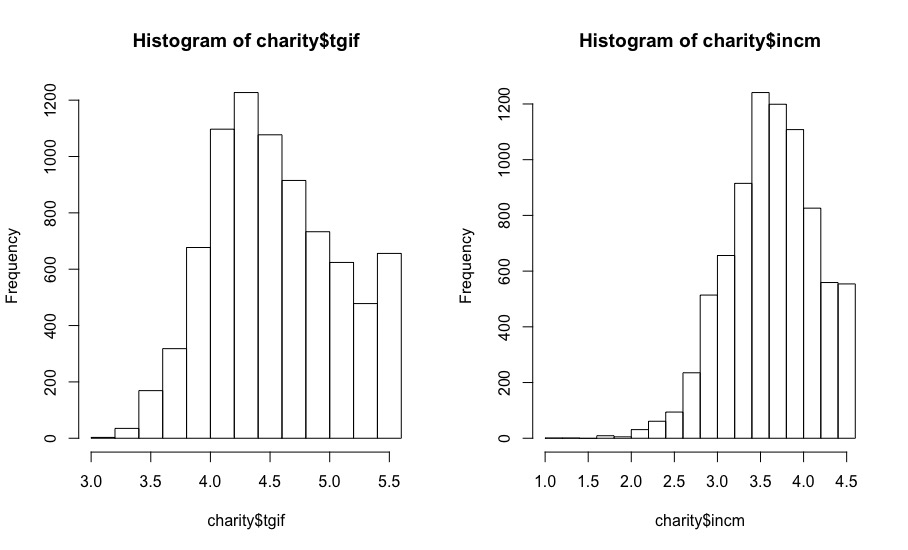
The average gift amount was truncated at the max usual value of 26.55 due to so many outliers existing at the top of the plot.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Maximum** |
| 1 | 7 | 12 | 14.99 | 20 | 39.50 |

After being truncated at the max usual value, all variables were put onto histograms that you can observe below. All of the variables show that they have non-normal data and are skewed to the right.



Due to the non-normal data, a couple of variables needed to be transformed to confirm to a normal distribution. Both the dollar amount of lifetime gifts to date (TGIF) and the median family income in potential donor’s neighborhood variables (INCM) had a log transformation performed.



Correlations between variables that serve as direct/indirect measurements of potential donor’s wealth were calculated and some significant correlations could be noticed. The first was between the average home value in potential donor’s neighborhood (AVHV) and INCM. This correlation was 0.71 between the two variables. The second most significant was the positive correlation of 0.84 between AVHV and the average family income in potential donor’s neighborhood (INCA). The next significant correlation was between AVHV and the percent categorized as low income in potential donor’s neighborhood (PLOW) at -0.69. Other significant variables include a 0.80 correlation between INCM and INCA. INCM also highly correlates with PLOW at -0.83. INCA has a high correlation with PLOW at -0.71.

Correlations among variables that describe potential donor’s interactions with charity were calculated. There are some correlations that should be noted. The first is between TGIF and the dollar amount of the largest gift to date (LGIF) at 0.08. TGIF also has a low correlation with the dollar amount of the most recent gift (RGIF) at 0.082, and with the average dollar amount of gifts to date (AGIF) at 0.067. Another correlation to consider is between the lifetime number of promotions received to date (NPRO) and TGIF at 0.88. LGIF correlates with RGIF and AGIF at 0.83 and 0.81 respectively. RGIF correlates with AGIF at 0.75.

**Pre-Model Process:**

Before analysis is to begin, the charity data was split into training, validation and testing sets. ID was omitted from those sets because it will not be needed in the model development phase. The training data set has 3,984 rows and 24 columns. The validation set as 2,018 observations. The test set contains 2,007 records and the validation set currently has 2,018. The training, validation, and test training sets were all standardized.

Exploratory data analysis was performed on the training dataset and frequency tables were created between DONR and the categorical variables. Even though the number of children is not a categorical variable, a table between the number of children and DONR was created. DONR by region was first to be created. Region five is the remainder with 14% of donors.

DONR by Region 1: 23% of donors and 20% of population

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1,627 | 1,541 |
| 1 | 362 | 454 |

DONR by Region 2: 45% of donors and 34% of population

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1,553 | 1,092 |
| 1 | 436 | 903 |

DONR by Region 3: 9% of donors and 12% of population

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1,675 | 1,817 |
| 1 | 314 | 178 |

DONR by Region 4: 9% of donors and 13% of population

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1,635 | 1,812 |
| 1 | 354 | 183 |

DONR by homeownership was also put into frequency tables. Of the 1,995 donors, approximately 1,947 are homeowners which is 97% of the data. Of the 3,519 homeowners in the set, 1,947 or approximately 55% are donors.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 417 | 48 |
| 1 | 1,572 | 1,947 |

DONR by the number of children yielded that 60% of donors have no children.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 194 | 1,201 |
| 1 | 203 | 201 |
| 2 | 770 | 381 |
| 3 | 494 | 162 |
| 4 | 237 | 44 |
| 5 | 91 | 6 |

DONR by the household income category yielded the following results.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 1 | 183 | 29 |
| 2 | 322 | 145 |
| 3 | 202 | 205 |
| 4 | 609 | 1,226 |
| 5 | 309 | 274 |
| 6 | 190 | 70 |
| 7 | 174 | 46 |

DONR by gender showed that approximately 40% of donors are male and 60% are female.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 769 | 805 |
| 1 | 1,220 | 1,190 |

Donor by wealth rating can be shown in the table below.

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 92 | 5 |
| 1 | 83 | 5 |
| 2 | 71 | 26 |
| 3 | 110 | 31 |
| 4 | 138 | 70 |
| 5 | 113 | 79 |
| 6 | 121 | 150 |
| 7 | 97 | 141 |
| 8 | 676 | 881 |
| 9 | 488 | 607 |

The correlation between the donor and the number of months since last donation shows that they are negatively correlated at -0.15

**Results**

**Model Building**

A classification model was attempted to be developed in order to predict the DONR variable using variables except for ID and DAMT. The first to be fit was a logistic regression model. There were four logistic regression models fit for this particular data set. The best model had a low error of 0.1253 and an AIC value of 1216.2.

The second model that was generated was a logistic regression with GAM. After the model was ran with all variables and had an AIC of 2789.143, we included some transformations and added a natural spline for TLAG. Significant variables were the only variables kept in the final GAM model. Interaction terms for correlated variables were added to the model. This improved the AIC to 2561.171. In both cases, the models ran better against the validation set than they did against the training set. The second GAM model had a low AIC value of 1209.219 against the validation set.

We built a model using LDA even though LDA models are not usually created with qualitative variables. The first model was ran using all of the variables in the model before using a reduced variable set for the second LDA model. The first model against the training set had an error of 0.1566. The error rate for the first LDA model was higher on the validation set than it was on the training set at 0.1699. The second model had a higher error rate on the validation set compared to the training set as well. The error for the second model on the validation set was 0.1372.

A QDA model was created to fit against the training and validation set. The first model for QDA included all variables and had an error of 0.1463. The second QDA model did not include all variables and had an error of 0.2073. The first model had a lower error on the validation set compared to the training set. However, the second model had a lower error rate on the training set rather than the validation set at 0.2244.

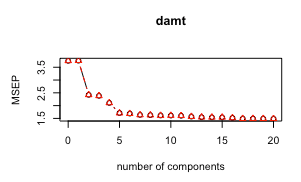
The next model created was the K-Nearest Neighbors model. The first model ran was able to correctly classify 77.8% of the data. To make the model better, the number of nearest neighbors was increased. The accuracy of the model was improved when more of the nearest neighbors were added. The new prediction accuracy was 79.5%. From there, the number of nearest neighbors was increased to five and resulted in 80.3% being correctly classified. Other models added to the mix were decision trees, bagging, random forest, and support vector classifier. One thing to note is that the dataset for the tree classifier was also used for the SVC model.

The models had their prospective total donors and profits calculated and compared to each other. The table below shows the expected number of donors and the amount of profit expected to make. The boosted model showed to have the highest predicted profit with $11,840.00.

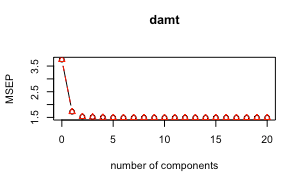
|  |  |  |
| --- | --- | --- |
| **Model** | **Donors** | **Profit** |
| Logistic | 1374 | $11,650.50 |
| GAM | 1386 | $11,655.50 |
| LDA | 1323 | $11,593.00 |
| QDA | 1330 | $11,506.50 |
| KNN | 1224 | $10,790.50 |
| Tree | 1165 | $11,140.50 |
| Bag | 1038 | $11,075.50 |
| Boost | 1243 | $11,840.00 |
| SVM | 1057 | $10,414.00 |

The next model needing to be created was to predict the DAMT variable using any of the variables except for ID and DONR. All candidate models were fit in this portion of model building. The first linear regression model was created with all variables included and then it was ran with just a subset of variables with an improved error rate. The criteria for the fit was to find the lowest standard error. The model with the lowest error was the linear regression model ran against the validation set with 0.1584. The Best Subset Model with K-Fold Cross-Validation was ran with a low standard error of 0.163.

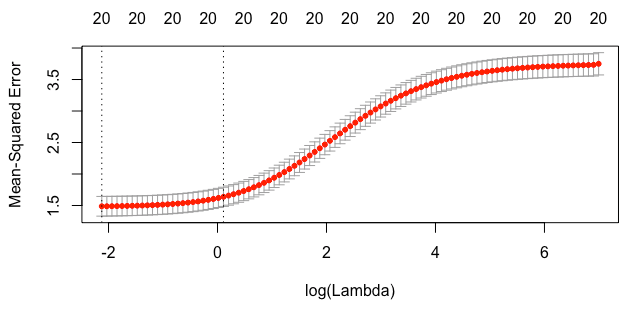
The next model to be produced was the Principal Components model. The model showed that there is an elbow in the graph at 5 and that the lowest point is around 20. These observations were noted and then the model was tested again using 20 components.



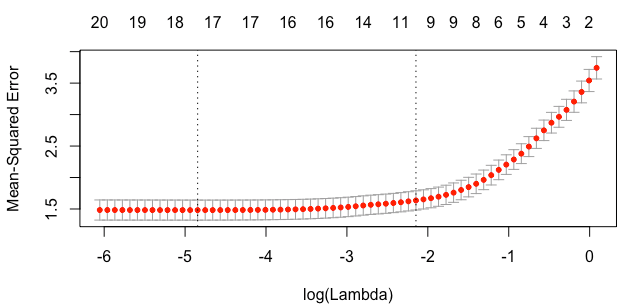
A Partial Least Squares model was created for the Ridge Regression performed on the model. In the initial summary of the model, there was a noticeable elbow in the graph at three with a minimal reduction afterwards. The elbow suggested that three components is sufficient. The standard error was low at 0.1644.



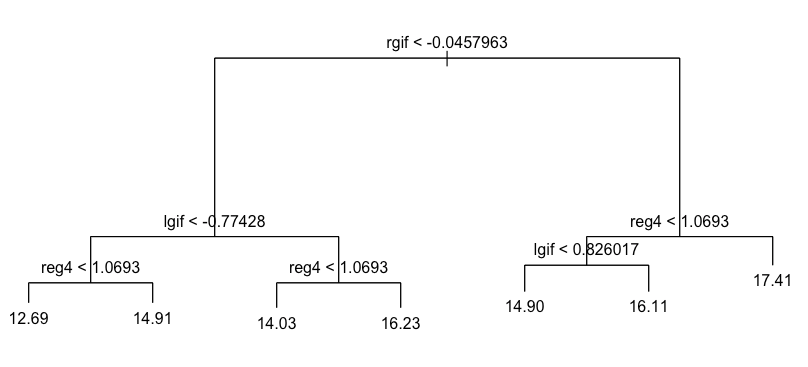
The next prediction model was Ridge Regression. Cross validation was used to select lambda. The model showed a low error of 0.1657. The graph can be seen below.



After Ridge Regression, a Lasso model was fit to the data. The same matrices and lambda grid from the Ridge Regression was used for the Lasso modeling. The standard error for the model was low at 0.1639. The chart can be seen below.



A decision tree was created for the data set as well. After the initial tree was created, cross-validation was performed to prune the tree. Using seven terminal nodes did not improve the MPE.



Other models added were a bagged model with an error of 0.177, a random forest model with an error of 0.1733, and a boosted model with an error of 1.544. Each of those models were considered in final model selection. A table was created to compare the MPE of the models. The boosted model is the preferable model to produce the lowest MPE.

|  |  |  |
| --- | --- | --- |
| **Model** | **MPE Values** | **Std. Error Values** |
| OLS | 1.605 | 0.1584 |
| CV Best Subsets | 1.702 | 0.1634 |
| PCR | 1.709 | 0.1640 |
| PLS | 1.747 | 0.1644 |
| Ridge | 1.725 | 0.1657 |
| Lasso | 1.713 | 0.1639 |
| Tree | 2.241 | 0.1920 |
| Bag | 1.726 | 0.1770 |
| Random Forest | 1.674 | 0.1733 |
| Boost | 1.544 | 0.1673 |

**Conclusion:**

There were many models developed in the research process. The best model for each depended on which measure of error that was trying to be achieved. The first problem asked to develop a classification model for the DONR variable using any of the variables as predictors except for ID and DAMT. The best model to use to maximize profit is the boosted model. This particular model came out to have the lowest error rate and predicting a total of 1,243 donors and a total of $11,840.00 in profit. There were other models that could have been utilized such as the GAM model that had the highest number of donors predicted at 1,386. However, the GAM model did not have as high of a total profit amount predicted as the boosted model.

The second objective was to develop a prediction model for the DAMT variable using any of the variables as predictors except for ID and DONR. The mean prediction error was used as the evaluation criteria for the model. The lowest mean prediction error was found in the boost model at 1.544 and a standard error of 0.1673. The next closest in having the lowest MPE value was the ordinary least squares model at 1.605 and a standard error value of 0.1584.