Final Group Project - MSDS 422-56
Summer 2018 - Dr. Nathan Bastian

Group 3 - Nate Belete, Hema Geddam, Tyson Prince

8/31/2018

Contents

Introduction

This assignment focuses on the efforts of a charitable organization to develop a plan for identifying and soliciting previous and potential new donors. The charity believes that by utilizing information specific to each donor, such as home location, neighborhood demographics, household income and wealth, and previous donation history, they can develop an effective marketing strategy to target individuals most likely to donate, and thereby maximize the effectiveness of their campaign. Our role will be to assess various methods that will help the charity determine how many donors to target and how much revenue and profit to expect.

Exploratory Data Analysis

A fundamental step in any modeling effort is an exploration of the data that will drive the prediction results. This analysis is necessary to ensure that we have an understanding of patterns and relationships that exist in the data, and that we can assess whether any adjustments should be made to the data to account for errors, anomalies, or missing information. The data set used in this assignment consists of a training portion, from which models will be created, a validation portion, against which models will be checked for performance, and a test portion, on which our best models will predict donation activity.

Below is a summary table of the original predictor variables in the data set, which will be used as the basis for predicting donation activity. This simple summary can reveal a few basic characteristics of the data. As a few examples:

- The first 9 variables appear to be categorical, such as ratings or counts, while the remainder appear to be continuous, such as dollar amounts or percentages

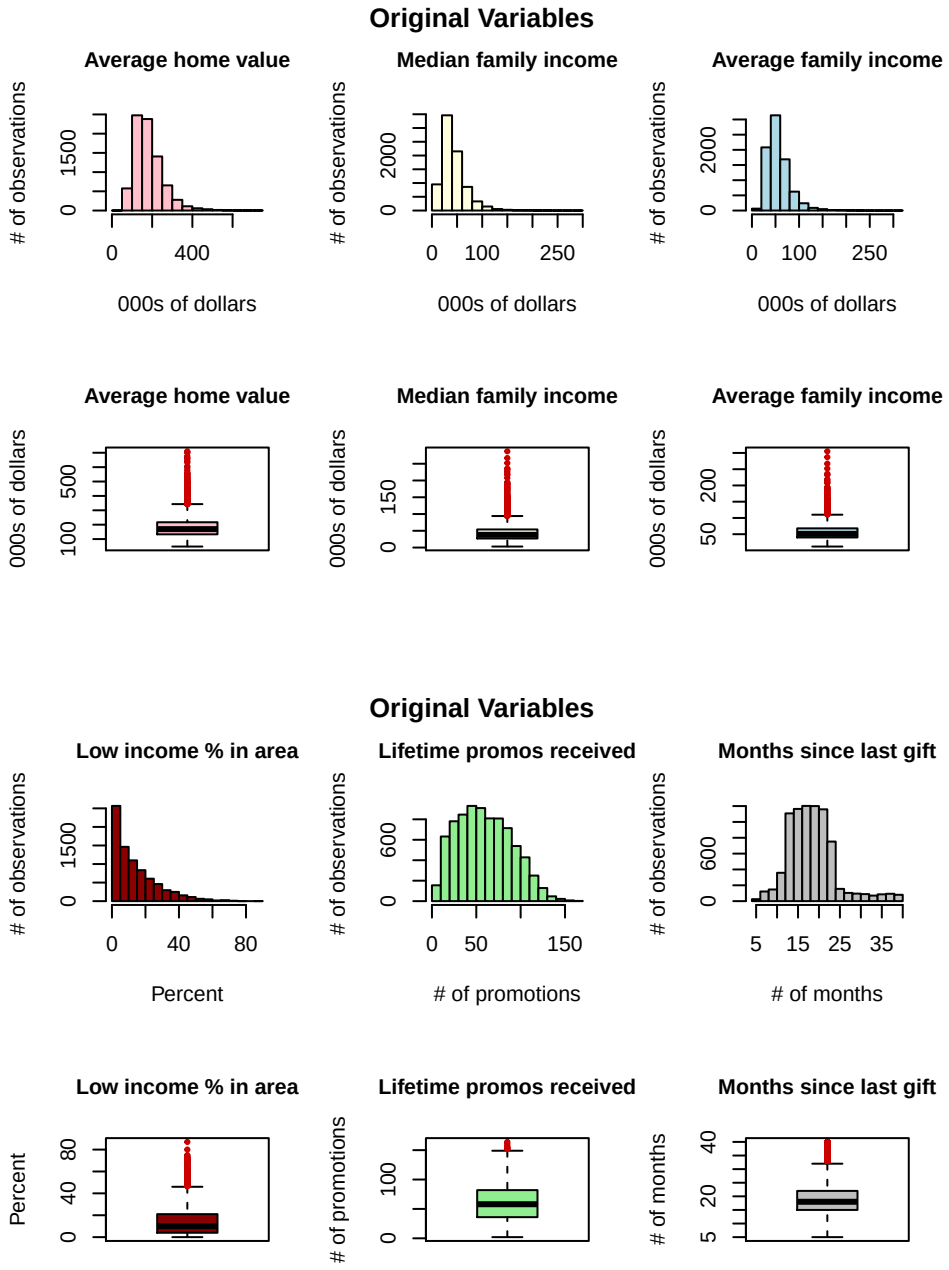
- None of the predictors appears to be missing values

- Outliers may be present, as indicated by large maximum values that are far higher than corresponding 3rd quartile values (as with predictors incm, tgif, and lgif, among others)

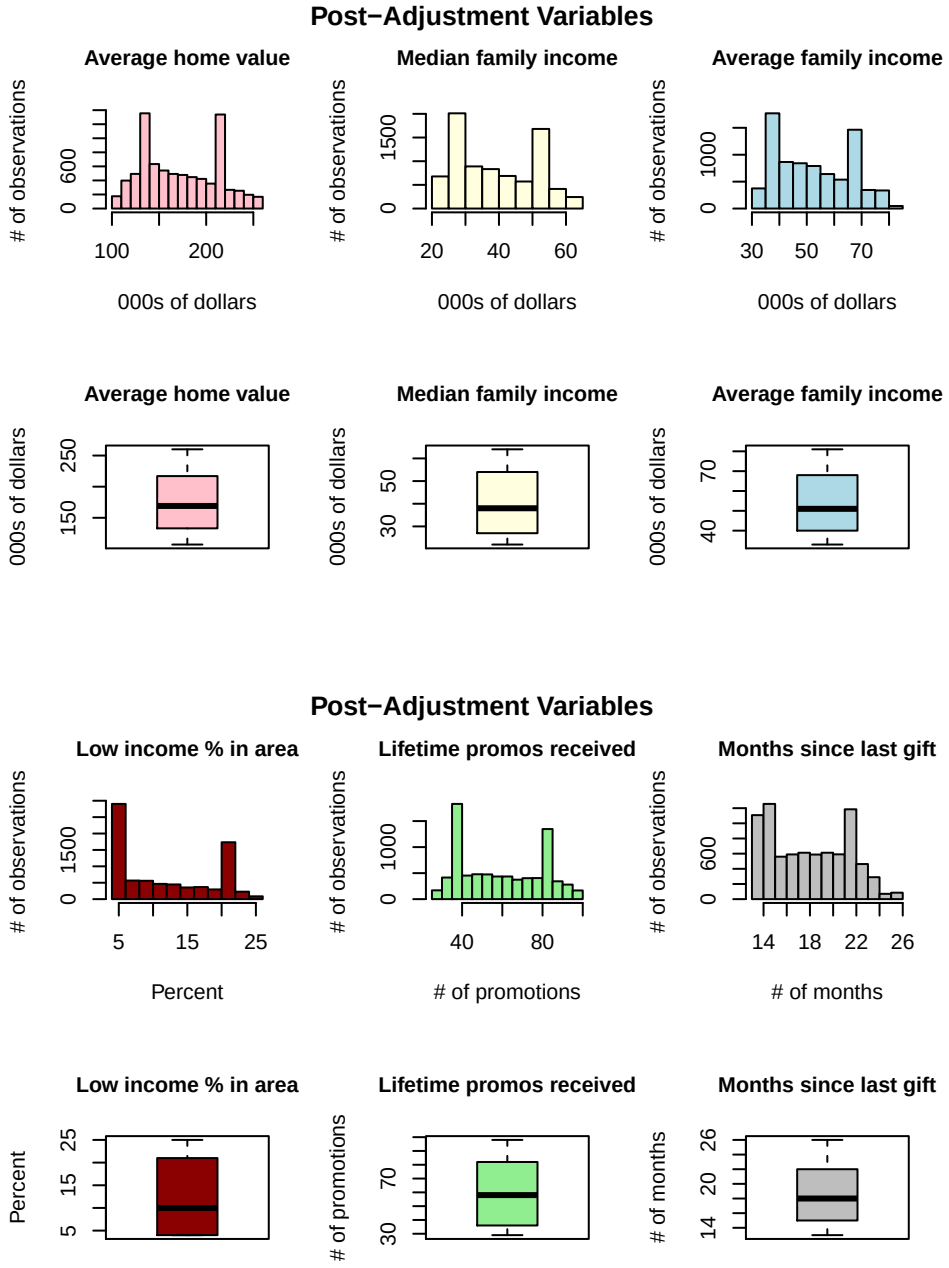
- Some variables appear to be Yes/No "dummy" variables split from an original categorical variable (as with reg1, reg2, reg3, and reg4)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
reg1	0	0	0	0	0	1
reg2	0	0	0	0	1	1
reg3	0	0	0	0	0	1
reg4	0	0	0	0	0	1
home	0	1	1	1	1	1
chld	0	0	2	2	3	5
hinc	1	3	4	4	5	7
genf	0	0	1	1	1	1
wrat	0	6	8	7	9	9
avhv	48	133	169	183	217	710
incm	3	27	38	43	54	287
inca	12	40	51	56	68	305
plow	0	4	10	14	21	87
npro	2	36	58	60	82	164
tgif	23	63	89	113	137	2057
lgif	3	10	16	23	25	681
rgif	1	7	12	16	20	173
tdon	5	15	18	19	22	40
tlag	1	4	5	6	7	34
agif	1	7	10	12	15	72

A graphical survey of some of the predictors confirms the presence of likely outliers. The histograms below reveal that several predictors are right skewed, while corresponding boxplots identify outliers in red at the upper end of each predictor's range. Outliers can be errors or may reflect actual characteristics of a data component. In the case of average home value, median family income, and average family income, most of the observations are within a range that would appear to be otherwise normal, while a small number of observations fall in the extreme upper end. The *low income %* predictor appears to reflect the fact that relatively few people in the donor pool live in areas of high poverty.



Outliers can cause difficulties in constructing models because of their influence on how future observations will be predicted. As just noted, however, they may be true data elements and should not be discarded carelessly. Rather, simple steps can be taken to preserve the observations while lessening their potential negative impact on prediction. In our data set, outlier influence has been lessened by imputing values for observations that fall at the extreme end. Specifcally, any value falling *below 80%* of the 1st quartile value has been imputed to the 1st quartile value, while any value falling *above 120%* of the 3rd quartile value has been imputed to the 3rd quartile value.

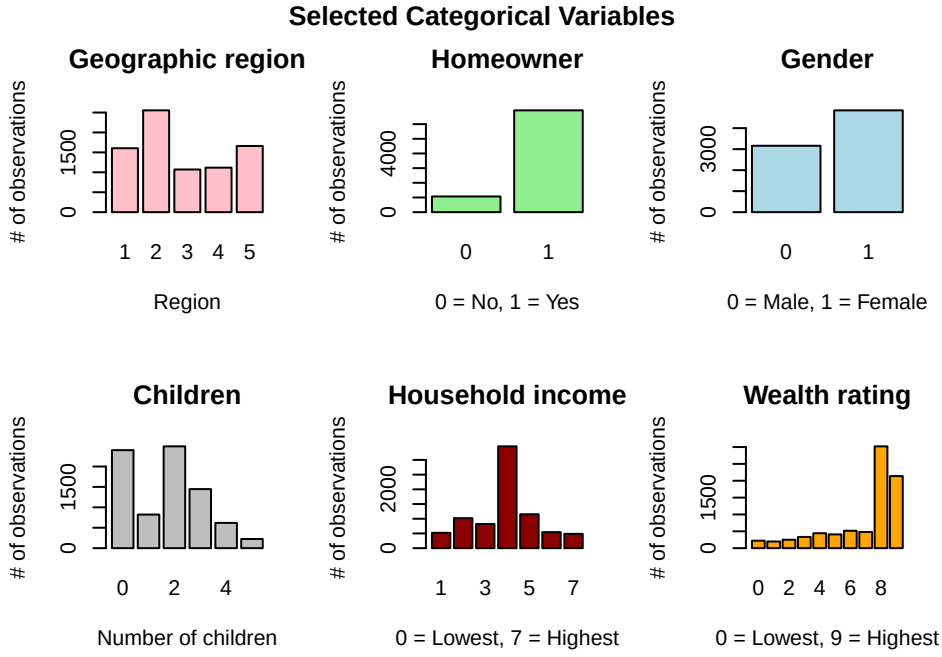


Categorical variable graphs reveal demographic characteristics of the donors in the data set, which the charity may utilize to focus specifically on high-potential groups. For example:

- The geographic region "dummy" variables have been combined (for visual clarity only) and show the highest concentration of donors in region 2

- Most of the donor list are homeowners, which may indicate a higher likelihood of donating

- Household income appears nearly normally distributed, while wealth rating is concentrated at the higher end, which may indicate that donors have means to give non-cash items like stock or real-estate



Data Prepation

To prepare our data set for the modeling process, we can take steps to break out the predictors into individual components, and we can create subsets of the original data set to make the entire process easier. All of this manipulation will occur behind the scenes and can be found in the accompanying R code. In this exercise we will do the following:

- Certain categorical variables will be split into individual "dummy" variables that provide a direct 0/1 (or "no/yes") indicator. As an example, the household income variable (hinc) can be divided into 6 individual indicator variables, each indicating whether the donor falls into that category. This allows the modeling process to consider specific groups of donors where appropriate.

- Certain continuous variables will be "binned," or split into a small number of groups. As an example, the total lifetime gifts variable (tgif), which ranges from **$23** to **$2,057**, can be simplified into a few groupings to make donor identification easier.

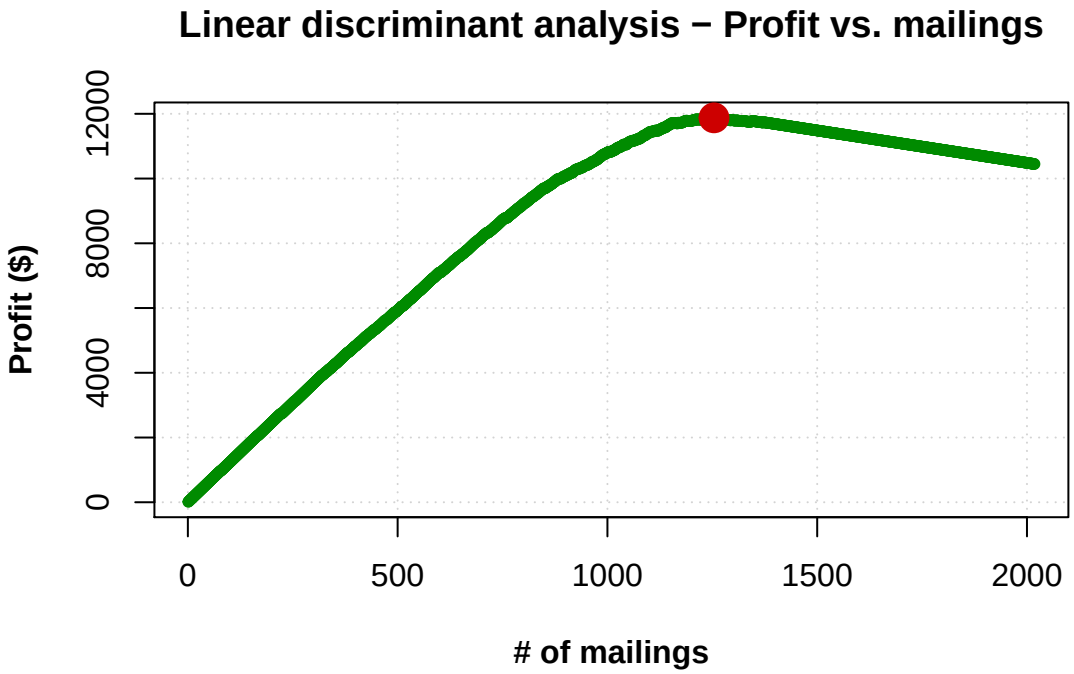
- The original full data set, which combines observations for training, validation, and testing, will be divided into subsets for ease of application. Additionally, the predictor variables will be standardized so that the modeling process is not incorrectly influenced by the scale of some variables over others.

Classification Models

The first set of models that we will evaluate will classify the potential donor list into those likely to donate vs. not likely to donate. Our goal is to find the model with the highest estimated profit, which is a function of the average donation amount of $14.50 and the cost per mailing of $2.00. We will use the validation data set to assess model performance.

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is similar to logistic regression in the sense that it estimates the probabilities that a given observation will fall into the response classes and often provides similar results. An important assumption in LDA is that the classes share a common covariance matrix. In the LDA model below, the largest profit based on validation data is achieved at **1,254** mailings, resulting in profit of **$11,876**.

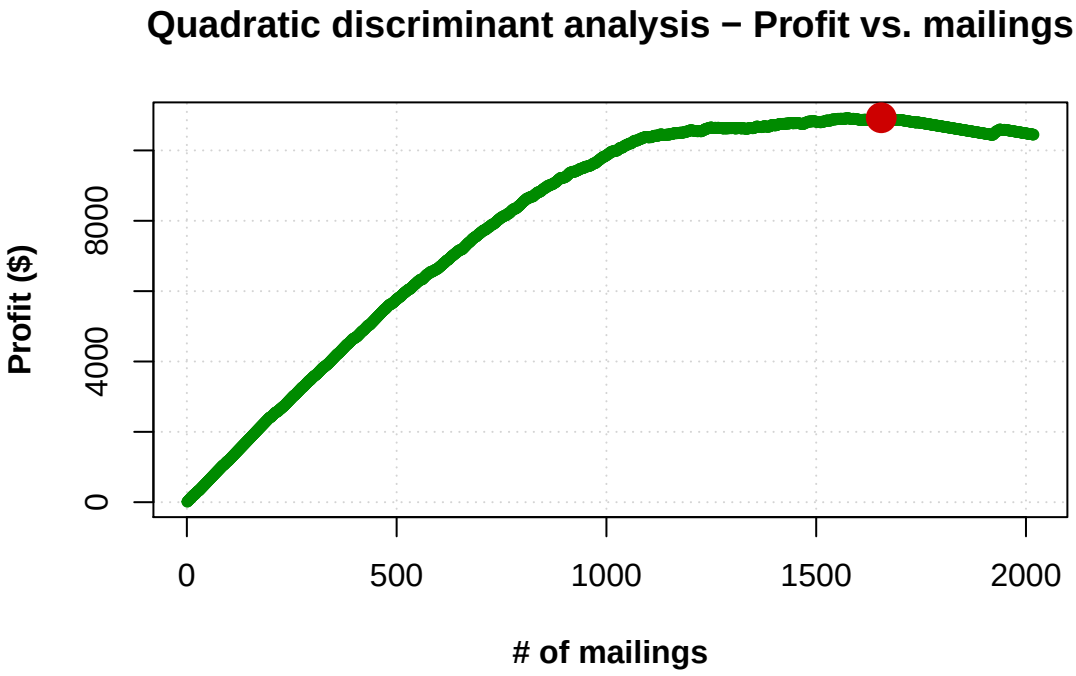


Model	Number of Mailings	Max Profit
Linear Discriminant Analysis	1254	11876

Model	Actual: 0	Actual: 1
Model predict: 0	757	7
Model predict: 1	262	992

Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) is similar to LDA, yet it relaxes the assumption of a common covariance matrix among classes. QDA can outperform LDA when the training data set is very large. In the QDA model below, the largest profit based on validation data is achieved at **1,655** mailings, resulting in profit of **$10,929**.

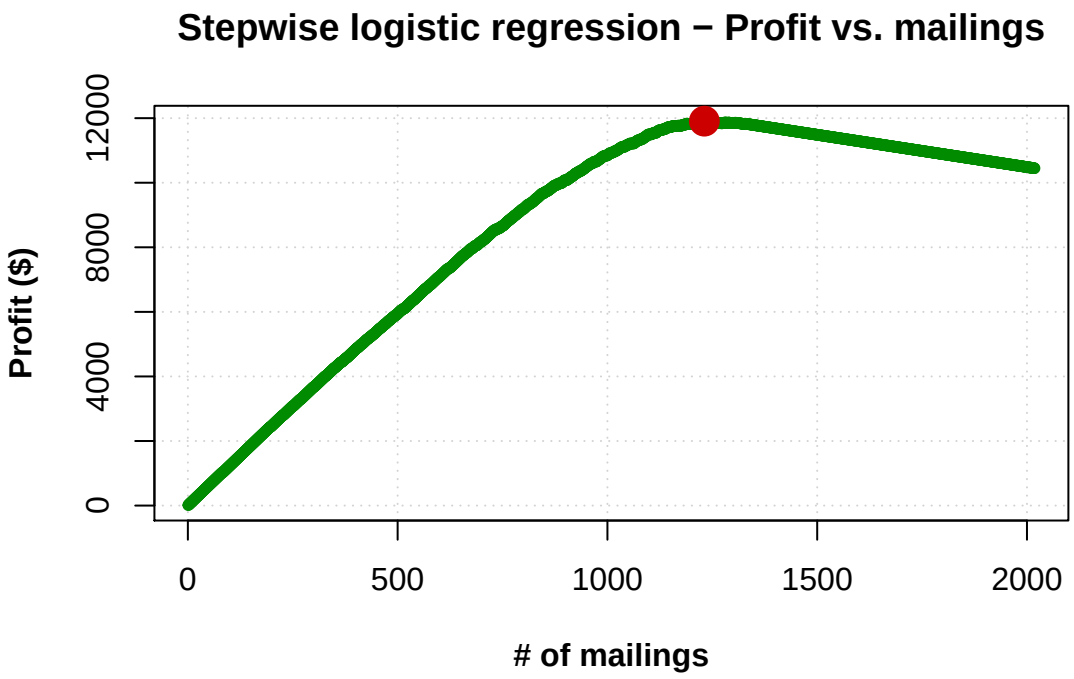


Model	Number of Mailings	Max Profit
Quadratic Discriminant Analysis	1655	10929

Model	Actual: 0	Actual: 1
Model predict: 0	346	17
Model predict: 1	673	982

Stepwise Logistic Regression

Stepwise logistic regression combines the logistic regression method of predicting class probabilities with the stepwise variable selection method. In many cases, stepwise variable selection can outperform a similar model where variables are selected manually. With stepwise variable selection, we start with a current model that includes no explanatory variables, only the y-intercept. We then consider a candidate model that is different than the initial model by the addition and subtraction of one explanatory variable. If the candidate model has an AIC value that is less than our current model, then we accept the candidate model as our current model. We then repeat the same steps until all explanatory variables are tested such that the current model has the lowest AIC value relative to all candidate models. In the stepwise logistic regression model below, the largest profit based on validation data is achieved at **1,231** mailings, resulting in profit of **$11,908**.

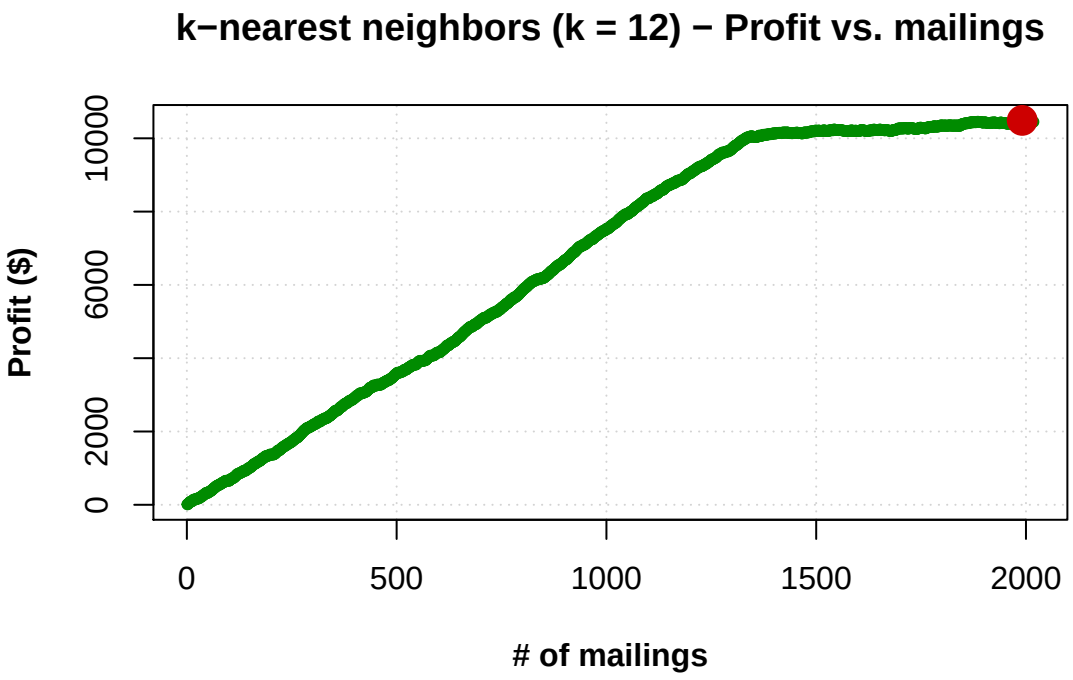


Model	Number of Mailings	Max Profit
Stepwise Logistic Regression	1231	11907.5

Model	Actual: 0	Actual: 1
Model predict: 0	779	8
Model predict: 1	240	991

k-Nearest Neighbors Model

The k-nearest neighbors (KNN) classification method attempts to classify an observation based on the already-classified observations around it. This method has flexibility in determining the class to be assigned by varying the number of neighboring observations to include. For example, at $k = 1$, only the single nearest observation will be used to classify the new observation. At $k = 15$, on the other hand, the nearest 15 observations will be used, with the class having the majority of observations being assigned to the new observation. In the KNN example below, the largest profit based on validation data is achieved at $k = 12$, corresponding to **1,991** mailings, resulting in profit of **$10,489**.



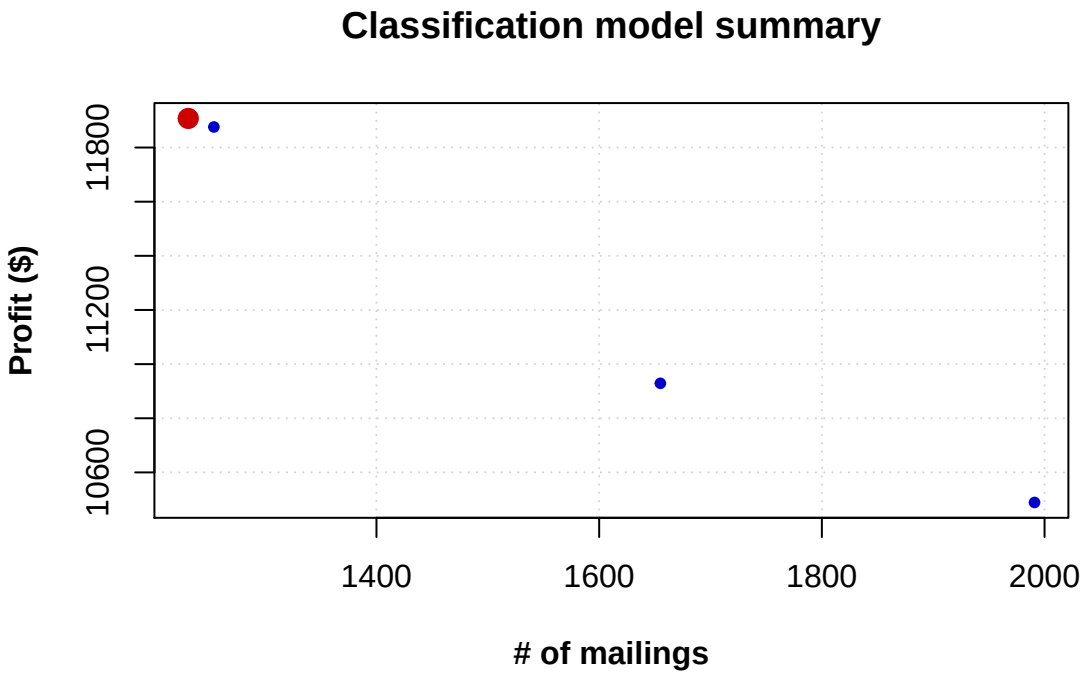
Model	Number of Mailings	Max Profit
k-Nearest Neighbhors	1991	10489

Model	Actual: 0	Actual: 1
Model predict: 0	554	120
Model predict: 1	465	879

Classification Model Summary

The classification models assessed have varying results for the number of mailings and the associated profit. As the summary table below shows, the stepwise logistic regression model appears to perform best on the validation data. The linear discriminant analysis model follows closely, as is expected given its similarity to logistic regression. The quadratic discriminant analysis and k-nearest neighbors models appear to be poorer fits to the data.

Model	Number of Mailings	Max Profit
Linear Discriminant Analysis	1254	11876
Quadratic Discriminant Analysis	1655	10929
Stepwise Logistic Regression	1231	11907.5
k-Nearest Neighbors	1991	10489



Regression Modeling

The second set of models that we will evaluate will predict the amount that each potential donor will give. Our goal is to find the model with the lowest mean squared error of prediction compared to the validation data set.

Multiple Linear Regression Model

Least-squares regression is one of the most popular modeling methods because of its ease of interpretation and broad applicability. In fact, least squares regression can often be a wise starting point, from which more complex methods can be explored if necessary. In our example of a multiple linear regression model containing all predictors, the mean squared error of prediction is **1.4314** and the standard error is **0.1514**.

Model	MSE	Standard Error
Multiple Linear Regression	1.4314	0.1514

Backward Selection Linear Regression Model

In the classification modeling section above, we explored the stepwise variable selection method in one of the models. Variable selection can also be applied in regression modeling. In this case backward selection is used as the variable selection method. In this method, we start with a current model containing all predictors. We then consider a candidate model that is different than the initial model by the subtraction of one explanatory variable. If the candidate model has an AIC value that is less than our current model then we accept the candidate model as our current model. We then repeat the same steps until the AIC value can no longer be improved by removing a variable. In our backward selection model, the mean squared error of prediction is **1.4326** and the standard error is **0.1505**.

Model	MSE	Standard Error
Backward Selection Linear Regression	1.4326	0.1505

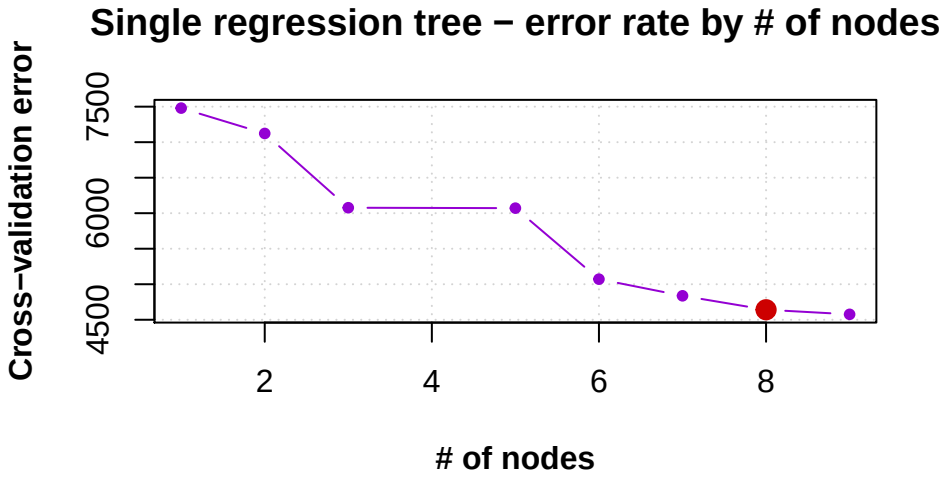
Stepwise Linear Regression Model

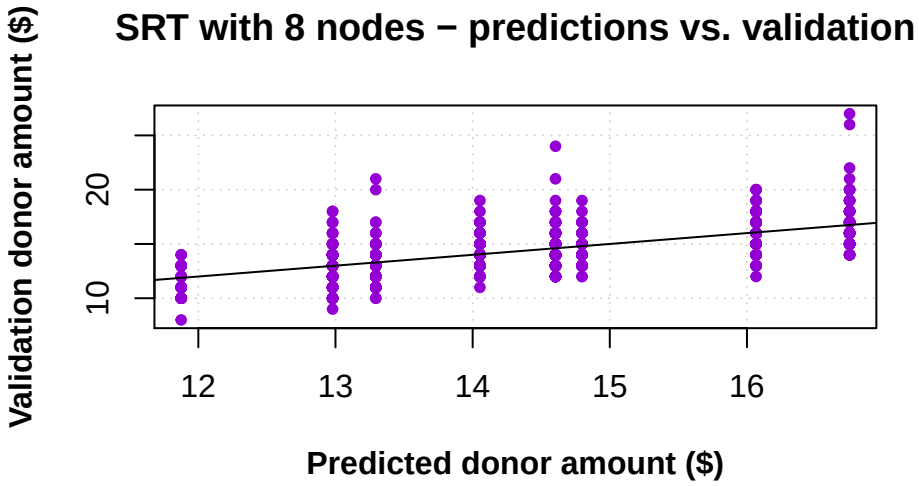
For the sake of comparison, the stepwise variable selection process can also be applied to the regression modeling process. The same rules apply as in the stepwise method in the classification modeling section. In our stepwise linear regression model, the mean squared error of prediction is **1.4559** and the standard error is **0.1493**.

Model	MSE	Standard Error
Stepwise Linear Regression	1.4559	0.1493

Single Regression Tree

The single regression tree method creates one decision tree from which to classify observations. At each split of the tree, a decision is made to classify observations such that the split produces a significant difference in the resulting classes. In the single regression tree below, we apply cross-validation methods to determine the ideal number of nodes, or splits, on the tree. The top graph implies that **either 8 or 9 nodes** results in the lowest error rate. The bottom graph shows a scatter plot of predicted donor amounts compared to validation set data. This single regression tree model with **8 nodes** produces a mean squared error of prediction of **2.5361** and a standard error of **0.2053**.

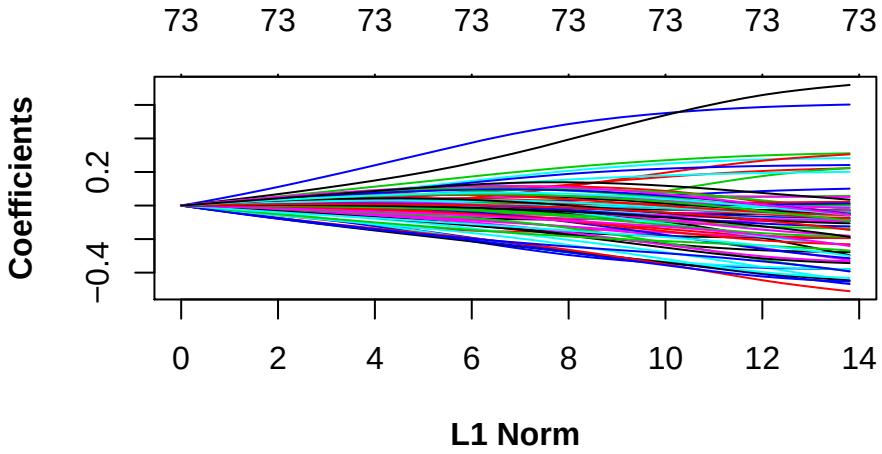


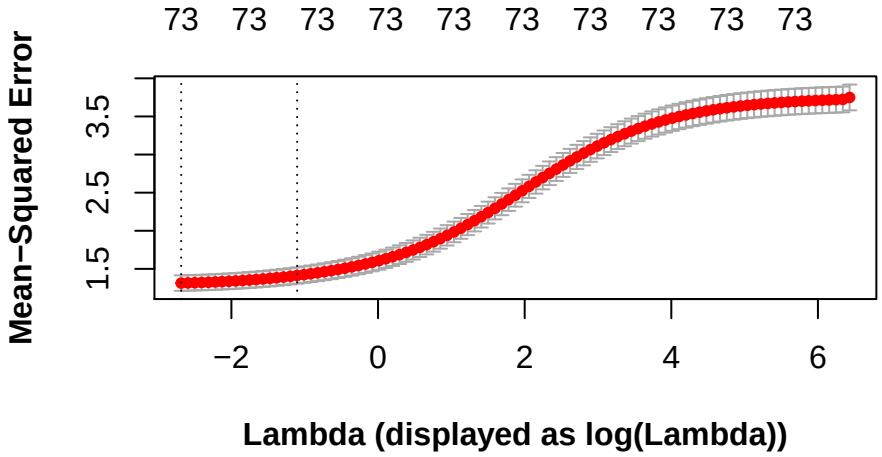


Model	MSE	Standard Error
Single Regression Tree	2.5361	0.2053

Ridge Regression

Ridge regression is an alternative to least squares regression that allows for a reduction in the variance of the model without a corresponding increase in bias. By varying a tuning parameter, *lambda*, the coefficients of some predictors are shrunk, reducing their impact on the response. The shrunken coefficients still exist, however, approaching zero but never reaching it, meaning that ridge regression does not perform variable selection. The two charts below show that the variable coefficients trend toward zero but never actually reach it, while the value of the tuning parameter, and thus the coefficient estimates, impact the mean squared error. In the ridge regression model below, the best value of *lambda* produces a mean squared error of prediction of **1.4654** and a standard error of **0.1580**.

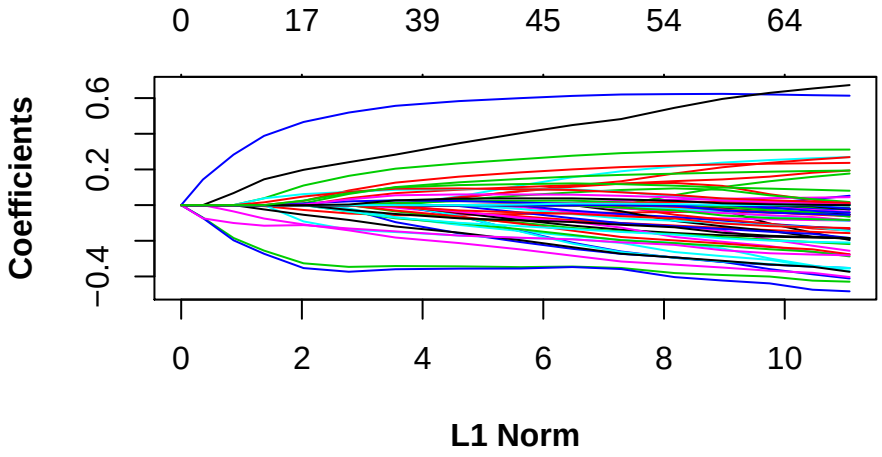


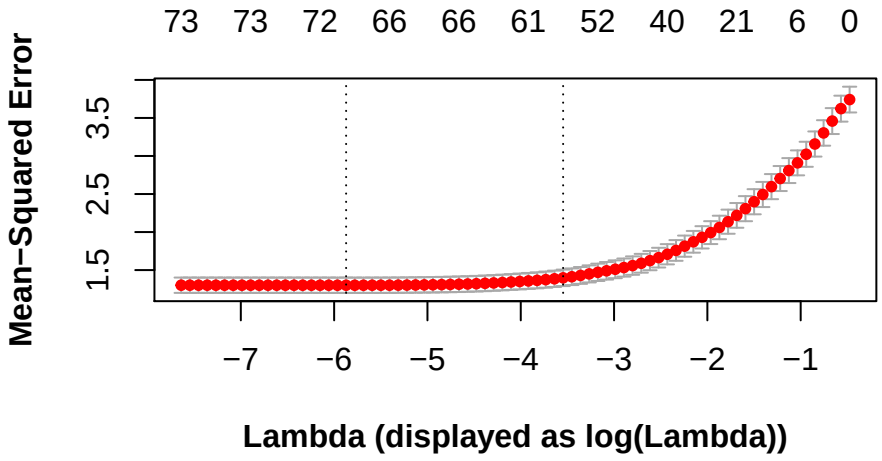


Model	MSE	Standard Error
Ridge Regression	1.4654	0.158

Lasso

The lasso method also serves as an alternative to least squares regression, but has an advantage over ridge regression in that it can shrink some variable coefficients all the way to zero, essentially eliminating the variables from the model. In this way, the lasso performs variable selection. The first chart below shows how coefficients trend toward and become zero at different points, while the second chart shows how the value of *lambda* impacts both the mean squared error and the number of variables in the model. In the lasso model below, the best value of *lambda* produces a mean squared error of prediction of **1.4624** and a standard error of **0.1579**.



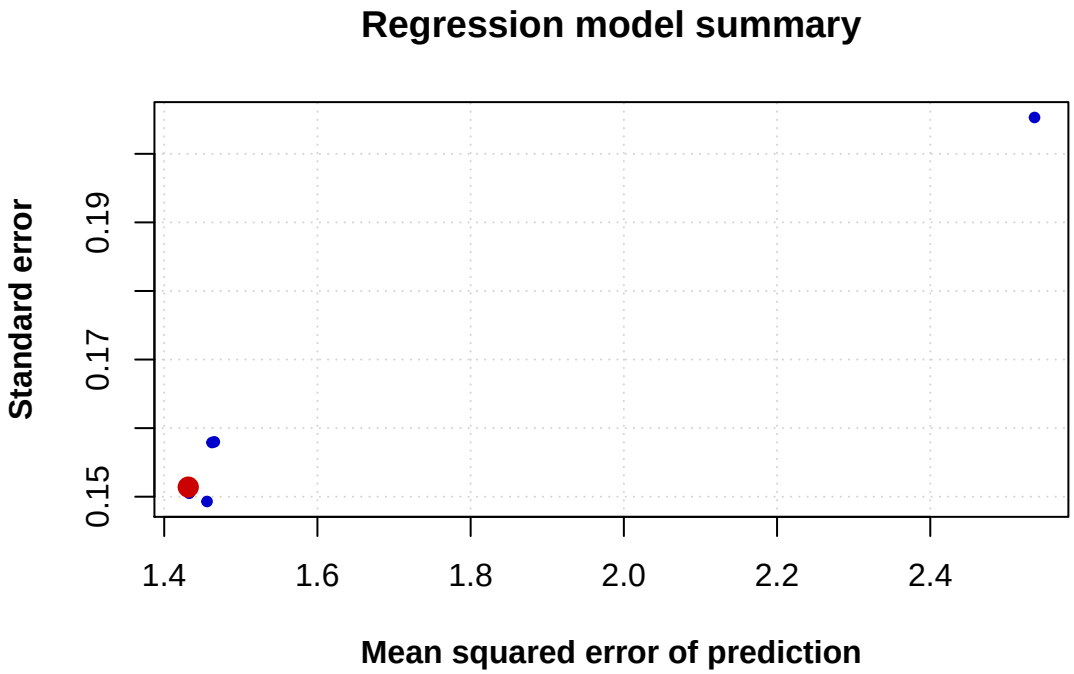


Model	MSE	Standard Error
Lasso	1.4624	0.1579

Regression Model Summary

The regression models assessed have varying results for the mean squared error of prediction and associated standard error. As the summary table below shows, the multiple linear regression model appears to perform best on the validation data. Other regression models fare similarly well, with only the single regression tree model showing noticeably worse results.

Model	MSE	Standard Error
Multiple Linear Regression	1.4314	0.1514
Backward Selection Linear Regression	1.4326	0.1505
Stepwise Linear Regression	1.4559	0.1493
Single Regression Tree	2.5361	0.2053
Ridge Regression	1.4654	0.158
Lasso	1.4624	0.1579



Model Selection

Based on evaluation of the classification and regression models constructed above, the best model in each section is:

- Classification: Stepwise logistic regression
- Regression: Multiple linear regression

Model	Number of Mailings	Max Profit
Stepwise Logistic Regression	1231	11907.5

Model	MSE	Standard Error
Multiple Linear Regression	1.4314	0.1514

Summary

Applying the best classification model and the best regression model to the test data results in a donor pool of **297 individuals** that the charity will contact. As the associated .csv output file shows, this results in an expected gross revenue of **$4,267**. Factoring in the **$2** cost of each mailing, the expected net profit that the charity can expect to receive is **$3,673**.

Test_Mailing	Freq
0	1710
1	297