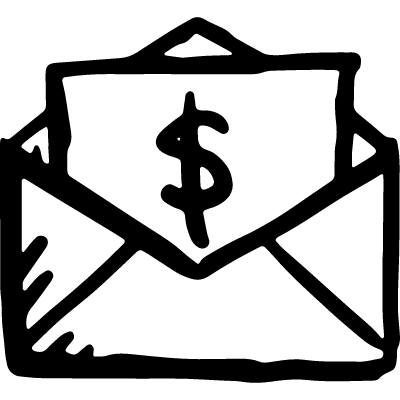
**Predictive Model for Donation Campaign**



MSIA 401: Predictive Analytics I

Fall 2016

Kristin Meier, Dustin Fontaine, Dylan Fontaine, Aditya Venkataraman

**Executive Summary**

Give a non-technical summary of your findings mentioning the key predictors of responders vs. non-responders and of the amount of contributions. This summary is for upper management and should not include any equations and as few statistics as possible. So don't mention things like R2 here. (About 1/2 page)

**Introduction**

Many non-profit organizations solicit contributions from past donors and this can be a tedious and time-consuming process. Furthermore, response rate for donations is low. It can be extremely beneficial to target donors based on the likelihood that they will actually donate, as well as the expected amount. One such non-profit is interested in developing predictive models to do just this. Donor data from 10/2001 to 12/2010 were provided, which include the donation responses to a solicitation sent in 10/2010. With the provided dataset on past donors, models were developed to predict the probability a given donor will donate, and the amount that will be donated. This paper describes the process of building such models and the results that follow.

To use the data for building predictive models, some pre-processing steps were taken, which are described in section [4]. This includes filling in missing values and calculating new variables that may serve as significant predictors. Section [4.1.3] covers exploratory data analysis, used to understand the provided data and to exploit any obvious relationships between the independent and dependent variables. Certain predictor variables are intuitive and expected to influence the expected donation of an individual, such as the size of dollar amount of their contributions and the number of times contributed. Others, such as solicitation code, may or may not play a significant role. Calculated variables, which will be described in further detail in section [4], such as whether donations have increased in quantity over time, are expected to have a positive relationship with probability of donation.

In order to predict whether or not a given donor will donate, logistic models were built. For predicting the donation amount, multiple linear regression models were built. Sections [4.2] and [4.3] respectively cover, in detail, the process for creating these models. The approach to model building was to begin with a simple model and progressively add quadratic and interaction terms as needed, all while eliminating insignificant predictor variables with each step. The logistic models were compared using the AUC (area under the ROC curve) and the linear models were compared using the Adjusted-R squared values. This section also includes model diagnostics, such as detection of influential observations and normalizing transformations.

Section [5] covers the model validation and how the chosen logistic and linear regression models were tested against the test data set. This includes the calculation of expected donations for each individual in the test data set and how well the top 1,000 predicted customers perform in terms of contributions.

Finally, section [6] will draw conclusions from the analysis.

**Model Fitting**

**Data** **Cleaning**

Before any models could be fit to the data, the data had to be cleaned and pre-processed. Missing values had to be filled in, multiple data sets had to be merged together, errors had to be corrected, and redundant information had to be deleted. The following paragraphs describe all of the data cleansing steps that were applied to the raw data before model fitting.

The first step in the data cleaning process was filling in all missing values in the CNDOL2 and CNDOL3 columns. These columns represent the dollar amounts of a person’s second most recent donation and third most recent donation. Since many people had only donated to the organization once, they did not have a second most recent donation or third most recent donation. The data set originally had missing values for these cases, so the missing values were set to “0”.

Next, the contribution and solicitation codes were categorized according to their type. The raw data set originally had these codes stored as four-digit numbers in the columns CNCOD1, SLCOD1, CNCOD2, SLCOD2, CNCOD3, and SLCOD3. The last four of those columns were dropped because they contained many missing values that could not be imputed. For values in the first two of those columns, corresponding code types (A, B, C, D, or M) were assigned and stored in new columns ContType1 and SolType1, allowing the contribution and solicitation codes to be used as a categorical variable in our models. The original columns that contained the four-digit codes were subsequently deleted.

The STATCODE column of the raw data listed the state (or territory) that each donor resided in. This column contained 61 unique levels, which was too many for this column to be a useful predictor. To make this information more useful, the state codes were grouped into ten different regions. The regions corresponded to 9 regions of the United States and an “Other” category for locations that were not in the continental United States. These regions were listed in a separate file which was then merged with the raw data set. After the merging took place, the STATCODE column was deleted and the Region column took its place.

A few columns in the data set were either meaningless or redundant. The ID column represented a donor’s id number, which would not be useful for prediction. This column was deleted. The columns CNDAT1, CNDAT2, and CNDAT3 contained the dates for a person’s most recent contribution, second most recent contribution, and third most recent contribution. The data set also had columns CNMON1, CNMON2, and CNMON3 that represented the number of months since a person’s most recent contribution, second most recent contribution, and third most recent contribution. For a given donation, if you know one of these two pieces of information, you can deduce the other since the current date minus the months since donation is equal to the contribution date. There would be a multicollinearity problem if both sets of variables were kept in the model, so the CNDAT columns were deleted.

Some errors were found in the column named CNMONF. This column represents the months since a person's largest contribution. The column contained mostly values between 0 and 162, but a few values of 1146 were present. The 1146 values are very high and unrealistic since they would mean that people donated to the organization over 90 years ago. The 1146 values are likely the results of a typo, so they were all replaced with the value 146.

After all of the preceding clean up steps were completed, there were still some missing values in the data set. The missing values occurred in columns CNMON2 and CNMON3 for people that did not have more than one donation. Since it is not possible to impute these values, the columns were kept to allow for the various models to be fit with and without the columns included in order to find the best possible fit.

**Added Calculated Variables**

In addition to the variables that were given in the raw data, a few calculated variables were added to the data set that may potentially be significant in the predictive models. The calculated variables avg, avgTime, don2, don3, and incr\_don are described below.

The variable avg represents the average of all of a person's donations. If the person only donated once, avg was set equal to CNDOL1. If the person donated twice, avg was set equal to the average of CNDOL1 and CNDOL2. If a person donated three times, avg was set equal to the average of CNDOL1, CNDOL2, and CNDOL3.

The variable avgTime represents the average amount of time between a person’s donations. If a person has donated more than one time, then avgTime is set equal to (CNMONF - CNMON1)/ (CNTMLIF-1). CMONF is the month of a person’s first donation and CNMON1 is the month of a person’s most recent donation. CNTMLIF is the total number of times that a person has contributed in their lifetime. If a person has only donated once in their lifetime, then avgTime is set equal to 0.

The variable don2 is a binary variable that represents if a person has donated at least twice to the organization (1 if they have, 0 if they haven’t). The variable don3 is also a binary variable; it represents whether a person has donated at least three times to the organization or not (1 if they have, 0 if they haven’t).

Lastly, the variable incr\_don is a binary variable that represents whether or not a person’s most recent donation was larger than their second most recent donation. If it is, this variable is set equal to 1, otherwise it is set equal to 0. If a person does not have two donations, then this variable is also set equal to 0.

**Division of Data into Training Set and Test Set**

The cleaned data was divided into a training set and a test set. The training set was used to fit all of our predictive models and the test set was used to test our predictive models to see how well they performed. Every third observation from the original data into the test set was placed into the test set, making training set to test set size ratio 2:1.

**Exploratory Data Analysis**

***Univariate Continuous Distributions***

Information is available for 66,134 different donations for a non-profit solicitation company (from the training set).  For each donator, there are features regarding latest contributions, first, second and third donations, months since a contribution was made, state and region, gender and total contributions. For each continuous variable, histograms were made to visualize variable centralities, spread, potential outliers and skewness. These general characteristics are demonstrated below in Figures 1 and 2 (Section 7).

*Target Dollars:* The mean target dollars, the response variable, is $2.38 and ranges from no donation up to $1500 donated. The distribution is strongly right skewed and centered at approximately 2 dollars. 73% of the donors did not donate in the fall 2010 donation cycle. The most commonly occurring donations are $2, $3, $5, $10 and $15 making up 23.6% of the whole training set. The $1500 donation would need to be further examined as a potential outlier or influential observation as it is significantly higher than any other donation.

*Latest Contributions:* The mean dollar amount for each of the latest contributions is $9.37, $9.25 and $8.95 respectively. For each subsequent contribution, each mean donation decreases slightly although the centrality and spread remains similar. Each distribution is strongly right skewed with the maximum values also decreasing for each subsequent latest contribution ($1000, $750, $600). The maximum donation for each of the three latest contributions could be examined as potential outliers that would need further testing to determine if those are actual outliers.

*Largest Contribution:* The largest contribution ranges from $2 to $1000 with a mean value of $10.53. The distribution doesn’t appear to follow any known distribution but appears to be bimodal with a peak at roughly $5 and another peak at $20. The distribution is slightly right skewed with no blatant outliers upon observation.

*First Contribution:* The distribution for the first contribution is bimodal with a peak at roughly $5 and another peak at $20. The mean amount for the first contribution is $8 with the minimum amount being nothing donated and the maximum amount being $100. Upon inspection, the first contribution appears to be less in value to the largest contribution. The distribution is slightly skewed to the right with no clear outliers from observation.

*Lifetime Contribution:* The cumulative lifetime contribution has a mean of $33.40 with a minimum value of $2 and a maximum value of $3750. The distribution has a long right tail and is strongly right skewed as well. There are no apparent outliers from inspection.

*Months Since First Donation was Made:* The mean number of months since the first donation was made is 8.97 months. The distribution is right skewed slightly with the maximum months donated being 126 months and the minimum months being right after.  There are two modes that occur at approximately 3 months and 25 since the first donation. From inspection, there appear to be no clear outliers

*Average Time Since First Donation:* The mean average time since the first donation is 4.21 months and the distribution appears to be bimodal with a mode at approximately 1 month and 5 months.  There appear to be no evident outliers. The maximum average time since first donation is 71 months and the minimum average time is 0 months since first donation.

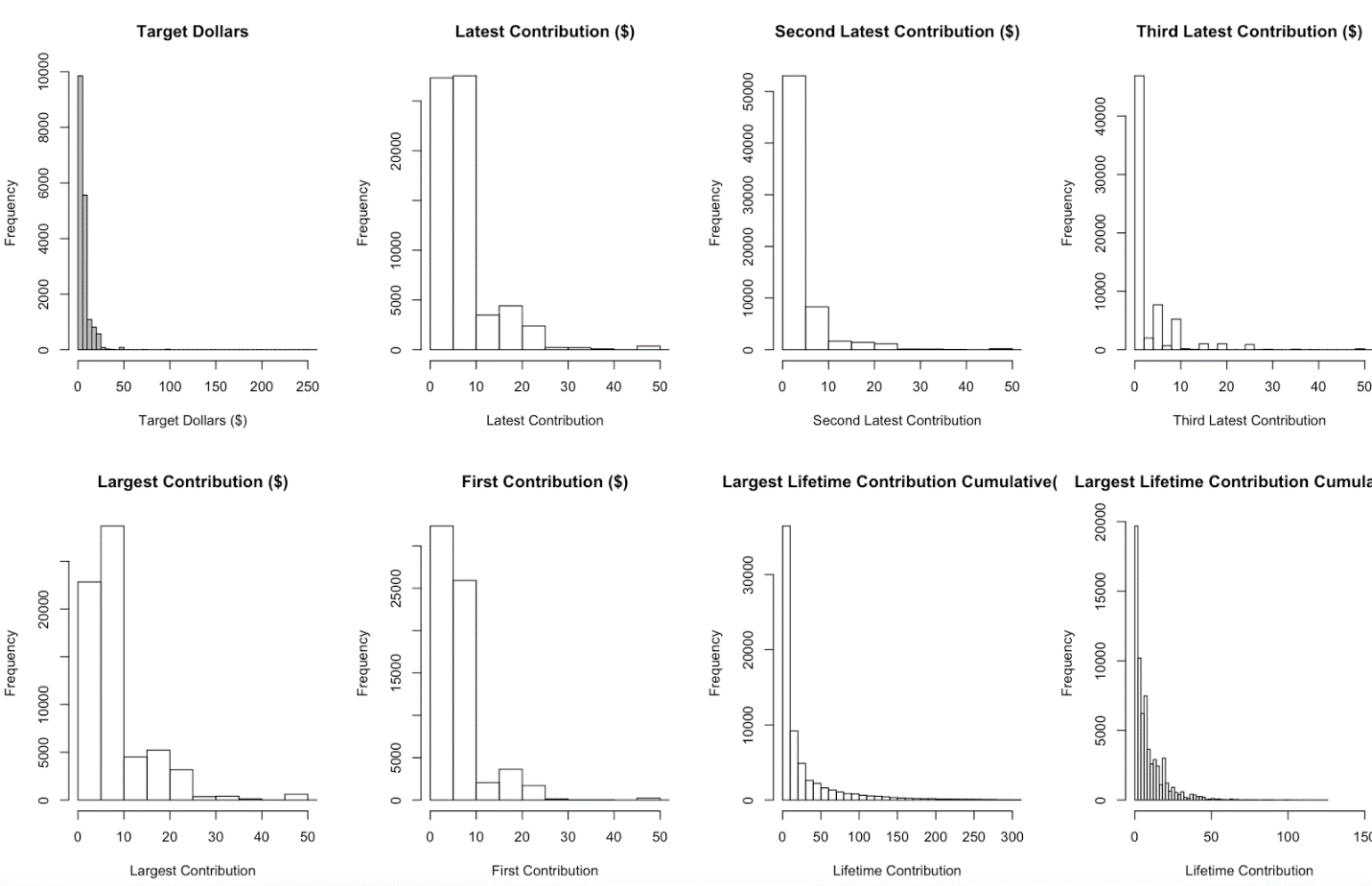


Figure 1: Move to Section ⅞

add( Table like in an “economist paper” summarizing all this)

***Univariate Relationships for Categorical Variables***

The main feature to summarize univariate categorical relationships was looking at counts and proportions per variable.

Add Kristin’s Table

***Bivariate Relationships***

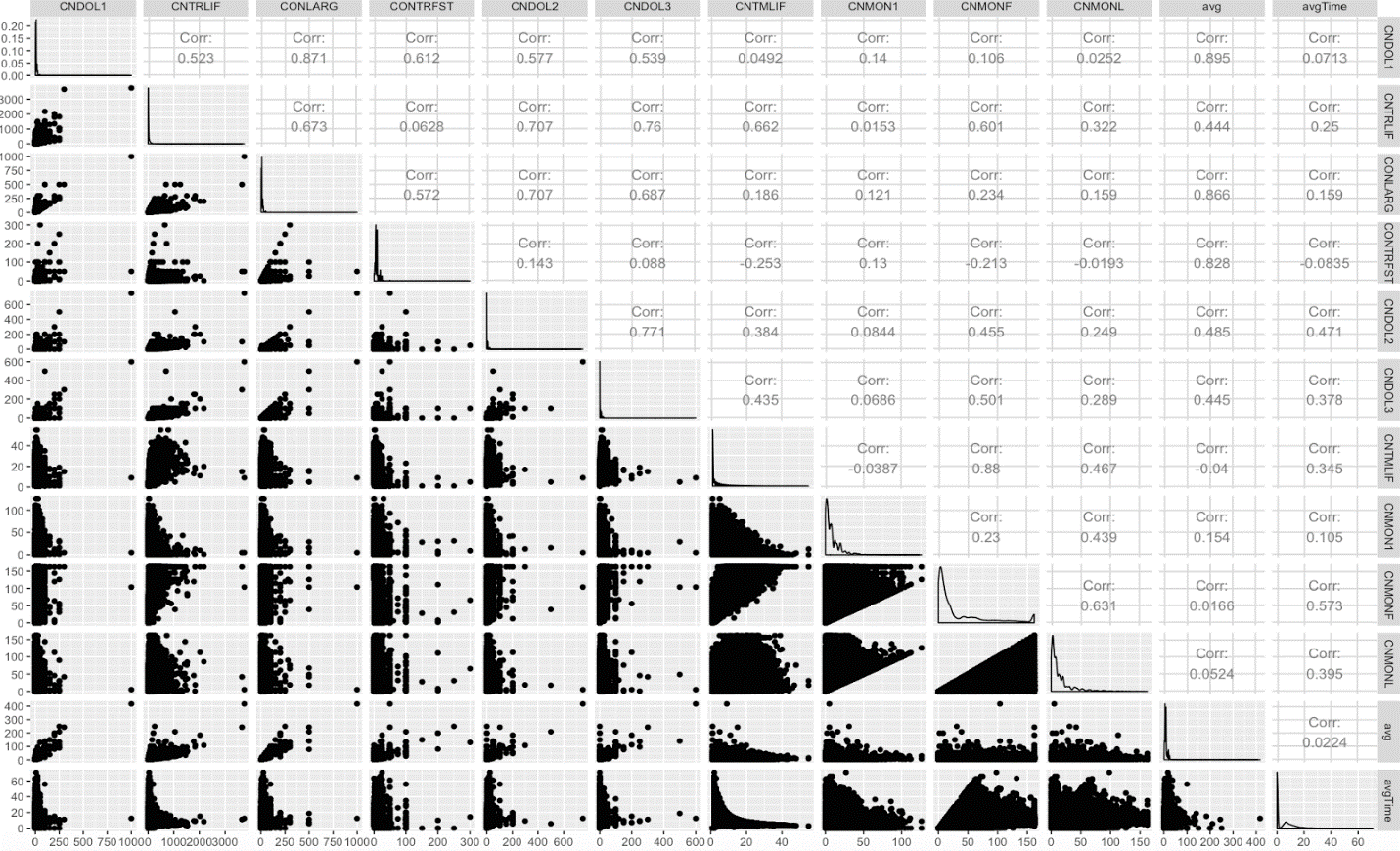


Figure 3: Bivariate Relationships amongst all continuous variables (move to appendix)

Figure 3 demonstrates the pairwise bivariate relationships for all continuous variables. The lower diagonal shows pairwise scatter plots and the upper diagonal shows pairwise correlations between respective variables. The diagonal itself shows the univariate distribution against time. There is strong evidence of multicollinearity present indicated by the pairwise scatter plot. Several of the pairwise correlations are greater than 0.7, namely that between the largest contribution and the first donation, second and third dollar donations with lifetime and largest donations, and months since first contribution to times contributed lifetime. In addition, several other pairwise correlations have moderate evidence of multicollinearity evidenced by a correlation greater than 0.3. Thus, it is worthwhile to explore pairwise inter

**Logistic Regression Model Overview**

A logistic regression model was created to estimate the probability that a person in the test set will donate in response to a future solicitation sent out by the organization. The response variable for the model was a binary variable called donated. If a person made a donation in the test time period from 10/2010 to 12/2010 then their value for donated was set equal to 1. Otherwise their value for donated was set equal to 0.

Many variations of the logistic model were created and tested in order to find the best fitting model. The first model created was very simple and over time more complex models were created. To evaluate the models, the two primary metrics used were AUC (the area under the curve on the ROC plot) and the CCR (correct classification rate). Below is a description of each variation of the logistic regression model and the reasoning for trying it out. A formula for each model is listed in the Appendix.

***Basic Logistic Regression Model***

The initial model included all of the variables that were present in the cleaned-up data set. No quadratic or interaction terms were included in this model and no transformations were made. Just simple first order terms were included. This model was used as a baseline to compare with the more complex models made later on.

|  |  |  |
| --- | --- | --- |
| Name: logModel | AUC: 0.716036 | CCR: 0.7530394 |

***Forwards and Backwards Stepwise Logistic Regression Model***

The next logistic regression models were created using forwards and backwards stepwise algorithms on the basic logistic regression model. Since not all of the variables used in the first model were significant, a potentially better model could be created by filtering down the selected variables to only those that are good predictors.

The forwards stepwise algorithm was run with the starting point being a blank model and the upper bound being the full basic logistic regression model. The algorithm added and removed variables one at a time and selected the best model based on the Akaike Information Criterion (AIC).

|  |  |  |
| --- | --- | --- |
| Name: forwards | AUC: 0.716004 | CCR: 0.7531906 |

The backwards stepwise algorithm worked the same way as the forwards stepwise algorithm, except instead of starting with a blank model, the backwards algorithm starts with the full model. Variables were removed one at a time and the best model was selected based on Akaike Information Criterion (AIC). The resulting model is the same as the model that resulted from forwards stepwise algorithm.

|  |  |  |
| --- | --- | --- |
| Name: backwards | AUC: 0.716004 | CCR: 0.7531906 |

The models that resulted from the stepwise algorithm have similar AUCs and CCRs as the original model, but include less predictors.

***Quadratic Terms Logistic Regression Model***

Based on the exploratory data analysis completed earlier, it was decided that some quadratic terms (squared terms) should be included in the model in addition to the original predictors. This could potentially increase the AUC and/or CCR of the model. For the first quadratic model, a squared term was added for each numeric variable in the cleaned data set. The resulting model is shown below. As you can see, it has better AUC and CCR values than the previous models.

|  |  |  |
| --- | --- | --- |
| Name: logModel.quad | AUC: 0.7194766 | CCR: 0.7550656 |

***Backwards Stepwise, Quadratic Terms Logistic Regression Model***

The backwards stepwise algorithm was then performed on the quadratic logistic model to rid the model of insignificant predictors. The resulting model has less terms than the full quadratic model, but near equivalent AUC and CCR values.

|  |  |  |
| --- | --- | --- |
| Name: backwards.quad | AUC: 0.7194993 | CCR: 0.7550051 |

***Interaction Terms and Quadratic Terms Logistic Regression***

Next, interaction terms were added to the model since there may be a complex relationship between the response variable and multiple predictors. All possible interaction terms for numeric variables were included in the model. This model resulted in the highest AUC so far, but the lowest CCR. The stepwise algorithms were not run on this model since it is so large.

|  |  |  |
| --- | --- | --- |
| Name: logModelInter | AUC: 0.7290657 | CCR: 0.7604186 |

***Quadratic Terms and Log Transformation Logistic Regression***

Next, log transformations were made on various predictor variables since they did not display homoscedasticity. All variables that included a dollar amount (CNDOL1, CNDOL2, etc.) were transformed within the quadratic model. The results for this model were recorded and then some of the log transformations were removed to see if a better model could be produced. All possible combinations of log transformations for dollar amount variables were tested and the best combination is shown below. The AUC of this model was fairly low, but its CCR was fairly high.

|  |  |  |
| --- | --- | --- |
| Name: logModel.quad.log | AUC: 0.720162 | CCR: 0.7566685 |

**Optimal p\* for Best Logistic Regression Model**

The best performing logistic model, which minimized CCR was logModel.quad.log. In order to classify observations 1 or 0 (donate or do not donate), a threshold probability, p\*, must be calculated. The optimal p\* value will minimize the number of incorrectly classified observations. For this model, the optimal p\* value is .5. If this model were to be put into production at the non-profit, they may associate a higher cost with predicting that a donor will donate when in reality they don’t, or a false positive. Thus, a higher cost may be associated with false positives. When using a cost function that follows this logic, Cost = 2\*FP + 1\*FN, where FP is the number of false positives and FN is the number of false negatives, the optimal p\* value becomes .37. Cost curves can be seen in [figure x].

**Multiple Regression Model**

A multiple linear regression model was built to estimate the dollar amount that a person in the test set would donate in the future time period.  The response variable used was TARGDOL.  To train the model, only donors with a TARGDOL > 0 were used.  This prevents the many records (73%) with a TARGDOL = 0 from biasing the predictions downwards.  We will be using this model in conjunction with the logistic model that predicts if a donor will donate, so this is a reasonable assumption.

An array of multiple linear regression models were built with various degrees of complexity.  Each model was evaluated by looking at its Multiple R-Squared and Adjusted R-Squared value.  Below is a description of each model.

***Basic Multiple Linear Regression Model***

The first model includes all variables present in the cleaned-up data set.  There were no interactions or higher order terms added. This model is the simplest and serves as a benchmark for the later models.

|  |  |  |
| --- | --- | --- |
| Name: lmModel | Multiple R-squared:  0.839 | Adjusted R-squared:  0.8387 |

***Backwards Stepwise Multiple Linear Regression Model***

This model uses the same variables as the basic multiple regression model, but a backwards stepwise algorithm is run to find the best model (based on AIC) by removing one variable at a time.  This was explored because not every variable in the first model was found to be significant.  The resulting model and its performance is shown below. The backwards stepwise model has similar performance metrics to the original, but some of the predictors were removed.

|  |  |  |
| --- | --- | --- |
| Name: backwards.lm | Multiple R-squared:  0.8388 | Adjusted R-squared:  0.8387 |

***Quadratic Terms Multiple Linear Regression Model***

Including quadratic terms could possibly provide more information than just their first order counterparts.  For each numeric column, an additional column was included that had the squared values of the numeric column.  The resultant dataset was fit to a linear model. As expected, the Multiple R-squared is higher since there are more variables than in the original model.  The Adjusted R-squared also increased, which is a positive sign.

|  |  |  |
| --- | --- | --- |
| Name: lmModel.quad | Multiple R-squared:  0.8804 | Adjusted R-squared:  0.8801 |

***Backwards Stepwise, Quadratic Terms Multiple Linear Regression Model***

A backwards stepwise was again fitted, this time including all of the squared terms.  This model was fit because many predictors in the initial quadratic model were insignificant.  As expected, many variables were removed.  The Multiple R-squared dropped very slightly, with the Adjusted R-squared remaining the same.

|  |  |  |
| --- | --- | --- |
| Name: backwards.lm.quad | Multiple R-squared:  0.8803 | Adjusted R-squared:  0.8801 |

***Interaction Terms and Quadratic Terms Multiple Linear Regression***

In addition to including the squared versions of numeric predictors, this time the interactions between numeric variables was also included. This model shows a great increase in both performance metrics.

|  |  |  |
| --- | --- | --- |
| Name: lmModel.inter | Multiple R-squared:  0.893 | Adjusted R-squared:  0.8922 |

***Removed Influential Observations, Multiple Linear Model***

The possibility of influential observations skewing the fit of the regression was explored.  The influence.measures function in the stats package was used to identify influential observations.  The simplest multiple linear regression model was fit with the remaining, uninfluential data points.

|  |  |  |
| --- | --- | --- |
| Name: lmModel.rminf | Multiple R-squared:  0.6745 | Adjusted R-squared:  0.674 |

This version of the model performed worse.  Both r-squareds decreased greatly.  This is surprising, but is not unexplainable.  There were a total of 1236 influential observations removed, a rather sizable chunk of the training set.  This much of a change in the training set can cause this much of a change in the output.

***Removed Influential Observations, Quadratic Multiple Linear Model***

The influential observations were also removed from the quadratic model.  This had a similar result as the previous model, the model fit decreased greatly from the fit without removing those observations.

|  |  |  |
| --- | --- | --- |
| Name: lmModel.quad.rminf | Multiple R-squared:  0.6852 | Adjusted R-squared:  0.6844 |

***Quadratic terms, Log of some terms, Multiple Linear Regression***

Before fitting, this time the log was taken of many of the numerical predictors.  These predictors were chosen based on their distributions.  There was some trial and error testing done to find the best subset of predictors to log.

|  |  |  |
| --- | --- | --- |
| Name: lmModel.quad.log | Multiple R-squared:  0.8675 | Adjusted R-squared:  0.8672 |

**Final Models Chosen**

From the array of logistic and multiple regression models that were fit to the data, one model of each type had to be selected to predict how much each person in the test set would donate. The logistic model calculated the probability that each person would donate and the multiple regression model calculated their predicted donation. Each person’s probability value was multiplied by their predicted donation to calculate their expected donation. Then, the donors with the top 1000 expected donations were selected and their actual donations were added together.

Since the selected models had to work together, it was not appropriate to select the two best models independently. The models had to be tested together so that the total donation from the top 1000 expected donors could be calculated. All possible combinations of the logistic and multiple regression models were tested together and the results from each test were recorded. The top five model combinations are shown below in Figure 4, and the full results are shown in the appendix.

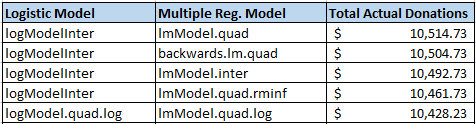


Figure 4: Top Five Model Combinations

As you can see, the best combination of models was both models that included quadratic terms and log transformations of some variables. The total expected donation from the top 1000 expected donors was $10,514.73.

Add- Graphs, QQnorms, VIF/Cooks distance etc

**Model Validation**

Explain how you validated the model against the test data set. Report the results about how well the model predicted the test set sales values and how well your top 1,000 predicted customers performed in terms of contributions. (About 4 pages)

Expected donation was $10,300

**Conclusions**

Draw conclusions about significant predictors, any key missing predictors which

would have improved the model, etc. (About 2 pages)

**References**

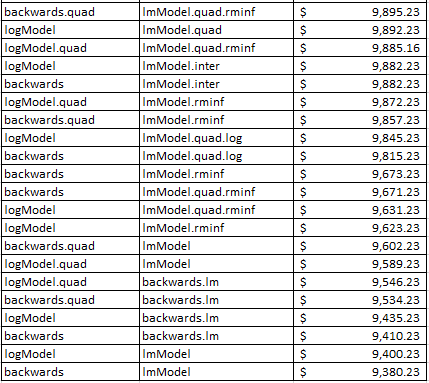
**Appendix (Printouts, Graphs)**

Formula for lmModel.inter

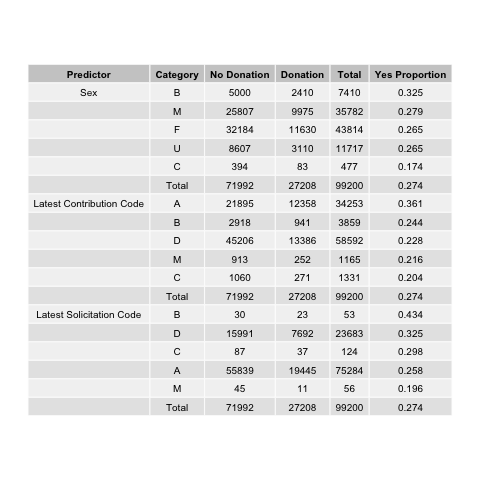
|  |
| --- |
| targdol ~ sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST +   sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF +   sq\_CNMONL + sq\_avg + sq\_avgTime + CNDOL1 + CNTRLIF + CONLARG +   CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + CNMON1 + CNMONF +   CNMONL + avg + avgTime + don2 + don3 + incr\_don + CNDOL1.CNTRLIF +   CNDOL1.CONLARG + CNDOL1.CONTRFST + CNDOL1.CNDOL2 + CNDOL1.CNDOL3 +   CNDOL1.CNTMLIF + CNDOL1.CNMON1 + CNDOL1.CNMONF + CNDOL1.CNMONL +   CNDOL1.avg + CNDOL1.avgTime + CNDOL1.don2 + CNDOL1.don3 +   CNDOL1.incr\_don + CNTRLIF.CONLARG + CNTRLIF.CONTRFST + CNTRLIF.CNDOL2 +   CNTRLIF.CNDOL3 + CNTRLIF.CNTMLIF + CNTRLIF.CNMON1 + CNTRLIF.CNMONF +   CNTRLIF.CNMONL + CNTRLIF.avg + CNTRLIF.avgTime + CNTRLIF.don2 +   CNTRLIF.don3 + CNTRLIF.incr\_don + CONLARG.CONTRFST + CONLARG.CNDOL2 +   CONLARG.CNDOL3 + CONLARG.CNTMLIF + CONLARG.CNMON1 + CONLARG.CNMONF +   CONLARG.CNMONL + CONLARG.avg + CONLARG.avgTime + CONLARG.don2 +   CONLARG.don3 + CONLARG.incr\_don + CONTRFST.CNDOL2 + CONTRFST.CNDOL3 +   CONTRFST.CNTMLIF + CONTRFST.CNMON1 + CONTRFST.CNMONF + CONTRFST.CNMONL +   CONTRFST.avg + CONTRFST.avgTime + CONTRFST.don2 + CONTRFST.don3 +   CONTRFST.incr\_don + CNDOL2.CNDOL3 + CNDOL2.CNTMLIF + CNDOL2.CNMON1 +   CNDOL2.CNMONF + CNDOL2.CNMONL + CNDOL2.avg + CNDOL2.avgTime +   CNDOL2.don2 + CNDOL2.don3 + CNDOL2.incr\_don + CNDOL3.CNTMLIF +   CNDOL3.CNMON1 + CNDOL3.CNMONF + CNDOL3.CNMONL + CNDOL3.avg +   CNDOL3.avgTime + CNDOL3.don2 + CNDOL3.don3 + CNDOL3.incr\_don +   CNTMLIF.CNMON1 + CNTMLIF.CNMONF + CNTMLIF.CNMONL + CNTMLIF.avg +   CNTMLIF.avgTime + CNTMLIF.don2 + CNTMLIF.don3 + CNTMLIF.incr\_don +   CNMON1.CNMONF + CNMON1.CNMONL + CNMON1.avg + CNMON1.avgTime +   CNMON1.don2 + CNMON1.don3 + CNMON1.incr\_don + CNMONF.CNMONL +   CNMONF.avg + CNMONF.avgTime + CNMONF.don2 + CNMONF.don3 +   CNMONF.incr\_don + CNMONL.avg + CNMONL.avgTime + CNMONL.don2 +   CNMONL.don3 + CNMONL.incr\_don + avg.avgTime + avg.don2 +   avg.don3 + avg.incr\_don + avgTime.don2 + avgTime.don3 + avgTime.incr\_don +   don2.don3 + don2.incr\_don + don3.incr\_don + donated + SEX +   ContType1 + SolType1 + Region |

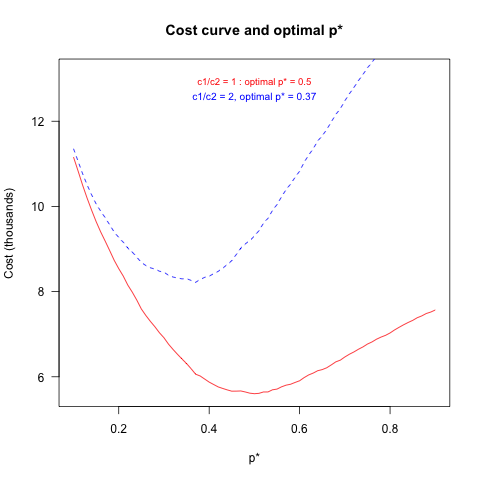
**Results from Every Combination of Models**





**Frequency Table for Categorical Variables**





**Logistic Model Formulas**

|  |  |
| --- | --- |
| **Model Name** | **Formula** |
| logModel | donated ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + CNMON1 + CNMONF + CNMONL + avg + avgTime + don2 + don3 + incr\_don + SEX + ContType1 + SolType1 + Region |
| forwards | donated ~ CNMON1 + CNMONL + CNTMLIF + CNMONF + ContType1 + CONTRFST +  SolType1 + incr\_don + SEX + don2 + avgTime + CNTRLIF + CNDOL2 +  CNDOL1 + CONLARG + don3 + Region |
| backwards | donated ~ CNMON1 + CNMONL + CNTMLIF + CNMONF + ContType1 + CONTRFST +  SolType1 + incr\_don + SEX + don2 + avgTime + CNTRLIF + CNDOL2 +  CNDOL1 + CONLARG + don3 + Region |
| logModel.quad | donated ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 + incr\_don + Region + sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST + sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + sq\_avg + sq\_avgTime |
| backwards.quad | donated ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNTMLIF + CNMON1 + CNMONF + CNMONL + avg + avgTime + don2 + don3 + incr\_don + SEX + ContType1 + SolType1 + Region + sq\_CNDOL1 + sq\_CONLARG + sq\_CNDOL2 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONL + sq\_avg + sq\_avgTime |
| logModel.quad.log | donated ~ log(CNDOL1) + CNTRLIF + log(CONLARG) + log(CONTRFST+1) + log(CNDOL2+1) + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 +incr\_don + Region + log(sq\_CNDOL1) + sq\_CNTRLIF + log(sq\_CONLARG) + log(sq\_CONTRFST+1) + log(sq\_CNDOL2+1) + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + log(sq\_avg) + sq\_avgTime |
| logModelInter | donated ~ sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST + sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + sq\_avg + sq\_avgTime + CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + CNMON1 + CNMONF + CNMONL + avg + avgTime + don2 + don3 + incr\_don + CNDOL1.CNTRLIF + CNDOL1.CONLARG + CNDOL1.CONTRFST + CNDOL1.CNDOL2 + CNDOL1.CNDOL3 + CNDOL1.CNTMLIF + CNDOL1.CNMON1 + CNDOL1.CNMONF + CNDOL1.CNMONL + CNDOL1.avg + CNDOL1.avgTime + CNDOL1.don2 + CNDOL1.don3 + CNDOL1.incr\_don + CNTRLIF.CONLARG + CNTRLIF.CONTRFST + CNTRLIF.CNDOL2 + CNTRLIF.CNDOL3 + CNTRLIF.CNTMLIF + CNTRLIF.CNMON1 + CNTRLIF.CNMONF + CNTRLIF.CNMONL + CNTRLIF.avg + CNTRLIF.avgTime + CNTRLIF.don2 + CNTRLIF.don3 + CNTRLIF.incr\_don + CONLARG.CONTRFST + CONLARG.CNDOL2 + CONLARG.CNDOL3 + CONLARG.CNTMLIF + CONLARG.CNMON1 + CONLARG.CNMONF + CONLARG.CNMONL + CONLARG.avg + CONLARG.avgTime + CONLARG.don2 + CONLARG.don3 + CONLARG.incr\_don + CONTRFST.CNDOL2 + CONTRFST.CNDOL3 + CONTRFST.CNTMLIF + CONTRFST.CNMON1 + CONTRFST.CNMONF + CONTRFST.CNMONL + CONTRFST.avg + CONTRFST.avgTime + CONTRFST.don2 + CONTRFST.don3 + CONTRFST.incr\_don + CNDOL2.CNDOL3 + CNDOL2.CNTMLIF + CNDOL2.CNMON1 + CNDOL2.CNMONF + CNDOL2.CNMONL + CNDOL2.avg + CNDOL2.avgTime + CNDOL2.don2 + CNDOL2.don3 + CNDOL2.incr\_don + CNDOL3.CNTMLIF + CNDOL3.CNMON1 + CNDOL3.CNMONF + CNDOL3.CNMONL + CNDOL3.avg + CNDOL3.avgTime + CNDOL3.don2 + CNDOL3.don3 + CNDOL3.incr\_don + CNTMLIF.CNMON1 + CNTMLIF.CNMONF + CNTMLIF.CNMONL + CNTMLIF.avg + CNTMLIF.avgTime + CNTMLIF.don2 + CNTMLIF.don3 + CNTMLIF.incr\_don + CNMON1.CNMONF + CNMON1.CNMONL + CNMON1.avg + CNMON1.avgTime + CNMON1.don2 + CNMON1.don3 + CNMON1.incr\_don + CNMONF.CNMONL + CNMONF.avg + CNMONF.avgTime + CNMONF.don2 + CNMONF.don3 + CNMONF.incr\_don + CNMONL.avg + CNMONL.avgTime + CNMONL.don2 + CNMONL.don3 + CNMONL.incr\_don + avg.avgTime + avg.don2 + avg.don3 + avg.incr\_don + avgTime.don2 + avgTime.don3 + avgTime.incr\_don + don2.don3 + don2.incr\_don + don3.incr\_don + targdol + SEX + ContType1 + ContType2 + ContType3 + SolType1 + SolType2 + SolType3 + Region |

**Multiple Regression Model Formulas**

|  |  |
| --- | --- |
| **Model Name** | **Formula** |
| lmModel | targdol ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 + incr\_don + Region |
| backwards.lm | targdol ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNTMLIF +  CNMONF + CNMONL + avg + don2 + don3 + incr\_don + ContType1 + SolType1 |
| lmModel.quad | targdol ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 + incr\_don + Region + sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST + sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + sq\_avg + sq\_avgTime |
| backwards.lm.quad | targdol ~ sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST +  sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONL +   sq\_avg + sq\_avgTime + CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + CNMONF + CNMONL + avg + avgTime + don2 + don3 + incr\_don + SEX + ContType1 + SolType1 |
| lmModel.rminf | targdol ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 + incr\_don + Region |
| lmModel.quad.rminf | targdol ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST + CNDOL2 + CNDOL3 + CNTMLIF + SEX + CNMON1 + CNMONF + CNMONL + ContType1 + SolType1 + avg + avgTime + don2 + don3 + incr\_don + Region + sq\_CNDOL1 + sq\_CNTRLIF + sq\_CONLARG + sq\_CONTRFST + sq\_CNDOL2 + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + sq\_avg + sq\_avgTime |
| lmModel.quad.log | targdol ~ log(sq\_CNDOL1) + sq\_CNTRLIF + log(sq\_CONLARG) + log(sq\_CONTRFST+1) + log(sq\_CNDOL2+1) + sq\_CNDOL3 + sq\_CNTMLIF + sq\_CNMON1 + sq\_CNMONF + sq\_CNMONL + I(log(sq\_avg)) + sq\_avgTime + log(CNDOL1) + CNTRLIF + log(CONLARG) + log(CONTRFST+1) + log(CNDOL2+1) + CNDOL3 + CNTMLIF + CNMON1 + CNMONF + CNMONL + avg + avgTime + don2 + don3 + incr\_don + SEX + ContType1 + SolType1 + Region |