PREDICT 422 Charity Project

Nikhil Agarwal
Northwestern University
PREDICT 422 | Section 58

Introduction

The intent of this project is to construct a classification and a predictive model that can predict potential donors (DONR) and the expected donation amount (DAMT) respectively. Over 8000 observations have been used to construct the candidate models. Furthermore, two groups of datasets (one using factors and the other not, with and without variable transformations) were used to construct over 12 types of models. The key metrics to determine the best models are maximum profit (for the response variable DONR) and average squared prediction error (for the response variable DAMT). This report provides a succinct overview of the models and results. Next steps are also discussed.

Analysis

Dataset Overview

The primary dataset consists of 8,009 observations with 20 potential predictors and two response variables. Table 1 is a data dictionary that provides a quick summary of the variables. Note that the data types listed are unmodified from the import process and have not been transformed.

Table 1: Data dictionary

Columns	Definition	Data Type
REG1	Region (There are five geographic regions; only 4 are needed for analysis since if a potential donor falls into none of the four he or she must be in the other region.	Binary
REG2	Same as REG1	Binary
REG3	Same as REG1	Binary
REG4	Same as REG1	Binary
HOME	(1 = homeowner, 0 = not a homeowner)	Binary
CHLD	Number of children	Integer
HINC	Household income (7 categories)	Integer
GENF	Gender (0 = Male, 1 = Female)	Binary
WRAT	Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and 0 being the lowest.)	Integer
AVHV	Average Home Value in potential donor's neighborhood in $ thousands	Numeric
INCM	Median Family Income in potential donor's neighborhood in $ thousands	Numeric
INCA	Average Family Income in potential donor's neighborhood in $ thousands	Numeric
PLOW	Percent categorized as “low income” in potential donor's neighborhood	Numeric
NPRO	Lifetime number of promotions received to date	Integer
TGIF	Dollar amount of lifetime gifts to date	Numeric
LGIF	Dollar amount of largest gift to date	Numeric
RGIF	Dollar amount of most recent gift	Numeric
TDON	Number of months since last donation	Integer
TLAG	Number of months between first and second gift	Integer
AGIF	Average dollar amount of gifts to date	Numeric
DONR	Classification Response Variable (1 = Donor, 0 = Non-donor)	Binary
DAMT	Prediction Response Variable (Donation Amount in $).	Numeric

For the model training process, the dataset had already pre-defined observations that were part of the training, validation, or testing datasets. As such, each of the observations were then split up according to the pre-defined classification. Table 2 provides a break out of each data collection and the number of observations in each one.

Table 2: Breakdown of training, validation, and testing data counts

Data Collection	Number of Observations	Percentage of observations
Training	3984	49.7%
Validation	2018	25.2%
Testing	2007	25.1%

Approximately, 50% of the observations are part of the training dataset. No variable within the dataset had missing values. Therefore, data imputation will not be necessary for this dataset. The next section provides an exploratory data analysis of the training dataset[1]. Outliers and multicollinearity analyses were considered out of scope for this report.

EDA of Training Dataset

Table 3 illustrates some basic statistics on the many numeric variables along with the response variable DAMT.

Table 3: Descriptive statistics on numeric variables

Variable	Mean	Maximum	Skewness[2]	Kurtosis[3]	Standard Deviation	Median
AVHV	185.1842	710	1.54803	7.372442	74.70008	171
INCM	44.2884	287	2.000311	11.10594	25.17556	39
INCA	57.13554	287	1.867308	9.955876	25.27604	52
PLOW	13.73268	87	1.381409	4.935768	13.09388	10
NPRO	61.62927	164	0.2796	2.385898	30.3397	60
TGIF	116.7447	1974	5.053572	73.31685	85.45911	91
LGIF	23.19302	642	7.87338	108.4112	31.27813	15
RGIF	15.54669	173	2.653386	18.7009	12.0772	12
TDON	18.81476	40	1.126973	5.425644	5.582317	18
TLAG	6.301958	34	2.415135	11.45665	3.598049	5
AGIF	11.65924	64.22	1.64415	7.953055	6.411263	10.22
DAMT	7.26004	25	0.102645	1.156039	7.378314	10

Interestingly, there are many variables that are not normally distributed and are heavily skewed. There is also evidence of outliers (when comparing the mean, median, and maximum values) amongst some of the variables. For instance, AVHV has a mean of 185.18, but the median is only 171 and the maximum value is 710.

[1] Note that an EDA is not done on the validation or testing datasets to ensure minimal influence on the training dataset modeling process

[2] A skewness of 0 indicates normal distribution

[3] A kurtosis of 3 indicates normal distribution

Table 4 shows the correlation matrix for all of the predictor variables versus the two response variables. The intent here is to discover which predictor variables tend to have most correlation with the response variables.

Table 4: Correlation matrix of response variables vs. potential predictors

Predictor Variable	**donr**	**damt**
reg1	0.056	0.047
reg2	0.247	0.211
reg3	-0.104	-0.084
reg4	-0.126	-0.089
home	0.289	0.288
chld	-0.531	-0.551
hinc	0.028	0.045
genf	-0.017	-0.019
wrat	0.249	0.243
avhv	0.119	0.118
incm	0.158	0.164
inca	0.139	0.142
plow	-0.143	-0.137
npro	0.136	0.141
tgif	0.116	0.127
lgif	0.026	0.082
rgif	0.015	0.085
tdon	-0.096	-0.088
tlag	-0.141	-0.133
agif	0.009	0.079

Surprisingly, there is no one predictor variable that has a strong positive or negative correlation with either DONR or DAMT. Nevertheless, the predictor CHLD has a medium negative correlation with the response variables. Intuitively, this makes sense as individuals with children may be less likely to donate and if they do, may have reduced donations. Another interesting note is the little impact that the variable GENF has on either response variable.

Note how the category region is already mutated into four[4] binary variables (REG1, REG2, etc.). Figure 1 illustrates the size of each region. Region 1 makes up about 20.5%, Region 2 makes up about 33.6%, Region 3 makes up about 12.3%, Region 4 makes up about 13.5%, and Region 5 makes up about 20.1% of the training dataset.

[4] If all 4 regions (REG1, REG2, etc.) are 0, then the observation is considered as part of Region 5. Consider these variables as indicator (i.e., dummy) variables

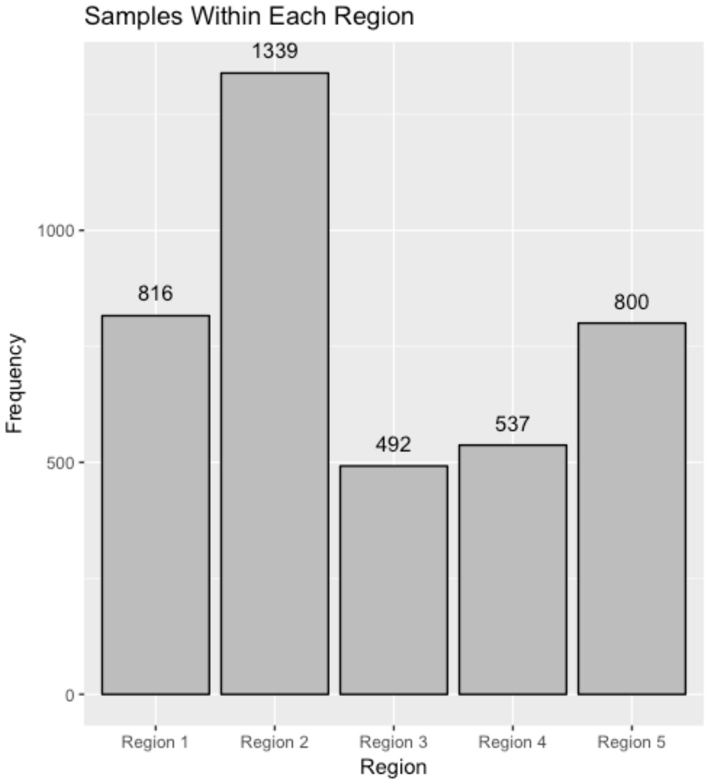


Figure 1: Training data observation counts by Region

Numerous histograms and boxplots were created for each numeric and integer related variables. For brevity not all of the histograms and boxplots are in this report. However, Figure 2 illustrates an example of the boxplot and histogram for the variable AVHV.

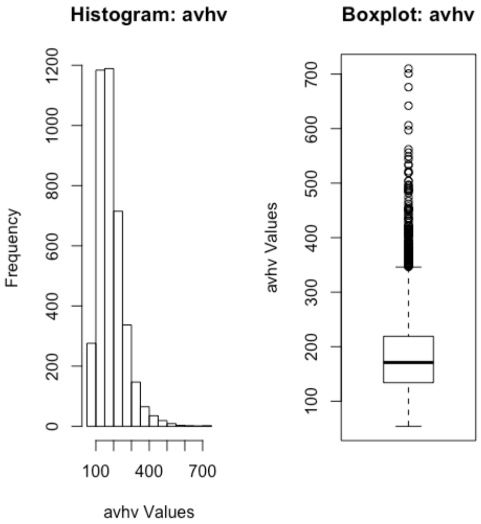


Figure 2: Example histogram and boxplot of variable AVHV

Note the skewness in the histogram as well as the presence of outliers for this predictor variable. Table 3 alluded to this skewness (1.54) earlier. For this project, outlier management was not conducted. For future steps, this may be a necessary step or if a more robust model is required than the one suggested in this exercise.

Variable Transformations

Many of the numeric variables (e.g., AVHV, INCA, etc.) have significant skewness and do not adhere to normality assumption. Figure 2 provided an excellent example of the skewness for variable AVHV. Furthermore, many of the numeric variables were also standardized so that they would have a mean of 0 and a standard deviation of 1. Table 5 summarizes the variable transformations applied to each of the variables, if any. Note that no standardization or transfixion was done for either response variable.

Table 5: Overview of datasets, predictor variable transformations, and data type

	DATASET 1				DATASET 2			
Predictor Variables	**Raw**	**Log Transformation**	**Standardized**	**Data Type**	**Raw**	**Log Transformation**	**Standardized**	**Data Type**
REG1	Y	N	N	Binary	Y	N	N	Binary
REG2	Y	N	N	Binary	Y	N	N	Binary
REG3	Y	N	N	Binary	Y	N	N	Binary
REG4	Y	N	N	Binary	Y	N	N	Binary
HOME	Y	N	N	Binary	Y	N	N	Binary
CHLD	Y	N	N	Factor	Y	N	N	Numeric
HINC	Y	N	N	Factor	Y	N	N	Numeric
GENF	Y	N	N	Binary	Y	N	N	Binary
WRAT	Y	N	N	Factor	Y	N	N	Numeric
AVHV	Y	Y	Y	Numeric	Y	Y	Y	Numeric
INCM	Y	Y	Y	Numeric	Y	Y	Y	Numeric
INCA	Y	Y	Y	Numeric	Y	Y	Y	Numeric
PLOW	Y	Y	Y	Numeric	Y	Y	Y	Numeric
NPRO	Y	N	N	Numeric	Y	N	N	Numeric
TGIF	Y	Y	Y	Numeric	Y	Y	Y	Numeric
LGIF	Y	Y	Y	Numeric	Y	Y	Y	Numeric
RGIF	Y	Y	Y	Numeric	Y	Y	Y	Numeric
TDON	Y	N	N	Numeric	Y	N	N	Numeric
TLAG	Y	N	N	Numeric	Y	N	N	Numeric
AGIF	Y	Y	Y	Numeric	Y	Y	Y	Numeric

In Table 5, there is also mention of two different datasets. The first dataset (Dataset 1), considers predictor variables CHLD, HINC, and WRAT as factors. The second dataset (Dataset 2), considers these variables as numeric. The same transformation was applied to the validation datasets as well. The use of these two datasets alludes to an assumption on whether a particular variable (e.g., CHLD, HINC, or WRAT) could be considered as classifications or numeric values. A log transformation was used since it presented more adherence to the normality assumption. Figure 3 illustrates the variable AVHV after it undergoes a log transformation.

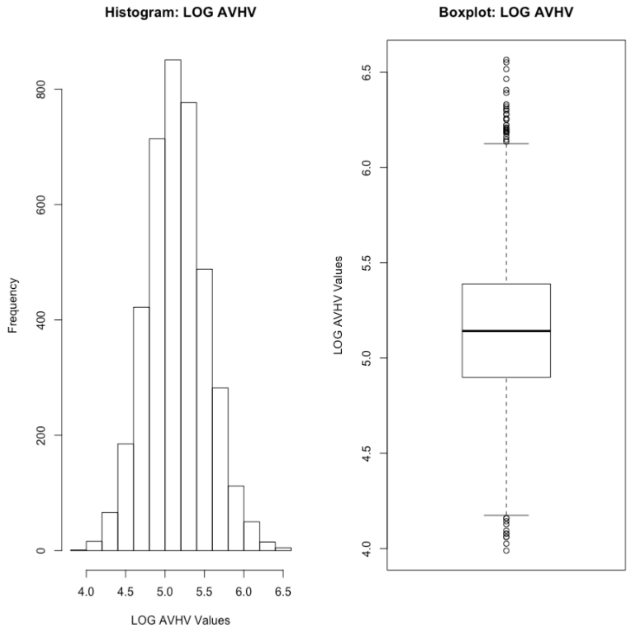


Figure 3: Example histogram and boxplot of log transformation of variable AVHV

After the log transformation of this variable, the skewness and kurtosis values were approximately 0.168 and 3.123 respectively. Another note to make here is that no variable that had undergone a log transformation was standardized (mean of 0 with a standard deviation of 1).

Analysis of the DONR Classification Models

For the classification model, five different models were constructed using the two different datasets (identified in Table 5): logistic, LDA, QDA, Boosting, and Random Forest. Table 6 summarizes all of the results for each type of model. In this section, succinct highlights of each type of model is provided. Prior to operationalizing any classification model, the resulting projections will need to be adjusted for the weighted sampling used in the training and validation datasets. In the **Results of the DONR Classification Models** section, justification is provided for the model chosen for each response variable. For each classification model, a plot like Figure 4 was constructed to help illustrate the idea that as the number of mailings increased, the profit would increase, but then decrease as well. Rather than using a threshold of 0.5 to convert the probabilities into a particular class (i.e., 0 or 1), the maximum profit was determined and then the appropriate probability threshold (that would yield the maximum profit) was used. In every model, the arbitrary value of 0.5 was found to be not as optimal as the calculated threshold.

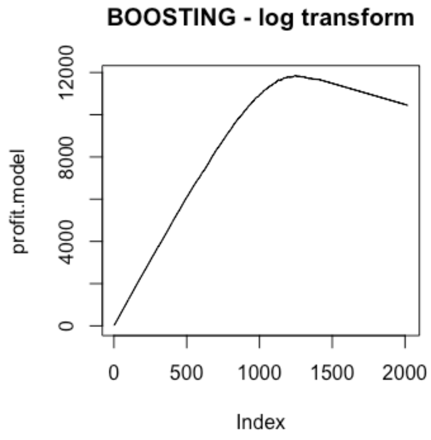


Figure 4: Example plot showing effect of profit vs. number of mailings

Classification Model – Logistic Regression Model

Six logistic models were created: 3 using each dataset described in Table 5. Each model used all predictors initially and then a stepwise automated variable selection (using BIC as the selection criteria) was used to eliminate non-statistically significant predictors. Using Dataset 1, the highest realized profit was about $11852.50 with an error rate of 13.1%. Using Dataset 2, the highest realized profit was about $11848.50 with an error rate of 14.3%.

Classification Model – LDA

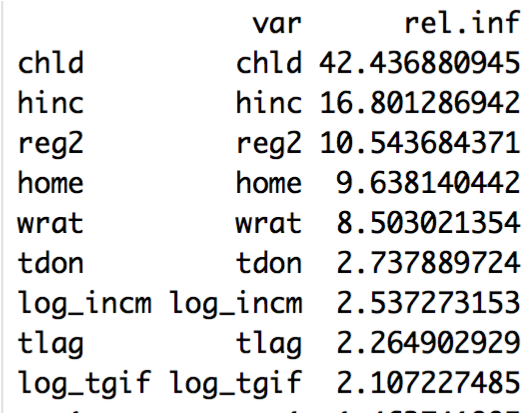
Similar to the logistic regression model, six linear discriminant analysis (LDA) models were constructed. Neither LDA model within each Dataset was superior to the best performing logistic model for each dataset. Using Dataset 1, the best performing LDA model predicted a maximum profit of $11843, used log transformations, and had an error rate of 13.6%. Using Dataset 2, the best performing LDA model predicted a maximum profit of $11368, used log transformations, and had an error rate of 21.8%.

Classification Model – QDA

Similar to the LDA models, six quadratic discriminant analysis (QDA) models were constructed. Neither QDA model within each dataset was superior to the best performing logistic model for each dataset. Using Dataset 1, the best performing QDA model predicted a maximum profit of $11041.50, used log transformations, and had an error rate of 28.5%. Using Dataset 2, the best performing QDA model predicted a maximum profit of $11270.50, used log transformations, and had an error rate of 21.3%. It is interesting to note that the error rate for the QDA model using Dataset 2 had a lower error rate than the QDA model using Dataset1 AND the respective LDA model (which had an error rate of 21.8%). Some of the high error can be attributed to the lack of higher polynomial transformations.

Classification Model – Boosting

All six boosting models, using all 20 predictors in both Datasets, outperformed the logistic regression models. For each boosting model, the relative influence plot and statistics were analyzed. Figure 5 illustrates a sample output of the relative influence statistics and Figure 6 illustrates the relative influence plot. These plots help one understand impactful predictors in the boosted model. Using Dataset 1, the best boosting model projected a maximum profit of $11852.50 with an error rate of 13.1%. Using Dataset 2, the best boosting model projected a maximum profit of $11848.50 with an error rate of 14.3%. Unlike QDA, the boosting models have a lower error rate and is in line with the LDA and logistic regression models thus far. One key advantage of these boosted models is the 10-fold cross validation that took place during their development.



```
              var     rel.inf
chld         chld 42.436880945
hinc         hinc 16.801286942
reg2         reg2 10.543684371
home         home  9.638140442
wrat         wrat  8.503021354
tdon         tdon  2.737889724
log_incm log_incm  2.537273153
tlag         tlag  2.264902929
log_tgif log_tgif  2.107227485
```

Figure 5: Example of relative influence statistics

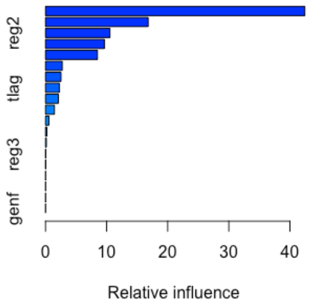


Figure 6: Example relative influence plot

Classification Model – Random Forest

Like the other strategies, six different random forest models were constructed. Unlike the other previous models constructed, the raw data using Dataset 1 (no transformations and no standardization) resulted in a projected profit of $11783.00 with an error rate 14.1%. Using Dataset 2, the log-transformed variables resulted in a projected profit of $11793.50 with an error rate of 15.1%. Compared to the boosting models, the random forest models have a similar error rate, but reduced maximum profit. Figure 7 illustrates an example plot that shows the importance of each variable. This plot was closely analyzed to help provide deeper insight into the impact of each predictor.

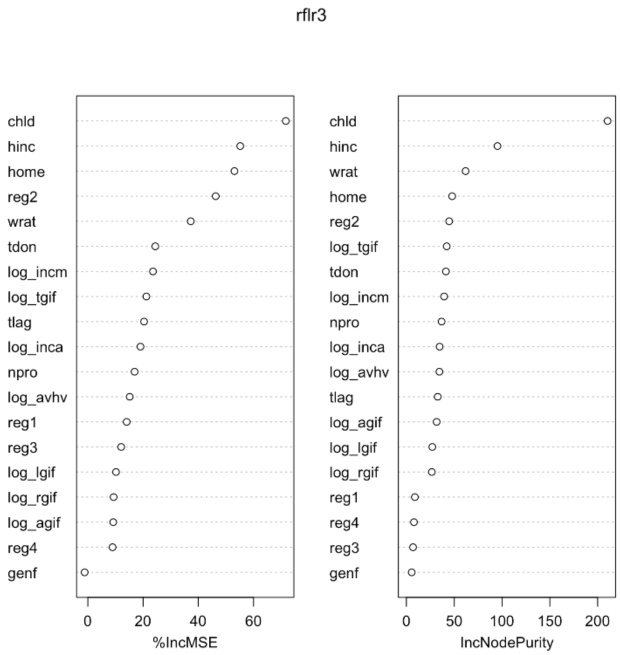


Figure 7: Variable importance plot example

Results of the DONR Classification Models

Recall from the DATA PREPARATION section that there are two different datasets. The key metric that is to be used, per the business requirements of this exercise, is maximum profit. From a pure metric standpoint, the boosting model for Dataset 1 (identifying certain variables as factors[5]) that uses log transformations is the preferred model. However, Dataset 2 considers many of the variables as numeric values (instead of factors) and this results in the model utilizing standardized[6] variables as the preferred model. Only $4 separates these two models, suggesting that both models and approaches may be plausible. Comparing the preferred model using Dataset 1 to the preferred model in Dataset 2, it's clear to see that the Dataset 1 model has a higher maximum profit with a slightly lower number of correct observations (1244 vs. 1275).

The chosen model is highlighted in a green cell in Table 6. Adjusting for the oversampling, this model classifies (using the test dataset) 1703 as non-donors and 304 as donors. Therefore, it can be expected to mail 304 donors with the highest posterior probabilities.

Table 6 summarizes the results of all of the models used in the classification step (response variable DONR) for each dataset.

[5] Recall that CHLD, HINC, and WRAT are considered factors

[6] Standardized variables have a mean of 0 and a standard deviation of 1

Table 6: Classification model results for response variable DONR

		Dataset 1 (with factors)			Dataset 2 (no factors)		
Model	**Transformation**	**Max Profit**	**Error Rate**	**Correct Obs.**	**Max Profit**	**Error Rate**	**Correct Obs.**
logistic	raw	11823.5	0.144	1273	11416	0.216	1397
logistic	standardized	11823.5	0.144	1273	11416	0.216	1397
logistic	log-transform	11848.5	0.143	1275	11362.5	0.222	1402
LDA	raw	11826.5	0.127	1228	11367.5	0.246	1472
LDA	standardized	11826.5	0.127	1228	11367.5	0.246	1472
LDA	log-transform	11843	0.136	1256	11368	0.218	1392
QDA	raw	10996.5	0.291	1527	11238	0.229	1399
QDA	standardized	10996.5	0.291	1527	11238	0.229	1399
QDA	log-transform	11041.5	0.285	1519	11270.5	0.213	1361
Boosting	raw	11850.5	0.132	1245	11846.5	0.143	1276
Boosting	standardized	11850.5	0.132	1245	***11848.5***	***0.143***	***1275***
Boosting	log-transform	***11852.5***	***0.131***	***1244***	11834	0.144	1275
Random Forest	raw	11783	0.141	1257	11732	0.138	1239
Random Forest	standardized	11777.5	0.145	1267	11718.5	0.144	1253
Random Forest	log-transform	11771.5	0.157	1299	11793.5	0.151	1288

Analysis of the DAMT Predictive Models

For the predictive models, seven different types of model were constructed using the two different datasets (identified in Table 5): linear regression, ridge regression, lasso regression, boosting, random forest, principal component regression, and partial least square regression. Table 7 summarizes all the results for each of these models. In this section, succinct highlights for each type of model is provided. In the **Results of the DAMT Predictive Models** section, justification is provided for the model chosen for each response variable. Recall that the key metric in choosing the appropriate model is the average squared prediction error using the validation dataset.

Predictive Model – Linear Regression

Six different linear regression models were constructed and each one underwent the stepwise automatic variable selection. BIC was the key metric for evaluation of variables entering or exiting the linear regression models. The intent of these regression models was to construct a baseline in which to compare the other models. As such, many of the required conditions (such as lack of multicollinearity) were not met. Intuitively, it was found that the linear regression utilizing the log transformations in both datasets yielded the best linear regression models (see Table 7).

Predictive Model – Ridge Regression

For this type of model, the information in the dataframe was converted into matrices in order to make use of the package `glmnet`. Furthermore, cross-validation was used to determine the best lambda value (i.e., the tuning parameter) across a large range of values (from intercept only to least-squares). For both datasets, it was found that the log-transformed data performed the best. However, neither of the best regression models were better than the linear regression models using the same transformed variables. In this type of model, coefficient values are not necessarily 0, but are very close. Hence, all of the predictors were used – regardless of their viability. Figure 8 illustrates a cross-validation plot that helps visualize the optimal tuning parameter, lambda. This value is calculated algorithmically instead of visual estimation. Figure 9 shows an example output of the coefficients. Note how none of the predictor variables are at absolute 0.

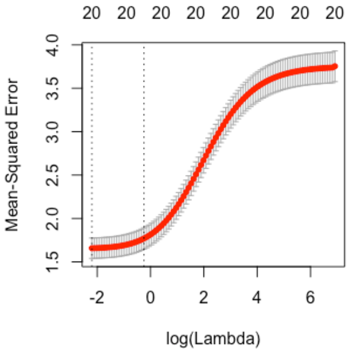
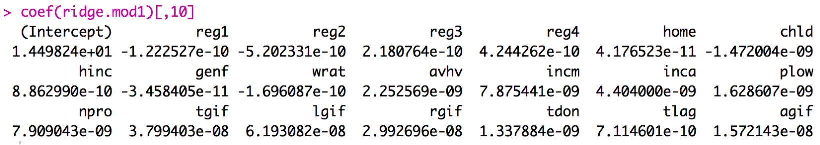


Figure 8: Cross validation plot for lambda

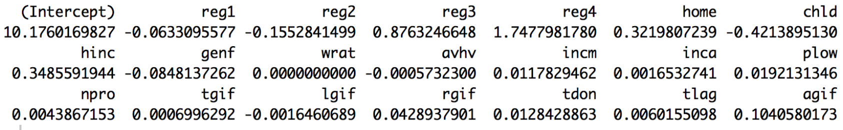


```
> coef(ridge.mod1)[,10]
  (Intercept)          reg1          reg2          reg3          reg4          home          chld 
 1.449824e+01 -1.222527e-10 -5.202331e-10  2.180764e-10  4.244262e-10  4.176523e-11 -1.472004e-09 
         hinc          genf          wrat          avhv          incm          inca          plow 
 8.862990e-10 -3.458405e-11 -1.696087e-10  2.252569e-09  7.875441e-09  4.404000e-09  1.628607e-09 
         npro          tgif          lgif          rgif          tdon          tlag          agif 
 7.909043e-09  3.799403e-08  6.193082e-08  2.992696e-08  1.337884e-09  7.114601e-10  1.572143e-08 
```

Figure 9: Example ridge regression coefficient values

Predictive Model – Lasso Regression

Similar to Ridge Regression, information in the dataframe was converted into matrices in order to make use of the package `glmnet`. Furthermore, cross-validation was used to determine the best lambda value (i.e., the tuning parameter) across a large range of values (from intercept only to least-squares). For both datasets, it was found that the log-transformed data performed the best. However, neither of the best regression models were better than the linear regression models using the same transformed variables. The key difference here was that the lasso regression resulted in coefficients being 0 for some of the predictors. A similar cross-validation plot (see Figure 8) to understand the impact of different lambda values was also constructed for each lasso model. Figure 10 illustrates an example coefficient list for a particular lambda value. In contrast to ridge regression, note how some of the lasso regression model (example) coefficients actually have a value of 0. In a way, lasso models are inherently conducting variable selection.



```
  (Intercept)          reg1          reg2          reg3          reg4          home          chld 
10.1760169827 -0.0633095577 -0.1552841499  0.8763246648  1.7477981780  0.3219807239 -0.4213895130 
         hinc          genf          wrat          avhv          incm          inca          plow 
 0.3485591944 -0.0848137262  0.0000000000 -0.0005732300  0.0117829462  0.0016532741  0.0192131346 
         npro          tgif          lgif          rgif          tdon          tlag          agif 
 0.0043867153  0.0006996292 -0.0016460689  0.0428937901  0.0128428863  0.0060155098  0.1040580173 
```

Figure 10: Example coefficient values for lasso model

Predictive Model – Boosting

A similar approach to the prior boosting models (for classification) was used. The lowest average square prediction error using Dataset 1 was 1.27 with either the raw (untransformed) data or the standardized data. Using Dataset 2, however, the lowest was with the standardized data. Nevertheless, this was quite close to the untransformed data as well. This boosting model, for Dataset 2, was the best performing model (in terms of average prediction error) of any type of model. One key advantage of these boosted models is the 10-fold cross validation that took place during their development. As done for the classification model, plots were also constructed to help understand the relative influences. For brevity, the plots for this specific model are not attached in this report.

Predictive Model – Random Forest

A similar approach to the prior boosting models (for classification) was used. The lowest average square prediction error using Dataset 1 was 1.65 with the standardized data. Using Dataset 2, however, the lowest was with the untransformed data. As done for the classification model, the variable importance plot was also constructed. For brevity, this plot is not attached in this report.

Predictive Model – Principal Component Regression

Six different models were constructed, each with cross validation. Furthermore, a validation plot was constructed to help determine the number of principal components to use for the predictive model (when testing against the validation dataset). This approach helped to ensure that the number of principal components used would result in a low MSEP (mean square error predicted). Each of the models ended up using a different number of principal components. The log-transformed variables led to the lowest average square prediction error for both datasets (see Table 7). Figure 11 is an example of a validation plot that visually shows the number of components and the MSEP. In this example, the greater the number of components, the better. However, such visual analysis is not ideal as it may be misleading. Therefore, a summary output was also done (see Figure 12). Note how 36 components may be sufficient versus 37. In this case, it was decided to go with 37 components.

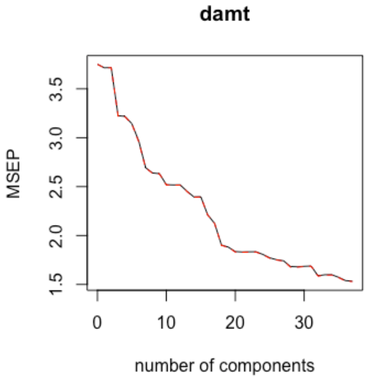
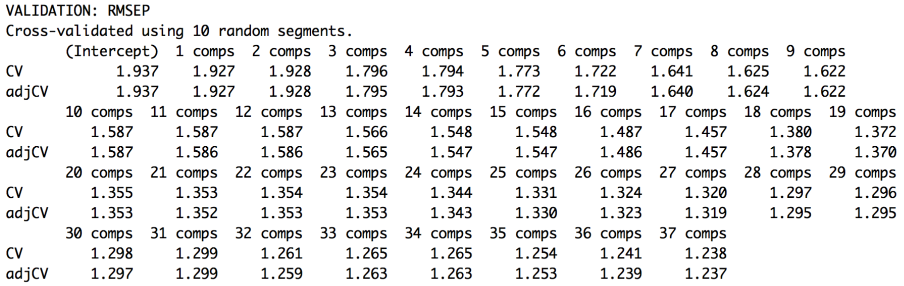


Figure 11: Validation plot for MSEP vs. principal component count



```
VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV           1.937    1.927    1.928    1.796    1.794    1.773    1.722    1.641    1.625    1.622
adjCV        1.937    1.927    1.928    1.795    1.793    1.772    1.719    1.640    1.624    1.622
       10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
CV        1.587     1.587     1.587     1.566     1.548     1.548     1.487     1.457     1.380     1.372
adjCV     1.587     1.586     1.586     1.565     1.547     1.547     1.486     1.457     1.378     1.370
       20 comps  21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps
CV        1.355     1.353     1.354     1.354     1.344     1.331     1.324     1.320     1.297     1.296
adjCV     1.353     1.352     1.353     1.353     1.343     1.330     1.323     1.319     1.295     1.295
       30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
CV        1.298     1.299     1.261     1.265     1.265     1.254     1.241     1.238
adjCV     1.297     1.299     1.259     1.263     1.263     1.253     1.239     1.237
```

Figure 12: Numerical output of cross-validated RMSEP vs. number of components

Predictive Model – Partial Least Squares Regression

Six different models were constructed, each with cross validation. Furthermore, a validation plot was constructed to help determine the number of components to use for the predictive model (when testing against the validation dataset). This approach helped to ensure that the number of components used would result in a low MSEP (mean square error predicted). Each of the models ended up using a different number of components. The log-transformed variables led to the lowest average square prediction error for both datasets (see Table 7). Similar visual plots (see Figure 11 and Figure 12) were constructed for these models as well.

Results of the DAMT Predictive Models

Recall that the key metric to decide the most appropriate model is Mean Square Error Prediction (MSEP). Using this metric, the best model was the linear regression using log transformations on Dataset 1 (see cell highlighted in green on TABLE XX). It is surprising to see that such a simple model outperformed the corresponding boosting models for the same dataset. One of the key advantages of this model is its simpler interpretability compared to the boosting models.

Table 7 summarizes the results of all of the models used in the prediction step (response variable DAMT) for each dataset.

Table 7: Predictive model results for response variable DAMT

		Dataset with factors	**Dataset with numeric**
Model	**Transformation**	**Prediction Error**	**Prediction Error**
Linear	raw	1.643	1.858
Linear	standardized	1.643	1.858
Linear	log-transform	***1.396***	1.629
Ridge Regression	raw	1.958	1.958
Ridge Regression	standardized	1.961	1.961
Ridge Regression	log-transform	1.734	1.734
Lasso	raw	1.877	1.877
Lasso	standardized	1.871	1.871
Lasso	log-transform	1.633	1.633185
Boosting	raw	1.527	1.539435
Boosting	standardized	1.527	***1.539267***
Boosting	log-transform	1.555	1.567842
Random Forest	raw	1.653	1.65574
Random Forest	standardized	1.65	1.668346
Random Forest	log-transform	1.673	1.681108
PCR	raw	1.653	1.866657
PCR	standardized	1.653	1.866657
PCR	log-transform	1.435	1.62172
PLS	raw	1.682	1.866785
PLS	standardized	1.682	1.864312
PLS	log-transform	1.414	1.622855

Conclusion and Next Steps

The boosted model for classifying donors (DONR) outperformed the other models, including the simpler and easier to interpret logistic models. On the other hand, the predictive model for determining the potential donation amount (DAMT) is a simple linear regression model that is simple to understand and outperforms the more complex boosted models. One key takeaway from this exercise is that complex models do not necessarily outperform the simpler models. The models demonstrated here are by no means an exhaustive list. Note how two different datasets (one using factors and another not) resulted in slightly different results. Next steps include (but certainly not limited to) further analysis into interaction terms, outlier management, and other types of variable transformations. This would also provide a more insightful opportunity to recraft the models here for further improvements.