Course Project

PREDICT 422, Winter 2018
Scott Burley

Introduction

In this assignment we are asked to develop machine learning models to improve the cost effectiveness of a charity's direct marketing campaigns to solicit donations. The dataset consists of 8009 observations, each representing a previous donor. We are asked to predict whether the donor will respond to a mailing (DONR), and what amount they will donate if so (DAMT).

The dataset is divided into 3984 training observations, 2018 validation observations, and 2007 test observations (for which we are not given the values of DONR and DAMT). The training and validation sets have been oversampled to have about equal numbers of donors and non-donors. We are told that only about 10% of the test set consists of donors.

Data dictionary:

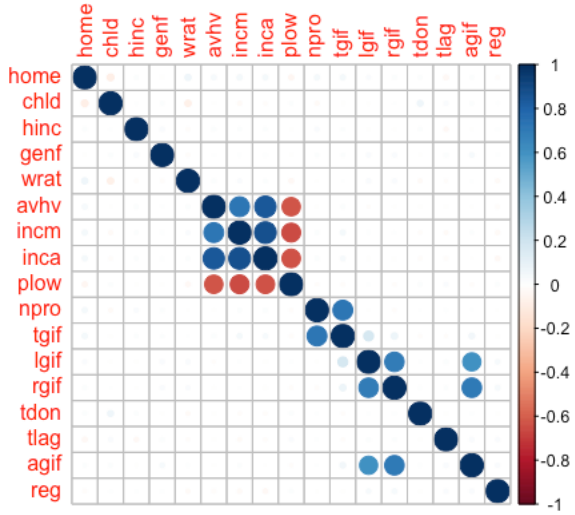
Variable	Description	Note
ID	ID number	
REG	Region	Categorical labels, 0 to 5. Condensed from individual variables (REG1, REG2, etc.)
HOME	Homeowner	True/False
CHLD	Number of children	
HINC	Household income	Categorical labels, 1 to 7
GENF	Female gender	True/False
WRAT	Wealth rating	Categorical labels, 0 to 9, with 9 having the highest wealth. Calculated from neighborhood population statistics, scaled by state.
AVHV	Average home value	By neighborhood
INCM	Median family income	By neighborhood
INCA	Average family income	By neighborhood
PLOW	Percent low income households	By neighborhood
NPRO	Lifetime promotions received to-date	
TGIF	Total lifetime gifts to-date	In dollars
LGIF	Largest gift to-date	In dollars
RGIF	Most-recent gift	In dollars
TDON	Months since last donation	
TLAG	Months between first and second gift	

Variable	Description	Note
AGIF	Average gift amount to-date	In dollars
DONR	Classification target variable	1 = donor, 0 = non-donor
DAMT	Regression target variable	Donation amount in dollars

Data Exploration

Cross correlations

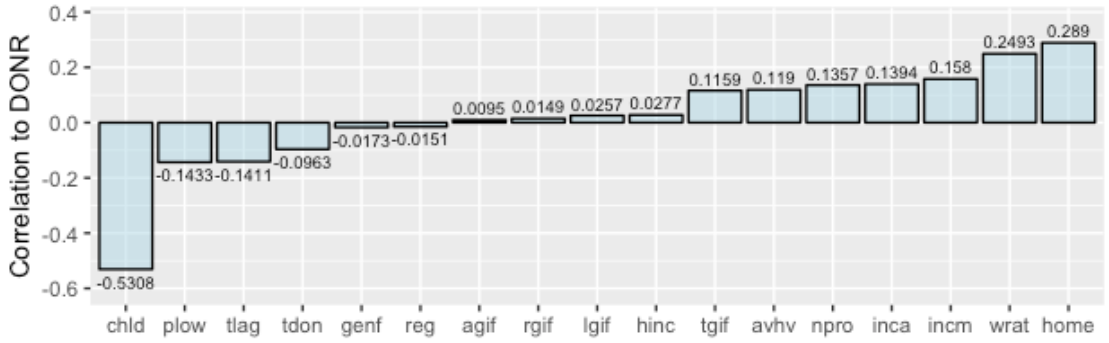
First we'll plot a graphic of the cross correlations between our predictors:



There is an obvious cluster of correlated variables: average home value, median income, average income, and (inversely) percent low income. These make intuitive sense as they all measure the wealth of the donor's neighborhood. In addition, there is a strong relationship between average gift, most-recent gift, and largest gift–most donors probably give about the same amount each time. Finally, there is a strong relationship between number of promotions and total gift.

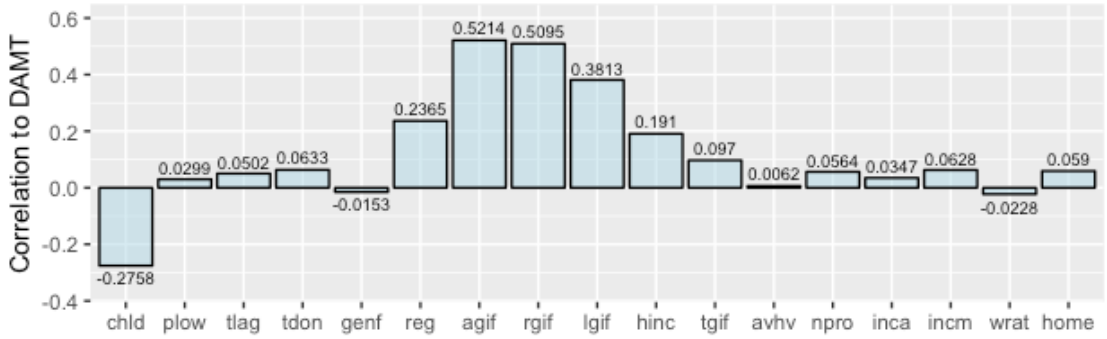
Correlations with targets

Now we'll look at how these predictors relate to our two target variables:



It looks like the strongest single predictor of whether a targeted donor will donate again is the number of children they have–more children translating to a lower probability of donation. Homeowners are more likely to donate, as are those with a higher wealth rating. The cluster of neighborhood wealth variables are also somewhat related to donation probability–donors in wealthy neighborhoods being more likely to donate.

Note that the correlation with region in this plot is meaningless, as the region numbers are presumably arbitrary.

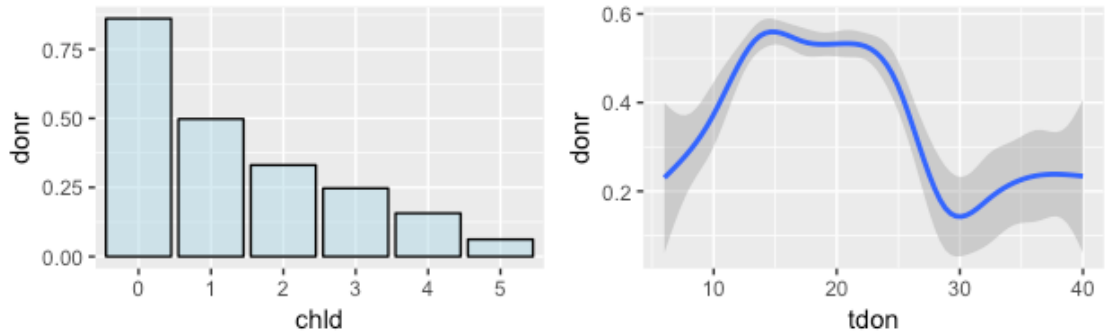


Looking at the amount donated (after filtering out those who did not donate), the three gift-size variables strongly correlate to larger donations. The number of children again seems to be important, with more children translating to smaller donations, though the effect is not as strong as with the DONR variable. Higher household income also correlates to larger donations.

Intriugingly, region seems to have a moderate relationship with donation size. Again, the actual correlation statistic is not meaningful, but it does suggest some difference in donation size between regions.

Predictor distributions

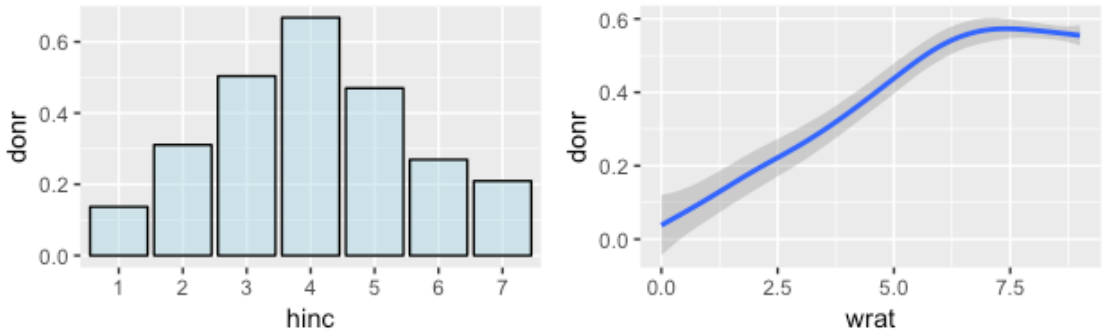
We'll now take a closer look at the distribution of some of the predictors with respect to the two target variables to get a better sense of the relationships between them:



As seen previously, there is a strong negative correlation between the number of children a donor has and the probability of donating again. However, donors with no children seem to be particularly more likely to donate than would be implied by a linear relationship. This also makes intuitive sense–donors with any number of children are probably more alike than donors with none. We'll try to capture this effect be creating a binary variable to indicate childless donors, CHLD0.

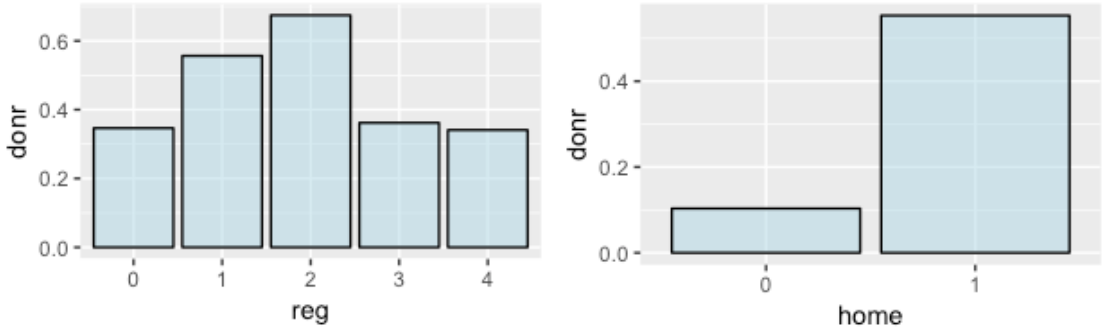
On further inspection, time since last donation shows a strong relationship with donation probability, though with an unusual distribution. Donors are much more likely to donate roughly 12-24 months after the last donation than either earlier or later. But again, this makes sense– donors are unlikely to donate again shortly after their last donation, while those who have not donated in over 2 years are unlikely to do so again.

Time since last donation has too many discrete values to create a dummy variable for each one. Some of the models we will fit later are flexible enough to handle this distribution. For those that aren't, we can create some bins to group donors with similar TDON values. After some testing, we decide to use three bins: 0-11 months, 12-23 months, and 24+ months.

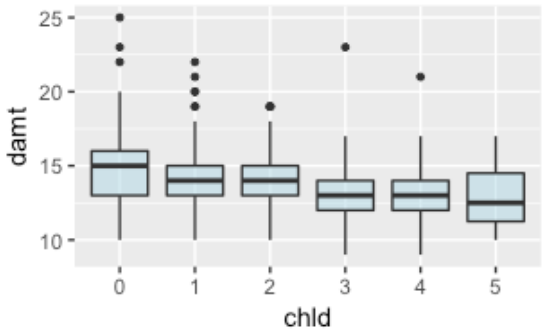


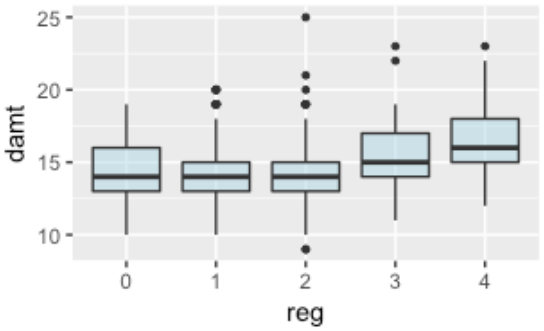
Household income also shows an interesting relationship with DONR that was not obvious from the correlation plot. It looks like middle income households are most likely to donate. Wealth rating has a more conventional positive relationship with DONR, though donation probability seems to taper off with higher ratings.

We can try to capture these distributions by modeling both of the predictors with dummy variables for each level, since HINC has only 7 levels and WRAT only 10.



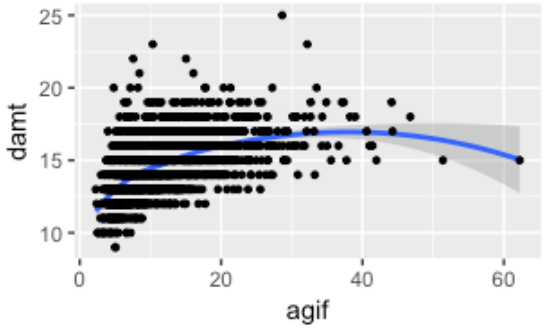
Here, we can confirm that region is an important predictor. Donors in region 2 and especially region 3 are more likely to donate again that in the other three regions. Plotting donation rates by homeownership, we confirm the strong correlation seen earlier–homeowners are more than 5 times as likely to donate.

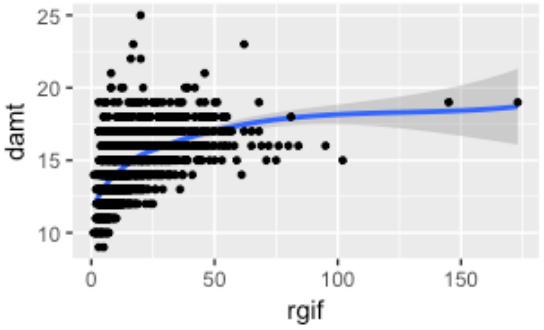


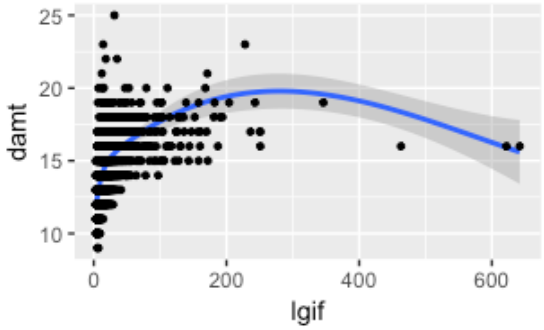


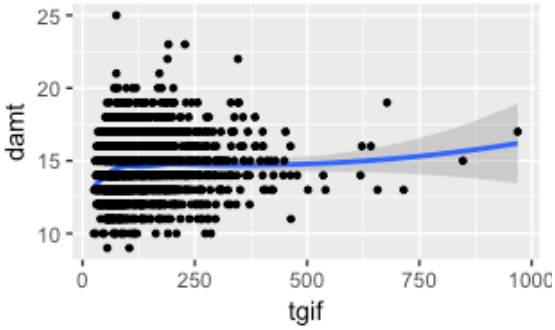
Switching to the DAMT target variable now, we see that there is a weak, but negative association between number of children and donation size. Interestingly, the smallest donations seem to be roughly the same between groups. Donors with few or no children have greater variance due to a minority of large donations among those groups.

Looking at donation size by region, it seems like regions 3 and 4 tend to make larger donations, with regions 1 and 2 making smaller ones. This is interesting, as we saw previously that regions 1 and 2 were those most likely to donate.

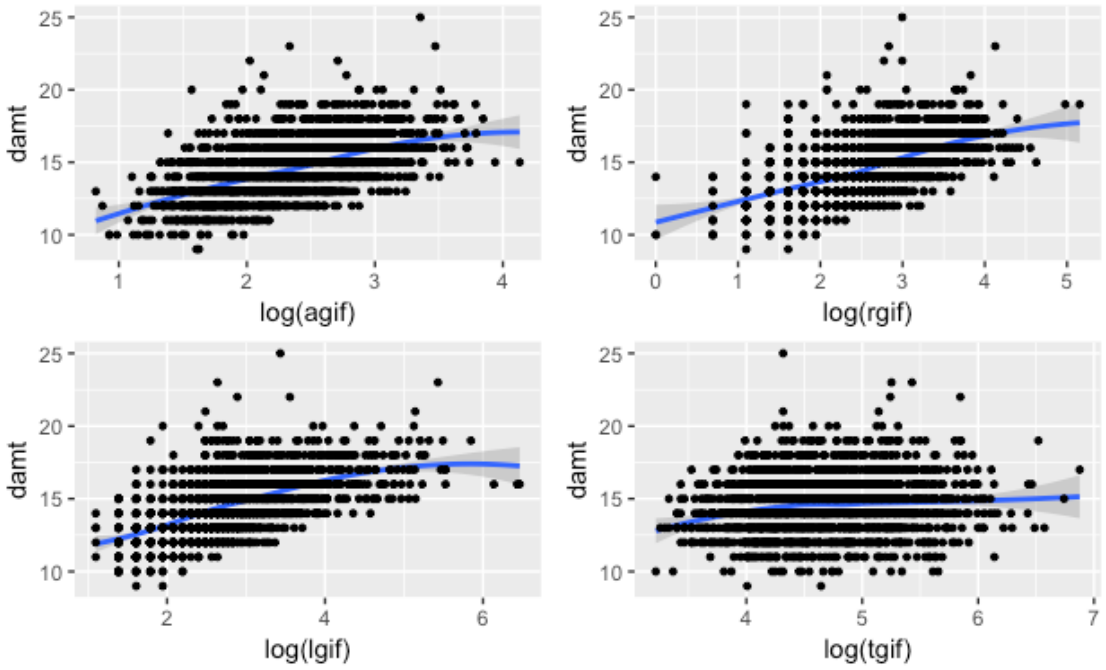








The three gift-size variables have strong, positive correlations to donation amount, while total gifts has a weaker but still positive relationship. However, these relationships appear to be non-linear. We might get better results by taking the natural log of these four variables.



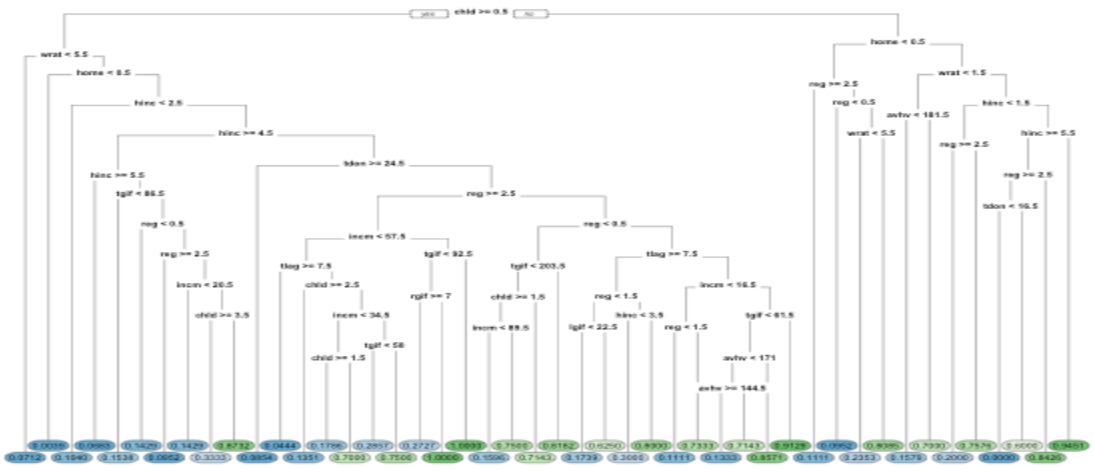
It looks like log transformation produces a more linear relationship for all four of these variables. We will use these transformations for modelling.

Model Fitting - DONR

Now we'll test fit a few different models to predict the probability of donation.

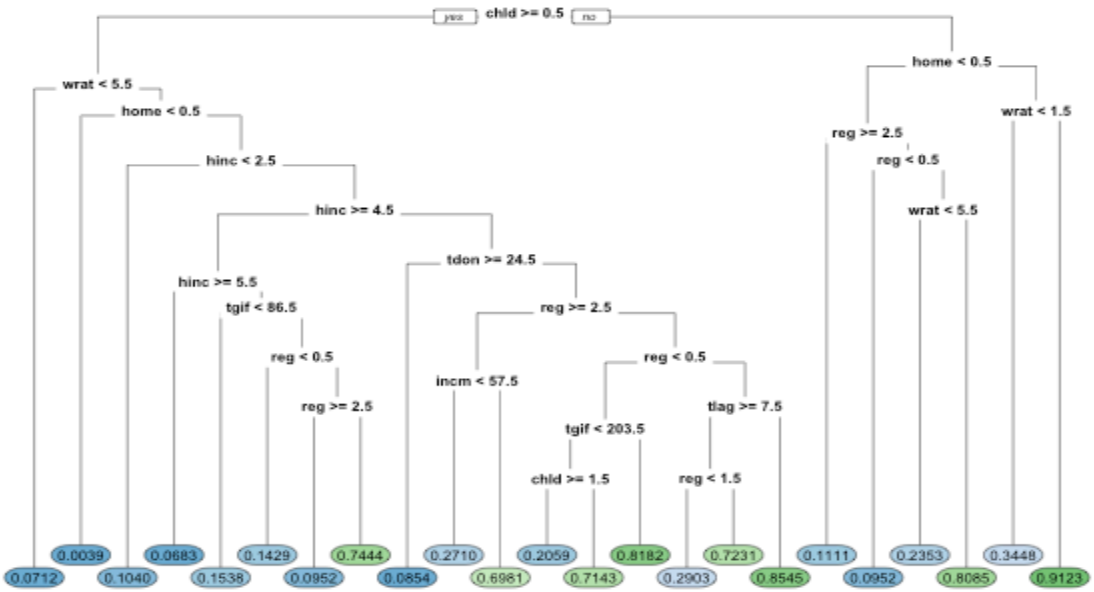
Classification tree

Our first model is a classification tree, with splits made to minimize Gini impurity between the resulting nodes. Classification trees can be useful for finding relationships between predictors that might otherwise not be obvious.



The initial tree is quite large and unwieldy, with 46 nodes. To prune it, we calculate the 10-fold cross-validated accuracy of possible sub-trees.

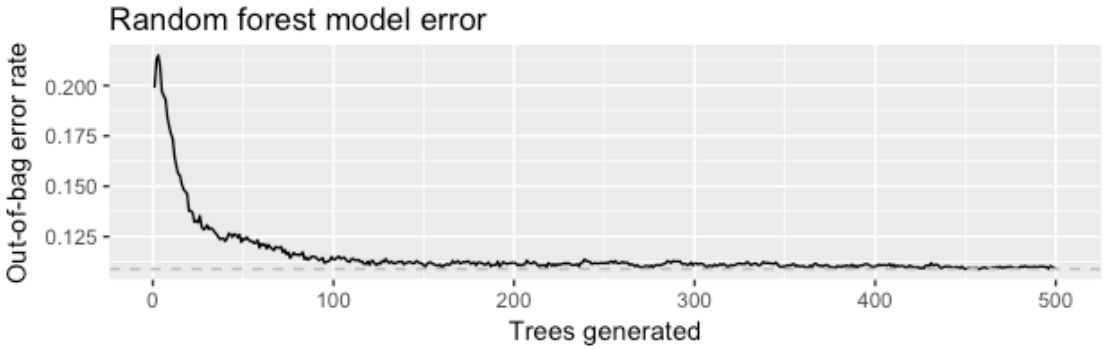
The minimum cross-validated error is found with a tree having 32 nodes. Common practice for pruning trees is to use the smallest tree with cross-validated error within one standard error of the minimum. This is the 11th tree, which has 23 nodes.



The smaller tree is somewhat more manageable to interpret, which is the strength of tree models. It includes no obvious surprises, such as variables which predict different outcomes for different subgroups.

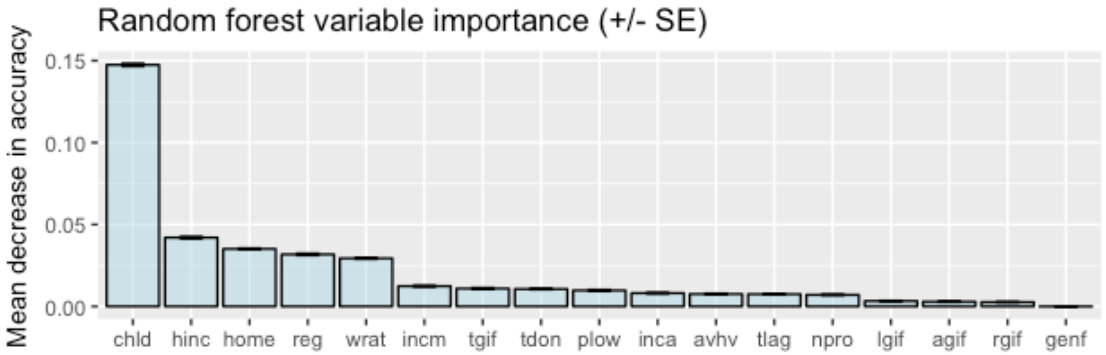
Random forest

Next, we'll generate a random forest model, which aggregates many tree models fitted to random subsets of the data. Since each component tree leaves out some observations ("out-of-bag" observations), we can measure its accuracy on that data and then aggregate to determine an overall error rate.



It looks like out-of-bag error does not diminish much after the first hundred trees or so. Therefore, the 500-tree model we've built is likely sufficient.

Random forest models are difficult to interpret "black boxes", but we can still measure how important each variable is to the model by refitting the model without it, and seeing how much the out-of-bag error increases.



Again, there are no surprises. It looks like the number of children a donor has is the single most important predictor, followed by household income, homeownership, region, and wealth rating. The four neighborhood wealth variables likely have low scores for importance due to their correlation with each other, which means that removing any one doesn't increase error by much.

Logistic regression

Next, we'll fit a logistic regression model. Since this is a linear model, which is less flexible than the tree-based models we've fit so far, we'll need to make sure to use the variable transformations discussed in the previous section. Namely:

- Log-transforming the donation size variables: TGIF, LGIF, RGIF, and AGIF
- Creating a dummy variable for donors with no children: CHLD0
- Binning the time since last donation variable, creating two dummy variables for those with high or low values of TDON: TDONH and TDONL

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.420e+01  1.951e+00 -12.400  < 2e-16 ***
## home         5.494e+00  3.506e-01  15.670  < 2e-16 ***
## chld        -9.594e-01  8.505e-02 -11.281  < 2e-16 ***
## hinc2        2.167e+00  4.608e-01   4.702 2.58e-06 ***
## hinc3        4.122e+00  4.678e-01   8.813  < 2e-16 ***
## hinc4        5.937e+00  4.556e-01  13.031  < 2e-16 ***
## hinc5        3.820e+00  4.549e-01   8.398  < 2e-16 ***
## hinc6        1.924e+00  5.035e-01   3.822 0.000133 ***
## hinc7        7.857e-02  5.363e-01   0.147 0.883520
## genf        -1.252e-01  1.296e-01  -0.966 0.333988
## wrat1        1.387e+00  1.107e+00   1.253 0.210138
## wrat2        3.228e+00  9.218e-01   3.502 0.000462 ***
## wrat3        2.439e+00  9.144e-01   2.667 0.007643 **
## wrat4        4.804e+00  8.769e-01   5.478 4.30e-08 ***
## wrat5        4.784e+00  8.765e-01   5.458 4.80e-08 ***
## wrat6        6.896e+00  8.703e-01   7.924 2.30e-15 ***
## wrat7        6.624e+00  8.777e-01   7.547 4.47e-14 ***
## wrat8        6.776e+00  8.443e-01   8.025 1.01e-15 ***
## wrat9        6.627e+00  8.459e-01   7.834 4.73e-15 ***
## avhv         4.931e-04  1.500e-03   0.329 0.742358
## incm         1.174e+00  2.557e-01   4.592 4.39e-06 ***
## inca         8.421e-02  3.724e-01   0.226 0.821084
## plow        -1.365e-02  8.836e-03  -1.545 0.122356
## npro         8.533e-03  4.371e-03   1.952 0.050904 .
## tgif         8.099e-01  2.296e-01   3.527 0.000420 ***
## lgif        -1.666e-01  1.853e-01  -0.899 0.368695
## rgif        -9.509e-02  1.747e-01  -0.544 0.586199
## tlag        -1.992e-01  1.981e-02 -10.054  < 2e-16 ***
## agif         3.392e-01  2.275e-01   1.491 0.136030
## reg1         1.855e+00  2.102e-01   8.825  < 2e-16 ***
## reg2         3.593e+00  2.136e-01  16.821  < 2e-16 ***
## reg3        -1.783e-01  2.431e-01  -0.733 0.463369
## reg4        -1.375e-01  2.386e-01  -0.577 0.564259
## chld0TRUE    3.750e+00  2.652e-01  14.139  < 2e-16 ***
## tdonLTRUE   -1.904e+00  3.293e-01  -5.783 7.35e-09 ***
## tdonHTRUE   -1.773e+00  2.143e-01  -8.273  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## AIC: 1729.9
```

This is a promising looking model, with many significant coefficients. It also performs well on the validation data, with a 90% accuracy rate:

```
##          actual
## predicted   0   1
##         0 895  84
##         1 124 915
```

This model still contains a few predictors that don't seem to be significant. We'll try to reduce it using a stepwise algorithm to minimize BIC (which tends to result in smaller models than AIC):

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -26.22833    1.48630 -17.647  < 2e-16 ***
## home          5.46144    0.34812  15.689  < 2e-16 ***
## chld         -0.94466    0.08453 -11.175  < 2e-16 ***
## hinc2         2.14309    0.46419   4.617 3.90e-06 ***
## hinc3         4.09249    0.47101   8.689  < 2e-16 ***
## hinc4         5.89560    0.45868  12.853  < 2e-16 ***
## hinc5         3.81490    0.45843   8.322  < 2e-16 ***
## hinc6         1.91102    0.50379   3.793 0.000149 ***
## hinc7         0.09757    0.53944   0.181 0.856472
## wrat1         1.26881    1.08928   1.165 0.244095
## wrat2         3.13664    0.91096   3.443 0.000575 ***
## wrat3         2.39424    0.89816   2.666 0.007682 **
## wrat4         4.66050    0.86207   5.406 6.44e-08 ***
## wrat5         4.64952    0.86152   5.397 6.78e-08 ***
## wrat6         6.77759    0.85599   7.918 2.42e-15 ***
## wrat7         6.49149    0.86262   7.525 5.26e-14 ***
## wrat8         6.64912    0.82960   8.015 1.10e-15 ***
## wrat9         6.49242    0.83091   7.814 5.55e-15 ***
## incm          1.52411    0.13219  11.529  < 2e-16 ***
## tgif          1.17806    0.11534  10.214  < 2e-16 ***
## tlag         -0.19644    0.01964 -10.004  < 2e-16 ***
## reg1          1.85880    0.20927   8.882  < 2e-16 ***
## reg2          3.55410    0.21183  16.778  < 2e-16 ***
## reg3         -0.17468    0.24207  -0.722 0.470526
## reg4         -0.13207    0.23780  -0.555 0.578649
## chld0TRUE     3.72660    0.26374  14.130  < 2e-16 ***
## tdonLTRUE    -1.84604    0.32568  -5.668 1.44e-08 ***
## tdonHTRUE    -1.74889    0.21200  -8.250  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## AIC: 1726.1
```

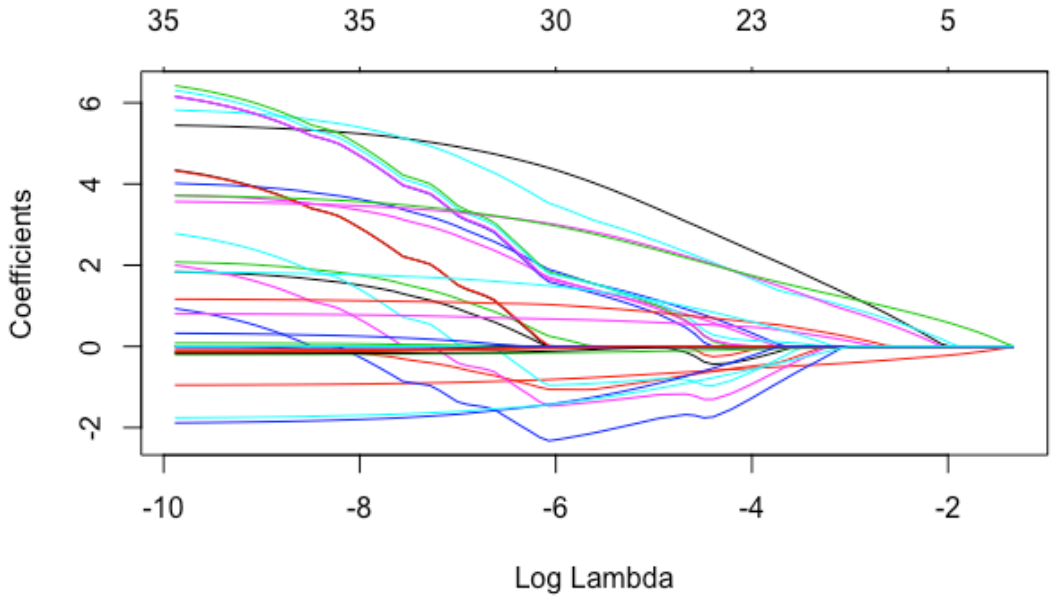
The stepwise algorithm removed 8 variables: average income, average home value, and percent low income (all strongly correlated to median income, which was kept); average gift, largest gift, and most recent gift (strongly correlated with each other, but no clear relationship to DONR); gender (no clear relationship); and number of promotions (very weak relationship).

This reduced model shows similarly strong performance on the validation data:

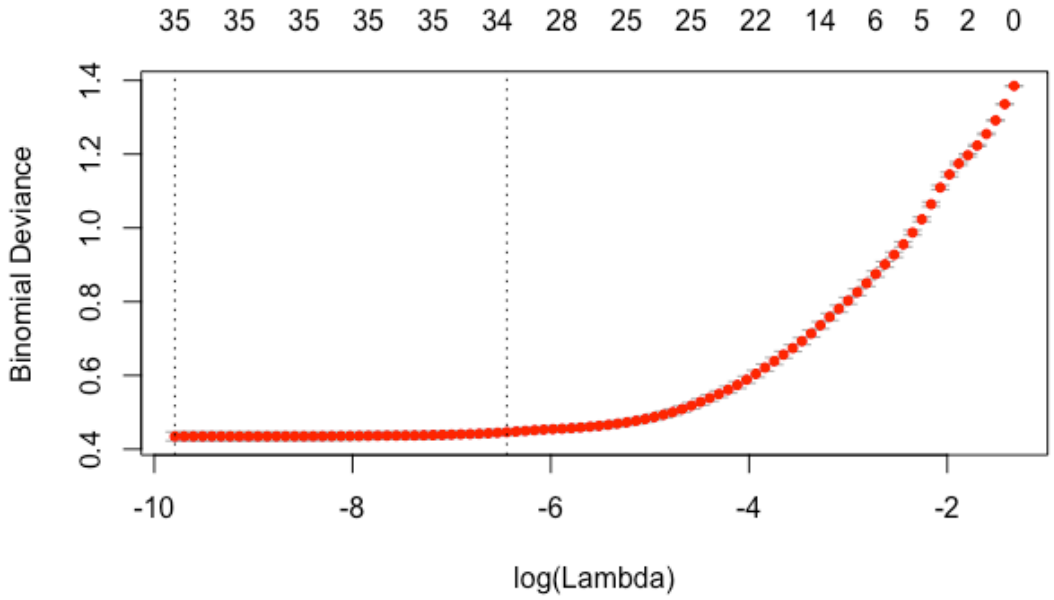
```
##          actual
## predicted   0   1
##         0 891  81
##         1 128 918
```

LASSO

We can also try reducing the number of predictors by using a LASSO model, which shrinks each coefficient by a paramater lambda, potentially shrinking some to zero (ie, removing them from the model). This plot gives the shrunken coefficients for different values of lambda:



To choose the best value for lambda, we use 10-fold cross validation to measure the model error at various points:

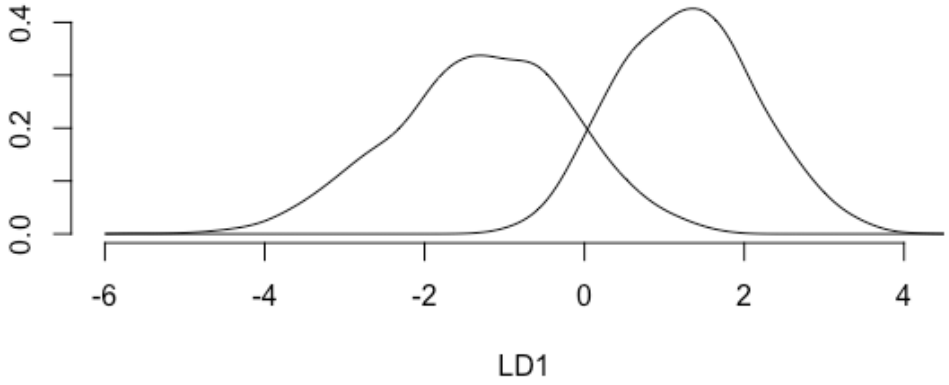


The lowest error is found with a lambda value very close to zero, indicated by the left dashed line on the plot above; essentially this is the full model. However, as with tree models a common practice is to use the largest lambda value within one standard error of the minimum, indicated the the rightmost dashed line.

Unfortunately, this removes only one variable (LGIF), so it's unlikely that our LASSO model will perform much differently from the logistic model we fit earlier.

Linear discriminant analysis

Logistic regression can sometimes perform poorly on test data when the two classes (donors and non-donors, in this case) are well separated. We should try fitting a model to see how it does on the validation data.



Performance on the validation data seems good, but slightly worse than the logistic model we fit previously.

```
##          actual
## predicted   0   1
##         0 872  65
##         1 147 934
```

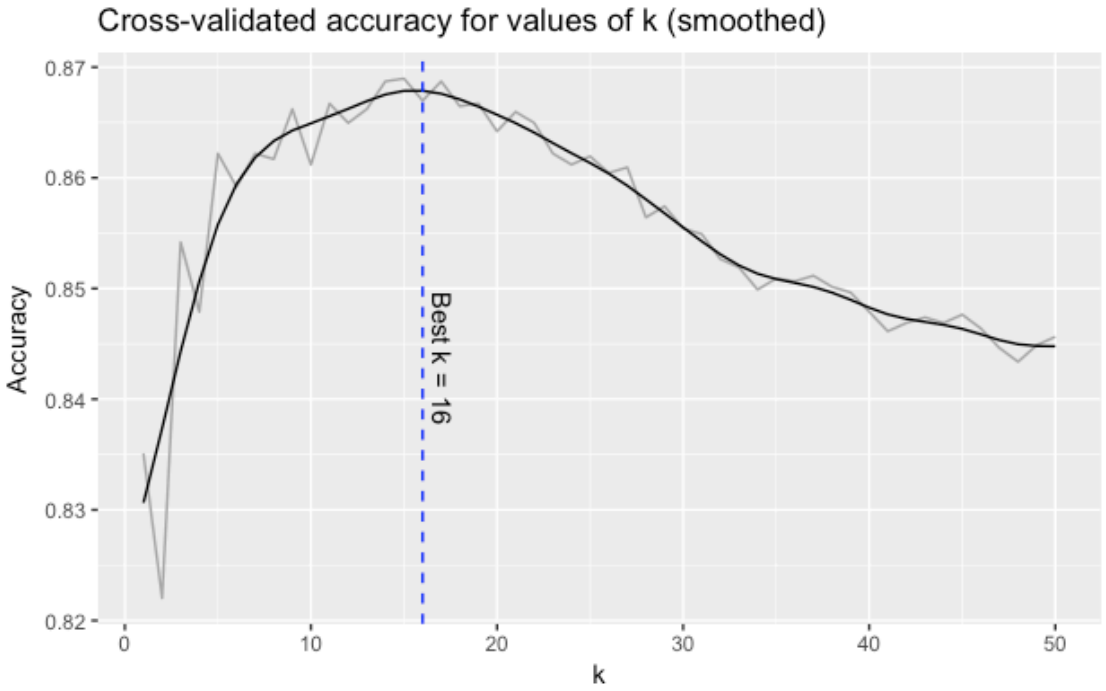
k-nearest neighbors

k-NN is an extremely flexible model for classification. However, it tends to perform poorly with large numbers of predictors, as each observation will have few close "neighbors" across the resulting many-dimensional hyperplane. To reduce the dimenstionality of our data, we will remove the 8 predictors that were de-selected by the stepwise algorithm in our logistic model. This leaves us with 11 predictors.

Still, the results on the validation data are not too promising for k = 3:

```
##           actual
## predicted   0   1
##         0 791  76
##         1 228 923
```

Perhaps k = 3 isn't the optimal value for k? We can use cross validation to try to find a better one:



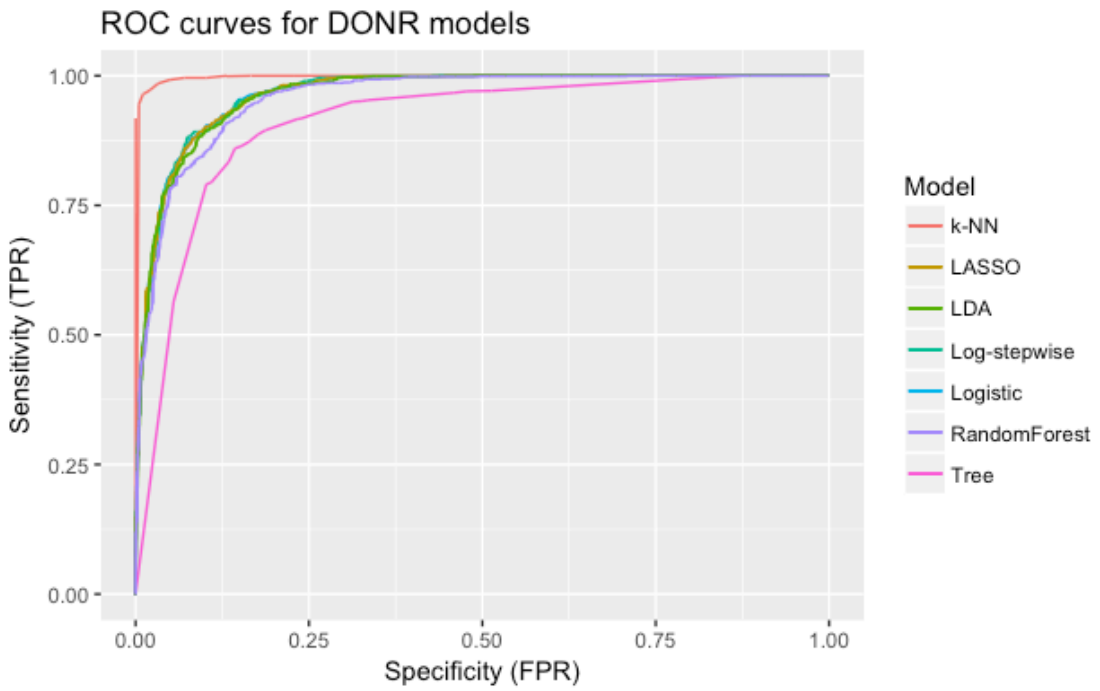
Applying this new value to the validation data produces markedly better results:

```
##     act
## pred   0   1
##    0 959   5
##    1  60 994
```

Validation - DONR

To assess the models we have fit, we'll generate predicted probabilities for DONR for each observation in the validation set. This is straightforward for most of the models, as they inherently generate class probabilities as their output. For the k-NN model, we'll use the proportion of the 16 nearest neighbors in class 1 (donated) as an ersatz predicted probability.

Once we've generated predicted probabilities for every observation, we can generate ROC curves to get a sense of how each model trades off sensitivity for specificity.



It looks like most of the models we've fit perform very similarly to each other. The tree model is clearly inferior, while the k-NN model does great. Calculating the area under each curve confirms that k-NN beats out the others:

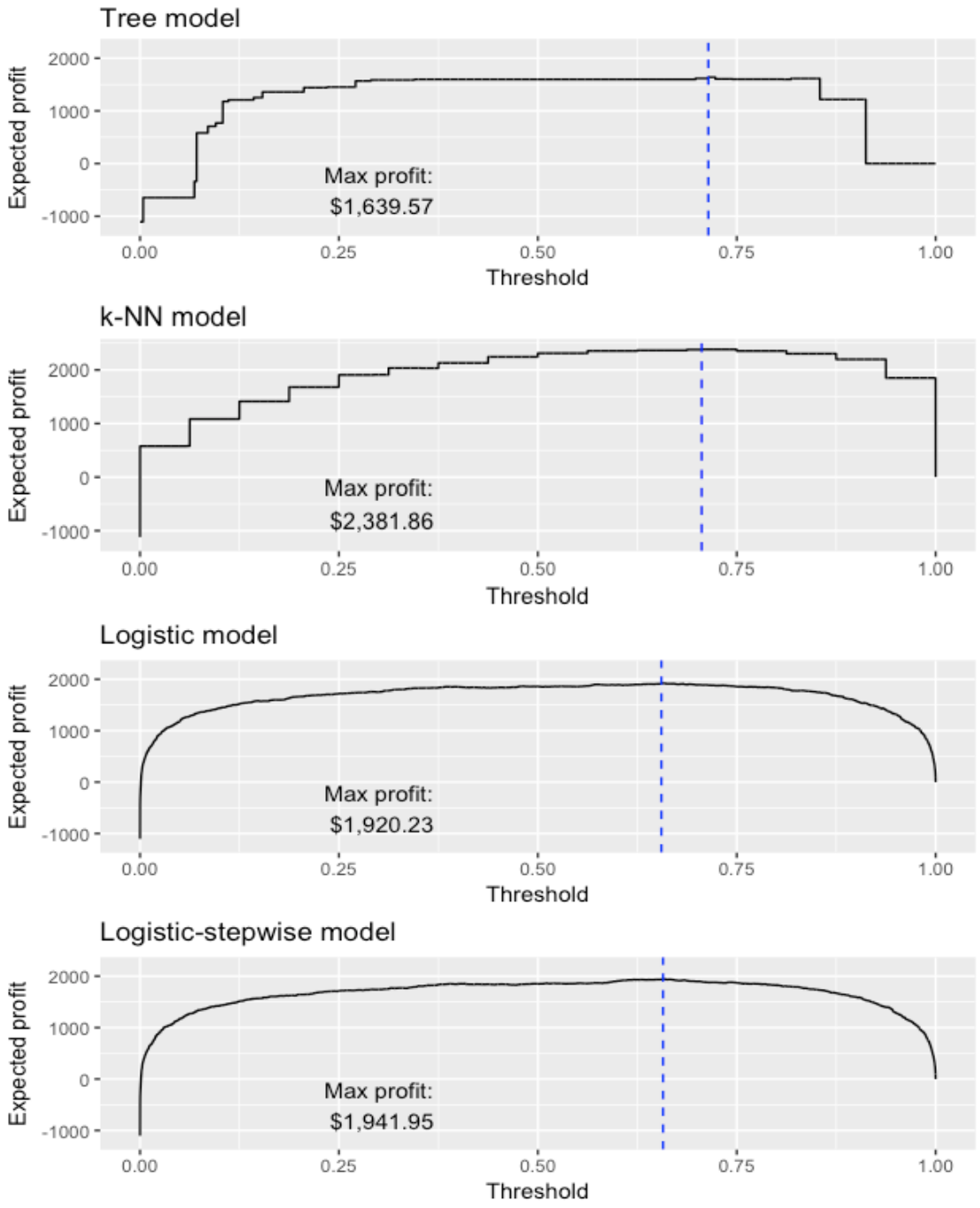
However, keep in mind that we are given an explicit profit function for evaluating our classifications. A mailing costs $2, while donations are assumed to be $14.50. Therefore, a true positive is worth $12.50 and a false positive is worth -$2. Since a true positive is worth several times more than a false positive costs, we should care more about increasing our true positive rate, and not so much about decreasing the false positive rate.

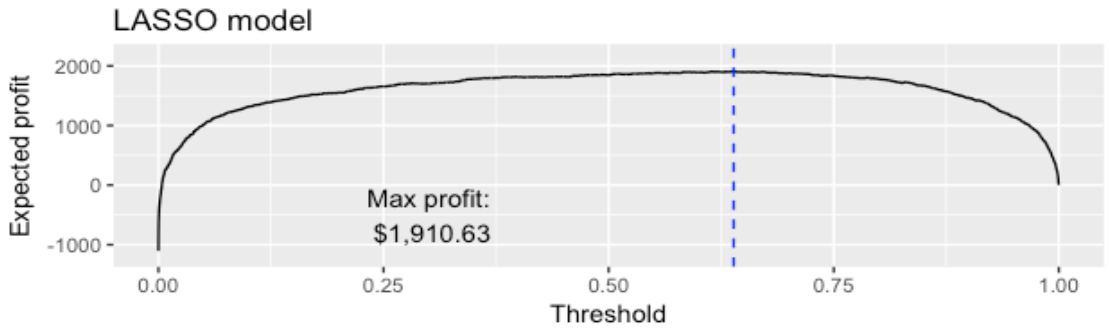
On the other hand, we are told that the training and validation sets have been oversampled to have equal numbers of donors and non-donors–only 10% of the test set consists of donors. Since there are so many more negatives than positives in the test data, keeping the false positive rate down becomes much more important.

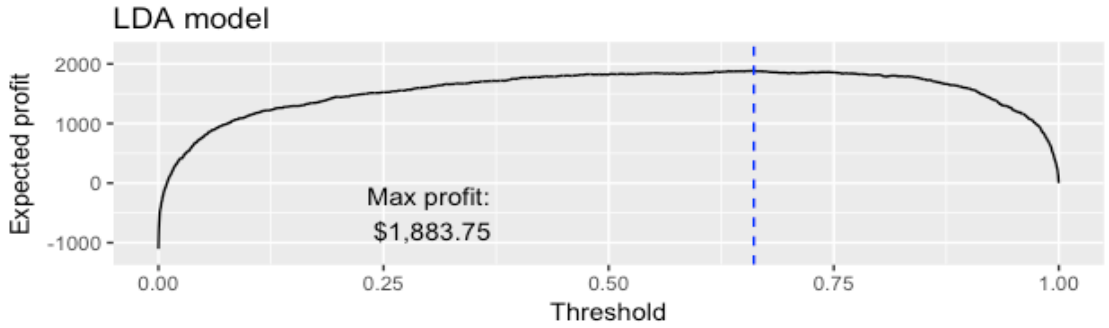
To determine the best performing model, we'll generate profit curves for each model using this profit function:

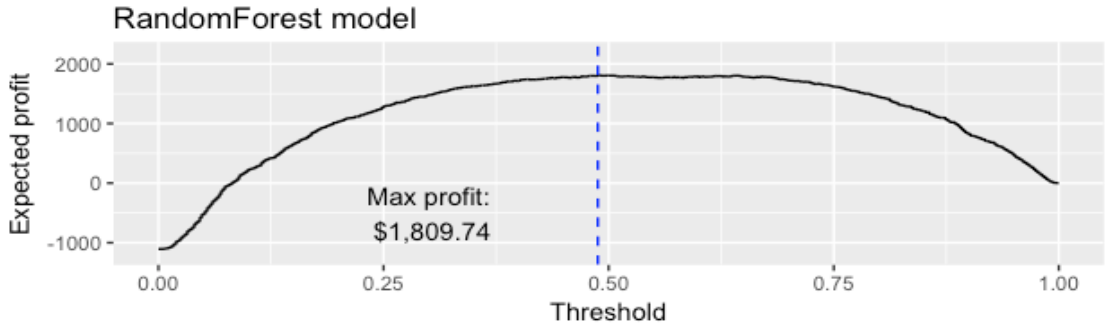
$$Profit(T) = 12.5 \times PR \times TPR_T - 2 \times (1 - PR) \times FPR_T$$

where PR is the positive rate in the test set (asssumed to be 10%), and T is a threshold variable. All previous donors with a predicted probability greater than T will receive a mailing, resulting in the true and false positive rates TPR_T and FPR_T, respectively. We calculate the profit function for all values of T between 0 and 1, inclusive, at invetervals of 0.0001, then multiply by the number of observations in the test set:









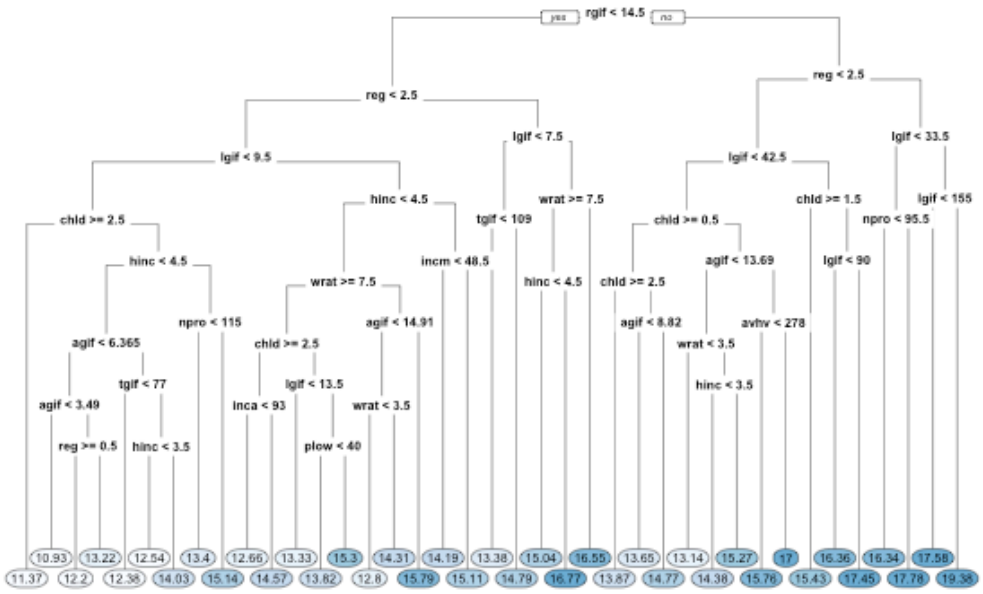
It looks lik k-NN is indeed the best model for maximizing profit. On a dataset the size of the test set with a 10% positive rate, we expect a profit of $2381.86 using a threshold of 0.7059 (ie, 12 or more of the 16 nearest neighbors must be donors). This is 94.94% of the theoretical maximum profit of $2508.75.

Model Fitting - DAMT

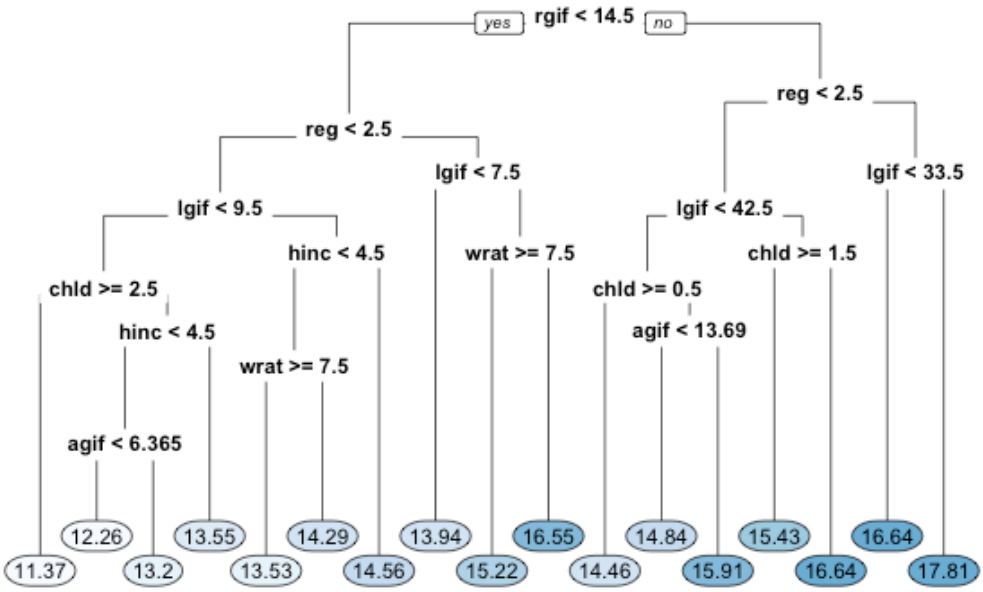
Next, we turn our attention to fitting models for our second target variable, donation amount.

Regression tree

As with DONR, we'll start by creating a regression tree for DAMT to see if there are any unusual relationships that we may have missed during exploratory analysis.



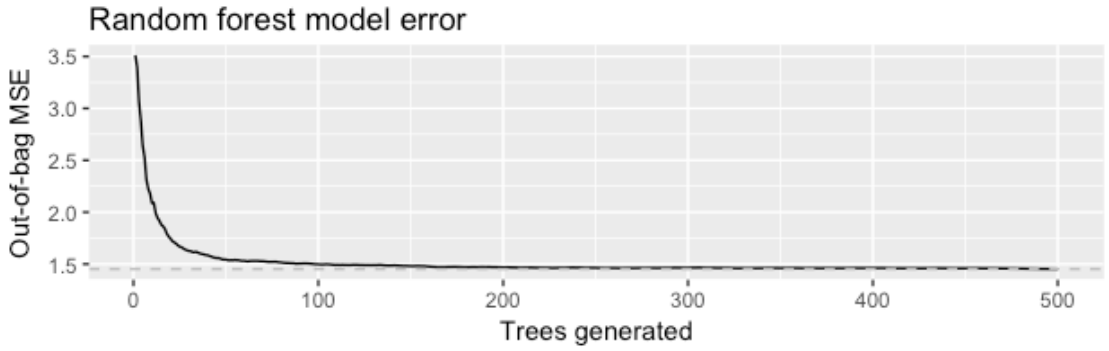
The lowest cross validated error is found with a subtree having 37 nodes. The smallest sub-tree within one standard error of the minimum error has 17 nodes, so we'll prune our tree to that size.



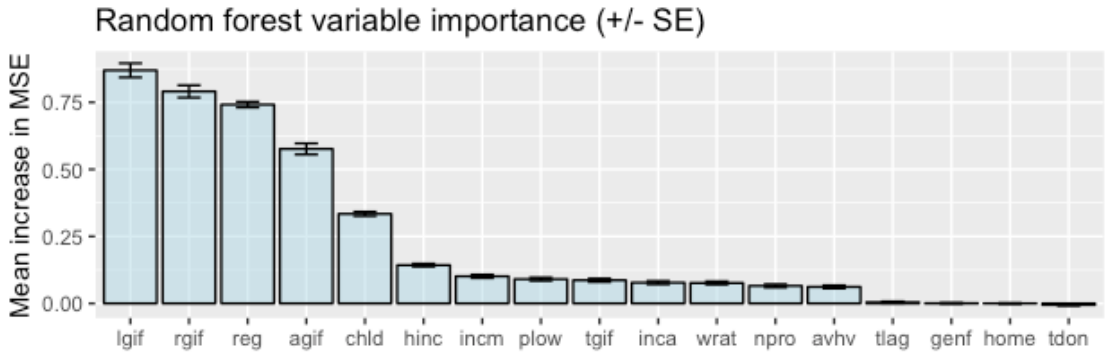
We find no unusual relationships in this tree–all predictors are behaving as expected.

Random forest

Next, we'll try fitting a random forest model.



Out-of-bag error seems to stop decreasing around 150 or 200 trees, so our 500-tree model is probably sufficient.



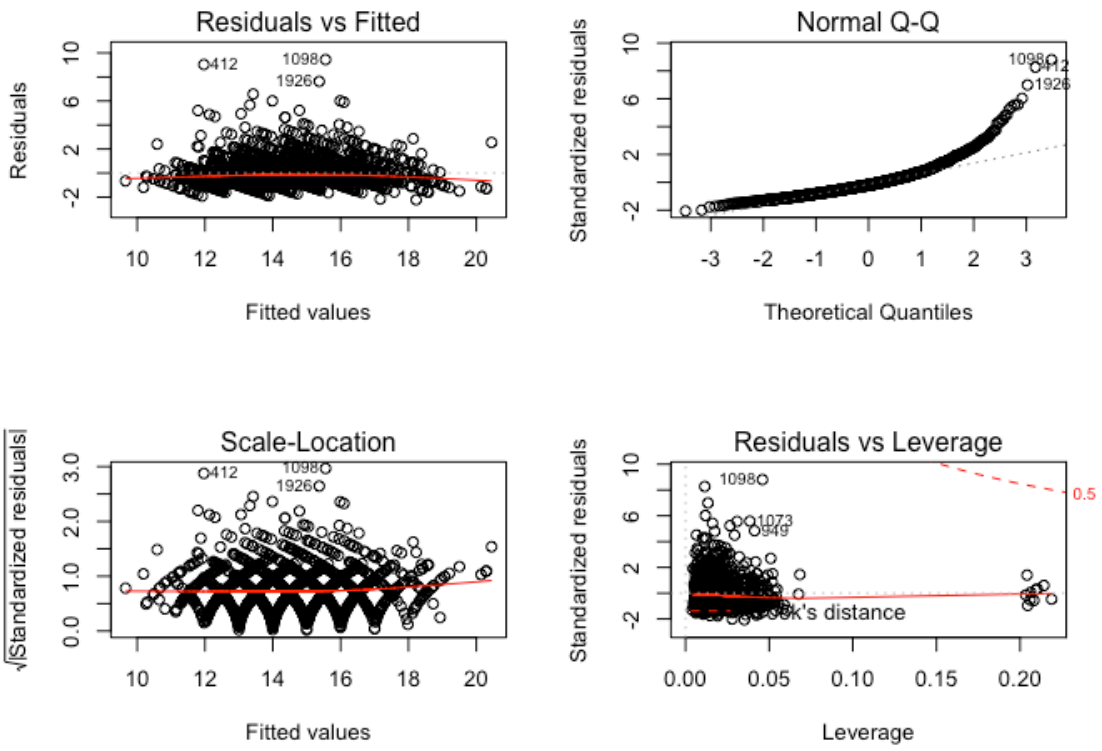
There are no big surprises for variable importance, though it does look like region is quite important. Recall that we found a modest correlation between region and donation amount, though the actual statistic was not meaningful due to the arbitrary labeling of regions. Household income turns out to be somewhat less important than first suspected, probably because its explanatory power is duplicated by the three gift-size variables (largest gift, most-recent gift, and average gift).

Interestingly, our correlation analysis found that average gift and most-recent gift were the two most strongly related of those three variables, with largest gift being of secondary importance. Here, we see that either largest gift or most-recent gift are more important than average gift.

Ordinary least squares

For predicting continuous variables such as DAMT, ordinary least squares regression should be the go-to choice as it's quick to calculate and often quite accurate. Here, we fit an OLS model to DAMT.

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0050017  0.8291608  -0.006 0.995188
## home         0.8351858  0.1657869   5.038 5.14e-07 ***
## chld        -0.4436619  0.0459298  -9.660  < 2e-16 ***
## hinc2        0.7215098  0.2247822   3.210 0.001350 **
## hinc3        1.0575356  0.2204225   4.798 1.73e-06 ***
## hinc4        1.5342030  0.2108819   7.275 4.98e-13 ***
## hinc5        1.9473722  0.2172090   8.965  < 2e-16 ***
## hinc6        2.0002473  0.2435839   8.212 3.91e-16 ***
## hinc7        2.2637672  0.2617047   8.650  < 2e-16 ***
## genf        -0.1125285  0.0504041  -2.233 0.025693 *
## wrat1       -0.3804257  0.6979337  -0.545 0.585764
## wrat2        1.5264797  0.5375980   2.839 0.004566 **
## wrat3        1.1383516  0.5304598   2.146 0.031998 *
## wrat4        2.1663641  0.5101799   4.246 2.28e-05 ***
## wrat5        2.2452034  0.5078896   4.421 1.04e-05 ***
## wrat6        3.1409636  0.5017768   6.260 4.73e-10 ***
## wrat7        3.0538483  0.5019871   6.084 1.41e-09 ***
## wrat8        2.1515128  0.4950132   4.346 1.45e-05 ***
## wrat9        2.1927360  0.4953378   4.427 1.01e-05 ***
## avhv         0.0004013  0.0005435   0.738 0.460397
## incm         0.8288312  0.1005324   8.244 3.00e-16 ***
## inca         0.0434867  0.1429500   0.304 0.761001
## plow         0.0309108  0.0035693   8.660  < 2e-16 ***
## npro         0.0009754  0.0016435   0.593 0.552918
## tgif         0.3088345  0.0891437   3.464 0.000543 ***
## lgif         0.5173235  0.0740505   6.986 3.86e-12 ***
## rgif         0.6785897  0.0682547   9.942  < 2e-16 ***
## tlag         0.0071380  0.0081333   0.878 0.380251
## agif         0.7261174  0.0888020   8.177 5.18e-16 ***
## reg1        -0.0383765  0.0851168  -0.451 0.652134
## reg2        -0.1081606  0.0804986  -1.344 0.179223
## reg3         0.9658313  0.1062090   9.094  < 2e-16 ***
## reg4         1.8558118  0.1049495  17.683  < 2e-16 ***
## chld0TRUE    0.0924065  0.1097531   0.842 0.399919
## tdonLTRUE    0.1007923  0.1418199   0.711 0.477351
## tdonHTRUE   -0.0316587  0.0934757  -0.339 0.734885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6783
## F-statistic: 121.1 on 35 and 1959 DF,  p-value: < 2.2e-16
```



This model has a reasonably strong fit (adjusted R^2 of 0.68) and most of the residual plots look good; the residuals are unbiased, homoskedastic, and have no concerning outliers. However, they do seem to be right skewed. This is not too surprising, as the data itself is right skewed (having a zero lower bound and several outliers to the upside). This means we cannot generate reliable prediction intervals, but this isn't a problem as we are mainly interested in the predictions themselves.

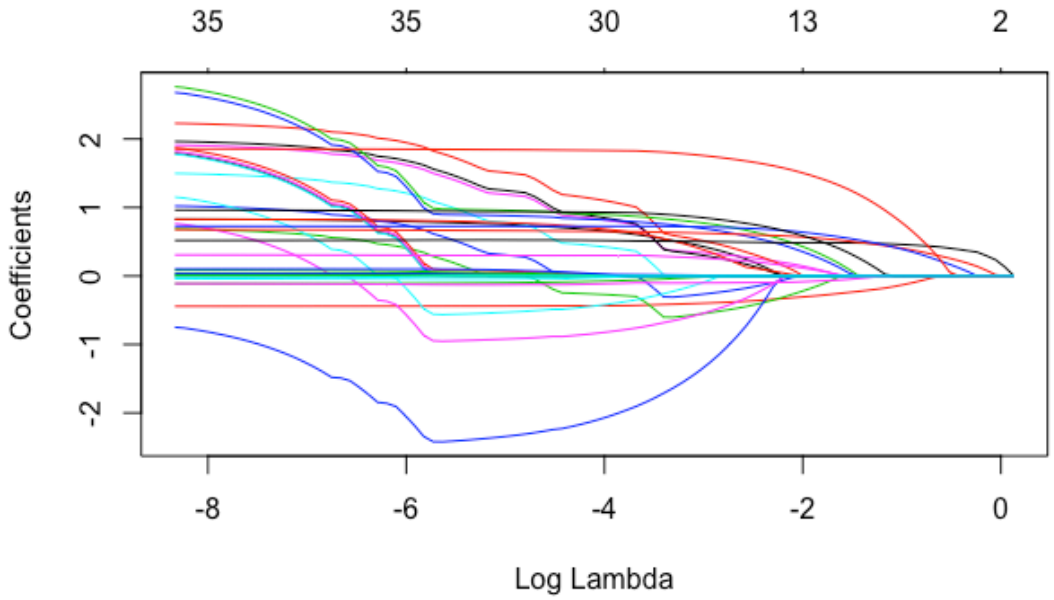
We'll also try using a stepwise algorithm to pare down the model:

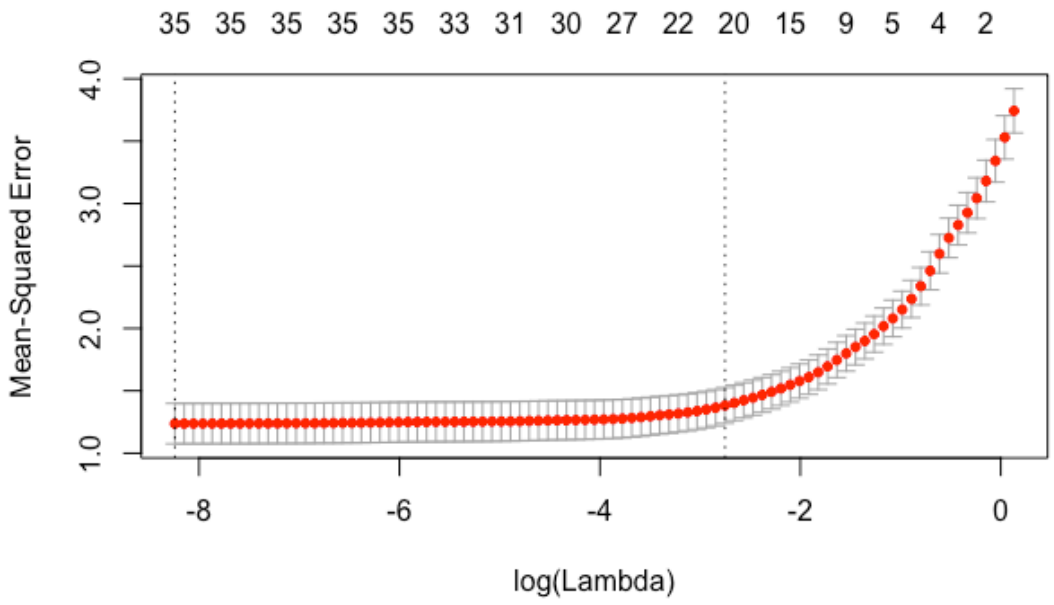
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008052   0.705029  -0.011  0.99089
## home         0.822419   0.164624   4.996 6.38e-07 ***
## chld        -0.479901   0.024594 -19.513  < 2e-16 ***
## hinc2        0.724436   0.224511   3.227  0.00127 **
## hinc3        1.053840   0.219768   4.795 1.75e-06 ***
## hinc4        1.530739   0.209704   7.300 4.17e-13 ***
## hinc5        1.945153   0.216358   8.990  < 2e-16 ***
## hinc6        1.996745   0.243401   8.204 4.17e-16 ***
## hinc7        2.263402   0.261482   8.656  < 2e-16 ***
## wrat1       -0.383966   0.696432  -0.551  0.58147
## wrat2        1.509826   0.537479   2.809  0.00502 **
## wrat3        1.146527   0.530306   2.162  0.03074 *
## wrat4        2.154137   0.509775   4.226 2.49e-05 ***
## wrat5        2.235959   0.507477   4.406 1.11e-05 ***
## wrat6        3.126804   0.501215   6.238 5.40e-10 ***
## wrat7        3.033440   0.501379   6.050 1.73e-09 ***
## wrat8        2.137782   0.494383   4.324 1.61e-05 ***
## wrat9        2.176076   0.494816   4.398 1.15e-05 ***
## incm         0.888930   0.079294  11.210  < 2e-16 ***
## plow         0.030359   0.003453   8.793  < 2e-16 ***
## tgif         0.354011   0.044762   7.909 4.30e-15 ***
## lgif         0.511338   0.072580   7.045 2.55e-12 ***
## rgif         0.675367   0.068088   9.919  < 2e-16 ***
## agif         0.728101   0.088339   8.242 3.05e-16 ***
## reg1        -0.035218   0.084861  -0.415  0.67819
## reg2        -0.108736   0.079787  -1.363  0.17309
## reg3         0.973161   0.105992   9.181  < 2e-16 ***
## reg4         1.854382   0.104865  17.684  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.6825, Adjusted R-squared:  0.6781
## F-statistic: 156.6 on 27 and 1967 DF,  p-value: < 2.2e-16
```

The algorithm removes eight variables, all of which were insignificant in the full model, were not selected by our regression tree, and were not strongly related to DAMT in our correlation analysis.

LASSO

We can also try to use LASSO to reduce the model, as we did with DONR.

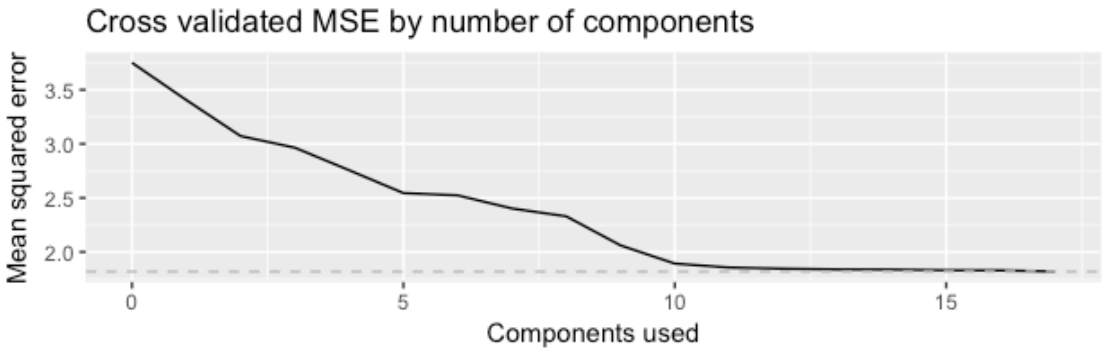




For DONR, the best lambda value (within one SE of the minimum, indicated by the rightmost dashed line) produces a signficantly reduced model. It shrinks to zero all eight of the predictors that were removed by the stepwise algorithm, as well as several dummy variables for household income and wealth rating that were significant in the OLS model.

Partial least squares

Since several of our predictors are strongly correlated, we can use partial least squares regression to capture only the latent factors they represent. We'll use 10-fold cross validation to choose the number of components to keep.

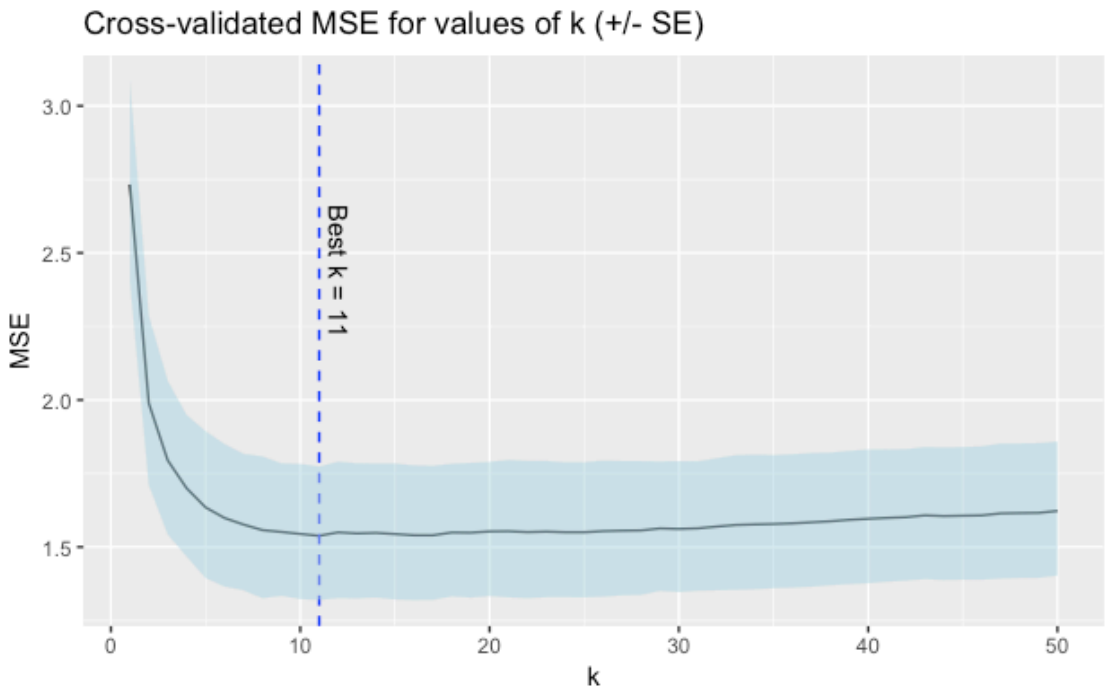


The lowest error is found with the full model, but error does not diminish much after the first 11 components. Therefore, we'll select the 11-component model for making predictions.

k-Nearest neighbors

Because k-nearest neighbors performed well for predicting donation probability, it would also be a good idea to try fitting a k-NN reression model for predictiong donation amount. Similar to what we did with DONR, we will remove the eight variables that were de-selected by the stepwise algorithm in the OLS model above, since k-NN is sensitive to dimensionality.

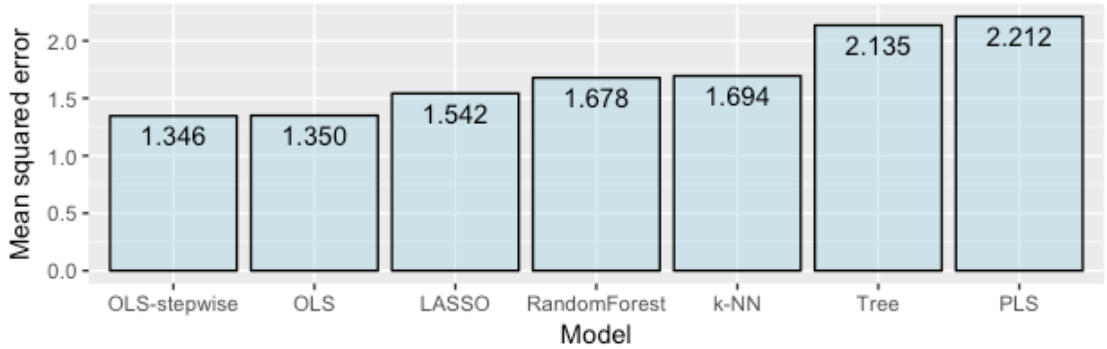
Using k = 3, we find a mean squared error of 2.0839 on the validation data set. That isn't terrible, but we saw that the PLS model fit in the previous section had MSEs below 2 on out-of-sample data. But once again, we can use cross validation to find the optimal value for k.



We find that k = 14 produces the lowest cross-validated error, which is close to the k = 12 value that we found for the DONR model. This model also looks like it will produce MSE values below 2, which is encouraging.

Validaton - DAMT

Validation for donation amount is much simpler than for donation probability. We generate predictions on the validation data using each model, then calculate mean squared error:



In this case, it looks like ordinary least squares regression, reduced using a stepwise algorithm, has produced the most accurate predictions.

Conclusions

We determine that the best model for maximizing profit is k-nearest neighbors. We find that an ordinary least squares regression model is most suitable for predicting donation amount.