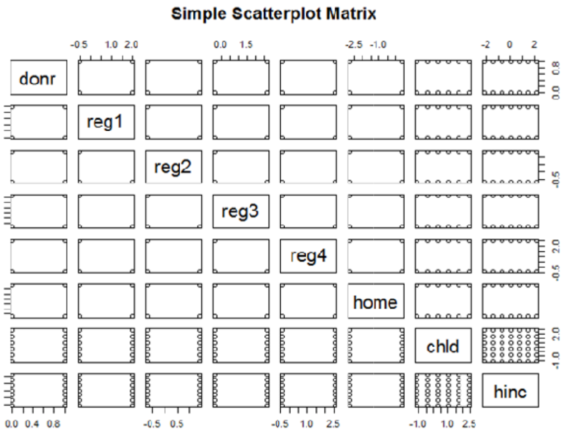
Tyler Smith
AMEX Example

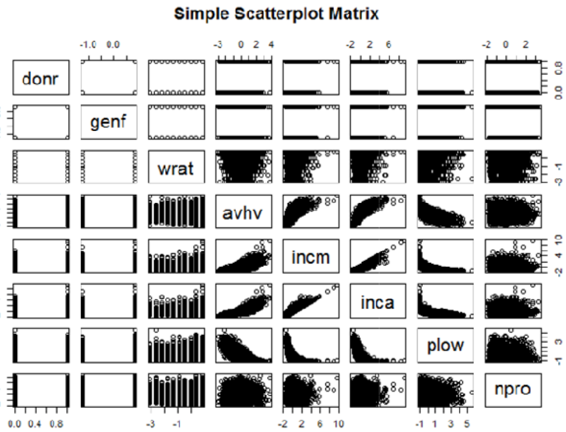
Donation Modeling: Practical Machine Learning

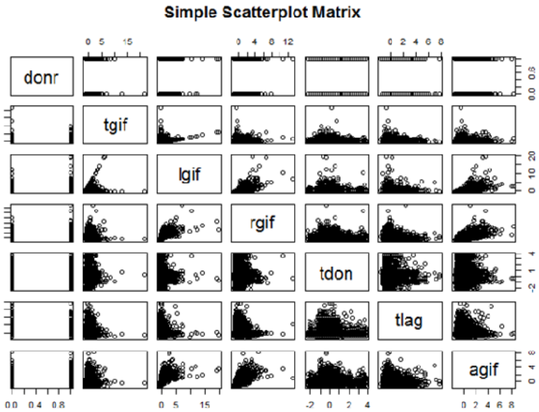
The primary purpose of this modeling exercise is twofold: first to predict whether or not an individual who received a mail solicitation would donate, and second to predict the amount people donated. Classification techniques are used in determining whether or not someone donated, and modeling techniques that are suited to continuous variables are used in predicting the actual amount donated.

Data Exploration

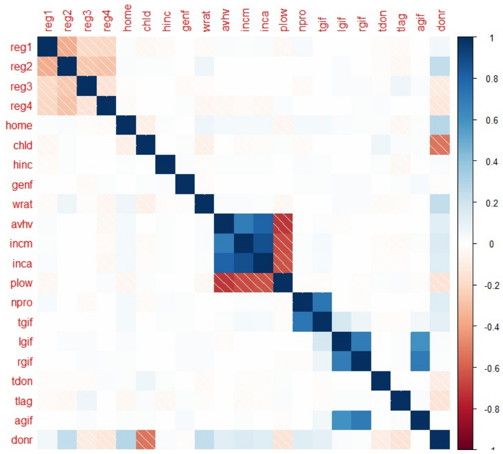
In order to make the data more amendable to modeling techniques the data was standardized. Doing this helped reduce extreme outliers and made the data easier to work with. Looking at a scatter plot matrix there seems to be little correlation between whether or not someone donated and the other predictor variables.



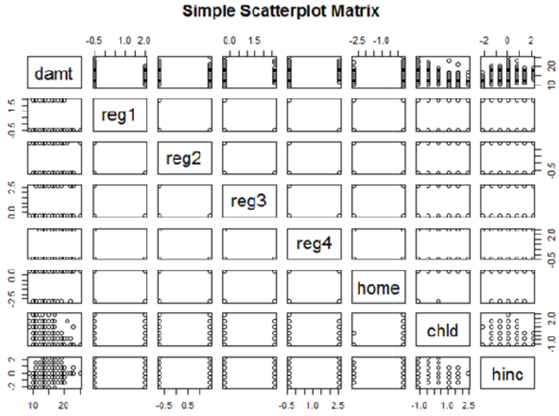


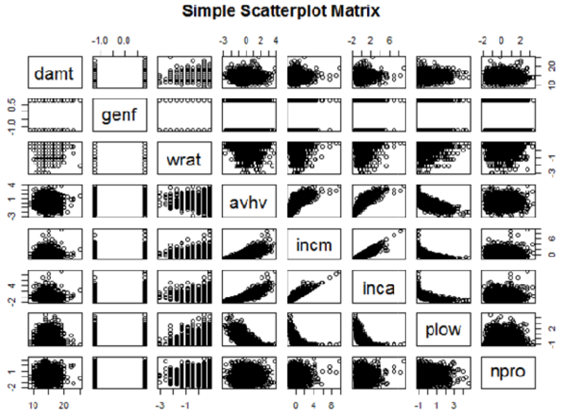


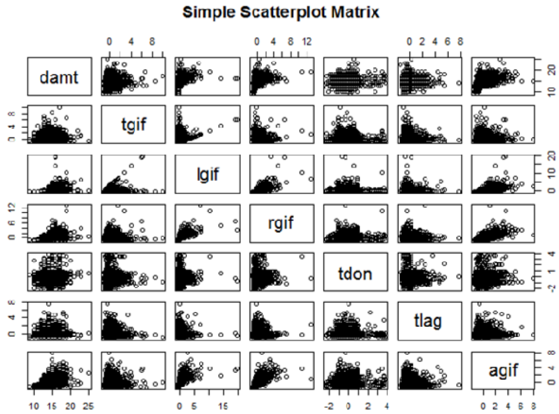
That said, a correlation matrix does seem to show some correlation in the data. It appears that owning a home is positively correlated to being a donor, while having a child seems somewhat negatively correlated to donating.



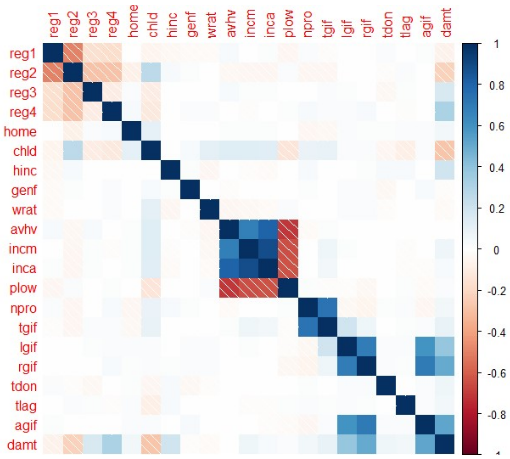
The scatter plot matrices for amount donated is a little more productive than for whether or not someone donated. It appears that amount donated correlates most to recent gift amount.







The correlation shows a little more detail on what variables are correlated with donation amount. Amounts previously donated have a stronger correlation with amount donated, and where someone is from also seems to correlated to higher donations.

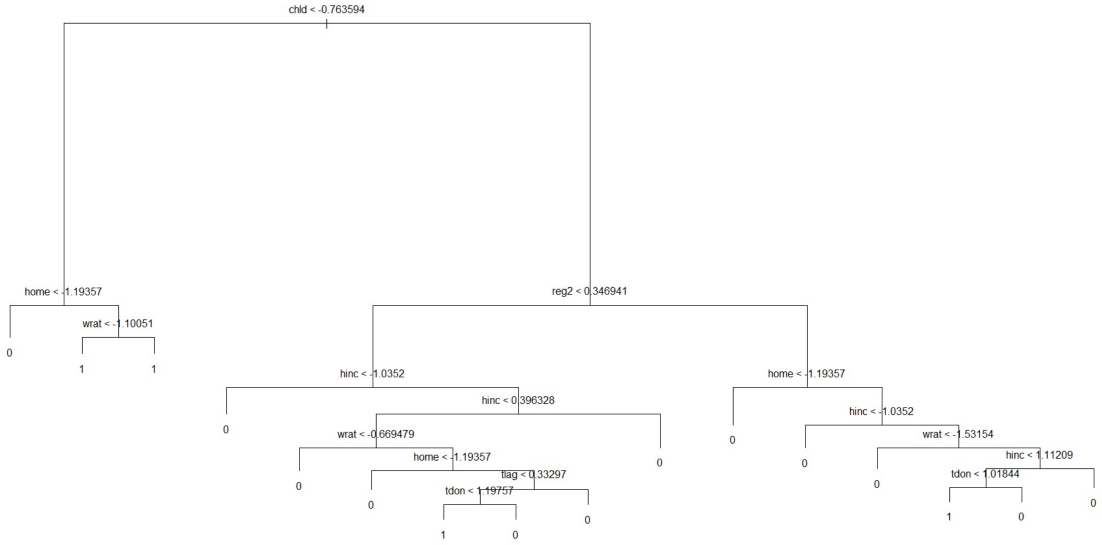


Normally a modeling project of this type requires much more effort in terms of data preparation. However, in this case the data appears to have been prepared. There seems to be very few outliers and no missing data. The data point for average home value has been log transformed, but the other data points have only been standardized.

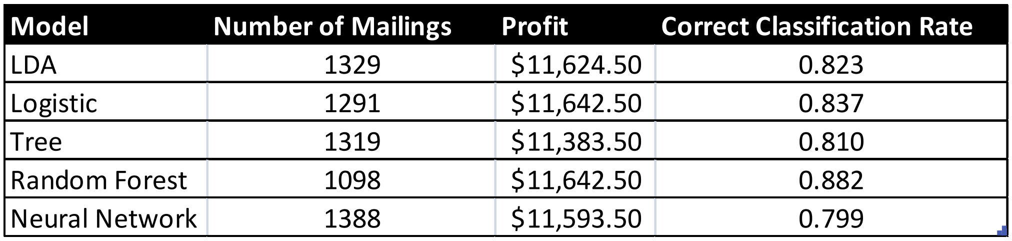
Classification Outcome Analysis

Five models were used in making prediction on who would donate. The models used were an LDA model, a logistic regression model, a classification tree model, a random forest model, and a neural network model. Each model had various difficulties associated with it. Creating the normal classification tree and neural network were quite time intensive and difficult to get to work. The main reason the tree was difficult was making sure the model understood the binary nature of the donor variable. Initially the output was difficult to convert back to 0's and 1's. Overall the neural network was not necessarily too difficult to use, but it was difficult to find a neural network that worked well, so quite a bit of time was used testing different version of the neural network model.

While the classification tree model did not necessarily yield the best results, the tree did help in defining what variables were most important. It was interesting to see that some of the variables that seemed important in the scatter plot and correlation matrices were not included in the results of the classification tree. This might be due to the classification tree linking important variables that might not be as observable through standard data exploration.



It was interesting that overall the models used for classification were fairly close. All of the models predicted an estimated profit above $11,000. The following table shows the results from all five classification models explored.



Model	Number of Mailings	Profit	Correct Classification Rate
LDA	1329	$11,624.50	0.823
Logistic	1291	$11,642.50	0.837
Tree	1319	$11,383.50	0.810
Random Forest	1098	$11,642.50	0.882
Neural Network	1388	$11,593.50	0.799

In order to better understand how well the models worked it helped to look at how often the models correctly classified mailing types. Using this information in conjunction with the goal of maximizing profit the Random Forest Classification model preformed the best. The Random Forest also had fewer classification error types than other models used; details for these error rates can be observed in the R file associated with this project.

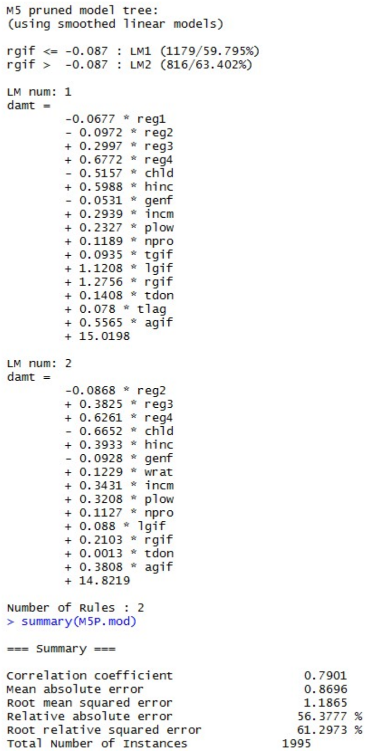
Continuous Outcome Analysis

Five models were also used in predicting the amount that would be donated; a linear regression model, a ridge selection model, a general additive model, an M5P regression tree, and another random forest. Making prediction in terms of amount seemed much easier than the classification models. The results seem to be distributed in a way that made sense, with no major outliers. This is reflected in the R code associated with this project. That said this may have been due to the test data set not including any really large donations.

Again the results for all models seemed to be fairly close. This helps confirm that the models explored were all focused on the correct information, and making similar projections. The following table shows the results of all five models.

Model	Mean Prediction Error	Standard Error
Linear Regression	1.868	0.170
Ridge Selection	1.867	0.170
General Additive	1.868	0.170
M5P Tree	1.679	0.164
Random Forest	1.654	0.172

The first three models seemed to return very similar results. The model that seemed to perform the best was an M5P regression tree. This method uses the same splitting functionality as the tree, but the end leafs are regression equations. Using the M5P method also showed which variable were most important in deciding how much would be donated.



```
M5 pruned model tree:
(using smoothed linear models)

rgif <= -0.087 : LM1 (1179/59.795%)
rgif >  -0.087 : LM2 (816/63.402%)

LM num: 1
damt =
        -0.0677 * reg1
        - 0.0972 * reg2
        + 0.2997 * reg3
        + 0.6772 * reg4
        - 0.5157 * chld
        + 0.5988 * hinc
        - 0.0531 * genf
        + 0.2939 * incm
        + 0.2327 * plow
        + 0.1189 * npro
        + 0.0935 * tgif
        + 1.1208 * lgif
        + 1.2756 * rgif
        + 0.1408 * tdon
        + 0.078 * tlag
        + 0.5565 * agif
        + 15.0198

LM num: 2
damt =
        -0.0868 * reg2
        + 0.3825 * reg3
        + 0.6261 * reg4
        - 0.6652 * chld
        + 0.3933 * hinc
        - 0.0928 * genf
        + 0.1229 * wrat
        + 0.3431 * incm
        + 0.3208 * plow
        + 0.1127 * npro
        + 0.088 * lgif
        + 0.2103 * rgif
        + 0.0013 * tdon
        + 0.3808 * agif
        + 14.8219

Number of Rules : 2
> summary(M5P.mod)

=== Summary ===

Correlation coefficient                  0.7901
Mean absolute error                      0.8696
Root mean squared error                  1.1865
Relative absolute error                 56.3777 %
Root relative squared error             61.2973 %
Total Number of Instances             1995
```

It was interesting to see that much of the regression equations from the model made sense. For example the model predicted that people with children would essentially be less willing to donate, which makes sense. The M5P model does the best job in predicting amount donated based on the metrics chosen.

Model Selection

Based on profit and the classification rate the random forest seems to be the best model for predicting if someone will donate. In terms of predicting the amount donated the M5P regression tree performed the best on the validation data set. I think one of the biggest obstacles to the models actually being right is the fact the validation set had a higher response rate than the test set. This calls into question what the actual results of the test set would yield.