Data Exploration

In this problem, we are given several variables with the aim of predicting whether or not someone will donate, and if so, how much they will donate. We have variables such as the region, number of children, wealth rating, and several other variables that are potentially useful in this predictive analysis. First, lets look initially at our response variable to get an idea of the data that we are working on. A quick look at our response variable shows it is roughly equally distributed between those who donated, and those who didn't. In Appendix A figure 1, we see that the donation amount of those who donated appears to be normally distributed. The mean is approximately $14.45, with the standard deviation being $1.98. The goal is to send mailers to only those who will donate more than $2.00, as that is how much it costs for us to create and send a single mailer. Figure 3 of appendix A displays the median family income. This plot is included as it likely holds high explanatory power. Our next step in the exploration is to determine if there is any collinearity among our variables. The correlation plot can be found in Figure 2 of Appendix A. While being messy, it gives us several key takeaways. The first being that there are several variables which are collinear, such as average home value and median family income.

Data Preparation

While we have no missing values in our sets, there are some necessary preparations we must first make. For example, some variables such as average home value, and dollar amount of lifetime gifts to date are highly skewed. To mitigate any effect this may have on our model, we will transform these variables by taking the log. Now that we have transformed our variables, we can begin the modeling process.

Build Models.

We will first start out by creating a model using backward subset selection to predict whether or not someone will be a donor or not. From there, we can take those individuals, and model the expected amount that they will donate. The backward step model is shown below.

$$i = observation$$

$$Y_i \sim Bern(p_i)$$

$$logit\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 x_{reg1,i} + \beta_2 x_{reg2,i} + \beta_3 x_{reg1,i} + \beta_4 x_{home,i} + \beta_5 x_{chld,i} + \beta_6 x_{hinc,i} + \beta_7 x_{wrat,i} + \beta_8 x_{avhv,i} + \beta_9 x_{incm,i} + \beta_{10} x_{npro,i} + \beta_{11} x_{tgif,i} + \beta_{12} x_{tdon,i} + \beta_{13} x_{tlag,i}$$

Equation 1 Binomial Logistic Regression Equation

Now that we have this model, we can predict whether or not someone will donate on the test set, build a confusion matrix, and tweak our thresholds so that we have the highest AUC. In Appendix B, Figure 1, you'll find the initial confusion matrix when we set the threshold to .5. Figure 2 in Appendix B shows our AUC to be .91, showing that our model is effective at predicting whether someone will be a donor or not. Some further analysis shows us that the threshold at which AUC is maximized on our test predictions is .55. Using this threshold, we

find our new AUC to be .915, just slightly better than our previous AUC. The updated confusion matrix can be found in Figure 3 appendix B.

With our logistic regression model, we can now take those who donated, and then see how much they are projected to donate using various models. To begin, we'll start with a simple linear regression model, utilizing all of the variables provided.

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{reg1,i} + \beta_2 x_{reg2,i} + \beta_3 x_{reg3i} + \beta_4 x_{home,i} + \beta_5 x_{chld,i} + \beta_6 x_{hinc,i} + \beta_7 x_{wrat,i} + \beta_8 x_{avhv,i} + \beta_9 x_{incm,i} + \beta_{10} x_{npro,i} + \beta_{11} x_{tgif,i} + \beta_{12} x_{tdon,i} + \beta_{13} x_{tlag,i} + \beta_{14} x_{reg4,i} + \beta_{15} x_{genf,i} + \beta_{16} x_{plow,i} + \beta_{17} x_{lgif,i} + \beta_{18} x_{rgif,i} + \beta_{19} x_{tdon,i} + \beta_{20} x_{agif,i}$$

Our saturated model yields a mean squared error of 1.55, and an AIC of 6306.7. Next, we will consider only the significant variables from our saturated model. To investigate this performance, we'll check to see if there are any outliers or influential points. Figure 4 in appendix B shows the plot of our Cooke's distance values. Of note, there are no concerning outliers to remove from our set.

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{reg1,i} + \beta_2 x_{reg2,i} + \beta_3 x_{reg1,i} + \beta_4 x_{home,i} + \beta_5 x_{chld,i} + \beta_6 x_{hinc,i} + \beta_9 x_{incm,i} + \beta_{10} x_{npro,i} + \beta_{11} x_{tgif,i} + \beta_{12} x_{tdon,i} + \beta_{14} x_{reg4,i} + \beta_{15} x_{genf,i}$$

This simpler model yields a mean squared error or 1.508, and an AIC of 6298, which proves to be just slightly better than our previous model. There is no need to check for outliers as we have used the same data to train this model.

Lastly, we'll use forward step selection to select our variables for our final model.

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{reg1,i} + \beta_2 x_{reg2,i} + \beta_3 x_{reg3,i} + \beta_4 x_{home,i} + \beta_5 x_{chld,i} + \beta_6 x_{hinc,i} + \beta_9 x_{incm,i} + \beta_{10} x_{npro,i} + \beta_{11} x_{tgif,i} + \beta_{12} x_{tdon,i} + \beta_{14} x_{reg4,i} + \beta_{15} x_{genf,i} + \beta_{16} x_{plow,i} + \beta_{17} x_{lgif,i} + \beta_{18} x_{rgif,i} + \beta_{19} x_{tdon,i} + \beta_{20} x_{agif,i}$$

This model yields an AIC of 6297, and a MSE of 1.55, which is higher than the previous model. This model is remarkably similar to our saturated model, but yields a slightly lower MSE. This model yields a slightly worse mean absolute error of .89.

Model selection.

Initially, just looking at our models, we see that Model 2 has the second lowest AIC, and lowest MSE. Since all of these models are nested, and linear regression models, we can conduct an F test to compare the models. When we compare our saturated model to our linear model, our F-test yields a p-value of .85. This signifies that model 2 is not necessarily a better model. When we do the same for model 3, we find the p-value to be .80. This is a better result than model 2, but not significant enough to prefer it. This final model yields a MSE of 1.55, and a MAE of .88.

Now that we have selected a model, we can conduct our analysis to determine the level of profit that our firm will achieve by sending mailers only to the predicted donors. Our model predicts that there will be 331 donors in the actual test set given. After accounting for mailing costs, we expect our profit to be $4095.20. As we begin to understand what our model says, it is important to first recognize that we fail to address any issues with collinearity in this model, as we have included all of the variables we were given. However, our model states that region 1 and region 2 appears to have the most significant effect on donation amount. We find that homeowners also tend to donate more than non-homeowners. Our model states that for each child someone has, the donation amount is expected to decrease by $2.37. It also states that those who live in wealth areas tend to donate more money. The model also states that the greater the median income of a neighborhood, the greater the expected donation will be. Lastly, the number of months since last donation is also a significant variable. This is logical considering those who have recently donated will likely donate again.

Future Work

While this analysis for the firm is somewhat useful, there are several areas in which this model lacks. First being there is only one logistic regression model to determine whether or not someone will donate. Realistically, there should be multiple logistic regression models, and maybe even a neural network, or a quasi-binomial model depending on the dispersion parameter. Additionally, there should be further work done on the linear regression models. For example, interaction terms would be a great addition to include in future models.

Appendix A

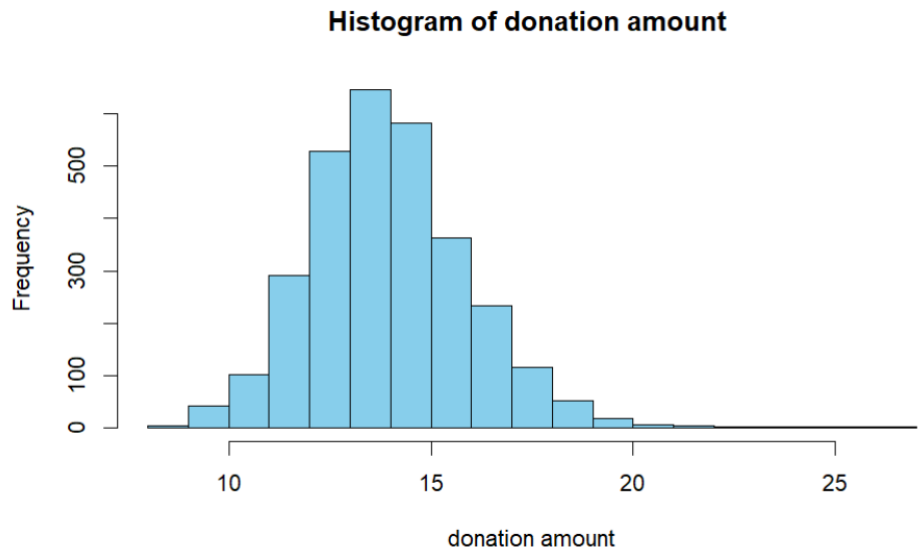


Figure 1Histogram of donation amount

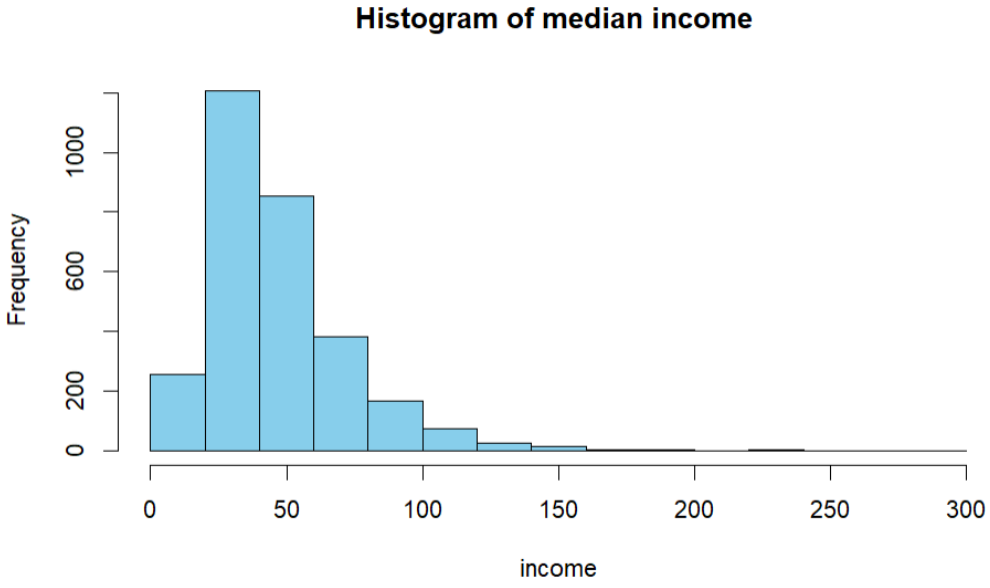
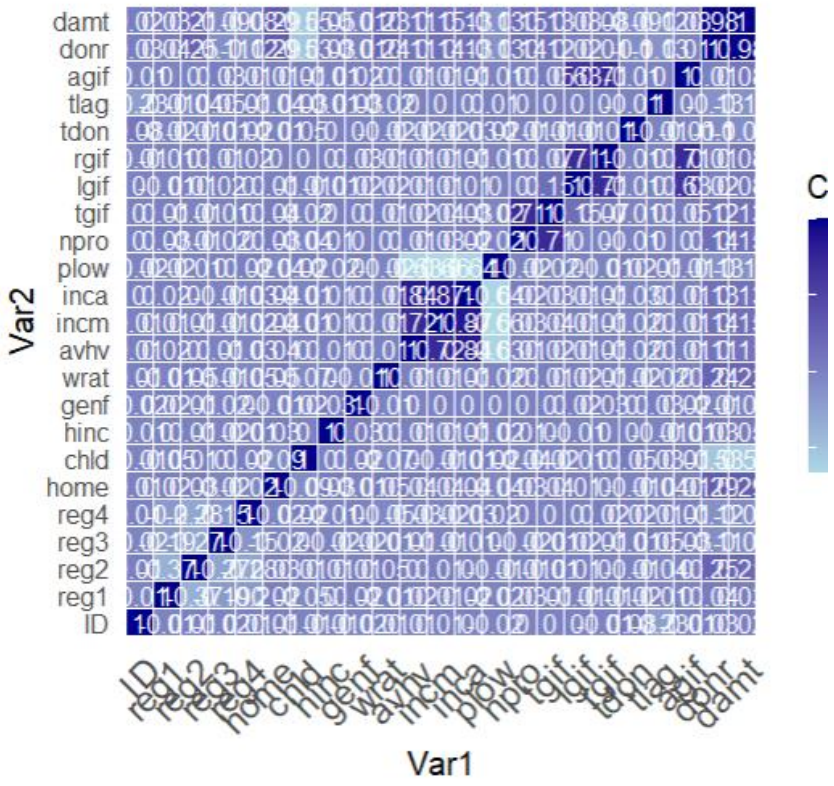


Figure 3 Histogram of median family income

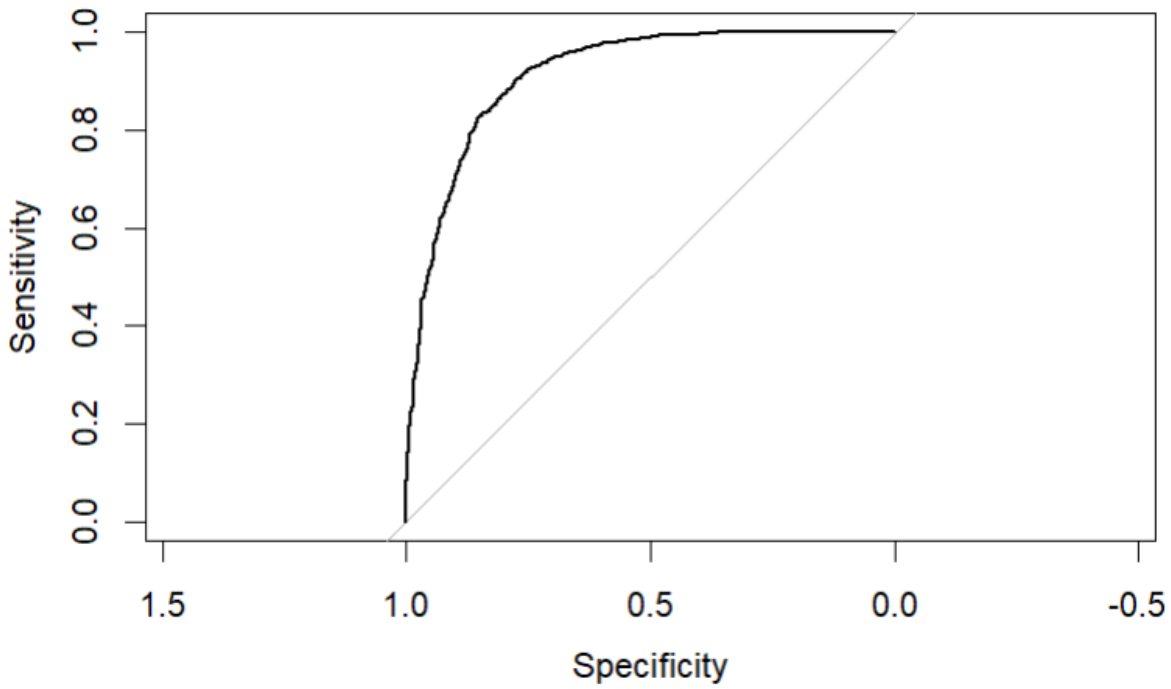


2Correlation plot

Appendix B

	Reference	
Prediction	0	1
0	828	137
1	191	862

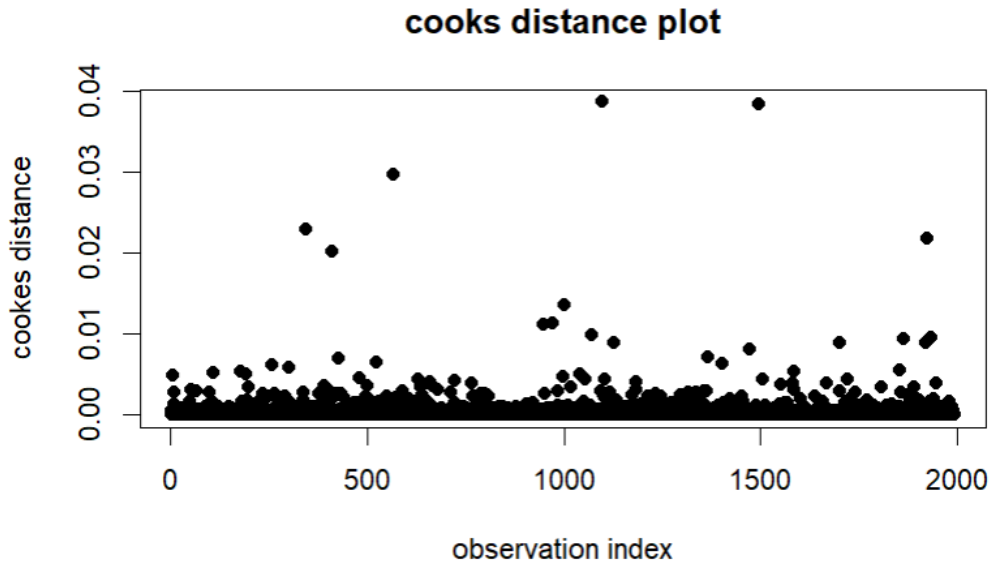
Figure 1: Initial Confusion Matrix for donors.



3ROC Curve

Prediction	Reference 0	1
0	854	164
1	165	835

Figure 4Optimized Confusion Matrix



5 Cookes distance plot.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.189957   0.047363 299.601  < 2e-16 ***
reg1        -0.038804   0.039626  -0.979  0.32758
reg2        -0.074053   0.042939  -1.725  0.08476 .
reg3         0.327051   0.040405   8.094 9.96e-16 ***
reg4         0.635806   0.041596  15.285  < 2e-16 ***
home         0.238225   0.060728   3.923 9.05e-05 ***
chld        -0.604395   0.037950 -15.926  < 2e-16 ***
hinc         0.501934   0.039843  12.598  < 2e-16 ***
genf        -0.063174   0.028496  -2.217  0.02674 *
wrat        -0.001583   0.041509  -0.038  0.96959
avhv        -0.056103   0.054302  -1.033  0.30165
incm         0.289597   0.059094   4.901 1.03e-06 ***
inca         0.046769   0.068895   0.679  0.49732
plow         0.235295   0.047488   4.955 7.86e-07 ***
npro         0.136824   0.044397   3.082  0.00209 **
tgif         0.058889   0.046039   1.279  0.20100
lgif        -0.055205   0.038431  -1.436  0.15103
rgif         0.516382   0.043862  11.773  < 2e-16 ***
tdon         0.072643   0.034931   2.080  0.03769 *
tlag         0.022708   0.033666   0.675  0.50007
agif         0.671843   0.040479  16.597  < 2e-16 ***
```

Table 1 Saturated Model results