MSIA 401 Project, Fall 2016

Jonny Burge, Sai Haran, Kate Larsen, Craig Ng

I. Executive Summary

Through binary logistic and multiple linear regression analysis, we determined that data points concerning extremes, rather than averages, proved most useful in discriminating between responders and non-responders, and predicting the magnitude of their donation. Key predictors in identifying responders included the size of a household's first, largest, and latest contribution, as well as the time elapsed since their last donation. Additionally, analysis of whether or not a household was adhering to their historical frequency of donations proved very useful.

For the given responders, key predictors in estimating the amount of contributions included the latest contribution amount, the largest contribution amount, and the months since the largest contribution. Additional predictors of high importance were the types for the solicitation and contribution codes. Once both models were developed and tested, we used both models to predict the top 1,000 donors with the highest contributions.

II. Introduction

Our approach to this project initially focused a significant amount on exploratory analysis. Our goals were to understand the data in both the greater context and in individual attribute distributions. After developing a strong understanding of the data at hand, our goal was to brainstorm potential significant predictors and any additional information that would be beneficial. We seeked to develop models that accurately attributed factors to a donor's contribution history. We approached this effort by

developing the logistic and multiple regression models in parallel. In developing each model, we followed an iterative approach in selecting predictors, building the model, and assessing the fit through various measures. Our subsequent and final goal was to then combine the two best models together to most accurately predict the top 1000 donors by greatest contribution. Throughout this process, we maintained a focus on identifying significant model alterations, as well as measuring the different solutions.

Prior to analysis, our a priori hypotheses focused on the temporal nature of the data. We thought that columns containing most recent data such as the last month contributed and the last contribution amount would be significant in our model. We used these ideas during to later think through possible interactions that incorporate such attributes.

We segmented the project into different phases as detailed in the following sections from data exploration to final model development. We began with data exploration to understand not only the context of each variable, but also the available observations. Throughout our analysis, we began to examine outliers and any unique relationships between the variables. In the next phase of the project, we focused efforts on determining possible sets of predictor variables. Employing a range of variable selection techniques enabled the group to identify influential features and remove those that did not seem as significant. Throughout this process, we also developed new features and scores to capture information available within the data.

Subsequently, we began building the logistic and multiple regression models. Model fitting was an iterative process as we would cycle through different sets of

predictors, interactions, and transformations to evaluate potential options. For the logistic regression model, we ran stepwise regression to evaluate which predictors would most likely be significant in the final model. In addition to the existing variables, we also created and evaluated new columns such as binned variables, interactions, and ratios that showed new patterns and relationships in the data. Using various subsets of these predictors, we built different logistic models and evaluated each by examining AUC values. We concluded with a model that included various original predictors, interactions, and newly calculated columns.

We followed a similar process for the multiple linear regression model by first conducting exploratory data analysis on the training set and then creating different models from a more defined attribute list. In exploratory data analysis, we identified outliers, examined multicollinearity, and evaluated newly created attributes. Next, we employed variable selection methods, such as stepwise and lasso regression, to gain better insight into significant predictors. Using this information and previous factor analysis, we developed a range of multiple regression models using different subsets of predictors, interactions, and log transformations. Throughout the process, we evaluated R^2 and MSE as measures of performance and ultimately concluded with a model with the highest expected value of contribution when given the test data.

III. Model Fitting

Before addressing specifics regarding the steps and methods used to fit our models, we will provide an overview of how we cleaned, imputed and manipulated the data set to prepare it for model fitting.

Exploratory Analysis

Initial inspection of the dataset revealed the presence of some missing values. These missing values, however, were not due to randomness, but rather due to no valid value existing in the given context. Specifically, for the instances in which the donor had fewer than three lifetime contributions, there were no values for features pertaining to the nature of their second or third most recent contribution.

Imputation

We performed very little imputation on the original dataset. As previously mentioned, all of the missing values were absent because they pertained to donations which had never occurred, and thus could not be observed. As a result, it does not make sense to impute these values. Since no reasonable imputation could be performed, these features could not included in our models. Instead new features were derived using the data in columns with missing values.

Derived Features

After observing the original data, we believed that there was additional information which was not represented in the existing features. In order to extract this information so that it could contribute to the fit of our models, we added some new features derived from those in the original dataset. The following section will discuss the

motivation and reasoning behind including the derived features in our model (see Figure 7 in the Appendix for the full table).

CNCOD_TYPE

The original dataset contained numeric contribution codes for each donor's previous three donations. The significance of these values, however, stems from the contribution type the numeric code represents. We created new features using the five corresponding alphabetic codes for the most recent contributions.

SLCOD_TYPE

A similar mapping was performed for the numeric values of SLCOD1, SLCOD2, and SLCOD3. The solicitation type was then stored in the new feature.

CNMON_MTH

The original dataset contains the dates of the previous three donations for each donor in terms of how many months ago the donation occurred. We were interested in seeing if including temporal data would reveal some sort of seasonality in contributions which would improve the models' performance. However, the original temporal features, would not provide this kind of insight. To address this need, we created new features containing the month the donation occurred for each donor's previous three contributions.

years_last_cntr

One of the donor attributes which intuitively seemed significant was the duration since the donor's most recent contribution. The feature CNDAT1 already contains how many months ago the last donation occurred, however we thought that there may be

some benefit from grouping this continuous variable. As a result, we grouped the values of CNDAT1 by intervals of two years.

CNMONF_BIN, CNMONL_BIN

Similarly, we grouped CNMONF, the duration in months since the donor's first contribution, and CNMONL, the duration in months since the donor's largest contribution, across two year intervals. These features were significant in our logistic and multiple linear models respectively.

avg_don_life

When considering what aspects of a donor's history would be useful in projecting the magnitude of future contributions, the donor's average contribution size came to mind. This value was easily derived from the donor's total lifetime contributions and number of lifetime contributions, both features in the original dataset.

avg_don_latest

Additionally, the magnitude of recent contributions may improve the model projecting the size of future donations. We calculated this by averaging every donor's three most recent contributions.

Division, Region

The original dataset provided state codes for each donor. We were interested in geographical trends pertaining to donation frequency and amount. In order to aggregate the donors by location we incorporated the notion of geographical divisions and regions as defined from the US Census data.

season

Since our objective was to predict the likelihood and magnitude of donations over the fall of 1995, there was no need to discriminate between donations occurring across the same month. This aggregation was performed by assigning each donation a seasonal classification.

CNTRLIF_bin, CONLARG_bin, CONTRFST_bin, CNTMLIF_bin, avg_don_life_bin, avg_don_latest_bin

We aggregated the features pertaining to total lifetime contribution amount, largest contribution amount, first contribution amount and total number of lifetime contributions, average donation over the donor's lifetime, and average donation of the donor's most recent contributions. These derived features did not prove to be significant and were ultimately dropped from our models.

freq

Another donor attribute that we intuitively thought was significant was the frequency at which that donor has contributed in the past. This feature was derived from the number of total contributions and the duration since the donor's first contribution. It is important to note that this value is of units months/donation. This allows it to be easily compared with other measurements of time, however must be inverted to be used a true frequency. This derived featured did prove to be significant and was included in both our logistic and multiple linear regression models.

due, rly_due

In addition to frequency, we were interested in the frequency's relation to the duration since the donor's most recent contribution. The desire to include this

relationship stemmed from the notion that if a donor historically contributed every few months but it has been a long time since their last contribution, they may be likely to donate soon. The features due and rly_due are binary values which represent whether the time elapsed since the donor's most recent contribution is greater than the typical duration between donations, or twice that value, respectively.

donate_prob

When considering the notion of a donor being 'due' we realized its relation to a similar problem we encountered in Healthcare. In our Healthcare class we had performed Poisson regression to determine the frequency at which kidneys of various arrived. From these frequencies we then computed the probabilities of these kidney arriving each month. In this instance we can use the cumulative density function to compute the probability of a donor contributing over the next four months given their historical frequency of donation. This evaluated using the equation $P(T \leq t) = 1 - e^{-\lambda t}$, where λ is the donor's donation frequency and $t = 4$ months. For our computation, the equation looks like $P(T \leq t) = 1 - e^{-(1/freq) * 4}$.

Training and Test Data

After reading the data into R, we wanted to preserve the current row order in order to have the correct test and training data sets. Therefore, we separated the data on indices before cleaning columns and added a column $SET to assign to test or train data. Once the data had been split to test and train, we then removed the $SET column from each data set. We implemented these steps first, before updating any columns or merging data, in order to keep the initial index order intact for our data sets.

Logistic Regression

Model

After cleaning the data, creating derived features, normalizing several continuous variables (CNDOL1, CNDOL2, CNDOL3, CONLARG, CONTRFST, CNTRLIF) and splitting the data into training and test sets, we began fitting the model by running a binomial glm on bin_TARGDOL against all non-binned predictors. Initially, we tested this model's goodness of fit using AUC as our main metric. When we discovered this may not be enough to determine how well our model achieves our end goal (improving final dollar value) we calculated correct classification rate on the test set using a p* of .5. While we picked this p* fairly arbitrarily, we discovered that it was still useful for determining goodness of fit as a comparative tool between models. We used these two metrics to rate every iteration of model moving forward.

After doing so, we ran a stepwise regression in both directions in an attempt to maintain fit while removing unnecessary predictors. The initial findings were that CNDOL3, Region, and avg_don_latest were not significant enough predictors to include in the model. Initial AUC and CCR hovered around .68 and .73, respectively. We felt that these were good enough baselines to move forward with the model in its current state.

Our next step was to determine if we had outliers or influential observations in the data, and if removing these observations would improve our fit. We thought it would be fruitless to search for outliers when the response variable was binary, so we moved onto influential observations. To find influential observations, we calculated leverage for

each observation and determined the observation influential using the rule of thumb, i.e. hii>2(p+1)/n. This ended up marking ~10000 observations as influential, which we thought was odd for a dataset consisting of only ~70000 records.

Running a new model without influential observations did not improve our goodness of fit metrics. AUC remained about the same while CCR dropped to .68. In response, we tried relaxing the rule of thumb to only find observations whose leverage exceeded 4(n+1)/n influential. Doing so reduced the number of influential observations to ~5000. Removing this subset of observations from a fit fared similarly to the previous attempt, so we decided to move on without implementing this measure in the model.

Next we tried replacing all continuous variables with their binned cousins. For example, instead of using CNDOL1, we used CNDOL1_bin, which was the former variable categorized into “bins”. To create the bins, we looked at the probability density functions of each predictor and assigned categories where we believed we could split the data. This gave us much less granularity in our predictors, but afforded us a chance to reduce influential observations’ effect on the target variable. For the majority of binned variables, this did not have the desired effect. We lost too much information in the binning process and lost much of our predictive power in the logistic model along with it. The binned variables that we decided to keep over their continuous counterparts, due to significance, were CNMONL_BIN and years_last_cntr.

The last step we took to improve the model was adding in interactions and ratios of predictors. The heuristic we used was to find individual predictors which had large standardized betas and find their interactions. Additionally we tried using variables

which we believed had a strong inverse correlation with the target variable (such as freq) as the denominator in several derived features. By iterating over several sets of features and letting stepwise regression sort out some of the less useful variables, we ended up using the following interactions: CONTRFST:CNTMLIF, CONLARG:CNMONL_BIN, and CNDOL1/avg_don_life.

At this point we were satisfied with the model we came up with and began validating it in combination with the multiple linear regression model, which we built in parallel. Since the final dollar value was the goal, getting the most discriminatory model via analysis of AUC or CCR was not the best metric to look at. We realized that even if a model is very discriminatory, it may be predicting more low-tier contributors (e.g. $5 contribution) rather than the big hitters (e.g. $200 contribution). Ideally, for this project's sake, we would have liked the logistic model to identify the top 1000 contributors in the test set as having a high probability to donate, even if that meant the other 29000 contributors were assigned a probability of 0. However we didn't think this was a very accurate parallel to "real-world" model fitting so we attempted to keep AUC and CCR high while getting total dollar value to an acceptable level.

To this end, we removed some derived predictors which had strengthened the model's discriminatory power in favor of increasing dollar value. The final model can be found in Figure 3 in the appendix, with final ROC curve and AUC found in Figure 5.

Multiple Regression Model

Initial feature selection

At the same time of developing the logistic model, we were also developing the multiple linear regression model. We first began with feature selection, examined outliers, and analyzed multicollinearity. We decided to drop any features related to the 2nd and 3rd most recent contributions (CNCOD2, CNCOD3, CNDAT2, CNDAT3, CNDOL2, CNDOL3, SLCOD2, SLCOD3, CNMON2, CNMON3) because of a large number of missing values. We visualized the missing data using the MISSMAP() function in the Amelia package, see Figure 2 in the appendix. The 2nd most recent donations had about 59,000 missing values and the 3rd most recent contributions had about 67,000 missing values. There was no way to impute the values, as we never considered removing the observations, since that would eliminate over ⅔ of the total observations, so we elected to remove the features instead. We believed we'd be able to capture some of the predictive power of the 2nd and 3rd most recent donations using the the total number of contributions and total lifetime contribution in dollars to calculate an average donation amount. We ended up dropping this feature because the vast number of donors only donated once, so the average donation amount ended up being equal to the most recent contribution, which would cause perfect multicollinearity. We also dropped the STATCODE feature because we transformed the data into Regions and Divisions per the U.S. Census in order to reduce the granularity of the categorical variable. We removed the CNCOD1 and SLCOD1 features because we merged the types from the separate file. We also removed the ID feature because it served no purpose with regard to the analysis, as it was just a counter for each observation.

We next conducted exploratory analysis prior on the data prior to fitting models. Our two main focuses during this analysis covered examining outliers and evaluating multicollinearity. Given the test and train data, we began examining the distribution for the target dollar column, specifically focusing on the extremes. We examined the maximum target dollar amount for the test and train set. Upon research, we found that the train set had a high extreme where one observation had $1500 as the target dollar amount contributed. There was a sharp drop in this amount from $1500 to the next highest amount of $250; furthermore, the summary of the column showed a mean of $25 and a median of $31.01. We classified this $1500 record as an outlier and removed it from the train set. We conducted similar analysis on the highest amount contributed in one's lifetime. The same observation with $1500 as the target dollar amount also held the maximum amount of life contributed in a lifetime at $3750.00. The summary of the column, with mean $43.93 and median $13.00, further showed this record as an extreme. Therefore, this one observation covering both of these cases was removed from the training set.

In addition to analyzing outliers, we also inspected for any multicollinearity. Once we had built the full model where all variables existed as predictors to the main response variable of target dollar contributed, we analyzed the multicollinearity of the attributes. We removed the following columns from our model where the VIF value were greater than 10: CNDOL1_bin, CNTRLIF_bin, CONLARG_bin, CONTRFST_bin, CNDOL2_bin, CNMONL_BIN, avg_don_life_bin, avg_don_latest_bin, avg_don_life, avg_don_latest. Multicollinearity with the binned columns most likely resulted from

having the original columns present in the model as well. We kept the continuous variables over the binned columns by choosing those with greater significance in the fit of the model. The remaining multicollinear columns most likely resulted from observations where there was only one contribution per donor, resulting in equal amounts for the first, average, greatest, and last contribution amount.

Models

Once we had a potential list of significant attributes and rather reliable observations, we began to build multiple regression models. We developed a range of different multiple regression models to include subsets of predictors, interactions, and log transformations; furthermore, we also ran stepwise and lasso regression to aid in variable selection prior to fitting our final multiple regression model.

We began by fitting a model with all the remaining features from the original data set, data transformations and derived features to develop a baseline expectation for the predictive power of the variables. We expected there to be multicollinearity because of the large number of data transformations and derived features, but we decided to leave them in for the full model to see which features emerged as significant. We did not check for multicollinearity at this point because we intended to use other methods (like stepwise regression) to reduce the number of features prior to checking if multicollinearity was still an issue.

We employed the stepwise regression technique on the full model to have the algorithm choose the features that we should use. We first set the null model as the model with just the intercept and the full model as the model with all of the features (as

discussed above) and ran the STEP() function in R in both directions with the null and full models as the lower and upper bounds. We ran the stepwise regression in concurrence with the lasso regression to be able to evaluate alternative models and found that the outputs were similar. Ultimately, we utilized the output of the lasso regression as the basis for our final model.

We used the lasso regression from the glmnet package in R to aid in variable selection. We set the number of folds as the default 3 and the standardize argument to false since our data was not standardized in to the model. Lasso regression was chosen over ridge regression due to the nature of the lasso acting as a variable selector. We used the model to determine which predictors coefficients would be set to 0 and which remaining variables would be selected. Upon review of the lasso output, we found that various columns received a close to 0 value for the coefficient. We removed the predictors from our future models that had close to 0 coefficient since these predictors would then not have as great an impact as the other attributes. In addition to reviewing the coefficients, we also reviewed the VIF values to determine multicollinearity in this model. A review of the VIF values and attributes showed that many of the dollar amount columns were closely related. This intuitively makes sense because many donors have only donated once. Therefore, the dollar amounts of a donor's highest, average, first, and most recent donation are all the same because in those cases, there has only been one donation. Upon this analysis, we removed the overlap by removing the avg_don_life (average lifetime donation) and avg_don_latest (average of three most recent donations) predictors from our future models.

Throughout our model fitting process, we experimented with different combinations of bins and newly created columns (ratio, due, etc) as predictors as well. We found that the bins to be less significant in the multiple regression model. Upon analysis of the significant columns, we decided to relevel the solicitation and contribution donation type code to “B” in the model. The code was originally defaulted to “A” as the reference level; however, in the model summary the distinction between A and the other levels, except B, were not as significant. After releveling the factor to “B,” each of the factor levels were then significant in the model. In addition to different predictors, we also evaluated different transformations of the response variable. After viewing the normal plot of the fully fitted model, see Figure 1 in the appendix, we applied a log transform to the response TARGDOL variable as we fit the different models. However, the fit worsened with each model as we added this transformation; for example, after adding the log transform to the final model, the adjusted R^2 value dropped to 0.4592. Similar results were obtained when log transforming the other models as well; thus, the final model does not have log transform.

Our final multiple regression model resulted in the following nine predictor variables, all of which were significant at the 0.005 level: CNDOL1, CNTRLIF, CONLARG, CONTRFST, CNMONL, CNCOD1_TYPE, SLCOD1_TYPE, CNMONF_BIN, freq. While the model’s adjusted R^2 value of 0.6157, Figure 4 in the Appendix, is not as high as in the other models we tested, we found that the fit best predicted the value of expected dollar amounts when given test data.

Since we trained the model on the subset of training set where the target dollar amount was greater than 0, so we first validated the model by first calculating the total difference between the predicted and actual values for the full test set and then the subset where the target dollar amount is greater than 0. Our understanding of the rationale behind training the model on a set that includes only positive donation amounts is that it increases the model's ability to predict the precise dollar amount if a given person were to donate. If we had included the observations with a target dollar amount equal to 0, it would remove the need for the logistic regression, as the presence of 0's would give the model some ability to predict whether a potential donor would donate at all (predicting a 0). Since the problem was broken down into calculating separate expected probabilities of donation and predicted donation amount, we decided that we should expand the testing to include both the full test set and the subset. The total difference for the full test set was $227,660 while the total difference for the subset was $506. We would expect the difference for the full set to be larger because the model was trained on a set that did not have any target dollar amounts equal to 0, so it would be less able to predict a 0.

We then examined the error rate with more granularity by first calculating the root mean square error using the subset of the test set where the target dollar amount is greater than 0 then calculating the root mean square error using the entire test set to gain an understanding of the error in the model at the individual prediction level. The root mean square error for the subset of the test data where the target dollar amount was greater than 0 is $5.32 and the root mean square error for the entire test set is

$10.44. The fact that the RMS for the subset is lower than the RMS for the entire test set makes sense because we trained the model on the subset of the training set where the target dollar amount is greater than 0.

The average target dollar amount for the subsetted data was $8.42 and the average target dollar amount for the full test set was $2.30. Due to the high percentage of non-donors (target dollar value equal to 0), the average for the full test set was substantially lower than the subset. Additionally, the presence of a large number of non-donors likely drove up the RMS because the model was trained on a set that didn't include any non-donors. Given this, the larger RMS and lower mean on the full test set do not indicate that the model performed poorly overall.

IV. Model Prediction

Using the logistic model and the multiple regression model, we predicted the test set donation values and compared these predicted values against the actual contribution amounts. To retrieve this expected value of TARGDOL, we ran the models on the test set of data and multiplied the output of the logistic model, which calculated the probability of a donor contributing, by the output of the multiple regression model, which predicted the dollar amount a donor would contribute. Selecting the top 1000 donors, ordered by expected value, and calculating the sum of their actual donations yielded a total value of $10,137.23 (see Appendix Figure 6). Throughout this prediction process, we experimented with different model features as described in the above sections in search of greater payoff values.

V. Conclusions

For the most part, the results obtained from our analysis supported our a priori hypotheses. Predictors which provided information about extreme behavior, such as most recent donation or largest donation, proved to be most useful when predicting both likeliness to donate as well as size of donation.

Both models utilized CNDOL1, CONLARG, CNTRLIF and either CNMONL or CNMONL_BIN. CNDOL1 and CONLARG were effective in identifying high yield households for fairly obvious reasons; those who had recent exposure to donation and those who had shown a propensity to donate large amounts of money would naturally be more likely to donate again, and in larger amounts. CNTRLIF represented either a proclivity to donating or contained the same information as CNDOL1 and CONLARG, meaning it only added more predictive power to both models. CNMONL was a bit of an oddity, as its positive coefficient indicated that those whose largest donation was farther in the past both were more likely to donate and were expected to donate more. This goes counter to our earlier notion that recency influenced the most current donation, but perhaps it represented the thought that households were more likely to donate if they hadn't donated a large sum in a long while.

The logistic model showed a noticeable inverse relationship between CONTRFST and bin_TARGDOL. We suspect that donors whose first contribution was very large may have shied away from future contributions due to their first donation

being sufficient for several years. This would prove useful in determining if a household would be likely to donate again.

Lastly we saw our derived feature, freq, and several of its offspring show strong predictive power in the logistic model. Though we didn't see many apparent patterns in frequency in the raw data, this feature may have been able to provide some information for those who happened to donate on a schedule. The other derived features, such as due and donate_prob, had similarly strong effects in pattern detection.

There are several predictors which we believe may have been able to increase both models' predictive power: income, education, children, and age. Of these, household income stands out as being most likely to be able to improve the models. One would think that a richer household would be more willing to part with their money in the form of a donation. Children and age both fall into a similar category; those with more children may be less able to donate to a cause, and older individuals may be either more exposed to the non-profit asking for donations, or more willing to donate, having less to spend money on. Finally, education may have been able to provide more demographic information that may help in both discrimination and ability to donate larger sums of money, as education and wealth are positively correlated.

The models fared well enough with just location, sex, and donation information, but we believe with additional demographic information we would achieve a better result.

Appendix

Figure 1: Normal Q-Q Plot for training data

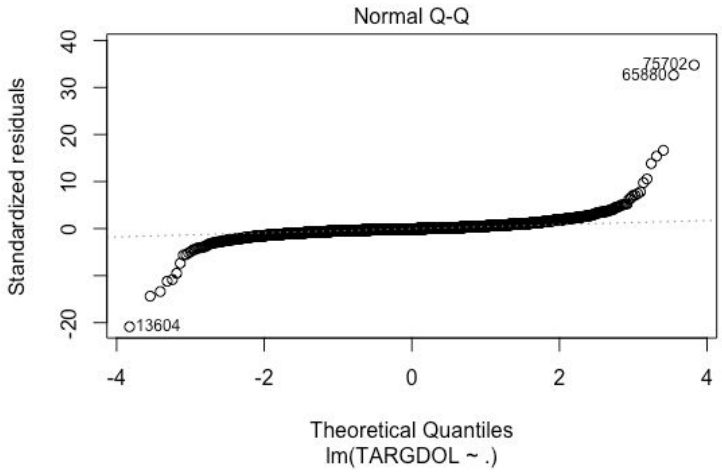


Figure 2: Missingness of full donation data set (CNDOL2 / 3 had N/A changed to 0)

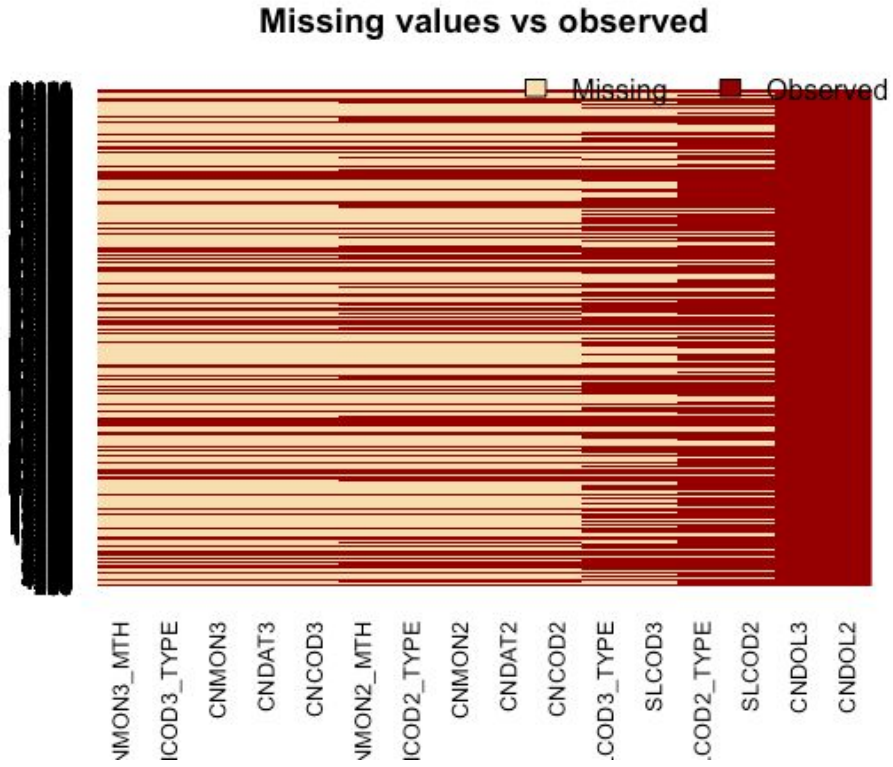


Figure 3: Summary output for final binary logistic regression model

```
Call:
glm(formula = bin_TARGDOL ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST +
    CNDOL2 + CNTMLIF + SEX + CNMONL_BIN + CNCOD1_TYPE + season +
    years_last_cntr + Division + freq + due + rly_due + donate_prob +
    CONTRFST:CNTMLIF + CNMONL_BIN:CONLARG + I(CNDOL1/avg_don_life),
    family = binomial, data = train_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0871  -0.8373  -0.5976   0.9367   3.0149

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.745195   0.086877  -8.578  < 2e-16 ***
CNDOL1                      0.111653   0.029117   3.835 0.000126 ***
CNTRLIF                    -0.046815   0.024581  -1.905 0.056841 .
CONLARG                    -0.086879   0.038835  -2.237 0.025277 *
CONTRFST                   -0.143690   0.024128  -5.955 2.60e-09 ***
CNDOL2                      0.077131   0.020669   3.732 0.000190 ***
CNTMLIF                     0.046132   0.004087  11.287  < 2e-16 ***
SEXC                       -0.501131   0.149762  -3.346 0.000819 ***
SEXF                       -0.236636   0.035387  -6.687 2.28e-11 ***
SEXM                       -0.181463   0.035947  -5.048 4.46e-07 ***
SEXU                       -0.265523   0.042574  -6.237 4.47e-10 ***
CNMONL_BIN(2,4]             0.684143   0.040212  17.014  < 2e-16 ***
CNMONL_BIN(4,6]             1.111212   0.049280  22.549  < 2e-16 ***
CNMONL_BIN(6,8]             0.996112   0.065995  15.094  < 2e-16 ***
CNMONL_BIN(8,10]            1.067067   0.089828  11.879  < 2e-16 ***
CNMONL_BIN(10,12]           1.058580   0.147894   7.158 8.21e-13 ***
CNMONL_BIN(12,14]           0.840126   0.248545   3.380 0.000724 ***
CNCOD1_TYPEB               -0.454418   0.055880  -8.132 4.22e-16 ***
CNCOD1_TYPEC               -0.226313   0.094707  -2.390 0.016866 *
CNCOD1_TYPED               -0.458689   0.030730 -14.926  < 2e-16 ***
CNCOD1_TYPEM               -0.531079   0.093090  -5.705 1.16e-08 ***
season(3,6]                -0.054741   0.036679  -1.492 0.135582
season(6,9]                 0.191207   0.028408   6.731 1.69e-11 ***
season(9,12]               -0.136278   0.032411  -4.205 2.61e-05 ***
years_last_cntr(2,4]       -0.349565   0.075876  -4.607 4.08e-06 ***
years_last_cntr(4,6]        0.312832   0.167693   1.866 0.062110 .
years_last_cntr(6,8]        0.101470   0.415871   0.244 0.807234
years_last_cntr(8,10]       0.661083   0.760255   0.870 0.384544
years_last_cntr(10,12]     -6.371545  68.506850  -0.093 0.925899
DivisionEast South Central -0.103524   0.048948  -2.115 0.034432 *
DivisionMiddle Atlantic    -0.030242   0.030690  -0.985 0.324424
DivisionMountain           -0.035474   0.043953  -0.807 0.419614
DivisionNew England        -0.049311   0.044363  -1.112 0.266340
DivisionPacific            -0.051848   0.032090  -1.616 0.106154
DivisionSouth Atlantic     -0.112618   0.030951  -3.639 0.000274 ***
DivisionUnavailable         0.004958   0.063581   0.078 0.937847
DivisionWest North Central -0.026644   0.043872  -0.607 0.543645
DivisionWest South Central -0.096093   0.040131  -2.394 0.016644 *
freq                       -0.045992   0.003714 -12.385  < 2e-16 ***
due                        -0.838949   0.040359 -20.787  < 2e-16 ***
rly_due                    -0.957419   0.083842 -11.419  < 2e-16 ***
donate_prob                 0.909589   0.099905   9.105  < 2e-16 ***
I(CNDOL1/avg_don_life)     -0.407266   0.107652  -3.783 0.000155 ***
CONTRFST:CNTMLIF            0.010243   0.002640   3.879 0.000105 ***
CONLARG:CNMONL_BIN(2,4]    -0.160477   0.037569  -4.271 1.94e-05 ***
CONLARG:CNMONL_BIN(4,6]    -0.250816   0.047072  -5.328 9.91e-08 ***
CONLARG:CNMONL_BIN(6,8]    -0.197765   0.056901  -3.476 0.000510 ***
CONLARG:CNMONL_BIN(8,10]   -0.262440   0.084826  -3.094 0.001976 **
CONLARG:CNMONL_BIN(10,12]  -0.469577   0.161596  -2.906 0.003662 **
CONLARG:CNMONL_BIN(12,14]  -0.500064   0.302837  -1.651 0.098684 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77703  on 66133  degrees of freedom
Residual deviance: 69955  on 66084  degrees of freedom
AIC: 70055

Number of Fisher Scoring iterations: 9
```

Figure 4: Summary output for final multiple regression model

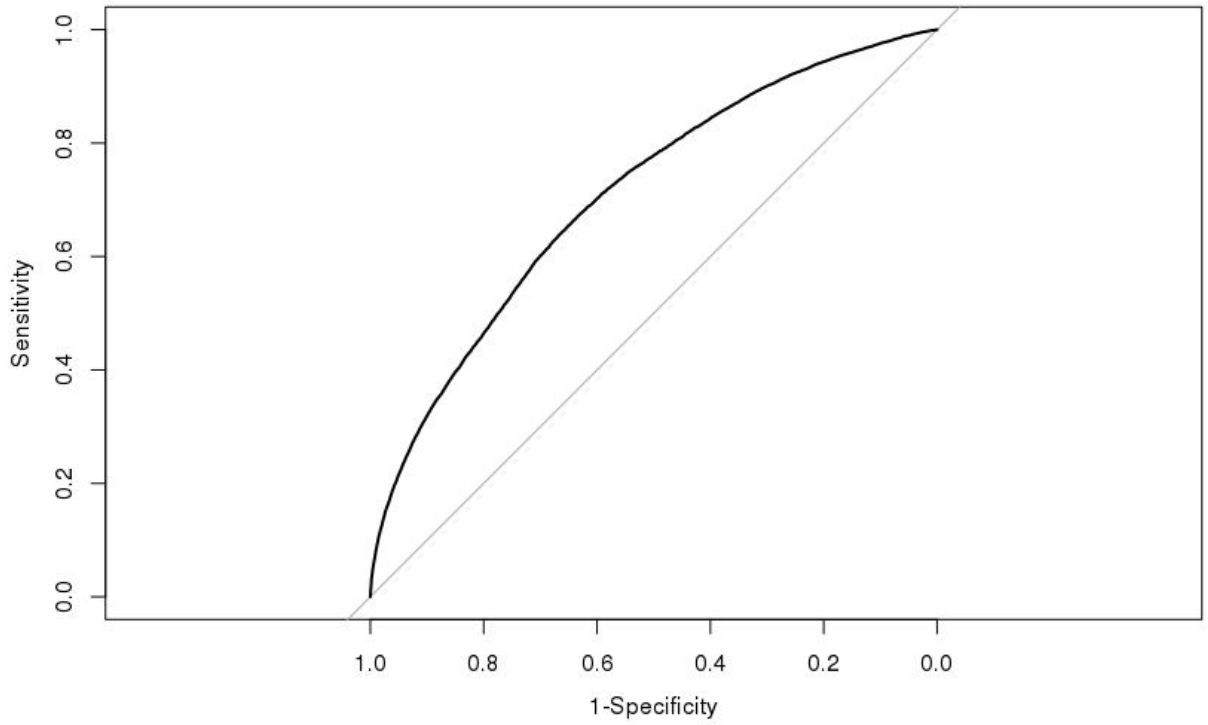
```
Call:
lm(formula = TARGDOL ~ CNDOL1 + CNTRLIF + CONLARG + CONTRFST +
    CNMONL + CNCOD1_TYPE + SLCOD1_TYPE + CNMONF_BIN + freq, data = train_set[-which(train_set$TARGDOL ==
    1500), ])

Residuals:
    Min      1Q  Median      3Q     Max
-68.290  -1.717  -0.370   1.143 174.680

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -7.803183   1.351122  -5.775 7.81e-09 ***
CNDOL1             0.659849   0.010160  64.945  < 2e-16 ***
CNTRLIF            0.007365   0.000810   9.093  < 2e-16 ***
CONLARG            0.104084   0.009327  11.159  < 2e-16 ***
CONTRFST           0.089558   0.010257   8.731  < 2e-16 ***
CNMONL             0.020638   0.001885  10.949  < 2e-16 ***
CNCOD1_TYPEA       1.653507   0.203015   8.145 4.05e-16 ***
CNCOD1_TYPEC       1.183133   0.413949   2.858 0.004266 **
CNCOD1_TYPED       1.973633   0.235776   8.371  < 2e-16 ***
CNCOD1_TYPEM       1.693377   0.427033   3.965 7.35e-05 ***
SLCOD1_TYPEA       6.554102   1.344861   4.873 1.11e-06 ***
SLCOD1_TYPEC       6.450190   1.604139   4.021 5.82e-05 ***
SLCOD1_TYPED       6.852832   1.351105   5.072 3.97e-07 ***
SLCOD1_TYPEM      11.118459   2.399526   4.634 3.62e-06 ***
CNMONF_BIN(2,4]   -0.575134   0.166271  -3.459 0.000543 ***
CNMONF_BIN(4,6]   -0.994276   0.166315  -5.978 2.30e-09 ***
CNMONF_BIN(6,8]   -1.197467   0.206760  -5.792 7.09e-09 ***
CNMONF_BIN(8,10]  -1.133490   0.211913  -5.349 8.96e-08 ***
CNMONF_BIN(10,12] -1.486601   0.238746  -6.227 4.87e-10 ***
CNMONF_BIN(12,14] -1.397081   0.207403  -6.736 1.68e-11 ***
freq               0.042766   0.008238   5.191 2.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.826 on 18117 degrees of freedom
Multiple R-squared:  0.6161,    Adjusted R-squared:  0.6157
F-statistic:  1454 on 20 and 18117 DF,  p-value: < 2.2e-16
```

Figure 5: ROC curve for binary logistic model




```
Call:
plot.roc.default(x = train_set$bin_TARGDOL, predictor = binlog$fitted.values,     xlab = "1-Specificity")

Data: binlog$fitted.values in 47995 controls (train_set$bin_TARGDOL 0) < 18139 cases (train_set$bin_TARGDOL 1).
Area under the curve: 0.7082
```

Figure 6: Top contributor summary

```
> summary(top_contributors$TARGDOL)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00   10.14   15.00  200.00
> sum(top_contributors$TARGDOL)
[1] 10137.23
```

Figure 7: Derived Features

[1] "CNCOD1_TYPE"
[2] "CNCOD2_TYPE"
[3] "CNCOD3_TYPE"
[4] "SLCOD1_TYPE"

```
 [5] "SLCOD2_TYPE"
 [6] "SLCOD3_TYPE"
 [7] "CNMON1_MTH"
 [8] "CNMON2_MTH"
 [9] "CNMON3_MTH"
[10] "years_last_cntr"
[11] "CNMONF_BIN"
[12] "CNMONL_BIN"
[13] "avg_don_life"
[14] "avg_don_latest"
[15] "Division"
[16] "Region"
[17] "season"
[18] "CNDOL1_bin"
[19] "CNTRLIF_bin"
[20] "CONLARG_bin"
[21] "CONTRFST_bin"
[22] "CNDOL2_bin"
[23] "CNDOL3_bin"
[24] "CNTMLIF_bin"
[25] "avg_don_life_bin"
[26] "avg_don_latest_bin"
[27] "freq"
[28] "due"
[29] "rly_due"
[30] "donate_prob"
```