Charity Direct Marketing Campaign

Final Group Project

Predict 422 – DL SEC 55, 2017FA

Critical Thinking Group 7

Conor Carey, Daniel Colvin, and Matt Thompson

Northwestern University

Table of Contents

1. Problem

A charitable organization wishes to maximize profits during their next fundraising campaign. The organization seeks a machine learning model that identifies which potential donor will contribute during this campaign, so they can be targeted.

The goal for this project is to develop a classification model, using data from the most recent campaign, which effectively supports the charitable organizations goals. Additionally, we are to develop a prediction model, based on the same data, to predict the gift amounts from donors.

We are to use modeling techniques from Predict 422 such as tree methods and linear models for this project – but we are not limited only to these techniques. The metrics used for judging model performance haven't been specified, it is incumbent upon us to specify which models we recommend and why.

2. Significance

The results from this project will facilitate the decision on which potential donors to contact during the upcoming direct marketing campaign. The cost of a mailing request is two dollars and, on average, ten percent of those who receive the mailing request respond by sending fourteen dollars and fifty cents. Based on this, it is not cost effective to just contact every potential donor because we would spend fifty-five cents on average, yielding a negative profit. Thus we need to identify the potential donors most likely to donate to maximize the amount of money collected for the charity. At the same time we need to identify those who are least likely to donate, so they aren't contacted, to minimize resources expended without return

The conclusions from the analysis will be limited to the given dataset, but the methodology can be applied to many cases. Taking this beyond predicting the likelihood of someone donating, models such as the one in this project can be used to predict a person's future actions such as using a coupon, voting, or refinancing a loan. When these predictions are accurate, it will allow someone to efficiently market to influence that person's decision.

3. Data

The data for this project was provided by the charitable organization which we are supporting. The raw data consisted of 24 variables: 20 explanatory variables, two predictor variables, one partition variable, and one ID variable. The ID variable is just a unique key that identifies each case or data point, which should not be used in any of the models. The explanatory variables describe things such as household income, gender, dollar amount of largest gift to data, months since last donation, and other information that impacts donations. The first predictor variable is

if they are a donor or not, and the second is the donation amount if they are a donor. Figure 1 shows a table of the explanatory and predictor variables, including their description.

Variable	Description
ID	ID number (reference only)
REG1-REG5	Geographic Region
HOME	Homeowner (1=Yes, 0=No)
CHLD	Number of children
HINC	Household income (7 categories)
GENF	Gender (0 = Male, 1 = Female)
WRAT	Wealth Rating
AVHV	Average Home Value in donor's neighborhood
INCM	Median Family Income in donor's neighborhood
INCA	Average Family Income in donor's neighborhood
PLOW	Percent categorized as "low income" in donor's neighborhood
NPRO	Lifetime number of promotions received to date
TGIF	Dollar amount of lifetime gifts to date
LGIF	Dollar amount of largest gift to date
RGIF	Dollar amount of most recent gift
TDON	Number of months since last donation
TLAG	Number of months between first and second gift
AGIF	Average dollar amount of gifts to date
DONR	Classification Response Variable (1 = Donor, 0 = Non - donor)
DAMT	Prediction Response Variable (Donation Amount in $).

Figure 1: Data Dictionary

In addition to the variables in Figure 1, there is a partition variable that identifies if the data point is part of the training set, validation set, or test set. Roughly 50% of the data is in the training set, which should be used to build the models. Roughly 25% of the data is in the validation set, which should be used to estimate performance of the models when operationally employed. The final 25% of data is the test set. The two predictor variables are unknown for the test set – we are to predict them using our models.

During exploratory data analysis (EDA), several issues with the data were observed. The first issue is variable type. All explanatory variables are "int" or integers in R. This is likely not appropriate since some variables are categorical. For example Household income (HINC) is specifically labeled as having 7 categories. If those categories don't have the same properties of a continuous variable then linear models might not perform as well as possible. There are two binary variables, gender (GENF) and homeowner (HOME), which are also integers but probably should be treated as categorical. Two additional variables, wealth rating (WRAT) and number of children (CHLD), are implied to be categorical as well. Finally, there are 4 variables to describe region (REG 1-4) that are just dummy variables for the categorical variable region. We are not correcting this issue for two reasons; provided code and severity of issue. The project sponsor, Prof B., provided code to standardize the explanatory variables – this code will not work if the

variables are not numeric. This issue really impacts linear models, which still may be adequate for prediction.

The next issue is correlation of explanatory variables. The correlations of all numeric variables (including the ones described in the previous paragraph) can be seen in Figure 2 below. The larger and darker the circle, the more correlated the two variables are. Correlation is a measure or how related, linearly, two variables are. With linear models we would like the explanatory variables to not be correlated with each other.

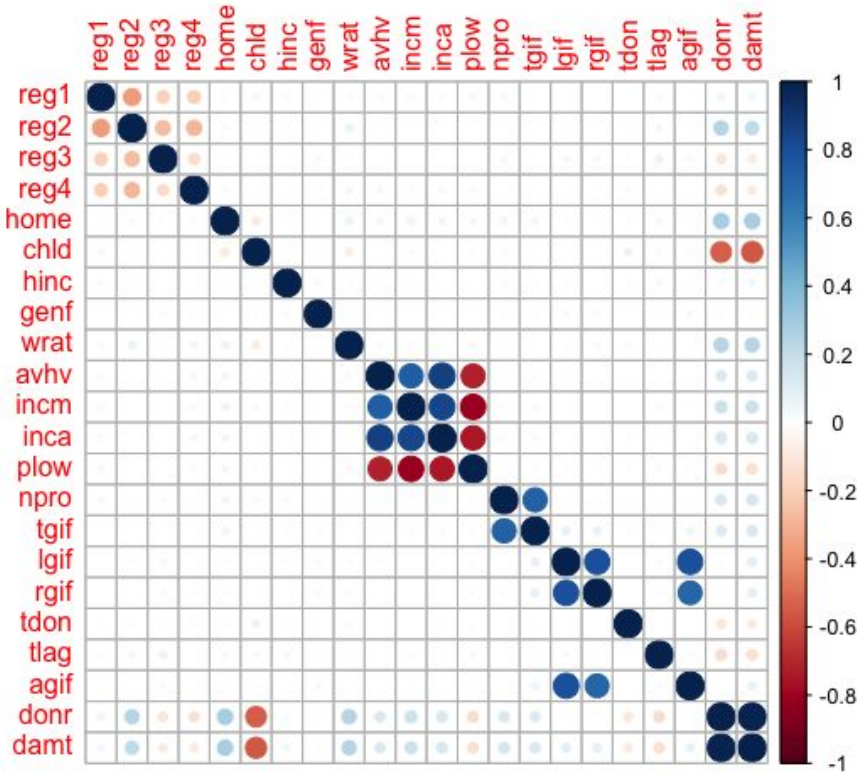

Figure 2: Variable Correlations

Depending on your training, correlations greater than 0.6-0.8 are cause for concern. We have 10 such correlations in our dataset:

- AVHV and INCM = 0.73
- AVHV and INCA = 0.84
- AVHV and PLOW = -0.63
- INCM and INCA = 0.87
- INCM and PLOW = -0.66
- INCA and PLOW = -0.64
- NPRO and TGIF = 0.71
- LGIF and RGIF = 0.70
- LGIF and AGIF = 0.62
- RGIF and AGIF = 0.71

Logically, all of these make sense that they are correlated so highly. AVHA, INCM, INCA, and PLOW all describe income/value of the donor's neighborhood. LGIF, TGIF, and RGIF all describe the amount of previous donations of the potential donor. When making variable

selections it will be important to see if both variables from any of the pairs above are included and if there is multicollinearity.

Most of the variables had outliers, Figure 3 shows a distribution of each variable, and we corrected them in two major ways, we either transformed the variable or we capped the outliers. Our preference was to transform, but sometimes we were unable to find a suitable transformation. We knew the predictor variables did not have to be normally distributed but we knew there were benefits if they were. For transformations we considered log, square root, and an "optimized" power transform using the BoxCox() function in R. When we were unable to successfully perform a transform we capped the data at the 5th and 95th percentile. The logic here was that our prediction model will predict the average or normal case, so it won't be sensitive to outlier prediction.

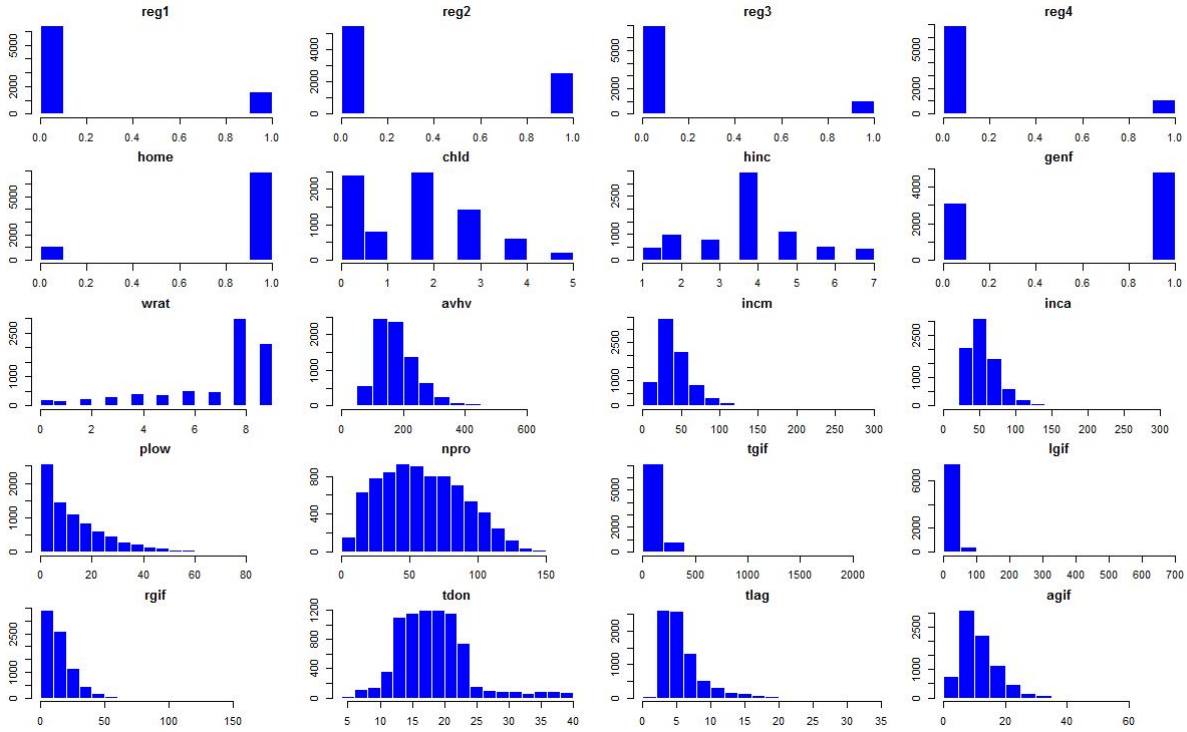

Figure 3. Explanatory Variable Histograms

Our response variable for prediction, Donation Amount (DAMT), was slightly skewed but we chose not to perform a transformation on it. If a linear model was chosen for the prediction model then a transform might be necessary to meet the normality assumption. In this case we recommend a log transform due to the simplicity to back transform predictions - we don't necessarily believe this was the most appropriate transform but it would be valid and easy to implement.

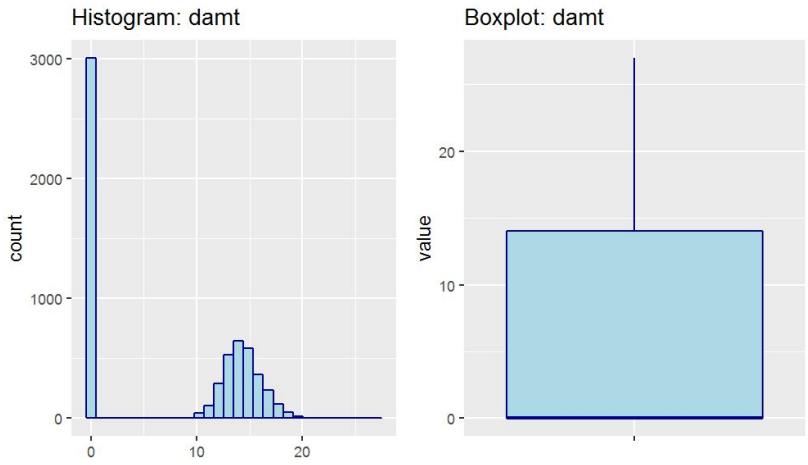

Figure 4. Histogram & Box Plot of Donation Amount

4. Classification Models

This section describes the types of models created for best classifying the response variable. The performance of these models will be summarized in a later section.

4.1 Logistic Regression

The logistic regression model aims to classify a response variable by indicating the probability, or odds, of it being a zero or a one. We manually selected, using a process similar to backwards automatic variable selection, which variables to include using a 95% confidence threshold. We built models using the explanatory variables as given, transformed variables as discussed in the outlier section, and non-linear variables (limited to second order). All logistic models had similar performance but our best logistic model used ten variables, including those which were transformed as well as one quadratic variable (the squared values of "TDON").

4.2 Linear Discriminant Analysis (LDA)

Linear discriminant analysis attempts to split the classified responses with a straight line and calculates estimates of the probabilities using Bayes' Theorem. As with the logistic regression model, we built models using the explanatory variables as given, transformed variables as discussed in the outlier section, and non-linear variables (limited to second order). Again each model had similar performance but the best linear discriminant analysis model used the transformed variables and added quadratic terms of some of the predictor variables.

4.3 Quadratic Discriminant Analysis (QDA)

Quadratic discriminant analysis is similar to LDA in that it also separates the classes and uses Bayes' Theorem to estimate the probabilities, however it assigns each class its own covariance matrix and the separation line has a quadratic, nonlinear curve to it. We followed that same process as describe in LDA above. Our best QDA model contained transformed predictor variables but did not contain any quadratic terms.

4.4 K-Nearest Neighbors

K-nearest neighbors attempts to classify the response based on the majority class surrounding an observation. Our k-nearest neighbors models were constructed using scaled values of the predictor variables, so that no predictor variable outweighed another. Furthermore, the best k-nearest neighbor used k = 7 nearest neighbors.

4.5 Bagging

Bagging, or bootstrap aggregation, is an ensemble method which creates many random sub-samples of the training data (with replacement), generates a decision tree on each sub-sample, and calculates the average prediction. Our best bagging model used all unscaled, non-transformed predictor variables. We limited our bagging model to five hundred decision trees.

4.6 Random Forest

A random forest is an ensemble of many decision trees, like bagging. What differentiates it is that every new split in each decision tree reconsiders every predictor variable. This often provides an advantage over bagging because it is a less greedy approach. Our best random forest model used the same variables as with bagging and the average of five hundred decision trees.

4.7 Boosting

Boosting can potentially make further improvement from bagging because it avoids overfitting by learning slowly from each decision tree. It does this by creating decision trees on the residuals instead of on the response and applying this to the next decision tree as a weighting factor. Boosting models can be tuned further by specifying the response distribution and the interaction depth (the number of splits to perform on each tree). We initially build boosting models with 5,000 trees and considered depths ranging from 1 to 10. Performance appeared to level off after a depth of 6. We then considered additionally trees, first 8,000 then 10,000 with marginal improvement in performance. Our best boosting model used ten thousand trees, an interaction depth of six, and specifying the distribution as "Bernoulli".

4.8 Support Vector Classifier (SVM)

A support vector classifier separates the response classes using a hyperplane, which maximizes the space between response classes. A threshold parallel to the separation, called the margin, identifies whether a classification is estimated with confidence. An important parameter in this model is the type of kernel, or the shape of the separation (linear, polynomial, or radial). Also, a cost parameter varies the distance of the margin to the separating hyperplane. Our best SVM model used a radial kernel and a cost of five.

5. Prediction Models

Prediction models were created to predict the "DAMT" response variable, which describes the dollar amount of donations.

5.1 Multiple Linear Regression (MLR)

Multiple linear regression is a simple approach which uses the relationship between the response and the predictor variables to make a prediction. If there are improved correlations when combining multiple predictors, this model could perform well. Our best multiple linear regression model used exhaustive subset selection to select the best variables based on the AIC metric.

5.2 Ridge Regression

Ridge regression is similar in form to multiple linear regression except that it weights the coefficients using a shrinkage parameter. The shrinkage parameter keeps all predictor variables originally in the model but the weights can make the coefficients very close to zero. Our best ridge regression model used all of the predictor variables.

5.3 Lasso Regression

LASSO also tries to improve the least squares estimator by adding constraints to the value of the coefficients. This is done to allow only those explanatory variables that are large contributors to the dependent variable to have large coefficients. This prevents a variable with minimal impact on the outcome of having a very large positive/negative coefficient which could dramatically impact the response. Our best lasso regression model used all of the predictor variables.

5.4 Other

In addition to the prediction models described above, we also created a k-nearest neighbors, random forest, and SVM models to predict "DAMT". Descriptions of these models are found in section 4.

6. Performance / Accuracy

Classification Models

We evaluated each classification model's performance by training them on the training subset and testing them on the validation subset. We used three metrics to compare performance: area under the curve (AUC), response accuracy rate, and total estimated profit. Figure 5 below summarizes the performance of our models described in section 5.

AUC is one the standard metrics for classification analysis used to determine which models predict classes best. AUC cannot exceed 1.0 and we prefer models that have an AUC as close to 1.0 as possible. The curve referred to in AUC is the ROC curve, and we used the ROCR package in R to perform the calculation.

Accuracy rate is similar to AUC because it cannot exceed 1.0 and is an indicator of model performance. We calculated accuracy rate using the following equation on the validation set

$$accuracy\ rate = \frac{total\ \#\ correctly\ classified}{total\ \#\ validation\ points}$$

Estimated profit is the profit we would make if we followed the models recommendations, mailing to those we predict to donate and ignoring those we predict won't donate. We assumed the average donation would match the historical average of $14.50.

Name	Description	AUC	Accuracy Rate	Estimated Profit
Logistic 1		0.914	83.80%	$11,407
Logistic 2	variable transformed	0.913	83.89%	$11,365
Logistic 3	non-linear terms	0.921	84.34%	$11,425
Logistic 4	transform & nonlinear	0.922	84.39%	$11,426
LDA1		0.914	83.99%	$11,368
LDA2	variable transformed	0.914	83.55%	$11,357
LDA3	non-linear terms	0.922	84.19%	$11,370
LDA4	transform & nonlinear	0.923	84.49%	$11,402
QDA1		0.905	83.50%	$11,238
QDA2	variable transformed	0.913	83.25%	$11,269
QDA3	non-linear terms	0.884	81.67%	$11,000
QDA4	transform & nonlinear	0.890	80.53%	$10,885
KNN	K=7	0.798	79.63%	$10,811
Bagging	500 trees	0.953	89.05%	$11,715
Random Forest	500 trees	0.959	89.10%	$11,736
Boost1	5000 trees, depth 4	0.961	89.20%	$11,836
Boost2	5000 trees, depth 2	0.949	87.17%	$11,763
Boost3	5000 trees, depth 6	0.965	90.39%	$11,877
Boost4	5000 trees, depth 8	0.966	90.63%	$11,871
Boost5	10000 trees, depth 6	0.970	90.88%	$11,922
SVM1	kernel = linear	0.835	83.50%	$10,546
SVM2	kernel = polynomial	0.835	83.40%	$10,795
SVM3	kernel = radial	0.880	88.01%	$11,132

Figure 5: Summary of classification model performance.

The values are color coded based on their performance compared to each other, so dark green is the best, red is the worst, and yellow is between the two. Based on this chart we can see that overall the boosting models performed best, and the ten thousand, interaction depth = 6 model had the highest AUC, response accuracy rate, and total estimated profits. Based on these results, we would recommend this boosting model as the best classification model, even though it is likely the least interpretable model as well.

Prediction models

Similar to classification model performance evaluation, we trained each prediction model on the training subset and tested it on the validation set to compare models. Our main metric for prediction accuracy was the mean squared error (MSE) due to its simple execution across the various types of models. Figure 5 below summarizes the performance of our models described in section 5.

Name	Description	MSE
MLR1	All variables	1.623
MLR2	Subset selection	1.624
Lasso	All variables	1.627
Random Forest	All variables	1.704
Boost1	5000 trees, depth 6	1.437
Boost2	10000 trees, depth 6	1.384
SVM	kernel = radial	1.59

Figure 5: Summary of prediction model performance

The values are again color coded based on their performance compared to each other, so dark green is the best, red is the worst, and yellow is between the two. Based on this chart we can see that overall the boosting models performed best, and the ten thousand, interaction depth = 6 model had the lowest mean squared error. Based on these results, we would recommend this boosting model as the best prediction model, even though it is likely the least interpretable model as well.

7. Limitations

As with any model, given more time and resources we would like to be able to improve the predictability of our best model. Additionally we would like to explore other techniques that weren't specifically covered in this course.

Due to the provided code, we assumed variable types, specifically treating none as categorical, which likely limited the performance for the linear models, both classification and prediction.

8. Results

Our final boost classification and prediction models were applied to the test subset of the data. The output reveals that the organization should mail 324 donors for a total cost of $648. Based on these targeted donors, the donation amount should range from $10 to $18 (Figure 6), which yields approximately $5,230 in profits. Using these predictions, the mailing response rate increases from 10% (mailing random individuals) to approximately 16%, an increase of over 60%.

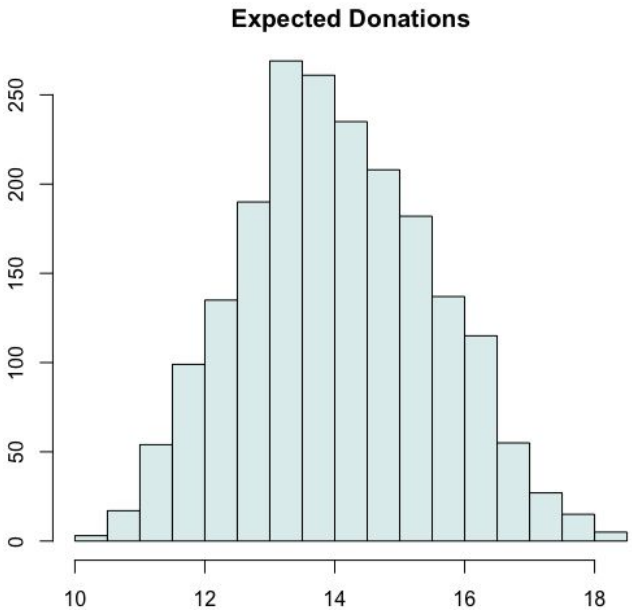

Figure 6: Distributions of test subset donation amounts

9. Conclusions

The goal of this analysis was to find the best classification model to maximize donation profits at a charitable organization and find the best prediction model that will predict the dollar amount of donations given. After exploring many types of models and optimizing the tuning parameters as best as time allowed, we are confident in our final resulting models. Even so, more time on this analysis would likely have allowed us to explore the data in even more detail and explore more model types and combinations to improve model classification and predictions. Future work would include applying our models within the organization.

10. References

James, Gareth et. al. 2017. *An introduction to Statistical Learning with applications in R*. New York: Springer Science+Business Media.