Bruckner, Funk, Sheets, Zimmerman
12/04/16

PREDICT 422 – SEC 56
Professor N. Bastian

Final Project

Introduction

In this report we develop machine-learning models to optimize a charitable organization’s direct marketing campaign. Our goal is two-fold: to identify likely donors using a classification model in order to maximize the net profit from the mailings, and to estimate the gift amounts from donors using a prediction model. We create several classification models, including logistic regression, logistic regression general additive model (GAM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (KNN), decision tree, boosted trees, random forests, and support vector machines (SVM), in order to determine which model best identifies likely donors. We also create several prediction models, including least squares regression, best subset selection with k-fold cross-validation, principal components, partial least squares regression, ridge regression, and lasso in order to predict the amount of donations.

Data Exploration

The data we are using for this report is the charity data set. This data set contains 3,984 training observations, 2,018 validation observations, and 2,007 test observations. The data has been weighted in order to make the training and validation samples each contain roughly an equivalent number of donors and non-donors. The charity data set includes two response variables, donr (donor) and damt (donation amount), for the classification models and prediction models, respectively. It also contains 20 predictor variables, including the geographic regions of previous donors, whether the donor is a homeowner or not, the number of children the donor has, the donor’s income, the donor’s gender, the amounts of previous donations, and various other wealth, income, and donation factors that may help create strong classification and prediction machine-learning models.

Prior to developing any models, we examine our data in order to see if there are any missing values or abnormalities in the data. Although the data set does not contain any missing values, several of the variables might benefit from transformations. We choose to apply logarithmic transformations to seven right-skewed variables: avhv, incm, inca, tgif, lgif, rgif, and agif. The top row of **Figure 1** shows the variables prior to transformation, while the bottom row shows the variables after transformation.[i] As the figure shows, the logarithmic transformations help normalize the distributions of these seven variables.

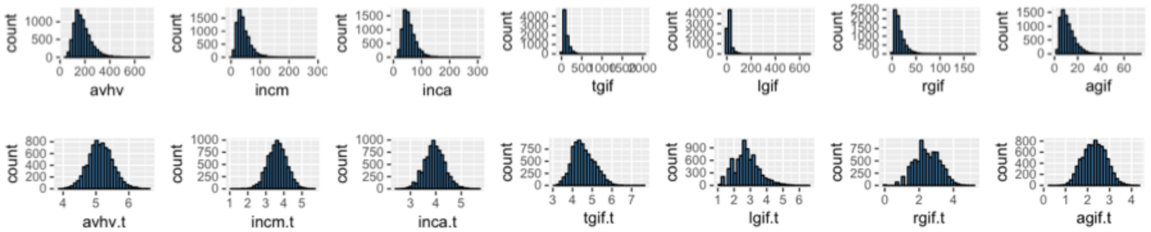

Figure 1: Transformation of Variables

[i] Please note: We keep the transformed variable names the same as the original variable names for our models, but we added a “.t” to the end of each variable name in Figure 1 to clearly distinguish between the original and transformed variables.

After transforming our variables, we create a correlation matrix to gain further insight into the data. **Figure 2** shows a visualization of the correlation matrix, while **Table 1** shows the values of the most statistically significant correlations.

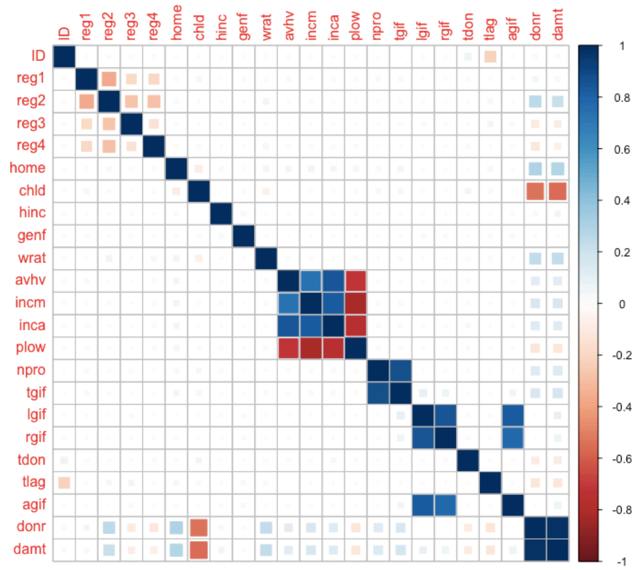

Figure 2: Correlation Matrix

Variable 1	Variable 2	Correlation
damt	donr	0.9817018
tgif	npro	0.8734276
rgif	lgif	0.8512241
inca	avhv	0.8484572
inca	incm	0.8296747
agif	lgif	0.8294224
plow	incm	-0.8120381
agif	rgif	0.7706645
plow	inca	-0.7510141
incm	avhv	0.7304313
plow	avhv	-0.7187952
damt	chld	-0.5531045
donr	chld	-0.5326077

Table 1: Significant Correlations

As **Figure 2** and **Table 1** reveal, certain variables have strong correlations. For instance, tgif (dollar amount of lifetime gifts to date) has a very strong positive correlation with npro (lifetime number of promotions received to date). Some variables, such as plow (percent categorized as “low income” in potential donor's neighborhood) and incm (median family income in potential donor's neighborhood in $ thousands), have a strong negative correlation. The correlations are something we bear in mind as we create our classification and prediction models.

Classification Models and Analysis

Model Technique 1: Logistic Regression

The first classification technique we use is logistic regression, a simple and robust technique for two response classes that is easily interpretable and applicable. The first of four logistic models includes all independent variables; the second model utilizes only those variables which are significant at p=0.05; the third model includes all independent variables as well as additional quadratic terms for numerical variables; and the fourth model is created using backward subset selection on the third model. The backward subset selection results in a 14-variable model. We use these 14 variables, shown in **Table 2**, for the fourth model and subsequent model testing as a comparison against models fit using all original variables. The most profitable model is the fourth model, yielding $11,668.00 in donations through 1,329 mailings. **Table 3** depicts the classification results and number of mailings for each of the four models.

Variable	Estimate	Signif.
(Intercept)	0.95601	***
reg1	0.66174	***
reg2	1.49590	***
home	1.38532	***
chld	-2.39491	***
I(hinc^2)	-1.09223	***
wrat	0.48869	***
I(wrat^2)	-0.42120	***
incm	0.69251	***
tgif	0.63898	***
I(tgif^2)	-0.06932	.
lgif	-0.21454	*
tdon	-0.30159	***
tlag	-0.55016	***
agif	0.15132	

Table 2: Backward Subset Selection Coefficients

Model	Validation Class			
		0	1	Number of Mailings
Model 1: All independent variables	0	600	21	1397
	1	419	978	
Model 2: All significant independent variables	0	580	20	1418
	1	439	979	
Model 3: All independent variables and additional terms	0	675	13	1330
	1	344	986	
Model 4: Backward subset selection	0	672	11	1329
	1	347	988	

Table 3: Classification Matrix with Mailings

Model Technique 2: Logistic General Additive Model

The GAM approach to modeling creates a function for each variable, allowing non-linearity in the model. This approach can be used for classification problems easily with a logistic model. The first logistic GAM model is fitted using all independent variables, and the second model fitted uses the 14 variables from the backward subset selection logistic model. The first model yields a profit of $11,389.00 with 1,396 mailings, and the second model outperforms the first, yielding a profit of $11,668.00 with 1,329 mailings. **Table 4** depicts the classification results and number of mailings for both models. Additionally, the second model's coefficients are identical to the backward subset logistic regression model's coefficients from the previous section.

Model	Validation Class			
		0	1	Number of Mailings
Model 1: All independent variables	0	601	21	1396
	1	418	978	
Model 2: 14 backward subset selection variables	0	678	11	1329
	1	341	988	

Table 4: GAM Classification Matrix with Mailings

Model Technique 3: Linear Discriminant Analysis

LDA is a classification method similar to logistic regression that utilizes Bayes' theorem as a classifier, and while it is theoretically more complex than the logistic model, in application it is similar in interpretability. The first LDA model is fit using all independent variables, yielding a profit of $11,362.00 from 1,395 mailings. A second LDA model fit using the 14 variables from the backward subset selection logistic model yields a higher profit of $11,647.50 from 1,361 mailings. The classification results and number of mailings for each model are in **Table 5**.

Model	Validation Class			
		0	1	Number of Mailings
Model 1: All independent variables	0	600	23	1395
	1	419	976	
Model 2: 14 backward subset selection variables	0	649	8	1361
	1	370	991	

Table 5: LDA Classification Matrix with Mailings

While the best LDA model performs well, it is not as profitable nor as accurate as the best logistic or logistic GAM models.

Model Technique 4: Quadratic Discriminant Analysis

QDA is a classification method almost identical to LDA, except that the covariance matrix is assumed to be specific to each class. This technique can have a more flexible fit and tends to perform well on large datasets. The first QDA model is fit using all independent variables, followed by a second model utilizing the 14 variables from the backward subset selection method used in prior sections of this study. Both models are much less accurate and significantly less profitable for similar numbers of mailings than previous modeling techniques. **Table 6** shows the classification results and number of mailings for each model. The second model required 1,430 mailings and yielded $11,074.50 in profits, requiring more mailings for less profit than the first model, which had a profit of $11,263.50 from 1,379 mailings.

Model	Validation Class			
		0	1	Number of Mailings
Model 1: All independent variables	0	607	32	1379
	1	412	967	
Model 2: 14 backward subset selection variables	0	550	38	1430
	1	469	961	

Table 6: QDA Classification Matrix with Mailings

Model Technique 5: K-Nearest Neighbor Classification

Although not necessarily the most interpretable of algorithms, the KNN classification method is strong, as it is one of the few completely non-parametric methods available for classification. Due to the lack of insight into the exact decision boundary for whether an individual is a donor or not, this technique is certainly one that should be tested. For this process, three models are fit utilizing different values of K: 1, 10 and 100. Although the classifier with K = 100 performs the best, we expect this since it will likely have a much higher bias towards the training data set. Despite this, it outperforms the other two classifiers on the validation data, maximizing profit at $11,299.50 through 1,390 mailings.

Model	Validation Class			
		0	1	Number of Mailings
K = 1	0	738	164	1116
	1	281	835	
K = 10	0	709	59	1250
	1	310	940	
K = 100	0	600	28	1390
	1	419	971	

Table 7: KNN Classification Matrix with Mailings

A tradeoff in profit and accuracy exists between the K=10 and K=100 models. Although the K=100 model is most profitable, the K=10 model is most accurate, as **Table 7** shows. The larger profit ultimately results in many more wasted mailings as well.

Model Technique 6: Decision Tree

Tree-based methods are simple and useful for interpretation, which is why we decide to use a simple decision tree to classify whether or not an individual is a donor. Since the process of building a decision tree includes a technique for pruning the resulting tree, all 20 predictor variables are passed in for model fitting. Pruning the tree ensures the original tree is not limited in performance due to over-fitting by the training data. The initial decision tree fit results in 1,319 mailings and a profit of $11,383.50. It is shown in **Figure 3**.

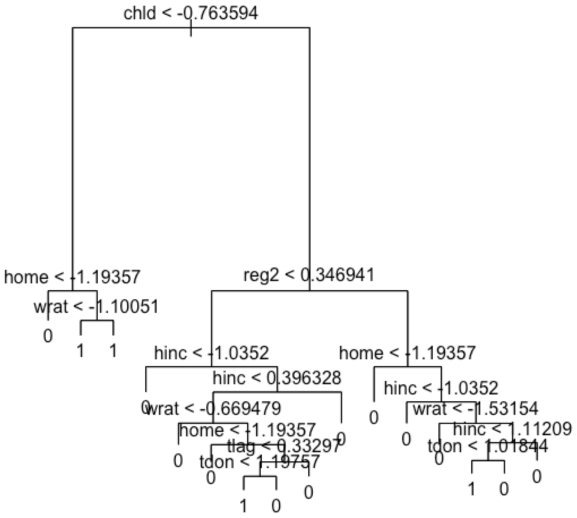

Figure 3: Decision Tree

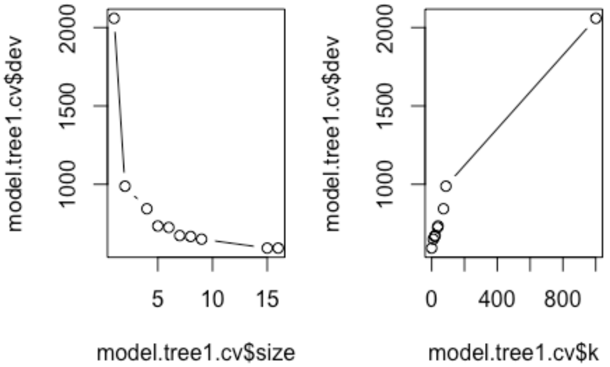

Figure 4: Decision Tree Plots

Although **Figure 4** demonstrates that the best decision tree is of size 10, the resulting tree from fitting that model is the exact same as the original decision tree model fit, capping this technique off at a $11,383.50 profit.

Model Technique 7: Random Forest

Having observed that the single, pruned decision tree still has not maximized profit, we try a method for aggregating several decision trees to enhance the performance of a single tree.

Three common techniques can achieve this—bagging, boosting, and random forests. Of these, the first of two techniques we test for this analysis is the random forest method. The model fit results in only 1,040 total mailings, maxing out its profit at $11,028.00. Although this technique is demonstrating that it is not the best in terms of profit, **Figure 5** below is incredibly useful in that it provides a clear visualization of the variables that are contributing the most to whether or not an individual will donate. Whether the donor has children is enormously important to their likelihood of donating.

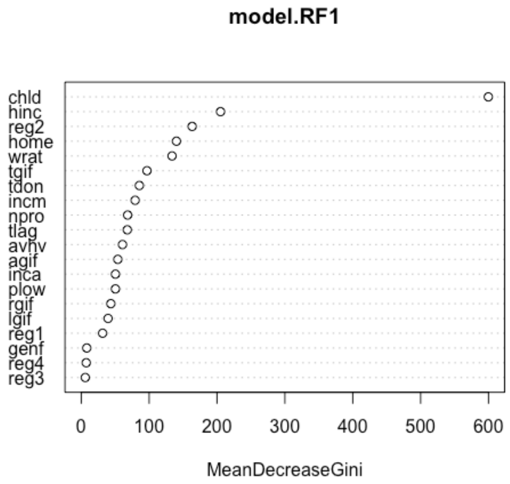

Figure 5: Mean Decrease in Gini Plot

Model Technique 8: Boosted Tree

With the resultant random forest performing so poorly in contrast to the individual decision tree, we attempt one additional multiple tree aggregation method. This method, referred to as boosting, learns more slowly and tends to reduce the likelihood of over-fitting as well. The boosted tree results in a total of 1,344 mailings, soaring the resulting profit to $11,594.50. Of all the tree-based classification methods we use, this is by far the best performer. In **Figure 6** below, we observe a similar visualization of the variables based on their relative influence on the classification of a donor.

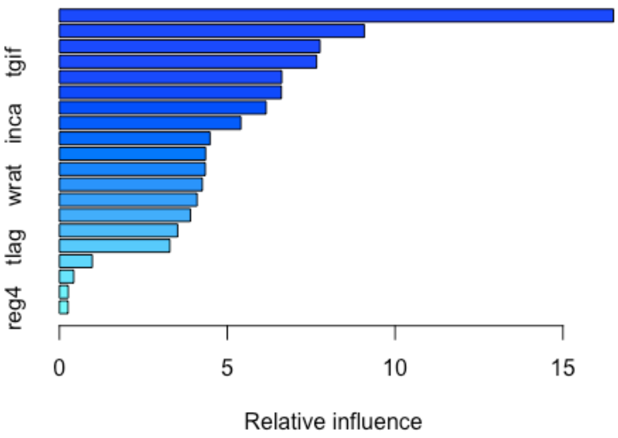

Figure 6: Relative Influence Plot

Table 8 below shows that the most important variable is how many children the potential donor has, which echoes the conclusion of the other two tree-based models.

Variable	Relative Influence
chld	16.5050114
agif	9.0830404
avhv	7.7627473
tgif	7.6615290
hinc	6.6249256
npro	6.6128775
incm	6.1582047
inca	5.4056824
tdon	4.4863957
plow	4.3549452
lgif	4.3471496
wrat	4.2586813
reg2	4.1002435
rgif	3.9026428
home	3.5211727
tlag	3.2892911
reg1	0.9795926
genf	0.4274159
reg3	0.2622024
reg4	0.2562489

Table 8: Variable Relative Influence

Model Technique 9: Support Vector Machine

The last classifier we test on this data is the SVM. We first run the SVM without any tuning by utilizing a linear kernel and then running it with cross-validation to determine optimal parameters. We then tune it using a more likely applicable radial kernel. As expected, the first model tanks in comparison to all previously fit classification models with a maximum profit of $10,459.00, which is mostly due to suggesting far too many mailings at 2,013. The tuned model using the radial kernel performs substantially better, raking in a potential profit of $11,536.00 on the validation data through only 1,366 mailings.

Prediction Models and Analysis

Model Technique 1: Least Squares Regression

We fit several least squares regression models by finding the models with the best Adjusted R-Squared, Mallows' Cp, or BIC values. The model with the best Adjusted R-Squared value contains 15 predictors; the model with the best Mallows' Cp contains 14 predictors; and the model with the best BIC contains 11 predictors. **Table 9** below is a comparison of how each of these least squares regression models performs on the validation data.

Model	MPE	Standard Error
Adjusted R-Squared	1.558254	0.1608325
Mallows' Cp	1.558577	0.1606044
BIC	1.562970	0.1604673

Table 9: Least Squares Regression Model Comparison

In terms of MPE, the Adjusted R-Squared model outperforms the other two models. However, the Mallows' Cp model has a nearly identical MPE and is a simpler model than the Adjusted R-Squared model since it uses one fewer predictor. It also has a slightly better Standard Error. For these reasons, we consider the Mallows' Cp model the best least squares regression model.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	14.18091	0.04033	351.616	< 2e-16	***
reg2	-0.05640	0.02848	-1.98	0.0478	*
reg3	0.33532	0.03221	10.410	< 2e-16	***
reg4	0.66223	0.03305	20.036	< 2e-16	***
home	0.22695	0.05556	4.085	0.0000459	***
chld	-0.59560	0.03424	-17.394	< 2e-16	***
hinc	0.51062	0.03645	14.009	< 2e-16	***
genf	-0.05967	0.02613	-2.283	0.0225	*
incm	0.43912	0.04425	9.924	< 2e-16	***
plow	0.39674	0.04788	8.287	< 2e-16	***
tgif	0.20163	0.02673	7.543	6.93E-14	***
lgif	0.38867	0.05856	6.637	4.13E-11	***
rgif	0.47472	0.05093	9.321	< 2e-16	***
tdon	0.07110	0.03205	2.218	0.0266	*
agif	0.38598	0.04847	7.963	2.81E-15	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Table 10: Coefficients for Mallows' Cp Model

Table 10 shows the coefficient estimates for the Mallows' Cp model. We see that all of the predictors are statistically significant.

Model Technique 2: Best Subset Regression with k-Fold Cross Validation

For this model, we use 20-fold cross-validation to find the best subset selection model with the lowest mean cross-validation error. **Figure 7** shows that the model with 18 variables has the lowest mean cross-validation error, and **Table 11** shows the errors.

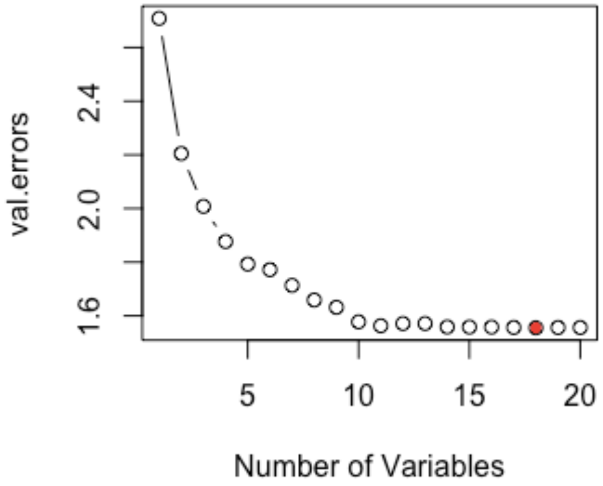

Figure 7: Best Subset Selection Using Cross-Validation

MPE	Standard Error
1.555443	0.1611711

Table 11: Ridge Regression Results

Table 12 shows the model coefficients. All the variables except reg1, wrat, inca, and tlag are statistically significant.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	14.18688	0.04360	325.424	< 2e-16	***
reg1	-0.03719	0.03620	-1.027	0.3044	
reg2	-0.08488	0.03932	-2.159	0.031	*
reg3	0.31578	0.03701	8.531	< 2e-16	***
reg4	0.64187	0.03808	16.854	< 2e-16	***
home	0.22316	0.05568	4.008	0.0000636	***
chld	-0.59284	0.03487	-17.001	< 2e-16	***
hinc	0.51069	0.03654	13.975	< 2e-16	***
wrat	0.02196	0.03807	0.577	0.5641	
genf	-0.05986	0.02614	-2.290	0.0222	*
incm	0.41271	0.05625	7.338	3.16E-13	***
inca	0.03616	0.04951	0.730	0.4652	
plow	0.40391	0.04952	8.157	6.03E-16	***
tgif	0.20043	0.02677	7.488	1.05E-13	***
lgif	0.38749	0.05862	6.611	4.91E-11	***
rgif	0.47543	0.05096	9.329	< 2e-16	***
tdon	0.07232	0.03208	2.255	0.0243	*
tlag	0.02516	0.03090	0.814	0.4157	
agif	0.38728	0.04850	7.985	2.36E-15	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 12: Coefficients for Best Subset Selection Using Cross-Validation

Model Technique 3: Principal Components

In order to choose the number of principal components, we looked at **Figure 8**. The lowest MSEP value is around 20 components, but the “elbow” of the plot appears around 5 components. We fit models using 5, 15, and 20 components in order to determine the optimal number of components.

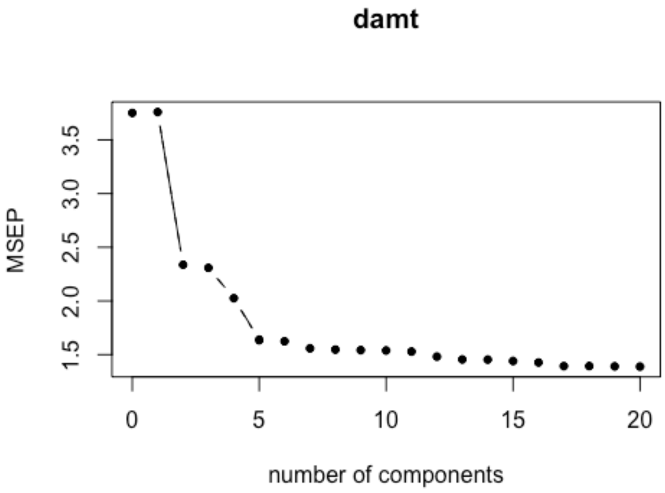

Figure 8: Principal Components Plot

As **Table 13** shows, the model using 20 principal components performs best, as it has the smallest MPE. However, we could choose the model containing 15 components if simplicity was a priority rather than having a slightly better MPE.

Number of Components	MPE	Standard Error
5	1.812705	0.1688532
15	1.598671	0.1607489
20	1.556378	0.1612150

Table 13: Principal Components Model Comparison

Model Technique 4: Partial Least Squares

Similar to the principal components method, the partial least squares technique requires us to select the number of appropriate components based on the following plot.

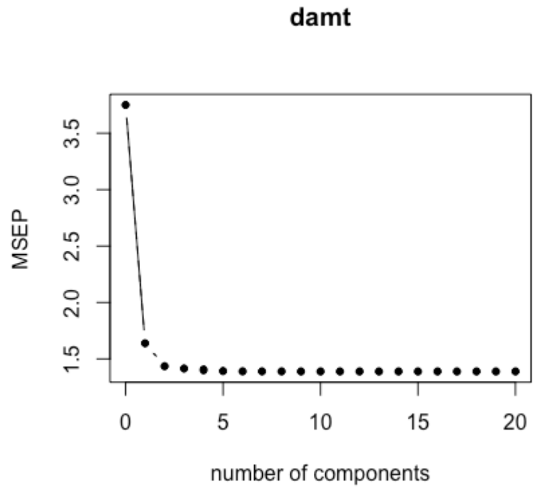

Figure 9: Partial Least Squares Plot

As seen in **Figure 9**, there is a large drop down to 3 components, but not much movement after that. This leads us to believe that 3 components are sufficient. The results for this model are in **Table 14** below.

Number of Components	MPE	Standard Error
3	1.592151	0.1613484

Table 14: Partial Least Squares Results

Model Technique 5: Ridge Regression

For the ridge regression model, we first select the best lambda. The MSE is smallest when log(Lambda) is smallest, as shown in **Figure 10**. The first vertical line from the left of the plot signifies the smallest MSE, while the second vertical line represents one standard deviation from the minimum MSE. The number 20 at the top of the plot indicates that all 20 predictor variables are present in the model regardless of the lambda value.

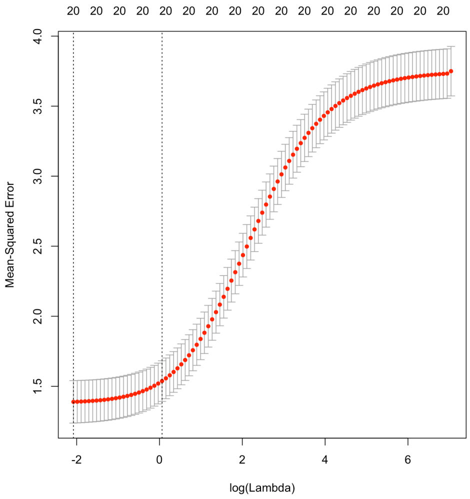

Figure 10: Ridge Regression Cross-Validation Plot

We set the best lambda equal to 0.1252296 to estimate the coefficients in **Table 15.**

Variable	Estimate
(Intercept)	14.22344033
reg1	-0.06495925
reg2	-0.12240428
reg3	0.28014764
reg4	0.58541542
home	0.20548728
chld	-0.54971631
hinc	0.47980195
genf	-0.05763863
wrat	0.01231801
avhv	-0.01751383
incm	0.30239430
inca	0.06359597
plow	0.30357092
npro	0.02575266
tgif	0.16534000
lgif	0.39461843
rgif	0.44561459
tdon	0.07087720
tlag	0.02834013
agif	0.38205327

Table 15: Coefficient Estimates

We then use this model to make the predictions and compute the errors in **Table 16**.

MPE	Standard Error
1.572113	0.1627705

Table 16: Ridge Regression Results

Model Technique 6: Lasso

The last prediction technique we use is lasso. The second vertical line from the left in **Figure 11** indicates the number of predictors in the model is 10, as suggested by the numbers at the top of the plot.

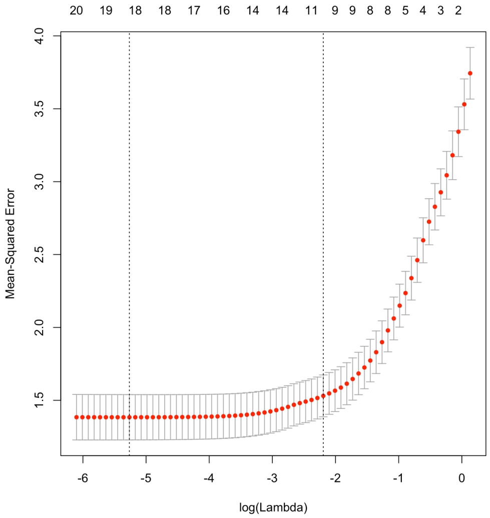

Figure 11: Lasso Cross-Validation Plot

We select the best lambda using cross-validation. In this case the best lambda is 0.005174504, which produces the coefficient estimates shown in **Table 17.**

Variable	Estimate
(Intercept)	14.338428658
reg1	.
reg2	-0.051092397
reg3	0.203743788
reg4	0.525828946
home	.
chld	-0.430556714
hinc	0.338995574
genf	.
wrat	.
avhv	.
incm	0.005835249
inca	.
plow	.
npro	.
tgif	0.076711998
lgif	0.380005233
rgif	0.399168041
tdon	.
tlag	.
agif	0.332079318

Table 17: Coefficient Estimates

Using this lambda, we make the predictions and compute the errors shown in **Table 18**.

MPE	Standard Error
1.562046	0.1611463

Table 18: Lasso Results

Results

Classification Models

Nine classification techniques yield several strong candidates to accurately identify donors while maximizing profit. **Table 19** below shows the results of each classification technique.

Model Technique	Number of Mailings	Profit
Logistic (all variables)	1397	$11,387.00
Logistic (significant variables only)	1418	$11,359.50
Logistic (all variables + additional terms)	1330	$11,637.00
Logistic (best subset)	1329	$11,668.00
Logistic GAM (all variables)	1396	$11,389.00
Logistic GAM (best subset)	1329	$11,668.00
Linear Discriminant Analysis (all variables)	1395	$11,362.00
Linear Discriminant Analysis (best subset)	1361	$11,647.50
Quadratic Discriminant Analysis (all variables)	1379	$11,263.50
Quadratic Discriminant Analysis (best subset)	1430	$11,074.50
K-Nearest Neighbors (K=1)	1116	$9,875.50
K-Nearest Neighbors (K=10)	1250	$11,130.00
K-Nearest Neighbors (K=100)	1390	$11,299.50
Decision Tree (unaltered and pruned)	1319	$11,383.50
Decision Tree (best subset)	1373	$11,333.50
Random Forest	1040	$11,028.00
Boosted Tree	1344	$11,594.50
Linear SVM (untuned)	2013	$10,459.50
Radial SVM (tuned)	1366	$11,536.00

Table 19: Classification Modeling Results

The logistic and logistic GAM techniques produce both profitable and accurate models, with a moderate number of mailings required. The logistic GAM best subset model is slightly more accurate at 0.8255699 than the logistic best subset model, which has an accuracy of 0.8225966. **Table 20** below shows these two models together for ease of comparison and accuracy calculation.

Model	Validation Class			
		0	1	Number of Mailings
Logistic (best subset)	0	672	11	1329
	1	347	988	
Logistic GAM (best subset)	0	678	11	1329
	1	341	988	

Table 20: Logistic Best Subset v. Logistic GAM Best Subset

While the LDA models are comparable in profit and accuracy, they are a bit more complex in theory and thus difficult to explain. By examining the number of mailings and their estimated profits on the validation data set as decision metrics, we determine that the model with the best performance on the validation data is the logistic GAM model that uses the subset of predictors identified from the backwards variable selection technique. This GAM model results in the same model fit as the original logistic model with backward selection. The logistic GAM best subset model, however, boasts the additional benefit of being highly interpretable and more accurate. Therefore, we will use this model to classify individuals as donors in the test set.

Prediction Models

Model Technique	MPE	Standard Error
Least Squares Regression (Mallows' Cp)	1.558577	0.1606044
Best Subset with k-Fold Cross-Validation	1.555443	0.1611711
Principal Components (20 components)	1.556378	0.1612150
Partial Least Squares (3 components)	1.592151	0.1613484
Ridge Regression	1.572113	0.1627705
Lasso	1.562046	0.1611463

Table 21: Regression Results

All six prediction techniques resulted in strong candidates for predicting donation amounts. As shown in **Table 21** above, least squares regression with Mallows' Cp is one of the strongest in terms of MPE and strongest in terms of Standard Error. Although best subset regression with k-fold cross validation and principal components both have slightly better MPE scores than least squares regression with Mallows' Cp, they are far more complex models with 18 and 20 variables, respectively. Using 14 variables, least squares regression with Mallows' Cp has far fewer predictors than either of the two models with a better MPE. The lasso model at 10 predictors is less complex than least squares regression with Mallows' Cp, but its simplicity is at the expense of accuracy. Overall, because it outperforms its competitors in accuracy while maintaining a relatively simple composition, the least squares regression with Mallows' Cp is the model we will use to predict donation amounts in the test set.

Conclusion

We will compile and deliver the results of this study to the customer and schedule a consultation to determine the customer's priorities. While classification methods that maximize profit exist, there may be tradeoffs between accuracy and profitability. The logistic GAM best subset model is a strong candidate that can satisfy the customer's desire for both accuracy of donor classification and maximized profit, while being a very interpretable and applicable model in a business setting. The results of this study strongly suggest this to be the ideal candidate for classification of donors and maximization of profit. The results of the prediction portion of this report suggest least squares regression with Mallows' Cp as a strong candidate for the prediction of donation amount. This candidate had the best Standard Error and one of the lowest MPE scores, and it is a relatively simple and interpretable model with wide application. This is a strong candidate for the customer to use in predicting donation amounts. We believe the combined logistic GAM best subset and least squares regression with Mallows' Cp models will make a powerful impact on the customer's ability to hone the accuracy of their mailings and maximize profits and donations.