PRED-422 Sec 55 Final Group Project
Justin Neumann, Daren Purnell, Natalie Walker
3 December 2017

Introduction

To assess and improve the cost-effectiveness of their direct marketing campaigns, a charitable organization desires a machine learning model to assist in predicting overall expected gift amounts from donors. A properly designed classification model could predict response rates, minimize excessive mailings, and appropriately target potential donors, thus maximizing the organization's profit from the marketing campaign. Leveraging recent mailing records, this organization compiled (3984) observations of (20) variables that relate to determining whether an individual will donate to their cause and, if so, the dollar amount of said donation.

This paper utilizes the provided *charity* dataset to create a two-stage approach for profit maximization. First, focus on classifying candidates based on whether they respond, or donate. It is desired to exceed the recent mailing records response rate of 10%. Second, create a prediction model that identifies the expected gift amounts from donors, based on the donor-classified observations only. It is desired to meet or exceed the recent mailing records average donation of $14.50. However, the desired end state is to maximize the expected profit from the entire mailing campaign. Each mailing costs $2.00 to send and, using the average donation amount and response rate identified above, the most recent campaign netted the charity organization -$0.55 per mailing. Overall, the resultant models should increase the net profit for each mailing.

Through the Exploratory Data Analysis (EDA), multiple variable transformations were considered and implemented in determining the expected donation amount from donors. Additionally, the EDA considered the relevancy of each variable to donor classification and donation prediction.

Following EDA, the first step of a two-stage approach is explored. Various models, to include gradient boosted tree, linear discriminant analysis, quadratic discriminant analysis, K-nearest neighbors, generalized linear model (logistic regression), classification tree, bagging, random forest, ensemble stacking, and least squares regression were examined and compared to determine the best possible model for binary prediction of whether an observation classifies as a donor or not. The objective of this classification building is to maximize the gross profit that our model achieves on the validation data set. The gross profit is the maximum in a cumulative sum of the ordered donation probability times the average donation amount of $14.50 minus $2.00 of mailing cost.

The second step, a predictive model for the donation amount of classified donors, is explored. Several model types and transformations were leveraged in the process, namely the Ordinary Least Squares (OLS), best subset selection techniques for OLS, ridge regression, lasso regression, principal component regression, partial least squares, gradient boosted trees, generalized linear models, random forests, K-nearest neighbors, several types of discriminant analyses and other tree based methods.

This paper concludes with the selection of the single best model in each step, though alternative candidates that performed slightly worse are also discussed. Those best two models, one for classification and one for regression, predict the donation probability and the donation amount found in the .csv file that accompanies this disquisition. Additional estimations are also made, like the average amount of money spent in the training, validation, and the test data sets.

Exploratory Data Analysis

The *charity* dataset consists of (8009) observations and (24) variables. Of the (24) variables, (4) are not considered predictors. The *ID*, *donr*, *damt*, and *part* variables serve to identify the observation identification number (*ID*), the donor classification (*donr*), donation amount (*damt*), and dataset partition identification (*part*). Each observation receives a (0) or (1) for donor classification (*donr*), with (1) indicating that the observation reflects a respondent who agreed to donate to the charity. For each observation classified as a donor, the donation amount (*damt*) is the dollar integer value of the donation amount.

The partition identification (*part*) assigns each observation to (1) of (3) data subsets: *train*, *test*, or *valid.* The *train* set contains 49.7% of the observations, the *test* set contains 25.1% of the observations, and the *valid* contains 25.2% of the observations. This data set is nearly complete with the only missing values identified in the *donr* variable and *damt* variables that were partitioned into the *test* portion of the data set. To facilitate creating the classification and prediction models, the remainder of the EDA will focus on the *train* portion of the dataset.

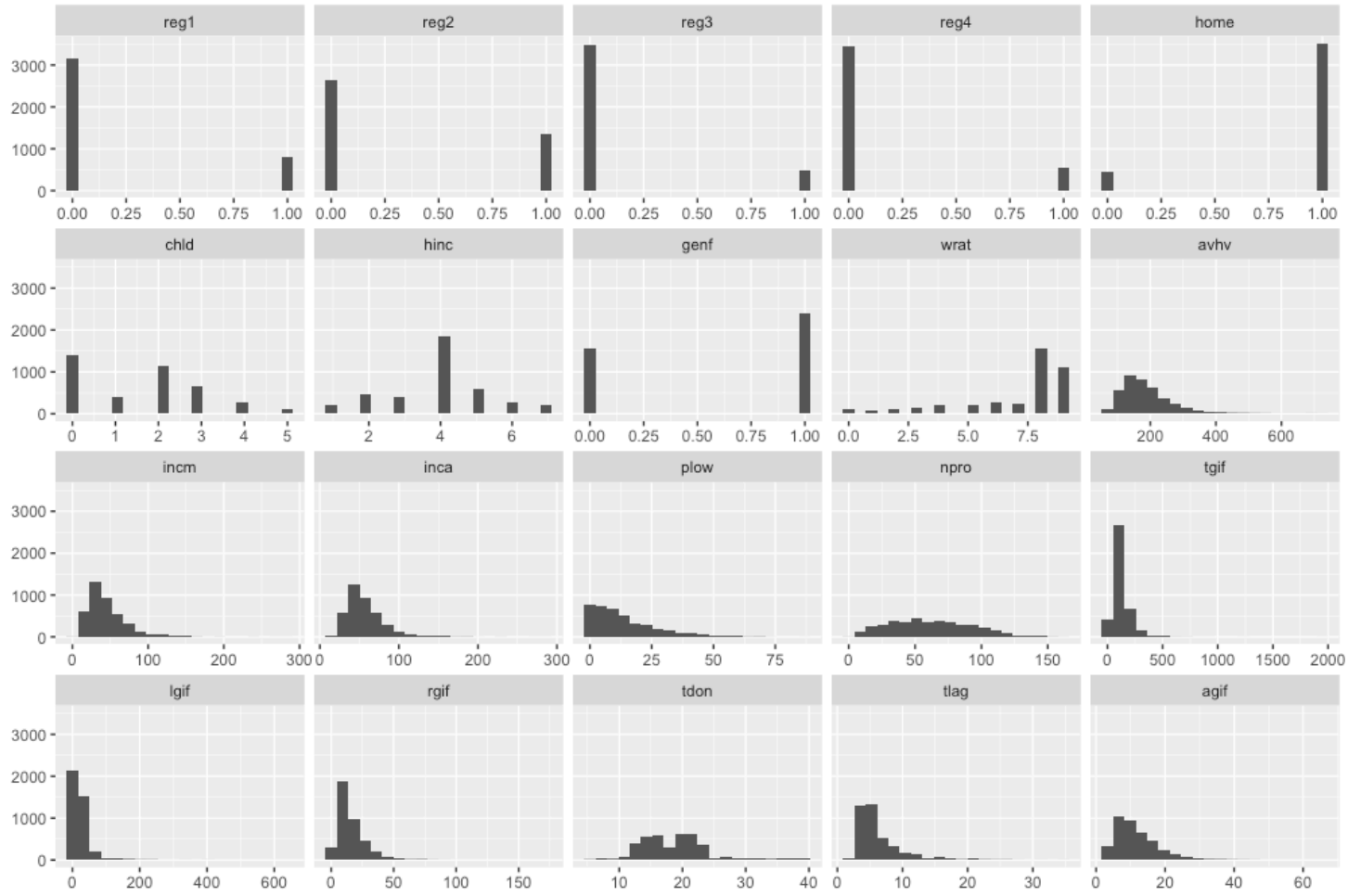
(3984) observations and (24) variables comprise the *train* dataset. The (20) predictor variables included in the data are identified in Figure 1-1 below.

Figure 1-1

Name	Type	Description		Name	Type	Description
reg1	Categorical	Geographic region where a potential donor resides (5 regions)		**wrat**	Categorical	Wealth Rating category (9 bins)
reg2	Categorical			**avhv**	Numerical	Avg home value (thousands)
reg3	Categorical			**incm**	Numerical	Median family income (thousands)
reg4	Categorical			**inca**	Numerical	Avg family income (thousands)
home	Categorical	Homeowner (Y/N?)		**plow**	Numerical	% of low income in neighborhood
chld	Numerical	# of Children		**npro**	Numerical	# promotions received to date
hinc	Categorical	Household income category (7 bins)		**tgif**	Numerical	Amt of lifetime gifts to date ($)
genf	Categorical	Male or Female		**lgif**	Numerical	Amt of largest gift to date ($)
tdon	Numerical	# months since last donation		**rgif**	Numerical	Amt of most recent gift ($)
tlag	Numerical	# months between 1st and 2nd gift		**agif**	Numerical	Avg amount of gifts to date ($)

Histograms. The *ggplot2* package in R was utilized to plot the histograms of each variable, as shown in Figure 1-2 below.

Figure 1-2

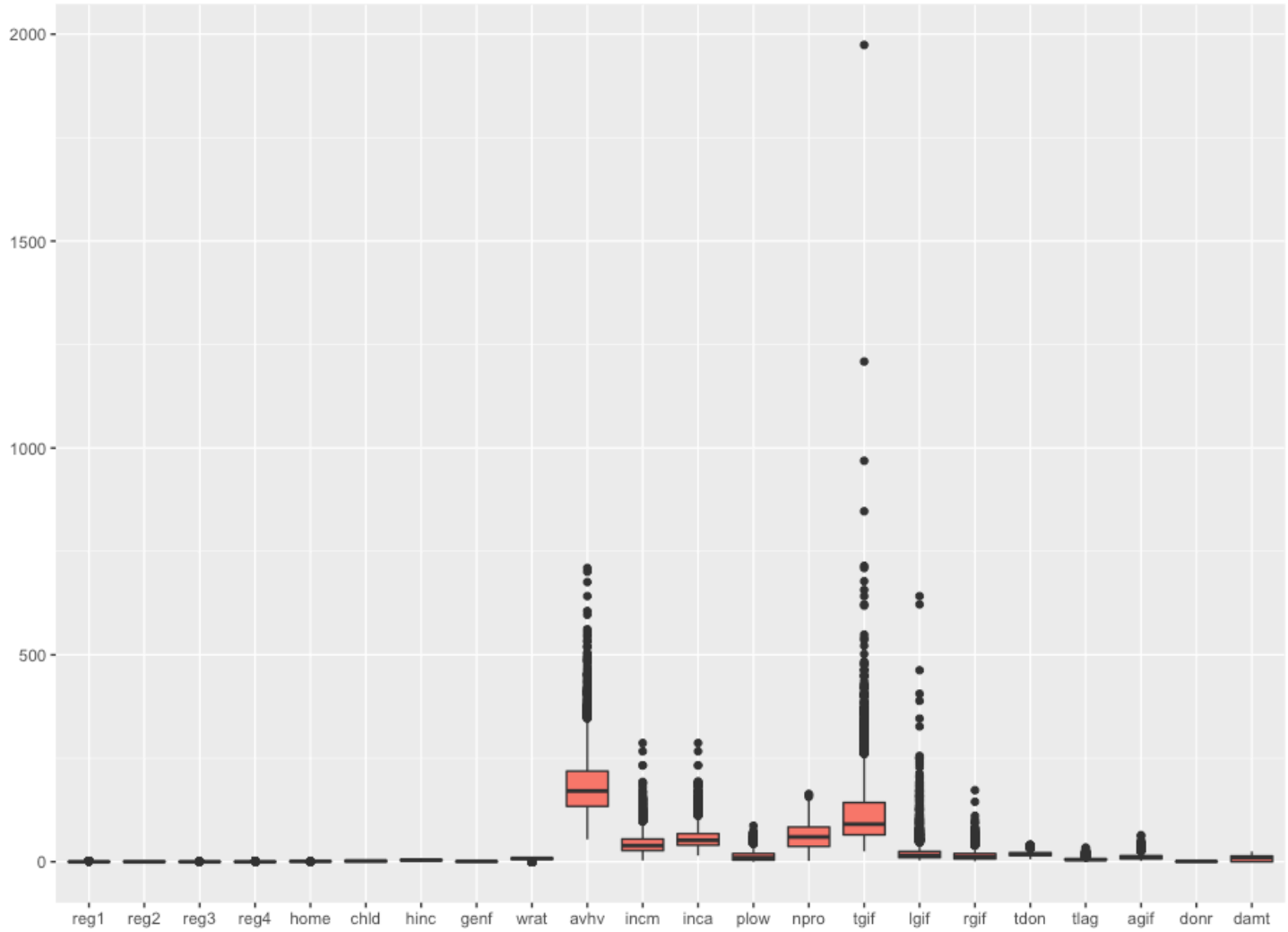


It is noted that many of the variables are categorical. The continuous variables *avhv*, *incm, inca,* and *agif* have clear left skew and potential outliers. *plow* maintains a strong left skew and is not normal. *npro* appears normal, but might be uniform. The variables *tgif, lgif,* and *rgif* have left skew, high kurtosis, and are potentially normal. *tdon* displays two potential clusters with some outliers. And, *tlag* shows strong left skew with potential outliers.

Scatterplots. In addition to the histograms, scatterplots of each predictor variable were created to determine distribution of the observations. Most continuous variables display a uniform appearance with high outliers and nonlinear trends. The only variable displaying a random distribution was *npro*.

Boxplots. Finally, boxplots for each variable were created to identify outliers and skewness and are displayed in Figure 1-3 below.

Figure 1-3



Correlation Coefficients. Further, the Pearson correlation coefficients for each variable related to the *donr* variable were calculated. The coefficients ranged from 0.0095 at the minimum to 0.2493 at the maximum. Moving into the classification modeling, consideration will be given to dismissing *agif* as a predictor based on its low correlation with *donr* (0.0095), however multiple modeling techniques leverage non-linear relationships and interactions where *agif* might provide enhanced performance. For linear modeling techniques, its dismissal will be considered, but all predictors, regardless of correlation coefficient should be included in non-linear modeling processes.

Transformations. The variables with high skew and non-normal distributions were considered for transformations. Logarithmic, square root, sine, and cosine transformations were considered to normalize the continuous variables for better predictive modeling. Figure 1-4 below displays the skew of each variable pre-transformation, the selected transformation, and the resulting skewness of the transformed variable.

Figure 1-4

	skew <dbl>	transformation <fctr>	newskew <dbl>
avhv	1.548030	log.avhv	0.1685455
plow	1.381409	sqrt.plow	0.1824484
npro	0.279600	sqrt.npro	-0.2569347
tgif	5.053572	log.tgif	0.4600263
tdon	1.126973	sqrt.tdon	0.4923123
tlag	2.415135	log.tlag	1.4036872

Principal Components Analysis. Principal Component Analysis (PCA) was conducted on the *train* data set to identify potential trends and groupings within the predictors. However, the PCA returned twenty (20) components mostly centered around zero (0), indicating a lack of redundancy and variance within the predictors.

Key EDA Takeaways. The *train* data set consists multiple variable types and ranges. It contains mostly non-normal predictor variables that contain mediocre linear correlation with the *donr* and *damt*. The categorical variables lack linear correlation as well. Many of the predictors include high outliers that might benefit from trimming, but trimming was not considered due to the ordinal nature of those variables. The biggest takeaway from the EDA process is the lack of redundancy in the predictor variables, indicating that modeling techniques utilized should allow for high dimensionality or incorporate a variable selection metric.

Step 1: Classification Modeling

Over (10) models were considered for the donor classification model, using the *train* data set. The *validation* data set values were compared against the predictions from the classification models and the gross profit value was calculated. Additionally, the Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) was also calculated for each model. All the models examined were compared based on the resultant gross profit due to the financial implications for the organization applying these techniques.

A comprehensive table of all models examined, their associated gross profits, and respective AUC values is located at the end of this section for review, however only the top (3) performing models are discussed in detail in this section of the paper. These models (in order of gross profit) include a heteroscedastic discriminant analysis model, a stacking ensemble model, and a quadratic discriminant analysis model.

Heteroscedastic Discriminant Analysis. A heteroscedastic discriminant analysis model utilizing all twenty (20) predictor variables was calculated using the R package *caret* and examined to determine the gross profit that is achievable with such a model. The Heteroscedastic Discriminant Analysis (HDA) model generalizes linear discriminant analysis by allowing unequal sample covariances, or heteroscedastic data. A rigorous 10-fold cross-validation approach in conjunction with the self-defined 'gross profit' metric function was followed and subsequently combined with an automatic tuning process for the three parameters. Parameter 'lambda' for regularization towards equal covariance was optimized to 1 (complete heteroscedasticity), parameter 'gamma' is a shrinkage parameter towards diagonal covariance matrices and was optimized to 1 (no shrinking). Additionally, the parameter 'newdim' determines the dimension of the discriminative subspace and was optimized to (3). The class distributions are assumed to be equal in the remaining dimensions. The calculated validation set gross profit was 15494.83 making this model the best performing examined in this analysis, by far

Stacking Ensemble. A stacking ensemble was utilized consisting of three base models and a (top-level) stacker. The base models included the HDA, the Flexible Discriminant Analysis (FDA) and the Quadratic Discriminant Analysis (QDA). The stacker was also an HDA model. This set up achieved a validation set gross profit of 13758.37. The individual scores of the base models were calculated by the validation set gross profit and are 15494.83 for the HDA, 12780.32 for the QDA, and 12447.71 for the FDA. Note that again, a rigorous 10-fold cross-validation approach in conjunction with the self-defined 'gross profit' metric function was followed and combined with an automatic tuning process for base model and stacker parameters. Because the ensemble model score failed to exceed the scores of its individual models, it is assumed that the stacker was unable to rationalize the different predictions with the given labels. The R software package *caretEnsemble* integrated with the R learning package *caret* to create an ensemble model in a selective process that started with over (12) different base models trying various stackers. The QDA used was without tuning parameters, the FDA used the optimal parameters 'degree' equal to (1) and 'nprune' equal to (16).

Quadratic Discriminant Analysis. The Quadratic Discriminant Analysis (QDA) was the third best performing classification model examined. All (20) predictor variables were utilized to calculate a final model. Utilizing the QDA method allowed us to assume a quadratic decision boundary between the donors and non-donors with the data that exceeded the performance of the linear modeling methods. This QDA model achieved a validation set gross profit of 13017.3.

Overall Classification Model Evaluation. Figure 2-1 below identifies all the considered models, their associated maximum gross profit, and AUC. This figure shows the models in descending order of gross profit calculation.

Figure 2-1

Model	Max Gross Profit	AUC
Heteroscedastic Discriminant Analysis	15494.83	0.7473
Stacked Ensemble	13758.37	0.9647
Quadratic Discriminant Analysis	13017.3	0.9101
Linear Discriminant Analysis	12631.39	0.9414
Flexible Discriminant Analysis	12447.71	0.965
Gradient Boosted Tree	12162.07	0.9478
Logistic Regression	11812.05	0.9421
Classification Tree	11790.19	0.9153
Bagged Tree	11774.45	0.9529
Random Forest	11399.13	0.8964
K-Nearest Neighbors	11297.97	0.5635

The average probability of achieving a positive classification was calculated for each relevant classification model and plotted against the maximized gross profit metric, as shown in Figure 2-2. This provides insights into the very existence of the relationship between the positive class average probability and the maximum gross profit on the validation set. The discriminant analysis model successes in this paper allude to the weak separation between the donor and non-donor class in the data and the presence of non-linear relationships between the response variable and the regressors. It also assures that the HDA model, though a leverage point in terms of this relationship, is clearly not something exceptional, but in line with the other models evaluated. Figure 2-3 immediately adjacent to Figure 2-2 displays a table of the average probabilities for the highest performing models.

Figure 2-2

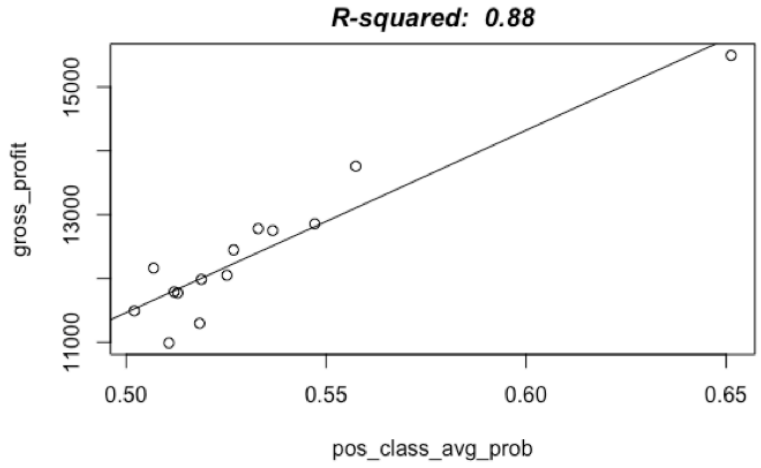


Figure 2-3

Model	Avg Probability
Quadratic Discriminant Analysis	0.5330
Flexible Discriminant Analysis	0.5269
Heteroscedastic Discriminant Analysis	0.6513
Stacked Ensemble	0.5574

Step 2: Regression Modeling

A large variety of regression models were tested on different set ups of data. In addition to using the data in its plain form, many modifications were explored that incorporated normalized features, added prediction probabilities from AUC-focused or gross profit focused models, predictions over all potential donors, and more. The specific types of models investigated included gradient boosted trees, generalized linear models, random forests, K-nearest neighbors, lasso regression, ridge regression, transfer lasso, Bayesian lasso, and more. Additionally, best subset selection was implemented using Residual Sum of Squares (RSS), Adjusted R^2, Bayesian Information Criterion (BIC), and Mallow's Cp to evaluate the inclusion of individual regressors in the predictive model. Mean Squared Error (MSE), which is equivalent to Mean Prediction Error (MPE), was used as the score to evaluate these models,

A **comprehensive table of all models examined** and their associated MSEs is located at the end of this section for review, however only the top (3) performing models are discussed in detail in this section of the paper. These models (in order of gross profit) include a extreme gradient boosting , a stacking ensemble model, and a Bayesian Lasso model.

Extreme Gradient Boosting (XGBoost). An extreme gradient boosting model utilizing all (20) predictor variables was used and examined to determine the MSE that is achievable with such a model. Regarding statistical properties, such as statistical metrics, this model tended to be very robust. However, it also provided the possibility of working with categorical features that are numeric in nature. The tuning parameters determined from a 10-fold cross-validation on the training share of the data were 'nrounds' of 100, 'max_depth' of 1, 'eta' of 0.3, 'gamma' of 0, 'colsample_bytree' of 0.8, 'min_child_weight' of 1 and 'subsample' of 1. This model achieved a top MSE of 1.41 on the validation set and is by far more superior than any other model presented in this section.

Stacking Ensemble. A stacking ensemble was utilized consisting of three base models and a (top-level) stacker. The base included the extreme gradient boosting (XGB) calculated using the *caret* package in R, the Bayesian Lasso and the ridge regression. The stacker was a Generalized Linear Model (GLM). This set up achieved a validation set MSE of 1.43. The individual scores of the base models were calculated by the validation set gross profit and are 1.41 for the XGB, 1.846 for the Bayesian Lasso and 1.849 for the ridge regression. Because the ensemble model score failed to exceed the scores of its individual models, it is assumed that the stacker was unable to rationalize the different predictions with the given labels. The software package *caretEnsemble* was integrated with the machine learning package *caret* in the R language to

develop an ensemble model with a selective process that started with over (9) different base models and tried various stackers. The tuning parameters for the XGB are stated in the model above, the tuning parameter for the Bayesian Lasso was 'sparsity' of 0.3. The tuning parameter for the ridge regression was 'lambda' equals 0.1.

Bayesian Lasso. The Bayesian lasso model was the third best achieving model examined, with an MSE of 1.846. The original input features of this data were combined with the gross profit oriented classification probabilities for the donation for model calculation. No normalization was applied because this resulted in an increase of MSE. The tuning parameter as part of the caret package interface was 'sparsity' equals 0.3.

Overall Regression Model Evaluation. Figure 2-4 below identifies all the considered models and their associated MSE. This figure shows the models in ascending order of MSE.

Figure 2-4

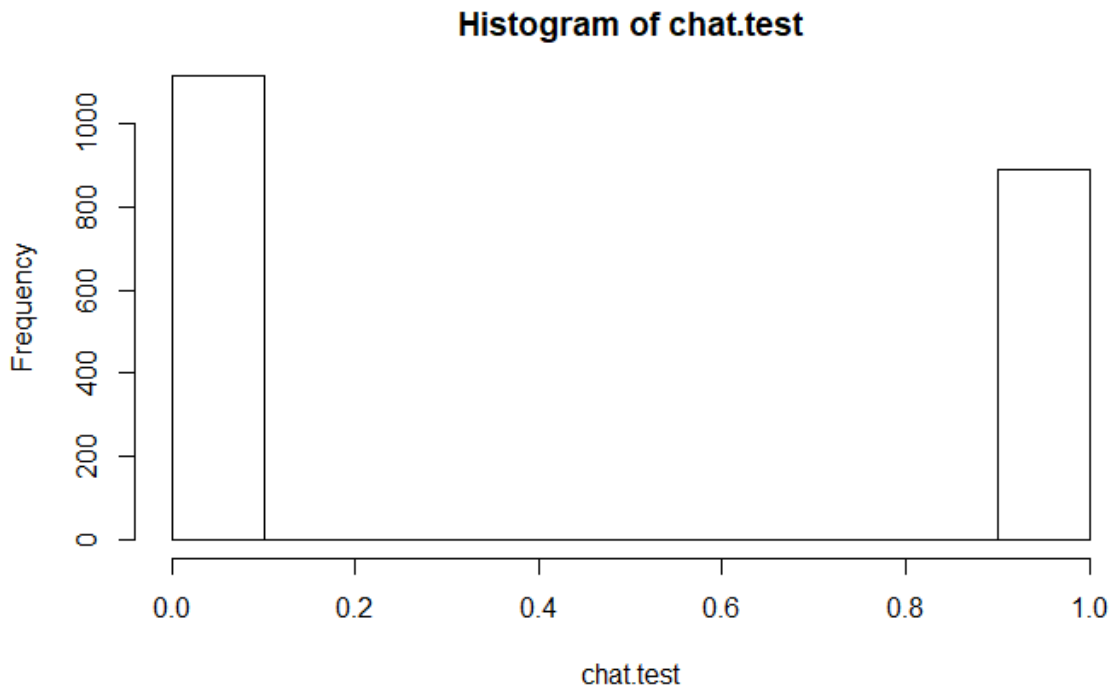
Model	MSE
Extreme Gradient Boosting	1.4128
Stacked Ensemble	1.4392
Bayesian Lasso	1.8463
Ridge Regression	1.85
Ordinary Least Squares Regression	1.8675
Generalized Linear Model	1.8827
K-Nearest Neighbors	2.1961
Transfer Lasso	28.1114
Lasso	28.1296
Partial Least Squares	28.1526
Principal Components Regression	28.1828
Best Subset Selection OLS	28.229

Summary of Best Models and Performance Metrics

As identified by the results above, the best classification model for the *charity* data set is the **Heteroscedastic Discriminant Analysis** (HDA) and the best predictive model is the **Extreme Gradient Boosting** (XGB). Using these models, identical learning processes were applied to the whole training and validation data, in order to predict the test data. Following, information regarding the number of mailers sent and the average predicted donation amount is provided to further explain the results. Additionally, limitations of the current results, to include mail "send-out" decisions with predicted donation amounts of $2.00 or less are identified.

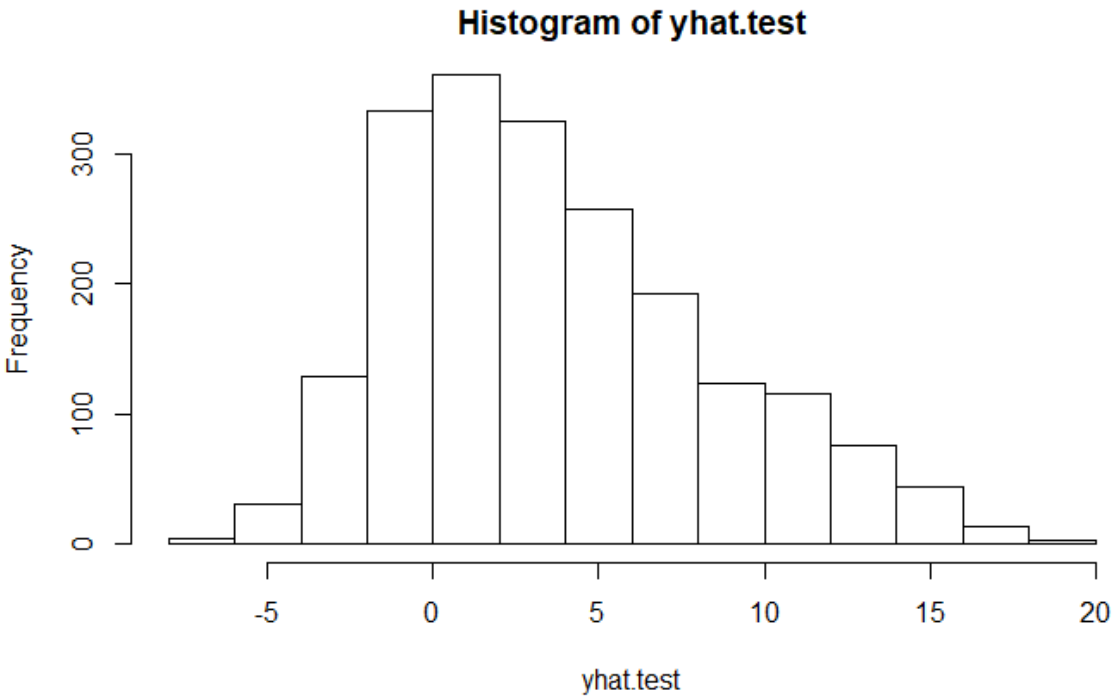
Using the best donation probability predictions, a test-adjusted send out threshold (in number of emails) was calculated based on the validation set number of emails. This resulted in (890) emails sent to specific donors and (1117) emails not sent. Figure 3-1 below displays this threshold.

Figure 3-1



The best donation amount predictions yielded an average of $3.66 per **potential** donor. Because the predictions were taken directly from the best model without modification, donation amounts were allocated to no-go donors. Additionally, go decision donors were identified in disregard of whether they contribute $2.00 or less in profit (which would be a loss business for each). Figure 3-2 provides a graphical depiction of the donation amount distribution.

Figure 3-2



The histogram identifies that a significant amount of predictions result in a loss of money, if identified as a go-decision donor. Specifically, given our final solution comma-separated-file, (182) actual donors bring no profit with an average donation amount of $0.00386, resulting in a loss of $363.30.

Conclusion

Through a two-stage approach, this paper sought the best classification and predictive models to predict both identification of an observation as a donor and also the identified donor's resulting donation amount, in an attempt to reduce excessive mailings and maximize gross profit of the charitable organization. (3984) observations in the training *charity* data set were leveraged to calculate the Heteroscedastic Discriminant Analysis (HDA) model that produced $15,494.83 in maximum gross profit on the validation data set. This same data set was utilized to create the Extreme Gradient Boosting (XGB) predictive model that produced an average $3.66 per potential donor on the validation data set.

While multiple iterations of ensemble models were considered for the classification model, the best solution (the HDA model) arose from a singular modeling technique that lacked iterative resampling or stacking (though 10-fold cross-validation was applied). Additionally, the HDA model identified for classification suggests that, as noted in the EDA, multiple nonlinear relationships and interactions exist within the donor classification relationship. Similarly, multiple iterations of stacking ensembles were considered for the regression modeling portion of this analysis. However, the best solution again resulted from a singular advanced modeling technique that leveraged tree-based, iterative resampling and creation of a single model from that sequential series.

When these models were combined with the implementation of the go decision rule (derived from our validation set), overall profit for the charitable organization was determined to be $11,125 (when applying the same gross profit logic as used with the validation set). Additional consideration should be given to different transformations or techniques that might improve the expected $363.30 loss on a specific share of the go-decision donors, those that are predicted to donate $2.00 or less. However, the designed models effectively predict response rates, minimize excessive mailings, and appropriately target potential donors, resulting in a maximization of the organization's marketing campaign profit.