NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Mailing Campaign Response Modeling

Yining Feng

March 7, 2021

1. Introduction & Problem Statement

XYZ Company is planning on developing a model-based method for targeting customers who are likely to respond to a new mailing campaign. XYZ's database of customers which contains a large number of variables relating to sales and campaign results would be leveraged to first assess which variables have the greatest 'importance' in determining both the chance of response to a mailing campaign as well as the likely spend of each customer. The reduced dataset will then be used to fit three classification models to predict chance of response, and three regression models to predict likely spend. Optimal models are selected and subsequently used in order to estimate the expected net revenue from conducting a new targeted mailing campaign, accounting for the cost of mailing each customer.

2. Data Exploration and Pre-processing

2.1 Data Description

The XYZ's customer database includes 30,779 records and 554 features of customer data. The features capture customer sales, results from previous mailing campaigns, and Experian properties which provide additional insights about each customer. The features are broken up into 345 character, 48 integer and 161 numeric variable types.

Due to the scale of data, a subjective assessment of variable relevance is conducted prior to employing any pre-processing or modelling routines. The subjective assessment carried out by assessing the descriptions for variables contained within the provided data dictionary and grading each by its perceived ability to predict both the chance of response and likely spend. This resulted in 227 variables being graded as having a 'low' relevance, which are subsequently excluded from the dataset. Note that a summary data file of each variable grade is available on request. The remaining variables can be broken up into 119 character, 48 integer and 160 numeric variable types.

From an initial look at the data, it is noteworthy that the compiled R data frame fails to distinguish between numeric and factor variables. As such, prior to performing any data exploration, all character class variables are converted to factor type and retained all other variables as numeric type. All ‘ANY_MAIL_x’ and ‘RESPONSEx’ variables are then manually converted to factor type, since these variables contain no character-based observations yet are observed to be categorical in nature.

2.2 Exploratory Data Analysis

Univariate Data Analysis

As part of the univariate data analysis, summary statistics for all of the 160 retained numeric variables are calculated and observed. The majority of numeric variables do not suffer from missing values. However, many variables have a minimum value of zero, suggesting zero-inflated data. Histogram and box plots are also generated and reviewed for a large subset of numeric variables, with total amount spent (TOTAMT) selected for further discussion (Figure 1). Note that zero value observations are removed prior to generating each plot.

It is apparent that the variable suffers from a heavy positive skew, which is in-fact a common attribute over the majority of numeric variables. The result is a number of observations which could be classified as outliers.

Bivariate Data Analysis

Since XYZ company intends on building a prediction model to determine both the chance of response to a mailing campaign (RESPONSE16) and the likely spend of each customer (TOTAMT16), there is an interest in identifying variables which have explanatory power over these two variables. As such, the Pearson correlation coefficient will be calculated and reviewed for all numeric variables against the numeric response variable TOTAMT16. Correlations for the 10 most correlated numeric variables against TOTAMT16 are shown in the table below.

Variable	Pearson correlation coefficient
YTD_SALES_2009	0.3723
YTD_TRANSACTIONS_2009	0.2938
LTD_SALES	0.2218
LTD_TRANSACTIONS	0.2067
PRE2009_TRANSACTIONS	0.1587
PRE2009_SALES	0.1445
TOTAL_MAIL_13	0.1182
TOTAL_MAIL_14	0.1178
TOTAL_MAIL_15	0.1178
SUM_MAIL_12	0.1158

Table 1. *Correlations vs. TOTAMT16 (Top 10 Correlations).*

None of the variables have reported a strong correlation with TOTAMT16, with the greatest absolute correlation being reported by total sales through to 2009 (YTD_SALES_2009) and total transactions through to 2009 (YTD_TRANSACTIONS_2009) at 0.37 and 0.29, respectively. Finally, bar plots are used to explore the relationship between the categorical response variable (RESPONSE16) and each numeric variable. Two of these plots have been selected for further discussion (Figure 2).

It is evident that there are indeed recognizable differences in both the mean and distribution of a number of numeric variables depending on whether they are associated with a positive or negative response to an advertising mailing campaign.

2.3 Data Pre-processing

As part of the data pre-processing routine, imputing data for missing observations (~11% of the dataset) is the first important step. This is initially attempted using the *rfImpute* function from the Random Forest package in R, however processing time eliminated this as a viable option. Instead, each variable is looped over and observations for numeric variables are imputed with the variables' median value, and at the same time, observations for character variables are imputed with the variables' most common value. Those variables which require imputation are copied and renamed to include the suffix 'IMP' while the original (non-imputed) equivalent is removed from the dataset. Note that while 89 of the retained categorical variables are identified as having missing values, only one of the retained numeric variables, ECHVPCT, is identified as having missing values.

The next step focuses on dealing with outlier observations for numeric variables. In many cases, the task of identifying outlier observations can be a subjective practice. As such, for this assessment, a statistical approach is taken to target those observations which fall outside the 1st and 99th percentile range. Observations which meet these criteria are replaced using the *squish* function as part of the scales package in R, effectively resulting in a newly created set of trimmed variables. Trimmed variables added to the dataset and can be recognized by the suffix 'T99'.

The final step is to create dummies for each of the retained factor variables. All factor variables need to be converted to dummies, however many of those variables have a high-level count and a number of levels with relatively low occurrence. As such, dummies are selectively created for only those factors which have 10 or less levels. Dummy variables are named to include the prefix 'DUM_', along with a suffix to represent the factor level. Note that k-1 dummies are created, where k is the original number of levels for each variable. This resulted in the creation of 351 dummies.

2.4 Variable Importance

The data processing routine produces a data frame of 586 numeric variables. With such a large dataset, it is clear that any subsequent model estimation would benefit from a further reduction in variable count. To achieve this, the *varImp* function is leveraged as part of the caret package in R to calculate the variable importance according to both response variables. In both cases, variable importance is calculated by fitting a Random Forest model, with 'importance' measured by the mean decrease in node impurity. Bar plots of the 20 most important variables for both response variables are shown in Figures 3 & 4.

Some commonalities between variable importance plots are noticeable, with TOTAMT, YTD_SALES_2009 and PRE2009_SALES all within the top-10 rank for both response variables. Figures 3 & 4 also reveal a fairly quick drop-off in variable importance beyond the first five variables. Based on these results, top-50 ranked variables by importance are selected and passed through to model estimation phase. It is hypothesized that this reduction would provide a suitable trade-off both in terms of accuracy and performance.

3. Model Estimation

3.1 Classification Modelling: Chance of Response

For this assessment, three classification-based models are employed to predict the chance a customer will respond to a mailing campaign. These models include a Naive Bayes, Random Forest and a Lasso and Elastic-Net Regularized Generalized Linear Model (GLMnet) classifier. In each case, the training function is leveraged as part of the caret package with a 3-fold cross-validation sampling method, which is applied to a 70% subset of training data and tested against a

30% subset. Default parameters are used for each model. The in and out-of-sample Receiver Operating Characteristic (ROC) Curves are shown for each model in Appendix A. The Naive Bayes classifier managed to deliver an in-sample Area Under the Curve (AUC) of 0.85, while its out-of-sample AUC was 0.82. Clearly, the classifier has avoided over-fitting the training data and has managed to maintain similar classification performance over both the training and test sets. The Random Forest classifier on the other hand, reports an in-sample AUC of 1.00 and an out-of-sample AUC of 0.85, which suggests that it greatly suffers from over-fitting. Finally, the GLMnet classifier generates an in-sample AUC of 0.87 and out-of-sample AUC of 0.86. The out-of-sample confusion matrix for each model is shown below.

Naïve Bayes			Random Forest			GLMnet		
	Pred: 0	Pred: 1		Pred: 0	Pred: 1		Pred: 0	Pred: 1
Actual: 0	3503	246	Actual: 0	3992	452	Actual: 0	4004	416
Actual: 1	496	218	Actual: 1	7	12	Actual: 1	20	38

Table 2. *Confusion Matrix: Classification Model Comparison*

It is observed that while the Random Forest and GLMnet classifiers are able to produce similar out-of-sample performance according to their ROC curves, the Random Forest classifier has done so by providing less true positive and true negative values. This is confirmed by observing the performance metrics in the table below, which shows that the GLMnet classifier is able to obtain a superior true positive rate and true negative rate.

	Naive Bayes	Random Forest	GLMnet
Accuracy	0.8337	0.8972	0.9026
95% CI	(0.8225, 0.8446)	(0.8879, 0.9059)	(0.8936, 0.9112)
Kappa	0.2793	0.0419	0.1284
Sensitivity	0.4698	0.0259	0.0837
Specificity	0.8760	0.9983	0.9950
Pos Pred Value	0.3053	0.6316	0.6552
Neg Pred Value	0.9344	0.8983	0.9059
Prevalence	0.1040	0.1040	0.1014
Detection Rate	0.0489	0.0027	0.0085
Detection Prevalence	0.1600	0.0043	0.0130
Balanced Accuracy	0.6729	0.5121	0.5394

Table 3. *Performance Metrics: Classification Model Comparison*

Based on the view of the performance metrics above, the GLMnet model has demonstrated a superior AUC, accuracy, sensitivity and specificity compared to the other models. As such, the GLMnet classifier would be employed to predict the chance of response.

3.2 Regression Modelling: Amount Spent

Three regression-based models are deployed to predict the amount a customer will spend. These models include a Multiple Linear Regression (MLR), Random Forest and extreme Gradient Boost linear regression estimator. Note that for the MLR, a stepwise variable selection technique is used based on the Akaike Information Criterion (AIC). As with the previous classification models, a 3-fold cross-validation sampling method is employed, and the same 30/70 split is maintained between test and training data subsets.

The in and out-of-sample actuals versus predictions for each model are shown in Appendix A. Each model seems to struggle with both outlier observations and the zero-inflated predictor data. The model assessment can be extended by observing the model fit statistics for each in the below table. Note that negative response values are taken as zero for the statistics shown in the table below. That is, negative spend amounts are interpreted to be zero.

	MLR	Random Forest	Grad Boost
Training set			
MAE	52.74	28.42	18.96
MSE	27243.39	9856.1	2005.58
RMSE	165.06	99.28	44.78
R^2	0.18	0.7034	0.9396
$Adj\ R^2$	0.1782	0.7019	0.9393
Test set			
RMSE	177.96	179.68	189.49
R^2	0.1366	0.1199	0.0211
$Adj\ R^2$	0.1321	0.1098	0.0099

Table 4. *Performance Metrics: Regression Model Comparison*

The performance metrics above indicates that each regression model has performed quite poorly over the test set of data. It is hoped however, that combining the predictions with chance of response will aid in dealing with the zero-inflated response data. For this assessment, the MLR regression model would be adopted to predict the amount spent as its training performance metrics are among the most favorable.

4. Customer Scoring

For the final part of this assessment, a customer score is created based on the combined predictions of the Random Forest classification and MLR model discussed above. This function is to represent the expected value from conducting a new advertising campaign, based on predictions against a subset of customers who have not yet been mailed. The customer score function is shown below.

$$Customer\ Score\ =\ P(response)\cdot\ E(net\ revenue)-\ Cost\ of\ Mail$$

For the above function, '*P(response)*' represents the probability of response as predicted by the chosen classification model, Random Forest. '*E(net revenue)*' represents expected net revenue, which is assumed to be 10% of the amount spent as predicted by the chosen regression model, MLR. And finally, the '*cost of mail*' represents the cost of mailing customers as part of a new advertising campaign which is assumed to be equal to $3.00 per customer. The sum of customer scores represents the expected value from conducting a new advertising campaign.

The customer score above is used to propose four possible marketing strategies. The first strategy, ALL_MAIL involves mailing all customers who have not yet been mailed, regardless of the probability of response or expected net revenue. This would obviously be a costly strategy, considering the cost of mailing all customers. The second strategy, ALLSCORE_MAIL involves mailing only those customers who return a positive customer score according to the above function. For the third strategy, HIGHPROB_MAIL, only those customers who are predicted to have a probability of response greater than or equal to 0.7 are mailed. Note that this strategy ignores the customer score and may capture customers who have a negative expected return when accounting for their predicted spend amount. Finally, the fourth strategy, HIGHVAL_MAIL, involves mailing only those customers who have a predicted spend amount of greater than or equal to $600 ($60 net revenue). This strategy also ignores the customer score and may capture customers who have a negative expected return when accounting for their probability of response.

Table 5. *Customer Score Summary*

strategy	criteria	no. mailed	expected value	value per customer
ALL_MAIL	ANY_MAIL_16 = 0	15,857	-$15,381.29	-$0.97
ALLSCORE_MAIL	Customer Score >= 0	1,111	$4,866.18	$4.38
HIGHPROB_MAIL	P(response) >= 0.7	19	$506.92	$26.68
HIGHVAL_MAIL	E(net revenue) >= 60	20	$830.40	$41.52

It should be no surprise that the greatest expected value comes from the strategy which involves targeting all customers with a positive customer score. However, there are also viable strategies from mailing only those customers with a high probability of response or a high predicted spend amount. These two strategies are able to achieve a much greater expected value per customer. It may be that the most effective marketing strategy would be to target those customers as flagged by HIGHPROB_MAIL and HIGHVAL_MAIL in the first instance. And then, depending on the success of that campaign, proceed to target the remaining customers flagged by ALLSCORE_MAIL. Note that a list of un-mailed customer account numbers according to the above strategies is available on request.

There are some interesting observations when reviewing the customers which are flagged by ALLSCORE_MAIL. The majority of these customers are flagged as 'active' customers, who are homeowners, with incomes between $50-$150k, and are educated with a bachelor's degree or higher. There are also some other interesting observations when comparing the predicted customer response against those who have previously been mailed. Firstly, of the 14,922 customers who were previously mailed, 1,440 customers did in-fact respond (~10% response rate). The chosen classifier, however, suggests that only 71 of the 15,857 un-mailed customers have a probability of response greater than 0.5 (~0.4% response rate). In addition, of the customers who were previously mailed, the average spend amount of those customers was $342. The chosen regression model, however, suggests that the average spend amount for those customers with a probability of response greater than or equal to 0.5, is only $204. It may be that both the chosen classification and regression models are quite conservative in their predictions, or that those customers who were already mailed carried a higher probability of response and higher predicted spend.

5. Conclusion

For this assessment, three classification models and three regression models are adopted to predict chance of response and to predict likely spend, respectively. From the fitted models, a GLMnet based model is found to be superior in predicting chance of response, and the MLR model to be superior in predicting spend. Optimal models are selected and subsequently used in order to estimate the expected net revenue from conducting a new targeted mailing campaign, accounting for the cost of mailing each customer. This score is then used to propose four possible marketing strategies, ranging from mailing all customers who have not yet been mailed to mailing only those customers who have a predicted spend amount greater than or equal to $600. Results suggest a viable strategy to mail customers with a high probability of response and/or high expected spend in the first instance, and to follow this by mailing the remaining customers with a positive customer score.

Appendix

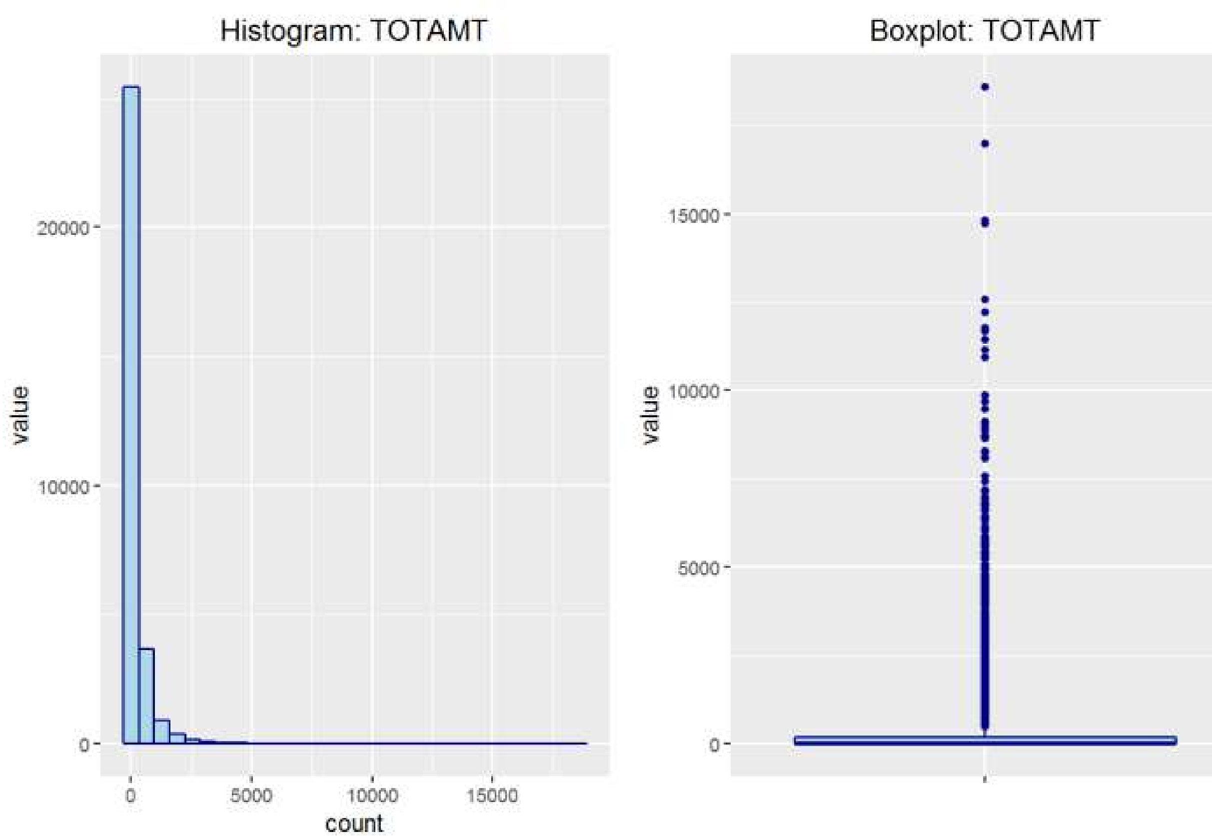

Figure 1. *Histogram and Boxplot: TOTAMT.*

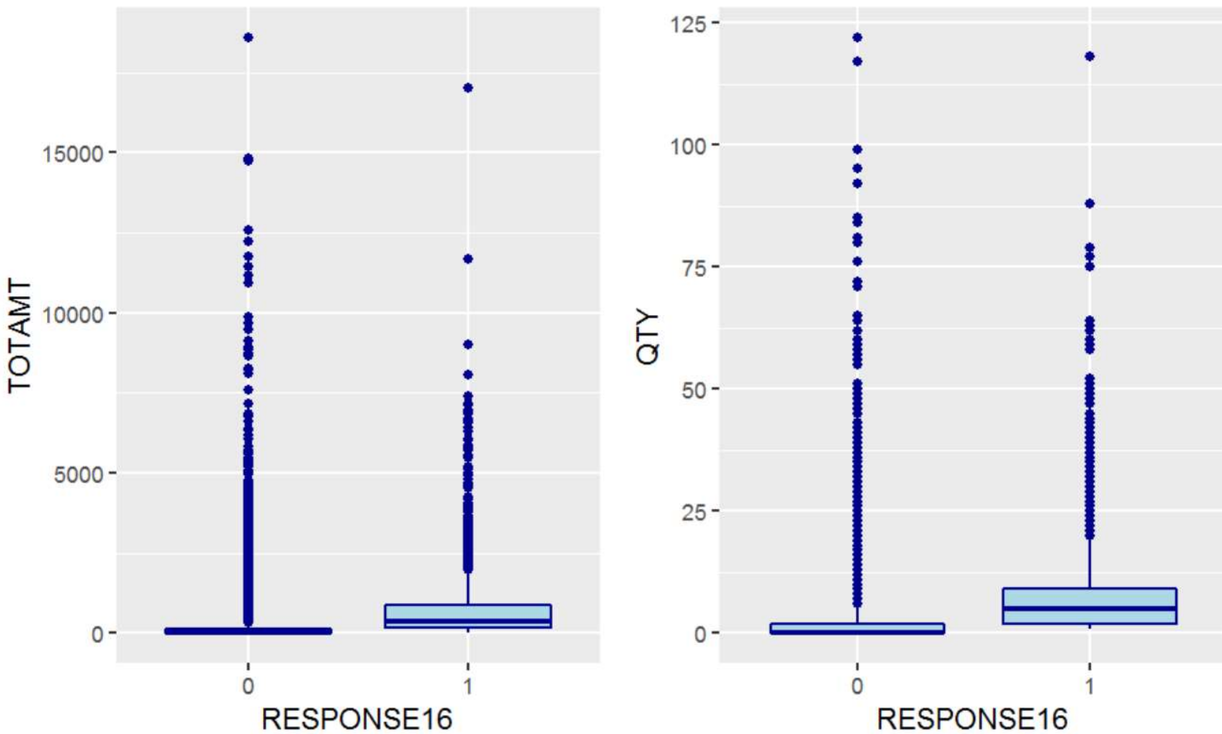

Figure 2. *Boxplot: RESPONSE16 vs. TOTAMT / QTY.*

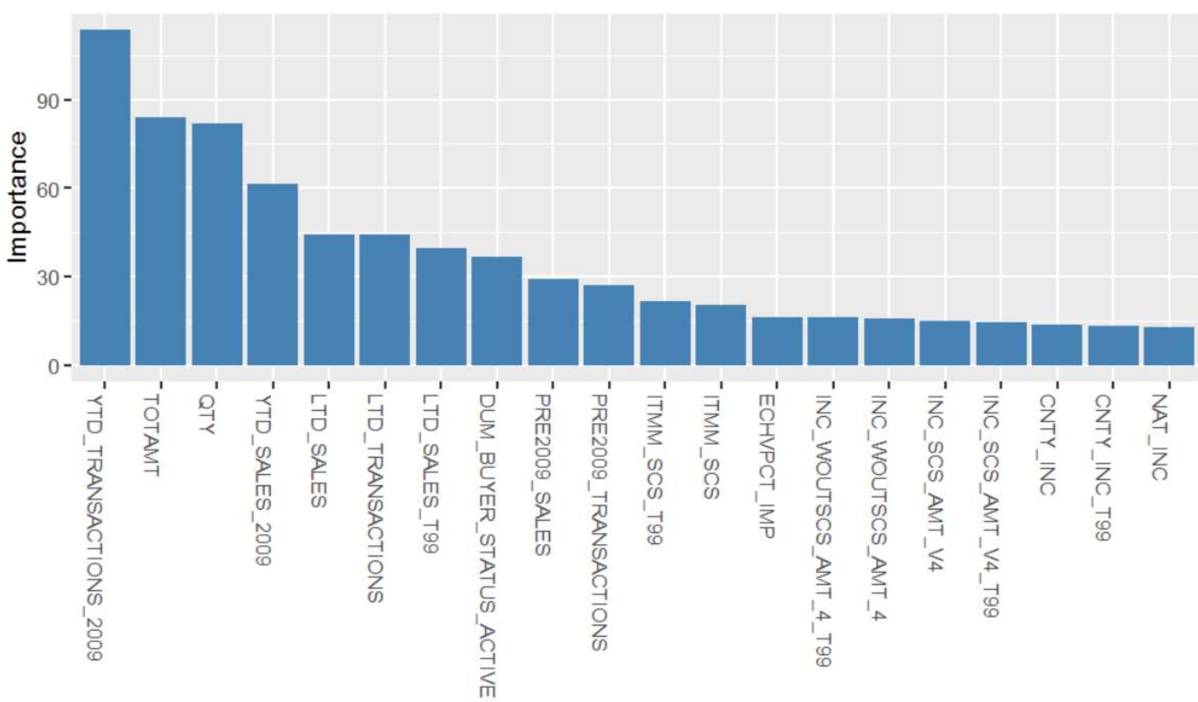

Figure 3. *Variable Importance: RESPONSE16.*

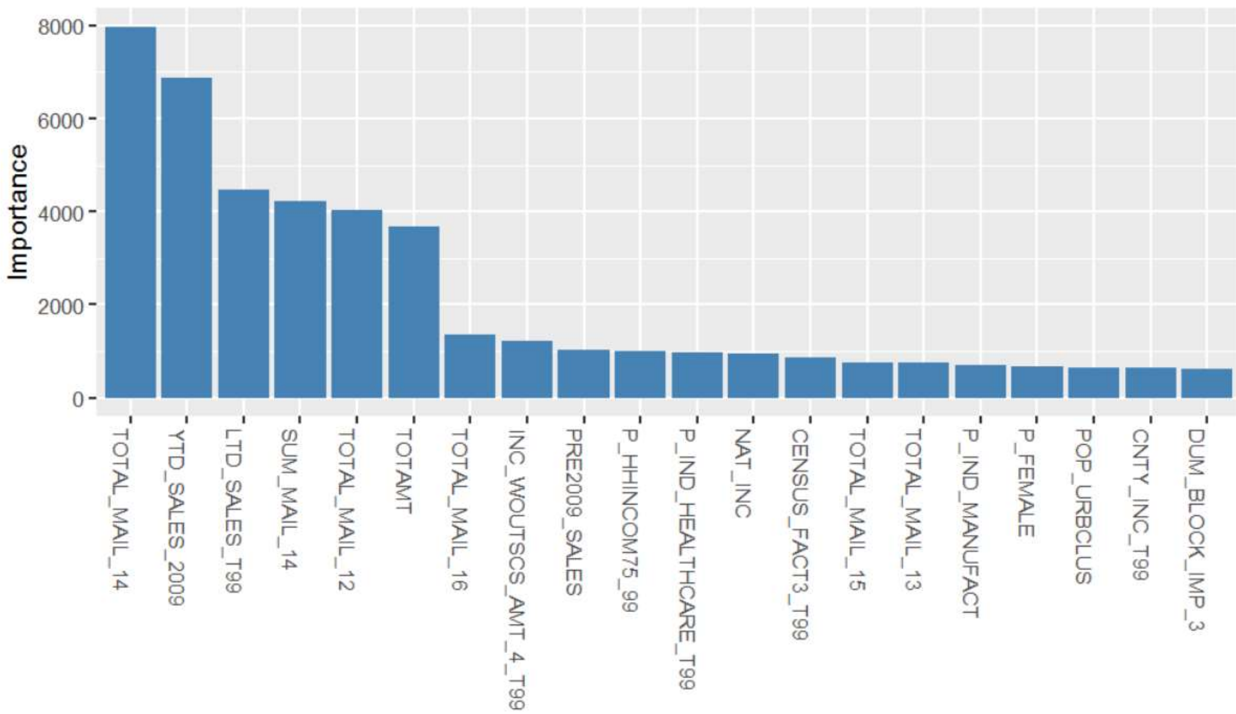

Figure 4. *Variable Importance: TOTAMT16.*

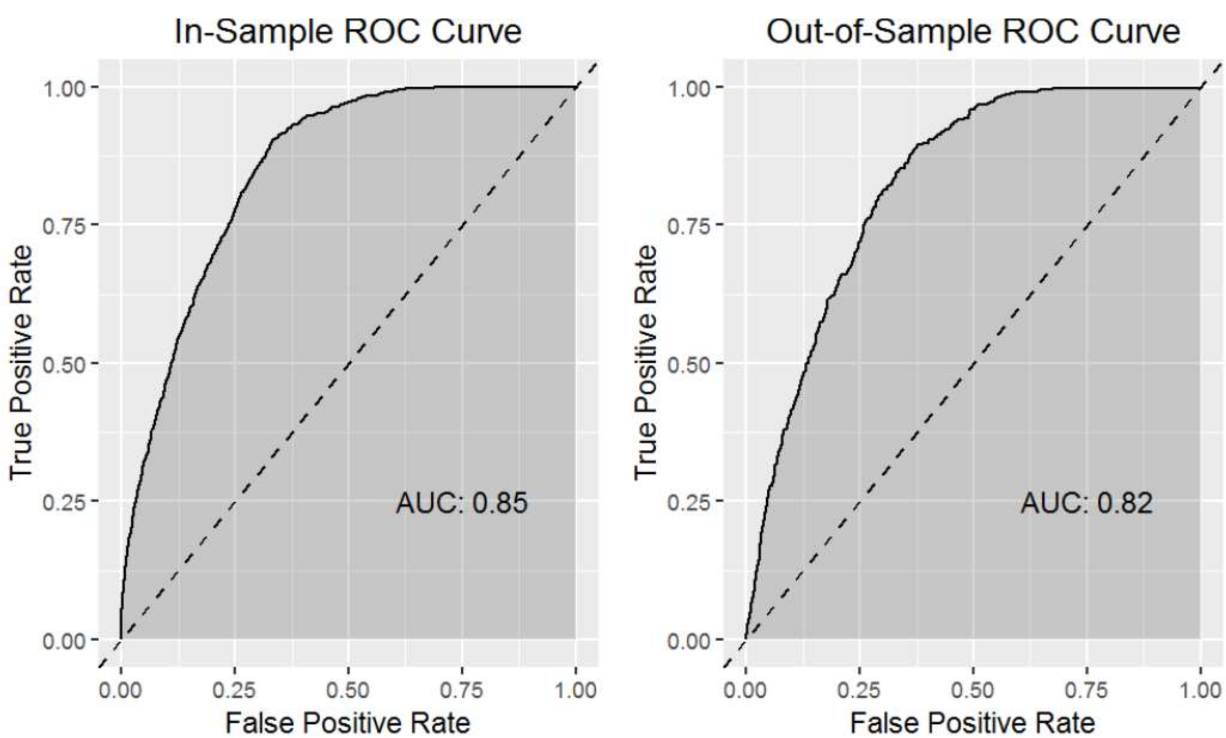

Figure 5. *ROC Curve: Naïve Bayes*

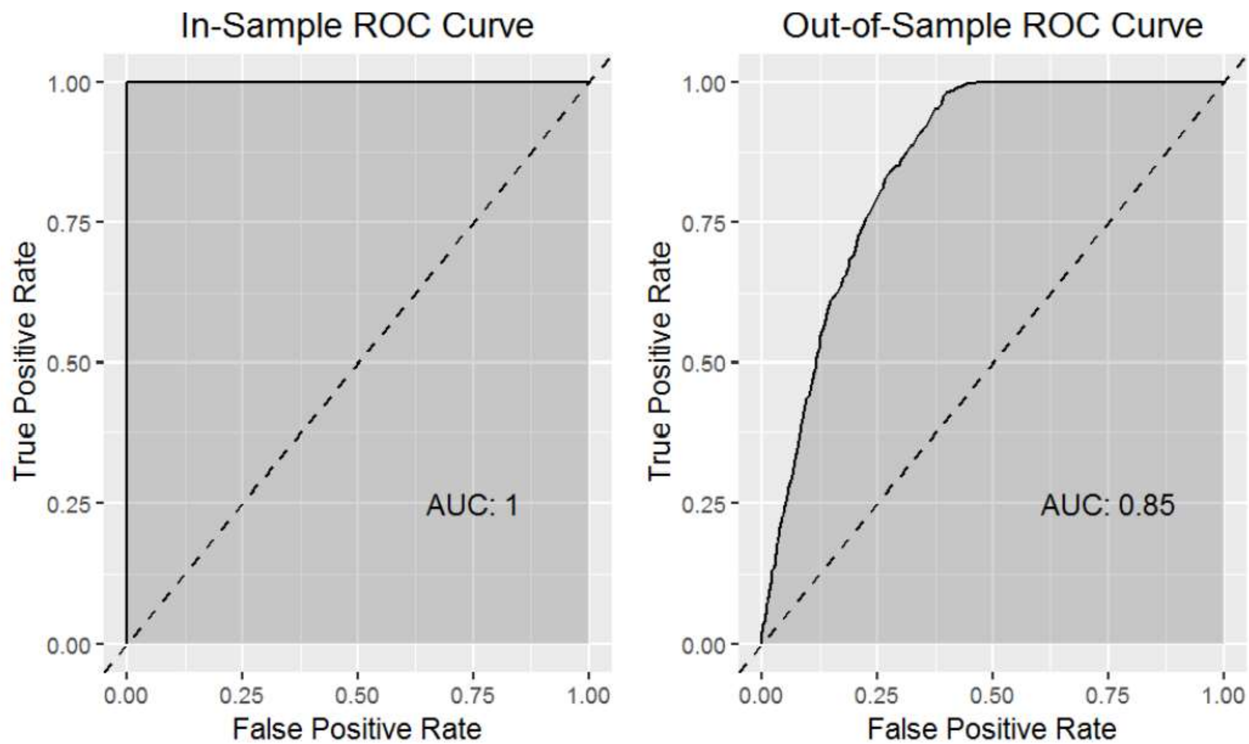

Figure 6. *ROC Curve: Random Forest*

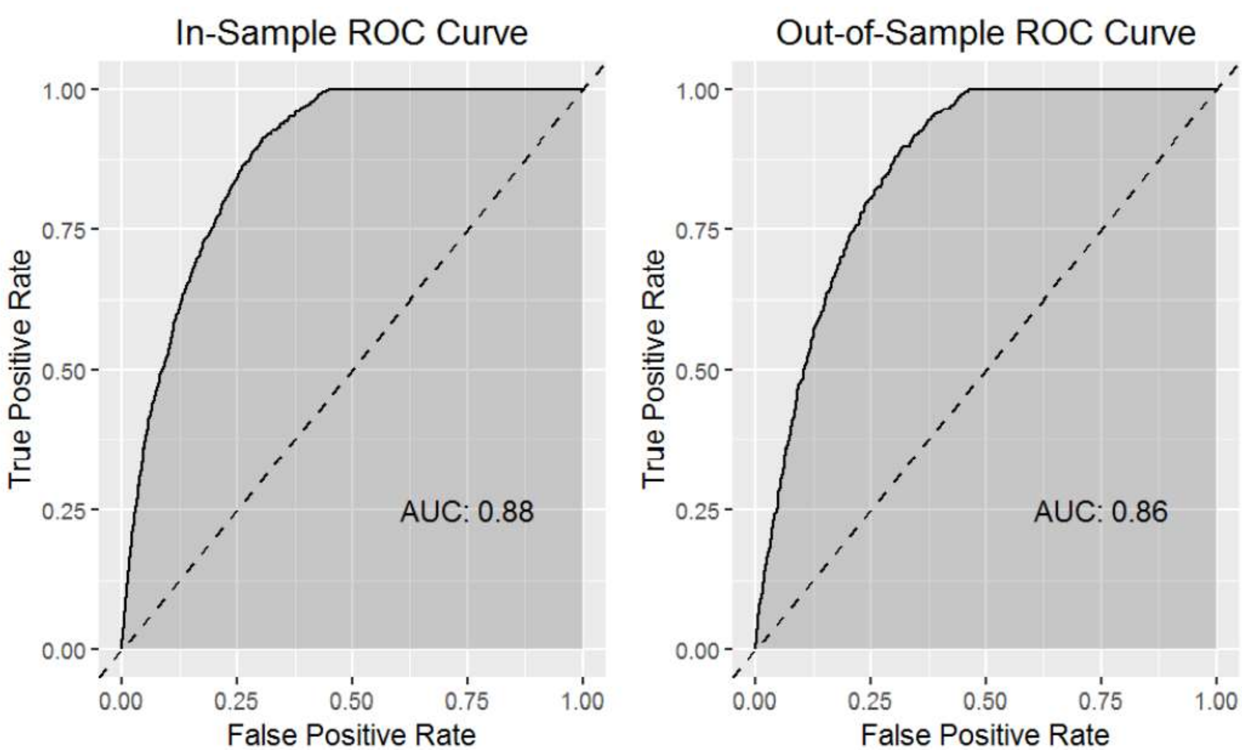

Figure 7. *ROC Curve: GLMnet*

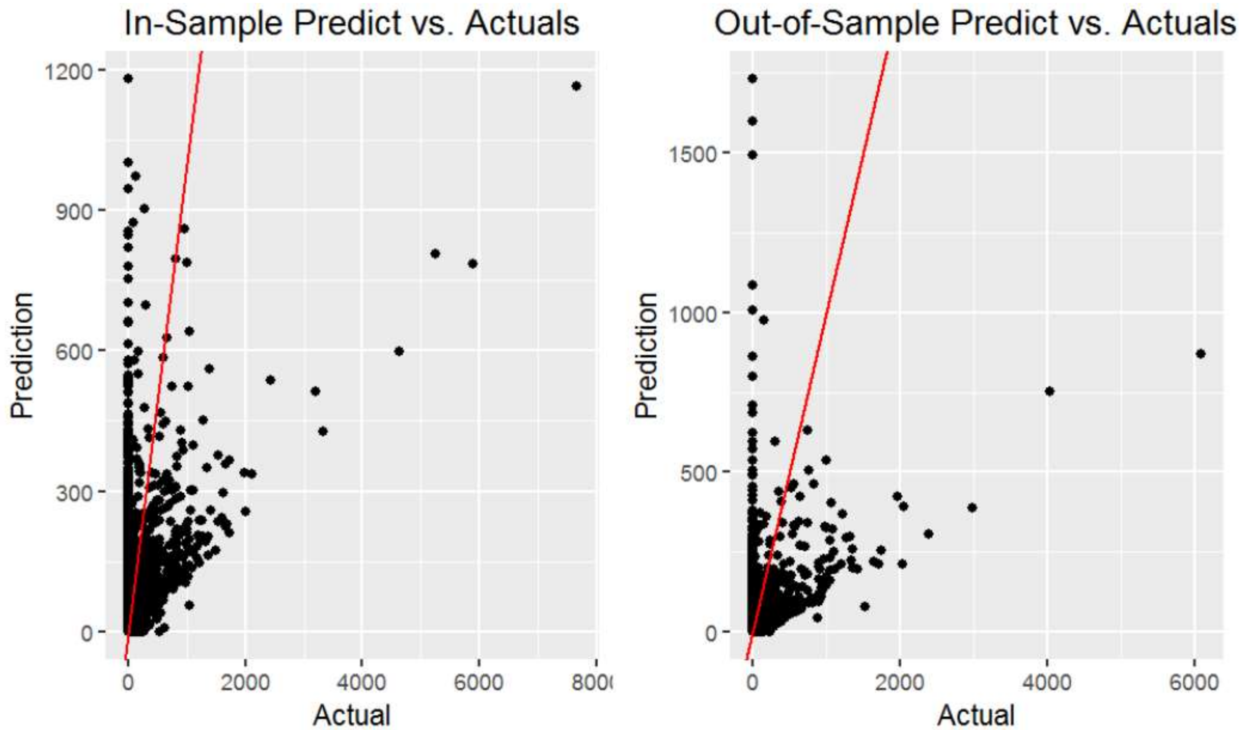

Figure 8. Actuals vs. Predictions: Multiple Linear Regression

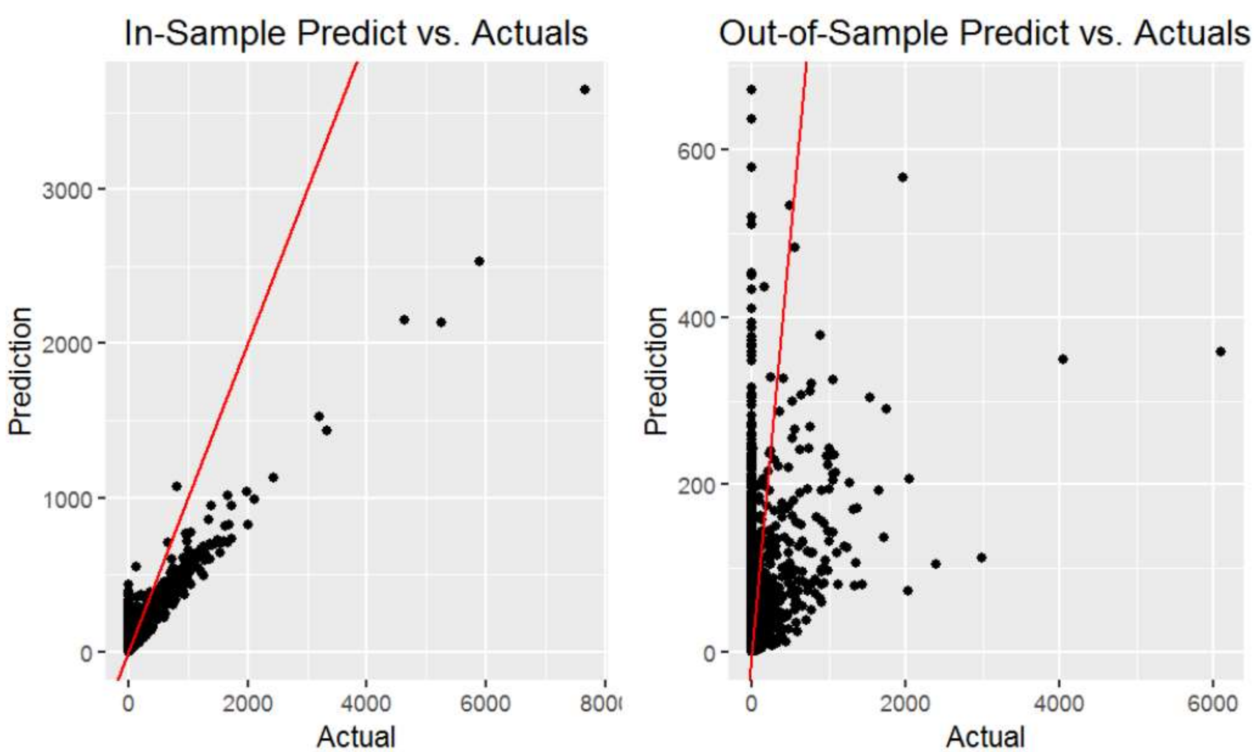

Figure 9. Actuals vs. Predictions: Random Forest

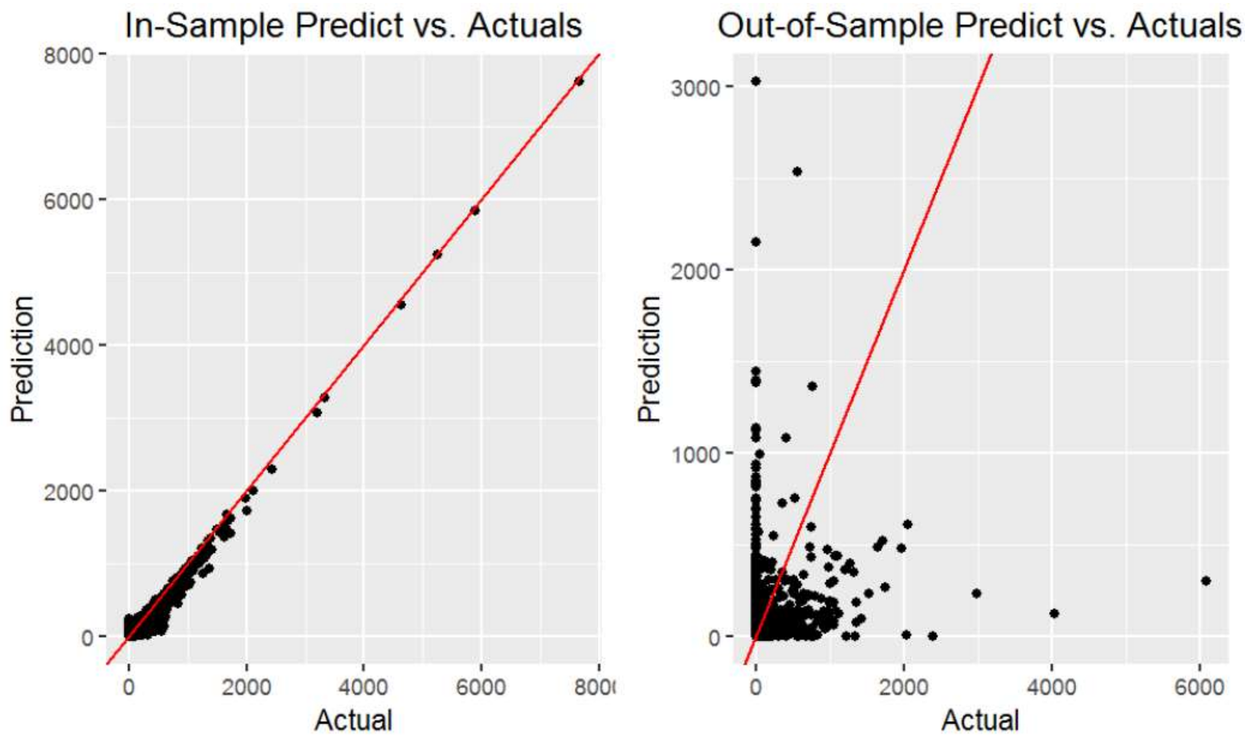

Figure 10. Actuals vs. Predictions: extreme Gradient Boost