**Part 1: Introduction**

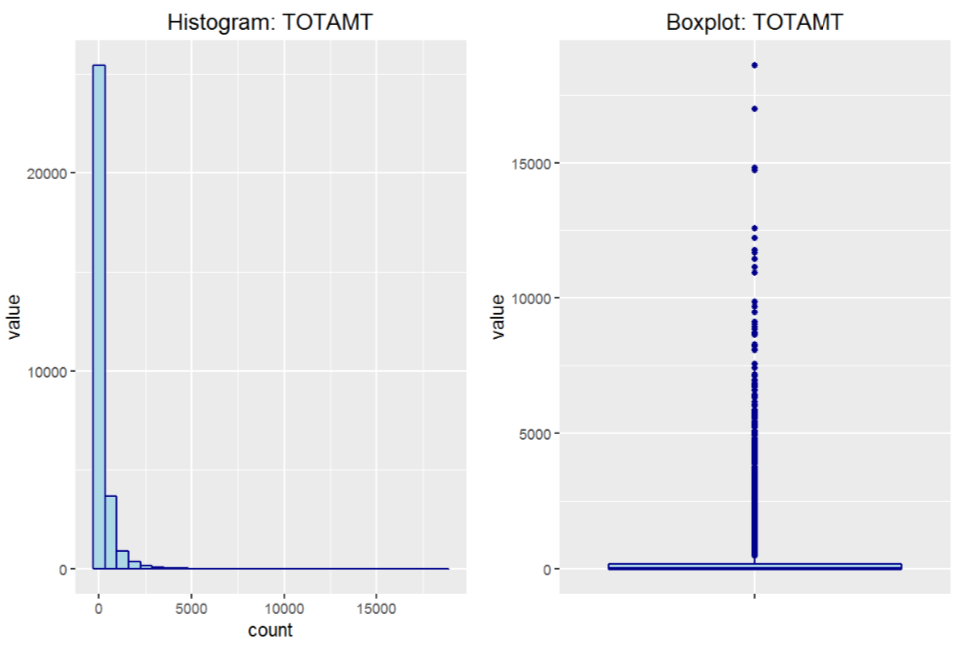
For this assignment, we are to develop and evaluate a predictive model to identify customers that are likely to respond to a mailing campaign. Once these customers are identified, we are to estimate the net profit that may result from targeting those customers as a part of the new mailing campaign. We will be using the XYZ database of customers which contains a large number of variables relating to sales, and prior campaign results. This data is used to assess which variables have the greatest ‘importance’ in determining both the chance of response to a campaign as well as the likely amount spent from each customer. Three classification models were developed to predict the chance of the response, and three regression models to predict likely spend. The best models are selected and subsequently deployed to estimate the expected net revenue from a new targeted mailing campaign, accounting for the cost of sending mail to each customer. The dataset used for this assessment is based on XYZ’s database of previous customers. The data set contains 30,779 records, and 554 features of customer data. The features that are in the data set capture sales, results from previous mailing campaigns, and Experian properties which provide information about each customer. The features are broken down into 345-character, 48 integer, and 161 numeric variables.

Due to the size of the data set, a subjective assessment of variable relevance prior to pre-processing modelling routines were performed. This was conducted 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 to spend. Another subjective analysis that was performed was to remove variables that could potentially force us to fall into a trap of predicting customer responses using a variable that’s another way to measuring the same response. This resulted in 227 variables being excluded from the data set. The remaining variables in the set could then be broken up into 119-character, 48 integer, and 160 numeric variables. An initial examination of the data revealed that the compiled R data frame did not distinguish between numeric and factor variables. Before any data exploration could be performed, we converted all character class variables to factor type and retained all other variables as numeric type. Then, we converted all ‘ANY\_MAIL\_x’ and ‘RESPONSEx’ variables to factor type, since the variables contained no character-based observations and were categorical in nature.

**Part 2: Exploration**

**2.1 Univariate Data Analysis**

The first form of data exploration that we performed was Univariate Data Analysis. As part of this analysis, summary statistics for the 160 retained numeric variable were calculated. The majority of numeric variables do not suffer from having missing values. One thing to note is that many variables do have a minimum value of zero, suggesting zero-inflated data. Histogram and box plots were generated and reviewed for a large subset of numeric variables. The total amount spent (TOTAMT) was selected for further review. No observations were removed before generating the histogram and box plots.



*Figure 2.1.1 Histogram and Boxplot: TOTAMT*

When examining the TOTAMT variable, we noticed that the variable has a strong positive skew, which is a common attribute over the majority of the numeric variables in the data set. The result is a number of observations When examining the TOTAMT variable, we noticed that the variable has a strong positive skew, which is a common attribute over the majority of the numeric variables in the data set. The result is a number of observations that could be classified as outliers.

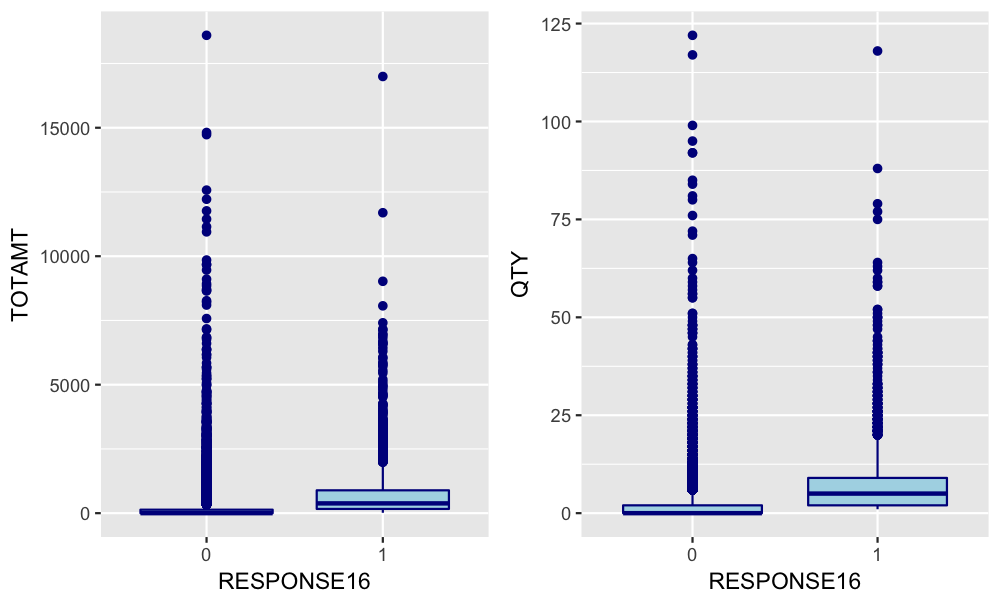
**2.2 Bivariate Data Analysis**

The second analysis that we performed was Bivariate Data Analysis. A prediction model to determine the chance of response to a mailing campaign (RESPONSE16) and the likely spend of each customer (TOTAMT16) will need to be developed. Therefore, it is advantageous to ourselves to identify variables that have explanatory power over the RESPONSE16 and TOTAMT16 variables. To do so, we calculated and reviewed the Pearson correlation coefficient for all numeric variables against our numeric response variable TOTAMT16. Correlations for the 10 most correlated numeric variables against TOTAMT16 are shown below.

|  |  |
| --- | --- |
| **Variable** | **Corr. Coeff.** |
| 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 2.2.1 Correlations vs. TOTAMT16 (Top 10)*

Even though the table above shows the top 10 correlated variables with TOTAMT16, there are no variables that have reported a strong correlation. The greatest absolute correlation is reported as being year to date sales for 2009 (YTD\_SALES\_2009) and year to date transactions for 2009 (YTD\_TRANSACTIONS\_2009) at 0.37 and 0.29 respectively. We then used bar charts to explore the relationship between the categorical response variable (RESPONSE16) and each of the numeric variables.



*Figure 2.2.1 Boxplot: RESPONSE16 vs. TOTAMT/QTY*

Looking at the tables above, we can notice that there are significant differences in both the mean and distribution of a number of the numeric variables depending on whether they are associated with a positive or a negative response to an advertising mailing campaign.

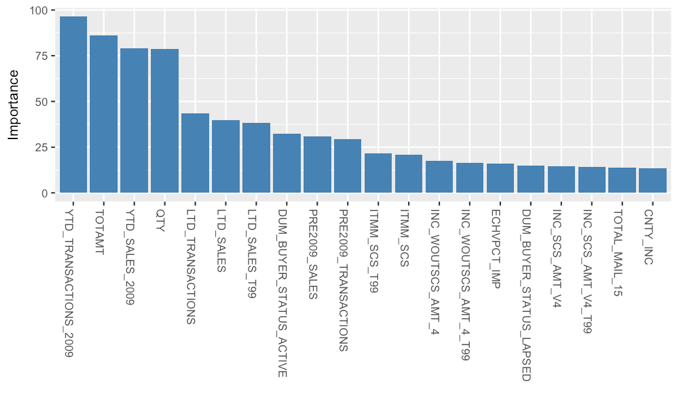
**Part 3: Data Preparation**

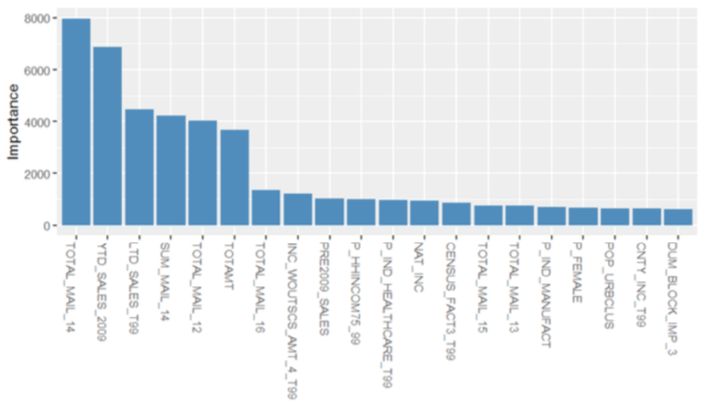
In preparing out data, we focused on imputing data for missing observations which accounted for approximately 11% of the data set. To save time with processing, we looped over each variable and imputed observations for numeric variables with the variables’ median value. While we did this we also, imputed observations for character variables with the variables’ most common value. The variables that required imputation were then copied and renamed with the suffix ’\_IMP’. The original values were the discarded from the data frame. 89 of the retained categorical variables were identified as having missing values however, only one of the numeric variables (ECHVPCT), had missing values.

The outlier observations for numeric variables were then handles. The task of identifying outlier observations can be subjective. To remove as much subjectivity as possible, we took a statistical approach and targeted observations that fell outside the 1st and 99th percentile. Observations that met this criterion were modified to create a set of trimmed variables. The trimmed variables were added to the dataset and are denoted by the suffix ‘\_Trim’. We then created dummy variables for each of the retained factor variables. The first option was to convert all factor variables to dummies but, we noted that many had a high count and a number of levels with low occurrences. Therefore, we only used dummy variables for factors that had 10 or less levels. All dummy variables include the prefix ‘DUM\_’, along with a suffix to represent the factor level. This resulted in the creation of 351 dummy variables.

**Part 4: Variable Relevance**

Data preparation created a data frame of 586 numeric variables. Variable count reduction was necessary to create accurate models moving forward. In order to do so, we utilized the varImp function which is part of the caret package in R to calculate the variable importance according to both response variables. Variable importance or relevance was calculated by fitting a Random Forest model, with ‘importance’ measured by the mean decrease in node impurity. Charts for the 20 most important variables for each response variables can be found below.

*Figure 4.1 Variable Importance: TOTAMT16 Figure 4.2 Variable Importance: RESPONSE16*



From the two charts we can observe that the two variable importance plots are similar with YTD\_SALES\_2009, TOTAMT, AND PRE2009\_SALES within the top-10 rank. There is also a quick drop-off in variable importance beyond the first five variables within both plots. From the results, we decided to move the top-50 ranked variables by importance, to our modeling phase. The reduction provides an educated balance of performance and accuracy.

**Part 5: Modeling**

**5.1 Classification Model: Chance of a Response**

We fit three classification models to predict the chance that a customer will respond to a mailing campaign. These models include Naïve Bayes, Random Forest and a Lasso Elastic-Net Regularized General Linear Model Classifier. For each model, we leveraged the train function as part of the caret package with a 3-fold cross-validation sampling method. We then applied it to a 70% subset of training data and tested it against a 30% subset. Default parameters were used for each model. The in and out-of-sample Receiver Operating Characteristic Curves (ROC Curve) are shown in the appendix. The Naïve Bayes classifier delivered an in-sample Area Under the Curve (AUC curve) of 0.85, white the out-of-sample AUC was 0.82. We did not overfit the training data with the classifier and maintained similar classification performance over both the training and test sets. The Random Forest classifier reported an in-sample AUC of 1.00 and an out-of-sample AUC of 0.85. This would suggest that it is dramatically over-fitting. The GLMnet classifier generated an in-sample AUC of 0.87 and an out-of-sample AUC of 0.86. The out-of-sample confusion matrix for each model can be seen below.

|  |  |  |
| --- | --- | --- |
| **Naïve Bayes** | | |
|  | Pred: 0 | Pred: 1 |
| Actual: 0 | 3503 | 246 |
| Actual: 1 | 496 | 218 |

|  |  |  |
| --- | --- | --- |
| **Random Forest** | | |
|  | Pred: 0 | Pred: 1 |
| Actual: 0 | 3,992 | 452 |
| Actual: 1 | 7 | 12 |

|  |  |  |
| --- | --- | --- |
| **GLMnet** | | |
|  | Pred: 0 | Pred: 1 |
| Actual: 0 | 4,004 | 416 |
| Actual: 1 | 20 | 38 |

*Tables 5.1.1 Confusion Matrices: Classification Models*

From the figure above, we noticed that while the GLMnet and Random Forest classifiers produced similar out-of-sample performance in regard to their ROC curves, the Random Forest classifier has done so by providing less true positive and true negative values. This is proven further when examining their performance metrics in the table below. This shows that the GLM classifier was able to obtain a superior true negative rate and true positive rate.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Naïve 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 5.1.2 Classification Model Comparison*

The performance metrics display that the GLMnet model has a superior AUC, sensitivity, accuracy, and specificity compared to all of the other models. The GLMnet classifier will be used to predict the chance of response from customers.

**5.2 Regression Model: Amount Spent**

The next models that were fit were the three regression-based models. These models will be used to predict the amount a customer will spend. A Multiple Linear Regression (MLR), Random Forest and eXtreme Gradient Boost linear regression estimator are also included in the model. A stepwise variable selection technique was used based on the Akaike Information Criterion (AIC) for the MLR. As with other classification models, we utilized a 3-fold cross-validation sampling method, and kept the same 30/70 split between the test and training sets.

The in and out-of-sample actual values versus the predictions for each model are shown in the appendix. Each model struggles with both outlier observations and the zero-inflated predictor information. The model assessment can be extended by observing the model fit statistics below. The negative response values were taken as zero for the statistics that are shown. All negative spend amounts should be treated as zero.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MLR** | **Random Forest** | **Gradient Boost** |
| 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 |

*Table 5.2.1 Performance Metrics: Training Set Regression Model Comparison*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MLR** | **Random Forest** | **Gradient Boost** |
| 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 5.2.2 Performance Metrics: Test Set Regression Model Comparison*

Each of the regression models performed poorly over the test set data. Combining the predictions with chance of response will aid in dealing with the zero-inflated response data. We will adopt the MLR regression to predict the amount spent as its training performance metrics were the most favorable.

**Part 6: Customer Scoring**

A customer score was constructed based on the combined predictions of the Random Forest classification and MLR model. The function represents 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 can be found below.

*CustomerScore = P(response) \* E(netrevenue) – CostofMail*

In the function above, ‘P(response)’ represents the probability of response as predicted by the chosen classification model, Random Forest. ‘E(netrevenue)’ is the expected net revenue, which is assumed to be 10% of the amount spent as was predicted by the chosen regression model, MLR. The ‘CostofMail’ is the cost of mailing customers as part of a new advertising campaign which is to be equal to $3.00 per customer. The sum of customer scores is the expected value from a new targeted advertising campaign.

The customer score formula above was used to propose four possible marketing strategies for XYZ to use. The first strategy is **ALL\_MAIL** which does exactly what the same suggests, in that they are to mail all customers who have not yet been sent a direct advertising mail piece, regardless of the response probability or the expected net revenue. This strategy is the most expensive. The second strategy is the **SCORE\_MAIL** strategy which is where XYZ will mail only those customers who have a positive score according to the above formula. The third strategy is the **HPROB\_MAIL** strategy where only those customers who are predicted to have a probability of response greater than or equal to 70% are send a mail piece. This strategy ignores the customer score and could potentially target customers who have a negative expected return when accounting for the predicted spent amount. The final strategy is the **HVAL\_MAIL** that only targets mailing customers who have a predicted spend amount of greater than or equal to $500 or $50 in net revenue. This final strategy also ignores the customer score, and may capture customers who have a negative expected return when accounting for the corresponding probability of response.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Strategy** | **Criteria** | **No. Mailed** | **Expected Value** | **Value per Customer** |
| ALL\_MAIL | ANY\_MAIL\_16 = 0 | 15,857 | -$42,630.25 | -2.69 |
| SCORE\_MAIL | Customer Score >=0 | 1,111 | $418.33 | $0.38 |
| HPROB\_MAIL | P(response) >= 0.7 | 19 | 352.06 | $18.52 |
| HVAL\_MAIL | E(net revenue) >= 50 | 20 | 652.07 | $32.60 |

*Table 6.1 Customer Scores*

The greatest expected value comes from the strategy that involves targeting all customers with a high customer scores. We will also find viable strategies from mailing only those customers with a high probability of response or a high predicted spend amount. The two strategies can achieve a much higher expected value per customer. It very well could be that the most effective marketing strategy would be to target those customers flagged by HPROB\_MAIL and HVAL\_MAIL in the first wave of mailing. Afterwards, depending on the success, XYZ should proceed to target the remaining customers flagged by ALLSCORE\_MAIL. We can also note a couple of observations when looking at the customers which were flagged by the ALL\_MAIL tag. The majority of customers were flagged as ‘active’ customers, who are homeowners with incomes between $50,000-$150,000. These same customers are educated with a Bachelor’s degree or higher. Other observations can be noted when we compare the predicted customer response with those who have previously been mailed. Of the 14,922 customers who were previously mailed, 1,440 customers did in-fact respond around 10% of the time. The classifier suggests that only 71 of the 15,857 un-mailed customers have a probability of response greater than 0.5, which is around a 0.4% response rate.

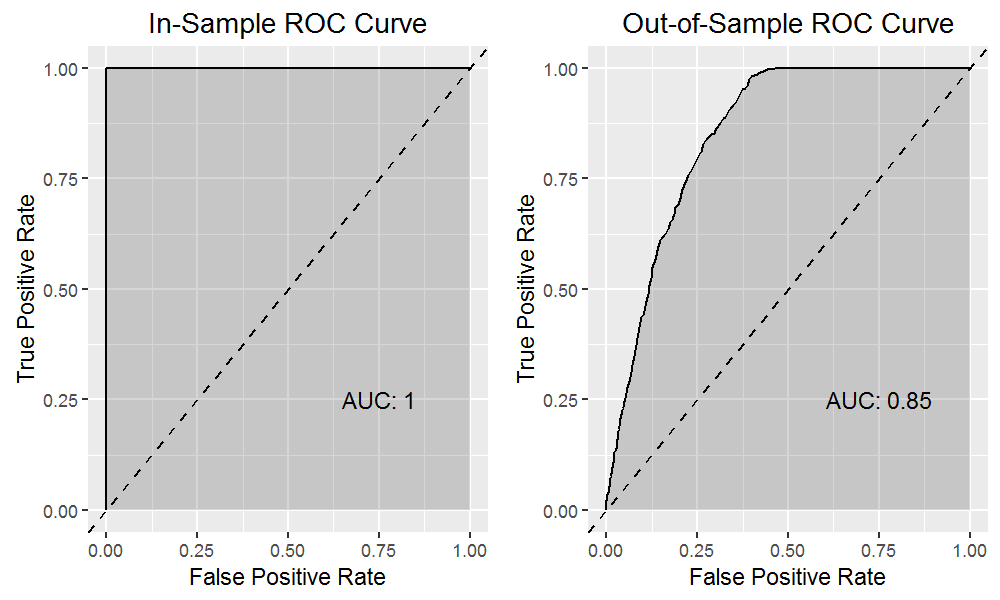
Of the customers who were previously sent a direct mail piece, the average spend amount for those customers with a response probability greater than or equal to 0.5, is only $204. It is possible that both the chosen classification and regression models are quite conservative in their predictions. It could also be the case that customers who have already been sent a direct mail piece carry a higher probability of response and a higher predicted spending amount. Either way, XYZ should explore future options for direct marketing pieces and maybe more market research is required to reach out to the 15,857 customers who have not been mailed.

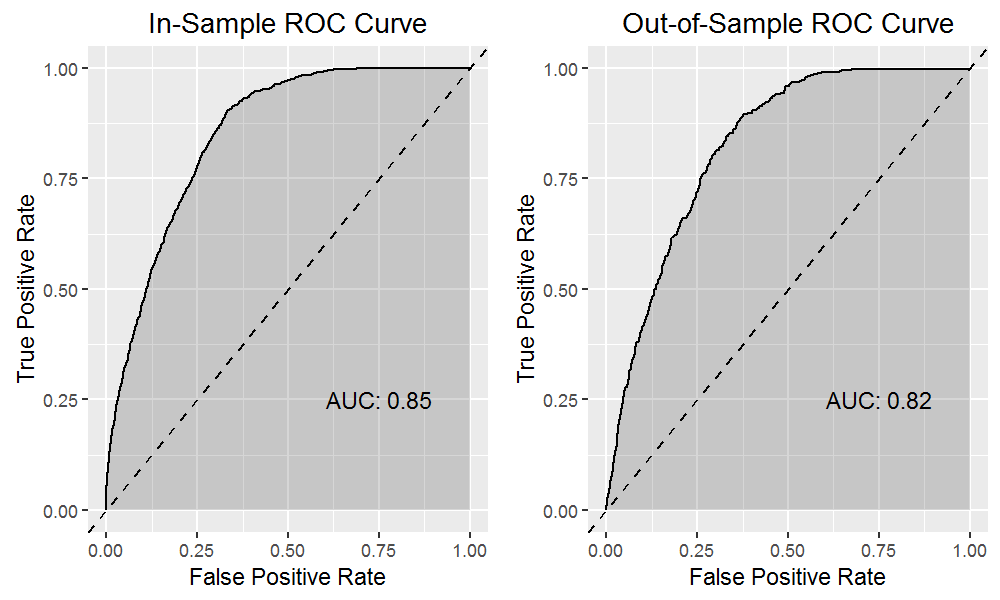
**Part 6: Customer Scoring**

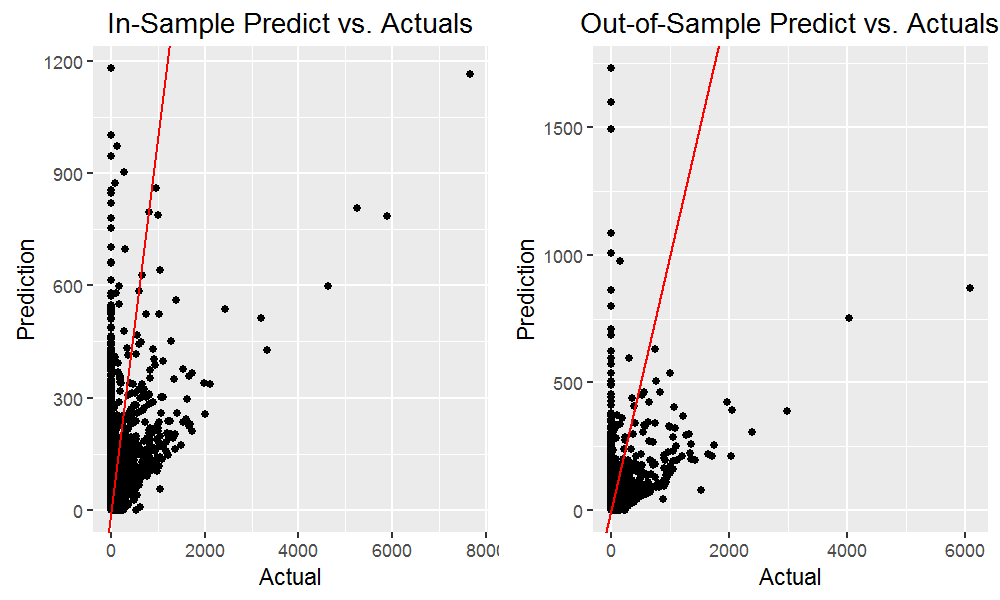
We fit three classification models to predict the probability of customer response, and three regression models to predict the customer likely spend. From these models, the GLMnet model was superior in predicting the probability of response. The MLR model was the superior model in predicting total spend. These models were selected and utilized to estimate the expected net revenue from a new targeted direct mailing campaign which accounted for the cost of mailing each customer. The scores that were calculated were used to propose four marketing strategies which ranged from mailing all customers who had not been mailed to mailing only customers who have a predicted spend amount greater than or equal to $500. These results suggest that the best strategy is to mail customers with a high probability of response and/or a high expected spend amount in the first instance, and then to follow this by mailing the remaining customers with a positive customer score. In the future, XYZ will need to expand their customer base with more customers with a high probability of response and/or a high expected spend in the first instance to maintain this marketing model.

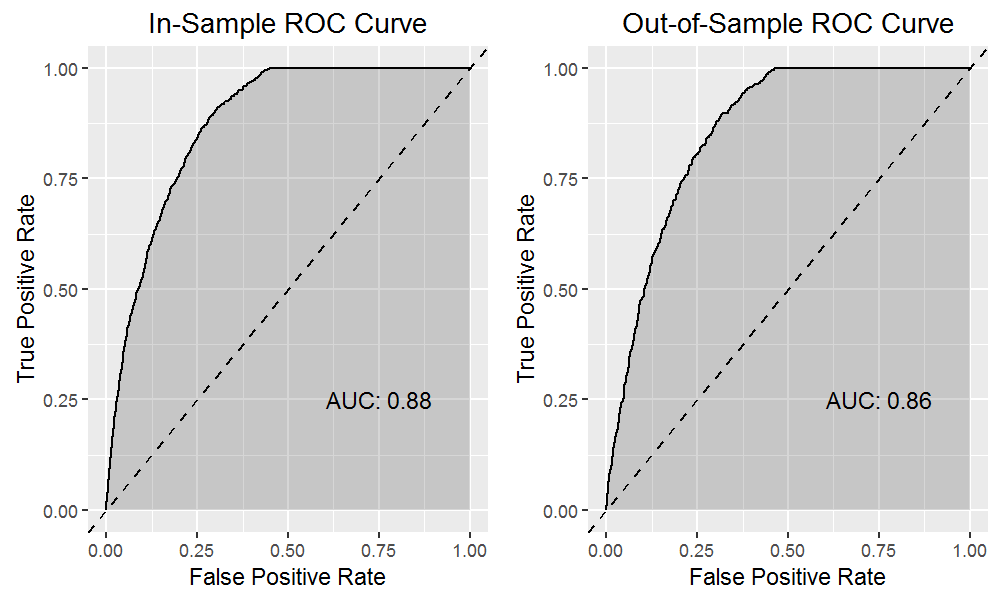
**Part 7: Appendix**

*Figure 7.1 ROC Curve: Naïve Bayes Figure 7.2 ROC Curve: Random Forest*

**

**

*Figure 7.3 ROC Curve: GLMnet Figure 7.4 Actuals vs. Predictions: MLR*



*7.5 Actuals vs. Predictions: Random Forest Figure 7.6 Actuals vs. Predictions: XG Boost*

