Nate Bitting
PREDICT 450 - Marketing Analytics
Solo 3 – XYZ Company Database Campaign Response Modeling

Data Preparation and Exploratory Data Analysis

The overall objective of this assignment is to develop a predictive model to identify which customers are likely to respond to a mailing campaign and, once identified, estimate the corresponding net profit that may result. To do this, we will be leveraging the dataset provided by XYZ company, which includes 30,779 records of customer data for previous mailing campaigns for their customer accounts. There are 554 features in this dataset including customer sales, previous mailing campaign results, and Experian features that provide additional insight about XYZ's customers. To develop the model, XYZ has elected to use the latest mailing campaign results, the 16th campaign they have conducted to date. There are 14,922 customer records from the 16th mailing campaign with a response rate of 10%. Due to the large number of variables, we will have to explore the features to understand what might be valuable to include in a predictive model.

Before digging into each variable, we first need to understand the data types present in the dataset. Table 1 below shows a frequency table for the number of variables by class (or data type).

Table 1: Frequency Table of Variable Class

character	integer	numeric
345	48	161

Given the large number of character variables, we will need to consider variables that prove useful in our model for predicting response. Our first assumption to make is that customers who have responded to previous campaigns or have purchased in previous years (preceding the year the campaign took place) are more likely to respond to future campaigns. To do this, we will create several new variables as follows:

- **TOTAMT_before_16** = cust_16_df$TOTAMT – cust_16_df$TOTAMT16
- **previousPurchase** = ifelse(cust_16_df$TOTAMT_before_16>0, 1, 0)
- **previousResponseCount** = (cust_16_df$RESPONSE0 + cust_16_df$RESPONSE1 + cust_16_df$RESPONSE2 + cust_16_df$RESPONSE3 + cust_16_df$RESPONSE4 + cust_16_df$RESPONSE5 + cust_16_df$RESPONSE6 + cust_16_df$RESPONSE7 + cust_16_df$RESPONSE8 + cust_16_df$RESPONSE9 + cust_16_df$RESPONSE10 +

cust_16_df$RESPONSE11 + cust_16_df$RESPONSE12 + cust_16_df$RESPONSE13 + cust_16_df$RESPONSE14 + cust_16_df$RESPONSE15)

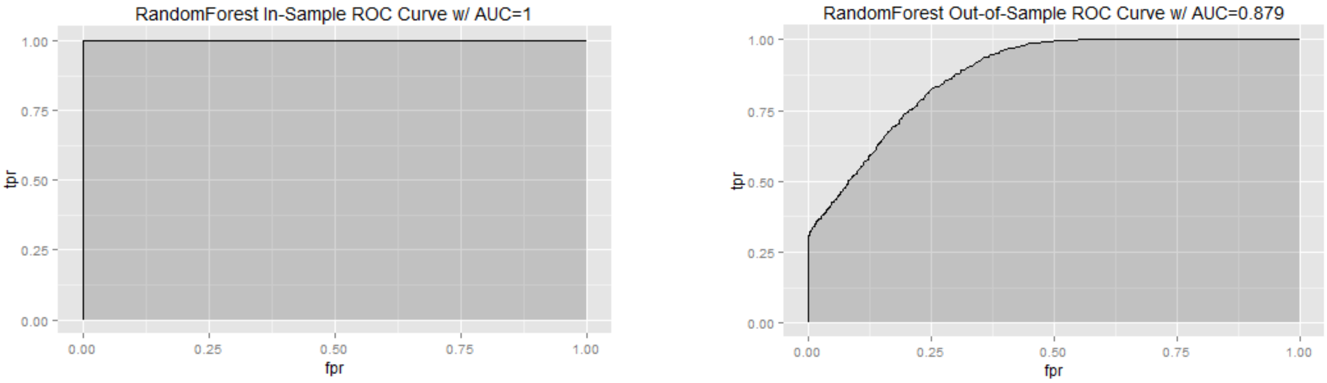
- **previousResponse** = ifelse(cust_16_df$previousResponseCount>0, 1, 0)
- **relativePurchase** = cust_16_df$TOTAMT_before_16 / mean(cust_16_df$TOTAMT_before_16, na.rm=TRUE)

These newly created variables will be extremely valuable in our modeling as customers who have responded to previous mailing campaigns or have made previous purchases before the campaign in question should have a higher likelihood of responding to future campaigns. We chose to exclude any sales or transaction fields for 2009, the same year the 16th mailing campaign occurred. The majority of the remaining variables in the dataset were factors with varying levels. For those variables, we simply created dummy variables for all levels (N-1). Lastly, we constructed a training dataset for model estimation and a test dataset to assess model performance. For this exercise, we took a random sample of 65% of the records for the training set and 35% for the test set.

Response Model Estimation and Evaluation

From a modeling perspective, we chose to use Random Forest (Model A) from the randomForest package, Logistic Regression (Model B) using glm, and a Lasso and Elastic-Net Regularized Generalized Linear Model from glmnet (Model C). To assess model performance, we will leverage ROC, AUC, and analysis of a confusion matrix for all models. Table 2 below provides the ROC and AUC results of the Random Forest model.

Table 2: Model A - Random Forest ROC and AUC

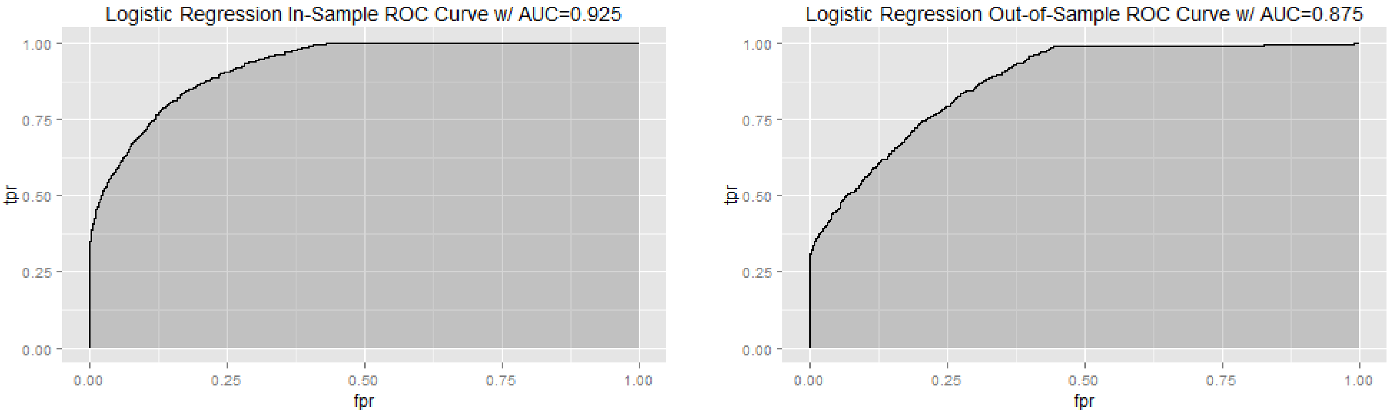


RandomForest In-Sample ROC Curve w/ AUC=1

RandomForest Out-of-Sample ROC Curve w/ AUC=0.879

Our next modeling approach is to use logistic regression to estimate customer response to the mailing campaign. We chose logistic regression as it is a commonly used model for binomial response classification. Table 3 contains the ROC and AUC results for our logistic regression model.

Table 3: Model B – Logistic Regression ROC and AUC

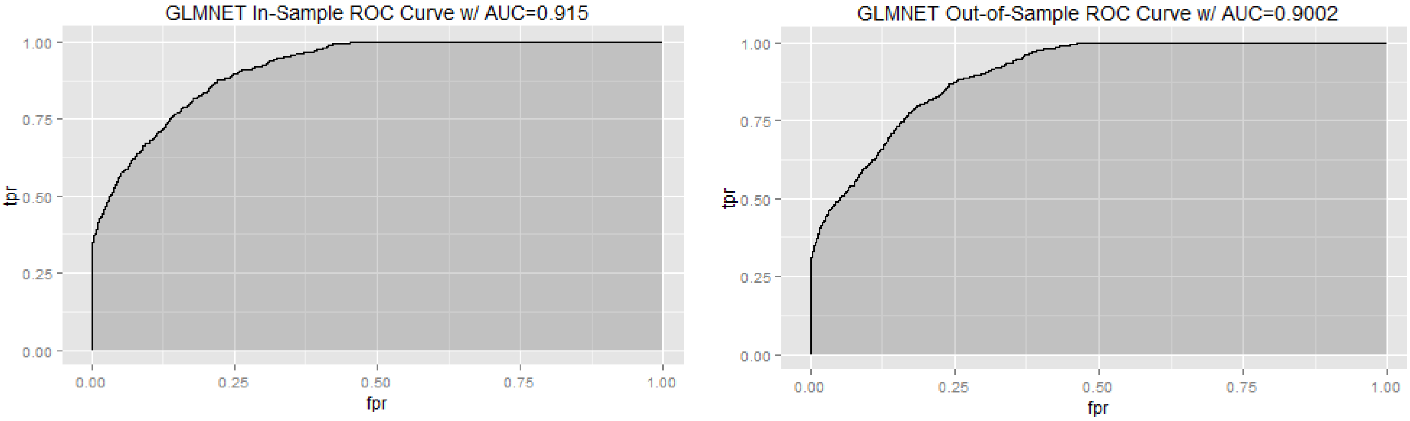

Logistic Regression In-Sample ROC Curve w/ AUC=0.925

Logistic Regression Out-of-Sample ROC Curve w/ AUC=0.875

When comparing the logistic regression model results in Table 3 to the random forest model results in Table 2, we can see the Model A has slightly higher AUC than Model B.

Our final modeling approach leverages the Lasso and Elastic-Net Regularized Generalized Linear Model from glmnet (Model C). In Table 4 below, we provide the ROC and AUC results for Model C.

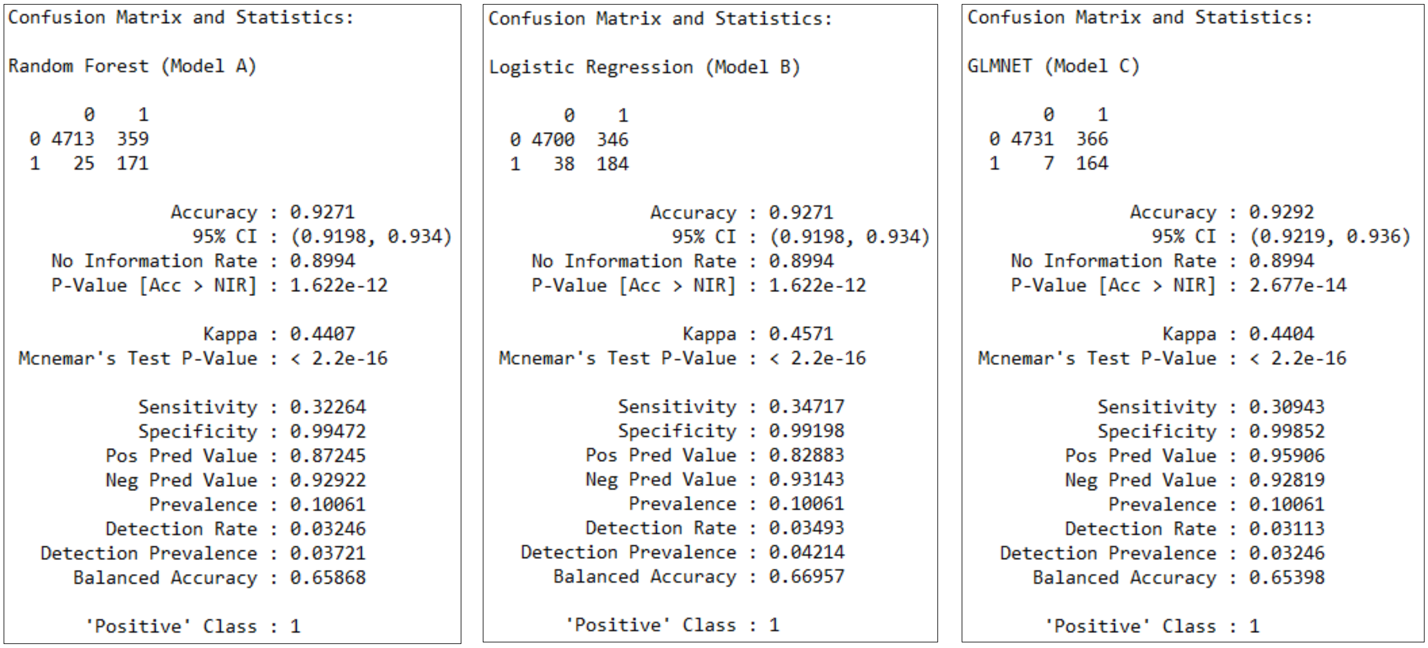
Table 4: Model C – Lasso and Elastic-Net Regularized Generalized Linear Model ROC and AUC


GLMNET In-Sample ROC Curve w/ AUC=0.915

GLMNET Out-of-Sample ROC Curve w/ AUC=0.9002

Model C's AUC is higher than both Models A and B with 0.9002 compared to 0.879 and 0.875, respectively. The AUC measure is a one method for model selection, but we will also leverage the results of the confusion matrix to assess the accuracy for each model as shown in Table 5 below.

Table 5: Confusion Matrix Results for Models A, B and C



```
Confusion Matrix and Statistics:

Random Forest (Model A)

      0    1
 0 4713  359
 1   25  171

               Accuracy : 0.9271
                 95% CI : (0.9198, 0.934)
    No Information Rate : 0.8994
    P-Value [Acc > NIR] : 1.622e-12

                  Kappa : 0.4407
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.32264
            Specificity : 0.99472
         Pos Pred Value : 0.87245
         Neg Pred Value : 0.92922
             Prevalence : 0.10061
         Detection Rate : 0.03246
   Detection Prevalence : 0.03721
      Balanced Accuracy : 0.65868

       'Positive' Class : 1
```

```
Confusion Matrix and Statistics:

Logistic Regression (Model B)

      0    1
 0 4700  346
 1   38  184

               Accuracy : 0.9271
                 95% CI : (0.9198, 0.934)
    No Information Rate : 0.8994
    P-Value [Acc > NIR] : 1.622e-12

                  Kappa : 0.4571
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.34717
            Specificity : 0.99198
         Pos Pred Value : 0.82883
         Neg Pred Value : 0.93143
             Prevalence : 0.10061
         Detection Rate : 0.03493
   Detection Prevalence : 0.04214
      Balanced Accuracy : 0.66957

       'Positive' Class : 1
```

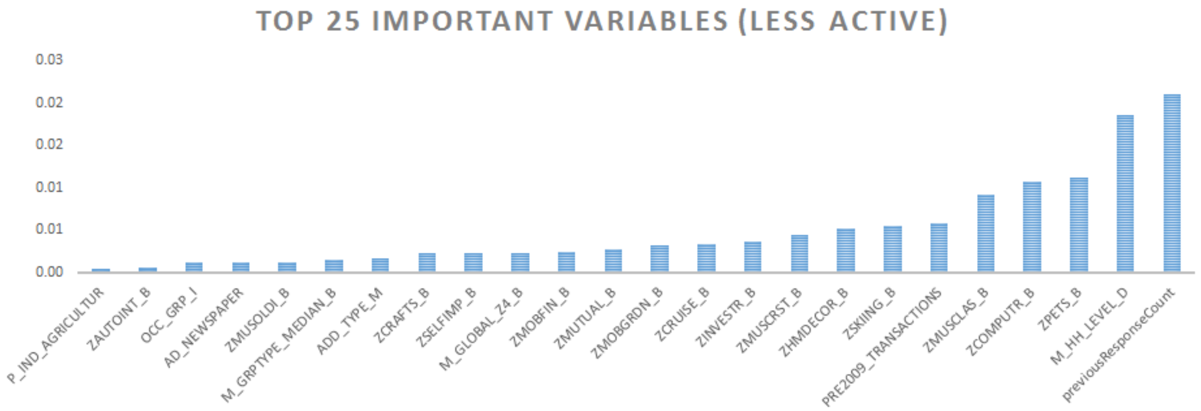
```
Confusion Matrix and Statistics:

GLMNET (Model C)

      0    1
 0 4731  366
 1    7  164

               Accuracy : 0.9292
                 95% CI : (0.9219, 0.936)
    No Information Rate : 0.8994
    P-Value [Acc > NIR] : 2.677e-14

                  Kappa : 0.4404
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.30943
            Specificity : 0.99852
         Pos Pred Value : 0.95906
         Neg Pred Value : 0.92819
             Prevalence : 0.10061
         Detection Rate : 0.03113
   Detection Prevalence : 0.03246
      Balanced Accuracy : 0.65398

       'Positive' Class : 1
```

Based on the results in Table 5, you can see the accuracy is exactly the same for both models A and B, but model C has the highest overall. We can see that precision (Pos Pred Value) of Model A is almost 5 points higher than Model B, but Model C's precision is almost 9 points higher than Model A. Specificity (when the model predicts no, how often was it actually no) for Model C is also the highest out of all three models as well. Generally, it is common practice to choose a model with the highest AUC, but a higher accuracy is even more desirable. In our case, Model C has both the highest AUC (0.9002) and highest accuracy from the confusion matrix results (0.9292). Therefore, we have elected to use the Lasso and Elastic-Net Regularized Generalized Linear Model (Model C) for our response model.

Response Model Interpretation

Now that we have selected Model C as our response model, we will now try to glean as much insight as possible for XYZ to better understand their customers. As you can see in Figure 1, customers who have responded previously has a higher probability of responding to a mailing campaign.

Figure 1: Variable Importance Plot for Model C



We actually removed the most influential feature, active, as it was skewing the chart for the next 24 important variables. This makes logical sense that active customers would have a high probability of responding to a campaign, but it is good our model confirmed this assumption for us. M_HH_LEVEL_D refers to the household level classification system and "D" indicates respondents from rural/low-income communities and it appears they have a higher probability of responding to the mailing campaign. Other interesting insights we can glean from these results is that customers who responded were financially savvy in that the Experian data indicated customers who invested or had mutual funds had a higher probability to respond. Lastly, from a marketing perspective, it is important to note that out of all forms of advertisement, only newspaper advertisements had a positive affect towards the probability of response than any other method including magazine, web, and TV.

Net Profit Model Estimation

Now that we have selected a final model to use to predict response to the next mailing campaign, we now need to develop a model to estimate the expected net profit (10% of expected revenue less cost to mail each customer) from those who respond. To do this, we will explore two model options including Random Forest for regression (Regression Model A) and Multiple Linear Regression (Regression Model B).

Regression Model A: Random Forest for Regression

Figure 2: Regression Model A w/ Outliers

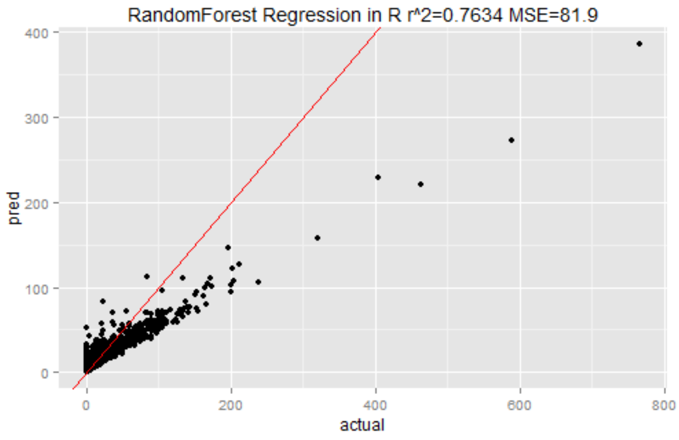
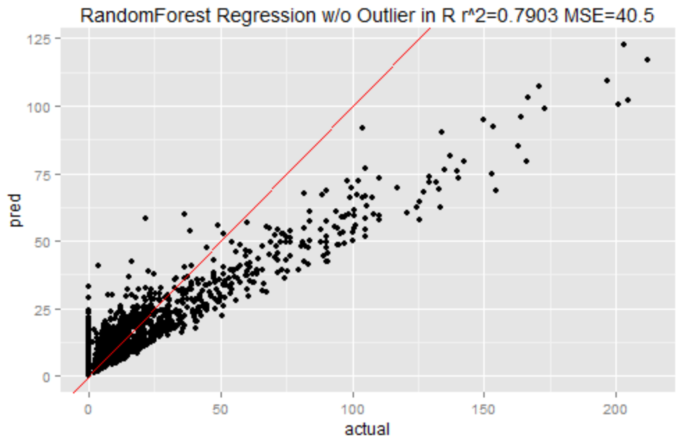


Figure 3: Regression Model A w/o Outlier



As you can see from Figure 2, there appears to be outliers that exist in the training dataset from the output of the Random Forest residual plot. Upon investigating the residuals, we uncovered a very large purchase from a particular customer that was skewing the results. With then re-ran the model excluding this record, which significantly improved the overall model fit as shown in Figure 3. The R2 went from 76% up to 79% and the mean square error (MSE) dropped from 81.9 down to 40.5. Figure 3 also suggest that our model predicts a lower net profit than actuals, which is a more conservative estimate to provide to XYZ.

Our next model approach to estimate net profit is using Multiple Linear Regression. Figures 4 and 5 provide the prediction vs actual results for net profit using a model that includes all observations and one excluding outliers, respectively.

Regression Model B: Multiple Linear Regression

Figure 4: Regression Model B w/ Outliers

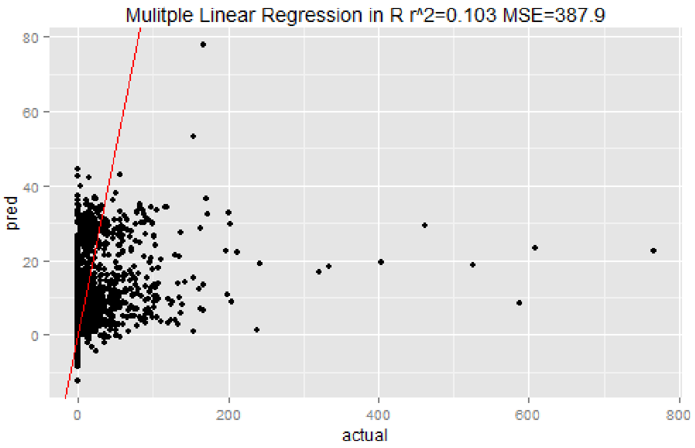


Figure 5: Regression Model B w/o Outlier

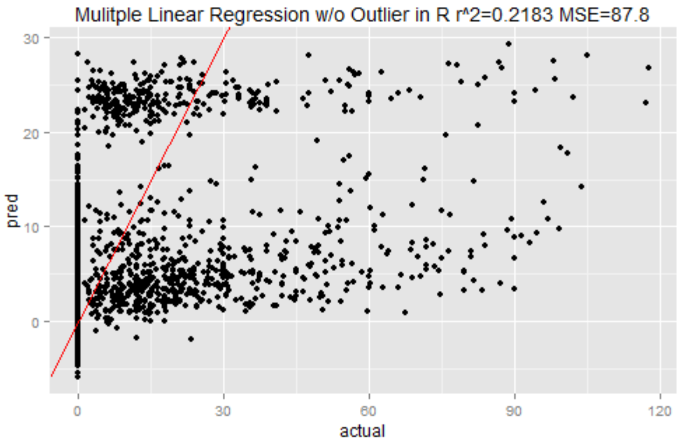


Figure 4 indicates multiple linear regression did not perform very well against the data and Figure 5 did not significantly improve after removing outliers from the dataset. We can clearly see there were many observations with large residuals.

Net Profit Model Evaluation

Table 6 was constructed to compare the overall results of Models A and B using various measures including R-squared, AIC, BIC, Mallow's Cp, mean squared error (MSE), and mean absolute error (MAE).

Table 6: Net Profit Model Selection

		Regression Model A: Random Forest	Regression Model B: Multiple Linear Regression
Training Sample	R-squared	94.8	21.83
	AIC	36565.9610	43837.5091
	BIC	38218.1621	45488.7158
	Mallow's Cp	459.9989	459.9995
	MSE	40.8081	87.8067
	MAE	2.4517	5.0282
Test Sample	MSE	89.3899	143.8978
	MAE	2.9782	4.1709

As shown in Table 6, the Random Forest regression model performed significantly better according to the results of all measures. The R-squared was higher; AIC, BIC, Mallow's Cp, MSE and MAE were all lower for the training sample of Model A. The MSE and MAE were also both lower for Model A vs Model B using the test dataset as well. Thus, we will use the Random Forest regression model to estimate net profit.

Customer Score

Now that we have selected the Lasso and Elastic-Net Regularized Generalized Linear Model (Model C) for our response model and the Random Forest (Regression Model A) for predicting net profit, we will now generate a score combining the estimates from both models. The score will be equal to the probability of a customer responding to the next mailing campaign multiplied by the expected net profit (estimated revenue less cost of sending out the mailers) as shown in Formula 1 below. This will give us a weighted score that we can then sort in descending order to see high potential customers who should generate revenue for XYZ. This score will be calculated against all customers who were not targeted in the 16th campaign.

Formula 1: Customer Score

$$\boldsymbol{Customer\ Score} = P(response)\ x\ E(Net\ Revenue) - \$1.00\ (cost\ of\ mailer)$$

Conclusion

Upon calculating the score against all customers who did not participate in the 16th campaign, we were able to identify 2,073 out of 15,857 (13%) customers who would yield an expected net profit greater than $0. By sending catalogues to these customers, the total mailing costs would be $2,073 and the expected net profit would be $13,591. In order for XYZ to better understand the characteristics of these target customers we have outlined some of the demographic and interesting takeaways for these 2,073 customers as follows:

- 74% are repeat customers and have responded to previous campaigns
- 83% are homeowners
- All customers reside in 30 zip codes (98% in 22)

- 40% reside in 3 zip codes (60093, 60091, 60067)
- All are *active* customers
- 56% have a regular Visa and 49% have a regular MasterCard
- 46% have household incomes between $50 - $150K
- 14% have household incomes between $175 - $200K
- 20% have household incomes greater than $250K
- 81% of customers live in rural areas and have an average commute greater than 3 hours
- 71% have a bachelor's degree or higher

By acquiring more data from Experian, XYZ can leverage these key facts for targeting new customers who possess similar characteristics as highlighted above.

Although XYZ may not be able to get all of the information needed to predict a score for future customers, we recommend trying to develop a similar model only using data XYZ can obtain for new customers. For example, XYZ could develop additional response models only using the features they are able to obtain for net new customers using the historical customer data already available. While it may not be perfect, it could yield some insight as to what features in the Experian dataset hold predictive power in estimating response and net profit.

The entire purpose of this assignment was to provide a prioritized list of customers for XYZ to target in the 17th mailing campaign (the 2,073 customers mentioned earlier). To develop a test to see how well our models are performing, we could identify a randomly selected subset (maybe 10%) of these customers and send them catalogues. We could then measure the response of these subset of customers before committing to a full blown mailing campaign that would cost XYZ a decent amount of investment dollars. If the test is successful, we can then yield higher confidence in our modeling approach and provide XYZ with a more scientific way for targeting their customers in the future.