**Creating Loyal Customers**

Increasing Customer Participation

Springboard DSC Capstone 1

Lisa Moller

**1 Introduction**

For nonprofits to retain and grow their donor base, they need to understand donor characteristics and behaviors. With the large amount of data now available, donor analytics can help harness the vast amount of information available and have a substantial impact on an organization’s planning, budgeting and forecasting. Donor analytics help nonprofits…

**Target Current Donors**: By better understanding effective campaign strategies, organizations can target current donors with their preferences. Several examples include…

-       How the donor is contacted: if a group of donors shows that email is more effective than a physical mailing, then the organization can use that information to their advantage.

-       Which events or campaigns to leverage for each group:  There may be certain donors who love going to happy hour events, and others who enjoy volunteer events. By better understanding what drives each donor to engage, i.e. – their lifestyle and preferences, nonprofits can better engage with donors on the donor’s level.

-       Pinpoint timing: By better understanding the donor base, the nonprofit also has insight into when their different segments of donors are most likely to donate.

-       Set “donor levels” or specific campaigns: Nonprofits would have a better idea of specific “asks”. For example, Bernie Sanders often talked about his average $27 donation. This gave people a specific amount to think about when they went to donate. Or another example is the Silver/Gold/Platinum membership level structure.

**Find New Donors**: Understanding the donor base allows nonprofits to grow their donor base by finding donors with similar interests and drives. The deeper knowledge allows a more focused approach to reaching out to new donors.

**Budget More Effectively**: The organization can manage costs based on predicted Lifetime Value of the donor.

The client targeted by this project is a hypothetical clothing store, “Threads”. Threads sent out a mail marketing promotion and tracked which customers responded to the promotion, along with other behavioral metrics, such as the number of visits to the store, amount of money spent, product bought and response to previous promotions. Threads would like to better understand the behavior of the customers in general and according to their response, or lack thereof, to the promotion. By better understanding customer behavior, Threads could use that information to achieve two goals, 1) increase participation in future promotions and 2) increase sales. The aim of this analysis is to help Threads accomplish those two goals.

**2 Dataset**

The dataset used in this project came from the case study discussed in chapter seven of the book “Data Mining Methods and Models”, by Daniel Larose (2006). The dataset was retrieved from this website: [http://www.dataminingconsultant.com/DMMM.htm](http://www.dataminingconsultant.com/DMMM.htm" \t "_blank).

The data is at the customer level, tracked by a customer ID. The data has zip code, the number of purchase visits, total money spent, the average amount spent per visit, percent to total sales for each product type, amount spent at four stores, whether the customer is a credit card holder or Internet shopper, the percent of returns, and the response rate for other promotions.

The one issue that arose was that there was no specific documentation provided for the dataset. The book referenced the features with general descriptions and in no particular order. Therefore, general descriptions of the columns from the book had to be manually matched to the columns in the dataset. There were two columns with no clear descriptions, but they did not prove to be significant features.

**3 Data Wrangling**

Generally Data Wrangling is finding, structuring, cleaning and adding to your data, validating and preparing for analysis. Because I am using the dataset from the source mentioned above, much of the finding/structuring/cleaning/validating was already completed. A fairly clean dataset was provided.

To get a sense of the data, I counted features and rows, detailed the type of each feature, highlighted the categorical columns, looked for missing values, and provided summary statistics for numerical columns.

**3.1 Data Structure**

There are 51 features and 21,740 rows. Of the 51 features, five are categorical and 46 are non-categorical. There are no null values.

**3.2 Summary Statistics**

The function ‘describe’ gives the count of each column, minimum and maximum values, the mean and quantiles (this is where it is helpful to know the data type of each of the columns because some columns are not relevant for this summary – i.e. categorical).

This function is helpful to see if there are any unexpected negative values or other outliers that should be explored. The summary also helps provide a general sense of the data. For example, the range of money customers spent: the minimum is $.99, while the maximum is $22,511.49.

**4 Exploratory Analysis**

Exploratory analysis helps form an intuition about the dataset. We kept our ultimate goal in mind, to better understand the behavior of the customers who responded to the promotion and use that information to increase participation in future promotions, while we searched for trends and patterns within the data.

Below are several interesting observations. More exploratory analysis completed for the dataset can be found at: [https://github.com/LMoller/Springboard-Capstone-1/blob/master/DataStory.ipynb](https://github.com/LMoller/Springboard-Capstone-1/blob/master/DataStory.ipynb" \t "_blank).

**4.1 Exploring Response to the Promotion**

We divided the data into groups of customers who responded to the promotion, and customers who did not. There were 3,611 customers who responded to the promotion out of 21,740 customers for a responsive rate of 16.6%.

4.1.1 Amount Spent at Each of the Four Threads Stores

Data was gathered at four Threads stores. We wanted to see if one of the stores had a higher percentage of customers who responded to the promotion. Perhaps the promotion was particularly effective at one of the stores. In Figure 1, we show the amount of money spent per store.

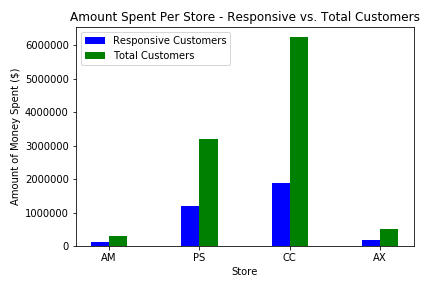


Figure 1: Amount of money spent at each store

The proportion of sales from the responsive customers to the sales from all customers looks fairly uniform for each store. While more than half of the money spent by responsive customers was at store CC, that amount looks to be proportionate to total sales.

4.1.2 Total Money Spent by Responsive and Non-Responsive Groups

Promotions are often meant to bring in new customers, or they can spur the current customers into coming back in. To explore the effect of the promotion in question, we plotted amount of total money spent by customers in the responsive and non-responsive groups in Figure 2.

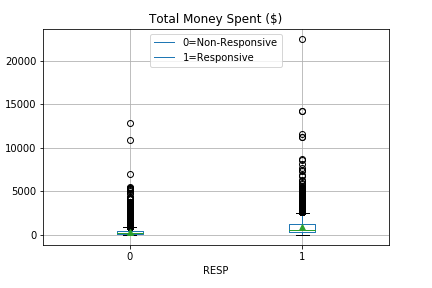


Figure 2: Amount of Total Money Spent by Responsive and Non-Responsive Groups

The customers who responded to the promotion, on average, spend more money than those who did not respond to the promotion. The mean amount of money spent for the non-responsive customers is close to zero. In fact the “box” of the boxplot, which represents the up to 75% of the data, is difficult to see because it is such a small number. The box for the responsive group is a bit more visible, showing a larger and higher range of spending.

4.1.3 Number of Product Classes Purchased

Threads tracked how many classes or types of products their customers purchased. This metric may help give them a sense of how much customers value their brand, or how much time customers spend in the store. We again looked at the difference in behavior between the responsive and non-responsive customers.

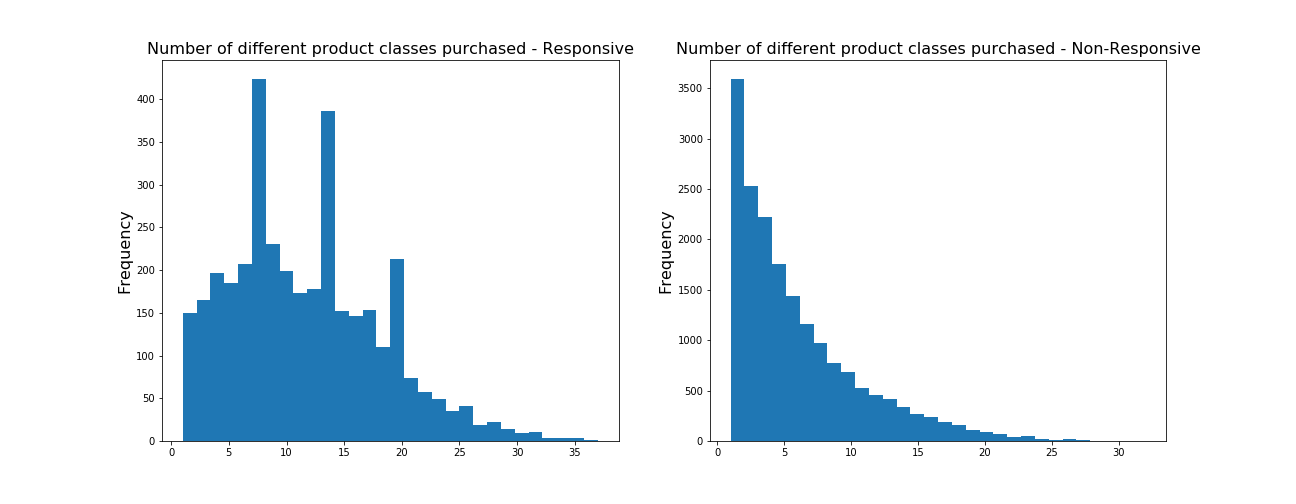


Figure 3: Product Classes Purchased

Figure 3 above shows the responsive customers have bought many more classes of products that the non-responsive customers. The histograms show the number of product classes bought by customers along the x-axis, and the frequency along the y-axis. It would be interesting to see if the promotion helped customers increase the number of classes of products, or if the people who have already bought a broader range of products are more likely to use the promotion.

Overall, responsive customers showed the following characteristics. They…

1.     have more purchase visits

2.     spent more money than non-responsive customers

3.     spent more money when they purchase items

4.     were long-time customers

5.     did not leave as much time between visits as non-responsive customers

6.     bought a larger range of products

7.     were often Internet-shoppers

Most of these trends seemed fairly expected for loyal shoppers. Loyal shoppers would be familiar with a brand for a long time; they trust the brand, so they would spend more money on the store. They would also check in fairly consistently as they need new items or to check if there are deals. Being an Internet shopper also makes sense - customers know what they like and what size they wear in the brand. It's easier to buy online when you know what you prefer before you shop.

The trend that most surprised me above was responsive customers bought a larger range of products. It makes sense that customers would build trust for the brand with one or two categories of products to start, and then expand once they trust and like the brand.

**4.2 Exploring Overall Customer Patterns**

To get a better idea of the customer behavior in general, we explored the customer set as a whole.

4.2.1 Types of Clothing

Some of the metrics tracked were the percent of sales of different products, for example, types of clothing. We set up a scatterplot to see if there were correlations in the percent of sales between the different product types.

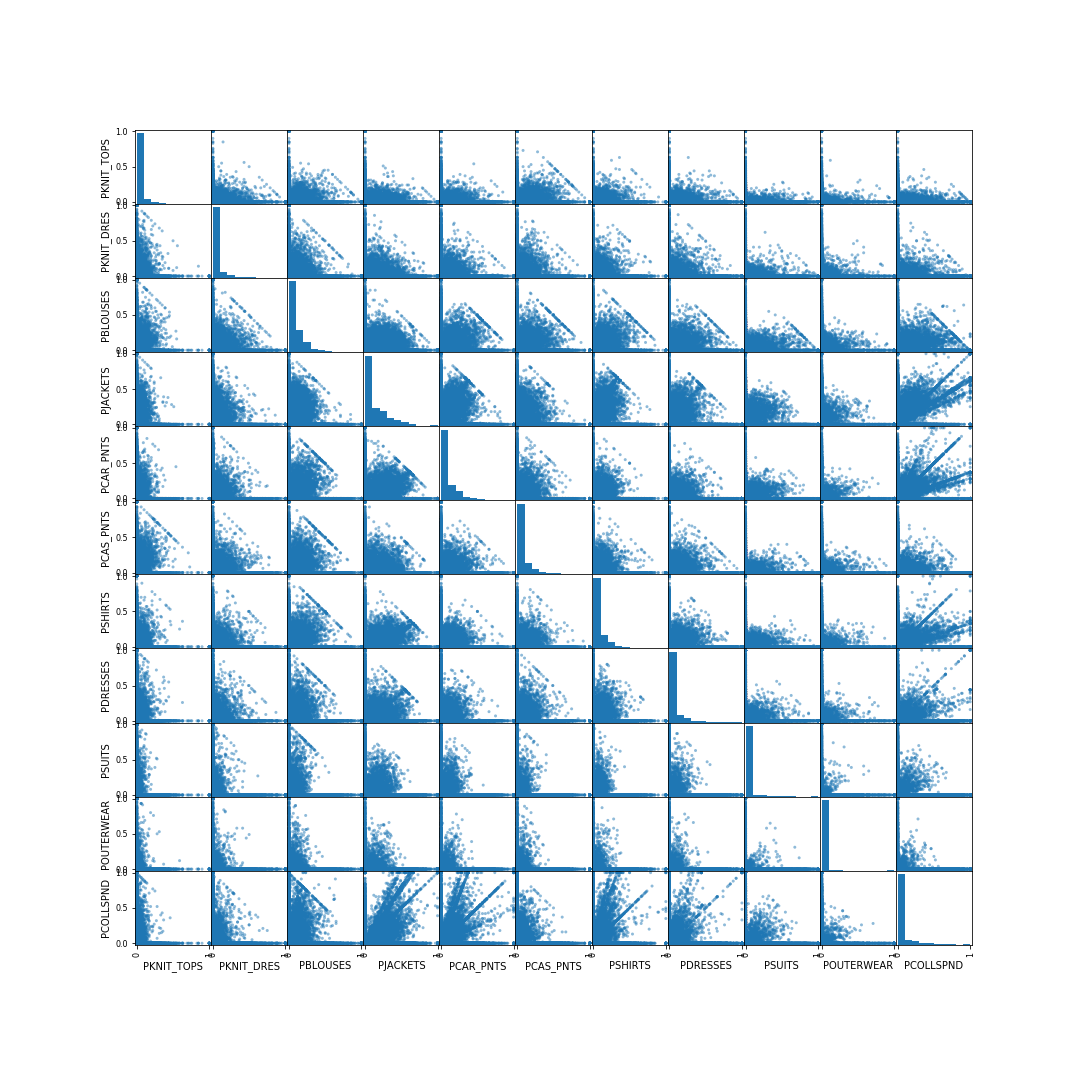


Figure 4: Percent of Sales by Product Type

In Figure 4, we can see that the Collectibles Line seems to be correlated to jackets, career pants, shirts, dresses, suits and outerwear. It would be interesting to know what is in the Collectibles line, but that information was not available. We do see a correlation between a few of the other classes, for example, suits and outerwear. We will continue to dig into the data to see if more patterns arise.

4.2.2 Percent of Sales by Clothing Type

We calculated the means of each clothing types’ percent of sales.

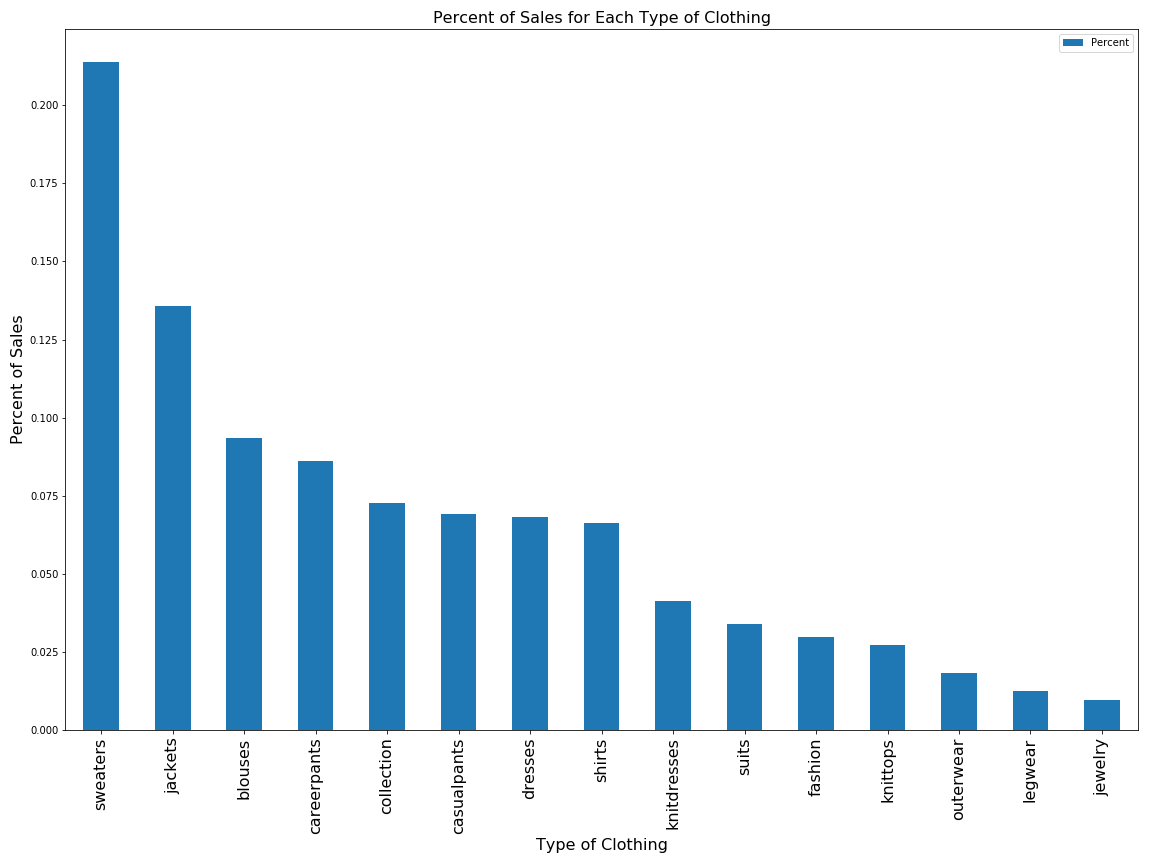


Figure 5: Percent of Sales for Each Clothing Type

Sweaters and outerwear are the largest portion of sales. For another plot, we looked at the percent of sales by zip code. The two zip codes with the highest sales were in Minnesota and Pennsylvania – two areas that can get quite cold. A metric to keep in mind if looking deeper into this data would be the last one/three/six months of sales. The time data may give a better picture also of weather patterns, and if the data is skewed because of the time of year, etc.

**5 Data Analysis**

After completing the exploratory work we verified our observations through statistical inference. After that, we used machine learning to parse the data, learn from it, and make a prediction about whether customers would respond to the promotion or not. By aggregating all of the information, from the exploratory work, statistical inference, and machine learning, we can help Threads make an informed decision on how to interact with their customers.

**5.1 Statistical Inference**

5.1.1 Testing the Significance of a Relationship

During the exploratory analysis, there seemed to be a significant relationship between the number of purchase visits, and the customers’ response to the promotion. Customers who responded to the promotion had an average of 10.92 purchase visits, while customers who did not respond to the promotion had an average of 3.90 purchase visits. We verified whether the observation was statistically significant using a two-sample t-test.

Generally the z-statistic is used if the sample size is over 30 and you know the population standard deviation. The t-statistic is used if you do not know the population standard deviation and the sample size is less than 30. While our sample size was over 30, we did not know the standard deviation for the population, so used the t-statistic. We used the two-sample t-test because we were looking at a categorical variable with the Responsive and Non-Responsive groups.

The p-value was very low (8.26 x 10-273) so we rejected the null hypothesis (that the Responsive and Non-Responsive customers had the same number of purchase visits). Therefore the difference we observed between the frequency of visits in the Responsive and Non-Responsive groups is statistically significant.

Frequent customers love to get great deals on products they trust and enjoy. Promotions may be a good way to keep them coming in and keep the product at the front of their mind.

5.1.2 Correlation

The percent of purchases of Jackets and Career Pants seemed to be correlated. As people had a higher percentage of Jacket purchases, they also had a higher percentage of Career Pant purchases. We tested the correlation using a Hypothesis Test to determine if the correlation was statistically significant.

The very small p-value (7.877 e-37) led us to reject the null hypothesis that there was no correlation between the purchase of Jackets and the purchase of Career Pants. The p-value told us there was a very small chance the results were observed by chance. There was a statistically significant relationship between the percent of purchases a customer has of Jackets and their percent of purchases of Career Pants. As customers spend more on Jackets, they spend slightly more on Career Pants as well.

**5.2 Machine Learning**

5.2.1 Prepare Data

The data had to be slightly modified for input into the machine learning models.

1.     We changed the categorical variable VALPHON from Y/N to 1/0.

2.     CLUSTYPE column was a categorical column with 50 different categories. We used get\_dummies() to flatten categories so each category had its own column with a 1/0 value.

3.     We removed the customer ID and target variables from the dataset.

5.2.2 Train/Test Split

To ensure we found the best parameters for the model and its ability to predict well on unseen data, we split the data into a training set and a test set. To do this, we used train\_test\_split().

The first parameter in the train\_test\_split is X.

X = all values from the prepared data

y = each customers’ response to the promotion (the target variable)

X and y had the following relationship…

1.     They have the same number of rows

2.     For a given row i of matrix X, the label that corresponds to that data point is exactly the value of vector y at that row

3.     The number of rows of X is the number of data points in the dataset

4.     The number of columns of X is the number of features of each data point in the dataset

By default, train\_test\_split splits to 75% train and 25% test data.

5.2.3 Models

We calculated the accuracy and a classification report for each model created (Figure 6). A classification report shows precision, recall and F1-Score.

**Precision**: Precision is the number of true positive results divided by the number of total (true and false) positive results. It measures the quality of the classification.

**Recall**: Recall is the number of true positive results divided by the true positives and false negatives (items that should've been in the positive class, but were not). This is a measure of completeness.

**F1 Score**: The F1 score is a weighted average of the precision and recall. An F1 score reaches its best value at 1 and worst score at 0.

These metrics helped compare the performance of models with each other.

We started with a logistic regression model and the model did fairly well. For the non-responsive group, the precision and recall were 86% and 98%, respectively. The responsive group did not do quite as well, the precision was 65% and the recall was only 21%, meaning it missed many of the customers who should have been marked as responsive.

To try and improve performance, we then tuned hyperparameters. In Logistic Regression, the most important parameter to tune is the regularization parameter C. The regularization parameter is used to control for unlikely high regression coefficients, and in other cases can be used when data is sparse, as a method of feature selection. We created a function to perform K-fold cross-validation and applied a scoring function to each test fold. Once we had the “best” C, we retrained/retested the data. We did not see an improvement.

From there, we moved on to exploring a random forest model. We did not see much of an improvement, so tuned the random forest model. There were many parameters to choose from, so we used Scikit-Learn’s GridSearchCV method to evaluate combinations from a defined grid. The parameters in the grid were n\_estimators, max\_features and max\_depth.

The recall for the responsive group increased some, but further tuning most likely would not help, since there is a large imbalance between the responsive and non-responsive groups, so the models have a difficult time accounting for the small portion of responses (responsive group is 16-17% of the total dataset).

To account for the imbalanced classes, we resampled the data and retrained/tested. Resampling attempts to balance the class size.

**Oversampling** randomly replicates the minority class to increase the population. We oversampled using the SMOTE (Synthetic Minority Oversampling Technique) algorithm. SMOTE finds the k-nearest neighbors of the minority dataset and uses those to randomly create similar observations. Oversampling can be done before or after the train/test split. The concern with oversampling before the split is "bleeding" the synthetic data into the test set. This would train the model to better predict the test set than a new dataset. For this reason, we performed oversampling after the train/test split, on the training set.

**Undersampling** randomly samples subsets of the majority class to decrease the population size. We undersampled using the Random Undersampling technique.

We performed both Logistic Regression and Random Forest on an Oversampled dataset and an Undersampled dataset to compare the results.

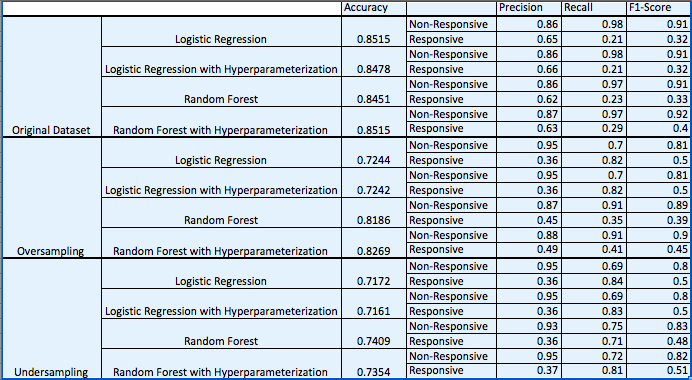


Figure 6: Accuracy and Classification Report Metrics for Each Model

5.2.4 ROC Curves

The metrics in Figure 6 were helpful to understand how the models were performing, but often can be difficult to determine if there is a clear “winner.” To compare the models using one standardized value instead of several metrics, we graphed the Receiver Operating Characteristic (ROC) curves. ROC curves compare the True Positive rate of the model to the False Positive rate. By looking at the True and False positive rates, we can prioritize what is important to get out of the model. A client's needs have a large impact on which model to choose. While all clients will want to be conscious of how they spend their money, each client will have a different comfort level with the proportions of True Positive and False Positive results. In this case, Threads is a large clothing store that presumably has a larger budget. They may be more comfortable with a larger False Positive rate than a small nonprofit.

With this in mind, we proceeded with the Logistic Regression with Hyperparameterization model from the Oversampled test run. The chosen model had one of the highest True Positive rates. By picking a model with a high True Positive rate, we have a higher likelihood of reaching customers who will respond to the promotion. The tradeoff is that the chosen model also has a higher False Positive rate than some of the other models. This means that we will also be reaching out to customers who are not likely to respond to the promotion.

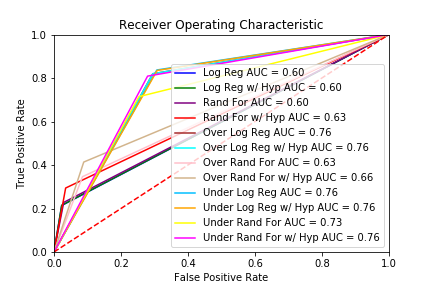


Figure 7: ROC Curves

5.2.5 Feature Importance

For the final piece of the evaluation, we looked at which features were the most important in the chosen model. The most important features are the features that have the largest impact in predicting whether the customer will respond to the promotion or not. We examined the coefficients of the chosen model fit on standardized parameters. Standardized parameters are useful when comparing parameters across different units. The parameters are normalized so they can be easily compared to each other. The 12 “most important” features were…

'LTFREDAY' - Lifetime average time between visits

'FRE' - Number of purchase visits

'FREDAYS' - Number of days between purchases

'CCSPEND' - Spending at the CC store

'PSSPEND' - Spending at the PS store

'MON' - Total net sales

'DAYS' - Number of days the customer has been on file

'RESPONDED' - Number of promotions responded to in the past year

'STYLES' - Total number of individual items purchased by the customer

'RESPONSERATE' - Promotion response rate for the past year - indicates which customers have ever responded to a marketing promotion before

'CLASSES' - Number of different product classes purchased

'CLUSTYPE' - Microvision lifestyle cluster type

The most important feature in the chosen Logistic Regression model was the 'Lifetime average time between visits'. The feature was a negative indicator, meaning as the average time between visits increased, the likelihood of the customer responding to the promotion decreased. The next six important features in the model had to do with the number of purchase visits, days the customer had been on file or spending amounts, including spending at individual stores. The four features after that were the customer's response to other promotions, the number of total products purchased, and the number of product classes the customer had purchased.

**6 Evaluation/Recommendations**

Now that we have a better understanding of customer behavior, and we have a model to predict who is most likely to respond to promotions, we recommend the following.

1. Threads should use the chosen Logistic Regression model to target the customers who are most likely to respond. This will help Threads engage with their customers in the most efficient and effective way.
2. Focus on the behaviors we identified in the exploratory analysis as characteristic of responsive customers, and also the most important features within the model and use those to target the “next level” of customers.
   1. From the exploratory analysis: the frequency of visits, the range of products bought, and Internet activity.
   2. From the model, the lifetime average time between visits, number of purchase visits, number of days between purchases, spending at the CC store, and spending at the PS store.

**7 Next Steps/Improvements**

1. Improving the model. It would be interesting to focus on some of the important features we identified, and train the model on those features only. It may help remove some of the “noise”.
2. Providing Threads a better understanding of who their customers are will allow them to focus marketing promotions on different "types" of customers, and allow them to reach out to potential new customers. Customers benefit in that they receive more curated content based on their preferences.
3. A survey of customer motivations and preferences would be very helpful.