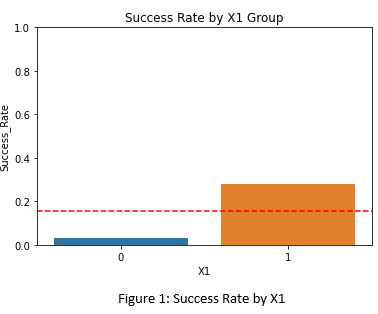
**Spotify Take Home Assignment – Doc Report**

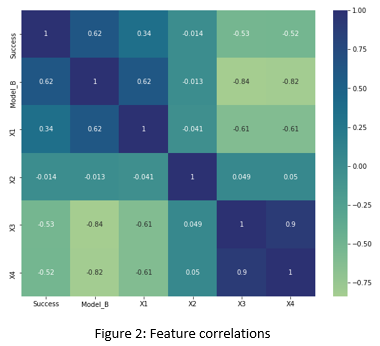
**Background**

This analysis is focused on examining email campaigns and predicting whether or not they will be successful. We have an existing model (“Model B”) that predicts whether or not each email campaign will be a success. However, we have four new features (X1, X2, X3, X4) and want to determine if these features can help us improve upon Model B. Therefore, a new model (referred to as “Model A”) will be built in order to test whether or not the features are valuable.

**Data Cleaning**

 In order to analyze the data, it was important to ensure that the data was clean. After inspection, it was observed that the data had no null values, no “out of place” values (such as strings in a numerical field), and no significant outliers. X1 was a Boolean field with the values distributed evenly between “True” and “False”. X2, X3, and X4 appeared to be normally distributed with a mean centered around 0. It was also noted that roughly 15% of the Success values were positive, with the remainder being negative. Although this represents a class imbalance, the imbalance is insignificant enough that it is likely not necessary to take special steps (such as under or oversampling) to address the imbalance.

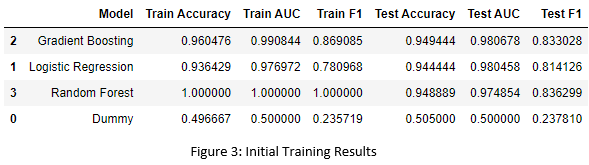
**Exploratory Data Analysis**

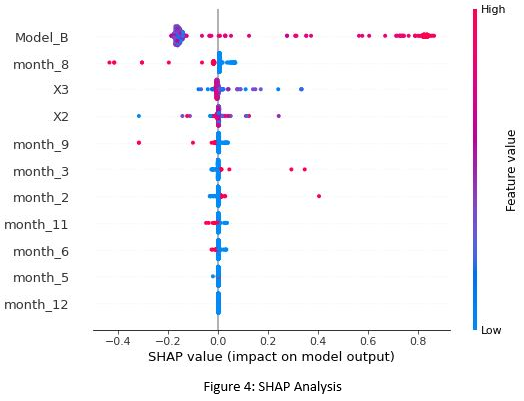
The next step was to look at the relationship between each of the new features and the target variable (Success). There tended to be higher success rates when X1 was true (Figure 1). X2 did not appear to be a good predictor of success rate. However, the variable was kept because we cannot be certain that it will not add value during the modeling stage. Finally, X3 and X4 had a, more or less, identical relationship to the success rate. Lower X3/X4 values were associated with a higher success rate. Therefore, X1, X3, and X4 were likely to be useful predictors of email campaign success. However, X3 and X4 appeared to be highly correlated. In order to further inspect the relationships a correlation heatmap was created (Figure 2). This provided more evidence that X1, X3, and X4 were likely to be good predictors of success. It also supported the observation that X3 and X4 were highly correlated. Additionally, it was observed that X1 had a moderately strong (negative) correlation with both X3 and X4. After accounting for multicollinearity, it is likely that we will only need to use one of X1, X3, and X4 (as the features may be redundant).

We also performed feature engineering using the date field. Since only month and year were captured, it was not possible to inspect the data at a finer grained level than the month. However, month of the year, quarter, and number of days between the email date and the most recent date were calculated. It appeared that there was some seasonality, as the 1st quarter (January – March) had the highest success rate and quarter 3 (July – September) tended to have low success rates.

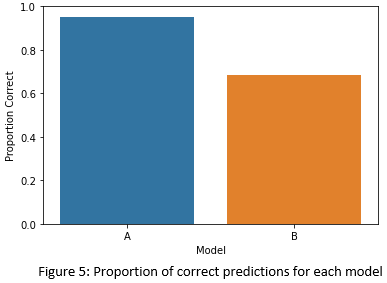
**Preprocessing and Modeling**

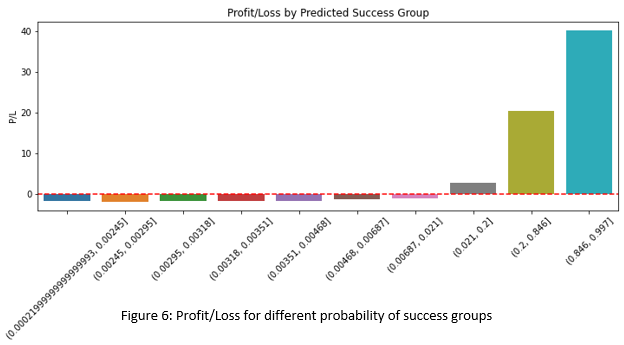
In order to preprocess the data, the feature matrix (X) and target vector (y) were separated, a 70/30 train/test split was performed, and categorical features were encoded using One Hot Encoding. Additionally, multicollinear features were removed until no features with a variance inflation factor greater than 2 remained. As expected, X1 and X4 were dropped. The final set of features contained X2, X3, and indicators for 8 different months of the year. We also ensured that the “Model\_B” column was not lost, as we explicitly want to use the Model\_B output as input to Model\_A.

 After preprocessing, we tested three different models (in addition to a baseline model). Logistic Regression, Random Forest, and Gradient Boosting models were selected due to their high interpretability. The results can be seen in figure 3. The Gradient Boosting model had the best performance and was, therefore, selected for hyperparameter tuning. The Tuned Gradient Boosted Model was the best model in terms of Accuracy, ROC-AUC score, and F1-score and was selected as Model A.

**Conclusions**

After building Model A, feature impact was examined using SHAP Analysis (Figure 4). It was noted that the Model B output was by far the most important feature. Higher probabilities output by Model B tended to be associated with higher probabilities output by Model A. Therefore, if Model B predicted success, it was more likely that Model A would predict success. For this reason, using Model B and Model A as an ensemble will likely be the best modeling strategy (we will see below that Model A is a significant improvement over Model B). Month 8 (August) was the second most important predictor. The month of August was associated with lower probability values, meaning that we are less likely to have successful email campaigns during the month of August. This agrees with the seasonal analysis we did during the exploratory data analysis stage. Feature X3 was the most important of the remaining features. Low values for X3 tended to have more impact on probability (in either the positive or negative direction). Higher X3 values tended to have less of an impact on probability. Much like X3, X2 tended to impact the model output but there was not much directionality associated with high or low X2 values. Finally, we are more likely to see successful campaigns in February and March. Conversely, June, September, and November tended to be associated with less successful campaigns.

After completion of the modeling stage, Model A and Model B were compared in order to determine if Model A was able to improve upon Model B. Using a decision boundary of 0.5, the predictions made by each model were compared. Model B made correct predictions for roughly 68% of email campaigns. However, Model A made correct predictions for roughly 95% of email campaigns (figure 5). Therefore, Model A definitely showed improvement over Model B.

 Since Model A was known to be an improvement over Model B, we used Model A to assess value and to develop a strategy for email campaigns. We sorted all of the test cases by predicted probability of success in descending order. We then divided the test cases into 10 groups of roughly equal size (deciles). Using the number of actual (not predicted) successes per group, the fact that a single email campaign costs 1 cent, and the fact that a successful campaign generates 25 cents in value we determined the profit/loss if we were to market to each of the 10 groups. It was found that a profit can be generated if we move forward with email campaigns where Model A’s predicted probability of success is 0.21 or higher (Figure 6). Although this seems like a low probability threshold, we need to keep in mind that the revenue generated by a successful email campaign is 25 times the cost of the campaign. For this reason, it is still worth targeting groups with a seemingly low predicted probability of success. The 0.21 threshold corresponds to targeting roughly the top 30% of email campaigns in terms of Model A’s predicted probability of success. Based on Model A’s predictions, this top 30% can be expected to generate a profit.