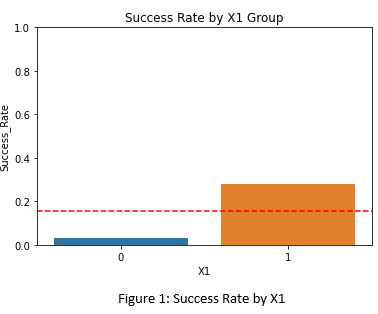
**Spotify Take Home Assignment – Doc Report**

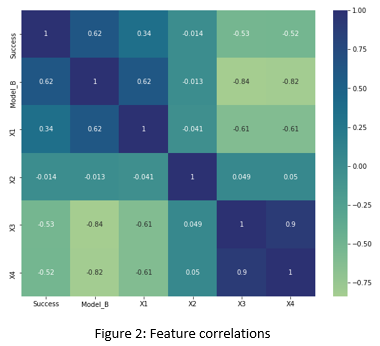
**Background**

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).

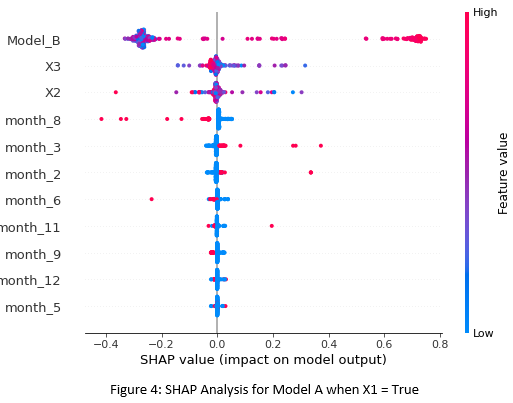
We also performed feature engineering using the date field. Since only month and year were captured, it was not possible to inspect the data at a finer grained level than the month. However, month of the year, quarter, and number of days between the email date and the most recent date were calculated. 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.

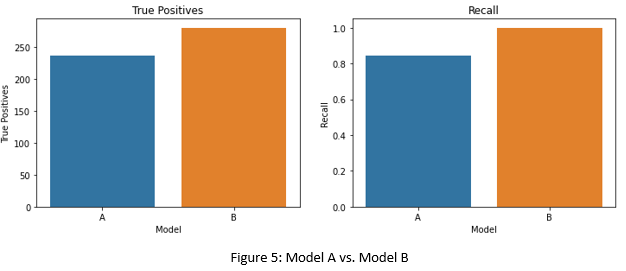
**Preprocessing and Modeling**

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.

Two different modeling scenarios were tested. First models were trained using the entire training dataset. Next, because X1 was dropped but appeared to be a good predictor of success, we tried creating separate models for email campaigns when X1 was true vs. when X1 was false. For each of these modeling scenarios, four different models were tested: Logistic Regression, Random Forest, Gradient Boosting, and XG Boost models were selected due to their high interpretability. For each scenario, the Random Forest model had the best performance and was selected as the final model. Recall was used to evaluate the model because the value of a successful campaign is 25x the value of an unsuccessful campaign. However, F1-score was also considered because we wanted to ensure that the model was reasonably balanced and was not “over predicting” successful email campaigns. It was found that the best scenario for Model A was to use separate models depending on whether X1 was true or false.

**Conclusions**

After building Model A, feature impact was examined using SHAP Analysis when X1 is true (Figure 4) and when X1 is false. When X1 was true, 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. This is why if Model A was selected, an ensemble model where Model B feeds into Model A would likely be the best strategy. Feature X3 was the most important of the remaining features. Lower 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. Month 8 (August) was the fourth most important predictor. The month of August was associated with lower probability values, meaning that we are less likely to have successful email campaigns in August. Finally, we are more likely to see successful campaigns in February and March. These findings agree with the seasonal analysis that was performed during exploratory data analysis. Model B was also the most impactful feature when X1 was false and, once again, higher Model B probabilities had a positive impact on Model A’s output probability. The month of February also showed some importance when X1 was false. In February, email campaigns tended to be more successful.

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 (Figure 5). The number of successful campaigns predicted by each model were examined. Model B identified more true positives and also had a better recall score. Recall is relevant because the value generated by a successful campaign is 25x the cost of an unsuccessful campaign. Therefore, Model B appeared to be the better model.

In order to assess value and develop an email campaign strategy for both Models A and B, the test cases were sorted by predicted probability in descending order. The test cases were then divided 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. For both models it was optimal to market to the top 30% of email campaigns in terms of predicted probability. It was determined that $63.60 would be generated using either Model A or Model B if emails were sent to the top 30% of predicted probabilities. Therefore, once again, we can conclude that Model A is not offering any improvement over Model B. Model B should continue to be used because it is a less complicated model and may be more robust in terms of generalizing to unseen data. When working with Model B, we should move forward with all email campaigns with a predicted probability greater than 0.618. These are the email campaigns that tend to produce a profit (Figure 6).

