Ayan Karim

Final TakeHome Model

**Description of Data:**

This data set contains demographic information on customers and labels whether they’ve joined the loyalty program. More specifically, it includes a mixture of categorical (binary), continuous and discrete variables on gender, age, whether there’s a card on file, the purchase amount and the time elapsed since last purchase for each customer.

The original data set contained 12,000 rows of data and 7 columns, however after cleaning the data and under sampling to balance our classes, we have 9,741 rows of data and 6 columns that we actually use for modeling.

**Approach:**

I treated this investigation as a classification project where I would model the data, based on features, to predict whether a customer will join the loyalty program or not. After exploring the data set it became obvious that there are many unrealistic values for the continuous variables. For example, they all had variables less than 0, which is ridiculous for purchase amount, age or days since last purchase. Furthermore, any age less than 18 seems a bit weird for joining a loyalty program. So, I decided to filter out all values less than the first quantile (25%) for each of the continuous variables.

There was strong right skew for each of the continuous variables, as well as a significant class imbalance for ‘loyalty’ labels. This is bad for training our model, so to clean the data further, I created reduced data frame that under-samples the “False” labels for “loyalty”. Then, I normalized the continuous variables using a Box Cox transformation.

For training, I split the under sampled data-frame into training and test sets and ran multiple baseline classification models to test their accuracy with a for-loop. Using these accuracies, I decided that gradient boosted classifier and random forest classifier perform the best. I ranked my features based on significance to our models and removed insignificant features. At the very least it made the models more efficient. I also compounded the “purch\_amt” variable because it was the most significant to our model. This makes sense because people who buy more would probably want a loyalty membership. This improved by the model a little bit in cross validation scores.

Finally, I evaluated my models further by generating a precision, f1-score and recall table. From this table the Gradient Boosted Classifier seemed the most consisted.