Ayan Karim

Final TakeHome Model

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.

The first step was to explore the dataset. After running a couple descriptive statistics, 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. Then I removed all the null values.

Visualizing the data showed a strong right skew for each of the continuous variables, as well as a significant class imbalance for the ‘loyalty’ column, which indicated whether the customer joined the loyalty program. 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. I visualized the correlation of our continuous variables and their distributions to make sure they were independent, and normal.

Finally, I split the under sampled data-frame into training and test sets. I ran baseline classification models and tested their accuracy on the hold-out group through a for-loop, so that I can quickly compare the baseline performance of multiple classification models. Using these accuracies, I decided that gradient boosted classifier and random forest classifier perform the best. So, I continue with these.

I ranked my features based on significance to our models and removed insignificant features. The model performance didn’t really improve after this, but was no worse off, so 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.

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.

The accuracy of the models isn’t great, but they are good for baseline modeling. More features and data would help improve this model