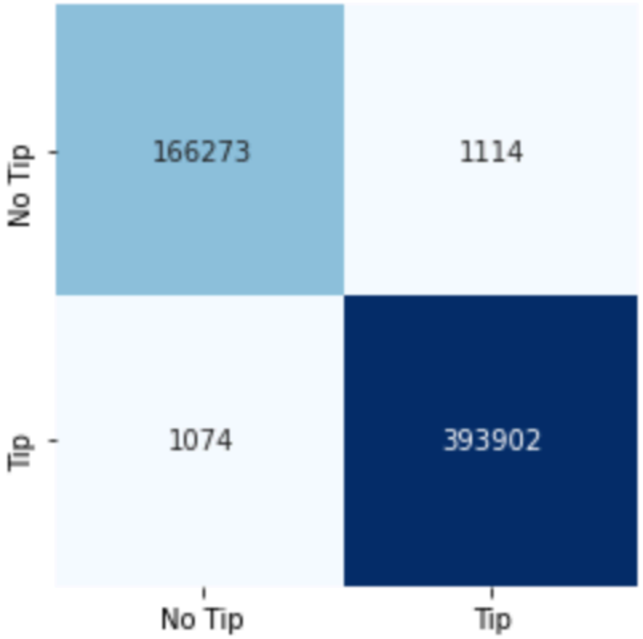
Classification MVP

With a boom in personal driving vehicles over the past couple of years, the competition for drivers in NYC has never been higher. Couple this with the astronomical amount a medallion costs, and a driver needs every possible advantage in their pocket to maximize their earnings. It is my belief that a predictive model can be used to gauge if a ride will result in a tip or not. To that end, I hope to focus on precision. I hope to build a model that minimizes the number of false positives, allowing riders to ensure they can earn more per ride.

The Data was larger than originally anticipated. For just yellow taxi cabs alone, there are 83 million observations for the 2019 calendar year. While I am attempting to learn to use Google Cloud Services, I am working on a smaller model in the interest of time. Initially models were not performing well. After diving deeper into the data, and confirming with the data source online, I realized that cash payments never record tip amounts. After removing all cash sales from the dataset, we are closer to a final model.

Below is a confusion matrix from a baseline logistic regression for only yellow cabs from January 2019, with predicted values along the y axis, and axis. The model was fit on a training set, the predictions for this matrix were done on the test set.

These results already have precision quite high, with a log loss of .01 but there is more to be done. In addition to compiling more data on Google Cloud, I have additional features such as time of day, and day of week I plan to introduce. I will also be combining the data frames I have collected for yellow cabs and for-hire-vehicles. I will also use different ensembling methods to gather feature importance, as that is what will be most useful to the drivers.


No Tip
Tip
No Tip
Tip