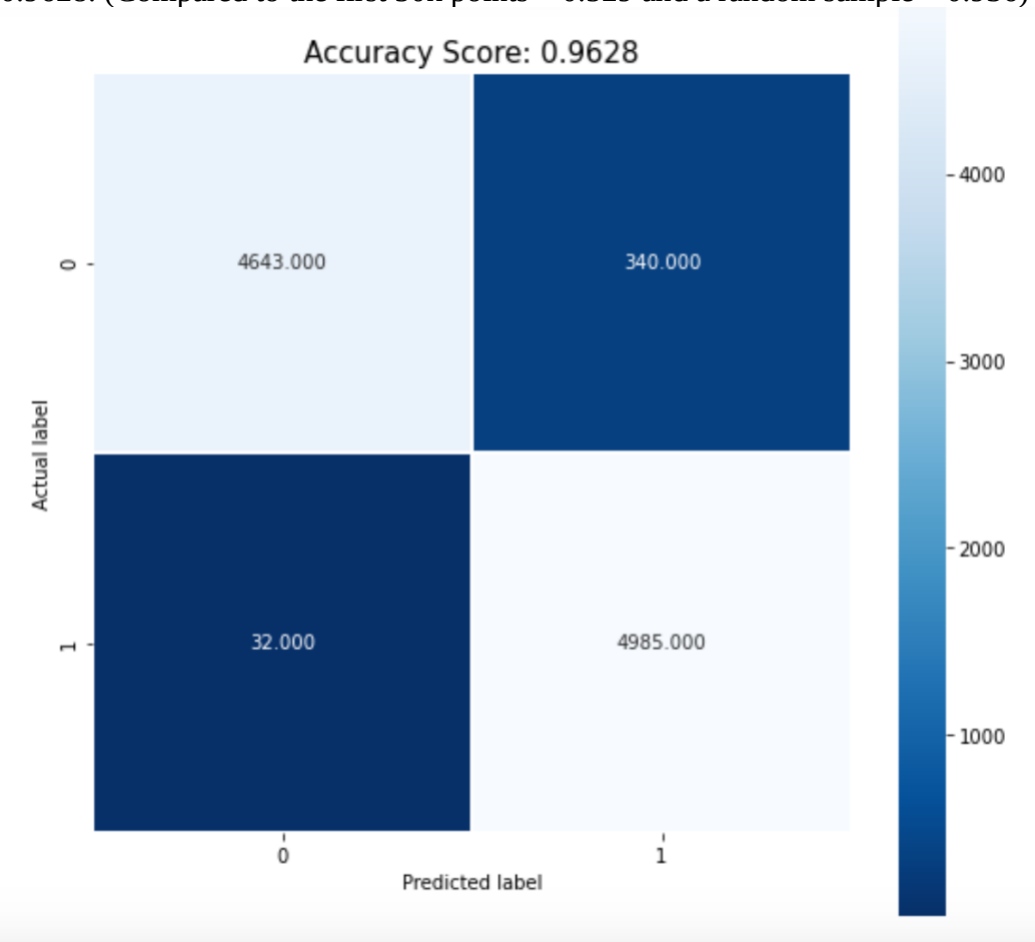
Classifying How NYC Taxi Rides Changed Between 2017 and 2019 Due to Uber

The goal of this project is to build a model that can use features such as trip distance, pickup location, tip amount, passenger count, etc. to classify a taxi ride as having occurred in 2019 vs. 2017. These two years were chosen because 2017 was when Uber caught up to taxi cabs in terms of popularity, and 2019 is when Uber has far surpassed cabs.

My initial baseline model, a logistic regression, was used on the last 50,000 points on the dataset of approximately 1 million data points that was obtained from the website. It had a score of 0.9628. (Compared to the first 50k points – 0.529 and a random sample – 0.936)



For the rest of my project, I need to determine which models would best fit my data, score them on classification metrics, and create visualizations to communicate the best model.