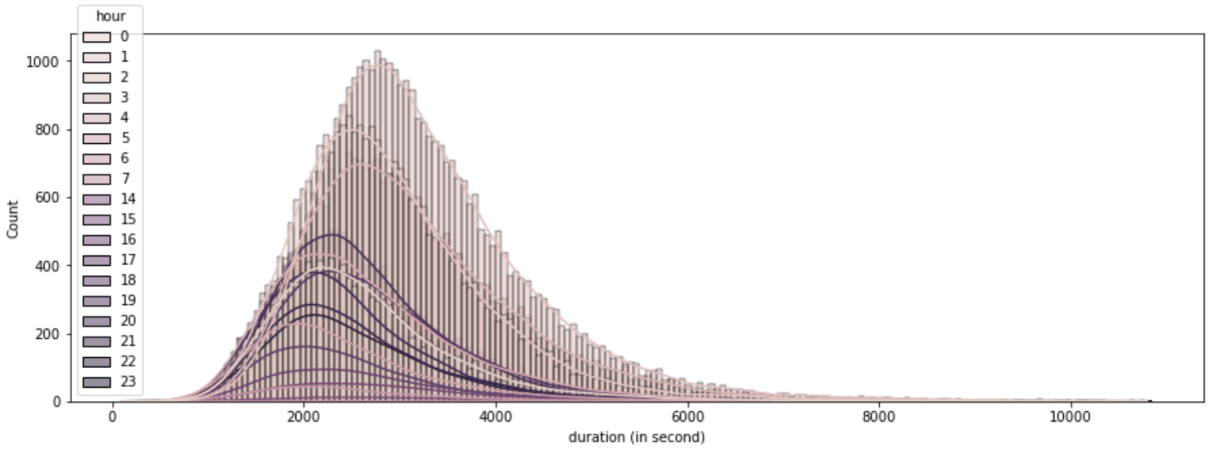
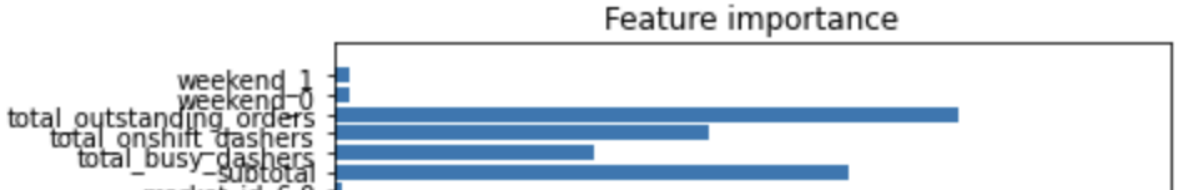
DoorDash challenge
Applicant: Yu-Ting Shen

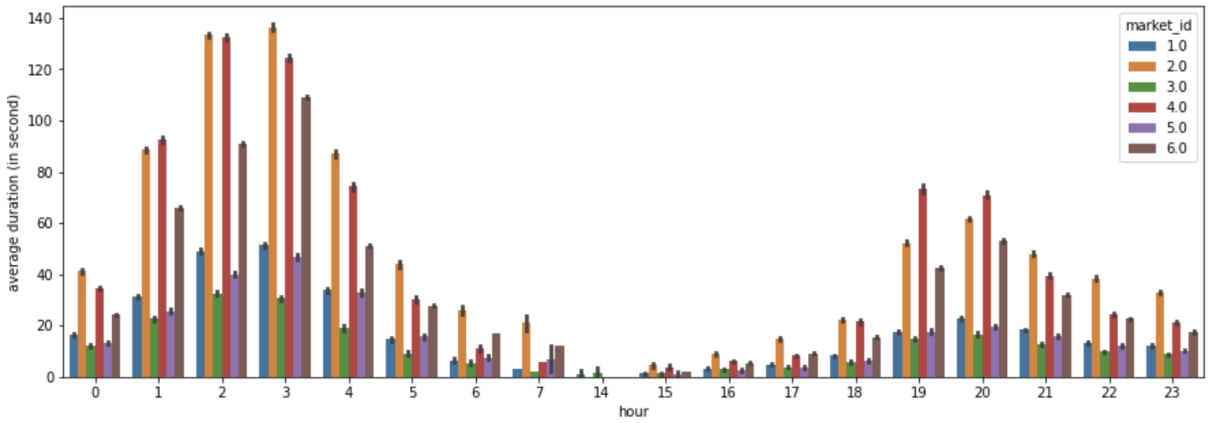
Summary:

From the EDA we know the order time is an important feature which affects the delivery duration. The orders in different hours have significantly different delivery duration.



From 1 am to 4 am is the period which has most onshift dashers, however, more than 80% of the dashers are busy dashers resulting in a lot of outstanding orders. Given limited dashers, if there are more outstanding orders, then the new orders have a higher possibility to be late. The feature importance plot also demonstrates the total_outstanding_orders is an important feature The following plots show the feature importance and the total_outstanding_orders with respect to hours in different markets.





Improvement:

I think if we can know the following information, then the model can be improved and could provide more accurate predictions.

- The latitude and longitude of stores and customers.
  If we could cut the map into several smaller pieces, then we could use all the dashers' data to estimate the traffic in each smaller piece. Given the lat-lon, we could find the corresponding piece on the map and use the traffic into the consideration.
- How many orders or the maximum orders a dasher is handling at a time.
  A dasher could deliver several orders at a time, it is equivalent to a batch job for a dasher. If a dasher delivers more orders at a time, then the outstanding orders will be picked up and delivered sooner.
- Type of the foods.
  If there are more customers ordering the same type of food, then a dasher could pick up several orders at the same time and deliver these in one run. For example, if a pizzeria gets 10 orders about the same time, then these 10 pieces of pizza will be ready about the same time. A dasher can pick up and deliver all of them

How to access new model’s performance before replacing the previous model:

Because DoorDash collects a lot of historical data, we can use the historical data from several marketplaces and different time windows as inputs. Then we use the new model and production model to predict the results using these inputs. If the new model’s results are all better than the production model’s results, then we can consider replacing the production model.