Phase 1

Two other business metrics:

1. participant_acceptance_rate: number of accepted booking allocation by participant divided by number of created booking allocation
2. cancel_rate_after_dispatch: percentage of booking cancelled (either by customer or driver) after dispatching a driver

Findings:

Experiment B performs better than experiment A in terms of **higher booking conversion rate**, **higher driver acceptance rate** as well as **lower cancel rate after dispatch**. Experiment A outperforms experiment B by **shorter mean pickup time**.

Experiment B allocates booking to driver who are more likely to accept, less likely to cancel and more likely to complete.

Experiment A allows customers to wait for shorter time to be picked up.

Normally high driver acceptance rate is at expense of low pickup time. Though longer customers wait, more likely they will cancel the booking, the cancel rate after dispatch is still lower for experiment B. Since the current business objective is to increase the number of completed orders, booking conversion rate is weighted more than mean pickup time.

Hence, I would recommend experiment B over A to business decision maker.

Phase 2

Business problem formulated:

Allocate booking made by customer to one of the 10 candidate driver to ensure high conversion rate.

ML problem formulated:

Predict probability of each candidate driver completing the booking, allocate to the one with highest probability.

Instead of having absolute 0 or 1 as independent variable, I chose to find the probability of each candidate driver having 0 or 1. This is because each booking has 10 candidate drivers, in the case of 0 or more than 1 driver are predicted to complete the trip, one single model would not be sufficient or straight forward to tell us which one to allocate.

Hence, I chose to predict the probability of each class of every candidate driver, and allocate the booking to the one with highest probability of completing the trip.

Approach:

Due to the nature of the model, logistic regression is used here to solve model the problem.

Besides ML model, an alternative non-ML model is also tried as a benchmark to compare the performance of ML solution. The rule based model uses single feature to make prediction, it is chosen based on histogram analysis, a higher density of failed trip when chosen feature exceeds a certain threshold.

ML solution does outperform non-ML model in terms of accuracy.

Insights:

Feature engineering places important role in model improvement. Simple aggregated features for drivers and geolocation have large impact to models.

Parameters *total_failed* and *total_completed* are observed to have highest coefficient magnitude, followed by *trip_distance*. Model performance increases a lot by adding *total_failed* and *total_completed* at later stage. *total_failed* has negative coefficient, indicating the more driver failed to complete trip after allocation in the past, the more likely he will fail to complete trip this time as well. *total_completed* has positive coefficient. We can derive that whether a driver is able to complete the trip or not is highly dependent on his trip history and user behaviors.

Another feature *is_peak* which indicates whether the trip is created at peak hour defined by GOJEK official website. It's interesting to find out it has relatively high positive coefficient, indicating a driver is more likely to complete the trip if the booking is made during peak hour.

Proposal for future improvement:

The training dataset is a bit imbalanced, where the ratio of 0 to 1 is 1/3. The model generated tends to have high false positive rate. Though *class_weight* parameter is tuned during training to assign more weights to 0 class, the model accuracy is still not improved. So other methods such oversampling and cost-sensitive learning could be explored in future, and probably using different scoring metric.

Also more features can be created based on customer, driver and geolocation data for modelling, such as tag on pickup location to indicate whether it's CBD or commercial or residential area.