questions_project

Please write down your questions/anything you do not understand about the dataset. If you know the answers to other people's questions, feel free to share! Ideally, we can finish this list by Friday.

Action:

What is the wave/wave_index?

A wave is a series of action performed by a courier, including a series of delivery and pickup. In a wave, a courier would have a bundle of orders (tracking_id). He would need to decide which item to pick-up first, whether to keep picking up items or to deliver some item first.

What is courier_wave_start_lng/lat?

It is the longitude and latitude of the courier when he begins the wave. It should be a matter of consideration given the bundle of actions to perform. Specifically, which item to pickup first, and what action should be taken right afterwards. Supposedly, in one wave, the longitude and latitude should be unique.

Why can two actions PICKUP and DELIVERY happen at the same geographic location? Or is there some kind of delay in the update of geographic location because according to expect_time, the time elapsed between PICKUP and DELIVERY should be about 10 minutes if we take the unit of expect_time as seconds (by the way, is it seconds?).

Pick-up and delivery locations are stored in the "distance" dataframe, and you would need to access them using courier information, wave information and tracking id information. The courier_wave_start_lng/lat is not a priority in my perspective, but you are also welcome to prove me wrong.

Regarding the time, it is using Unix timestamp. You can try and find ways to turn it into datetime object, or you can simply use them, it does not make a difference. The numerical difference corresponds to seconds, for example 1587159105 and 1587159106 is one second apart.

What does DELIVERY mean? Does it mean the courier has delivered the food to the customer or does it mean the courier starts to deliver?

Both DELIVERY and PICKUP are supposed to mean the timestamp of the "finishing" of that action.

Distance:

What is the source and what is the target? I do not understand why two food deliveries are switching between being the source and the target.

Answer: This dataframe gives you the possible combination between actions.

For example, we have order A and B, and there are three actions: Assign, Pick-Up and Deliver.

First of all, the type of action determines what point we are using. It doesn't matter if it is the source or the target. Assign takes the location of the courier, pick-up takes the location of the shop, and deliver takes the location of the customer.

Then, this dataframe gives you all the combinations you could perform in a certain wave, and their corresponding distance. For example,

- one possible action sequence is pick-up A and then pick up B. Then the data will be
    - (Current Wave ID, A's tracking ID, PickUp, Restaurant A's Lng, Restaurant A's Lat, B's tracking ID, PickUp, Restaurant B's Lng, Restaurant B's Lat, distance between Restaurant A and B.
- Another possible sequence can be pick-up A and then delivery A. Then, the data will be
    - (Current Wave ID, A's tracking ID, PickUp, Restaurant A's Lng, Restaurant A's Lat, A's tracking ID, Delivery, Customer A's Lng, Customer A's Lat, distance between Restaurant A and Customer A.

Chaining these actions would give a "wave" of actions performed by the courier. There are many combinations, but one rule should apply: an order must be assigned before picked up, and must be picked up before delivered. In Distance data set, it gives the complete combination of all possible pairs.

grid_distance is the distance between the source and the target?

Answer: Yes.

I think the unit for grid_distance is meter because, for the same vendor, it shows the grid_distance is 1.

Answer: I believe that distance is measured by meter. The exact unit does not matter that much as you can always perform standardization.

Why is the grid distance between two places different from two directions?

Answer: About the grid distance: this is a distance provided by maps. It means that it is the shortest distance that a courier would travel to a location. It guarantees that the courier would follow accessible roads, instead of "flying" over buildings. A road that leads to one place may not be accessible from the other direction, my assumption is that they have taken into consideration traffic rules and controls.(One-way roads, for example)

Order:

What is estimate_pick_time? How is it estimated?

Answer: Estimate_pick_time is generated by the platform. It is the time that the platform tells the restaurant.
That is up to you to find out. If you need a hint, I recommend using some information to try and find out. I would consider the current weather, the courier info (speed, capacity), the order amount, the distance. I do not think there is any part where you need to predict estimate_pick_time yourself, but definitely it will help with your presentation if you can do some Explanatory Data Analytics in your presentation.

On the other hand, estimate_pick_time should be useful for the prediction of actual pick up time.

Why does the same vendor with the same shop_id have different aoi_id's? Is the same vendor not in the same area? Perhaps because one vendor can have different restaurants all over the city? But the geographic location shows that the same vendor is in the same place.

Answer: Allow me to clarify: aoi is the unique identifier for the delivery point, not the vendor.

Pick_lng and pick_lat are the same as the geographic location in Distance when the type is PICKUP.

Deliver_lng and deliver_lat are the same as the geographic location in Distance when the type is DELIVERY.

Orders are confirmed almost exactly when they are created.

What are we really being asked? "The goal is to build a model to predict the delivery man's next move, based on his historical decision and current status."

Answer: This is a difficult question, and my expectation is: do what you want with the data. This is a practice where you would need to utilize many skills, such as exploring the distribution/patterns in the data, probably plotting something, making assumptions, cleaning up the data, and after all that, you may be able to construct a model. The exploratory data analytics part is just as important as building a model. And because this is a real-life dataset, we want you to practice using data to make observations as well as recommendations. In short, imagine you are part of a Data Analytics team for a consulting firm, or a city planning organization, or the delivery company itself: what will you do with the data. I think starting from a certain perspective, or taking up a certain "role", would help you construct your presentation. It is okay to be creative.

If you are asking what you are asked for the Kaggle submission, you are required to "fill in the blanks" of the "action_20200229.csv", as I've masked some of the timestamps. And then, keep only the expect_time column, and add a unique ID column in front, in the form of "1,2,3, etc.". Submit only this file.

There are some compromises I must make in doing this. Originally, you should construct from scratch: what action to take at each point, along with their expected time. However, the Kaggle autograder does not allow me to customize my own grading metric like that, so I used only the expected time column. This actually makes it easier for you to check your decision, as the "real" sequence is already there in the file. But make sure you do so yourself, and present your thoughts in your presentation. This is important, and don't be afraid if you have different sequences, this happens from time to time. The important part is presenting how you did this.

To what degree of accuracy is the professor expecting for MAE?

Answer: You would need to figure that out yourself by running a baseline model first before you do anything with the data.

Isn't MAE Mean Absolute Error? Why then is the formula missing 1/n

Answer: Yes, it is. But here the dataset has a known "n", so it doesn't matter. You can add it in your optimization if you want, it is okay. But I assume you will be using existing metrics in packages, so you would not encounter any problem with this.

Where is the example submission? I cannot find it on Kaggle

Answer: I will upload it to NYUClasses shortly. I wasn't aware that you cannot access it.

####