OCADO TECHNOLOGY INTERNSHIP TASK

Stack used : Python, Jupyter Notebook, SQL

INFO : All the analysis are already made in previous PDF. In this PDF I focus only on elaborating on asked questions.

Python code: LINK

Name: Wiktor Łach

Location : Kraków

1. The current prediction algorithm is very naive. It calculates the mean from all collected data and applies it to every future order. We need to explore alternative ideas. One of them is predicting delivery times per sector. Describe how you would validate this hypothesis using available data.

As we could see in the analysis part segment played an important role in delivery times. To validate this hypothesis first I would make predictions based on the sector. Then I would use graphs to visualize the predictions and actual delivery times. I would also calculate the average absolute error for predictions, to see if it is a better algorithm I would compare it to the previous one (by comparing absolute error).

2. Using the data, propose some alternative method/algorithm that will predict delivery times more accurately. Describe the methodology to validate the new algorithm.

Basing on the fact that there were some noticeable correlation between some data types and delivery duration I would use the information about: number of products in order, sector, day of the week and time of the day. For each of these values I would calculate the average. At the end I would calculate the mean of those values and it would be my final prediction. I could also use the information about which driver will deliver the order, but I would have to get an insight info why the differences between drivers' time are so significant. The model that I created would be simple, but still way better than the one that has been used before.

If I could use more advanced algorithms including machine learning I would go for the decision tree and I would use the same data as in the model above. Using this model for sure I would get better results, but the whole algorithm would be way

more complicated, but results would be still explainable. If company didn't care about the explainability of model and the precision in predicted delivery time was the most important factor I would go for random forest algorithm, so I could avoid overfitting and get more precise results.

3. Why could some deliveries take more time? For example, some buildings don't have elevators etc. Describe your ideas.

- Number of ordered products – the delivery man might not be able to take every product at once, because of its weight or volume, so they would have to take multiple walks while delivering.
- Traffic jams – some areas of the city in certain time might have more traffic jams than others. Obviously the longer you stay in traffic the higher delivery time is.
- Parking problems in certain areas
- The delivery man might be tired at the end of his shift, and it might caused the spike in the last our of shift (as shown in one of the plots in previous PDF)
- Different types of delivery vehicles. Some of delivery man might use for example motor bikes that might be faster than cars in certain case.
- Waiting for the customer to collect the order.
- Wrong address given by customer.

4. What additional data would be worth collecting for future analysis of this domain?

For sure the distance between grocery and destination place is must have. Also getting information about the time for every segment of delivery would be very useful (for example packing products, driving a car, delivering from car to customer etc.) Talking with employees (especially delivery man) could provide some insights that would give some ideas what are the most important aspects to focus on.

5. What is the risk of over- or under-estimating the delivery times?

For over-estimating delivery times the biggest risk is that delivery-man would have too long breaks between orders and it would cause the waste of company money. Too many employees would be needed for delivering.

For under-estimating delivery times the risk is that delivery man won’t be able to deliver orders in time. It might cause the angriness in customers so as the result they might stop using the grocery service.