Part 3.

Building and verifying the hypothesis

Author:

Filip Bulanda

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.

I am going to assume, that we are using the same parameter, which is an average delivery time. Therefore, I would start with calculating the average delivery time for each sector. Before the actual calculations, it would be necessary to clean the data from erroneous data, and perhaps outliers (since average is sensitive to them). After obtaining the averages I would calculate the prediction times for the deliveries and compare the predictions with real delivery times. For validation, I would probably calculate the values of errors like MAE (mean absolute error – tell us the average error between prediction and predicted value), RMSE (root mean square error – penalises larger errors more than MAE). All of that is possible given our current data and is very easy to get with a short Python script (or any other tool for statistical analysis like R or even Excel).

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

Based on the first report about data analysis done on the same data, I managed to find two features that added up to the delivery time the most, which were driverID and sectorID. I think, that a relatively simple model, that could give us decent result would be based on using both average delivery time for sectors and drivers. It would then take weighted average from those to values and return the result as our prediction. Since delivery times per driver seem to be more varied than delivery times per sector, average of drivers would receive higher weight.

To validate this method, the same methodology as in pt.1 could be followed. Calculate required values, calculate predictions and measure errors. Only significant difference would be the need to fine-tune the weights for the weighted average, checking whether or not have the errors improved or worsened after each weight change.

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

Since the delivery times varied significantly depending on sectors, it could be due to the difference in buildings in sector and their environment. That could mean lack of elevators, lack of parking adjacent to the building, or higher building (reaching random floor of skyscraper would take significantly longer than reaching random floor of small flat). If one sector is of a different specialisation (ex. industrial or commercial zone) then it also could

impact time for finishing the delivery. Additionally if a sector covers richer areas, the deliveries might be heavier (although this did not seem to be the case for this provided data) or contain gated communities.

When it comes to drivers (which was also observed to be one of the reasons for differences in delivery times) their differences in performance could come from multiple reasons. First of them could be company policy. If company implements some kind of gratification system for finishing deliveries faster, it could make some drivers do all they can to maximise their efficiency, while at the same time some could be completely unbothered. It could technically be also impacted by drivers personality. More melancholic drivers could be walking slower (some people tend to climb two stairs at a time, while others climb one).

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

When it comes to additional data, it could be worth to collect information correlated to arguments provided in pt.3. That would be data like:

- Customers' building information (floor of the apartment, accessibility to elevator, possible access restrictions)
- Delivery location data (distance from closest parking, parking accessibility)

If the delivery time also takes into account the time it takes to driver from local grocery store to customers house, then it would be helpful to gather additional data, like:

- Spatial data (traffic patterns, distance from store to customer)

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

I think, that both under and over estimating could pose some risks.

If our model was to perform a significant over-estimation, the customers could feel discouraged to choose our delivery, and might look for someone who could provide it faster, or just go shopping themselves. Depending on how the company operates and assigns schedules, over-prediction could lead to scheduling errors, leaving drivers with lots of idle time.

On the other hand under-prediction could pose significant risks. Customers could use our service because of short delivery times but later be annoyed that their expectations were not met. Such situations could lead to losing significant customer base, and prevent company from gaining new loyal customers. If the under-prediction was significant, drivers could feel constantly rushed, which would be not only mentally unhealthy but also dangerous if they are driving in urban areas. As with over-prediction it could lead to errors in driver schedules, depending on company schedule handling.