Delivery Time Prediction – Building and Verifying the Hypothesis

Natalia Klinik

1. **Introduction.**

Currently, delivery time is predicted as an average of all data. This approach is very naive, as it does not take into account local differences. It results in *planned_delivery_durtion* oscillating around 3 minutes for basically all orders.

One of the hypotheses is to predict the delivery time by sector. In order to validate, the following steps should be taken:

- Divide the data into sectors (grouping by *sector_id*);
- For each sector, calculate:
    - average actual delivery time,
    - average prediction error
- Compare these values with the overall average to see if the sector averages better predict actual times.

If the prediction error decreases after introducing the sector predictions, it means that the hypothesis is accurate and worth implementing in the system. If it does not improve the results, we could also consider checking the delivery not only by sector, but also by driver, as it intuitionally influences the delivery time.

2. Proposal for an alternative prediction algorithm.

Instead of a single average, it would be reasonable to create a simple prediction model based on several features, such as:

- sector of delivery,
- driver,
- total order weight,
- number of products,
- day of the week/holidays - weekend and holiday deliveries may be different),
- time of day (e.g., peak hours).

It is important to remember that categorical features (such as *sector_id*, *driver_id*, *day_of_the_week*) need to be encoded into numerical ones in order to be used by the model.

When it comes to holidays, adding a simple yes/no encoding should work. Of course, this also needs to be mapped to numeric values.

If we do not want to extract the exact hour from the date, again a simple yes/no column indicating whether the delivery was made during peak hours should be sufficient.

The algorithm can take the form of:

- **Linear regression** – if we need a method that is fast, easy to interpret.
- **Decision tree** – can be more accurate, as handles non-linear relationships well.
- Simple **Random Forest** model – to handle more complex data.

Model construction and validation:

- Divide the data into a training set and a test set.
- Build the model on the training data and test its effectiveness on the test set, for example by using root mean squared error (RMSE) or mean absolute error (MAE).
- Compare the results with the current approach (constant mean) to see whether the prediction is more accurate.

3. What influences the longer delivery time?

Despite the lack of additional data, potential factors can be identified:

- **Lack of an elevator in the block** - increases the time it takes to deliver purchases, as the driver need to climb several flights of stairs.
- **Order weight** – our data actually shows that heavier orders take longer to be delivered (as analyzed in the *part2.ipynb* file); it does not however mean that smaller orders always take less time than the heavier ones.
- **Zones with heavy traffic** – traffic is probably one of the main features that can cause delivery delays, so distinguishing traffic zones could help the analysis.
- **Zones with limited parking spots** – delivery trucks usually require a bigger parking spot than the regular ones, so if the zone is lacking them, delivery might take longer, as the driver needs to find the spot/walk some distance on foot.
- **Zones with poor signage** – it will take more time to deliver if the driver cannot find the building.
- **Peak hours** – even in zones with little traffic, during peak hours the delivery might take longer.
- **Weather conditions** – difficult/unusual conditions can affect driver speed.

4. What additional data is worth collecting?

To improve prediction accuracy, we should consider collecting:

- **Building type** (e.g. block / single-family house / office building);
- **Elevator information** (yes / no);
- **Floor number**, especially in buildings without elevators (it will give us inside on the number of floors to be covered on foot by the driver);
- **Weather data** (e.g. rain, snow, fog);
- **Parking information** (e.g. number of spots for delivery trucks);
- **Zone and peak hours in this specific zones**, as those might differ in different districts;
- **Average driver's speed** (e.g. based on GPS).

5. Risks of erroneous forecasts.

Underestimating delivery time can lead to overworked drivers, customers dissatisfied with delays, or errors in efficient route planning.

Overestimating delivery time can lead to inefficient use of vehicles and resource utilization (drivers may have 'empty windows' between orders).

It is therefore, it is important to create an accurate and adaptive model. It should be able to learn from new data and changing conditions.