Spis treści

1. Validating the Hypothesis: Sector-Based Predictions

To validate the idea that delivery time predictions would be more accurate when calculated per sector, I would:

- Group the delivery data by sector_id.
- Calculate the average actual delivery duration for each sector.
- Compare these averages with the global average used in the current model.
- For each order, calculate the prediction error:
    - once using the global average,
    - and once using the sector-based average.
- Evaluate the Mean Absolute Error for both methods.

If the sector-based model consistently yields a lower prediction error, it would confirm the hypothesis that delivery duration is sector-dependent and that the model should consider location-specific features.

2. Proposed Algorithm and Validation Methodology

Proposal:

Use a regression model (e.g., linear regression or decision tree) that takes multiple features into account, such as:

- Sector
- Planned delivery time (e.g., hour of day, weekday)
- Route length (if available)

- Courier history or experience (if available)

Validation:

1. Split the dataset into training (80%) and test (20%) sets.
2. Train the model on the training set.
3. Evaluate the model on the test set using metrics like MAE or RMSE.
4. Compare performance against the baseline models (global and sector-based averages).

3. Why Some Deliveries Take More Time

Potential factors:

- No elevator in the building – additional time required for stairs.
- Gated communities or restricted zones.
- Heavy traffic during rush hours.
- Limited parking space or long walking distance.
- Bad weather (rain, snow, etc.).
- Address issues (unclear signage, incorrect location).
- Inexperienced courier unfamiliar with the sector.

4. Additional Data Worth Collecting

To improve future delivery time prediction, the following data should be considered for collection:

- Building type (house vs. apartment)
- Presence of an elevator
- Floor number
- Parking availability/distance to entrance
- Day of the week and time of day
- Traffic conditions at delivery time
- Delivery distance in kilometers/meters
- Historical delivery delays or reschedules

Incorporating these variables would allow the model to better reflect real-world conditions.

5. Risks of Over- and Under-Estimating Delivery Times

Both overestimating and underestimating delivery durations can lead to operational issues:

Underestimation risks:

- Missed delivery windows
- Driver delays and cascading schedule problems
- Lower customer satisfaction

Overestimation risks:

- Idle driver time
- Fewer deliveries per shift
- Inefficient use of resources

Both scenarios reduce reliability and trust in the system. A well-calibrated model should strike a balance to support accurate planning.