Spis treści

1. Introduction

This report presents an evaluation of the current delivery time prediction model and proposes improvements based on available data. The current approach, which uses a global average delivery time for all orders, is overly simplistic. This section explores an alternative hypothesis, suggests a better predictive method, identifies delivery factors, proposes new data to collect, and discusses the risks of inaccurate predictions.

2. 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 (MAE) 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.

3. Proposed Algorithm and Validation Methodology

As an alternative to the current naive approach, I propose using a simple linear regression or decision tree model that considers multiple features:

- sector_id
- total_weight of the order
- number_of_products
- driver_id
- time_of_day (morning, afternoon, evening)

Methodology to validate the model:

- Split the dataset into training (80%) and test (20%) subsets.
- Train the model on the training set.
- Predict delivery durations on the test set.
- Evaluate accuracy using metrics such as MAE (Mean Absolute Error) or RMSE (Root Mean Square Error).
- Compare the new model's performance to the current global average baseline.

4. Why Some Deliveries Take More Time

Some deliveries may take more time due to real-world constraints not captured in the dataset:

- Lack of elevators in buildings (requires climbing stairs)
- Difficulty accessing the entrance (e.g., intercom, security gate)
- Poor GPS accuracy or hard-to-find addresses
- High floors in apartment buildings
- Parking difficulties (distance from vehicle to entrance)
- Customer delays (not answering, not home)
- Weather conditions or traffic congestion

These factors can significantly impact the actual time it takes to deliver an order.

5. 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.

6. 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.

Conclusion

This report outlines a pathway toward smarter and more accurate delivery time predictions by using sector-based averages and regression-based approaches. Incorporating additional real-world variables and minimizing data errors will further improve the prediction model's effectiveness and operational value.