# Monika_Juzek_research

May 11, 2025

## 1 Part 3: Building and Verifying Hypotheses for Improved Delivery Time Prediction

**Introduction**

In Part 2 of this analysis, I explored historical delivery data and identified several factors that appear to influence the actual delivery duration. These factors include the delivery sector, the assigned driver, the time of day, and, to a lesser extent, the characteristics of the order itself (weight and number of items).

Building upon these findings, Part 3 focuses on formulating and validating hypotheses for improving our current, simplistic delivery time prediction algorithm.

### 1.1 Validating Prediction by Sector

**Line of Thinking:**

My analysis in Part 2 strongly indicated that delivery times are not uniform across different geographic sectors. I observed significant variations in median delivery durations when comparing sector to sector. This leads me to hypothesize that predicting delivery times based on the specific sector of an order will be more accurate than the current method of using a single average across all deliveries.

**Proposed Validation Method:**

To test this hypothesis, I propose the following steps:

1. **Establish a Baseline (Global Average):**
    - I will first calculate the overall average actual delivery time from our historical dataset.
    - Then, I will evaluate the performance of this global average as a predictor on a subset of our historical data that I will designate as my "validation" set (mimicking future, unseen orders). I will use metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to quantify the prediction error. This will serve as my baseline performance.
2. **Implement Sector-Based Prediction:**
    - Next, I will group our historical delivery data by `sector_id`.
    - For each distinct sector, I will calculate the average actual delivery time specifically for deliveries within that sector.
    - Using the same "validation" dataset as in step 1, I will now predict the delivery time for each order based on the average delivery time calculated for the sector to which that order belongs.

3. **Evaluate Sector-Based Prediction:**
   - I will then calculate the MAE or RMSE for the sector-based predictions on the same "validation" dataset.
4. **Compare Performance:**
   - Finally, I will compare the prediction error metrics (MAE or RMSE) obtained in step 1 (global average) with those obtained in step 3 (sector-based prediction).

**Expected Outcome:**

If my hypothesis is correct, I anticipate that the sector-based prediction method will yield a lower MAE or RMSE compared to the global average method. This would indicate that predicting delivery times per sector results in a smaller average prediction error and, therefore, more accurate delivery time estimates. This improvement would demonstrate the value of considering the delivery location when making predictions.

### 1.2 Proposing an Alternative Prediction Algorithm: A Feature-Based Approach

**Proposed Algorithm:**

Building upon the insights gained from our exploratory data analysis, I propose a more sophisticated algorithm for predicting delivery times. This approach will leverage several key features that demonstrated a correlation or variation with the actual delivery duration. These features include:

- `sector_id`: To account for the inherent differences in delivery times across geographic areas.
- `driver_id`: To capture individual driver performance variations.
- Temporal Features:
  - `hour of the day`: To model the impact of traffic and time-dependent factors.
  - `day of the week`: To account for potential day-specific patterns in delivery times.
- Order Characteristics:
  - `total number of items`: Exhibiting a weak to moderate positive correlation with delivery duration.
  - `number of unique products`: Showing a weak positive correlation with delivery duration.

**Methodology for Validation:**

To validate the efficacy of this feature-based algorithm, I will employ the following methodology:

1. **Data Preparation:** The historical delivery data will be preprocessed by extracting the aforementioned relevant features. Subsequently, the dataset will be partitioned into a training set (used to train the predictive model) and a testing set (used for unbiased evaluation of the model's performance on unseen data).

2. **Model Training:** A suitable machine learning model capable of learning complex relationships between the input features and the target variable (`actual_delivery_duration`) will be selected and trained using the training dataset. Potential candidate models include:

   - Linear Regression with interaction terms: To capture potential synergistic effects between features (e.g., driver efficiency in specific sectors).
   - Tree-based ensemble methods (e.g., Random Forest, Gradient Boosting): Known for their ability to model non-linear relationships and feature interactions effectively.

3. **Model Evaluation:** The trained model will then be used to generate delivery time predictions for the orders in the testing dataset. The performance of the model will be quantitatively assessed using appropriate evaluation metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provide a measure of the average magnitude of the prediction errors.
4. **Performance Comparison:** The prediction error metrics obtained from the feature-based model will be compared against the metrics calculated for both the current naive global average method and the simpler sector-based prediction approach (validated in the previous step).

**Expected Outcome:**

I anticipate that the feature-based prediction model will demonstrate a statistically significant reduction in MAE and RMSE compared to both the baseline global average method and the sector-based prediction. This improvement in predictive accuracy will validate the hypothesis that incorporating a broader range of relevant features leads to more reliable and precise delivery time estimations.

## 1.3 Why Some Deliveries Take Longer?

Several things beyond just the order and location can slow down deliveries:

- **Building Issues:** No elevator, confusing layout, security checks.
- **Item Handling:** Big, heavy, or fragile items need more care. Special instructions from customers also add time.
- **Outside Factors:** Traffic, bad weather, parking problems.
- **Driver Factors:** Unfamiliarity with the area, efficiency, customer availability.

## 1.4 What additional data would be worth collecting for future analysis of this domain?

To predict delivery times even better in the future, it would be helpful to collect more information, such as:

- **Building Details:** Type of building (house, apartment with/without elevator, office), delivery floor.
- **Parking:** How easy it is to park and how far to walk to the customer.
- **Delivery Notes:** Any special instructions from the customer.
- **Item Details:** Size and weight of each item.
- **Precise Location:** Exact GPS coordinates of the delivery.
- **Driver Info:** How experienced the driver is.

## 1.5 The Problem with Getting the Guess Wrong

Both under-estimating and over-estimating when it comes to delivery times can cause problems as:

- **Under-estimating:** Customers get annoyed when their delivery is late, customer service gets busy with complaints, drivers feel rushed, and people might stop trusting our delivery times.

- **Over-estimating:** Customers might think our deliveries are too slow and choose other services, our drivers might have too much free time, and we might look less efficient than our competitors.