Building and verifying the hypothesis

Szczepan Polak

Hypothesis validate

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.

Calculating delivery times for each sector, can help improve the accuracy of the prediction.
In order to confirm this hypothesis, I would first check what the correlation is between average run times and sectors.
In the event that the analysis of the data indicates some relationship between delivery time and sector, it would mean that it is worth trying to use this data in the construction of the algorithm.
Because of the historical data we have, the next step would be to create an algorithm and test it on a validation set.
In this case, the base algorithm would be the previous one based on the mean, which should also be tested on the validation set.
The results should then be compared using metrics suitable for regression, e.g. MSE (Mean Squared Error) or MAPE (Mean Absolute Percentage Error).
If a comparative analysis of the results obtained indicates that the new model/algorithm more accurately predicts delivery times, this will mean that the hypothesis is fully validated.

New algorithm ideas

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

Creating a new algorithm could involve creating it manually or using machine learning models suitable for the regression task, such as a decision tree or random forest. However, each option will first rely on data collection and analysis. Based on the analysis, it would be useful to use information about the sector and also the time of day of delivery to make more accurate predictions. Undoubtedly, the data also indicate that a significant share of delivery time is contributed by the employee, in particular the time of physical delivery of the shipment recognized in the data as STOP time. Perhaps it would be useful to collect other information, including information on employees, which is discussed in the next sections.

Personally, I would suggest using a machine learning model, probably as interpretable as possible, which will give the best results (ranging from linear regression to a decision tree or random forest), mainly due to the difficulty of interpreting neural network predictions.

I would perform validation of the new algorithm exactly as described in the previous section:

- data analysis
- creation of a model / algorithm
- prediction on the validation set
- prediction error analysis, comparison to baseline results on validation set

Why could some deliveries take more time?

For example, some buildings don't have elevators etc. Describe your ideas.

- Parking problems which takes time.
- Weather (e.g. rain, or snow) → longer travel time, issues with physical delivery at the end.
- Traffic jams (time of day) → longer travel times.
- Weight/size of cargo.
- Employees (employees generally have different characteristics that affect their capabilities and performance at work, perhaps a difference in age, physicality and more which can affect longer delivery times especially the physical segments.)
- several deliveries in a single pass

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

- It will undoubtedly be interesting to include analysis of employee data, such as age, seniority, level of athleticism, etc. (with rules, ethics, of course) to solve the mystery of why some of them, on average, deliver shipments faster than others.
- Employee working time
- Vehicle type
- Weather conditions
- building type, presence of elevator

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

If the delivery time is overestimated, the negative impact will be much less. Undoubtedly, the customer, then, may be surprised by faster delivery which will affect satisfaction, although it is worth remembering that if the overestimation is significant, and this information is passed to the customer immediately we make a bad first impression. Of the negatives, this can impact problems in the fleet management and delivery system, although it seems to be less than in the opposite situation.
If the transit time is underestimated, the result is worse because, the delay affects customer dissatisfaction, and also introduces chaos in the system and the need to reschedule future plans.