BUILDING AND VERYFINING THE HYPOTHESIS

Natalia Wrześniak

I. Main goals

Brief description of other possible way of predicting delivery times. Below report is connected to analysis report [1] and exploration of possible reasons for longer deliveries as well as potential over- or under- estimating.

II. Assumptions and methodology

Note that as in the previous report [1] The most important assumption is the way of calculating actual delivery length. I am taking into account only segments with type 'DRIVE' as I assume that this is the most important factor while delivering. I simply subtracted 'end time' and 'start time' I did not use STOP segments as I am not sure if all activities which take places during break are connected to 'delivery' actions such as taking stairs, waiting for a customer to open the door etc. or there are just driver's breaks.

III. Analysis

1. Alternative idea of prediction delivery duration – predicting delivery times per sector

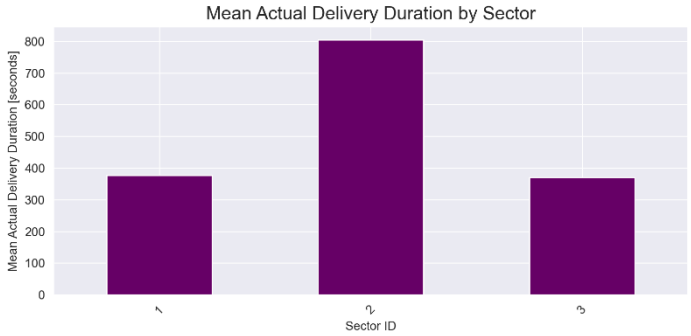

Chart 1 Mean Actual Delivery Duration by Sector

As shown in [Chart 1] there is a reason to suspect that predicting delivery times per sector is a better option than currently used algorithm. Longer (upper than an hour) delivery durations are appearing pretty much only in sector '2'.

After applying algorithm that uses data and mean values for particular sector it could be validated by counting error calculated as dividing predicted time by actual time and make a histogram showing the distribution of these errors. They certainly will be lower than in current method.

2. My idea of predicting delivery times is to use machine learning model.

Because of the fact that in the dataset there is quite a lot historical data good idea it will be to use machine learning model or even deep learning model. I suggest using Gradient Boosting Regression algorithm, which often performs well with tabular data and can handle complex relationships between features.

All tables ('orders', 'products', 'orders_products' and 'route_segements') from dataset should be merged into one table based on their common keys – 'order_id' and 'product_id'.

Data should be splited into training and testing sets. And then model should be fitted to train data and hyperparameters connected to chosen model should be tuned in order to achieve better performance on unseen data (testing set).

Validation method would be to count Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and compare them to actual true values.

Of course, this method could be updated all the time while receiving actual delivery times for new orders and improve model performance.

3. Some deliveries could take more time than the others – why?
    - Lack of elevators
    - Problems with accessing the building (for example guarded housing estate)
    - Problems with contacting the customer
    - Lack of available parking spots
    - Poor road quality
4. What additional data would be worth collecting for future analysis of this domain?
    - Clear separation between driver's stop for delivery issue or for any other reasons, for example: 'DRIVE', 'STOP', 'STOP_DELIVER'
    - Is it a single-family house or apartment building?
    - If apartment building does it have an elevator?
    - Maybe divide sector into more subsectors?
    - Size of the packages (orders) – maybe if unique size delivery takes more time?
5. What is the risk of over – or under – estimating the delivery times?

Overestimating	Underestimating
Less efficient work – too little deliveries could be planned for each day	No physical ability to deliver all orders by drivers which could lead for customers disappointment
Addressee could not be on site	Addressee could not be on site
	No enough breaks for drivers as they will be all day in hurry – not ethical and dangerous
Decreased customer satisfaction	Decreased customer satisfaction

Table 1 Risk of over - or under - estimating the delivery times

IV. Conclusions

1. Simple alternative idea of prediction delivery duration is to predict times per sector as there is a significant difference between times in analyzed three sectors.
2. My suggestion is to use machine learning model to improve prediction performance.
3. Collecting new data types could lead to better accuracy while predicting deliver durations.

V. Extra Sources

[1] ANALYSIS OF DELIVERY TIME PREDICTIONS report, Natalia Wrześniak