I don’t know if I understood the table called “route_segments” correctly, because the delivery time for each order calculated by me is much bigger than expected delivery time in table “orders”.

I calculated the delivery time as a difference between value of “segment_end_time” in a row where segment_type = “STOP”, and value of “segment_start_time” in a row where segment_type = “DRIVE” and which is just below the row with “STOP”, because in my opinion there weren’t any orders which were counted in 3 segments intentionally.

Cleaning data

Before any work on dataset, the data must first be cleansed, which I did as follows:

1. First of all, I deleted duplicated rows.
2. Secondly, I deleted all the rows with segment_type = “STOP” and order_id = NULL. I wasn’t sure if I should take it into consideration during this task, but I decided to not do it, because I can’t be sure that data in these rows are correct.
3. After it I deleted rows with all of the orders where segment_end_time is lower than segment_start_time. I considered just replacing data in these 2 columns in selected rows, but again I wasn’t sure if then these data would be correct or not.
4. At the end I noticed that there were orders in which delivery took more than 4 hours, so I deleted them, because these data were probably also incorrect.

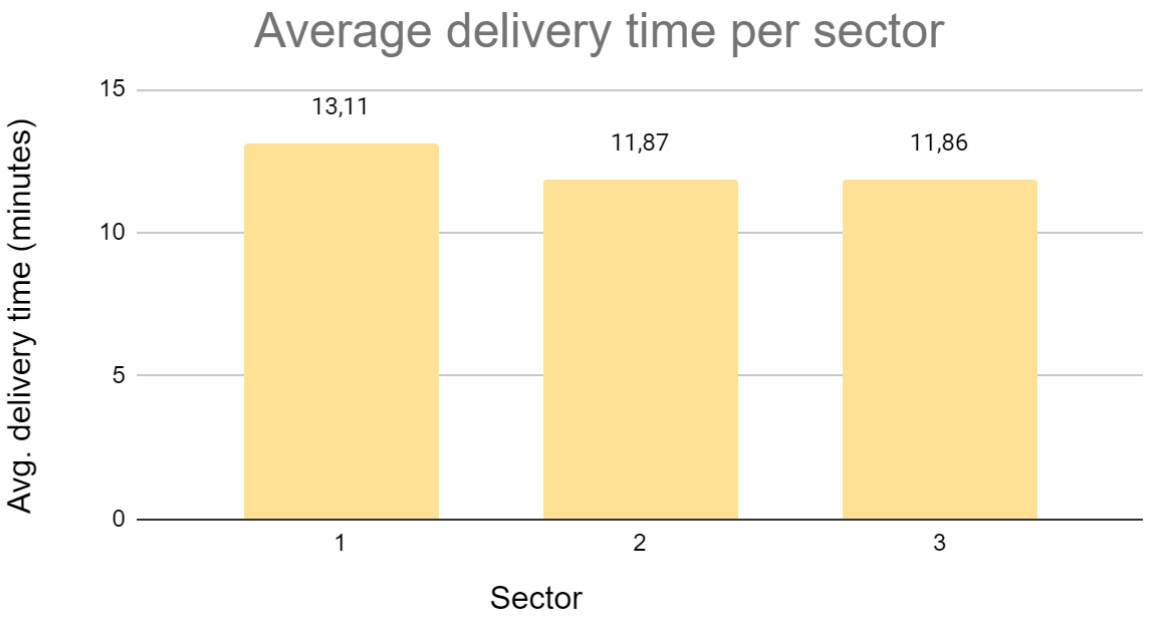
Task 1

All of the orders were delivered in 6 - 25 minutes, but most of them were delivered in 11 - 16 minutes.

Task 2

The prediction error for most of the deliveries were in 5 - 15 minutes which is a lot if most of the orders were delivered in 11 - 16 minutes.

Task 3

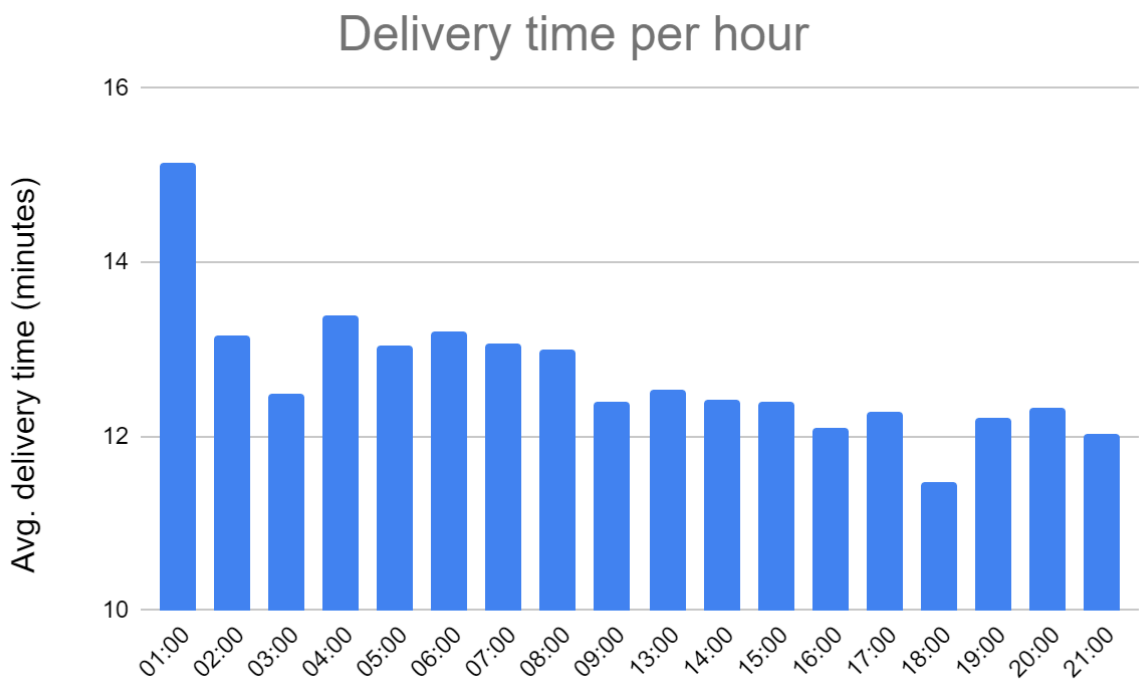


There is a difference in average delivery time between sector 1 and others. It's about 10%, so it's a noticeable difference and it's good to count it in expected delivery time.

Task 4

Delivery time during the day

I considered if there is any difference in delivery time due to the time of the day, because this may be correlated with traffic jams or some other events.



The graph above shows that on average, during the day, the delivery time decreases. I just want to point out that there is a small sample for orders delivered at 1:00. and 13:00, so there may be some anomalias.

Let's divide this graph into 2 parts: 1:00 - 9:59 and 13:00 - 21:59. After that, we see that there is a difference in average delivery time between morning (12,4 min) and afternoon (13,23 min) shift. The difference is less than 7%, so it's not much, but we can also take it into consideration during the counting expected delivery time.

It's probably not because of the traffic jams, because I would expect they occur between 15:00 - 17:00, but it may happen, because of the people who are more tired in the morning shift which is normal.

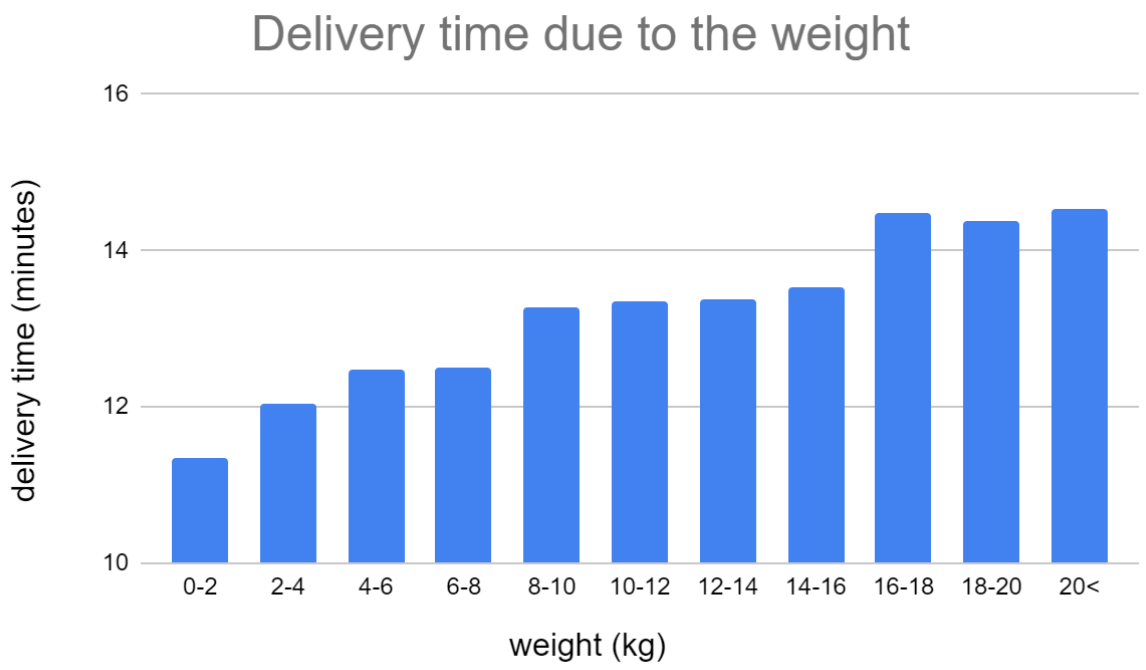
I also noticed that especially in the morning shift, on average, the longer delivery time is in the first hour of shift and shortest in the last one. In the first one we have a small sample, but the difference is close to 14% compared to the avg in 2:00 - 8:59; and in the last one it's less than 5%.

Summary

In my opinion, the only factor which may be valuable at this moment due to the time of the day is a difference between morning and afternoon shift, but if the company collects more data I will check again the difference between the first hour of the shift and the rest.

Delivery time due to weight

I considered if there is any difference in delivery time due to the weight of the order.



The graph shows that there is a significant difference because of it. The heavier the order, the longer it takes to deliver. Let's divide this graph into 4 parts:

1. orders lighter than 4 kg
2. orders 4 - 8 kg
3. orders 8 - 16 kg
4. orders heavier than 16 kg

Order's weight	**< 4 kg**	**4 - 8 kg**	**8 - 16 kg**	**16 kg <**
Average delivery time (minutes)	11,83	12,49	13,57	14,78
Number of orders	683	775	717	31

The table above shows that the difference in average delivery time due to the weight of the order is significant.

Summary

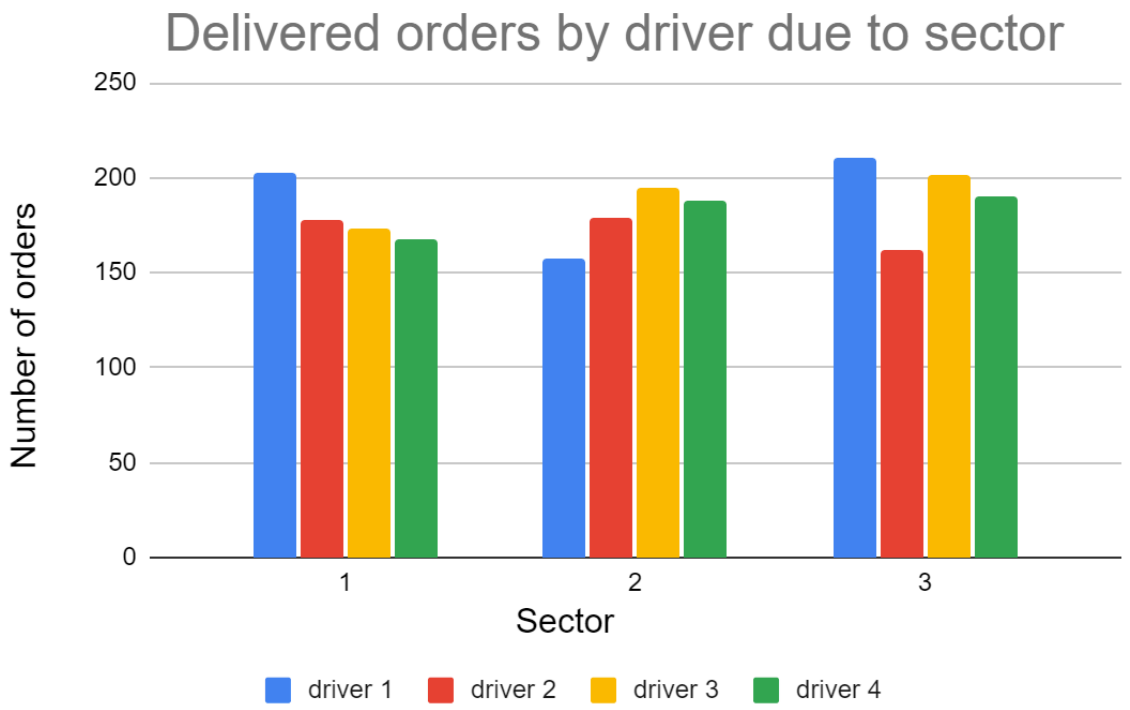
In my opinion, this factor may be really valuable for prediction quality improvement.

Delivery time due to the driver

I considered if there is any difference in delivery time due to the driver who delivers the order. And below there is a table with the data I get.

driver_id	1	2	3	4
Average delivery time (minutes)	11,08	12,13	13,42	14,50

In the table above we clearly see that some drivers deliver orders faster and others slower. To be sure that these data are valid and are not distorted by a different amount of orders delivered in different sectors, I checked how many orders each driver delivered in each sector.



I think that data are not distorted, because as you can see the fastest driver delivered the most orders (in comparison to the others) in the sector 1 where the delivery time on average takes 10% longer than in other sectors

Summary

In my opinion, this factor will definitely be really valuable for prediction quality improvement if it's possible to know which driver will deliver the food in the moment when the system calculates the expected delivery time.