Ocado Intern Recruitment Task

Jakub Michnicki

April 7, 2024

1 Technology stack

- Python
- numpy
- seaborn
- pandas
- matplotlib
- mysqlconnector

2 Part 2.

1. Generate a histogram showing the actual delivery length with 1 minute granularity (rounded up).

Figure 1: Parcel delivery length

Most of the parcels are being delivered within the 25 minute mark. There are some outliers which could be possibly dealt with to make the estimation more accurate.

2. Generate a histogram showing prediction error (difference between planned and actual delivery times).

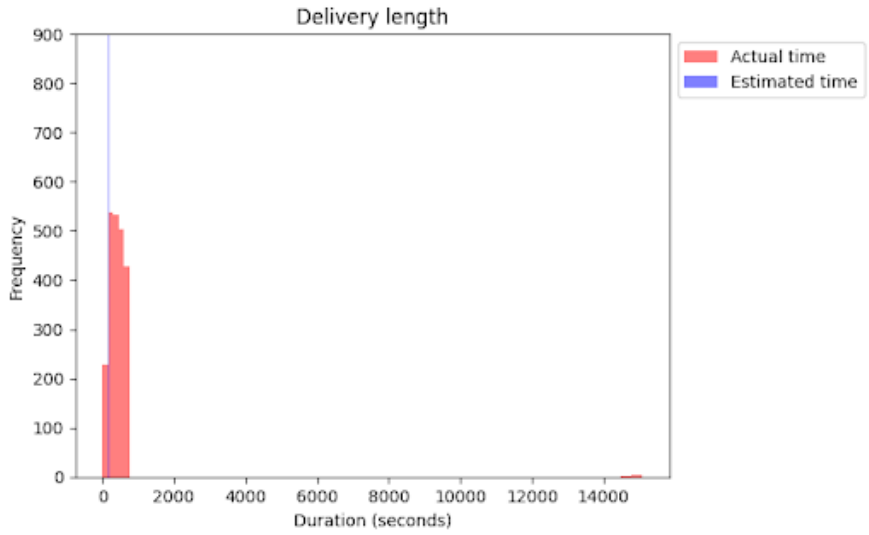

Figure 2: Parcel delivery length - actual vs average

Creating this comparison shows that the estimated time lies withing the range of the actual time but let's zoom into it.

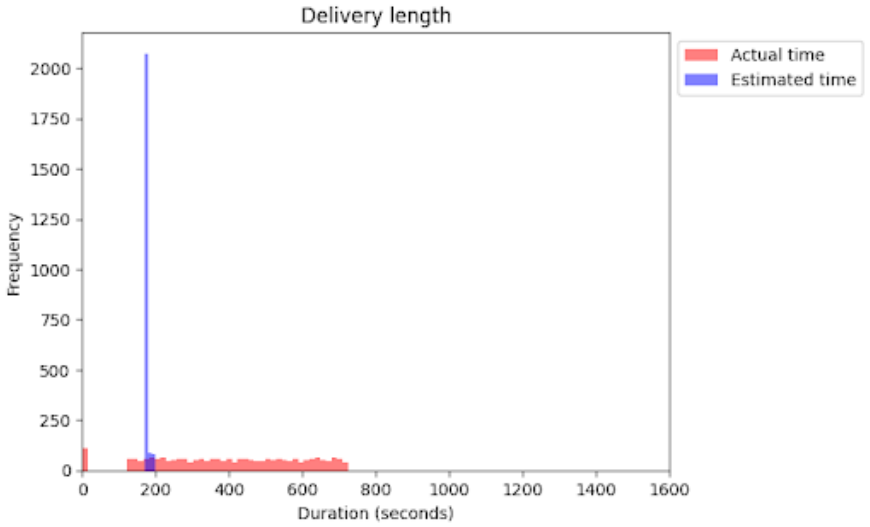

Figure 3: Delivery length - actual vs average x-limited

It is visible that the estimated time is located in one spot which makes the estimation very inaccurate and that will need some tweaking.

3. We received insight from our drivers that delivering in one of the sectors is significantly longer than in other sectors. Generate a chart to visualize this hypothesis.

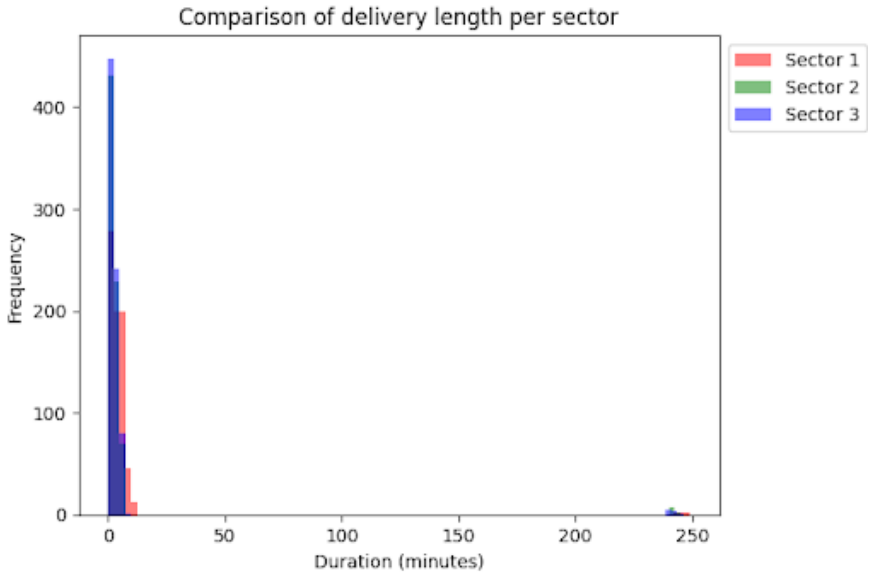

Figure 4: Whole dataset

It is visible that there is some red area expanding to the right but it is not that clearly visible because of the outliers so let's skip them for now.

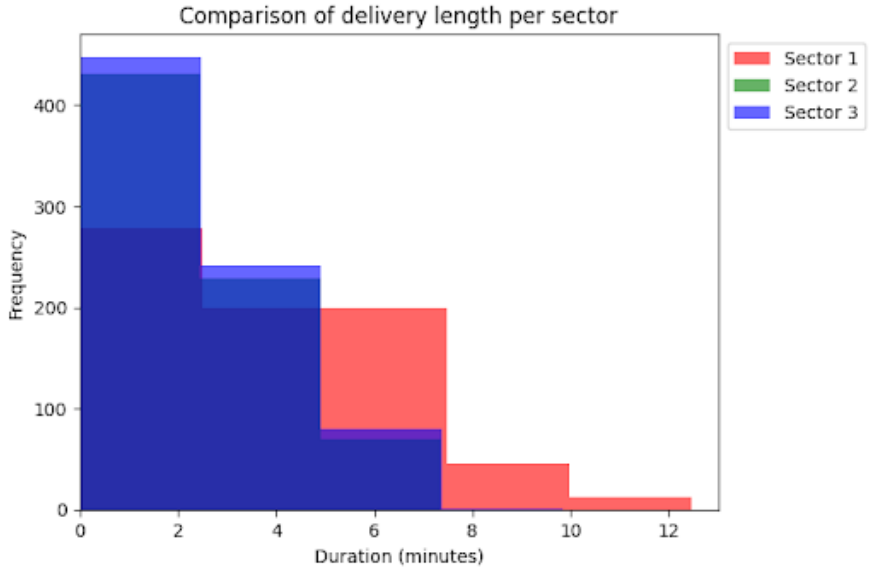

Figure 5: Dataset x-limited

As shown above, sector 1 (red area) is taking way more time than either sector 2 or 3 so this feature has an impact and is correlated with our main feature - parcel delivery time. This should be probably the feature to pass to when creating a model to predict the duration in the future.

4. Play with the data by grouping, aggregating and remodelling it. Are you able to find any correlations or trends that could be valuable for prediction quality improvement? Describe briefly your findings and visualise them on charts.

Considering all of the variables present in the database, I do not think that things like:

- order_id

- product_id

would have any relationship with the shipping time.

Things I cannot evaluate more on:

- quantity
- weight
- customer_id

Quantity and weight of delivered parcels make the car heavier but it shouldn't really impact the time of arrival of the package but rather the fuel consumption. Customer ID might be of use since some of the customers might create delays but I do not think the amount of them is so great for it to actually matter.

Variables that I think might be worth investigating are:

- driver_id (perhaps the driver drives slower than others)
- segment_start_time (might show whether the time of the day makes the traffic more congested)

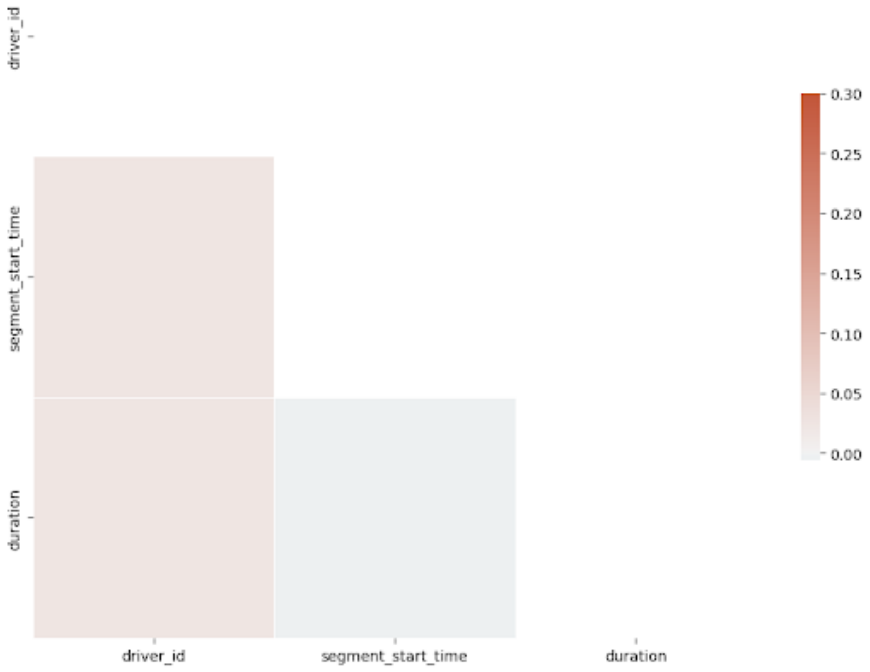

Figure 6: Correlation matrix of selected features

The correlation matrix shows that there is little to none correlation between selected features and the delivery time so sticking to only sectors should be a good idea (this idea should be verified with correlation matrix too).