Data Analysis

Vitaly Kozhemyak

January 2022

1 Modeling results

To compare quailty of different models RMSE metric is used. We perform train-test splitting and use the test dataset to obtain the following results.

Model	**RMSE**	**R^2**
Ridge regression	0.43178	0.98734
LightGBM	0.40342	0.98895
Random Forest	0.44910	0.98631
Dense neural network	0.55436	0.97914
Baseline	0.41872	

Table 1: Result table

2 Top opportunity to improve the upfront pricing precision

In addition to main features:

```
['gps_confidence', 'predicted_distance',
 'predicted_duration', 'eu_indicator',
 'overpaid_ride_ticket', 'dest_change_number',
 'prediction_price_type', 'change_reason_pricing',
 'entered_by']
```

Figure 1: Main features

I suggest we use features such as

```
['day_of_week', 'is_weekend',
 'part_of_day', 'speed',
 'isairport_2000', 'isairport_6000']
```

Figure 2: Additional features

As I mentioned before in *Data_Analysis.ipynb* that the next features `['isairport_2000', 'isairport_6000']` could be derived from GPS-coordinates. Also if I knew pickup and dropoff coordinates I would add:

- weather conditions,
- the most visited places,
- good/bad neighbourhoods,
- some data from similar ride-hailing applications.