LVIV·2018 DATA SCIENCE <SUMMER>SCHOOL

Project:P02. X fare forecasting

Rajesh Tamang; Dmytro Maliarenko

Project description

➢ **Model for fare estimate, which outputs most accurate recommended fare**

Dataset

Input - data of X city for Y month

Constraints - locations (pickup and dropoff)

ride distance

suburb rides

search time

seasonality

Output

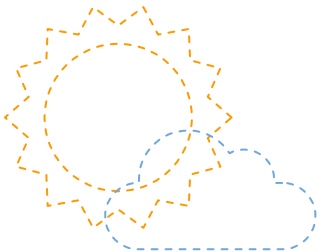
Predict fare estimate, which outputs most accurate recommended fare for maximisation of supply and demand balancing.

Predictive modeling → Regression task

Feature Engineering

➢ **We added weather data to our model**

→ uklon



In [28]:
```
weather.tail()
```

Out[28]:

	floor_date	Temp,C	Weather	Wind,km/h	Humidity	Barometer,mbar
4411	2017-10-31 21:30:00	3	Clouds	24	0.60	1008
4412	2017-10-31 22:00:00	3	Clouds	24	0.60	1008
4413	2017-10-31 22:30:00	3	Clouds	28	0.65	1009
4414	2017-10-31 23:00:00	3	Clear	28	0.65	1009
4415	2017-10-31 23:30:00	2	Clear	28	0.70	1009

In [29]:
```
weather.describe()
```

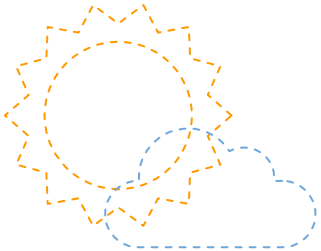
Out[29]:

	Temp,C	Wind,km/h	Humidity	Barometer,mbar
count	4416.000000	4416.000000	4416.000000	4416.000000
mean	15.276042	12.468976	0.721587	1015.846241
std	7.707456	7.541801	0.187891	8.560472
min	-3.000000	0.000000	0.200000	978.000000
25%	9.000000	6.000000	0.590000	1011.000000
50%	14.000000	13.000000	0.750000	1016.000000
75%	21.000000	17.000000	0.880000	1020.000000
max	34.000000	43.000000	1.000000	1039.000000

Source: https://www.timeanddate.com/weather/ukraine/kyiv/historic?month=8&year=2017

Feature Engineering

- **Holidays**



In [34]:
```
holidays.head()
```

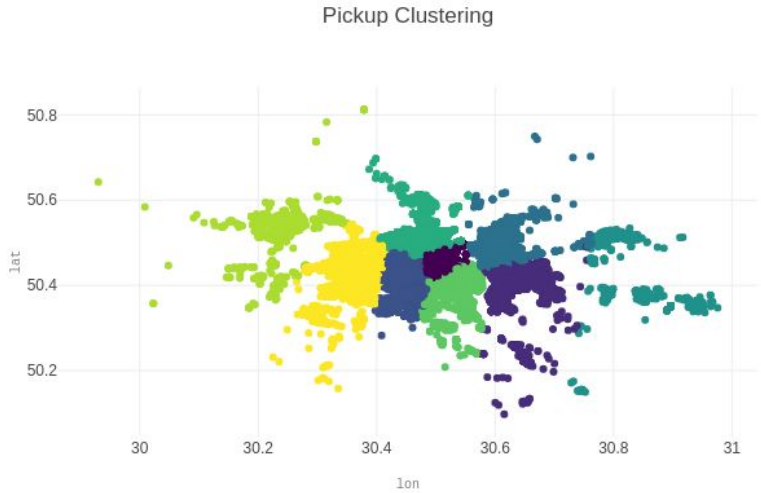
Out[34]:

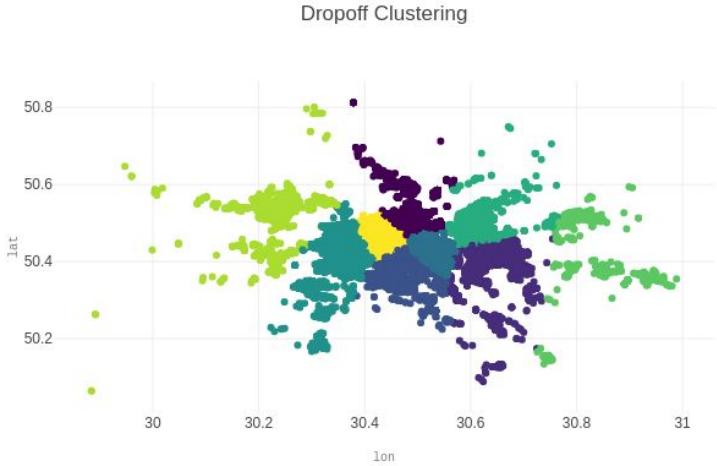
	date	isHolidays
0	2017-08-24	1
1	2017-10-14	1

Feature engineering:

Clustering for

- pickup location
- dropoff location
- ride





Models selection

- We started from simple Linear Regression (mape = 12%)
- Explored different models:
  - XGBoost
  - LightGBM
  - RandomForestRegressor
  - ExtraTreesRegressor

LightGBM demonstrated the best result and speed among other models on the validation subset

LightGBM

Regression performance accuracy

LightGBM:

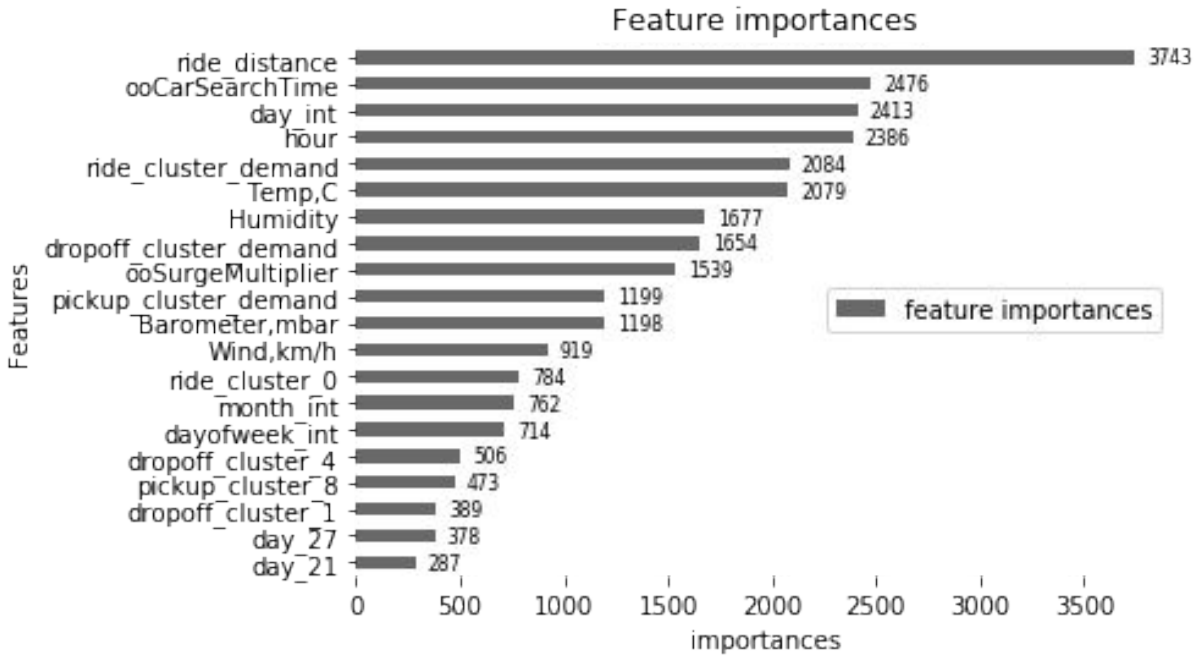
In [83]:
```
# The mean squared error
print("Mean squared error: %.2f" % mean_squared_error(y_test, y_pred))
```

```
Mean squared error: 300.03
```

In [84]:
```
# The mean absolute percentage error
print("Mean absolute percentage error: %.3f" % mean_absolute_percentage_error(y_test, y_pred) + '%')
```

```
Mean absolute percentage error: 6.961%
```

Feature importances




LVIV • 2018
DATA
SCIENCE
<SUMMER>SCHOOL
JULY 16 -
AUGUST 2
Thanks!
LVIV • 2018