ITMO

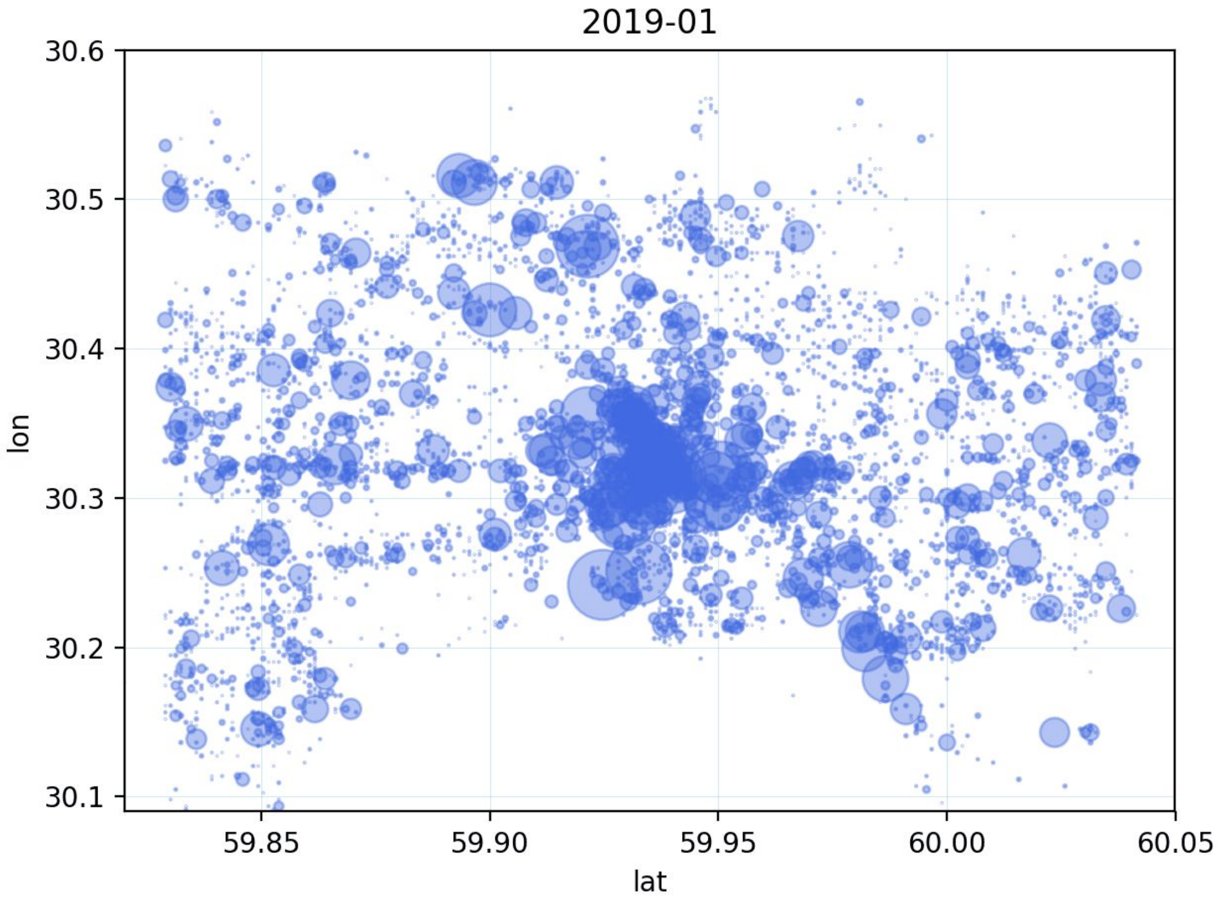
Machine Learning for Industrial Data

Laboratory task №1

Prusskiy D.A., J42322c
Barkovskii V.V., J42332c

2022

Laboratory task problem description



Posts count scatter map

Goal - posts count prediction in time series second dimensional coordinate data.

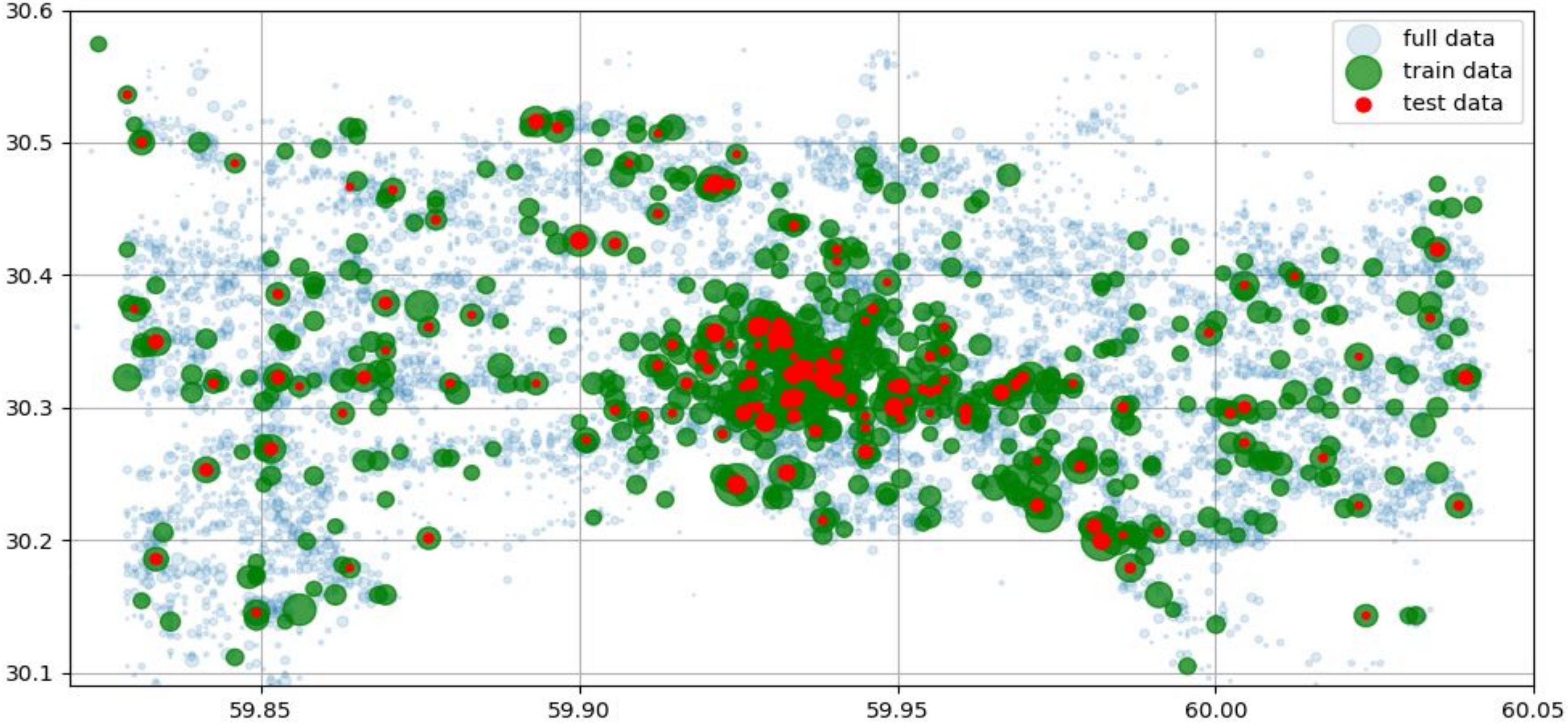
We had the data with posts description over the year, and therefore we had to predict the posts count per coordinate cell for the month ahead.

Step-by-step laboratory work execution

1. Data analysis and preprocessing;
2. Feature extraction;
3. Model hyperparameters selection;
4. Test data evaluation.

Preprocessing

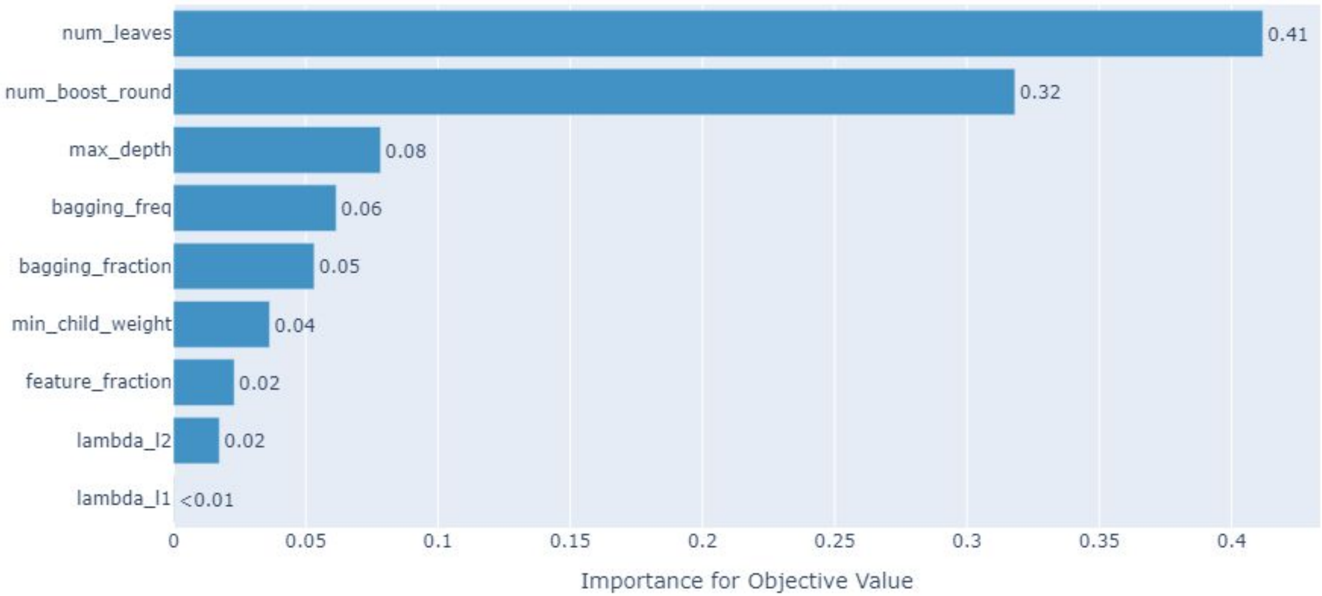
1. Removing coordinate outliers;
2. Removing small polygons;
3. Log(target).

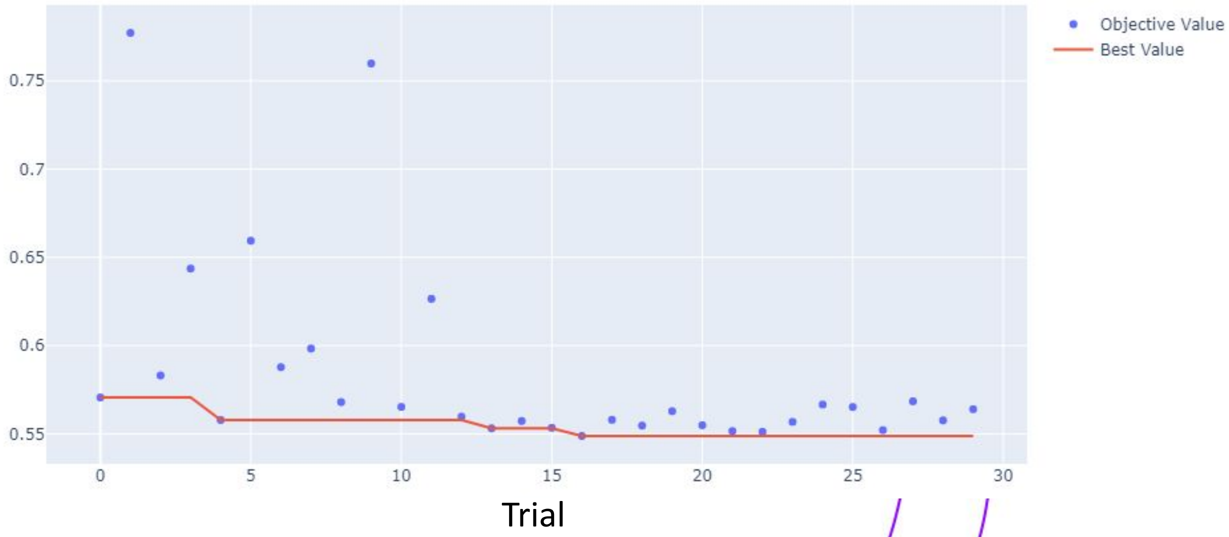


Features

1. Latitude and longitude;
2. Posts count and aggregated features per poly calculating;
3. Features lagging;
4. Datetime feature extraction:
   a. *sin + cos* for yearly, weekly and daily seasonalities;
   b. timestamp for overall trend.

Hyperparameters selection





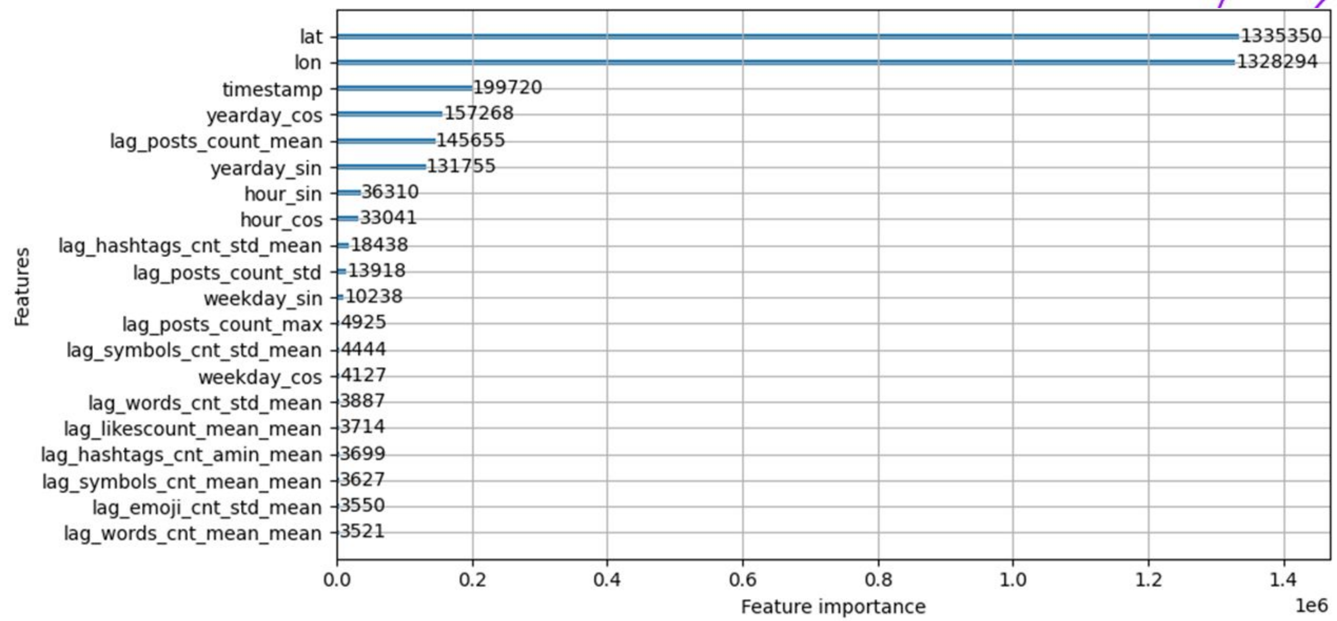
1. Optuna framework was used to select optimal LGBM model hyperparameters.

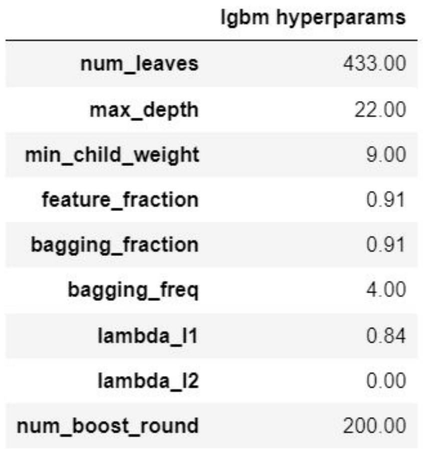
2. Optimization metric was RMSE, and the last month of train dataset was used as validation part.

3. Count of trials was 30.

4. The posts count was logarithmic to normalize target before training

Model description





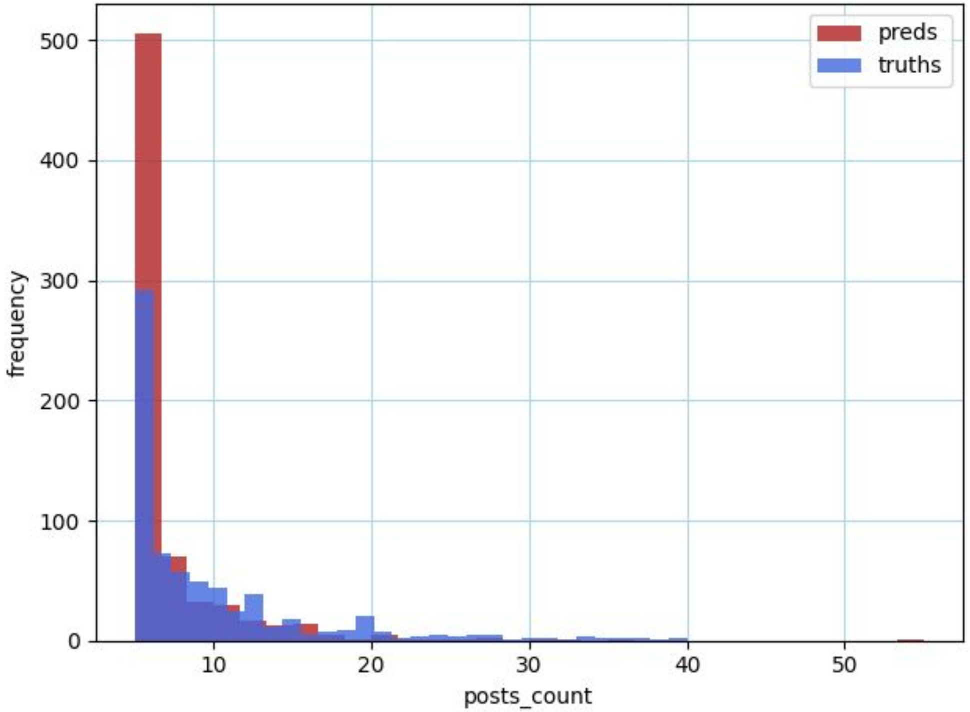
	lgbm hyperparams
num_leaves	433.00
max_depth	22.00
min_child_weight	9.00
feature_fraction	0.91
bagging_fraction	0.91
bagging_freq	4.00
lambda_l1	0.84
lambda_l2	0.00
num_boost_round	200.00

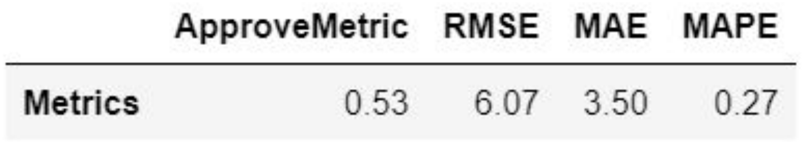
1. The LGBM model was used to predict posts count.

2. The model with setted up hyperparameters was trained on the whole train dataset.

3. The most important features are lon, lat and date-time domain features.

Test data evaluation





	ApproveMetric	RMSE	MAE	MAPE
Metrics	0.53	6.07	3.50	0.27

1. Posts count per cell were predicted on test data set. Minimum posts count was selected 5.

2. The target approve metric was received **0.53**.

3. Other metrics are represented on the table.

Thank you for your attention!

ITsMOre than a UNIVERSITY