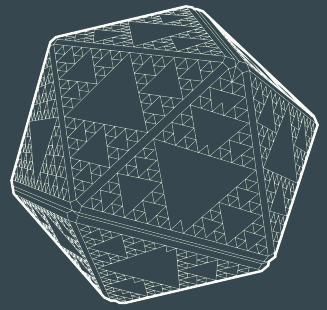
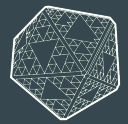


B I G

D A T A



L A B

vodafone

Vodafone

Base station internet traffic prediction

Dima Mashchenko

Business objective:

Optimization of investment into the network infrastructure

Expected result:

The list of nodes that require upgrading in 6 month

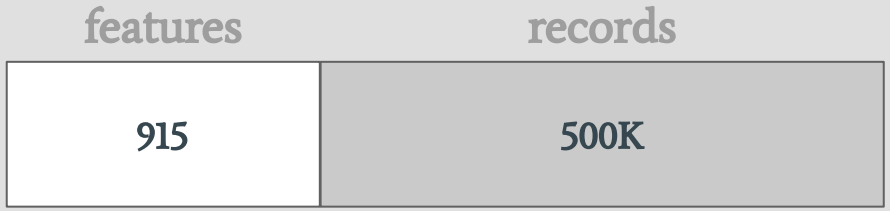
Problem solving approach:

Prediction of traffic consumption for each mobile subscriber

Key metric: **RMSE** - root mean squared error

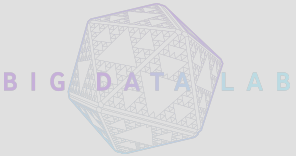
Data provided:

- **subscriber data mart**
- **mobile internet traffic for the last 5 months**



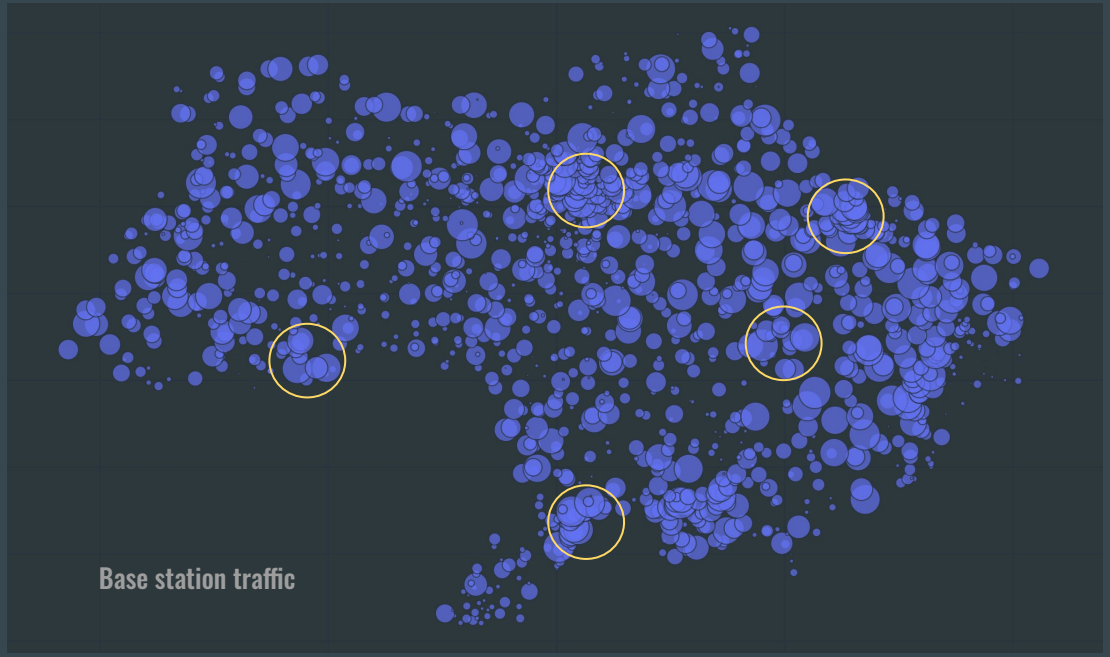
Please keep in mind: the data has been previously preprocessed for security reasons

Extra dataset & features:

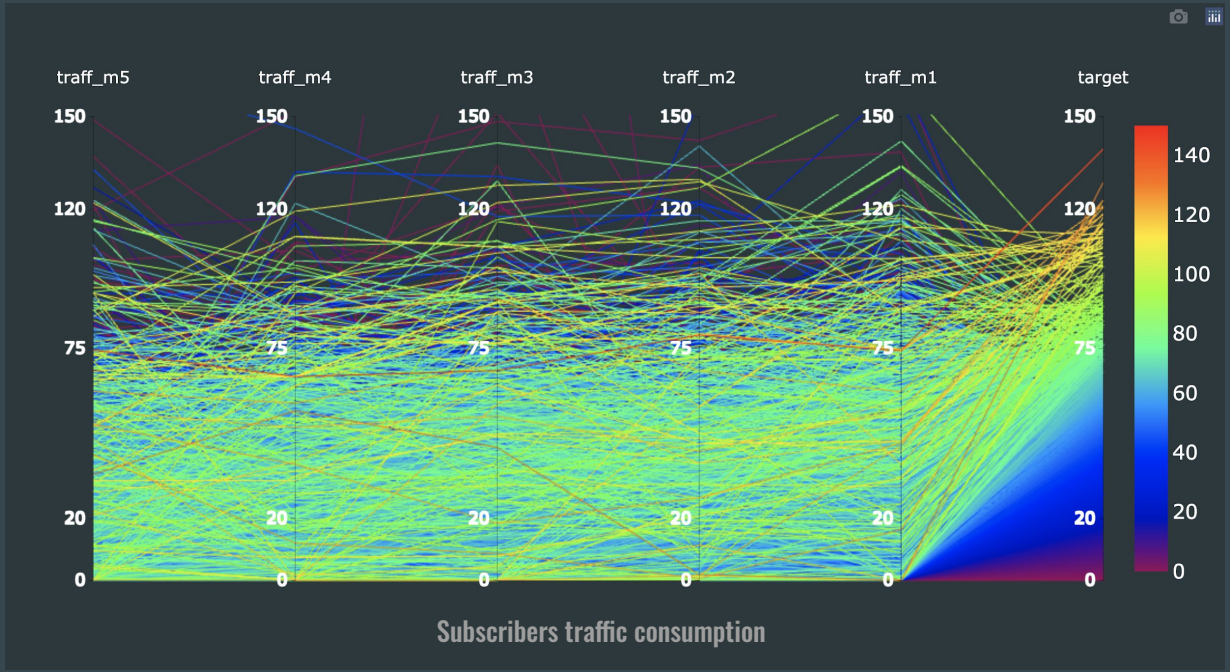


- dataset
  - geo dataset: [city, lat, lon, capital, population]
- features
  - traffic statistics: [min, max, std, td, mean, median]
  - traffic statistics by city: [min, max, mean, median, td]

Data exploration:

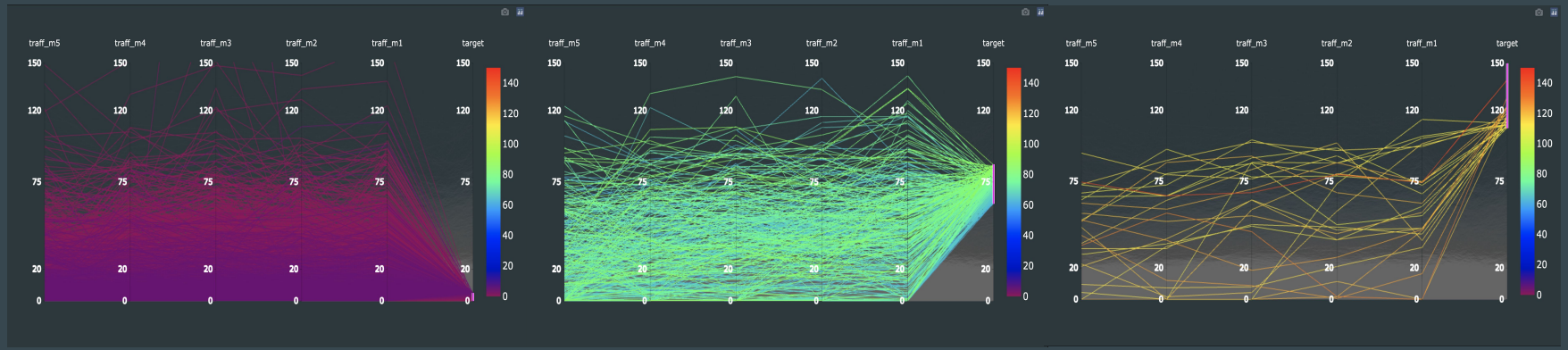


Data exploration:



Data exploration:

BIG DATA LAB



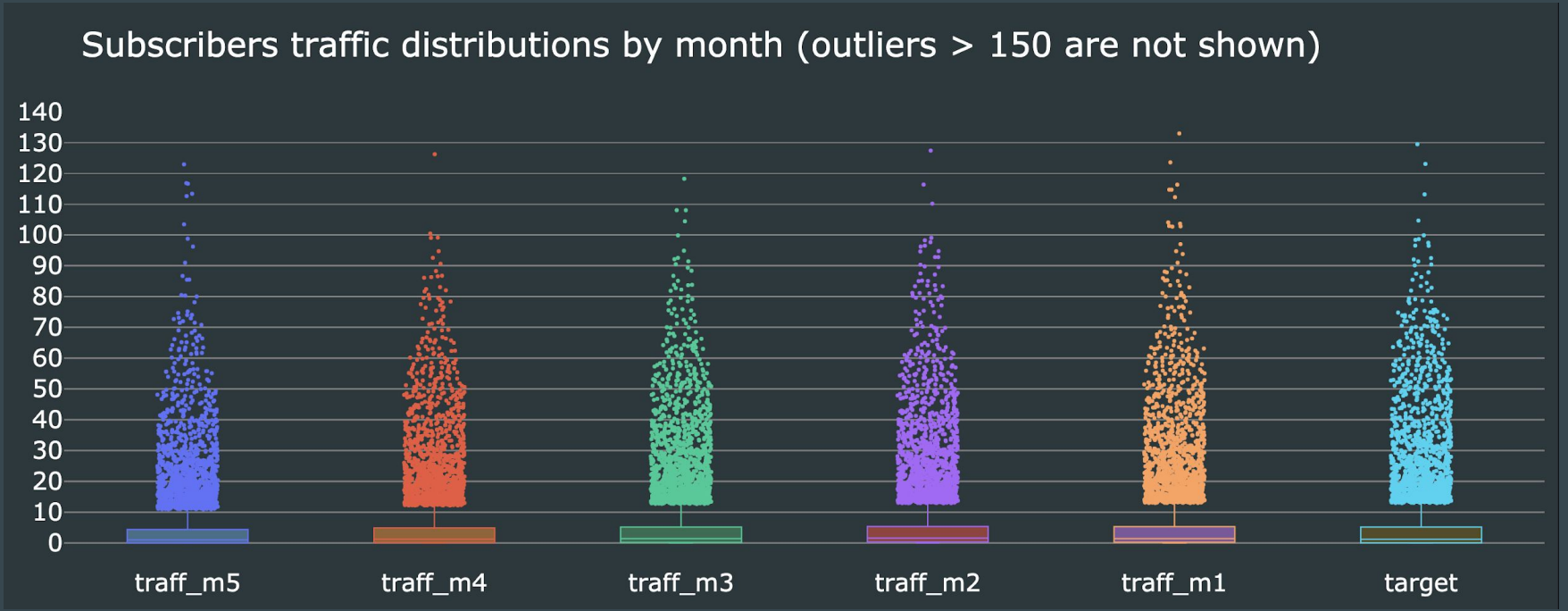
Low consumers

Middle consumers

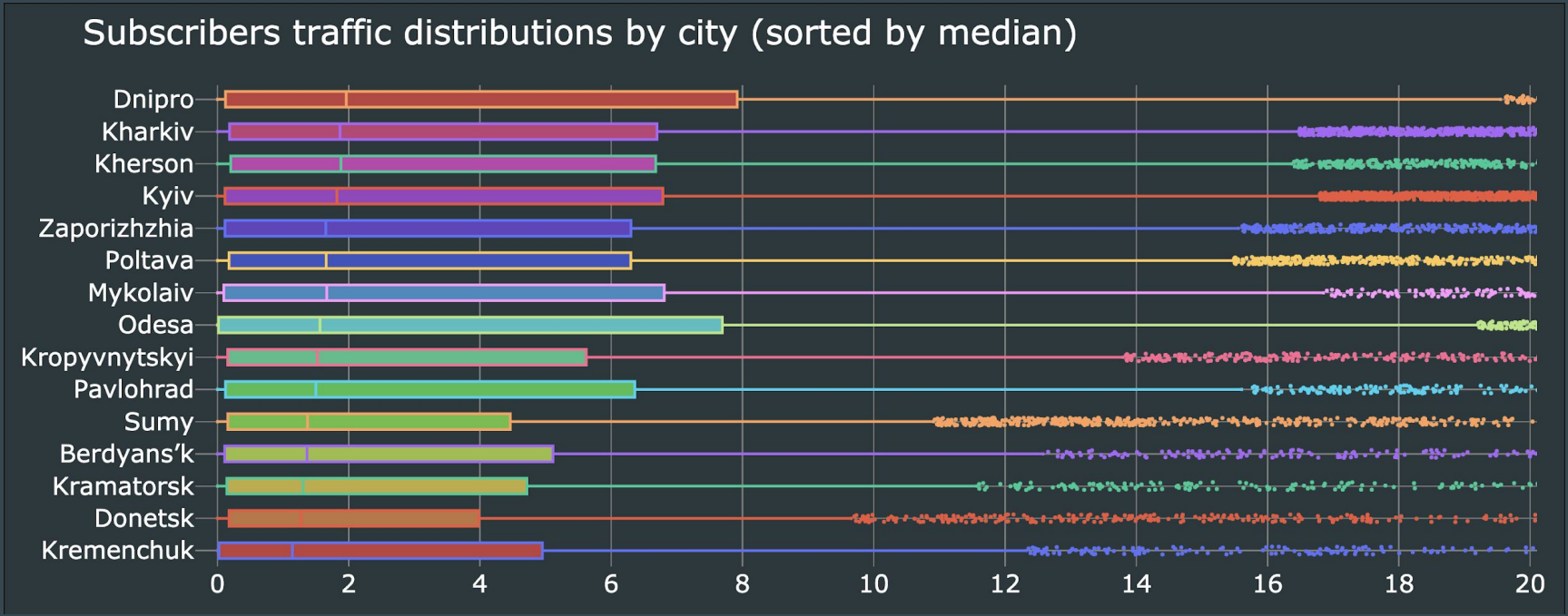
Top consumers

vodafone

Data exploration:



Data exploration:



Baseline:

Feature selection algorithm: **Boruta (custom implementation)**

- automatic feature selection
- works well with big feature set
- based on LGBMRegressor

Base estimator: **LGBMRegressor**

- handle missing values
- fast
- efficient

Base RMSE score: **9.393 (traffic mean as a predictor)**

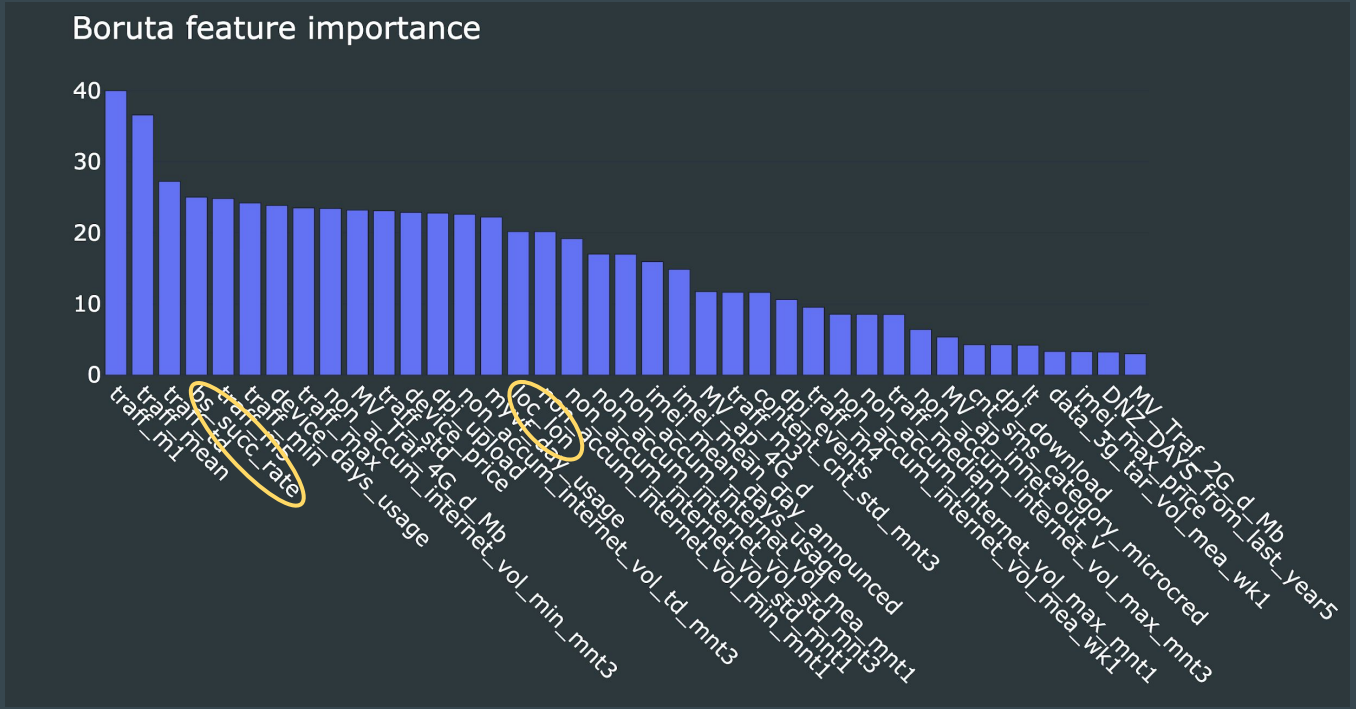
Feature selection:

BIG DATA LAB

	rmse	r2	max error	feat. count
Boruta + city cluster stats	8.652	0.447	137.2	44
Boruta	8.660	0.446	137.2	37
All initial + all city	8.636	0.447	137.5	967
All initial	8.645	0.448	140	915
Base	9.393	0.348	139	5

vodafone

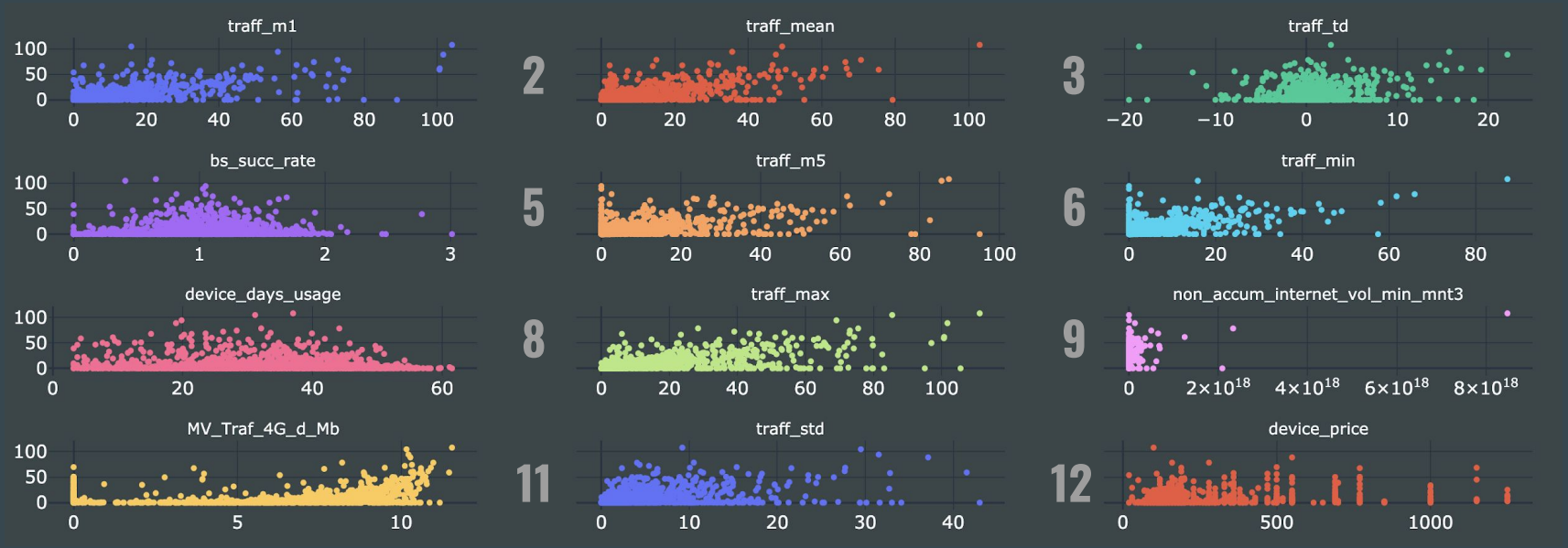
Feature selection:



Feature selection:

BIG DATA LAB

Top 12 boruta features



vodafone

Model selection:

BIG DATA LAB

	rmse	std	
Lasso	8.582	0.02	
ElasticNet	8.637	0.03	
GradientBoostingRegressor	8.379	0.04	too slow
HistGradientBoostingRegr	8.373	0.03	
LGBMRegressor	8.376	0.03	

5-fold shuffle split cross validation score on train dataset

vodafone

Model evaluation:

BIG DATA LAB

	rmse	r2	max error	
Lasso	8.766	0.433	184.5	
ElasticNet	8.750	0.435	158.9	
HistGradientBoostingRegr	8.659	0.446	137.6	
LGBMRegressor	8.663	0.446	137.2	

Tested on validation dataset

vodafone

Hyper-parameters tuning:

BIG DATA LAB

	rmse	r2	max error
HistGradientBoostingRegr	8.623	0.451	137.2
LGBMRegressor	**8.610**	**0.452**	**137.9**

Best params:

colsample_bytree: 0.8
learning_rate: 0.015
max_depth: 5
min_child_samples: 200
min_split_gain: 0.05
n_estimators: 200
num_leaves: 30
reg_lambda: 0.3
subsample_for_bin: 100000

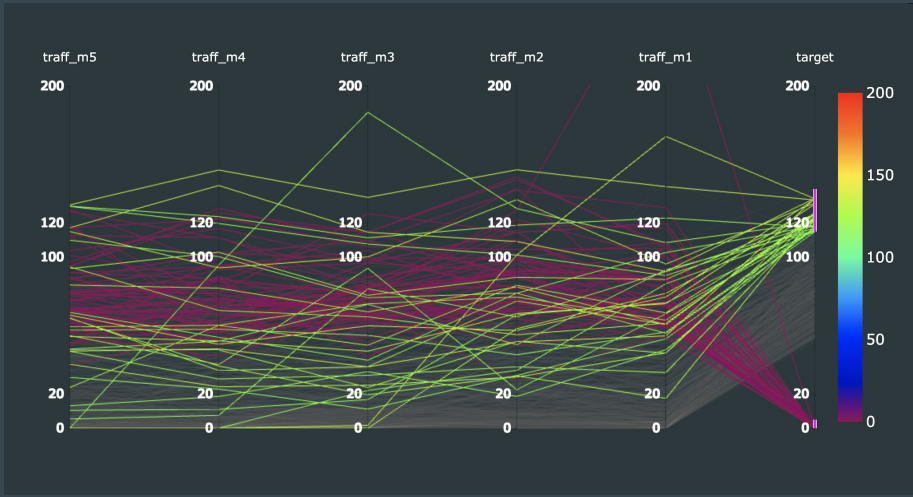
Performed by GridSearchCV. Searched on train. Evaluated on validation.

vodafone

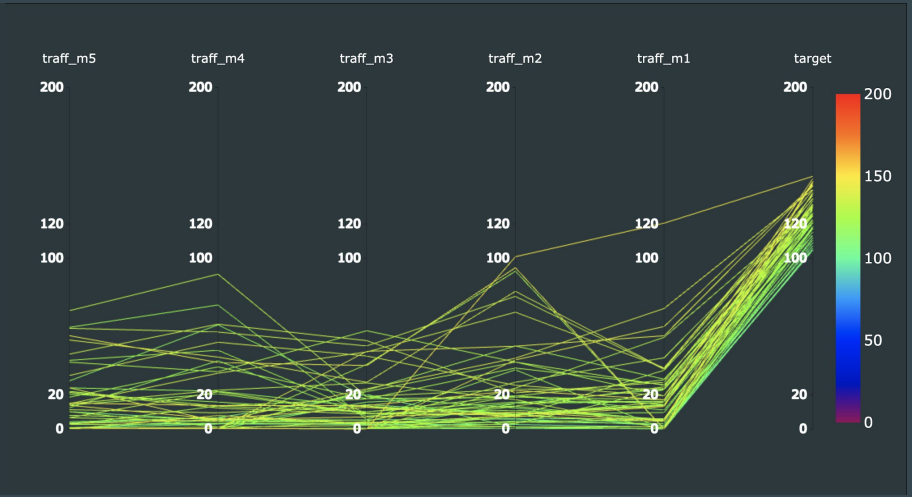
Residuals exploration:

BIG DATA LAB

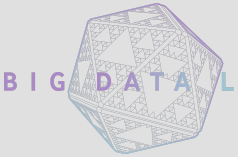
50 < residuals < 100



residuals > 100



vodafone

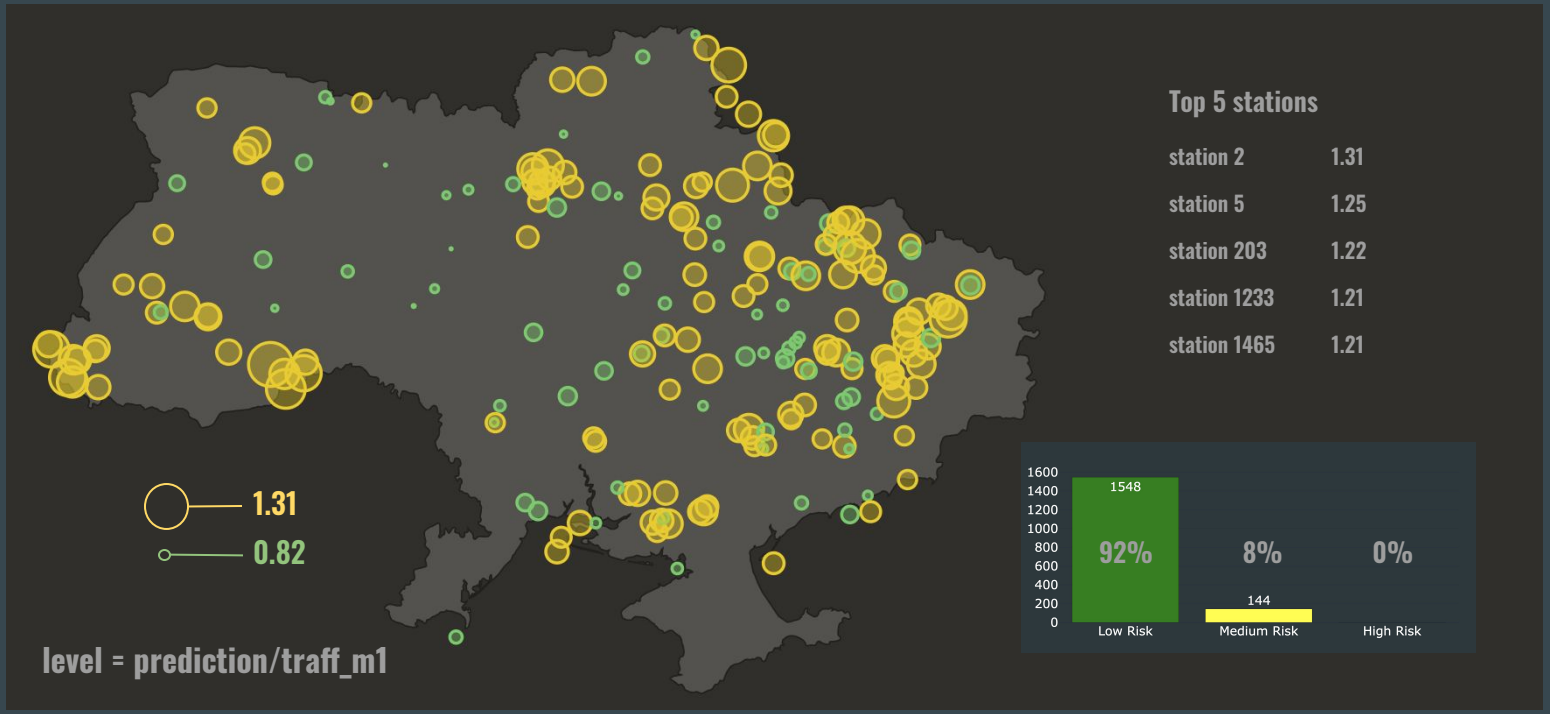


… and what about the business objective?

vodafone

Stations consumption level:

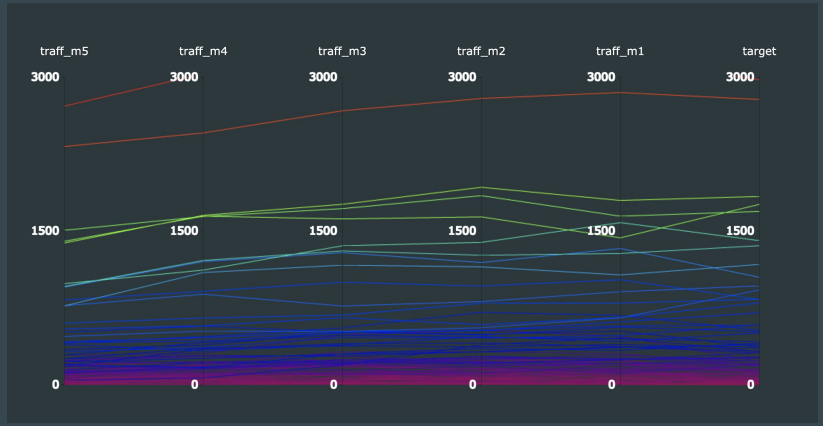
BIG DATA LAB

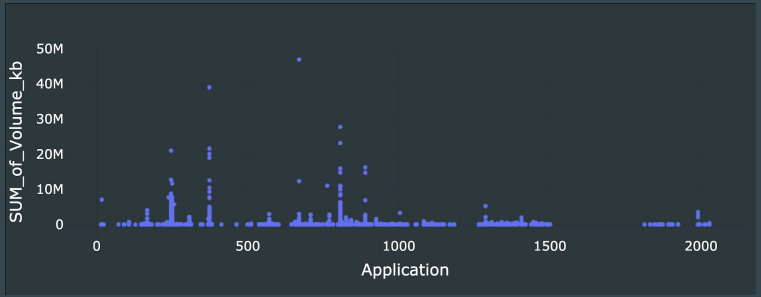


vodafone

Ideas for improvements:

- track
  - traffic directly on the base station
  - subscribers count trend for each station
- use more
  - traffic points (target - 12)
  - device data (applications)
  - subscribers related data (age)





Dima Mashchenko

Thank You

https://github.com/dmashchenko/vodafone-diploma

mashchenkod@gmail.com