Test task
ML modelling

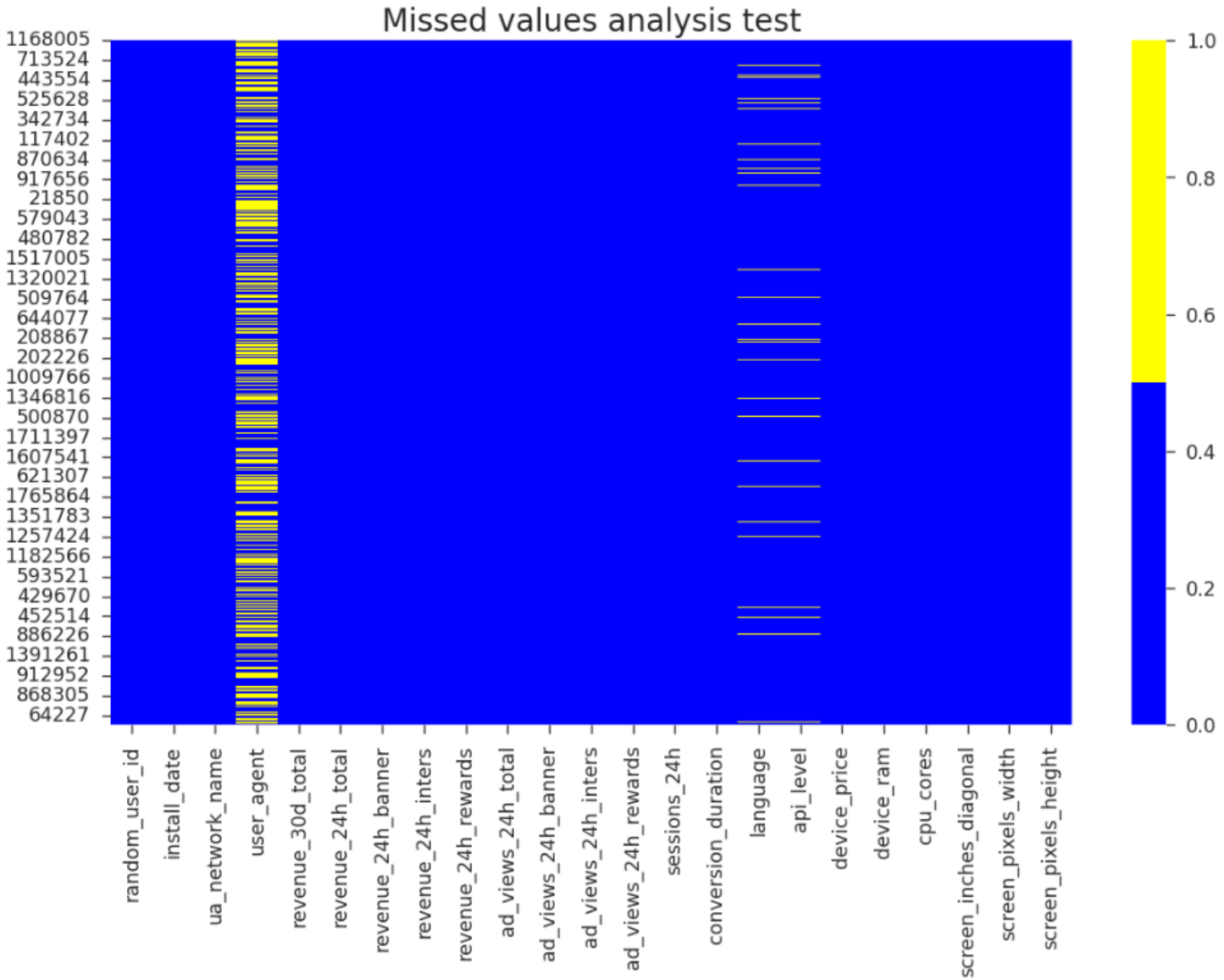
Timoshenko Gala

24.04.2023

Pipeline

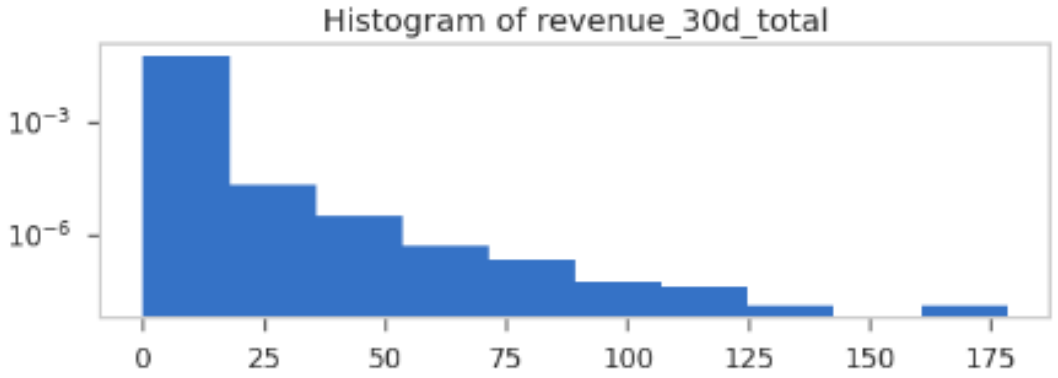
- Step1.ipynb. EDA(cleaning data from Null, outliers, duplicates, making graphs, using statistics & correlation)
- Step2.ipynb.Baseline model. Linear Regression.
- Step3.ipynb. LGBMRegression
- Step4. Choose the best model.

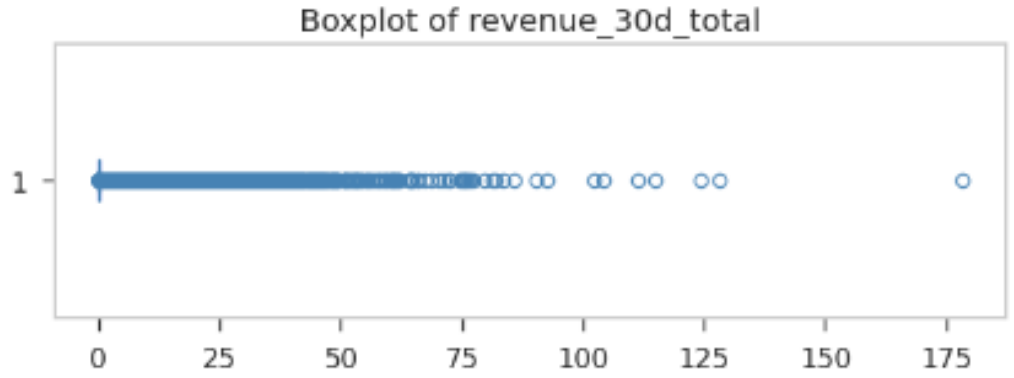
EDA

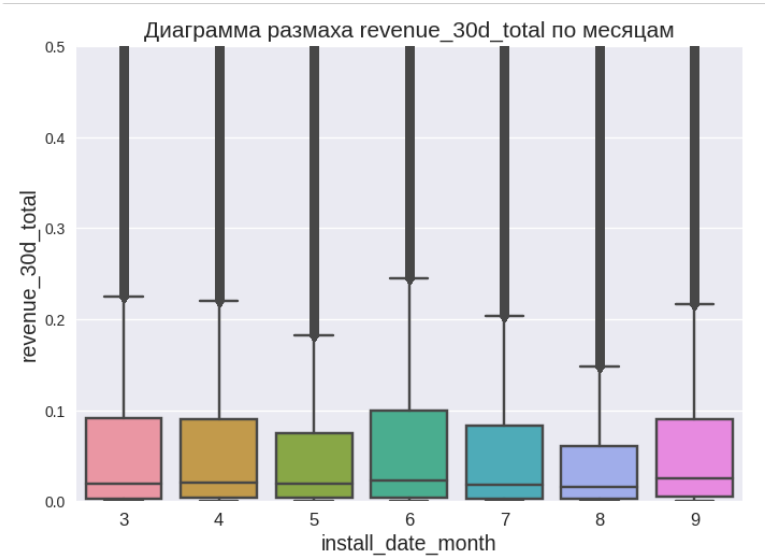


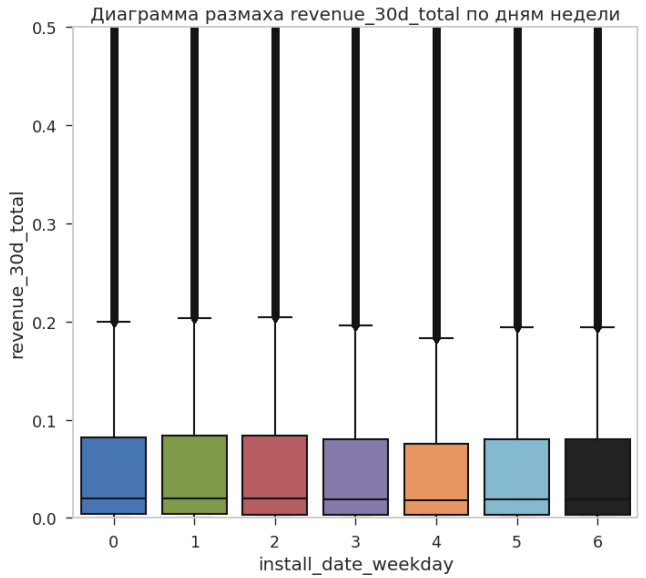
name	% miss values	actions
api_level	2.81	Fill top freq value (30)
language	2.81	Fill top freq value (en)
user_agent	40.77	New feature OS (Android11-top) Linux only
country_code	100	Drop column

Revenue_30d_total target column

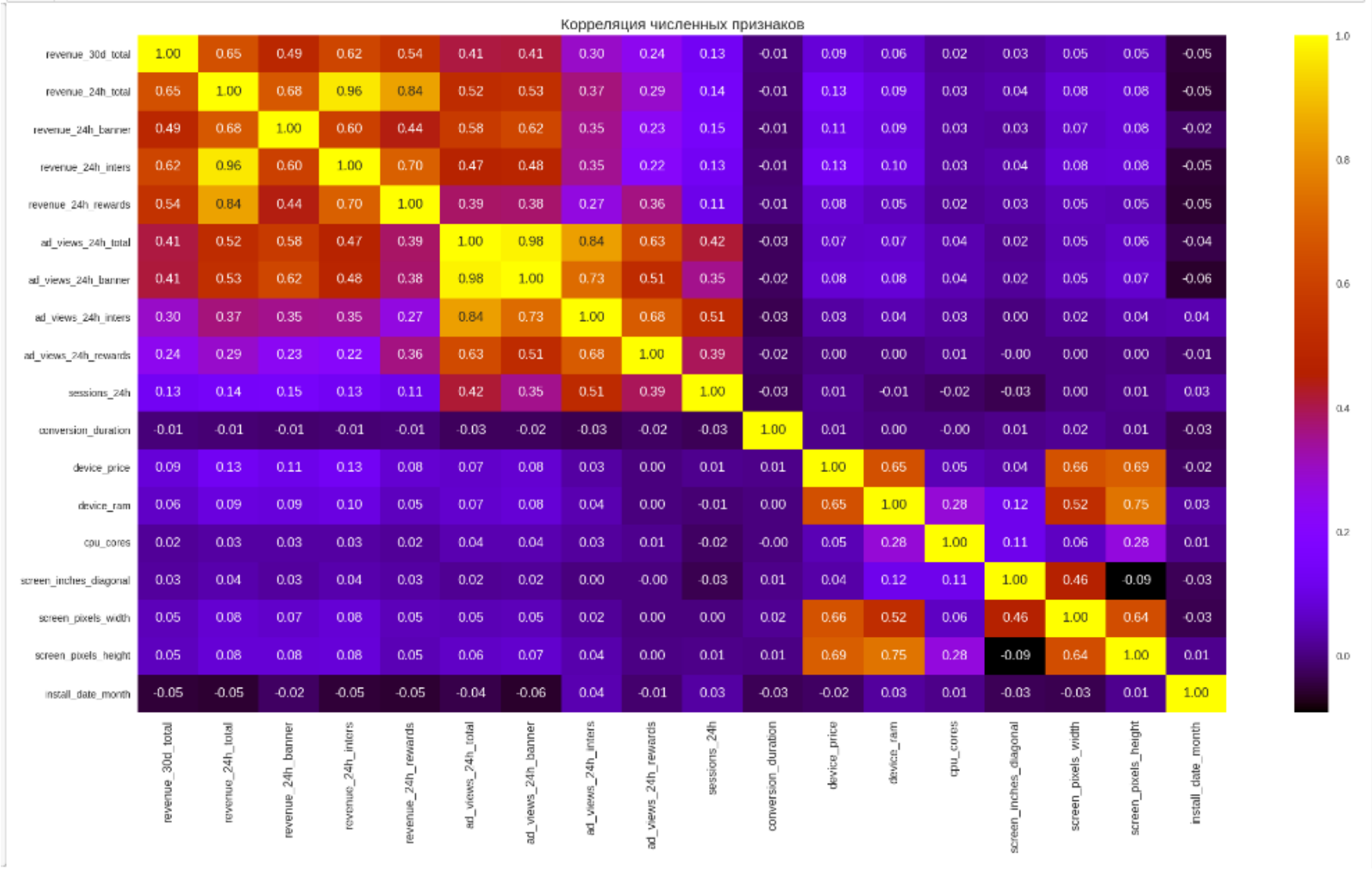








Correlation



Not bad linear correlation

- revenue_… and ad_views_…
- Device_ram, device_price, screen_height

Feature generation

1. install_date_month
2. date_day
3. user_agent - Android OS type
4. install_date_weekday
5. screen_inches_diagonal
6. screen_pixels_width
7. screen_pixels_height
8. revenue_30d_total_median_per_os
9. revenue_30d_total_median_per_lan
10. square_number columns
11. sqrt_number columns
12. log_number columns

Feature std_scaler

1. revenue_24h_rewards
2. revenue_24h_total
3. revenue_24h_banner
4. revenue_24h_inters
5. api_level
6. sessions_24h
7. screen_inches_diagonal
8. ad_views_24h_reward
9. ad_views_24h_total
10. device_price
11. screen_pixels_width
12. conversion_duration

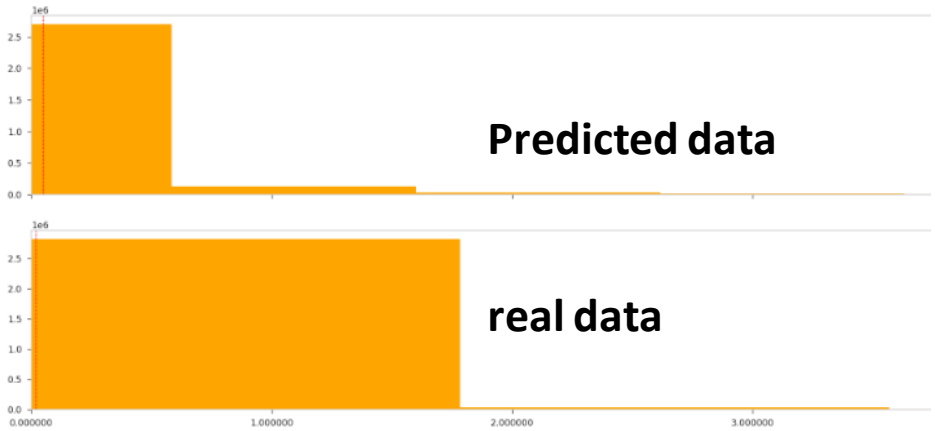
LinearRegression model

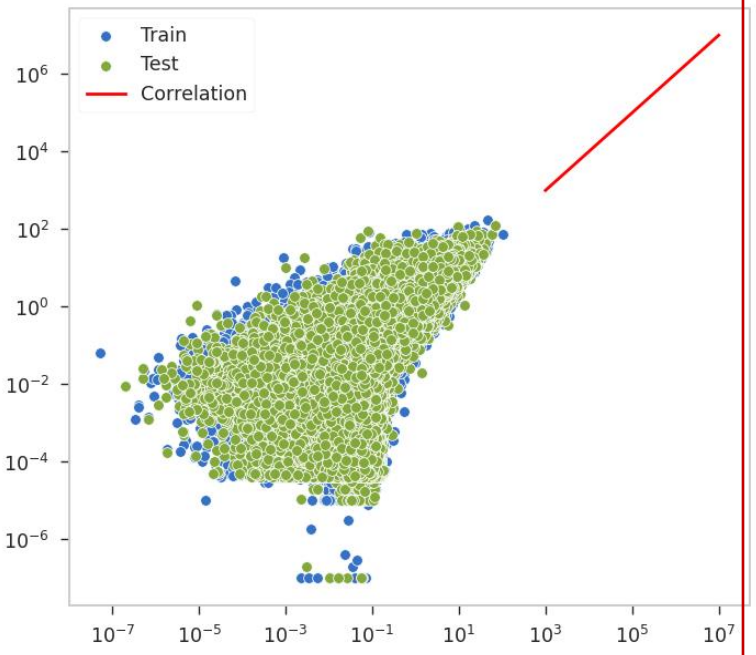
Train: 0.4348, Test: 0.4399
Crossval [0.4201; 0.4328; 0.4273; 0.4726]

Train.csv

[25]:

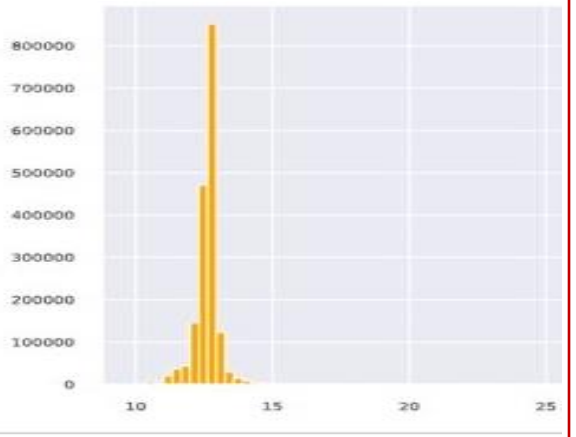
	revenue_30d_total	prediction	relev_error_formule	smape_error
count	2887999.00000	2887999.00000	2887999.00000	2887999.00000
mean	0.20299	0.17664	8.01041	1.50032
std	1.05468	0.61467	22.57462	0.60083
min	0.00000	-0.29877	-1.00000	0.00000
25%	0.00350	0.00673	-0.53174	1.10382
50%	0.01985	0.04782	0.00000	1.79120
75%	0.08597	0.12082	4.50689	2.00000
max	178.39592	110.38463	179.99970	2.00000





Test.csv

mean	12.63827
std	0.49833
min	10.39411
25%	12.49510
50%	12.68738
75%	12.78726
max	42.37554

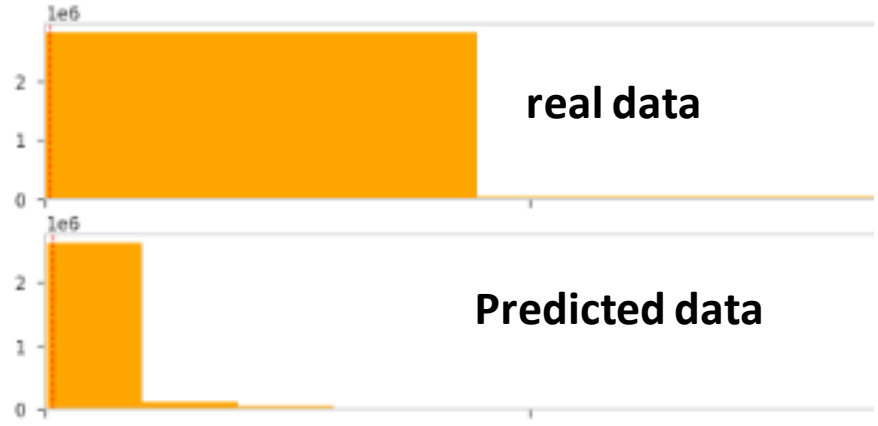


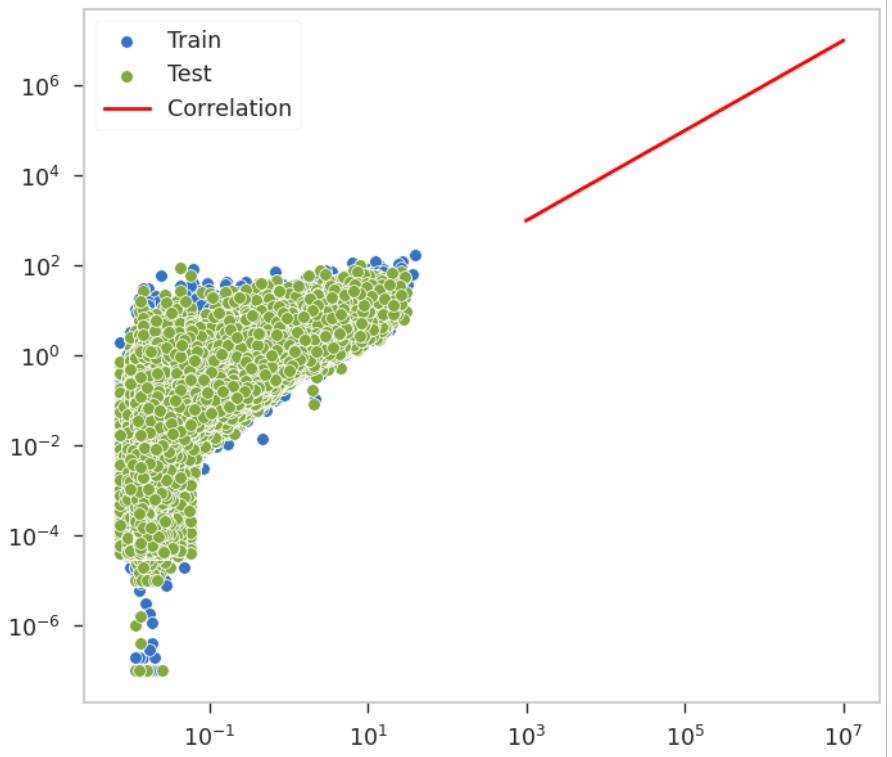
LGBMRegression model

train: 0.508991 test: 0.527053
train: 0.481179 test: 0.502642
train: 0.471994 test: 0.500058
train: 0.466292 test: 0.500101

Train.csv

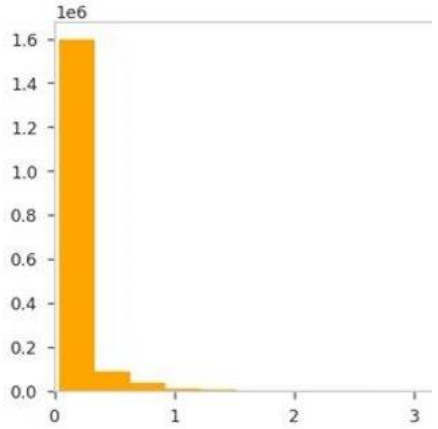
	revenue_30d_total	prediction	relev_error_formule	smape_error
count	2887999.00000	2887999.00000	2887999.00000	2887999.00000
mean	0.20299	0.17664	7.41275	1.31841
std	1.05468	0.60382	22.13789	0.62263
min	0.00000	0.00747	-0.99990	0.00000
25%	0.00350	0.01687	-0.57514	0.81149
50%	0.01985	0.03248	0.00000	1.48156
75%	0.08597	0.10371	3.54256	1.90616
max	178.39592	39.55039	179.99980	2.00000



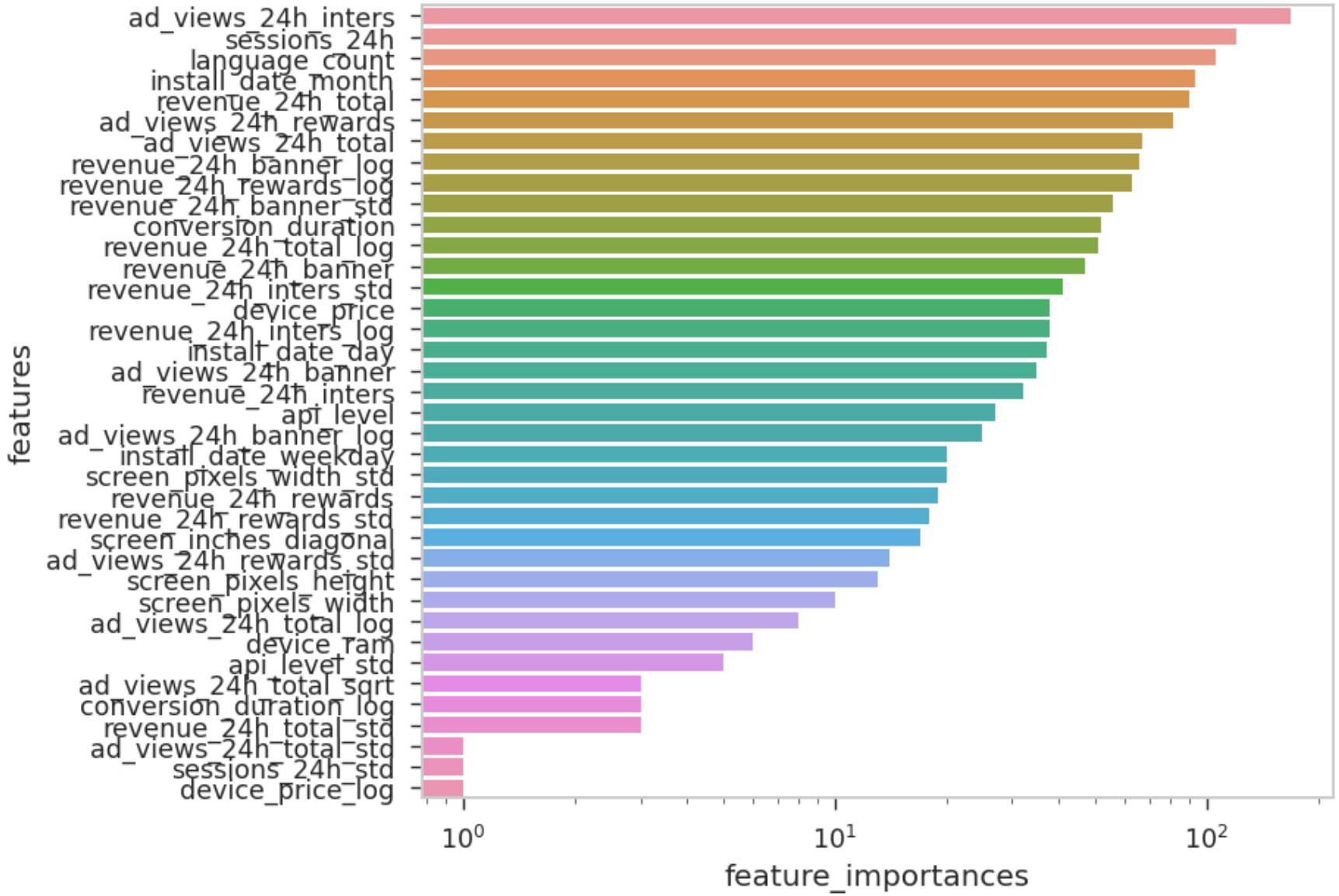


Test.csv

mean	0.17228
std	0.41612
min	0.03277
25%	0.06011
50%	0.07221
75%	0.13221
max	29.75212



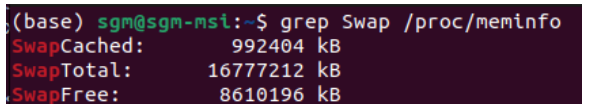
Feature importance



Choose the best model

Linear regression	LGBMRegression
bad in prediction 0 min-values	bad in prediction max-values
Set metric Rel_err can occur inf values (division to 0), better to choose SMAPE=0.6, Rel_err_std=20%	SMAPE=0.6, Rel_err_std=22%
Predicted range on test[min=10; max=42]	Predicted range on test[min=0.03; max=29]
Predicted range on train[min=0; max=110]	Predicted range on train[min=0.07; max=39]
Cross_val Train: 0.4348, Test: 0.4399	Cross_val train: 0.508991 test: 0.527053

- Choosing between Linear regression & LGBMRegression is hard, cause the value-range is diff in both case.
- I've tried to run XGBoost, CatBoost, MLP (NeuralNetwor) but Kernel was dead quickly, I swapped all the memory, but it didn't work



- If you could provide computing power I'll try extra runs
- Rel_err is smaller in Linear Regr, but I'm confused about the set range in test.csv
- So I suggest to choose LGBMRegression at the moment and continue on looking the best version