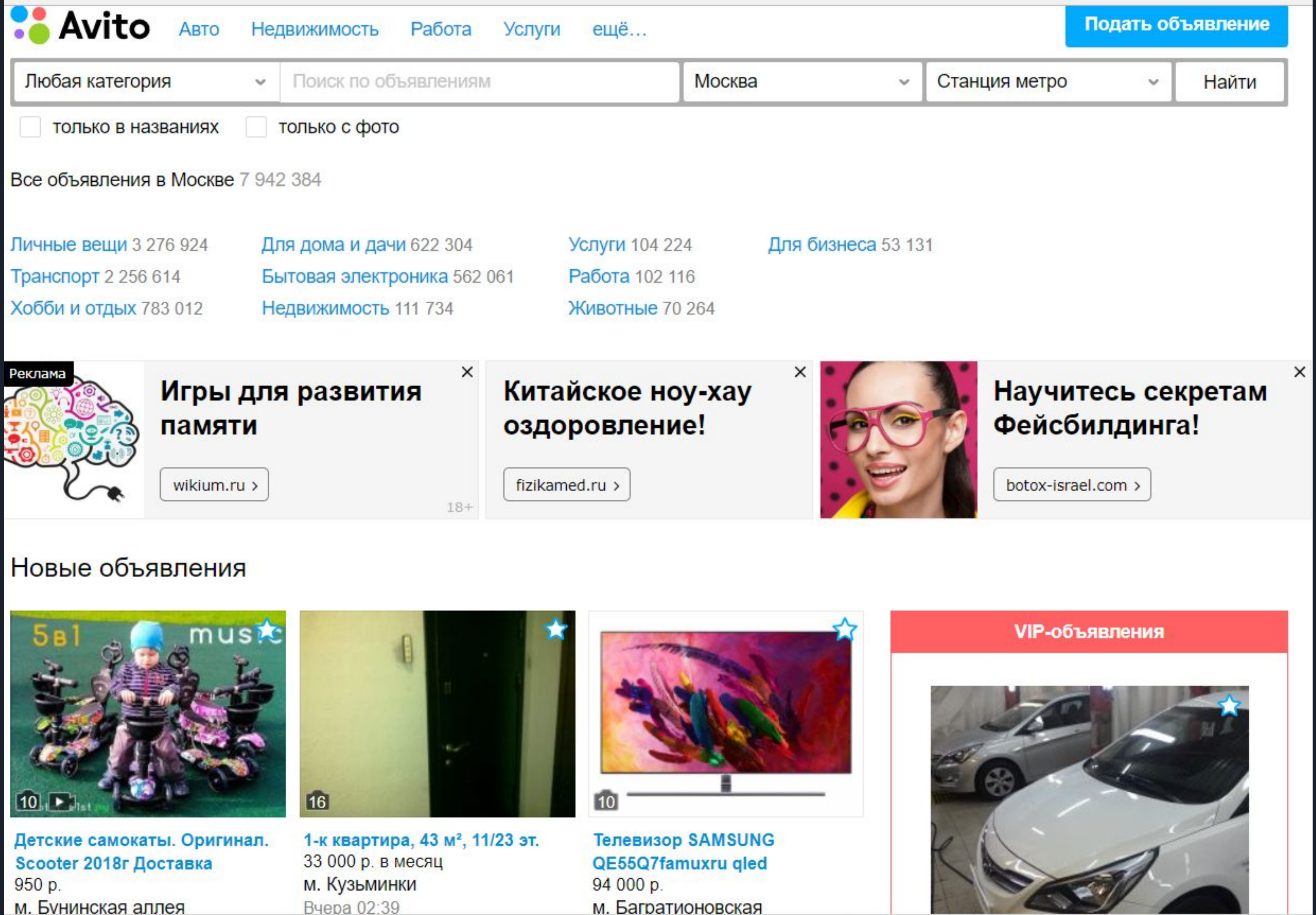
Avito: Retail website deal probability predictor

June 13th, 2018

Stephan, Matan, Itai and Ilai


Avito
Авто
Недвижимость
Работа
Услуги
ещё…
Подать объявление
Любая категория
Поиск по объявлениям
Москва
Станция метро
Найти
только в названиях
только с фото
Все объявления в Москве 7 942 384
Личные вещи 3 276 924
Для дома и дачи 622 304
Услуги 104 224
Для бизнеса 53 131
Транспорт 2 256 614
Бытовая электроника 562 061
Работа 102 116
Хобби и отдых 783 012
Недвижимость 111 734
Животные 70 264
Реклама
Игры для развития памяти
wikium.ru ›
18+
Китайское ноу-хау оздоровление!
fizikamed.ru ›
Научитесь секретам Фейсбилдинга!
botox-israel.com ›
Новые объявления
Детские самокаты. Оригинал. Scooter 2018г Доставка
950 р.
м. Бунинская аллея
1-к квартира, 43 м², 11/23 эт.
33 000 р. в месяц
м. Кузьминки
Вчера 02:39
Телевизор SAMSUNG QE55Q7famuxru qled
94 000 р.
м. Багратионовская
VIP-объявления

Given data - categorical/numerical features

	item_id	user_id	region	city
0	b912c3c6a6ad	e00f8ff2eaf9	Свердловская область	Екатеринбург
1	2dac0150717d	39aeb48f0017	Самарская область	Самара
2	ba83aefab5dc	91e2f88dd6e3	Ростовская область	Ростов-на-Дону
3	02996f1dd2ea	bf5cccea572d	Татарстан	Набережные Челны
4	7c90be56d2ab	ef50846afc0b	Волгоградская область	Волгоград
5	51e0962387f7	bbfad0b1ad0a	Татарстан	Чистополь

price	item_seq_number	activation_date	user_type
400.0	2	2017-03-28	Private
3000.0	19	2017-03-26	Private
4000.0	9	2017-03-20	Private
2200.0	286	2017-03-25	Company
40000.0	3	2017-03-16	Private
1300.0	9	2017-03-28	Private

Given data - Free text

category_name	param_1	param_2	param_3	title	description
аксессуары	одежда	джинсы	26	colins	хорошем состоянии.

Believable and Informative Description Copy

Description:
AMAZING WATCH FOR SALE!!!!

DON'T MISS THIS DEAL. IT'S THE DEAL OF THE CENTURY!!

Unlikely

Description:
I have an adjustable Chaleur D'Animale Watch for sale.

It's never been worn and still in the original box. Battery included.

Informative

Description:
fancy watch for sale

no low ball offers, cash and carry

Poor Quality

Given data - Images

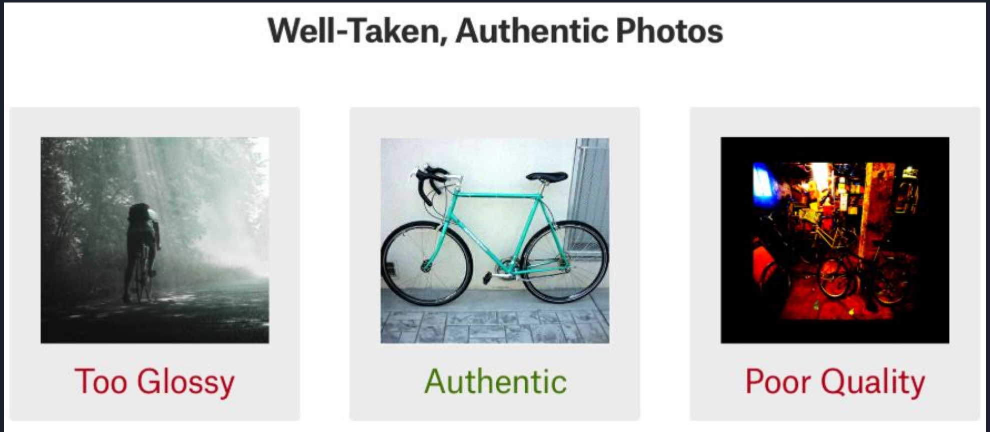


image	image_top_1
d10c7e016e03247a3bf2d13348fe959fe6f436c1caf64c...	1008.0
9c9392cc51a9c81c6eb91eceb8e552171db39d7142700...	692.0
b7f250ee3f39e1fedd77c141f273703f4a9be59db4b48a...	3032.0
6ef97e0725637ea84e3d203e82dadb43ed3cc0a1c8413...	796.0
54a687a3a0fc1d68aed99bdaaf551c5c70b761b16fd0a2...	2264.0
eb6ad1231c59d3dc7e4020e724ffe8e4d302023ddcbb99...	796.0
0330f6ac561f5db1fa8226dd5e7e127b5671d44d075a98...	2823.0
9bab29a519e81c14f4582024adfebd4f11a4ac71d323a6...	567.0
75ce06d1f939a31dfb2af8ac55f08fa998fa336d13ee05...	415.0
54fb8521135fda77a860bfd2fac6bf46867ab7c06796e3...	46.0

Given data - goal: predict **deal probability** based on the **ad parameters, text and images**

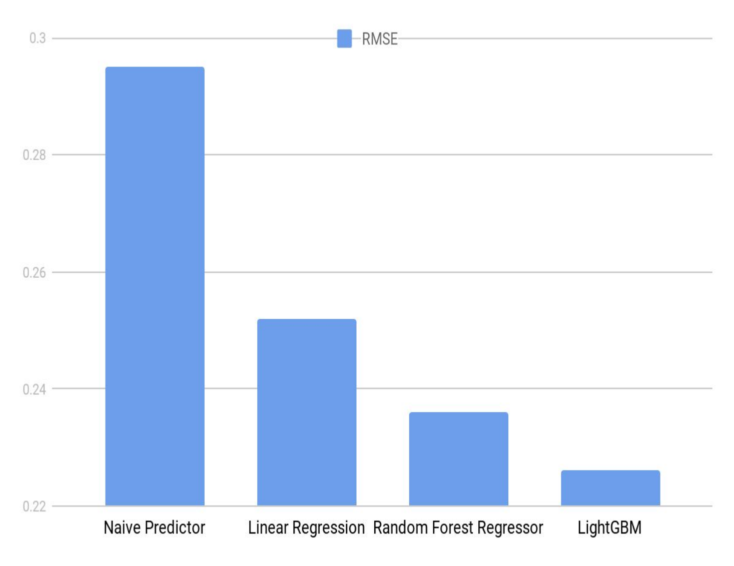
deal_probability
0.12789
0.00000
0.43177
0.80323
0.20797
0.80323
0.00000

Models and Evaluation - On validation set

Evaluation: minimize *RMSE*

Models:

- **Naive Random Prediction** - 0.295
- **Linear Regression** - 0.259
- **Random Forest Regressor** - 0.236
- **CatBoost** - 0.235
- **LightGBM** - 0.234



Competition - current state

1800 - 1250: 0.23XX

1200-1000: 0.225X

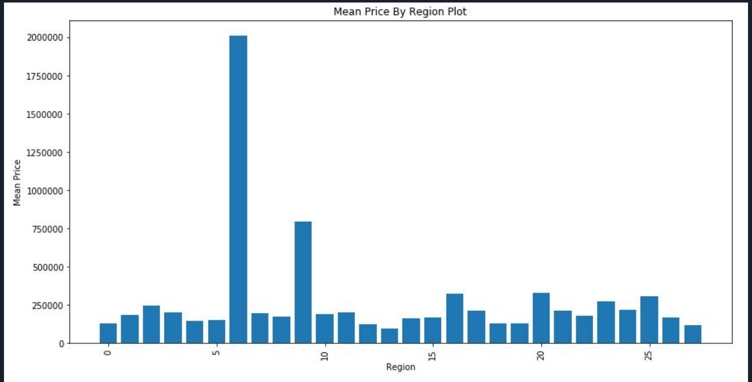
1000-100: 0.225X-0.220X

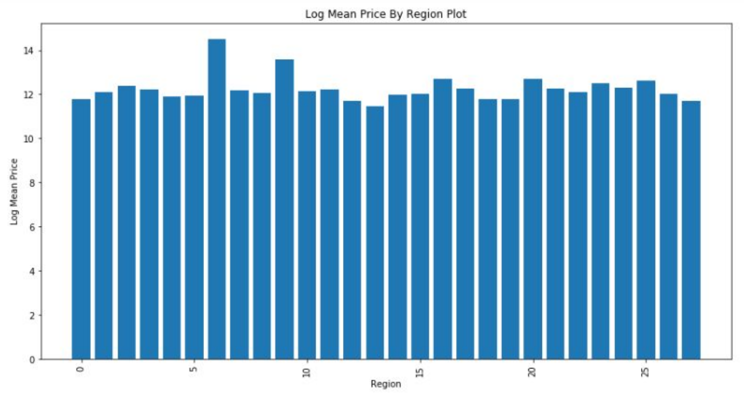
100-20: 0.2200-0.2180

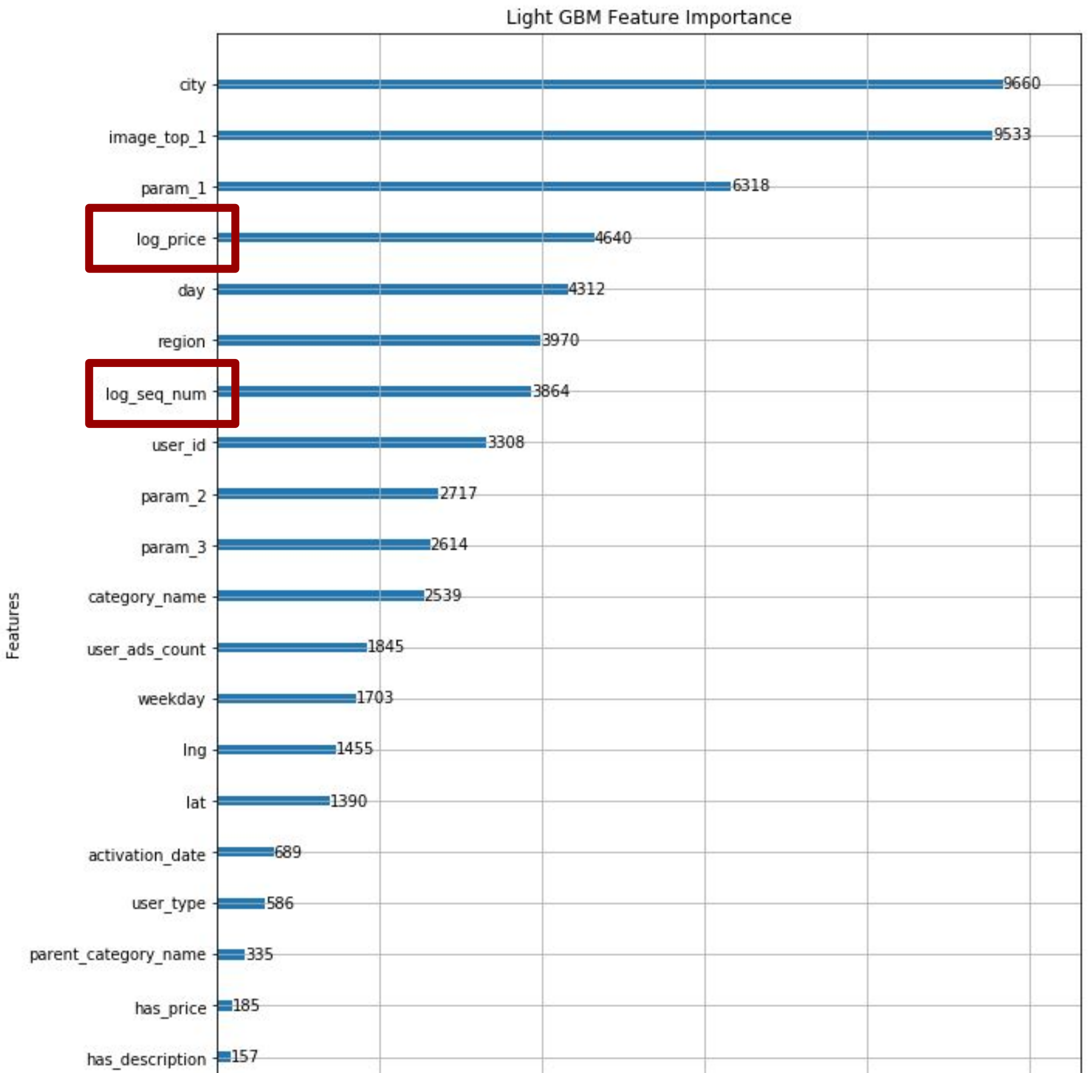
20-1: 0.2180-0.2122

LGBM - Basic feature enrichment

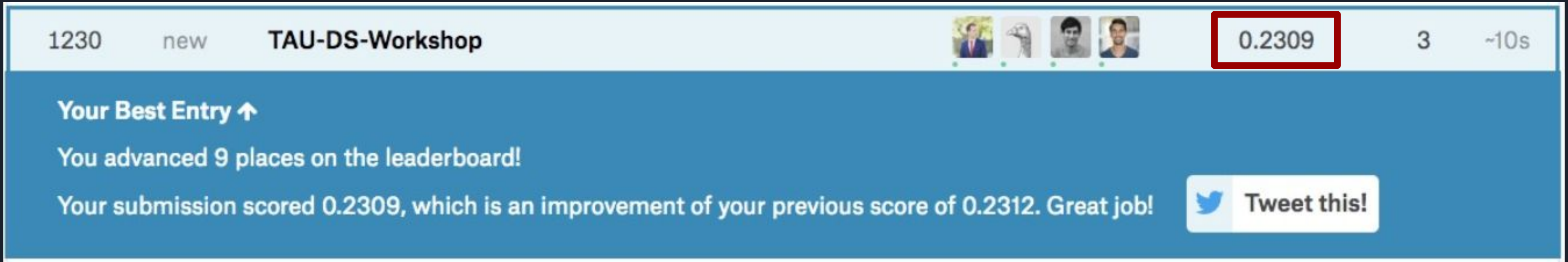
- Activation date → weekday, month, year.
- Log transform on
  - Price
  - Sequence number
- User ads count
- Add boolean features as “has_params”
- Data cleaning - replace missing price with median price, image_top_1 with -1 etc.







Light GBM Feature Importance
Features
city
9660
image_top_1
9533
param_1
6318
log_price
4640
day
4312
region
3970
log_seq_num
3864
user_id
3308
param_2
2717
param_3
2614
category_name
2539
user_ads_count
1845
weekday
1703
lng
1455
lat
1390
activation_date
689
user_type
586
parent_category_name
335
has_price
185
has_description
157



Out of 1800 participants (66% percentile)

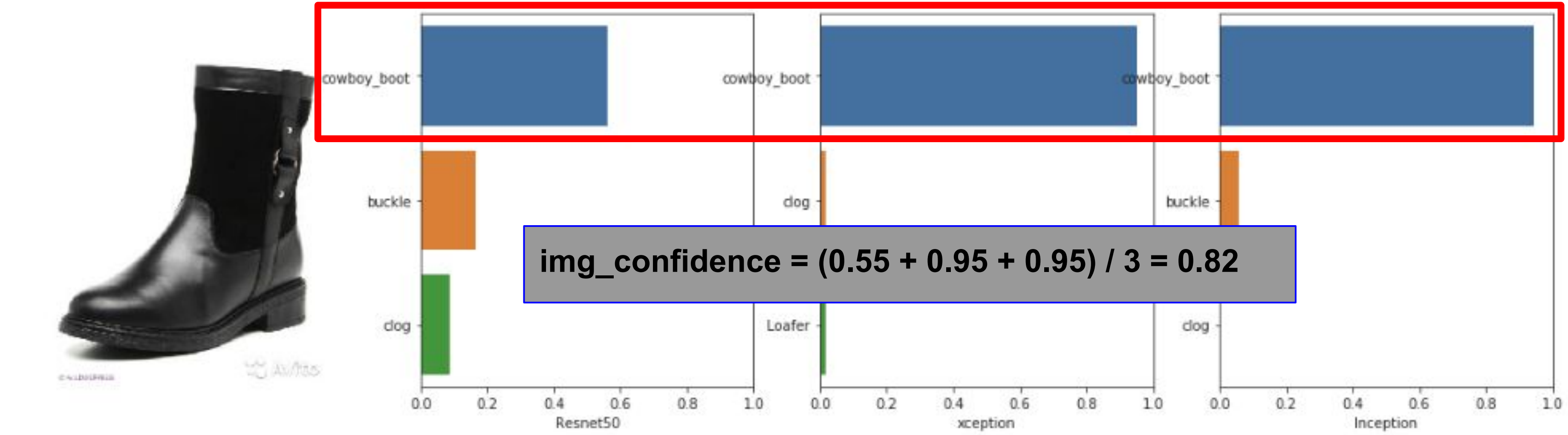
LGBM - Add image features

<u>Basic guideline:</u>

The quality of the ad image significantly affects the demand volume on an item

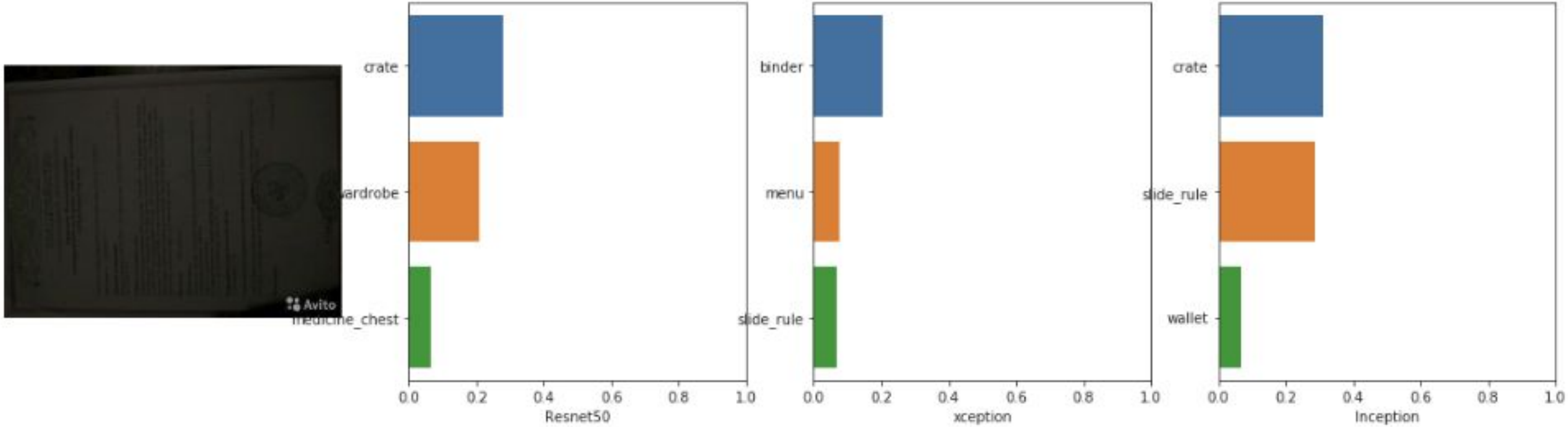
- Features: size, colorfulness, dominant color, average color
- Image Quality? sharpness, luminance
- Image “Confidence”: average of top-1 probability tags from three models (Resnet50, Xception, Inception), may work as a proxy for image quality

Feature eng. - Images Examples



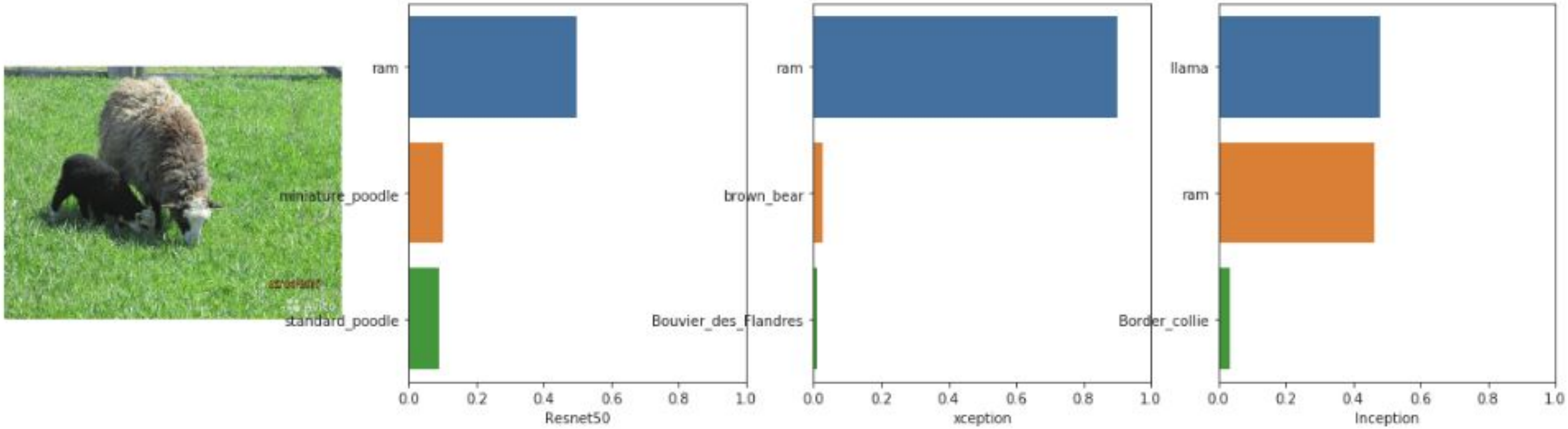
	img_size	img_sharpness	img_luminance	img_colorfulness	img_dominant_color	img_color_avg	img_blue_std	img_green_std	img_red_std
0	172800	360.226204	2969.34262	1.499964	[254, 254, 254]	[172, 172, 172]	103.977177	103.837343	103.825449

Feature eng. - Images Examples

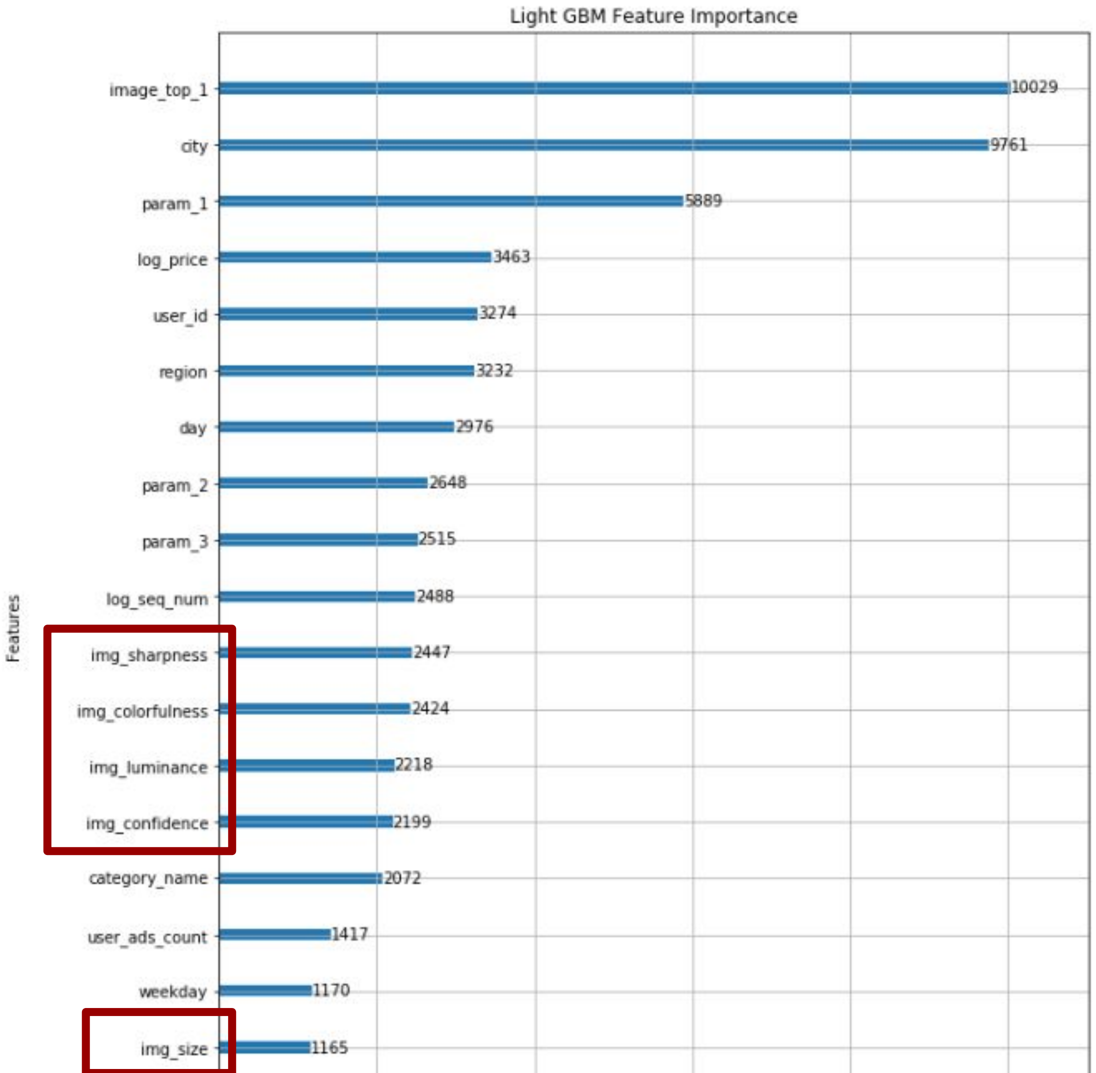


	img_size	img_sharpness	img_luminance	img_colorfulness	img_dominant_color	img_color_avg	img_blue_std	img_green_std	img_red_std
0	172800	65.503883	508.862134	4.474525	[31, 36, 39]	[24, 29, 32]	10.741228	10.866127	11.341529

Feature eng. - Images Examples



	img_size	img_sharpness	img_luminance	img_colorfulness	img_dominant_color	img_color_avg	img_blue_std	img_green_std	img_red_std
0	172800	1922.065405	2432.776622	48.306307	[108, 181, 140]	[90, 153, 117]	35.680399	49.146307	40.220842



Light GBM Feature Importance
Features
image_top_1
10029
city
9761
param_1
5889
log_price
3463
user_id
3274
region
3232
day
2976
param_2
2648
param_3
2515
log_seq_num
2488
img_sharpness
2447
img_colorfulness
2424
img_luminance
2218
img_confidence
2199
category_name
2072
user_ads_count
1417
weekday
1170
img_size
1165

LGBM - Add text features

	price	item_seq_number	image_top_1	deal_probability	title_word_count	description_word_count	merged_params_word
price	1.000000	0.061099	0.035071	-0.010853	0.065333	0.049828	-0.
item_seq_number	0.061099	1.000000	0.093324	-0.036068	0.132158	0.120281	-0.
image_top_1	0.035071	0.093324	1.000000	0.188871	0.247162	0.183789	-0.
deal_probability	-0.010853	-0.036068	0.188871	1.000000	0.017285	-0.001158	-0
title_word_count	0.065333	0.132158	0.247162	0.017285	1.000000	0.308280	-0.
description_word_count	0.049828	0.120281	0.183789	-0.001158	0.308280	1.000000	-0.
merged_params_word_count	-0.020851	-0.056616	-0.557352	-0.116995	-0.166665	-0.162763	1.
description_sentence_count	0.028791	0.125823	0.167183	-0.016130	0.251324	0.854852	-0.
description_words/sentence_ratio	0.017774	-0.008236	0.045569	0.050242	0.088518	0.092528	-0.
title_capital_letters_ratio	-0.033504	-0.055070	0.128419	0.021500	-0.303308	-0.010537	-0.
description_capital_letters_ratio	-0.002779	0.024389	0.087672	0.002238	0.029260	0.038988	-0.
title_non_regular_chars_ratio	0.099775	0.191032	0.189312	0.022199	0.434856	0.183368	-0.
description_non_regular_chars_ratio	0.005346	0.089587	0.083366	-0.011636	0.141690	0.291877	-0.
title_num_of_newrow_char	NaN	NaN	NaN	NaN	NaN	NaN	
description_num_of_newrow_char	0.011555	0.106225	0.159598	-0.024514	0.197092	0.755349	-0.
title_num_adj	0.006099	0.027831	-0.077893	-0.044298	0.291463	0.079471	0.
title_num_nouns	0.044343	0.107833	0.257261	0.014511	0.817634	0.298755	-0.
title_adj_to_len_ration	-0.013958	-0.019843	-0.152274	-0.049400	-0.019432	-0.024772	0.
title_noun_to_len_ration	-0.025251	-0.037511	-0.025477	-0.017093	-0.356624	-0.057076	0.
description_num_adj	0.066113	0.124427	0.137847	-0.001019	0.300760	0.898495	-0.
description_num_nouns	0.050675	0.120833	0.199486	0.004715	0.312669	0.970856	-0.
description_adj_to_len_ration	0.000882	-0.018710	-0.137510	-0.040210	-0.078614	-0.127597	0
description_noun_to_len_ration	0.006111	0.005324	0.061534	0.024627	0.029115	-0.047335	-0.
title_sentiment	-0.010653	0.017676	0.006652	-0.006089	0.070308	0.025541	-0.
description_sentiment	-0.011110	0.031508	0.006608	-0.020147	0.022263	0.104806	-0.

Feature eng. - Text, POS tagging

tagged_title	tagged_description	title_num_adj	title_num_nouns	title_adj_to_len_ration	title_noun_to_len_ration
[(Кокоби, S), ((, NONLEX), (кокон, S), (для, P...	[(Кокон, S), (для, PR), (сна, S), (малыша, S),...	0	3	0.0	1.000000
[(Стойка, S), (для, PR), (Одежды, S)]	[(Стойка, S), (для, PR), (одежды, S), (,, NONL...	0	2	0.0	0.666667
[(Philips, NONLEX), (bluray, NONLEX)]	[(В, PR), (хорошем, A=n), (состоянии, S), (,, ...	0	2	0.0	1.000000
[(Автокресло, S)]	[(Продам, V), (кресло, S), (от0-25кг, S)]	0	1	0.0	
			1	0.0	

description_adj_to_len_ration	description_noun_to_len_ration
0.142857	0.571429
0.000000	0.571429
0.117647	0.470588
0.000000	0.666667
0.000000	0.500000

The russian tagger rags sentences using the Russian National Corpus tagset:

http://www.ruscorpora.ru/en/corpora-morph.html

Here are some of the most important tags:

- S – noun
- A – adjective
- NUM – numeral
- A-NUM – numeral adjective
- V – verb
- ADV – adverb

Feature eng. - Text, Sentiment analysis

title_sentiment	description_sentiment
150000.000000	138477.000000
0.013922	0.196082
0.189638	0.558819
-1.000000	-1.000000
0.000000	0.000000
0.000000	0.000000
0.000000	1.000000
1.000000	1.000000

title_sentiment	description_sentiment
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	-1.0

Feature eng. - Free text - Significance

	price	item_seq_number	image_top_1	deal_probability	title_word_count	description_word_count	merged_params_worc
price	1.000000	0.061099	0.035071	-0.010853	0.065333	0.049828	-0.
item_seq_number	0.061099	1.000000	0.093324	[illegible]	[illegible]	[illegible]	-0.
image_top_1	0.035071	0.093324	1.000000	0.188871	0.247162	0.183789	-0.
deal_probability	-0.010853	-0.036068	0.188871	[illegible]	[illegible]	[illegible]	-0
title_word_count	[illegible]	0.132158	0.247162	0.017285	1.000000	0.308280	-0.
description_word_count	[illegible]	[illegible]	0.183789	-0.001158	0.308280	1.000000	-0.
merged_params_word_count	[illegible]	[illegible]	-0.557352	-0.116995	-0.166665	-0.162763	1.
description_sentence_count	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.854852	-0.
description_words/sentence_ratio	[illegible]	[illegible]	0.045569	0.050242	0.088518	0.092528	-0.
title_capital_letters_ratio	[illegible]	[illegible]	0.128419	0.021500	-0.303308	-0.010537	-0.
description_capital_letters_ratio	[illegible]	[illegible]	0.087672	0.002238	0.029260	0.038988	-0.
title_non_regular_chars_ratio	0.099775	0.191032	0.189312	0.022199	0.434856	0.183368	-0.
description_non_regular_chars_ratio	0.005346	0.089587	0.083366	-0.011636	0.141690	0.291877	-0.
title_num_of_newrow_char	NaN	NaN	NaN	NaN	NaN	NaN	
description_num_of_newrow_char	0.011555	0.106225	0.159598	-0.024514	0.197092	0.755349	-0.
title_num_adj	0.006099	0.027831	-0.077893	-0.044298	0.291463	0.079471	0.
title_num_nouns	[illegible]	[illegible]	0.257261	0.014511	0.817634	0.298755	-0.
title_adj_to_len_ration	[illegible]	[illegible]	-0.152274	-0.049400	-0.019432	-0.024772	0.
title_noun_to_len_ration	[illegible]	[illegible]	[illegible]	[illegible]	-0.356624	-0.057076	0.
description_num_adj	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.898495	-0.
description_num_nouns	[illegible]	[illegible]	0.199486	0.004715	0.312669	0.970856	-0.
description_adj_to_len_ration	[illegible]	[illegible]	-0.137510	-0.040210	-0.078614	-0.127597	0
description_noun_to_len_ration	0.006111	0.005324	0.061534	0.024627	0.029115	-0.047335	-0.
title_sentiment	-0.010653	0.017676	0.006652	-0.006089	0.070308	0.025541	-0.
description_sentiment	-0.011110	0.031508	0.006608	-0.020147	0.022263	0.104806	-0.

Counts, ratios

POS tagging

Feature eng. - Text, aggregative: TF-IDF

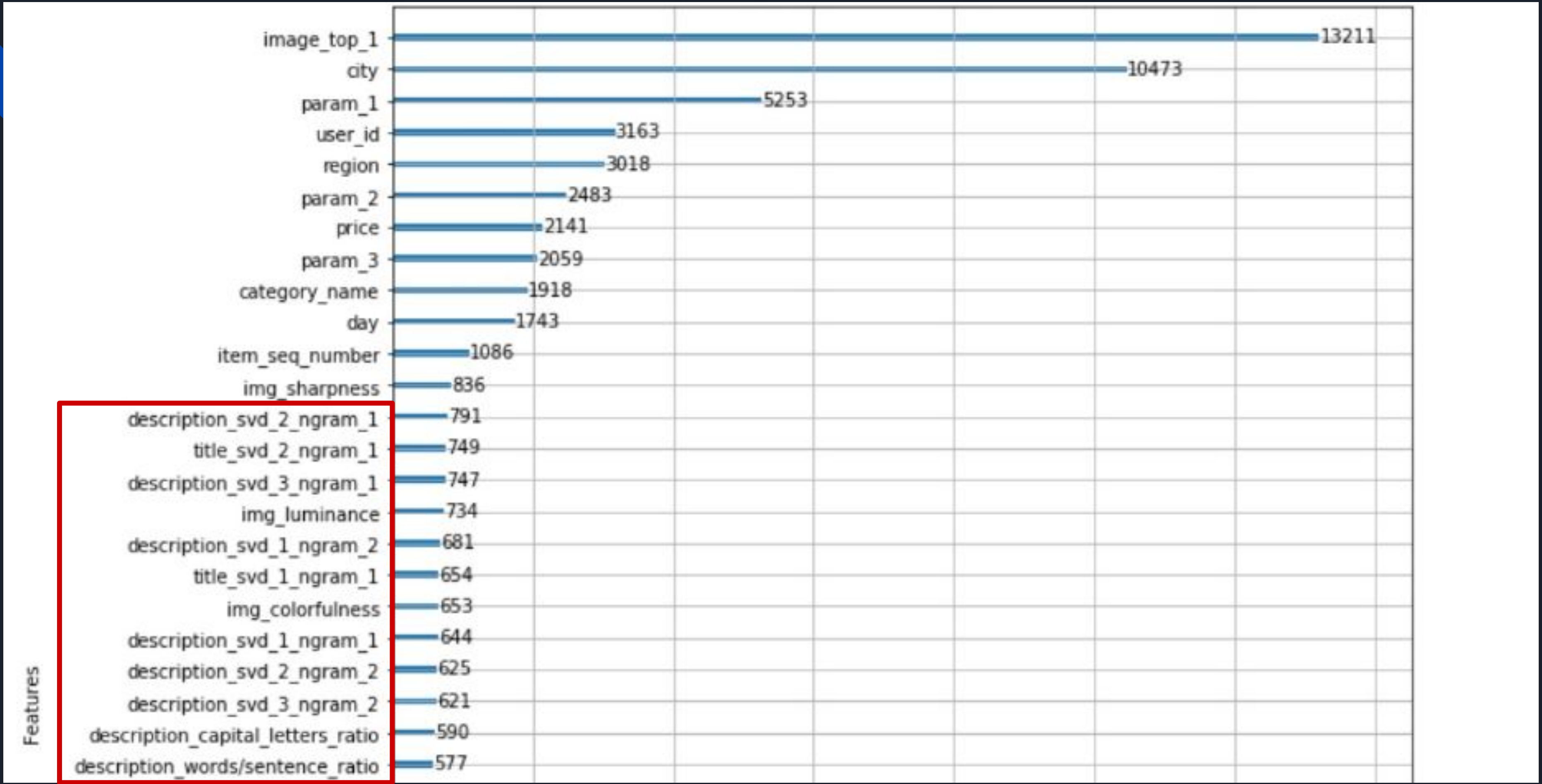
1. How important a word is to a document in a collection or corpus?
2. CountVectorizer (Term Frequency) → Inverse Document Frequency transform.

```
def tfidf_main(train_df, test_df, col_name, n_comp):
    ### TFIDF Vectorizer ###
    tfidf_vec = TfidfVectorizer()
    full_tfidf = tfidf_vec.fit_transform(train_df[col_name].values.tolist() + test_df[col_name].values.tolist())

    ### SVD Components ###
    svd_obj = TruncatedSVD(n_components=n_comp, algorithm='arpack')
    svd_obj.fit(full_tfidf)

    # Train
    train_tfidf = tfidf_vec.transform(train_df[col_name].values.tolist())
    train_svd = pd.DataFrame(svd_obj.transform(train_tfidf))
    train_svd.columns = ['%s_svd_%s_ngram' % (col_name, i+1) for i in range(n_comp)]
    train_df = pd.concat([train_df, train_svd], axis=1)
```

LSA. Like PCA, no normalization (works with sparse matrices)


image_top_1
13211
city
10473
param_1
5253
user_id
3163
region
3018
param_2
2483
price
2141
param_3
2059
category_name
1918
day
1743
item_seq_number
1086
img_sharpness
836
description_svd_2_ngram_1
791
title_svd_2_ngram_1
749
description_svd_3_ngram_1
747
img_luminance
734
description_svd_1_ngram_2
681
title_svd_1_ngram_1
654
img_colorfulness
653
description_svd_1_ngram_1
644
description_svd_2_ngram_2
625
description_svd_3_ngram_2
621
description_capital_letters_ratio
590
description_words/sentence_ratio
577
Features

Add Aggregative features: Users and Dates

aggregated_features.csv

Download

aggregated_features.csv

0 B

Preview (first 100 rows)

user_id	avg_days_up_user	avg_times_up_user	n_user_items
00000077ff21	12.5	2.0	2
000006497719	19.0	2.0	1
00000b4d72f6	3.0	1.0	1
00000d642d7e	13.0	1.0	2

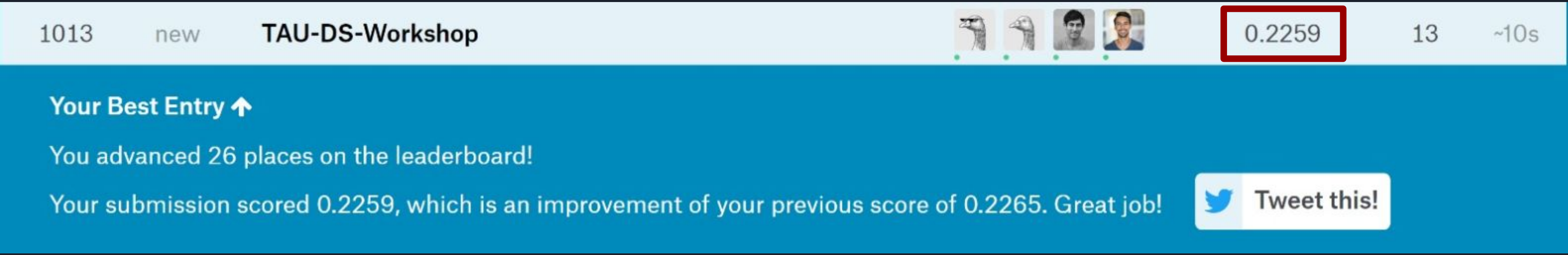
All data sets (train+test+active+periods - duplications) → aggregated activation dates per user.

Feature selection

- Eliminating highly correlated features:
    - Text features
    - Dates
    - Geo
    - Categorical features and has_x
- User_id

Models and Evaluation - LightGBM

- Grid search on hyperparameters:
  - boosting type: **gbdt** (Gradient Boosting Decision Tree), rf (Random Forest), dart (Dropouts meet Multiple Additive Regression Trees)
  - learning_rate: 0.03, **0.05**, 0.07, 0.1
  - num_leaves: 12, 16, 32, **64** (max leaves in a single tree)
  - feature_fraction: 0.5, **0.9**, 1 (randomly select part of features on each iteration)
- Cross validation (KFold)



Out of 1800 participants (45% percentile)

Competition - current state

1800 - 1250: 0.4 - 0.2300

1250-1000: 0.225X

Feature engineering
Hyperparameter Tuning

1000-100: 0.225X-0.220X

Better feature selection
Better hyperparameter tuning
Other Model: NN

100-20: 0.2200-0.2180

20-1: 0.2180-0.2122

Other Model: NN

Models - CatBoost

CatBoost is an open-source gradient boosting on decision trees library with categorical features support out of the box for Python, R (https://catboost.yandex)

- Very similar to LightGBM, we tried to tune the parameters similarly to LightGBM but with inferior results. It is also **much** slower. Results at the time:
  - LightGBM: 0.22534
  - CatBoost: 0.22585

Models - NN

- We’ve created a basic NN:
    - Vanilla MLP: 100K X 256 X 32 X 1, sigmoid activation.
    - Vectorized text features
    - DictVectorized all other (categorical features)
    - Have not run on whole data yet (problem feeding sparse matrices to Input layer)

Next milestones

- Input layer: feature encodings.
    - Better encode cat features (onehot encodings? Keras embeddings?)
    - Better encode continuous features
    - Better encode text features (Add TF-IDF encodings or LSTM language model hidden state as a feature)
- **Ensemble Neural nets (Different losses, Different inputs)**
    - Ensemble with Lgbm and neural nets?