Booking Prediction in Airbnb

1. Summary

This challenge is to predict whether a listing in Airbnb will be booked in the future. It is a binary classification problem based on a data set of 180 thousand samples, each of which consists of 45 features and the booking label (0 or 1). Finally the factorization machine model achieves rather good result (AUC = 0.862 in the testing data set). Building model consists of three main steps: first to analyze the data set, I used pandas and hive to get all the information of each feature, second to do the feature selection and extraction, after that I got 10 thousand features including the listing ID, finally to train the model and evaluate the performance of the model. The FM model is used widely in our company. I also tried the LIBFM with the default parameters and the AUC is 0.831.

2. Data Analyzing / Cleansing

Data analyzing helped me to understand the data set, which was extremely important for feature selection and extraction. I used pandas and hive to do the analyzing, which consisted of the percentage of missing data, the histogram distribution of each feature, statistics information of each feature, etc. Then I decided how to tackle with missing features and outlier feature values.
I use analysis.py and feature_select.hive to analyze feature.

Related Files Description

tmp / feature_miss_percentage.txt	==>	missing feature information
tmp / hist_pic	==>	histogram of each feature
tmp / feature_statistics.txt	==>	statistic information
tmp / feature_detail.txt	==>	the number and ratio of positive samples group by each feature value (do the discretization if necessary)

3. Feature Selection / Extraction

I reviewed those features one by one and selected 20 features (accessory 1) which were relevant to the booking label (dim_is_requested). As can be seen in the file named feature_detail.txt, The relation between the booking label (dim_is_requested) and the feature is high if the booking ratio changed a lot with the feature value, otherwise, the feature is useless and should be ignored.
The process of feature extraction is shown in the script named fm_feature_process.py.

4. Model Training / Evaluating

I used two versions of Factorization Machine, one of them were developed by our company and the other were LIBFM by National Taiwan University. Usually I prefer to use Factorization Machine than Logistic Regression, especially when I want to build the model quickly, because LR usually needs more feature combination. I evaluated the model by AUC (accessory 2), and I would select more features (sometimes maybe less feature or change the way of feature processing) including complex combined features, train the model, and evaluate it iteration by iteration, if it didn't perform good enough.

5. Code Description / Dependencies

analysis.py use pandas to get all the information and write into tmp
booking_predict_original_dataset.hive create hive table
feature_select.hive calculate the counting of all samples, positive samples, positive ratio group by each feature value

fm_feature_process.py fist to tackle with the missing feature and some special feature, second do the feature conversion, third to ignore those outlier value, finally to hash the feature to a number

generate_signature.py hash algorithm

train_fm / run.sh first step to train the model, second step to predict the testing data set by the model, finally evaluate the model by AUC

train_fm / train_fm.sh train the model by using Train-1.0-SNAPSHOT-jar-with-dependencies.jar, use config.xml to tune parameters and to add or reduce a feature

train_fm / predict.sh input the data set and predict the label by the specific model

train_fm / auc.sh calculate AUC

spark
hive
pandas
Train-1.0-SNAPSHOT-jar-with-dependencies.jar FM model developed by our company

6. Market Recommendation

A. Model Applying

- Use the booking predict model to rank the listing in the search page of airbnb, and it can contributes to the rise of GMV, but more attention should be paid to the diversity of the page and cold start problem, to prevent that the hot listings dominate.

- In the similar listing recommendation page, consider both the similarity and the booking prediction to give a proper ranking.

B. Findings

- The most useful features are those that are relevant to the popularity of the listing. The top listings get a lot of impressions on airbnb, which makes them having good transaction, and the impression and transaction create a snowball effect. I tend to pay more attention to those popular listings when I use airbnb and I prefer to choose those listings with a lot of comments and orders, which give me more information. The following are the features that mentioned above. 【m_total_overall_rating, m_reviews, m_checkouts, occ_occupancy_trailing_90_ds, p2_p3_click_through_score, listing_m_listing_views_2_6_ds_night_decay, r_kdt_listing_views_0_6_avg_n100, days_since_last_booking】

- Then it comes to those features like price, Wi-Fi, pictures, location and so on, and they are very relevant to the transaction too. The booking probability varies with the latitude and longitude of the house, but this feature is not used in the current version of model, because the feature needs some more detailed work (like feature cross). Here are the features, 【m_effective_daily_price, image_quality_score, dim_has_wireless_internet, price_booked_most_recent, dim_lat, dim_lng】

- The room size and person capacity are not that important 【dim_room_type, dim_person_capacity】. It is not true that “more pictures, better transaction”, and some house with expensive clean fee have quite good performance 【m_professional_pictures, m_pricing_cleaning_fee】.

- The amount of transaction varies a lot with time. 【ds_night_day_of_week, ds_night_day_of_year】

- The transaction does not rise with the ascent of users of searching, it has a drop in the middle and I have not found the proper explanation. 【general_market_m_unique_searchers_0_6_ds_night, general_market_m_contacts_0_6_ds_night, general_market_m_reservation_requests_0_6_ds_night】

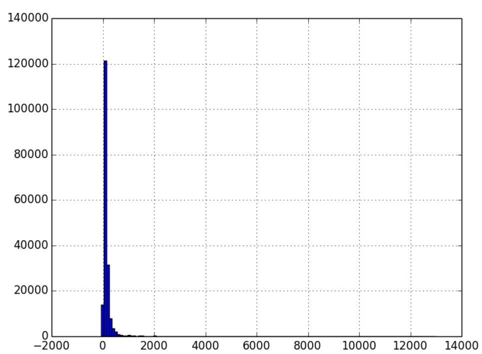
C. Thoughts

- In my opinion, airbnb is more like an assistant to those users who want to have a fantastic trip. Instead of recommending any listings that users might be interested in, it is more important to catch the users' eyes in time and to meet their needs.

Accessory 1

The following is the analysis of each feature. The number represents the feature index and the picture is the histogram and discrete feature distribution on positive and negative samples after feature processing.

5. m_effective_daily_price
Data missing: 0.0%
Abnormal data: negative number, too large number and too small number
Feature processing: take log(x + 3)



NULL	8	8	1.0
-2	2	2	1.0
0	3	3	1.0
1	81	8	0.09876543209876543
2	17	16	0.9411764705882353
3	36	32	0.8888888888888888
4	1689	1440	0.8525754884547069
5	29100	13950	0.4793814432989691
6	88181	27320	0.30981730758326625
7	47842	14386	0.30069813134902384
8	12641	2834	0.22419112411992723
9	3003	433	0.1441891441891442
10	1406	75	0.05334281650071124
11	153	3	0.0196078431372549
12	47	0	0.0
13	70	0	0.0

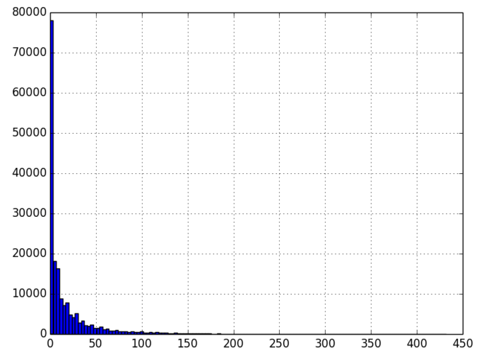
7. dim_market
Data missing: 0.0%
Discrete features

Paris	113704	31167	0.27410645183986493
Los Angeles	52698	20980	0.3981175756195681
San Francisco	17877	8363	0.46780779772892545

12. dim_is_instant_bookable
Data missing: 0.0%
Discrete features

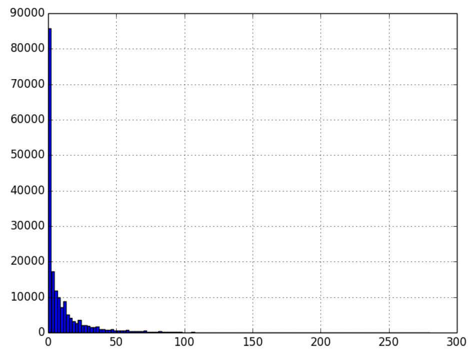
false	157427	47800	0.30363279488270756
true	26852	12710	0.47333531952927155

13. m_checkouts
Data missing: 0.101548754263%, ignore or delete samples with null data
Abnormal data: too large number
Data processing: log(x + 3)



NULL	187	80	0.42780748663101603
1	45443	6341	0.13953744251039765
2	39519	10257	0.2595460411447658
3	33647	11920	0.3542663536125063
4	29582	12759	0.4313095801500913
5	20789	10409	0.5006974842464765
6	11440	6439	0.5628496503496504
7	3411	2107	0.6177074171797127
8	261	198	0.7586206896551724

14. m_reviews
Data missing: 0.101548754263%, ignore or delete samples
Abnormal data: too large number
Data processing: log(x + 3)



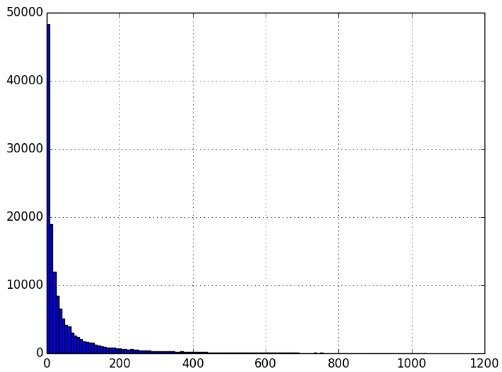
NULL	187	80	0.42780748663101603
1	56051	8964	0.15992578187721895
2	46894	14061	0.2998464622339745
3	35350	13844	0.3916265912305516
4	25016	11912	0.4761752478413815
5	14474	7807	0.5393809589608954
6	5516	3278	0.59427121102248
7	776	551	0.7100515463917526
8	15	13	0.8666666666666667

15. days_since_last_booking

Data missing: 20.5052457806%, as a new feature, the null data indicates that it never been booked

Abnormal data: too large number

Data processing: log(x + 3)



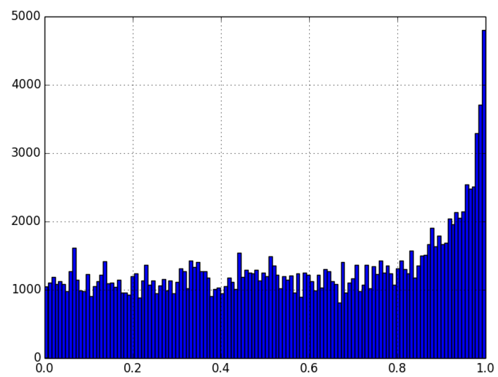
NULL	37836	3859	0.10199281107939528
1	9023	5414	0.6000221655768592
2	24181	13363	0.5526239609610851
3	25874	12422	0.48009584911494163
4	24756	10053	0.4060833737275812
5	21587	7149	0.33117153842590447
6	17545	4489	0.25585636933599315
7	12844	2282	0.1776705076300218
8	7654	1138	0.14868042853409982
9	2970	341	0.11481481481481481
10	9	0	0.0

17. image_quality_score

Data missing: 7.55913721572%, as a new feature, the null data indicates that the listing has no picture

Abnormal data: none

Data processing: floor(x * 10)



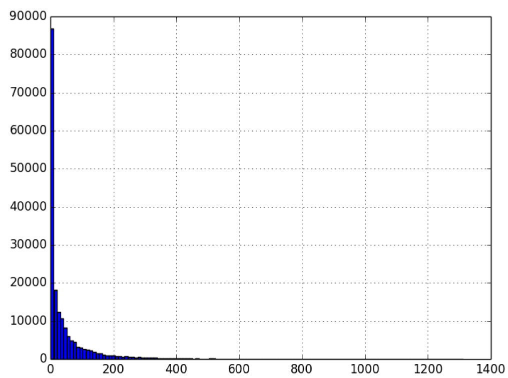
NULL	14011	4957	0.35379344800513884
0	14580	4042	0.27722908093278464
1	13880	3774	0.2719020172910663
2	14013	3996	0.28516377649325625
3	15395	4726	0.30698278661903217
4	15260	4530	0.29685452162516385
5	15132	4798	0.317076394395982
6	14243	4987	0.35013690935898334
7	15895	5511	0.3467128027681661
8	19261	7142	0.37080110066974714
9	32609	12047	0.36943788524640436

18. m_total_overall_rating

Data missing: 0.101548754263%, ignore or delete samples

Abnormal data: too large number

Data processing: log(x + 3)



NULL	187	80	0.42780748663101603
1	56614	9044	0.15974847210937224
2	5651	1474	0.260838789594762
3	25827	7255	0.2809075773415418
4	26240	9304	0.3545731707317073
5	27517	11247	0.40872914925318893
6	21353	10307	0.4826956399569147
7	14116	7617	0.539600453386228
8	5721	3457	0.60426498863835
9	1029	703	0.6831875607385811
10	24	22	0.9166666666666666

20. dim_has_wireless_internet
Data missing: 0.0%

0	11473	1851	0.1613353089863157
1	172806	58659	0.33945001909655914

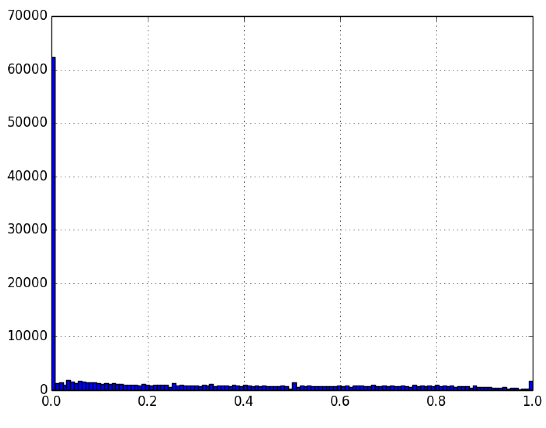
23. ds_checkin_gap
Data missing: 1.20609509742%, as a new feature, it has no obvious meaning

NULL	2221	877	0.39486717694732104
0	10032	3497	0.34858452950558216
1	8050	3571	0.4436024844720497
2	6675	3247	0.4864419475655431
3	5562	2772	0.49838187702265374
4	4998	2502	0.5006002400960384
5	4530	2308	0.5094922737306843
6	4231	2047	0.4838099740014181
7	137980	39689	0.2876431366864763

24. ds_checkout_gap
Data missing: 1.20609509742%, as a new feature, it has no obvious meaning

NULL	2221	877	0.39486717694732104
0	8782	2958	0.3368253245274425
1	6459	2846	0.4406254838210249
2	5170	2549	0.493036750483559
3	3989	2012	0.5043870644271747
4	3246	1616	0.49784349969192854
5	2706	1367	0.5051736881005173
6	2437	1233	0.5059499384489126
7	149269	45052	0.3018175240672879

27. occ_occupancy_trailing_90_ds
Data missing: 5.51350001086%, as a new feature, the null data indicates newly released listing
Abnormal data: none
Data processing: floor(x * 10), 0 as a specific value



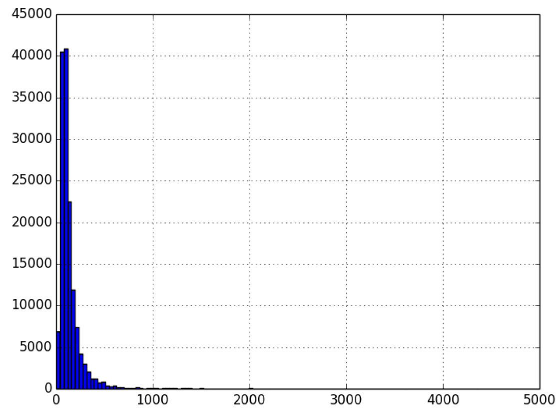
NULL	10218	3714	0.3634762184380505
0	79353	11041	0.13913777676962433
1	14304	4358	0.30467002237136465
2	12200	4649	0.38106557377049183
3	11102	4750	0.42785083768690324
4	9900	4663	0.471010101010101
5	10455	5375	0.5141080822572932
6	10218	5571	0.5452143276570758
7	10203	6088	0.596687248848378
8	9542	5930	0.621463005659191
9	5099	3344	0.6558148656599333
10	1685	1027	0.6094955489614243

30. price_booked_most_recent

Data missing: 20.5052457806%, as a new feature, the null data indicates that the listing never been booked

Abnormal data: too large number and too small number

Data processing: log(x + 3)



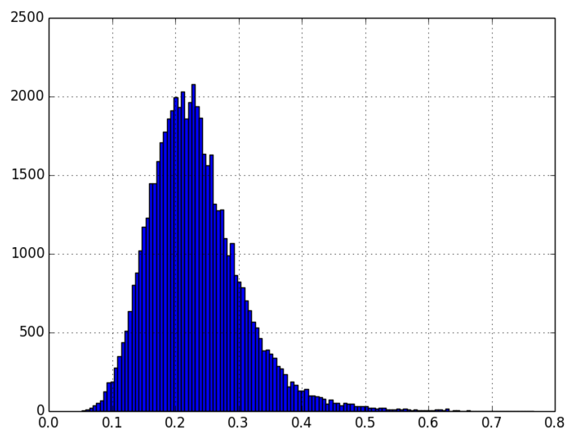
NULL	37836	3859	0.10199281107939528
1	41	4	0.0975609756097561
2	47	30	0.6382978723404256
3	108	41	0.37962962962962965
4	1415	743	0.5250883392226149
5	24801	11169	0.4503447441635418
6	67386	26406	0.3918618110586769
7	39014	14616	0.37463474650125594
8	10426	3053	0.29282562823709957
9	2232	501	0.22446236559139784
10	923	76	0.08234019501625135
11	46	8	0.17391304347826086
12	4	4	1.0

31. p2_p3_click_through_score

Data missing: 68.976584052%, as a new feature, the null data indicates that it did not appear in the search

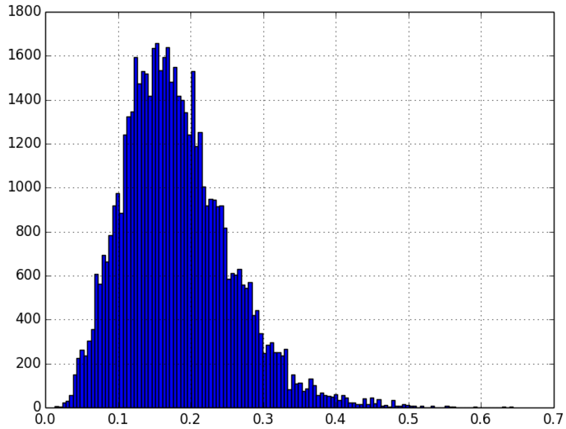
Abnormal data: none

Data processing: floor(x * 20)



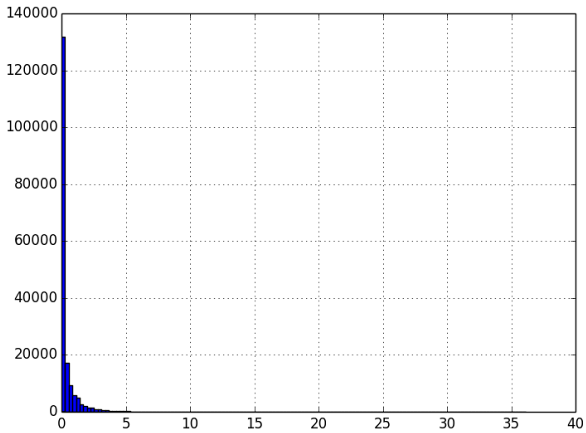
NULL	127110	39318	0.30932263393910786
1	568	149	0.2623239436619718
2	5465	1627	0.2977127172918573
3	14405	4879	0.33870183963901423
4	17064	6092	0.35700890764181903
5	10858	4164	0.3834960397863327
6	5201	2329	0.44779850028840607
7	2056	1016	0.49416342412451364
8	839	460	0.5482717520858165
9	381	235	0.6167979002624672
10	173	123	0.7109826589595376
11	75	53	0.7066666666666667
12	56	43	0.7678571428571429
13	14	10	0.7142857142857143
14	12	11	0.9166666666666666
15	2	1	0.5

32. p3_inquiry_score
Data missing: 70.1604144492%, as a new feature, the null data indicates that it did not appear in listing page
Abnormal data: none
Data processing: floor(x * 20)



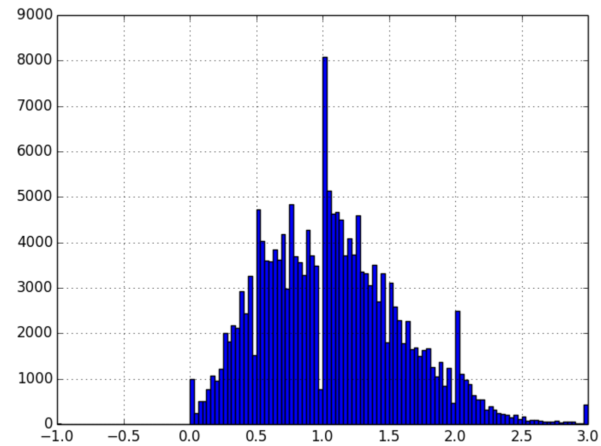
NULL	129290	40459	0.31293216799443113
0	602	124	0.2059800664451827
1	5748	1469	0.2555671537926235
2	14115	4783	0.33885936946510803
3	15049	5698	0.3786298092896538
4	10639	4186	0.3934580317699032
5	5329	2258	0.4237192719084256
6	2192	950	0.4333941605839416
7	756	327	0.43253968253968256
8	351	153	0.4358974358974359
9	156	67	0.42948717948717946
10	29	18	0.6206896551724138
11	17	12	0.7058823529411765
12	6	6	1.0

33. listing_m_listing_views_2_6_ds_night_decay
Data missing: 1.26365749289%, as a new feature, the null data indicates that it did not appear in listing view in past 6 days
Abnormal data: too large number
Data processing: log(100 * x + 3)



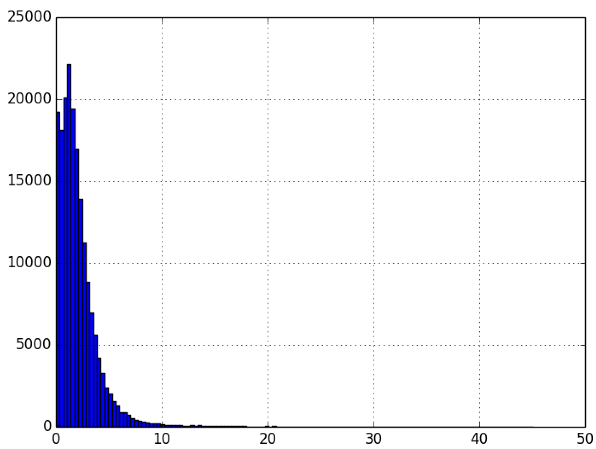
NULL	2346	1036	0.4416027280477408
1	107883	20656	0.19146668149754828
3	4958	1887	0.3805970149253731
4	18864	7465	0.3957273112807464
5	19814	9472	0.4780458261835066
6	14555	8586	0.5899003778770182
7	9883	6754	0.6833957300414853
8	4455	3421	0.7679012345679013
9	1281	1032	0.8056206088992974
10	212	175	0.8254716981132075
11	28	26	0.9285714285714286

39. kdt_score
Data missing: 0.0%
Abnormal data: negative number
Data processing: floor(x * 5)



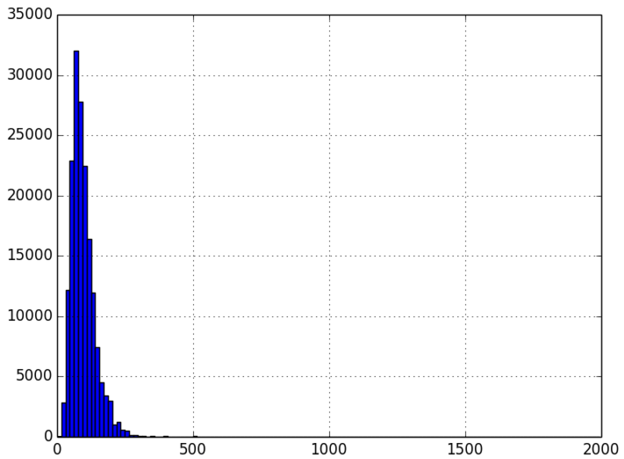
-5	6	1	0.16666666666666666
0	4400	1141	0.25931818181818184
1	11991	3172	0.26453173213243264
2	20633	5685	0.27552949159114043
3	24544	7108	0.28960234680573665
4	21252	6492	0.30547713156408807
5	32297	10136	0.3138371985014088
6	23107	8195	0.3546544337213831
7	17027	6556	0.38503553180243144
8	11034	4574	0.414536885988762
9	7206	3124	0.4335276158756592
10	6825	2812	0.41201465201465204
11	2025	811	0.4004938271604938
12	889	337	0.37907761529808776
13	394	127	0.3223350253807107
14	243	95	0.39094650205761317
15	406	144	0.35467980295566504

40. r_kdt_listing_views_0_6_avg_n100
Data missing: 0.000543041466646%, ignore or delete samples
Abnormal data: too large number
Data processing: log(x + 3)



NULL	1	0	0.0
1	53697	13579	0.25288191146618993
2	119297	41959	0.3517188194170851
3	10249	4578	0.446677724656064
4	996	378	0.3795180722891566
5	39	16	0.41025641025641024

45. r_kdt_m_effective_daily_price_booked_n100_p50
Data missing: 7.03564524187%, as a new feature, the null data indicates that it never been booked with same type of room in kdt
Abnormal data: too large number
Data processing: log(x + 3)



NULL	12975	2831	0.2181888246628131
1	7	3	0.42857142857142855
3	9	1	0.1111111111111111
4	2059	365	0.17727051966974258
5	33519	9195	0.27432202631343416
6	101165	33249	0.3286610982059012
7	33430	14516	0.4342207597965899
8	977	318	0.3254861821903787
9	127	30	0.23622047244094488
10	11	2	0.18181818181818182

Accessory 2

training dataset

```
show: 165614.0, clk: 54271.0, pctr: 0.351867524936, rctr: 0.327695726207, auc: 0.872365850773
```

testing dataset

```
show: 18401.0, clk: 6170.0, pctr: 0.354892978854, rctr: 0.335307863703, auc: 0.862483576882,
```