房源成交预测

1. 总结

这个问题是基于18万样本来预测房源是否在未来成交的二分类问题，通过对数据的分析处理和模型训练，最终的效果是选用fm模型，在测试集上auc = 0.862。目前的版本是这样的，只使用了通过特征选择得到的20组特征，在特征处理和离散化后得到1万个特征（包括房子id），没有做反复的特征迭代、复杂特征交叉（fm本身已经做了最简单的特征交叉）。fm实现是公司内部版本的，libfm也尝试过，用sgd默认参数的情况，auc = 0.831

2. 文件说明

README	项目总结
analysis.py	数据分析
booking_predict_original_dataset.hive	构建hive表
feature_select.hive	特征选择
fm_feature_process.py	数据清洗，特征处理
generate_signature.py	特征签名
tmp	数据分析+特征选择的中间结果
train_fm	训练和评估代码

3. 建模步骤

① 数据分析

用pandas、hive 分析每个特征的详细信息，为后续步骤（数据清洗、特征处理、特征选择）提供数据支持，包括：

pandas	执行 python analysis.py 即可在tmp文件夹下产生	
	feature_miss_percentage.txt	数据丢失率
	hist_pic	分布直方图
	feature_statistics.txt	均值方差等统计指标
hive	执行hive -f feature_select.hive统计每个特征在各个特征值上的总样本数、负样本数、正样本率	
	feature_detail.txt	结果保存

② 数据清洗和特征处理

清洗详细步骤见附件1。

③ 特征选择

根据feature_detail.txt，判断特征和目标的相关性，优选简单有效的特征

④ 特征工程编码，生成训练集、测试集

fm_feature_process.py

⑤ 模型训练，评估，迭代

只用了两个版本的fm，评估效果主要看auc（见附件2），模型迭代需要由少到多选择特征，验证单特征效果，最后加上交叉特征

4. 业务建议和思考

A. 应用模型

- 搜索房源页面，召回房源按照模型预测结果排序，有助于全站成交；在应用时要注意多样性和冷启动，避免全部推荐热门房源
- 相似推荐页面，召回房源按照相似度和成交预测值做综合排序，要注意相似度和成交概率的权衡

B. 其他

- 与成交相关性最强的特征，大部分是跟房源热度相关的特征，airbnb头部优质的房源曝光和成交都很充分，而且曝光、成交有滚雪球效应。作为用户我也有同样的感觉，很少人想做第一个吃螃蟹的人，选择一个评论多订单多的房源，我不会有太差的旅游体验：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
- 次强的特征是房源价格、wifi、图片，地理位置等，这些特征在我去日本选择房源都占了重要的因素，我从数据中发现成交率随经纬度变化比较大，不过经纬度特征的交叉离散化需要一些更为细致的工作，没有用在目前版本中：m_effective_daily_price，image_quality_score，dim_has_wireless_internet，price_booked_most_recent，dim_lat，dim_lng
- 房型、容纳人数并不那么重要：dim_room_type，dim_person_capacity。房间图片不是越多越有利成交，清洁费高反而成交不错：m_professional_pictures，m_pricing_cleaning_fee
- 周末效应，季节效应：ds_night_day_of_week，ds_night_day_of_year
- 过去6天搜索用户数等特征在中段有成交率陡降的情况，没找到合理的解释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

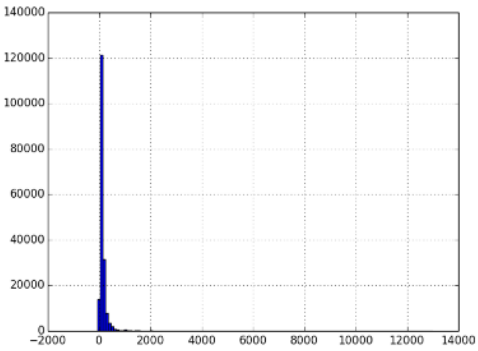
C. 感受

- 我的理解，airbnb更像是用户短期旅行、与世界沟通的助手，告诉你住哪里玩什么。不同于musically可以推荐任何用户可能感兴趣的内容，airbnb需要做的是及时抓住用户的短期需求，更贴心快捷地满足用户，宾至如归。

附件1

每个特征的详细分析，编号代表数据字段index，图分别是特征的直方图、每个特征值上的总样本数、负样本数、正样本率

5.m_effective_daily_price
数据丢失：0.0
异常数据：负数，大数字，小数字
特征处理：取log（x+3），待优化



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

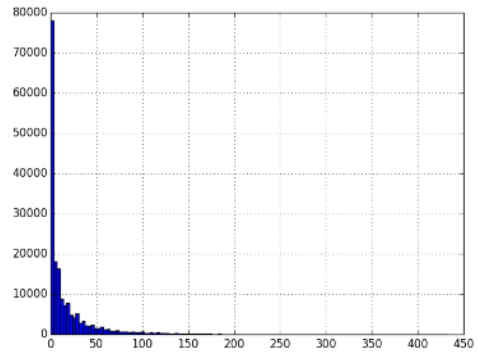
7.dim_market
数据丢失：0.0
离散特征

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

12.dim_is_instant_bookable
数据丢失：0.0
离散特征

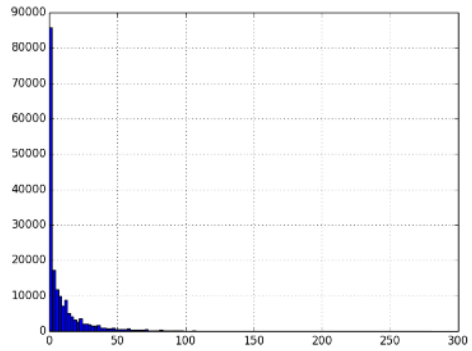
false	157427	47800	0.30363279488270756
true	26852	12710	0.47333531952927155

13.m_checkouts
数据丢失：0.101548754263，特征忽略或删除样本
异常数据：大数字
数据处理：log（x+3）



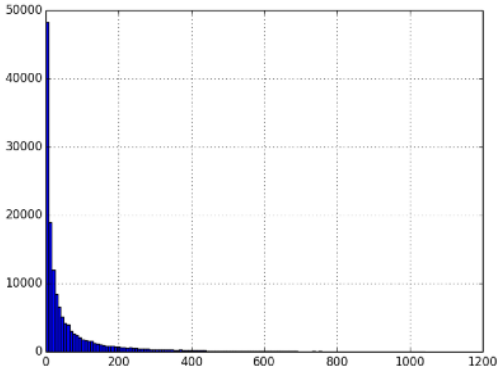
NULL	187	80	0.42780748663101603
1	45443	6341	0.13953744251039765
2	39519	10257	0.2595460411447658
3	33647	11920	0.3542663536125063
4	29582	12759	0.4313095801500913
5	20789	10409	0.5006974842464765
6	11440	6439	0.5628496503496504
7	3411	2107	0.6177074171797127
8	261	198	0.7586206896551724

14.m_reviews
数据丢失：0.101548754263，特征忽略或删除样本
异常数据：大数字
数据处理：log（x+3）



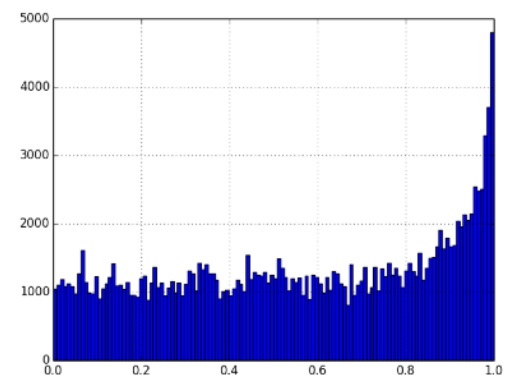
NULL	187	80	0.42780748663101603
1	56051	8964	0.15992578187721895
2	46894	14061	0.2998464622339745
3	35350	13844	0.3916265912305516
4	25016	11912	0.4761752478413815
5	14474	7807	0.5393809589608954
6	5516	3278	0.59427121102248
7	776	551	0.7100515463917526
8	15	13	0.8666666666666667

15.days_since_last_booking
数据丢失：20.5052457806，作新特征，表示从没被预定过
数据异常：大数字（不处理也可以）
数据处理：log（x+3）



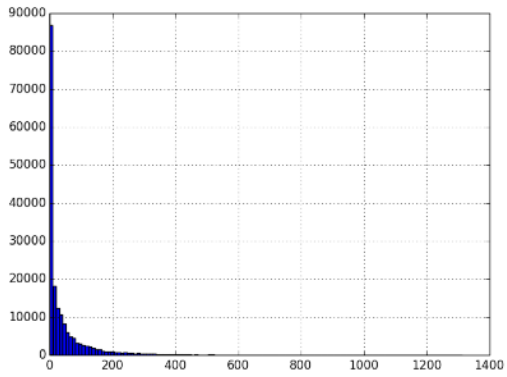
NULL	37836	3859	0.10199281107939528
1	9023	5414	0.6000221655768592
2	24181	13363	0.5526239609610851
3	25874	12422	0.48009584911494163
4	24756	10053	0.4060833737275812
5	21587	7149	0.33117153842590447
6	17545	4489	0.25585636933599315
7	12844	2282	0.1776705076300218
8	7654	1138	0.14868042853409982
9	2970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

17.image_quality_score
数据丢失：7.55913721572，作新特征，表示没有图片
数据异常：没有
数据处理：0～1值10等分



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
数据丢失：0.101548754263，特征忽略或者删除样本
数据异常：大数字
数据处理：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	0.5396004533862284
8	5721	3457	0.60426498863835
9	1029	703	0.6831875607385811
10	24	22	0.9166666666666666

20.dim_has_wireless_internet
数据丢失：0.0

0	11473	1851	0.16133530898631 57
1	172806	58659	0.33945001909655914

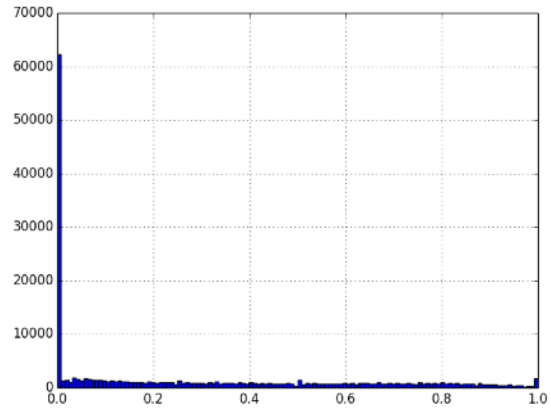
23.ds_checkin_gap
数据丢失：1.20609509742，作为新特征，没有明显含义

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
数据丢失：1.20609509742，作为新特征，没有明显含义

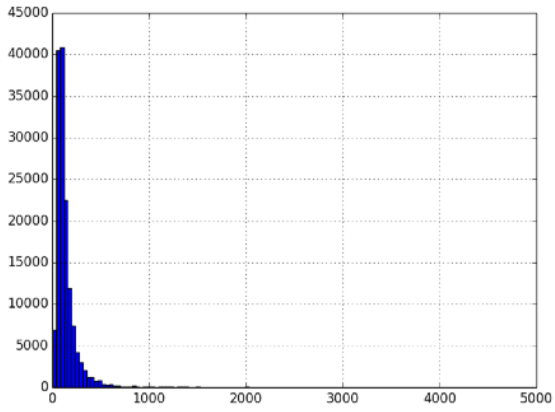
NULL	2221	877	0.39486717694732104
0	8782	2958	0.3368253245274425
1	6459	2846	0.4406254838210249
2	5170	2549	0.493036750483559
3	3989	2012	0.5043870644271747
4	3246	1616	0.49784349969192854
5	2706	1367	0.5051736881005173
6	2437	1233	0.5059499384489126
7	149269	45052	0.3018175240672879

27.occ_occupancy_trailing_90_ds
数据丢失：5.51350001086，作为新特征，表示新发布产品
数据异常：没有
数据处理：0～1值10等分，0单独分开



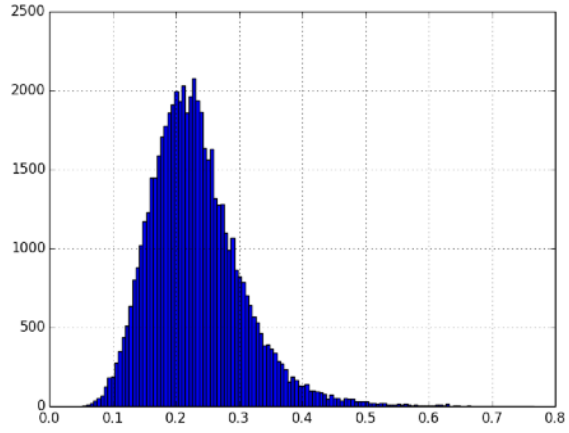
NULL	10218	3714	0.3634762184380505
0	79353	11041	0.13913777676962433
1	14304	4358	0.30467002237136465
2	12200	4649	0.38106557377049183
3	11102	4750	0.42785083768690324
4	9900	4663	0.471010101010101
5	10455	5375	0.5141080822572932
6	10218	5571	0.5452143276570758
7	10203	6088	0.596687248848378
8	9542	5930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

30.price_booked_most_recent
数据丢失：20.5052457806，作为新特征，表示从没有被预定过
数据异常：大数字，小数字
数据处理：log（x+3），待优化



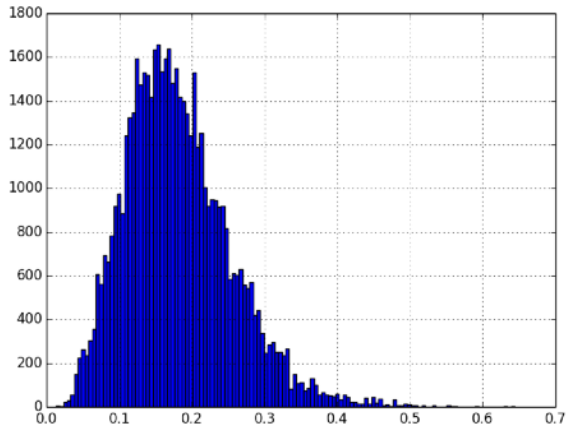
NULL	37836	3859	0.10199281107939528
1	41	4	0.0975609756097561
2	47	30	0.6382978723404256
3	108	41	0.37962962962962965
4	1415	743	0.5250883392226149
5	24801	11169	0.4503447441635418
6	67386	26406	0.3918618110586769
7	39014	14616	0.37463474650125594
8	10426	3053	0.29282562823709957
9	2232	501	0.22446236559139784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

31.p2_p3_click_through_score
数据丢失：68.976584052，作为新特征，表示没在搜索中出现过
数据异常：没有
数据处理：0~1值20等分，高分合并



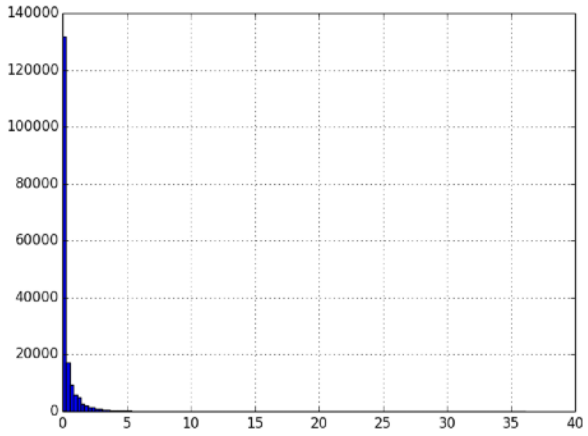
NULL	127110	39318	0.30932263393910786
1	568	149	0.2623239436619718
2	5465	1627	0.2977127172918573
3	14405	4879	0.33870183963901423
4	17064	6092	0.35700890764181903
5	10858	4164	0.3834960397863327
6	5201	2329	0.44779850028840607
7	2056	1016	0.49416342412451364
8	839	460	0.5482717520858165
9	381	235	0.6167979002624672
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
数据丢失：70.1604144492，作为新特征，表示从没在listing page出现过
数据异常：没有
数据处理：0~1值20等分，高分合并



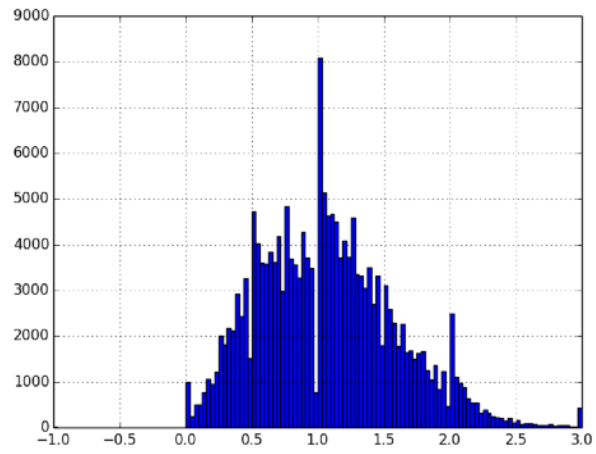
NULL	129290	40459	0.31293216799443113
0	602	124	0.2059800664451827
1	5748	1469	0.2555671537926235
2	14115	4783	0.33885936946510803
3	15049	5698	0.3786298092896538
4	10639	4186	0.3934580317699032
5	5329	2258	0.4237192719084256
6	2192	950	0.4333941605839416
7	756	327	0.43253968253968256
8	351	153	0.4358974358974359
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
数据丢失：1.26365749289，作为新特征，表示近6天没在listing view出现过
数据异常：大数字
数据处理：log（100*x+3）



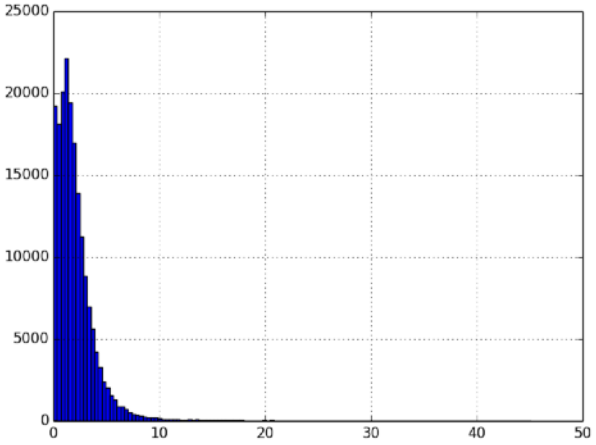
NULL	2346	1036	0.4416027280477408
1	107883	20656	0.19146668149754828
3	4958	1887	0.3805970149253731
4	18864	7465	0.3957273112807464
5	19814	9472	0.4780458261835066
6	14555	8586	0.5899003778770182
7	9883	6754	0.6833957300414853
8	4455	3421	0.7679012345679013
9	1281	1032	0.8056206088992974
10	212	175	0.8254716981132075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

39.kdt_score
数据丢失：0.0
数据异常：负数
数据处理：0~3值15等分，高分合并



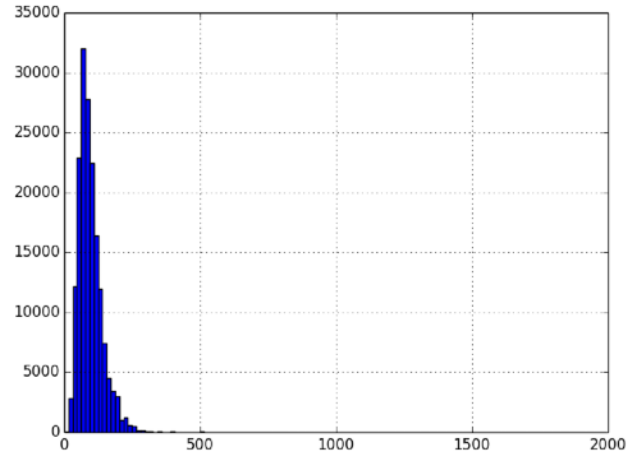
-5	6	1	0.16666666666666666
0	4400	1141	0.25931818181818184
1	11991	3172	0.26453173213243264
2	20633	5685	0.27552949159114043
3	24544	7108	0.28960234680573665
4	21252	6492	0.30547713156408807
5	32297	10136	0.3138371985014088
6	23107	8195	0.3546544337213831
7	17027	6556	0.38503553180243144
8	11034	4574	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
数据丢失：0.000543041466646，忽略特征或删除样本
数据异常：大数字
数据处理：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
数据丢失：7.03564524187，作为新特征，表示kdt里同房型没被预定过
数据异常：大数字
数据处理：log（x+3），待优化



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	0.236220472440944 88
10	11	2	0.18181818181818182

附件2:

训练集

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

测试集

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