说明文档

1. 最优成绩的结果文件见 final_predict.csv

2. 代码文件见 code 文件夹。

3. 算法思路：

特征工程 + xgboost

a. 数据预处理

数据分为线上和线下两个部分：

线下：1-17 天为训练集，18-24 天的指定时间片为验证集

线上：1-24 天为训练集，25-52 天为测试集

b. 特征工程

编号	Info Gain	解释
160	0.657431157	prev_gaps_5
34	0.123384164	prev_gaps_1
35	0.056180057	prev_gaps_2
191	0.025083145	aver_gap
161	0.017760746	prev_gaps_10
162	0.008791944	prev_gaps_15
163	0.006303087	prev_gaps_20
184	0.004781855	aver_request
3	0.004055107	minute
4	0.003594782	prev_requests_1
156	0.003409849	prev_requests_15
50	0.002544958	prev_gaps_17
197	0.002222981	aver_gap_30
208	0.002171024	PM
194	0.002109938	aver_gap_15
193	0.002016235	aver_gap_10
46	0.001444793	prev_gaps_13
192	0.001438387	aver_gap_5
48	0.001361367	prev_gaps_15
2	0.001301042	weekday

11	0.001264271	prev_requests_8
13	0.001192683	prev_requests_10
5	0.001064635	prev_requests_2
52	0.001060754	prev_gaps_19
41	0.001052598	prev_gaps_8
196	0.001022272	aver_gap_25
112	0.001	prev_ordernum_from_other_19
195	0.000977906	aver_gap_20
14	0.000957729	prev_requests_11
71	0.000945931	prev_driver_nums_8
206	0.000934027	tj_level_4_ratio
154	0.000924785	prev_requests_5
207	0.000892453	weather
47	0.000870687	prev_gaps_14
204	0.000863716	tj_level_2_ratio
45	0.000863428	prev_gaps_12
49	0.000850866	prev_gaps_16
6	0.000793545	prev_requests_3
17	0.000775493	prev_requests_14
203	0.000774877	tj_level_1_ratio
198	0.000771746	total_tj_cnt
37	0.000749064	prev_gaps_4
16	0.000737899	prev_requests_13
190	0.000733628	aver_request_30
8	0.000725003	prev_requests_5
7	0.000713873	prev_requests_4
155	0.000712812	prev_requests_10
42	0.000707607	prev_gaps_9
164	0.000680454	prev_gaps_25
199	0.000665484	tj_level_1_cnt
44	0.00063028	prev_gaps_11
12	0.000629225	prev_requests_9
165	0.000616753	prev_gaps_30
36	0.000616588	prev_gaps_3
74	0.000598114	prev_driver_nums_11
205	0.000576726	tj_level_3_ratio
51	0.000541657	prev_gaps_18
20	0.000536217	prev_requests_17
63	0.000534251	prev_gaps_30
9	0.000533141	prev_requests_6
178	0.000532924	prev_ordernum_from_other_valid_5
40	0.000511338	prev_gaps_7
1	0.000506565	district_id

202	0.000499224	tj_level_4_cnt
18	0.000496194	prev_requests_15
126	0.000496062	prev_ordernum_from_other_valid_3
21	0.000482698	prev_requests_18
91	0.000479056	prev_driver_nums_28
189	0.000472504	aver_request_25
43	0.000466631	prev_gaps_10
116	0.00046513	prev_ordernum_from_other_23
78	0.000458074	prev_driver_nums_15
23	0.000450822	prev_requests_20
94	0.00044947	prev_ordernum_from_other_1
88	0.000447944	prev_driver_nums_25
66	0.000446657	prev_driver_nums_3
25	0.000442231	prev_requests_22
157	0.000442051	prev_requests_20
39	0.000439892	prev_gaps_6
19	0.000439638	prev_requests_16
200	0.00042401	tj_level_2_cnt
15	0.000421898	prev_requests_12
29	0.00041647	prev_requests_26
105	0.000414418	prev_ordernum_from_other_12
110	0.000413672	prev_ordernum_from_other_17
106	0.000410938	prev_ordernum_from_other_13
10	0.000408584	prev_requests_7
187	0.00040364	aver_request_15
115	0.000399329	prev_ordernum_from_other_22
38	0.000398936	prev_gaps_5
86	0.000391305	prev_driver_nums_23
73	0.000390884	prev_driver_nums_10
89	0.000389864	prev_driver_nums_26
85	0.000387664	prev_driver_nums_22
55	0.000383748	prev_gaps_22
67	0.000381119	prev_driver_nums_4
188	0.000380459	aver_request_20
171	0.00037808	prev_driver_nums_30
82	0.000372902	prev_driver_nums_19
54	0.000371405	prev_gaps_21
56	0.00036495	prev_gaps_23
118	0.000343931	prev_ordernum_from_other_25
95	0.000343527	prev_ordernum_from_other_2
31	0.000343471	prev_requests_28
107	0.000343182	prev_ordernum_from_other_14
24	0.000338561	prev_requests_21

186	0.00033391	aver_request_10
30	0.000333653	prev_requests_27
99	0.000327703	prev_ordernum_from_other_6
60	0.000325292	prev_gaps_27
98	0.000325133	prev_ordernum_from_other_5
124	0.000324856	prev_ordernum_from_other_valid_1
28	0.000320983	prev_requests_25
97	0.00031736	prev_ordernum_from_other_4
64	0.000309162	prev_driver_nums_1
22	0.00030191	prev_requests_19
59	0.000299351	prev_gaps_26
119	0.000296848	prev_ordernum_from_other_26
201	0.000295201	tj_level_3_cnt
77	0.000294808	prev_driver_nums_14
33	0.000291593	prev_requests_30
68	0.000288679	prev_driver_nums_5
136	0.000288413	prev_ordernum_from_other_valid_13
173	0.000283046	prev_ordernum_from_other_10
92	0.000280177	prev_driver_nums_29
103	0.000274554	prev_ordernum_from_other_10
57	0.000273624	prev_gaps_24
101	0.000271906	prev_ordernum_from_other_8
53	0.000271424	prev_gaps_20
172	0.000270204	prev_ordernum_from_other_5
104	0.000269107	prev_ordernum_from_other_11
166	0.000267594	prev_driver_nums_5
62	0.000265667	prev_gaps_29
185	0.000258083	aver_request_5
100	0.000255252	prev_ordernum_from_other_7
61	0.000252542	prev_gaps_28
58	0.000251054	prev_gaps_25
90	0.000249782	prev_driver_nums_27
27	0.000249709	prev_requests_24
131	0.000249268	prev_ordernum_from_other_valid_8
128	0.000248232	prev_ordernum_from_other_valid_5
65	0.000242647	prev_driver_nums_2
121	0.00024107	prev_ordernum_from_other_28
96	0.000232806	prev_ordernum_from_other_3
26	0.00022837	prev_requests_23
80	0.000227266	prev_driver_nums_17
72	0.000224462	prev_driver_nums_9
129	0.000223017	prev_ordernum_from_other_valid_6
137	0.000215235	prev_ordernum_from_other_valid_14

176	0.000214716	prev_ordernum_from_other_25
102	0.000210562	prev_ordernum_from_other_9
32	0.000203554	prev_requests_29
70	0.000200439	prev_driver_nums_7
120	0.000199804	prev_ordernum_from_other_27
167	0.000197077	prev_driver_nums_10
170	0.000196892	prev_driver_nums_25
83	0.000195957	prev_driver_nums_20
133	0.000195114	prev_ordernum_from_other_valid_10
152	0.000191634	prev_ordernum_from_other_valid_29
113	0.000188069	prev_ordernum_from_other_20
159	0.000184022	prev_requests_30
183	0.000183071	prev_ordernum_from_other_valid_30
76	0.000182172	prev_driver_nums_13
180	0.00017985	prev_ordernum_from_other_valid_15
75	0.000178591	prev_driver_nums_12
177	0.000177964	prev_ordernum_from_other_30
122	0.000177721	prev_ordernum_from_other_29
175	0.000177306	prev_ordernum_from_other_20
109	0.000176489	prev_ordernum_from_other_16
87	0.000176014	prev_driver_nums_24
168	0.00017306	prev_driver_nums_15
111	0.000172781	prev_ordernum_from_other_18
134	0.000171834	prev_ordernum_from_other_valid_11
81	0.000171078	prev_driver_nums_18
146	0.000166397	prev_ordernum_from_other_valid_23
114	0.000165892	prev_ordernum_from_other_21
84	0.000159362	prev_driver_nums_21
181	0.000158431	prev_ordernum_from_other_valid_20
108	0.000153885	prev_ordernum_from_other_15
141	0.0001508	prev_ordernum_from_other_valid_18
93	0.000150106	prev_driver_nums_30
132	0.00014852	prev_ordernum_from_other_valid_9
151	0.000148392	prev_ordernum_from_other_valid_28
153	0.000146715	prev_ordernum_from_other_valid_30
79	0.000146664	prev_driver_nums_16
174	0.000146415	prev_ordernum_from_other_15
169	0.000143816	prev_driver_nums_20
135	0.000142981	prev_ordernum_from_other_valid_12
127	0.000142627	prev_ordernum_from_other_valid_4
69	0.000142583	prev_driver_nums_6
130	0.000138934	prev_ordernum_from_other_valid_7
143	0.000138079	prev_ordernum_from_other_valid_20

123	0.000135968	prev_ordernum_from_other_30
144	0.000134793	prev_ordernum_from_other_valid_21
149	0.000134351	prev_ordernum_from_other_valid_26
138	0.000130926	prev_ordernum_from_other_valid_15
158	0.000130654	prev_requests_25
142	0.000130271	prev_ordernum_from_other_valid_19
145	0.000120271	prev_ordernum_from_other_valid_22
179	0.000113956	prev_ordernum_from_other_valid_10
140	0.000112975	prev_ordernum_from_other_valid_17
125	0.000112828	prev_ordernum_from_other_valid_2
117	0.000111926	prev_ordernum_from_other_24
139	0.000108992	prev_ordernum_from_other_valid_16
148	0.000107488	prev_ordernum_from_other_valid_25
150	0.000101256	prev_ordernum_from_other_valid_27
182	9.79E-05	prev_ordernum_from_other_valid_25
147	7.63E-05	prev_ordernum_from_other_valid_24

c. 特征选取

经过试验，选取前 52 维特征可以达到最好的线下测试结果

d. 模型选取

采用 GBDT 算法，使用 xgboost 的 R 语言版本，单个模型

参数如下：

booster	'gbtree'
objective	reg:linear
eval_metric	mae
max_depth	7
colsample_bytree	0.9
min_child_weight	10

eta	0.01

4. 运行方法

1. 修改 get_data_dir.m 里面的原始 data 路径

2. 运行 get_unique_items.m 获取所有的 district_hash, driver_hash, passenger_hash,并转化为唯一数字 id,便于保存

3. 分别运行 read_raw_order_data.m, read_raw_traffic_data.m, read_raw_weather_data.m 读入所有训练和测试的原始数据，并缓存成 mat

4. 修改 get_null_driver_id.m 的 null_id 值，设置为 driver_hash 为 NULL 的数字 id

5. 运行 prepare_train_data 和 prepare_test_data,以及 add_more_feature_2_forall, add_more_feature_2_test 准备好特征

6. 运行 sample_train_feat_back 采样生成 train 和 validation 数据

7. 运行 xgboost， 得到特征 importance 排名，根据 importance 排名选择特征，运行 reduce_feature_dim 得到选取的特征

8. 根据选取的特征训练 xgboost 模型得到结果，运行 predict_with_period 和 parse_rst 对模型结果进行调整得到最终结果。