维保线下热门店预测V1.0

背景：在维保店粒度的热门店预测的基础上，爱萍姐提出，除店的预测单量排序外，还希望获得具体有哪些人选择这些门店。粲哥在此基础上出了一版司机粒度的维保意愿预测，根据意愿构建维保店排序，特征主要考虑基于geohash的轨迹特征，V1.0在此基础上增加了司机的特征，细化了轨迹特征，提高了预测效果与排序效果。

具体的工作：

特征添加：

粲哥第一版特征：

来源表格	特征
am_dw.mta_haixiu_pid_driving_info_other_result_1_daily	stay_time_cut ,geohash_6_nums ,geohash_5_nums ,geohash_6_hours ,geohash_6_avg_hour ,geohash_6_max_hour ,geohash_6_min_hour ,is_ordered(作为label，表示在该geohash下订单)
am_dw.dim_mta_store_info	mta_haixiu_store_geohash_6（用于筛选数据，而非作为特征）

在此基础上增加了司机特征，细化了轨迹特征，新增的特征为：

来源表格	特征
am_dw.mta_haixiu_pid_driving_info_other_result_1_daily（细化轨迹特征）	geohash_5_hours ,geohash_5_avg_hour ,geohash_5_max_hour ,geohash_5_min_hour ,rank_geo_6 ,rank_geo_5
am_dw.csum_driver_integrate_info（增加司机特征）	driver_type ,driver_join_model ,driver_verify_status ,is_auth_driver,order_finish_count_1d ,order_finish_count_1w ,order_finish_count_1m ,order_finish_distance_1d ,order_finish_distance_1w ,order_finish_distance_1m ,morning_peak_order_finish_distance_1d ,morning_peak_order_finish_distance_1w ,morning_peak_order_finish_distance_1m ,night_peak_order_finish_distance_1d ,night_peak_order_finish_distance_1w ,night_peak_order_finish_distance_1m
am_dw.dwd_mta_store_info_extend（细化轨迹特征）	num_store

特征说明：

以下特征均是对每天的每人统计	
stay_time_cut	在geohash块待的时间长短
geohash_6_nums	与该geohash7属于一个geohash6的数量
geohash_5_nums	与该geohash7属于一个geohash5的数量
geohash_6_hours	在该geohash6待的总小时数（向下取整）

geohash_5_hours	在该geohash5待的总小时数（向下取整）
geohash_6_avg_hour	在该geohash6待的平均时间点（小时）
geohash_5_avg_hour	在该geohash5待的平均时间点（小时）
geohash_6_max_hour	在该geohash6待的最大时间点（小时）
geohash_5_max_hour	在该geohash5待的最大时间点（小时）
geohash_6_min_hour	在该geohash6待的最小时间点（小时）
geohash_5_min_hour	在该geohash5待的最小时间点（小时）
rank_geo_6	该geohash6待的时间的排序
rank_geo_5	该geohash5待的时间的排序
以下特征直接来源于专快司机信息汇总表	
driver_type	司机类型（1：加盟司机；2：自营车司机；3：营运司机）
driver_join_model	合作模式（1：普通加盟车；2：自营长包车；3：直营京B；4：加盟豪华车；5：对公司司机）
driver_verify_status	当前状态（0：未审核；1：已审核；2：已锁定；3：已禁用；4：未完成注册；5：已删除；8：不知道
is_auth_driver	是否认证司机（1：是；0：否）
order_finish_count_1d	最近1天完成订单数
order_finish_count_1w	最近7天完成订单数
order_finish_count_1m	最近30天完成订单数
order_finish_distance_1d	最近1天完成订单公里数
order_finish_distance_1w	最近7天完成订单公里数
order_finish_distance_1m	最近30天完成订单公里数
morning_peak_order_finish_distance_1d	最近1天完成早高峰订单公里数
morning_peak_order_finish_distance_1w	最近7天完成早高峰订单公里数
morning_peak_order_finish_distance_1m	最近30天完成早高峰订单公里数
night_peak_order_finish_distance_1d	最近1天完成晚高峰订单公里数
night_peak_order_finish_distance_1w	最近7天完成晚高峰订单数公里数
night_peak_order_finish_distance_1m	最近30天完成晚高峰订单数公里数
以下特征来源于维保店信息表（包括线下）	
num_store	该geohash7有几家维保店（根据geohash7匹配）

数据：

从am_dw.mta_haixiu_pid_driving_info_other_result_1_daily中取出20180715 - 20180806所有司机每一天的所有轨迹，并拼接各个相关表形成特征表am_temp.xsc_mta_pre_data_xie

预处理：

采样：正样本数量较少只有3000+，而负样本非常多，训练与验证的总数据选用全部3000+的正样本与20万采样后的负样本（选择20180715-20180803数据用于训练，20180804-20180806作为验证）。

模型：

该问题的特征是二分类，数据极不平衡，且评价首要指标为排序正确，而非意愿预测正确

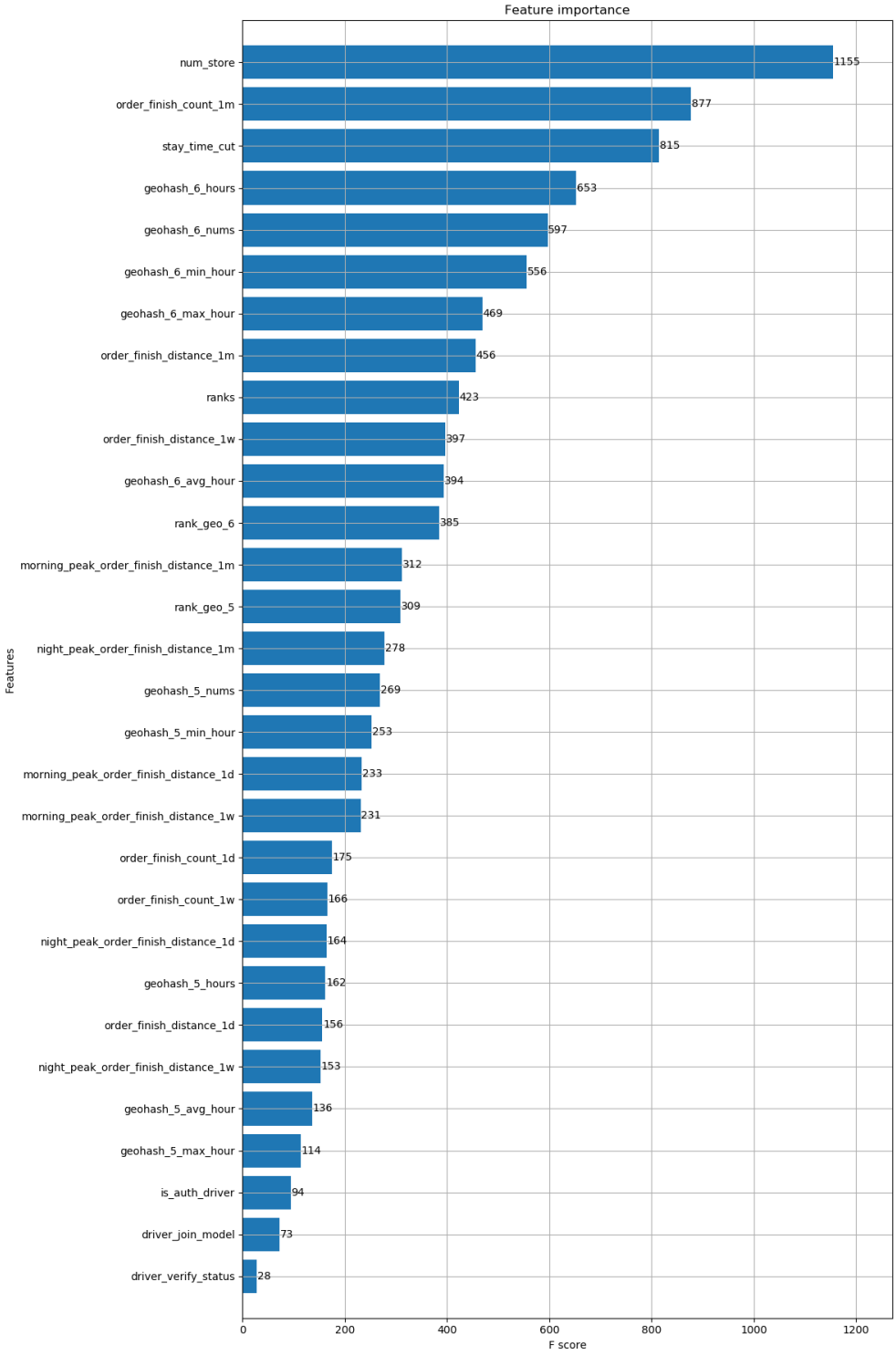
使用xgboost模型训练二分类问题；根据不平衡问题的特点设定参数（min_child_weight, scale_pos_weight, max_delta_step）；同时评价指标的要求是：选择阈值，保证准确率为1的情况下，False Negative的数量越大越好。

模型参数：

```
params={'booster':'gbtree',
 'objective': 'binary:logistic',
 'eval_metric': 'auc',
 'learning_rate':0.01,
 'max_depth':4,
 'lambda':10,
 'n_estimators':100000,
 'subsample':0.9,
 'colsample_bytree':0.9,
 'min_child_weight':0.1,
 'scale_pos_weight' :5,
 'seed':1,
 'nthread':4,
 'silent':1,
 'max_delta_step':10}
```

结果：

特征重要度：



可以看出新加入的特征起到了非常大的作用

排序对比：

我们选择20180715-20180803的数据来训练，选择2018-08-04，2018-08-05，2018-08-06这3天的数据来判断结果好坏。分别根据is_ordered和pred来排序。

因为只需要排序正确，因此我们提高准确率而压低召回率，保证排序的正确性。保证排序后，如果想要获得每个油站较为准确的预测人数，可以通过减小阈值实现。

选择阈值为0.985:

confusion matrix为[[339143 1], [378 255]]，False Negative为255

AUC: 0.9852， F1-score: 0.5737， Precision: 0.9961

	mta_haixiu_store_geohash_6	pred
0	wtw93p	55
1	wtw3ch	48
2	wtw33d	47
3	wtw6h5	25
4	wtw3dr	25
5	wtw3qr	18
6	wtw2f9	18
7	wtw3cb	15
8	wtw3tb	14
9	wtw3z5	12
10	wtw6hg	11
11	wtw3nk	11
12	wtw3vr	8
13	wtw3u5	6
14	wtw35k	2
15	wtw652	1

	mta_haixiu_store_geohash_6	is_ordered
0	wtw93p	105
1	wtw3ch	81
2	wtw33d	66
3	wtw6h5	59
4	wtw3qr	51
5	wtw2f9	44
6	wtw3tb	36
7	wtw3dr	35
8	wtw3cb	32
9	wtw6hg	25
10	wtw3z5	22
11	wtw3nk	22
12	wtw3u5	20
13	wtw3vr	15
14	wtw652	12
15	wtw35k	8

对每个店预测人数的分布：