以富聊为例基于因果推断的特征选择

datawalks空间: pai_test_xjp

代码空间: https://code.alibaba-inc.com/pai_biz_arch/feature_selection

富聊数据准备

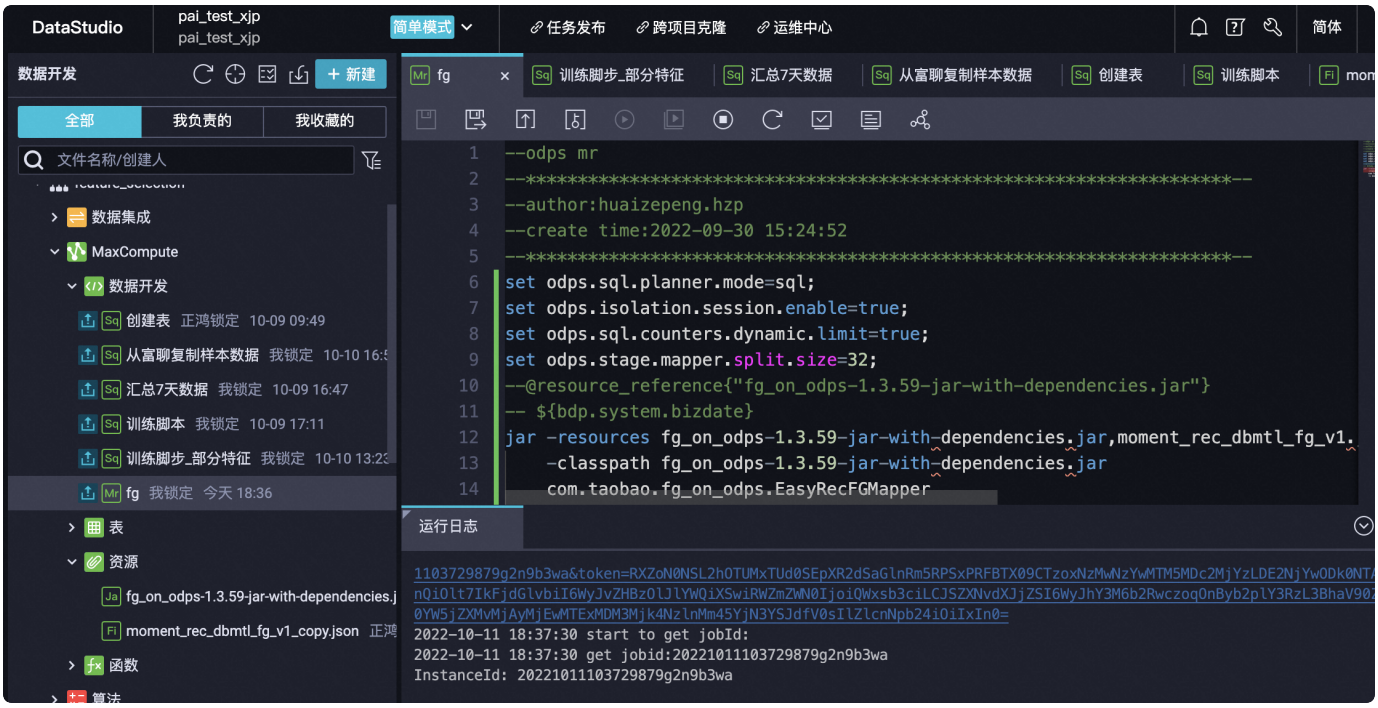
1 从客户项目空间拷贝数据至pai_test_xjp

说明：如果其他业务的datawalks空间包含空闲CPU计算资源（基于因果的特征选择算法不需要GPU），也可以直接在业务空间里进行以下流程。

示例：将所需的数据按日期分区拷贝至mc表moment_rec_dbmtl_rank_sample_v1（20220901-20220930）

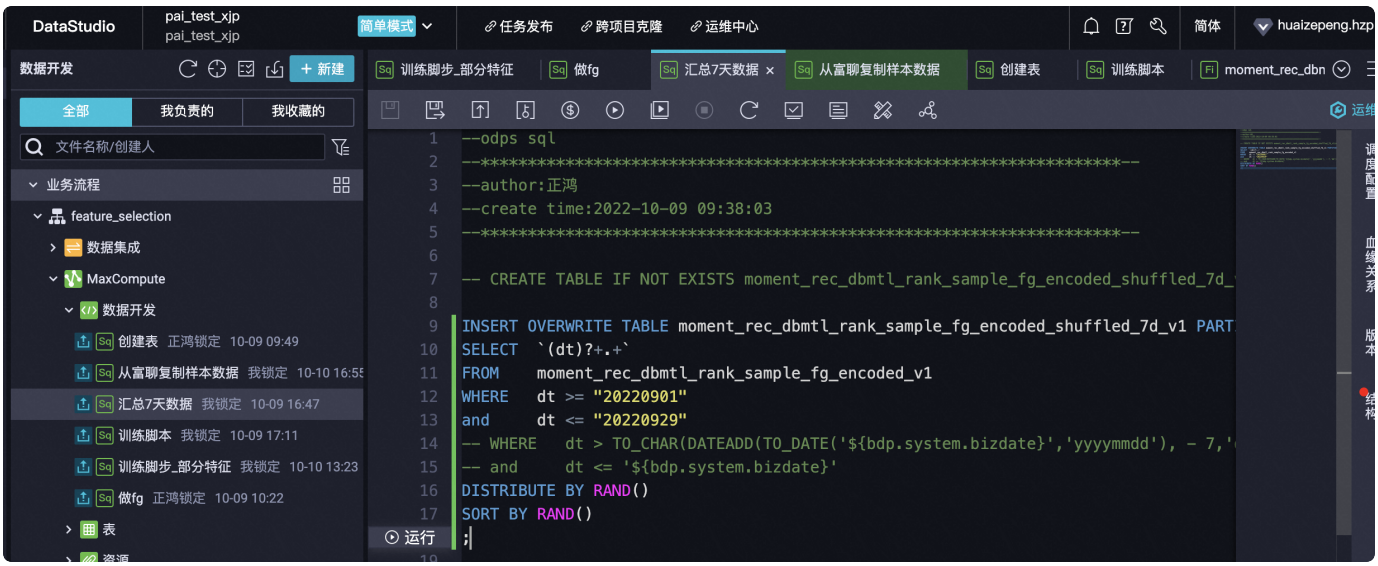
2 对一段时间数据进行fg操作，以适配后续模型训练接口

示例：fg编码之后数据同样安日期分区存入mc表moment_rec_dbmtl_rank_sample_fg_encoded_v1（20220901–20220930）



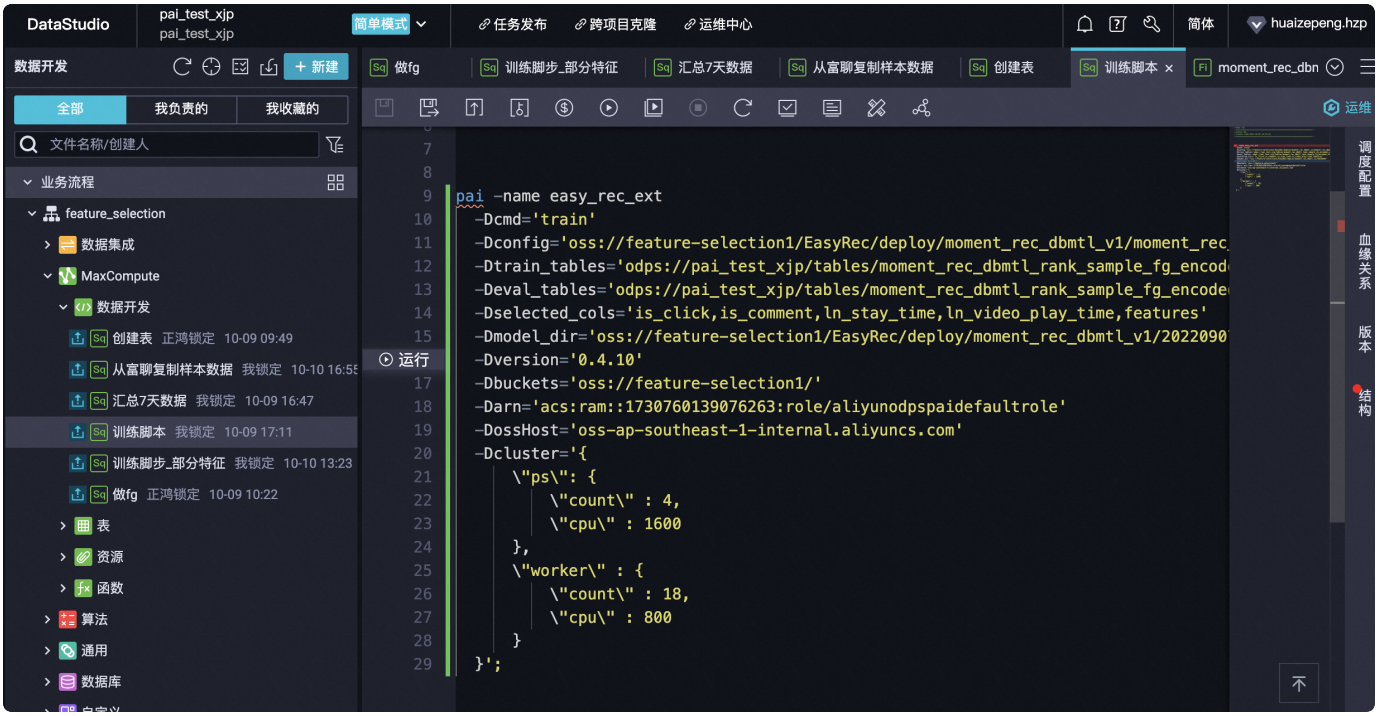
3 将训练数据shuffle并生成新表

示例：存入mc表moment_rec_dbmtl_rank_sample_fg_encoded_shuffled_7d_v1（20220901–20220929）



全量特征模型测试

示例：按照图示参数运行训练模型的节点



训练日志详见

https://logview.aliyun.com/logview/?h=http://service.ap-southeast-1.maxcompute.aliyun-inc.com/api&p=pai_test_xjp&i=20221013060423431gt4oedf03_76712f2f_6b34_479c_8a89_512ae061481d&token=NUgvWE41eTRVTFZaWVhYMU01eWttbno4SUJJPSxPRFBTX09CTzoxNzMwNzYwMTM5MDc2MjYzLDE2NjgyMzMwNjkseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN0IjoiQWxsb3ciLCJSZXNvdXJjZSI6WyJhY3M6b2RwczoqOnByb2plY3RzL3BhaV90ZXN0X3hqcC9pbnN0YW5jZXMvMjAyMjEwMTMwNjA0MjM0MzFndDRvZWRmMDNfNzY3MTJmMmZfNmIzNF80NzIjXzhhODlfNTEyYWUwNjE0ODFkIl19XSwiVmVyc2lvbiI6IjEifQ==

📎stderr.txt

全量特征带入排序模型的结果为

AUC（cilck）：0.8764

AUC（comment）：0.8483

```
[2022-10-13 22:48:56,663] [INFO] [76#MainThread] [tensorflow/python/estimator/estimator.py:2003]
Saving dict for global step 25000: auc_is_click = 0.87644935, auc_is_comment = 0.84829897,
global_step = 25000, loss = 0.6029204, loss/loss/cross_entropy_loss_is_click = 0.39436373, loss/loss/
cross_entropy_loss_is_comment = 0.04522771, loss/loss/l2_loss_ln_stay_time = 0.15809497, loss/loss/
l2_loss_ln_video_play_time = 0.005233984, loss/loss/total_loss = 0.6029204,
mean_squared_error_ln_stay_time = 0.15809102, mean_squared_error_ln_video_play_time = 0.005233996
```

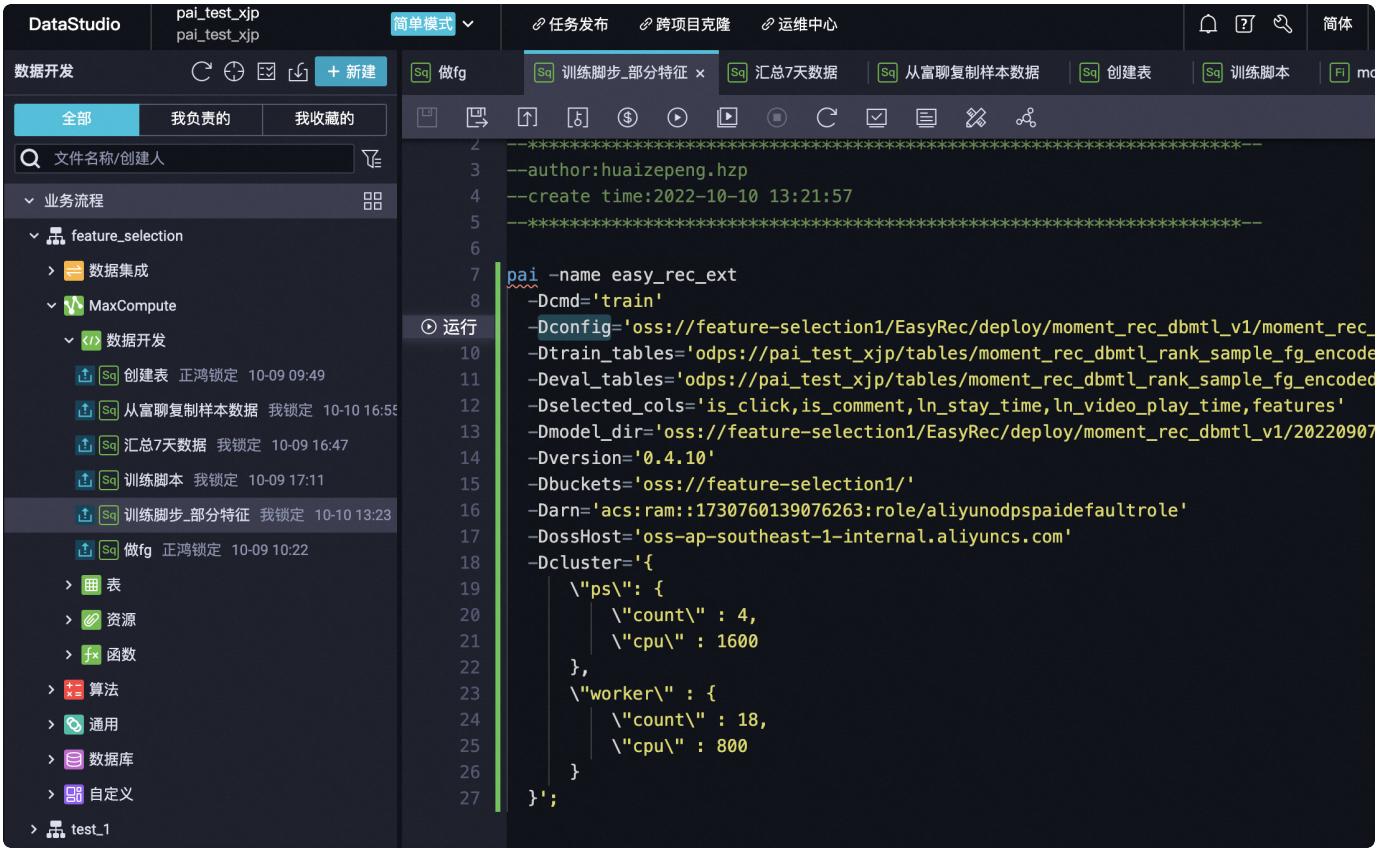
📎stderr.txt

特征子集模型测试

1 运行基于因果推断的特征重要性计算和选择代码

详见https://code.alibaba-inc.com/pai_biz_arch/feature_selection

2 上传新的config文件并在datawlaks里更改-Dconfig参数并运行节点



【第一组】

特征子集筛选及结果：

保留百分比70.47%

```
筛选的因果影响阈值为：  0.05
筛选之后的特征子集数量/全量特征：  284 / 403
字符串类型特征(全部保留)：  23
数值型特征(根据因果算法筛选)：  261
```

训练日志

https://logview.aliyun.com/logview/?h=http://service.ap-southeast-1.maxcompute.aliyun-inc.com/api&p=pai_test_xjp&i=20221016171159327glqledf03_eab0d39c_fa92_48c0_9d80_dd551da83d9f&token=MUNYK0RBUHp3RGJOMllrdFBUM3l0U0wyQzRvPSxPRFBTX09CTzoxNzMwNzYwMTM5MDc2MjYzLDE2Njg1MzIzMjQseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN0IjoiQWxsb3ciLCJSZXNvdXJjZSI6WyJhY3M6b2RwczoqOnByb2plY3RzL3BhaV90ZXN0X3hqcC9pbnN0YW5jZXMvMjAyMjEwMTYxNzExNTkzMjdnbHFsZWRmMDNfZWFiMGQzOWNfZmE5Ml80OGMwXzlkODBfZGQ1NTFkYTgzZDlmIl19XSwiVmVyc2lvbiI6IjEifQ==

特征子集带入排序模型的结果为

AUC（cilck）：0.8751（1-0.8751/0.8764=-0.15%，以下计算过程同理）

AUC（comment）：0.8620（+1.61%）

```
[2022-10-17 08:02:59,456] [INFO] [76#MainThread] [tensorflow/python/estimator/estimator.py:2003]
Saving dict for global step 25000: auc_is_click = 0.87512666, auc_is_comment = 0.8619958, global_step
= 25000, loss = 0.7633971, loss/loss/cross_entropy_loss_is_click = 0.40042582, loss/loss/
cross_entropy_loss_is_comment = 0.045534927, loss/loss/l2_loss_ln_stay_time = 0.15781039, loss/loss/
l2_loss_ln_video_play_time = 0.15962552, loss/loss/total_loss = 0.7633971,
mean_squared_error_ln_stay_time = 0.15780726, mean_squared_error_ln_video_play_time = 0.1596233
```

📎stderr (1).txt

【第二组】

特征子集筛选及结果：

保留百分比51.36%

```
筛选的因果影响阈值为：  0.1
筛选之后的特征子集数量/全量特征：  207 / 403
字符串类型特征(全部保留)：  23
数值型特征(根据因果算法筛选)：  184
```

训练日志

https://logview.aliyun.com/logview/?h=http://service.ap-southeast-1.maxcompute.aliyun-inc.com/api&p=pai_test_xjp&i=20221017165705819g78sedf03_f7d1103a_f8b7_4f78_8ca6_7768c755b86b&token=SkpYK0xTRVFyVHlzcENUNit1QndQQTFGN3Z3PSxPRFBTX09CTzoxNzMwNzYwMTM5MDc2MjYzLDE2Njg2MTc4MzMseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN0IjoiQWxsb3ciLCJSZXNvdXJjZSI6WyJhY3M6b2RwczoqOnByb2plY3RzL3BhaV90ZXN0X3hqcC9pbnN0YW5jZXMvMjAyMjEwMTcxNjU3MDU4MTlnNzhzZWRmMDNfZjdkMTEwM2FfZjhiN180Zjc4XzhjYTZfNzc2OGM3NTViODZiIl19XSwiVmVyc2lvbiI6IjEifQ==

特征子集带入排序模型的结果为

AUC（cilck）：0.8457（-3.5%）

AUC（comment）：0.8207（-3.25%）

```
[2022-10-18 09:12:14,941] [INFO] [76#MainThread] [tensorflow/python/estimator/estimator.py:2003]
Saving dict for global step 25000: auc_is_click = 0.84565073, auc_is_comment = 0.82074124,
global_step = 25000, loss = 2.2980962, loss/loss/cross_entropy_loss_is_click = 0.41666704, loss/loss/
cross_entropy_loss_is_comment = 0.047375597, loss/loss/l2_loss_ln_stay_time = 1.6553028, loss/loss/
l2_loss_ln_video_play_time = 0.17875202, loss/loss/total_loss = 2.2980962,
mean_squared_error_ln_stay_time = 1.6552429, mean_squared_error_ln_video_play_time = 0.17875153
```

📎stderr (2).txt

【第三组】

特征子集筛选及结果：

保留百分比22.33%

```
筛选的因果影响阈值为：  0.2
筛选之后的特征子集数量/全量特征：  90 / 403
字符串类型特征(全部保留)：  23
数值型特征(根据因果算法筛选)：  67
```

训练日志

https://logview.aliyun.com/logview/?h=http://service.ap-southeast-1.maxcompute.aliyun-inc.com/api&p=pai_test_xjp&i=20221018140853490goloedf03_e67bee1b_37ac_4aae_8c61_fc836a37ddd8&token=OFhaNVNZYWU2dklvLzBSeHhnb2Z5UFk3azhVPSxPRFBTX09CTzoxNzMwNzYwMTM5MDc2MjYzLDE2Njg2OTQxMzUseyJTdGF0ZW1lbnQiOlt7IkFjdGlvbiI6WyJvZHBzOlJlYWQiXSwiRWZmZWN0IjoiQWxsb3ciLCJSZXNvdXJjZSI6WyJhY3M6b2RwczoqOnByb2plY3RzL3BhaV90ZXN0X3hqcC9pbnN0YW5jZXMvMjAyMjEwMTgxNDA4NTM0OTBnb2xvZWRmMDNfZTY3YmVlMWJfMzdhY180YWFlXzhjNjFfZmM4MzZhMzdkZGQ4Il19XSwiVmVyc2lvbiI6IjEifQ==

特征子集带入排序模型的结果为

```
[2022-10-19 07:53:17,731] [INFO] [76#MainThread] [tensorflow/python/estimator/estimator.py:2003]
Saving dict for global step 25000: auc_is_click = 0.8423181, auc_is_comment = 0.81068414, global_step
= 25000, loss = 2.3382466, loss/loss/cross_entropy_loss_is_click = 0.42462486, loss/loss/
cross_entropy_loss_is_comment = 0.047913395, loss/loss/l2_loss_ln_stay_time = 1.6919271, loss/loss/
l2_loss_ln_video_play_time = 0.17378066, loss/loss/total_loss = 2.3382466,
mean_squared_error_ln_stay_time = 1.6918635, mean_squared_error_ln_video_play_time = 0.17377973
```

AUC（cilck）：0.8423（-3.89%）

AUC（comment）：0.8106（-4.44%）

📎stderr (3).txt

结果总结

注:括号里是相对变化	全量特征	特征子集1(阈值0.05)	特征子集2(阈值0.1)	特征子集3(阈值0.2)
特征个数	483（100%）	284（70.47%）	207（51.36%）	90（22.33%）
点击AUC	0.8764	0.8751（–0.15%）	0.8457（–3.5%）	0.8423（–3.89%）
评论AUC	0.8483	0.8620（+1.61%）	0.8207（–3.25%）	0.8106（–4.44%）