京东 JData 算法大赛【高潜用户购买意向预测】

梅尚健

PART1：代码运行说明

1. 总体解决方案

本文将高潜用户购买意向预测，抽象为一个二分类问题。从用户，商品，品牌，用户-商品，用户-品牌五个维度进行特征提取。将观察天未来 5 天有购买行为的用户-商品对标记为正样本，观察天过去 30 天至未来 5 天有交互行为但未购买的用户-商品对标记为负样本。由于正负样本比例极不平衡，采用了对正样本进行重采样及负样本进行下采样的方式来平衡正负样本比例。利用 xgboost 进行模型训练，最后利用 LR 对预测结果进行加权。取每个用户最高预测概率对应的 user-sku 对，取 top12000 作为最终输出结果。

2.实现方案技术栈

集市堡垒机(环境) + Hive(ETL) + Spark& Spark-Xgboost(Model)

3.代码运行说明

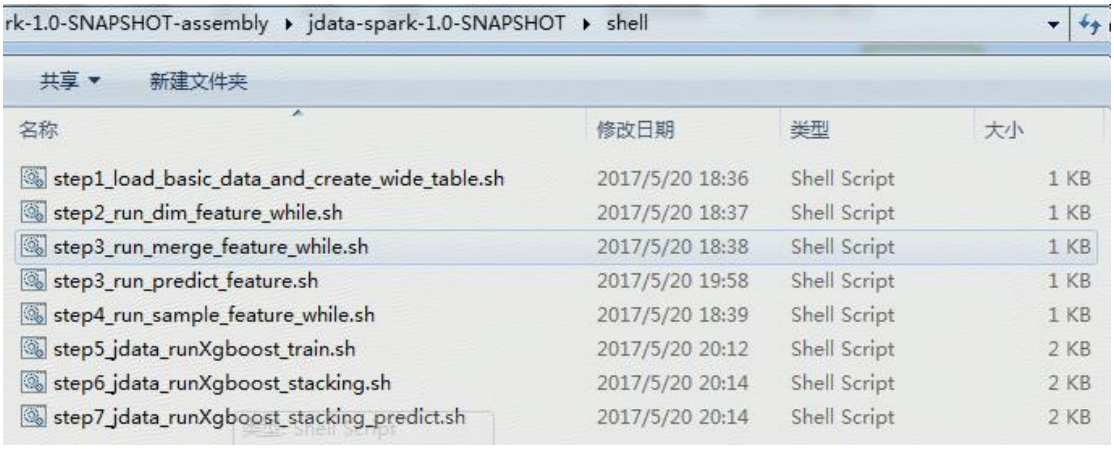
3.1 工程代码目录：JData-Spark

使用 maven 进行源码编译，成功编译后会生成 jdata-spark-1.0-SNAPSHOT-assembly.zip

jdata-spark-1.0-SNAPSHOT-assembly.zip 目录结构如下：

3.2 脚本运行

进入到 shell 子目录，目录下各步骤脚本如下：



各个步骤运行说明，各个步骤间有依赖，每个步骤运行完再运行下一个步骤**【首先需要将比赛数据集解压后手动放到 data 目录下】**:

3.2.1 step1_load_basic_data_and_create_wide_table.sh：加载数据集到 hive 表，并将 user/product/action 汇总加工成为一张基础数据宽表。

运行命令：nohup sh step1_load_basic_data_and_create_wide_table.sh > tmp.log &

3.2.2 step2_run_dim_feature_while.sh：加工 user/sku/brand/user_sku/user_brand 各维度特征，由于后续使用了滑窗集成，所以需要运行多份(part1-part12)。

运行命令：nohup sh step2_run_dim_feature_while.sh > tmp.log &

3.2.3 step3_run_merge_feature_while.sh：将各个维度的特征汇总为一张宽表，便于后续进行抽样及模型训练。由于后续使用了滑窗集成，所以需要运行多份(part1-part12)。

运行命令：nohup sh step3_run_merge_feature_while.sh > tmp.log &

step3_run_predict_feature.sh：预测指标加工

运行命令：nohup sh step3_run_predict_feature.sh > tmp.log &

3.2.4 step4_run_sample_feature_while.sh：对加工的特征宽表进行采样

运行命令：nohup sh step4_run_sample_feature_while.sh > tmp.log &

由于模型训练是在风控集市跑的 spark 任务，因此若要让脚本在其它集市可用，需要手动修改脚本及配置文件【影响步骤 3.2.5，3.2.6，3.2.7】

```
#!/bin/sh
numExecutor=50
executorMemory=20
driverMemory=10
etlDate=20160406
taskType=train_xgboost
spark-submit  \
--master  yarn-cluster \
--name  "SparkModelStackingMain" \
--class com.sjmei.jdata.xgboost.SparkModelStackingMain \
--properties-file ../conf/jdata/spark-defaults-jdata.conf \
--num-executors ${numExecutor}  \
--executor-memory ${executorMemory}g  \
--driver-memory ${driverMemory}g  \
--jars ../target/scopt_2.11-3.5.0.jar,../target/xgboost4j-0.7-jar-with-dependencies.jar,../target/xgboost4j-spark-0.7.jar,../target/velocity-1.7.jar \
--queue bdp_jmart_risk.bdp_jmart_risk_formal \
--files ../conf/graph-jdata.properties,../conf/hive-site.xml,../script/mid_result_script/model_blend_feature.sql,../script/mid_result_script/model_blend_pred_feat
../target/jdata-spark-1.0-SNAPSHOT.jar  \
--nWorkers 50 \
--initDate 2016-04-11 \
hdfs://ns2/user/mart_risk/dev.db/dev_temp_msj_risk_jdata_feature_train_v3_sample \
hdfs://ns2/user/mart_risk/sjmei/tests/xgboost/model_stacking_v3_${etlDate} \
hdfs://ns2/user/mart_risk/sjmei/tests/xgboost/result_stacking_v3_${etlDate} \
${taskType} \
>./jdata_runXgboost_blend_${taskType}.`date +%Y%m%d`.log 2>&1
```


3.2.5 step5_jdata_runXgboost_train.sh：利用 spark-xgboost 进行模型训练

运行命令：nohup sh step5_jdata_runXgboost_train.sh > tmp.log &

3.2.6 step6_jdata_runXgboost_stacking.sh：将 spark-xgboost 训练好的 6 个子模型，进行 stacking LR 融合
运行命令：nohup sh step6_jdata_runXgboost_stacking.sh > tmp.log &

3.2.7 step7_jdata_runXgboost_stacking_predict.sh：利用 3.2.6 训练好的融合模型，预测用户在 4.16-4.20 会购买的 user_sku 对
运行命令：nohup sh step7_jdata_runXgboost_stacking_predict.sh > tmp.log &
3.2.7 步骤运行完后，需要手动将生成 txt 格式结果从 hdfs 目录 down 下来，并按比赛要求的格式整理输出
hadoop fs –get hdfs://ns2/user/mart_risk/sjmei/tests/xgboost/result_stacking_v3_20160406

PART2：解题思路

1. 总体解决方案

本文将高潜用户购买意向预测，抽象为一个二分类问题。从用户，商品，品牌，用户-商品，用户-品牌五个维度进行特征提取。将观察天未来 5 天有购买行为的用户-商品对标记为正样本，观察天过去 30 天至未来 5 天有交互行为但未购买的用户-商品对标记为负样本。由于正负样本比例极不平衡，采用了对正样本进行重采样及负样本进行下采样的方式来平衡正负样本比例。利用 xgboost 进行模型训练，最后利用 LR 对预测结果进行加权。取每个用户最高预测概率对应的 user-sku 对，取 top 12000 作为最终输出结果。

2.特征工程

从用户、商品、品牌、用户-商品交互、用户-品牌交互 5 个维度，再对各个维度从不同周期(1/3/5/7/10/15/30 天)进行建模及特征提取。

用户维度(script/dim_feature_v3/dim_user_feature_etl.sql)：用户等级，性别，注册天数，年龄等级，浏览量，点击量，关注量，加购车量，下单量，取消关注量，点击购买率，点击加购率，点击关注率，浏览购买转化率，加购购买转化率，关注购买转化率，浏览 day/sku/brand 数，点击 day/sku/brand 数，关注 day/sku/brand 数，加购车 day/sku/brand 数，下单 day/sku/brand 数，取消关注 day/sku/brand 数，最近浏览/点击/购买/关注距今天数，平均每天浏览/点击/购买/关注量(sku/brand 数)，平均浏览/点击/购买/关注行为操作间隔天数
商品维度(script/dim_feature_v3/dim_sku_feature_etl.sql)：商品评论数，好评率，差评率，商品属性 1，属性 2，属性 3，浏览量，点击量，关注量，加购车量，下单量，取消关注量，点击购买率，点击加购率，点击关注率，浏览购买转化率，加购购买转化率，关注购买转化率，点击购买用户占比，点击加购用户占比，点击关注用户占比，浏览购买转化用户占比，加购购买转化用户占比，关注购买转化用户占比，浏览用户数，点击用户数，关注用户数，加购车用户数，下单用户数，取消关注用户数，平均每天浏览/点击/购买/关注量(用户数)，

平均每个用户浏览/点击/购买/关注量

品牌维度(script/dim_feature_v3/dim_brand_feature_etl.sql)：浏览量，点击量，关注量，加购车量，下单量，取消关注量，点击购买率，点击加购率，点击关注率，浏览购买转化率，加购购买转化率，关注购买转化率，点击购买用户占比，点击加购用户占比，点击关注用户占比，浏览购买转化用户占比，加购购买转化用户占比，关注购买转化用户占比，浏览用户数，点击用户数，关注用户数，加购车用户数，下单用户数，取消关注用户数，平均每天浏览/点击/购买/关注量(用户数)，平均每个用户浏览/点击/购买/关注量，商品热度(点击量*0.01+下单量*0.5+加购量*0.1-取消关注量*0.1+关注量*0.1)

用户-商品交互维度(script/dim_feature_v3/dim_user_sku_feature_etl.sql)：浏览量，点击量，关注量，加购车量，下单量，取消关注量，点击购买率，点击加购率，点击关注率，浏览购买转化率，加购购买转化率，关注购买转化率，浏览 day 数，点击 day 数，关注 day 数，加购车 day 数，下单 day 数，取消关注 day 数，最近浏览/点击/购买/关注距今天数，平均浏览/点击/购买/关注行为操作间隔天数

用户品牌交互维度(script/dim_feature_v3/dim_user_brand_feature_etl.sql)：浏览量，点击量，关注量，加购车量，下单量，取消关注量，点击购买率，点击加购率，点击关注率，浏览购买转化率，加购购买转化率，关注购买转化率，浏览 day 数，点击 day 数，关注 day 数，加购车 day 数，下单 day 数，取消关注 day 数，最近浏览/点击/购买/关注距今天数，平均浏览/点击/购买/关注行为操作间隔天数

交叉类特征(script/dim_feature_v3/feature_wide_table.sql)：用户-商品与用户浏览/点击/关注/加购/下单/取消关注占比，用户-品牌与用户浏览/点击/关注/加购/下单/取消关注占比 etc.

3. 样本选择及特征预处理

1. 将观察天未来 5 天有购买行为的用户-商品对标记为正样本，观察天过去 30 天至未来 5 天有交互行为但未购买的用户-商品对标记为负样本。
Eg.(将 2016-04-06~2016-04-10 有下单的用户-商品标记为正样本，将 2016-03-06~2016-04-10 有交互但未下单的用户-商品标记为负样本)。

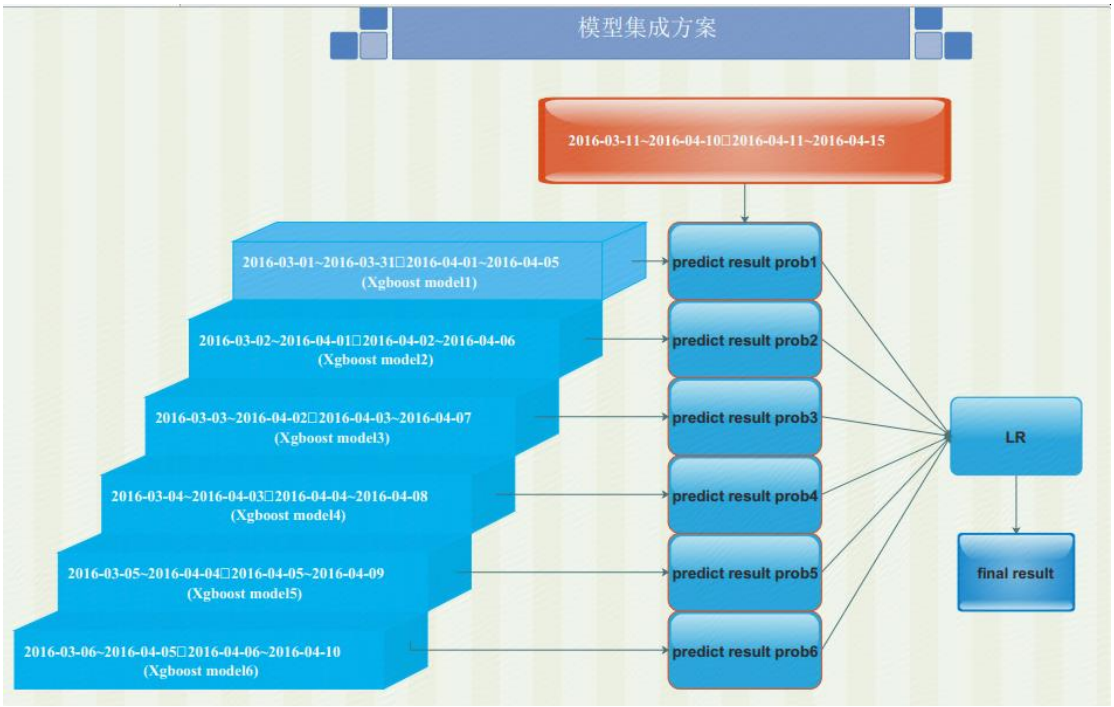
2. 由于正负样本比例极不平衡，正负样本比例约为 1500:2400000，采用了对正样本进行重采样及负样本进行下采样的方式来平衡正负样本比例。
具体采样方式为：将正样本复制 10 份，同时将负样本通过随机采样为 200000

3. 特征预处理：对于类型型特征：通过 Spark VectorIndexer 进行 one-hot 编码。对于连续型特征，通过 Spark Normalizer 进行归一化。

4. 模型集成方案

利用 spark-xgboost 进行滑窗模型训练，最后利用 LR 对各个预测结果进行加权。取每个用户最高预测概率对应的 user-sku 对，取 top 12000 作为最终输出结果。

单模型准确率 xgboost>gbdt>rf，因此只选择了 xgboost 以滑窗方式进行模型训练，未采用多模型融合 stacking，融合过程也只用了 lr 加权融合，其它方式待尝试。



5. 总结

第一次参加数据挖掘类比赛，作为一名新手，通过本次比赛，学习了数据挖掘的整个流程，同时也进一步熟悉了 spark ml 框架的使用。其次，更多的是在实践过程中体会到了自身的不足，要想打好比赛，必须源于对业务的深入理解及数据的细致分析，而这一点恰恰是做的最不好的。比赛中没有花很多时间对数据进行深入理解与细致分析。在特征处理，调参方面也做的很糙。 solo 比赛很累，思维也很受限，只知道堆特征+xgboost+lr 融合的方案，看到排行榜上其他同学的成绩都在噌噌地往上涨，而自己又不知道该如何优化涨分，成绩也一直停滞不前，以后要多向大牛学习以及多和别的同学一起交流学习。

作为一名 CS 专业毕业的人，对于使用各种数据挖掘工具进行技术实现不是什么问题，但其实对于数据挖掘来说，更重要的还是分析建模能力，对业务的感知能力，自己这方面还很欠缺，今后需要多多加强。

感谢公司举办的此次大赛，让我获益良多！