1、 精简版的特征、模型与结果

1）最少需要保留那些处理方法和特征，才能保证结果在 90%以上？

最好是保留全部的特征，我们小组对特征的重要性进行过排序，根据特征的重要性对特征进行删选，然后效果不好。

2）最重要的模型是哪个？

在这次比赛中，我们使用了两大类模型，一个是对用户进行预测的 User 模型(Umodel_0, Umodel_1, Umodel_2)，一个对用户和商品对进行预测的 US 模型(USmodel)。

开始我们只有一个 Umodel_0 模型，然后根据线上的结果，我们发现 Umodel_0 的模型不够好，然后我们就对 Umodel_0 增加一些特征，从而生成 Umodel_1 和 Umodel_2，进而提高 U 模型的效果。

2、可运行的代码：

1）训练（模型生成）部分

A. 必要注释。

2）预测部分（为了测试选手结果的真实性，请提供可编译/运行的预测代码）

A. 输入：原始数据，处理后数据或提取的特征；

京东比赛方提供的原始数据。放在项目的/data 目录下。

B. 输出：预测 csv，结果应与排行榜提交一致；

输出结果在 sub/best_result.csv

C. 说明文档：

a）描述编译/运行预测代码需要的资源；

软件：XGboost0.5 ,Anaconda3,Ubantu 16.04， jupyter

硬件：64G 内存，500G 磁盘

b）代码使用说明：如何才能运行提供的预测代码

运行 start.sh 即可。

3、特征工程

赛题提供了 user 表、 sku 表、comment 表以及 2016-02-01 ~ 2016-04-15 的用户商品行为数据，可以从 user 表中提取用户的基本信息特征，从 sku 表和 comment 表提取商品的基本信息以及商品的评价情况，从 2016-02-01 ~ 2016-04-15 的用户商品行为数据中提取到更加丰富的特征：用户相关的特征，商品相关的特征，用户-商品交互的特征。

1）关键特征

- user feature
  - 用户年龄特征
  - 用户性别
  - 用户等级特征（等级 2）
  - 注册时间与截止日期的时间间隔（天）
  - 用户前 1/2/3/7/14/28 天各行为次数
  - 用户的行为转化率

    （行为 4/行为 1 、行为 4/行为 2 、行为 4/行为 5、行为 4/行为 6、行为 3/行为 2）

- 用户购买 /加购/关注前浏览天数
- 用户购买 /加购/关注前浏览次数
- 用户平均访问时间间隔
- 用户六种行为的平均访问时间
- 用户前 1/2/3/7/14/28 天 6 种行为 0/1 提取
- 用户最早/最近一次行为时间距离最后日期的时间（精确到小时）
- 用户最后一次行为的次数
- 用户层级 2/3/7/14/28 各行为天数
- 用户各行为/总行为的比值
- 用户前 1/2/3/7 天访问 P 集合的商品数
- 用户前 14/28 天访问 P 集合的商品数/用户访问总体的商品数
- 用户购买每种 cate 的数量
- 用户子集全集的活跃天数
- 用户前 1/2/3/7/14/28 天购买/加购/关注/点击/浏览品牌数
- 用户点击各模块的数量（模块 14、21、28、110、210）/点击所有模块的数量
- 用户购买 cate8 的数量占购买数量的比率
- 用户子集行为与全集行为比值

- sku feature
  - 商品前 1/2/3/7/14/28 天行为次数总和
  - 商品类别特征独立编码
  - 商品行为的转化率
  - （行为 4/行为 1 、行为 4/行为 2 、行为 4/行为 5、行为 4/行为 6、行为 3/行为 2）
  - 商品购买/加购/关注前访问天数
  - 商品购买/加购/关注前访问次数
  - 商品平均访问间隔
  - 商品六种行为平均访问间隔
  - 商品前 1/2/3/7/14/28 天 6 种行为 0/1 提取
  - 商品的重复购买率
  - 商品最近一次行为的时间距离当前日期的时间
  - 商品最近一次行为的行为次数
  - 商品的层级 2/3/7/14/28 各行为天数
  - 商品各行为/总行为的比值
  - 商品从点击到购买的时间间隔
  - 商品前 1/2/3/7/14/28 天总购买/加购/关注/点击/浏览品牌数
  - 商品点击各模块的数量（模块 14、21、28、110、210）/点击所有模块的数量
  - 商品被购买前发生的 6 种行为次数的平均值、最小值、最大值
  - 商品的 6 种行为频率

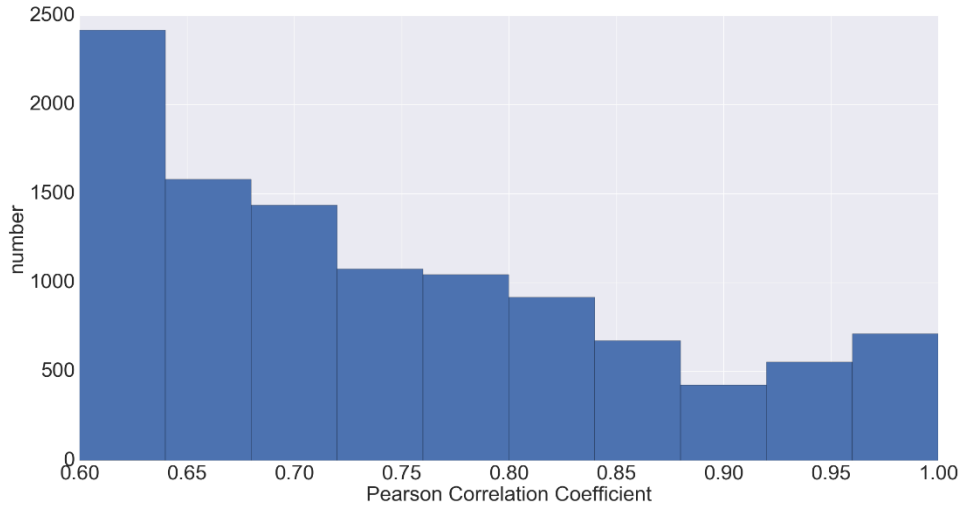
- user-sku feature
  - 该用户对该商品的行为总和
  - 用户商品行为衰减
  - 用户商品对的 1/2/4/5 行为/用户对应行为
  - 用户关注或加购，但是不购买，且加购或关注天数距离最后日期小于 10 天的记为 1，否则记为 0
  - 用户商品行为 0/1 提取
  - 用户商品最近一次行为时间距离最后日期的时间差
  - 用户商品最近一次行为次数
  - 用户商品 2/3/7/14/28 行为层级天数
  - 用户商品 6 种行为频数
  - 用户商品各行为/总行为
  - 用户商品各点击模块/总点击模块
  - 该用户购买该商品从点击到购买的时间间隔
  - 用户对该商品的行为比率（type i /type）
  - 用户购买该商品前 k 天的 6 种行为

2）特征思考与特征获取

分别对 user 、sku 、user-sku 三种 key 进行提取，各特征均是基于京东业务进行思考和提取的。首先，用户的特征，要从消费者自身考虑，用户的购买习惯，购买行为差异等等，因此需要对用户进行建模，提取用户的偏好。用户发生购买行为之前的一系列行为非常重要，比如：一些用户喜欢在购买之前先浏览详情页，点击详情页的某些模块，以及用户从浏览该商品到购买该商品的时间差；一些用户浏览该商品时，并不是急着需要，因此选择关注。对用用户这块，我们相当于构造用户画像。对于商品，类似于用户操作，刻画出商品的画像，比如，商品对于用户的喜爱程度，商品的好评率等等。同时，商品的类别也非常重要，比如说用户购买的是冰箱电脑，短时间类不会购买，由于主办方将数据处理后，我们通过数字独立编码，将其变成特征。对于用户商品特征，不同的用户对于不同的商品偏爱不同,这个 key 细致的刻画了该用户对于该商品的思想观念。这次比赛的特征均从中得到思路，得到解决方案。

3）特征相关关系

特征之间有一定的相关关系，在比赛中我们队伍使用皮尔逊相关系数(Pearson Correlation Coefficient)来衡量特征之间的相关性



横坐标代表相关系数的绝对值，越高代表相关性越大，模型的特征两两组合一共有50W 左右的可能取值。其中相关系数大于 0.6 的组合数量如上图所示。可以发现大约有10000+特征组合之间具有很强的相关性。我们尝试过去掉这些相关性很高的特征，但模型效果提升微小。

4）特征处理

A. 为什么选择这个模型？

我们使用的模型是 xgboost，它基本思想是把多个分类准确率较低的树模型组合起来，成为一个准确率很高的模型。近年来不断有队伍借助 xgboost 在比赛中 夺得冠军。

xgboost 的主要优点有：

1. xgboost 借助 OpenMP，能自动利用单机 CPU 的多核进行并行计算。xgboost 的并行是在特征粒度上的。
2. xgboost 自定义了一个数据矩阵类 DMatrix，会在训练开始时进行一遍预处理，从而提高之后每次迭代的效率。
3. xgboost 在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数、每个叶子节点上输出的 score 的 L2 模的平方和，使学习出来的模型更加简单，防止过拟合。
4. 剪枝：XGBoost 先从顶到底建立所有可以建立的子树，再从底到顶反向进行剪枝。比起 GBM，这样不容易陷入局部最优解。

B. 模型的训练方式？

首先，从下表所列的特征区间中提取 Umodel 的训练集、验证集的特征和标签，对 Umodel 进行 xgboost 模型训练以及线下验证，然后提取预测区间的特征作为训

练好的Umodel的输入，输出即为预测会购买用户的概率值排名；同理，可得USmodel的输出，即为预测会购买用户商品对的概率值排名。

	特征区间（提取 feature）	预测区间（提取 label）
Umodel 训练集	2016-03-09 到 2016-04-10	2016-04-11 到 2016-04-15
Umodel 验证集	2016-03-04 到 2016-04-05	2016-04-06 到 2016-04-11
Umodel 测试集	2016-03-15 到 2016-04-15	2016-04-16 到 2016-04-20
USmodel 训练集	2016-03-30 到 2016-04-10	2016-04-11 到 2016-04-15
USmodel 验证集	2016-03-24 到 2016-04-05	2016-04-06 到 2016-04-11
USmodel 测试集	2016-04-04 到 2016-04-15	2016-04-16 到 2016-04-20

C. 是否进行了模型融合？模型的融合方式？

在这次比赛中，我们基于 xgboost 训练了两类模型，一类是对用户进行预测的 User 模型(Umodel_0, Umodel_1, Umodel_2)，一类是对用户商品对进行预测的 US 模型（USmodel），然后对两类模型进行了融合。

首先，Umodel_0*0.4 + Umodel_1*0.3 + Umodel_2*0.3 得到三个 Umodel 融合后的预测会购买用户的概率值排名，取 top700。然后将这 top700 的用户与 USmodel 的预测会购买用户商品对进行 merge，最后将 merge 之后的结果与 USmodel 的 top325 进行合并去重，即可得到最终提交结果。

5）有趣的发现

A、使用了 MySQL 数据库的基本功能，自己编写了小黑管家程序对数据进行前期处理数据分析比对，历史同期比对等，这样可以针对一些特征的关联性进行了比对校验，例如 user_lv_cd 特征单独使用，要比 user_lv_cd 加上 sex 特征同时使用效果要好。

B、对数据进行了清洗，同一秒同一个人相同的操作视为无效操作直接去除。对用户年龄进行了转化，转化为-1，0，1，2，3，4，5，6 这样便于模型操作。没有注册日期的进行了处理赋值成-1，其他日期转成类别特征，注册日期距截至日期不同天数的赋个类别 1，2，3。

C、最突出的优势在于项目的分析，针对这个项目，我们从三方面进行了分析，一、业务，二、京东的系统以及运营方法，三、模型算法。通过使用小黑对数据汇总分析历史同期比对，得出了些强特征。我们对原始数据进行了全表的汇总记录分析，这为以后的分析处理节省了大量时间。