**1）概述**

**A. 采用的模型？**

在此次比赛中，针对数据量较大，训练速度较慢的情况，选择lightgbm作为基础训练模型。

最终，b榜所提交最优结果包含4个模型，按照F11（user+cate）、F12(user+cate+shop)分为两类，分别为：基于7天标签集的对F11进行预测的F11\_7模型，基于7天标签集的对F12进行预测的F12\_7模型；基于5天标签集的对F11进行预测的F11\_5模型，基于5天标签集的对F12进行预测的F12\_5模型.

**B. 重要的特征和数据清洗方式？**

**数据清洗：**

1.进过数据EDA发现，用户行为中的type=5的情况（加入购物车）仅在4.08号之后出现，无法在全集下进行该行为特征的提取，初步清洗进行type=5的数据的删除，后续单独进行特征提取。

2. 发现用户行为表中存在商品id(sku\_id)未在商品表中出现，针对这部分数据进行剔除。

3. 发现用户在cate:13下从未发生购买情况，对着两类品类数据进行删除。

4. 发现在vender\_id = 3666下从未发生购买行为，且占据整体数据的1/3，对其进行剔除，提高训练速度，并提高预测精度。

5.考虑到二月中旬为年假时间，aciton数据分布与其他月份存在较大分布，经过EDA后，发现数据在2月22日之后分布趋于稳定，故而在某些特征的构造上尽量避免使用2月22日之前的数据。

**特征工程：**

整体特征的构造包含三大类：第一类是基于原数据表的所有数据，进行统计特征的提取；另一类是基于用户的中期、短期行为记录，进过业务分析对电商实际场景的考虑提取和建立的相关行为统计特征，分别在： user\_id、 cate、 shop\_id、F11（user\_id+cate）、F12(user+cate+shop)5个维度进行构造；第三类是购物车特征。

主要包含：

**第一类：基于数据表的统计特征**

1. 基础用户表特征：

用户年龄段、性别、会员等级、城市等级、省的onehot特征；

用户所在省、市的出现频次；

用户注册时间距今时间差。

2. 基础行为表特征：

按照前3、5、7、21、30天进行滑窗统计：

基于user\_id、cate、shop\_id、F11、F12统计的用户各行为的总次数；

基于user\_id、cate、shop\_id、F11、F12统计的用户各行为的转化率（次数比值）；

按照前1、7天和全集进行滑窗统计：

基于user\_id统计的历史行为总次数及比值，与行为时间进行交叉（相除）；基于前1、7天和全集统计行为的变化（相减）。

基于行为时间的统计特征：用户活动天数、购买天数、最近活动时间、最近购买时间。

3. 基础店铺表特征：

店铺vip转化率特征（vip/fans）；

店铺开店时间距今时间差；

店铺所属品类的出现频次；

店铺下的购买次数的均值、总值。

4. 基础评论表特征：

品类下的商品数、品牌数、店铺数、评论数、好评数、差评数、好评/评论。

5. 多数据表交互统计特征：

用户行为涉及品类数、店铺数、商品数；

品类下用户转化率（购买用户/出现用户）、店铺转化率（购买店铺/出现店铺）、商品数、店铺数；

店铺下用户转化率（购买用户/出现用户）；

用户、品类、店铺出现频次的交叉（相除）。

**第二类：业务特征（按上述5个维度进行统计）**

1. 购买前浏览天数、关注前访问天数、购买前访问天数；

2. 平均访问间隔；

3. 六种行为下的平均访问间隔；

4. 购买频率；

5. 最近1/2/3/4/5/6/7/15/30天的行为的0/1提取；

6. 重复购买率；

7. 最近一次行为距今时间差；

8. 购买前浏览次数、关注前访问次数、购买前访问次数；

9.六种行为的交叉；

10. 最后一次行为的次数；

11. 浏览到购买的时间间隔、关注到购买的时间间隔；

12. 最早活动时间、最晚活动时间、总活动时间。

13. 按照3/7/15的时间粒度，对30天的行为进行0/1提取；

**以下为在user\_id维度的附加特征：**

1. 用户总购买、关注、浏览、评论的品类数、店铺数；

2. 用户在购买、浏览、关注行为下的F12维度行为次数/用户总；

行为次数（品类总行为次数、店铺总行为次数）。

**第三类：****加入购物车特征**

1. 加购（加入购物车）行为在5个维度下的次数及转化率（加购次数与其他行为次数的比值）；

2. 加购次数与购买、未购买行为的交叉；

3. sku维度的统计。

C. 特征是否经过处理？

部分特征数值大小存在归一化处理

D．使用的工具？

sklearn库下的最大最小归一化函数

**2）数据处理**

**A. 数据的处理？**

1. 主要为数据的清洗，如前；

2. 数据的缺失根据具体场景，使用众值、平均值、赋0等综合考虑，进行填充；

3. 7天标签集模型：使用后7天作为标签集提取label、标签集前30天作为用户集提取F12、标签集前所有时间作为特征集提取特征；

5天标签集模型：使用后5天作为标签集提取label、标签集前30天作为用户集提取F12、标签集前所有时间作为特征集提取特征

4. 在进行用户集提取后，按照正负样本（购买：未购买）比例1:30，对负样本进行下采样；

**B. 为什么要使用这些数据？数据的作用？**

1.使用全量数据(约两个月)构造统计特征，是为了对用户、商铺、品类的长效特征进行画像；

2.使用线上测试集日期前一个月左右的action出现的用户作为训练集，减少了线上线下的时间距离，使线上与线下的分布尽可能一致；考虑到用户的部分行为记录，未来一定时间内会影响用户的下一步行为，故而构造用户中期、短期行为特征所使用的action数据的时间需要和线上测试集尽可能的接近；

3. 提取前一个月的F12作为训练集索引，保证是近期用户的同时获取充足的训练集。

4. 进行训练集负样本的降采样，在保证了正样本数量的同时，获取最适合lightgbm模型训练的数据集，同时大幅提供训练速度。

**3）特征选择与获取**

**A. 哪些特征是关键特征？**

上述特征均为测试后的有效特征，其中基于5个维度提取的业务特征以及行为记录特征最为关键、其次为加购特征和多数据表的交互统计特征。

**B. 特征是如何想到和获取的？**

本赛题特征工程的基本思想为：结合基本的统计特征，通过使用户的行为记录，在长效、中期、短期三个时间维度，对用户、品类、商铺、F11(消费者和品类)、F12(消费者、品类、店铺）进行表示；同时合理利用type==5存在的最后八天数据，围绕type==5建立特征工程以及新的模型。

因此本次比赛的特征工程共分为三类。

第一类使用所有数据中各数据表相应的统计特征，目的是对相应用户、品类、商铺的长效特征进行刻画。

第二类特征的目的是根据用户中期、短期行为，对用户中期以及短期的兴趣特征进行刻画。结合赛题，进行了消费者、商品品类、店铺、F11(消费者和品类)、F12(消费者、品类、店铺）进行多个角度的挖掘。首先从消费者自身考虑，分析用户的购买习惯、行为习惯、购买的趋势、行为变化、行为积累从而产生购买等。主要考虑到用户行为的积累和趋势性，如用户在购买前其他行为的情况、用户的购买频率、购买间隔、其他行为到产生购买的时间和积累次数等等。从而完整得进行用户画像的构造，然模型有效学习用户行为从积累到变化的过程。之后从用户画像开始发散思维，结合赛题需求，加入品类画像、店铺画像、F11行为画像、F12行为画像的构造，从多个角度进行全面的赛题剖析，从而获取一个完备、有效的解决方案。

第三类特征是针对数据EDA后发现加入购物车行为与购买行为具有很强的关联。同时结合对业务场景的思考，考虑到加入购物车行为是用户行为积累过程中承上启下的关键一环。因此，针对加入购物车行为的缺失，对购物车行为进行单独提取，并进行该行为与其他行为的交互；同时针对该数据量小，对模型进行适应性修改，减少标签集天数，扩增购物车数据，获取更多的加入购物车行为信息，在b榜有一个千的提升。

**C. 特征之间是否有相关关系？**

第一类特征在各个数据表下提取，存在交互性但相互关联不大；第二类特征相互补充丰富，协同构建了五维画像，并且在构建画像上，后续特征可以在前面特征的基础上进行补充延伸，使得行为积累的过程被成功刻画出来。第三类特征补充了前面两类特征在加购购物车信息上的缺失，同时也是针对数据本身缺陷的一个解决方案。

**D. 特征是否经过处理？**

大多数特征经过了归一化处理

**4）模型选择与训练**

**A. 为什么选择这个模型？**

我们使用的是lightgbm模型，该模型与xgboost相比，有速度更快，准确率更高的优势。具体如下

1. 更快的训练速度和更高的效率：LightGBM使用基于直方图的算法。例如，它将连续的特征值分桶(buckets)装进离散的箱子(bins)，这是的训练过程中变得更快，同时加强了泛化能力。
2. 更低的内存占用：使用离散的箱子(bins)保存并替换连续值导致更少的内存占用。
3. 更高的准确率(相比于其他任何提升算法) ：它通过leaf-wise分裂方法产生比level-wise分裂方法更复杂的树，这就是实现更高准确率的主要因素。然而，它有时候或导致过拟合，但是我们可以通过设置max-depth参数来防止过拟合的发生。
4. 大数据处理能力：相比于XGBoost，由于它在训练时间上的缩减，它同样能够具有处理大数据的能力。
5. 支持并行学习

更快的训练速度以及可观的性能（一般情况与XGBoost相同），是我们选择lightgbm的最主要的原因。

**B. 模型的训练方式**

利用下表所列的标签集提取训练集标签，用户集提取训练集F12索引（F12模型）、F11索引（F11模型），特征集提取上述特征，然后进行拼接降采样，测试集进行索引和特征提取，然后进行lghtgnm模型训练，输出即为预测会发生购买的F12（F11）的概率值排名，取前k（可32000）个，然后进行结果的融合。

七天模型

|  |  |  |  |
| --- | --- | --- | --- |
|  | 标签集（提取label） | 用户集(提取索引) | 特征集（提取特征） |
| 训练集 | 2018-04-09到2018-04-16 | 2018-03-08到2018-04-09 | 2018-02-01到2018-04-09 |
| 测试集 | 无 | 2018-03-15到2018-04-16 | 2018-02-01到2018-04-16 |

五天模型

|  |  |  |  |
| --- | --- | --- | --- |
|  | 标签集（提取label） | 用户集(提取索引) | 特征集（提取特征） |
| 训练集 | 2018-04-11到2018-04-16 | 2018-03-10到2018-04-11 | 2018-02-01到2018-04-11 |
| 测试集 | 无 | 2018-03-15到2018-04-16 | 2018-02-01到2018-04-16 |

**C. 是否进行了模型融合？模型的融合方式？**

进行了模型融合，首先分别进行5天、7天标签集下的F11模型、F12模型的融合（利用F11模型结果对F12模型结果进行筛选），获取两个结果res\_7、res\_5；然后取res\_5前10000条与res\_7相组合，再去重，获取最优的预测结果。

**5） 说明文档**

**A. 描述编译/运行预测代码需要的资源和库以及版本备注；**

python3.6、datetime、pandas0.24.2、pickle、os、numpy1.16.2 、sklearn、lightgbm、dateutile、warnings

**B．代码使用说明：如何才能运行提供的预测代码。**

1. 建立data文件夹：放入解压后的训练数据。

2. 建立cache文件夹，存储中间结果。

3.建立res文件夹，存储最终结果

4. 运行get\_res.py，可以获得多组结果，其中merge5day\_10k.csv和merge\_2为最后一次B榜提交的两个结果（merge5day\_20k.csv效果可能更佳，尚未测试）。

**6）有趣的发现**

**A. 使用的小技巧？**

充分的EDA减少数据集有偏采样对模型带来的负面影响

训练数据的负采样

**B. 您觉得您最突出的优势是什么？**

1. 在长效、中期、短期三个时间维度以及用户、商铺、品类、F11、F12维度建立了全面的特征表示。

2. 对加入购物车特征的合理利用，对原有模型进行改进，同时对建立了能更好利用加入购物车特征的新模型与原模型互补，大幅提高了线上成绩。

**C. 每次成绩提升改进说明。**

1. 多个数据表的基础统计特征的不断添加：提升8个千

2. 数据清洗：提升两个千

3. 用户画像构建，用户积累特征的提取：提升4个千

4. 基于3，发散到其他维度，进行积累特征的提取：提升5个千

5. 针对购物车特征的提取：提升3个千

6. F11、F12双模型的构建与融合：提升两个千

7. 针对购物车特征，5天模型的构建与融合：提升一个千

**7）其他您想分享的**

在某些情况下使用Count map对category feature进行处理非常有效。

进行模型融合时，使用加权融合效果一般，使用新模型结果前k条对原始最优模型的结果进行剔除、筛选效果更佳。

结合赛题以及线上分数的反馈，发现在F11分数、F12分数以及召回率、准确率上的权衡，对最终分数影响很大。