**“京东杯”2019第六届泰达创新创业挑战赛**

**比赛攻略**

# 第一部分：赛题解读

## 题目解析

**目标**：提供2018-02-01到2018-04-15用户行为数据。需要预测 2018-04-16 到 2018-04-22用户对品类下店铺的购买。以 user\_id, cate 和 user\_id, cate, shop\_id F-SCORE 的加权组合作为最终评分。

**难点**：

1. 线下训练集和验证集的构造。

2. 特征的构造。

3. 样本正负比例严重失调，导致模型学习困难。

4. 用户购买某品类时在目标区间选择了没有交互的店铺购买。

5. 建模思路的选择。

**业务场景及其几种解决方案**：

根据赛题描述和业务场景可知，在目标区间购买的行为应该由两部分组成：一部分为复购行为，一部分为即时性购买。而目标商品的品类应该能够决定此类商品是以复购为主还是及时性购买为主。解题思路应该有如下几种：

1. 直接对 user\_id, cate, shop\_id, sku\_id 进行建模得到 user\_id, cate, shop\_id 组合。

2. 直接对 user\_id, cate, shop\_id 建模。

3. 对 user\_id, cate 进行建模，也对 user\_id，cate，shop\_id 进行建模，再组合两者结果。

4. 先对 user\_id, cate 组合建模进行召回再对 user\_id, cate, shop\_id 建模。

## 竞赛过程

首先对数据进行分析构造训练集和线下验证集合，以及构建正确的评价指标。通过对数据的分析、探索和尝试确定建模过程，最后选择了先对 user\_id, cate 组合建模进行召回再对 user\_id, cate, shop\_id 建模的思路。首先保证线上线下的一致性和稳定性然后通过对数据的 EDA 和对业务的理解提取特征提升模型性能。

# 第二部分：算法核心设计思想

2.1数据探索

**类别复购率分析：**

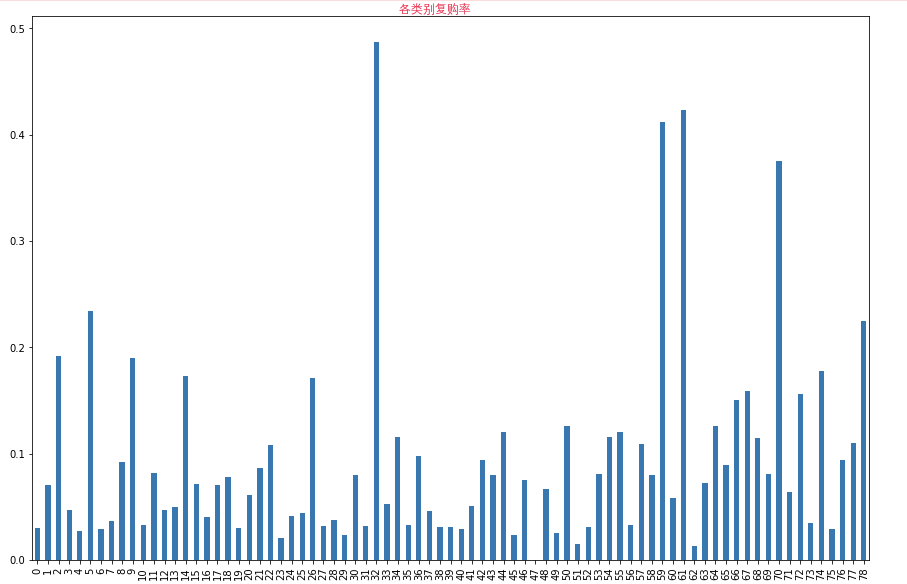


图2.1 各类别复购率

从图2.1可知，不同的类别在复购率上具有明显的差异。

**在目标区间购买样本在目标区间前的交互情况分析：**

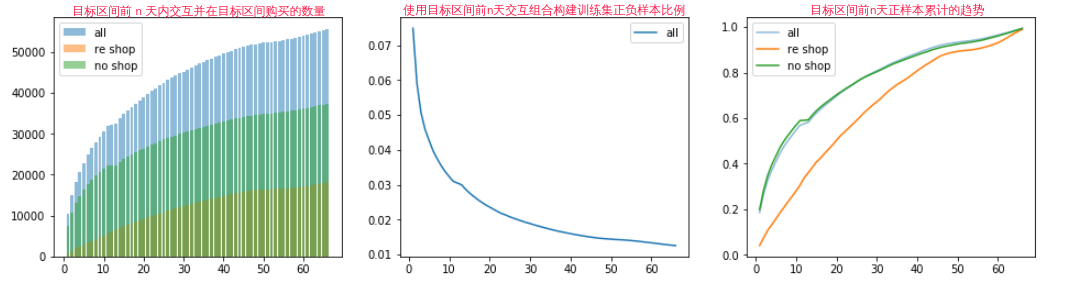


图2.2 user-cate相关分析



图2.3 user-cate-shop相关分析

由图2.2和图2.3可知：

1. 在目标区间的购买由两部分组成一部分为复购，一部分为即时性购买并且即时性购买远多于复购行为。
2. 复购和及时性购买增长不一致。及时性购买以近期的交互为主。
3. user\_id, cate 和 user\_id, cate, shop\_id 的增长情况基本一致。

**正负样本的分析：**

基于时间区间的考虑，我们选择了最佳的前30天和前40天从最后我们选择了正负样本对比分析，如下表所示：

表2.1 正负样本分布

| **时间区间** | **F11正样本** | **F11负样本** | **F12正样本** | **F12负样本** |
| --- | --- | --- | --- | --- |
| **30天** | 45195 | 2352912 | 25379 | 6101703 |
| **40天** | 49562 | 3055873 | 27664 | 8228463 |

从表2.1可知，采⽤40天相⽐30天的对正样本带来收益很小，但会导致负样本⼤幅度增加。并且从后面的尝试也发现前30天较40天预测效果更好一点。

综上数据简单分析，最后我们团队选择了使⽤前 30 天构建候选集。

**历史行为分析：**

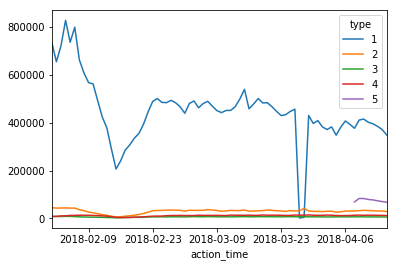


图2.4 历史行为相关分析

从图2.4可知，2-23 日之前的数据比较异常。猜测应该是受过年的影响。在 3-27，3-28 用户的浏览数据特别少而购买数量有所增加。猜测浏览数据丢失，并且期间可能有做促销活动。

**对某品类在目标区间选择了未交互的店铺分析：**

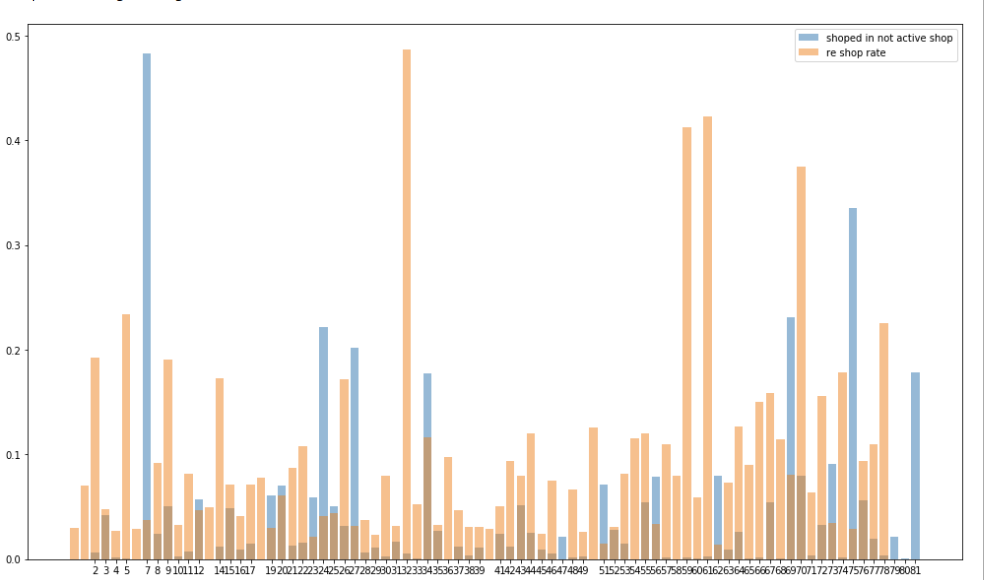
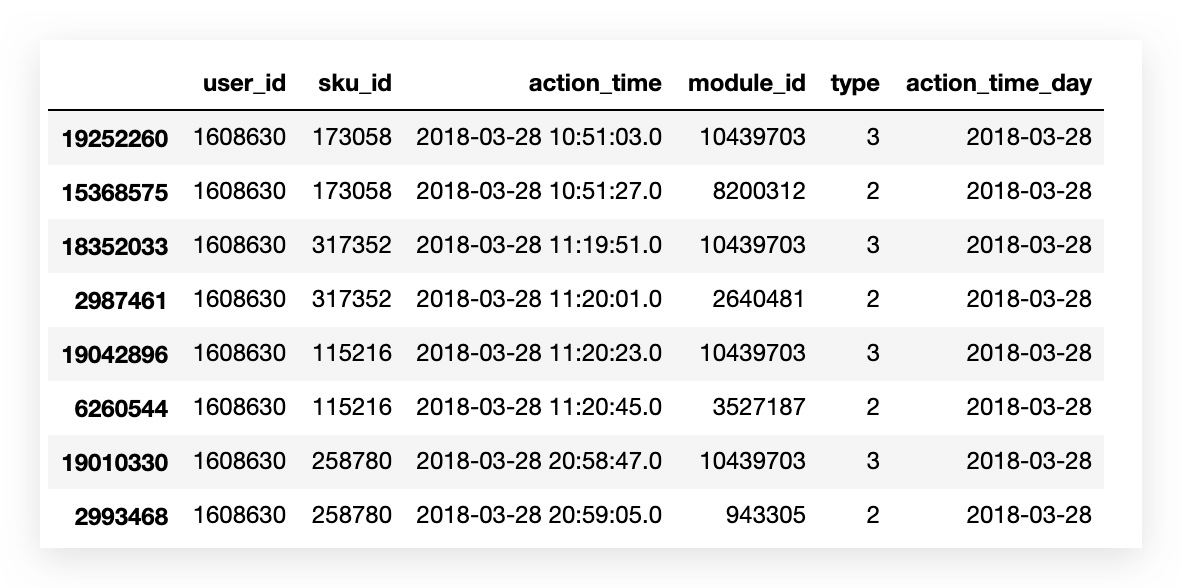


图2.5 店铺交互相关分析

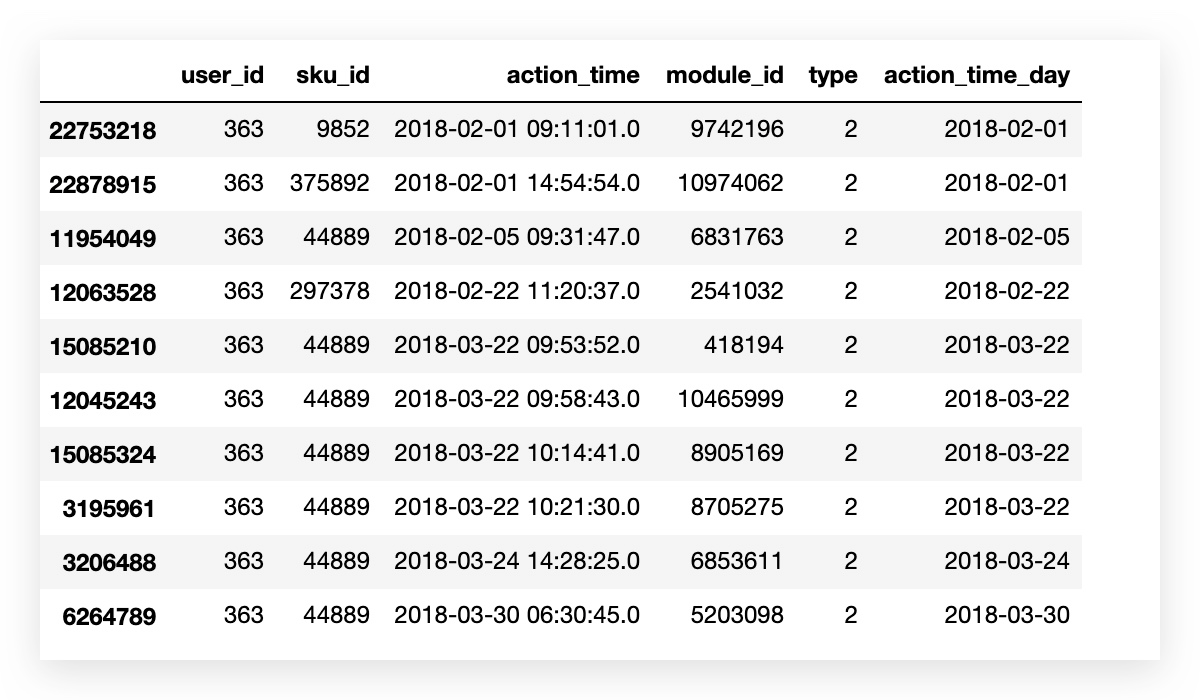
从图2.5可知，在未交互的店铺进行购买的行为，主要出现在复购率很低的类别上。

**其他发现：**

一些用户只在3, 28进行了购买，并在购买前进行了关注，如下所示，

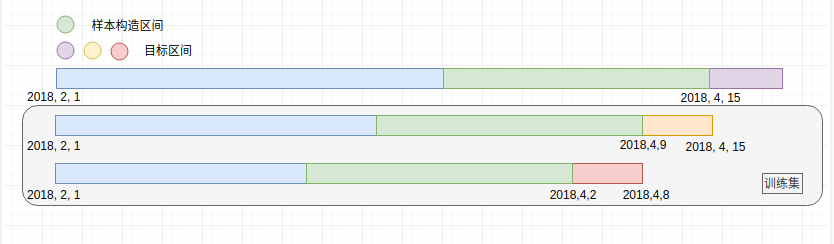


用户短时间内重复下单：



## 2.2构建数据集

综上数据探索，使用目标区间前30天交互过的组合构建候选集，并使用目标区间前的所有记录作为特征提取区间。线下使用 2018,4,2 - 2018,4,8号作为训练集，以2018,4,9-2018.4,15号作为时序验证集。数据集划分如下所示：



## 2.3解题方案

从上面对数据探索（图2.2，图2.3， 表2.1）可以看出 f12 的正样本量远远小于 f11 的正样本量。意味着不少的⽤⼾在⽬标区间选择了在没有交互过的店铺进⾏购买。如果直接对f12 建模会⾯临的问题如下：

1. 正负样本⽐例相差⼗分悬殊，导致训练困难。同时巨⼤的样本量可能导致内存不够。

2. 直接对 f12 进行建模会对 f11 的分数产⽣影响，限制 f11 的上线。

3. user\_id, cate 的交互信息比 user\_id, cate, shop\_id 信息更加丰富。并且 f11 的正负比例相对 f12 更加均衡导致 f12 的学习比 f11 更难。

所以我们考虑先**使用用户画像特征，类别画像特征以及用户类别的交互特征，对f11进行建模，预测在目标区间用户是否会对某类别进行购买。对目标区间为2018,4,9-2018.4,15的样本做5折交叉验证，取5折交叉验证的平均预测值作为预测结果。根据线上 f11的评价指标使用目标区间为2018,4,9-2018.4,15的样本确定最佳阈值，筛选出会在目标区间进行购买的user，cate组合作为f11 的结果。使用f11的结果对f12进行召回（去掉f11结果中没有的user，cate组合）构建f12的样本。在f11特征的基础上加入店铺画像特征，用户店铺交互特征，类别店铺交互特征对f12进行建模。和f11一样使用5折交叉验证，然后使用线上的评价指标使用目标区间为2018,4,9-2018.4,15的样本确定最优的阈值，得到最后的结果。（详细的建模过程见2.5模型设计）**这样做的好处如下：

1. user\_id 和 cate 交互已经能够确定用户对购买意图。在确定用户对该类别会购买的情况下可以类别和店铺的关系对店铺进行召回提高 f12 点正样本数。

2. 先使用 f11 进行召回能够大大降低 f12 的正负比例。使 f12 能够更容易学习。（通过线下测试使用f11 召回后对 f12 训练，f12 能够更快的收敛。并取得较好的效果。）

使用 f11 召回后f12 的正负样本比例如下：

|  | **F11 召回前** | **F11 召回后** |
| --- | --- | --- |
| 负样本数 | 6101703 | 531927 |
| 正样本数 | 25379 | 13728 |

## 2.4特征工程

以目标为导向从业务的角度出发进行构建。由2个部分组成。一个部分是画像特征，旨在描绘出不同的人，不同的品类和不同店铺的区别。以用户画像特征为例，在构建用户画像特征时除了基本特征以外我们希望能够描绘出用户的购物习惯，购物偏好、以及对平台的依赖度情况等。另一部分是使用划窗法构建交互特征。由于离目标区间越近正样本的数据的累积越快，所以我们使用了[1,3,7,15,30]的滑动窗口。描绘近期的活动特征。

***一些特别的处理如下：***

**没有使用购物车信息**

从数据探索中的图四可以看出只有部分加购物车的信息，考虑使用这部分特征会造成线上线下特征不一致，从而导致模型的稳定性变差。（计划使用构建子任务的方式利用这部分信息（如：例如将原问题从7天改为预测未来3天或4天用户对品类下店铺的购买，由于时间问题来不及尝试）

**提取画像特征时考虑2-23之前的信息**

从数据探索中的图四可以看出，受过年的影响， 2-23 号之前的数据分布比较异常。但在构建画像特征的时候还是考虑的这部分数据。如用户交互过的产品的好评率、购买过店铺的平均评级等。

**对3-27 和 3-28号的处理**

从数据探索中的图四可以看出，3-27 号 和 3-28 号的浏览数据疑似丢失了，而购买行为有增加的起伏。猜测可能是在搞活动。并且从图六可以看出有用户只在28号进行过购买（根据对业务对理解认为这类用户不是该平台的常用用户， 短期很有可能不会再次购买）。在构建某些特征的时候针对这个情况构建了一部分特征。如去除这两天后用户最后一次行为时间、用户最后一次购物是否在这两天，用户活跃度等。

**对连续下单行为进行合并**

从图七可以看出有用户会在短期内连续下单。在构建一些特征时候对一天连续购买算一次购买，例如用户购买间隔类特征。

**对商品好评率进行贝叶斯平滑**

由于评论数存在量级的差异，在计算好评率的时候会导致好评率的可靠性不高。例如评论数为1，好评数为1的商品。好评率高于评论数为100好评数为99的商品。

**top 召回**

虽然使⽤ f11 召回后使得 f11能够取得不错的得分，同时f12 的预测也变得容易且⽐较准确 。但是使⽤召回后 f12 的正样本变少了。就限制了 f12 的得分上线。从数据探索的图五可以看出在未交互的店铺进行购买的行为，主要出现在复购率很低的类别上。我们考虑这些可能属于⼤件商品例如电子产品，既然是⼤件商品，并且前⾯还没有交互过就说明了：

1. 客⼾不属于购买⼤件之前做很多功课的⽤⼾。

2. 客⼾很可能在销量很好的店铺购买（因为购买电子产品，在对此类产品不是很了解的情况下一般会选择市场主流产品进行购买）

根据这种想法提出了 **top召回**的解决⽅案：

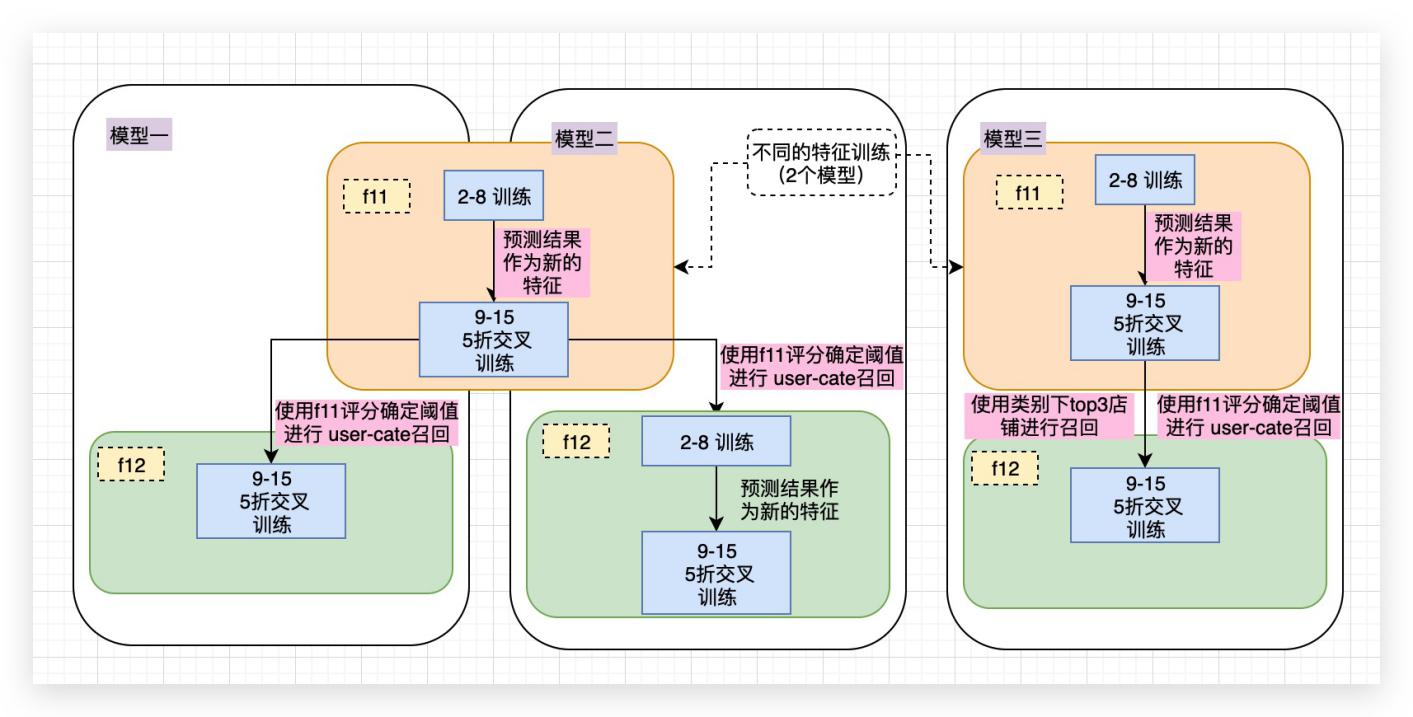
f11确定了类别，我们可以使⽤类别下销量最好的 top3 店铺召回构建候选集（考虑到⼤件特别是电⼦产品的市场变化很快，所以尝试过使⽤前⼀周、前两周、前⼀个⽉某类别销量top3 的店铺进⾏召回，通过观察验证集召回的效果，使用前一个月销量前三的店铺效果最好。）使⽤ top召回 后正负⽐例如下（在不大幅影响正负比例的情况下9-15号有⼤概2k 的正样本提升）：

|  | **top召回前** | **top召回后** |
| --- | --- | --- |
| **负样本** | 531927 | 559446 |
| **正样本** | 13728 | 15681 |

我们发现直接使⽤这这个方式效果并不好。（反思其原因是，这部分样本缺少交互特征，导致很多特征都是缺失的，模型没有足够的特征学到这一部分样本。这个时候这部分少量的样本就成了异常值反而影响模型的效果。）为了⽀持模型能够更好的学习到这部分少量的样本。构建了和 top召回相关的特征（如：⼾购买top店铺的 ⽐例，用户某类别top店铺的购买⽐例、店铺是否是前⼀周该类别下的top店铺等），线下效果得到了提升 f12 达到了0.88 左右。最后我们对 top 召回的模型结果使用较小的权重进⾏了加权融合（相当于对于top召回点这部分样本只选择置信度非常高的样本）。B榜获得了明显提升（3个百分点左右），也获得了最优的成绩。

## 2.5模型设计

我们使⽤ 2-8 训练，9-15作为线下时序验证集。使⽤时序验证和交叉验证确保模型的稳定性。训练期间使⽤ auc 作为模型的评估，然后再使⽤线上评价指标选取最佳阈值。后⾯发现 2-8 的分布和 9-15 分布不⼀致，导致使⽤ 2-8 号训练的效果并不好。最后使⽤ 9-15 号作为最终的线下训练区间。并使⽤迁移学习的思想利⽤ 2-8 号的信息。



**模型1:**

1. 使用 2-8 区间作训练集合，使用 9-15 验证集的到模型1，使用模型1对 9-15 和 16-22 进行预测作为新的特征。加入 9-15 和 16-22（f11）
2. 使用 9-15 进行 5 折交叉验证，得到5个模型。使用这5个模型对 9-15 和 16-22 进行预测取平均作为结果。使用9-15 的线上f11评价指标得分选择阈值。选出可能购买的 user-cate 组合。（f11）
3. 使用选出的 user-cate 进行召回构建 9-15和16-22 的 user-cate-shop （f12）候选集。使用 9-15 训练5折交叉验证，的到5个模型。使用这5个模型对 9-15 和 16-22 进行预测取平均作为结果。使用9-15的线上f12评价指标得分选择阈值。选出可能购买的 user-cate-shop 组合。（f12）

**模型2:**

1. f11 模型和模型1一样。
2. 使用选出的 user-cate 进行召回构建 2-8、 9-15和16-22 的 user-cate-shop （f12）候选集。使用 2-8 区间作训练集合，使用 9-15 验证集的到模型，使用模型1对 9-15 和 16-22 进行预测作为新的特征。加入 9-15 和 16-22（f12）
3. 使用 9-15 训练5折交叉验证，的到5个模型。使用这5个模型对 9-15 和 16-22 进行预测取平均作为结果。使用9-15的线上f12评价指标得分选择阈值。选出可能购买的 user-cate-shop 组合。（f12）

**模型3：**

1. 使用 2-8 区间作训练集合，使用 9-15 验证集的到模型1，使用模型1对 9-15 和 16-22 进行预测作为新的特征。（使用和模型1，模型2不一样的特征）加入 9-15 和 16-22（f11）
2. 使用 9-15 进行 5 折交叉验证，得到5个模型。使用这5个模型对 9-15 和 16-22 进行预测取平均作为结果。使用9-15 的线上f11评价指标得分选择阈值。选出可能购买的 user-cate 组合。（f11）
3. 使用选出的 user-cate 以及 cate top3 shop 进行召回构建 9-15和16-22 的 user-cate-shop （f12）候选集。使用 9-15 训练5折交叉验证，的到5个模型。使用这5个模型对 9-15 和 16-22 进行预测取平均作为结果。使用9-15的线上f12评价指标得分选择阈值。选出可能购买的 user-cate-shop 组合。（f12）

**模型融合：**

使⽤加权平均进⾏模型集成作为最后的模型 (权重⼤小：根据线下的得分⾃动选择。)

**0.4** \* 模型1 + **0.4** \* 模型2 + **0.2** \* 模型3

# 第部分：比赛经验总结

此类赛题数据集的构建成了⾄关重要的⼀步。可以使⽤对业务的理解确定⽅向，并快速画图对数据集进⾏探索，然后进⾏⼀些尝试确定最终的构建⽅式。也可以使⽤⼏种不同思路的构建⽅式最后再对结果进⾏融合。

构造稳定的线下验证集，时序任务可以用时序验证和 cv 来保证模型的稳定性。

一道题的解题思路往往是多种的，可以通过对数据多进行探索。根据数据的分布情况和一些解题经验确定一个好方向，往往能大大提高效率。当对某个情况不确定的时候要多尝试。

在对数据和赛题有一定的认识后，从业务场景出发，换位思考往往能够找到好的特征。

正负⽐例十分悬殊导致学习效果不好时，或者机器带不动的时候可以先召回再训练。通过召回后我们的模型不跑 cv，能够在16G内存的机器上跑完。

**组队:**

在比赛过程中多沟通不仅能提高工作效率还能产生一些好的想法。

**知识补充：**

提供一些我们没时间尝试的想法：

1. 使用不同的样本构造方式，最后对结果进行融合。如将复购和即时性购买做区分，分别建模。
2. 使用构建子任务的方式得到结果和原结果融合，如预测未来四天会购买组合；这样还能利用上购物车信息。
3. 使用推荐系统中常用的其他方式来进行召回， 针对不同对类别采取不一样的店铺召回方式等。
4. 开始提到的四种解题思路可以对其他的方案进行尝试，融合不同方案的结果。