**任务背景**

通过上一次你对营销短信的人群与时段的出色分析，阿普闪购的推广活动圆满达成了目标。当日订单量不断刷新新高，但与此同时，另一个问题也暴露出来：部分供应商对此次活动准备不足，导致生产跟不上了，很多商品发货周期变长，用户的整体满意度也受到了影响。

接下来，即将迎来下一次的大促节日，这次节日同以往不同，用户在开始活动的前三天内只能浏览商品、收藏或者添加购物车，不能下单，等三天过后才开始秒杀下单。为了不再重蹈之前准备不足的覆辙。你作为阿普闪购的数据分析师再一次临危受命。

**这一次，任务很明确：根据用户前三天预热期的行为记录，预测出每款商品的销量，这样可以提前知会供应商按照预测销量进行准备，避免之前准备不足的问题。**

**问题分析**

从任务说明中可以发现，我们的核心任务是从其他行为（点击、添加收藏、添加购物车）预测商品的销量。这和我们上一次的分析维度不同，这次主要是从商品维度进行分析。

我们有三个数据源，VIP 会员表和用户信息表基本都是用户的信息。而用户行为表中则包含了商品的信息，所以这次我们重点从用户行为表入手。但面临的挑战是：用户行为表是面向行为的，每一行代表一个用户的一次行为，可能是点击、可能是加入购物车，等等。我们首先需要从这个表中抽象出一个面向商品的表，然后再尝试建立预测销量的模型。

# 构建特征（特征工程）

### 为什么需要特征工程

在绝大多数需要建立模型的数据分析任务里，特征工程都是必不可少的环节。顾名思义，特征工程就是指准备特征、筛选特征的工作，只是这项工作具有一定的重要性和复杂度，所以也会叫作工程。

以线性回归模型为例，通过上一节的学习，我们都知道线性回归模型本质上是从找出自变量和因变量的关系。现实问题中，因变量是很好找的，因为往往因变量就是我们要预测的任务，比如在这个例子中，因变量就是我们的预测目标：商品的销量。

另一方面，自变量，也称为模型的特征，往往没那么明显。一个原因是因变量往往会受非常多的因素所影响，有直接的，也有间接的。另一个原因是我们拿到的数据集上往往没有现成的特征给我们。这需要我们从原始数据集中经过分析、变换之后才能得到我们想要的特征。

### 特征分析

特征工程的第一步，是根据我们已有的数据（哪怕不是直接可用的数据）以及我们的因变量，来预设出一组特征。第二步，就是从原始数据中计算出这些特征。第三步则是用这些特征对模型进行训练，如果训练出的模型有问题，则需要重新回到第一步，尝试重新尝试其他的特征。

我们这次的因变量是商品的销量，能够作为参考的数据是用户在预热期的三天的行为。用户在浏览商品时，不能下单的话，主要的行为就是：点击商品、把商品添加到购物车，以及收藏商品。那对于商品来说，商品的点击数、加购物车的次数，以及收藏的次数是否可以作为销量的特征呢？

判断特征是否有效可以从特征的值改变是否会影响因变量的值这个标准出发。这三个特征，不管是点击数，还是收藏数，还是加购物车的次数越多，就说明商品越受欢迎，则说明越可能有用户来购买，反之亦然。所以这三个特征都可以作为我们这次任务的特征。

除了商品维度的特征，商品所在店铺的特征是否对商品的销量有帮助呢？比如在一些商品数比较多的店铺，往往用户搜索进入的概率就大，那商品被用户看到的概率也就变大了。另一方面，店铺的 VIP 用户多寡似乎也能说明店铺是否优质，优质的店铺的商品理论上来说也更好卖一些。

综上所述，我们初步总结了五个可能对商品销量有影响的特征。

商品维度：

点击数；

添加到购物车的次数；

收藏数。

商品所在店铺的维度：

店铺的商品数；

店铺所拥有的 VIP 用户数。

### 商品特征计算

首先我们尝试计算商品的特征，从三张表的描述可以得知，在用户行为表中，有用户对商品的行为记录，所以我们从用户行为表入手来计算商品的特征。

基本的思路就是把用户行为表按照 item\_id 的维度聚合起来，然后点击数、加购物车数、收藏数和购买数都应该是聚合后的表的列。但目前用户行为表中没有这四列，相关的信息都由 action\_type 来表示了。

所以第一步，我们需要把 action\_type 的内容展开成四列来分别表示：点击、加购物车、收藏和下单。

规则就是如果 action\_type 是收藏时，则收藏列为 1，其他三列为 0；action\_type 为下单时，下单一列为 1，其他三列为 0, 以此类推。

这里就涉及我们在生成新列的时候，不仅是从现有列直接计算生成，而是需要有一定的逻辑判断（判断 action\_type ）的值，所以我们可以用 Series 的 apply 函数来实现，apply 函数可以把一个函数执行到指定的 Series 上，然后把函数的返回值作为新的 Series 返回。