目录

[写在前面 2](#_Toc412729545)

[赛题介绍 2](#_Toc412729546)

[赛题FAQ 3](#_Toc412729547)

[数据挖掘 初阶 4](#_Toc412729548)

[问题解析 4](#_Toc412729549)

[训练集和测试集的划分与构建 4](#_Toc412729550)

[数据清洗 5](#_Toc412729551)

[领域知识 —> 特征工程 6](#_Toc412729552)

[我的特征工程 8](#_Toc412729553)

[缺失值填充 11](#_Toc412729554)

[数据分布不一致问题 11](#_Toc412729555)

[单模型的调优 12](#_Toc412729556)

[正负比例失衡问题 12](#_Toc412729557)

[模型选择与融合 12](#_Toc412729558)

[比赛Top10的答辩PPT等学习资料 13](#_Toc412729559)

[数据挖掘 进阶 13](#_Toc412729560)

[附言 14](#_Toc412729561)

## 写在前面

之前写过[关于《天猫推荐算法大赛》的总结](http://weibo.com/p/1001603740163867054829)，但那并不适合给纯新手看，这里再针对性地进行整理，以方便新手理解。仍然以该赛题为例，讲解一个数据挖掘比赛的具体做法，层层深入。本次讲解假定读者对机器学习和数据挖掘有一定的了解，懂基础知识，比如《数据挖掘导论》、《机器学习实战》等，针对简单的数据集做过实验，推荐[《机器学习那些事》](http://wenku.baidu.com/view/1d893fcd8bd63186bcebbc51.html)。

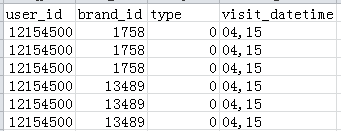
文章外链多为引申，如精力有限，先看本文。

## [赛题介绍](http://102.alibaba.com/competition/addDiscovery/gameIntroduce.htm)

本届赛题的任务就是根据用户4个月在天猫的行为日志，建立用户的品牌偏好，并预测他们在将来一个月内对品牌下商品的购买行为。我们会开放如下数据类型：

|  |  |  |
| --- | --- | --- |
| **字 段** | **字段说明** | **提取说明** |
| **user\_id** | **用户标记** | **抽样&字段加密** |
| **Time** | **行为时间** | **精度到天级别&隐藏年份** |
| **action\_type** | **用户对品牌的行为类型** | **包括点击、购买、加入购物车、收藏4种行为  (点击：0 购买：1 收藏：2 购物车：3）** |
| **brand\_id** | **品牌数字ID** | **抽样&字段加密** |

用户对任意商品的行为都会映射为一行数据。其中所有商品ID都已汇总为商品对应的品牌ID。用户和品牌都分别做了一定程度的数据抽样，且数字ID都做了加密。所有行为的时间都精确到天级别(隐藏年份)。——下图是部分样例数据截图和season1的数据集。



## [赛题FAQ](http://bbs.aliyun.com/read/154994.html)

1. 抽样方式

我们在做训练数据的时候，是首先从天猫全量用户中定了一个抽样比例，确定训练用户集User set。同样，在天猫全量品牌集合中，定了一个抽样比例，确定品牌集合Brand Set。大家拿到的这份数据，是User Set 中的用户在Brand Set 中的所有行为。

2. 测试集

问“如果一个用户在测试集中对一个品牌购买超过一次，是否要预测具体的购买次数”

答： 不需要的，只需要预测是否购买即可，不需要预测准确的数字。

3. 购买行为定义

* 问：点击“购买”就记录为一次购买，还是付款成功才算作一次购买？

答：付款成功才叫一次购买

* 问：代付款的话，购买行为计作购物人的还是代付人的？

答：如果用支付宝代付功能，这个不是很确定，等确定了再答复。

* 问：那如果有一个user，一次性买了n件商品（同一个brand），那么应该算是一次购买还是n次购买？

答：如果是一次性购买n个不同的商品（同个brand），会产生n条购买记录。

如果对同一个商品，分开购买n次(不同订单)，会产生n条购买记录。

P.S. 点击不一样，只要有发生点击，则不论点击了多少商品或品牌，都记一次。

* 问：对于一次购买（先加入购物车再购买），那么用户行为是否为“购买”，而没有“加入购物车”

答：不会，“加入购物车”与“购买”是独立行为，所以是一次“加入购物车”+一次“购买”（可模拟一下淘宝购物流程，帮助理解）

## 数据挖掘 初阶

问题解析

首先，我们要确定待解决的问题映射到数据挖掘，具体会是怎么样一个问题。如果连自己要解决的是什么问题都不清楚，那就别提怎么解题了。

根据官方描述，这次比赛要做的就是：根据用户4个月在天猫上对品牌的点击、收藏、购物车、购买等行为记录，预测第5个月哪些用户会购买哪些品牌。

显然，根据所给赛题和数据，我们需要解决的是监督学习中的分类问题，而且是二分类问题——即要判定 用户购买或未购买品牌。注意，这不是一个传统的推荐问题，因为数据是离线的，你提交的预测结果无法影响在线用户的决策；而且就所给字段来说（对象是brand，而且没有任何类目相关的信息），基本上只能在用户操作过的品牌里进行购买预测。对于推荐，Xavier Amatriain有个比较总结性的阐述： <http://www.slideshare.net/xamat/recommender-systems-machine-learning-summer-school-2014-cmu>

而数据挖掘项目里，所要研究的问题都会以***样本***为单位进行，分类问题里的类别标签则以样本在业务问题中的定义进行设置。

那么显然，这次赛题里的***样本由user\_id和brand\_id共同决定***，就如同著名的鸢尾花数据集中代表每一个样本的行号那样，与其他样本区别开来。而***样本的类别标签则由未来一个月中该用户是否购买该品牌决定***，通常来说，用1表示有购买的正样本，而0表示未购买的负样本。

### 训练集和测试集的划分与构建

而分类问题中，模型需要经过训练集的学习，才能用于测试集，而训练集和测试集在形式上的区别在于前者有类别标签，后者则需要模型输出相应的类别标签。

这里的问题是依据前4个月的用户操作记录预测第5个月的购买情况，所以通常来说，训练集的构建需要利用前3个月的数据，而其相应的类别标签，则来自于第4个月的购买情况；而测试集的构建则可用到4个月的完整数据。下面举个例子，用于解释训练集和测试集的具体表现形式。



训练集样本示例



测试集样本示例

训练集样本示例中的user\_id, brand\_id用于表示唯一的样本id，而feature\_1, feature\_2则用前3个月的数据构建而来（绝对不能引入未来的数据，否则会影响模型走入歧途）。在这里继续展开一下，本次比赛中，我划分训练集和测试所用的***分割日期***是第93天（将日期格式转换为相对的第几天），即训练集使用 1-92天的数据，而测试集使用93-123的数据。当然，你也可以使用不同的分割日期，甚至可以滑动分割日期，从而构造更多的训练集（测试集当然只有有一个）。

到这里，基本可以开始初步的训练和预测了——如果不考虑太多的话。

可以临时性的上几个简单的特征，比如，在***分隔日期***前brand的7天销量，或者用户对该brand的总点击次数等。详细部分另谈。

### 数据清洗

数据原本是经过了官方清洗的，所以不需要再清洗。

如果是异常性的数据，比如，点击量特别多的用户（疑似爬虫），也可将其设计到特征体系中（比如，构建特征——用户的总点击数），让模型自己学习；又或者是业务上的原因可能需要清洗，比如，用户购买之后发生的点击可能是查看发货状态等，但由于本次操作对象是品牌，而且用户点击的真实目的也不能确定，所以最好还是不要清洗这部分数据，而是应当设计相应特征，比如，用户购买该品牌之后对该品牌的点击次数。

以上例子仅供参考，针对不同的数据挖掘项目，需要根据其自身的数据、业务进行分析和处理。尽可能保留原数据，尤其是比赛的项目，通常都经过了数据清洗，草率的清洗数据很可能导致重要信息的丢失。

### 领域知识 —> 特征工程

都说特征工程重要，都说领域知识重要，那么为什么他们那么重要呢？

之前提到过，要想输出测试集的类别标签，需要模型在训练集经过学习，那么模型在训练集上学习的是什么？

我们知道，在给模型输入数据的时候，实际上模型用到的都是特征（及其对应的类别标签），而特征就是用于描述为什么该样本的类别标签就是如此，在本赛题中即用于描述某用户为什么购买/不购买某品牌。比如，用户购买某品牌的原因，是因为以前经常买，信得过，那么这里可以用【用户购买该brand的天数/用户访问该brand的天数】来刻画用户对该品牌的忠诚度。当然，并不是单一的特征就能描述所有情况，通常来说，需要针对所有可能的情形进行考虑，从而【深刻而全面】地刻画用户购买/不购买品牌的原因。这样，模型才能真正学到其中的【规律】，从而在测试集表现优异。

而这一过程，就称之为特征工程。显然，要做好特征工程，需要我们自身对【用户会否购买品牌】这一业务具有较深刻的理解，即领域知识，并将其***用模型可理解的方式表现出来***，通常来说就是用 数字或者字符串 表示，比如，用户购买品牌时的月份即用字符串表示，而用户购买品牌的次数则用数字表示。所有【领域知识】都应当想办法映射到特征体系，用于完善特征工程。

那么，具体怎么从【领域知识】映射到【特征工程】呢？

首先，要针对现有的领域知识进行梳理，并且针对梳理结果进行数据分析，以验证该领域知识是否符合真实业务。以下将部分举例说明具体映射过程。

* 统计用户在4个月中的购买量、点击量、收藏量、加购物车次数、有操作记录的天数、发生购买的天数、点击转购买率等等，并以点击转购买率做asc/desc排序，进行观察，可以发现点击量越高的，点击转购买率也往往越低，尤其点击量极高的用户，往往购买量为0。由此可以考虑将用户购买量、点击转购买率等一系列特征用于描述用户是否会在未来一个月发生购买。

甚至你可以发现，有些用户的点击是0，按理说这是不可能的，官方解释说确实存在数据丢失问题。因此，做好数据分析，不仅仅可以帮你理解业务，更能发现问题，从而解决问题，比如对缺失的值可以用中位数或者平均数来填充。

类似的统计分析可以同样针对品牌。

* 统计 用户从初次访问品牌到最终购买品牌的时间，可以发现绝大多数购买都是当天接触当天完成，越往后用户购买的可能性越低。由此可以推测用户对品牌的购买意愿是随着距离上次访问的时间拉长而衰减的，进而可以使用衰减函数来模拟该购买意愿。

当然，这里只是象征性地讲了几个比较典型的分析方向，利用自身已有的【关于购物的领域知识】可以帮你深刻理解数据，并进行相应分析乃至验证，甚至可以从分析结果中加深对业务、对赛题、对数据的理解，从而完善特征工程。

***一方面，利用领域知识对数据进行分析、实验验证，另一方面，从分析的结果、实验的反馈来更新领域知识。这是一个不断迭代的过程，需要成体系、且不断地完善。***

### 我的特征工程

这里将我比赛时用到的特征列上，仅供参考。

总而言之，特征的构建要从会否发生“购买”出发，既要考虑什么样的用户会买、什么样的品牌能卖，也要考虑哪些用户不会买、哪些品牌不好卖，更重要的是，要分析好用户与品牌之间的关系，找出已交互里用户会买的那些。

**1 品牌特征**

之所以先说品牌特征，是因为一开始重点研究是就是这个。一开始的时候，想着用户很大程度会受搜索排序的影响，为了启发思路特地去研究天猫的排序规则，可惜@樱木 说这是机密不能外泄，于是只能到网上找些[卖家自己研究出来的指标](http://news.iresearch.cn/Zt/123380.shtml)、以及一些可能借用过来的指标。

于是，前前后后借用的、受启发而想到的特征，基本如下：

**1.1 基本统计类特征**

A. 销量（分隔日期前3、7、15、30、124天，以下特征基本都是这么个周期，如无特别说明，谈论的各个指标均有特定周期）

B. 成交订单数（以同一天同一用户为一个订单）

C. 购买人数（周期内经过去重的购买人数）

D. 上面的3类特征分别再针对点击、收藏、购物车等操作进行统计。

**2.2 转化率类特征**

A．销量/点击数

B．成交订单数/点击订单数

C．购买人数/访问人数（这里考虑到部分点击数据丢失，就没以点击人数为分母了）

D．上面的3类特征分别再针对收藏、购物车等操作进行统计

**2.3 比值类特征**

A．返客率（某一周期内多次购买的用户数/总购买用户数）

B．老客户率（3天【或其他周期7、15、30】前曾购买过且在前3天内再次购买的用户数/3天内总购买用户数）

分别再针对收藏、购物车、点击等操作构建上述2个特征。

P.S.：这里返客率和老客户率均不含124这个周期。

C．跳出率（周期内只对品牌进行过1次点击操作的用户数/总用户数）

D．活跃度（周期内有3次及以上点击的用户数/总用户数）

E．人均点击数、人均购买量、人均收藏量、人均购物车量等

F．不在初次访问品牌时进行购买的订单数/总订单数（以此获得品牌在下个月被购买的可能性）

**2 用户特征**

用户特征的主要出发点是找出什么样的客户才可能买，结合之前对爬虫类用户的分析、以及对品牌特征构建的经验，得出如下特征：

**2.1 基本统计类特征**

A．购买量

B．成交订单数

C．购买品牌数（周期内去重）

D. 分别针对点击、收藏、购物车等操作再进行统计

**2.2 转化率类特征**

A．购买量/点击量

B．成交订单数/总订单数（一天访问一品牌为一订单）

C．购买品牌数/访问品牌数（品牌周期内去重）

D．分别再针对收藏、购物车等操作进行统计

**2.3 比值类特征**

A．用户活跃天数/当前日期（一般来说，越活跃的用户越不会买，大多特征源于对爬虫特点的分析）

B．发生购买的天数/总活跃天数

C．用户购买总量/购买天数

D．等等类似的一些可以反映用户购买可能性的特征

E．不在初次访问品牌时进行购买的订单数/总订单数（以此获得用户在下个月购买的可能性）

**3 用户-品牌特征**

**1 基本统计类特征**

A．截止到最后一天（比如交叉验证训练集的第61天）累计的点击（或衰减后点击）

B．同样的特征分别再针对收藏、购物车、购买等操作进行统计

C．截止到最后一天，用户对品牌进行过访问的次数、购买的次数、收藏的次数、购物车的次数（或经过时间衰减）

D．截止到最后一天，用户对品牌进行第一次访问（收藏、购物车、购买）到最后一次访问的时间间隔（或经过时间衰减）

E．用户最后一次访问（收藏、购物车、购买）距离分隔日期（如62）的时间间隔

**2 比值类特征**

A．购买该品牌次数/总购买次数

B．对该品牌的总点击（购买、收藏、购物车）数/访问次数

C．在对品牌A进行访问的那些天里，用户对A的点击数/那些天里的总点击数

D．在对品牌A进行访问的那些天里，A的销量/那些天里的所有品牌的总销量（这个特征在实现思路上消耗内存太大，未能实现）

E. 用户访问品牌A的天数/用户总活跃天数

F. 用户购买品牌A的天数/用户总购买天数

### 缺失值填充

构建特征的时候，往往会出现缺失值，比如，分隔日期距离用户上次收藏某品牌的间隔，如果用户从未收藏过该品牌，那么其值将为空，需要填充，而填充的原则即保持数据在分布上的一致性。比如，这里即应该填充一个较大的数，比如124（因为数据的总时间跨度也只有123天），用于表示在现有的123天内用户是没有收藏过该品牌的。其他的缺失值，也要做类似处理，使得填充的值能正确表示它原本该有的含义。

### 数据分布不一致问题

实际问题中，常常出现训练集、测试集数据分布不一致的问题，从而导致模型学到的“规律”无法有效应用到测试集，导致模型效果不佳。所以，往往需要针对训练集、测试集进行特殊处理。

Alex Smola 给出了一种基于logistic regression的处理方案：

<http://blog.smola.org/post/4110255196/real-simple-covariate-shift-correction>

### 单模型的调优

每个模型针对不同的训练集、测试集，都有其最优参数（局部最优或全局最优）。一般来说，由于抽样的随机性，甚至模型本身的随机性，很难做到全局最优，局部最优也很难。

以RF为例，通常来说，先将正负样本比例调到最优（由于RF的随机性，树越多则模型效果越稳定，因此在调优过程中，注意，不要将其随机性导致的些许差别错认是参数导致的，应当将树的数目设置的比较大，此时调参才具有可参考性），再调节单棵树的最大抽样量，每次随机从多少特征里选取，以及树的生长截止条件。

当然，不同的实现，可能存在不同的参数设置，需要读者自己根据所使用的工具进行理解、调优。

科学调参法（由微博@好东西传送门 整理，一个专门整理机器学习资料的微博，这里的外链当然要看！）：<http://ml.memect.com/archive/2014-12-21/long.html#3789989488121751>

### 正负比例失衡问题

抽样往往导致数据分布与全集数据分布不一致、重要信息缺失，但是又不得不进行抽样。

解决办法的总结（这里的外链当然要看！）：<http://weibo.com/p/1001603785752793219283>

### 模型选择与融合

根据前不久的[研究](http://jmlr.org/papers/v15/delgado14a.html)指出，下列模型在各数据集中表现最优：

The random forest is clearly the best family of classifiers (3 out of 5 bests classifiers are RF), followed by SVM (4 classifiers in the top-10), neural networks and boosting ensembles (5 and 3 members in the top-20, respectively).

通常来说，选择以上的模型就够用了，阿里这次比赛里常用的强大算法为RF和GBDT，辅以LR：

<http://www.52ml.net/15084.html>

不过，如果有时间和资源，可以一个个自己调试，主流算法的优劣及其适用场所介绍如下：

<http://www.cnblogs.com/tornadomeet/p/3395593.html>

当然，单个模型而言，优势有限，比赛中往往采用多个模型混合的方式。通常来说，模型与模型直接差异越大（保证单模型效果的前提下），混合模型效果越好。常见的混合方式如下：

<http://www.52ml.net/318.html>

而逻辑斯谛回归，常常具有特殊的用途，可以将brand\_id 进行dummy coding（自行谷歌），从而利用brand本身所蕴含的信息。

### 比赛Top10的答辩PPT等学习资料

官方资料（答辩PPT以及阿里官方平台教程等）：<http://bbs.aliyun.com/read/155044.html>

Oilbeater（比赛入门类）:

[阿里大数据竞赛非官方指南第一季终结篇](http://oilbeater.com/%E9%98%BF%E9%87%8C%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%AF%94%E8%B5%9B/2014/04/17/the-endof-season.html)

[阿里大数据竞赛非官方指南第三弹-- LR入门](http://oilbeater.com/%E9%98%BF%E9%87%8C%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%AF%94%E8%B5%9B/2014/04/04/the-bigdata-race-3.html)

[阿里大数据竞赛非官方指南第二弹](http://oilbeater.com/%E9%98%BF%E9%87%8C%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%AF%94%E8%B5%9B/2014/03/24/the-bigdata-race-2.html)

[阿里大数据竞赛非官方指南](http://oilbeater.com/%E9%98%BF%E9%87%8C%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%AF%94%E8%B5%9B/2014/03/16/the-setup-of-bigdata-race.html)

Beader: [阿里平台Map/Reduce入门](http://beader.me/2014/05/05/odps-mapreduce/)

## 数据挖掘 进阶

Kaggle大神们的经验总结

<http://blog.kaggle.com/2014/08/01/learning-from-the-best/>

如何做好特征工程

<http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/>

## 附言

有个说法，数据挖掘项目的表现，80%取决于特征工程，剩下的20%才取决于模型等其他部分；又说数据挖掘项目表现的上限由特征工程决定，而其接近上限的程度，则由模型决定。而特征工程，又往往取决于对问题的理解、对数据的理解，因此，数据挖掘项目（比赛）最终往往拼的是细节，是对数据更深刻的理解与处理。而这些东西，非常耗精力，当然，也需要有规划、有方法。