江离数据挖掘俱乐部

2019 安泰杯 跨境电商智能算法大赛

AliExpress消费行为预测

联合主办

上海交通大学安泰经济与管理学院

阿里巴巴集团AliExpress

天池平台

江离

做纯粹的数据挖掘。

——江离数据挖掘俱乐部

江离数据挖掘俱乐部（jiangliclub.com）创建于2019年7月。俱乐部以「寻找真正的数据挖掘技术」为宗旨，以「以艺为本，以德为先」为原则，以「用技术改造世界，让世界变得更加美好」为终极目标，从事和数据挖掘有关的活动。俱乐部内聚集着一群有着多年数据挖掘经验、数据挖掘功底深厚、对数据挖掘有深刻认识和独到见解的数据挖掘者。

数据挖掘竞赛是江离数据挖掘俱乐部的重要活动之一。俱乐部成员常通过参加数据挖掘竞赛来验证和完善自己的数据挖掘算法和解决方案，并提升自己的数据挖掘能力和经验。自建部以来，俱乐部曾数次派出代表队参加各大数据挖掘竞赛平台举办的数据挖掘竞赛，并取得了较为优异的成绩。

江离数据挖掘俱乐部

赛题

江离数据挖掘俱乐部

利用用户近十五天对商品的交互记录，预测用户接下来最有可能购买的30件商品。

$$score = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{30}\frac{s(b_j, k)}{k}$$

数据

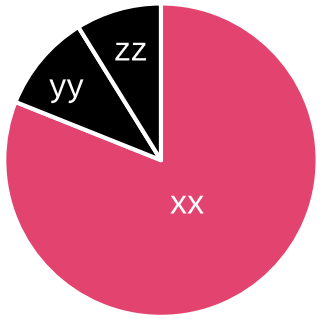
江离数据挖掘俱乐部

行为记录 5187万条+85万条

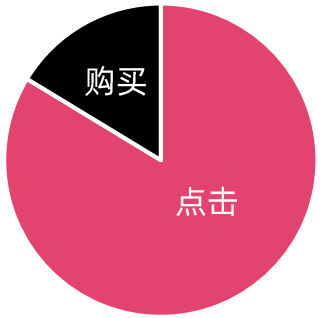
国家	xx
买家标识	817731
商品标识	4033525
时间	2018-06-12 07:12:58
单序	1
购买标志	1

商品 914万条

商品标识	4033525
类目标识	2153
店铺标识	49776
价格	180



各国记录数占比



点击与购买的记录数占比

线下测试

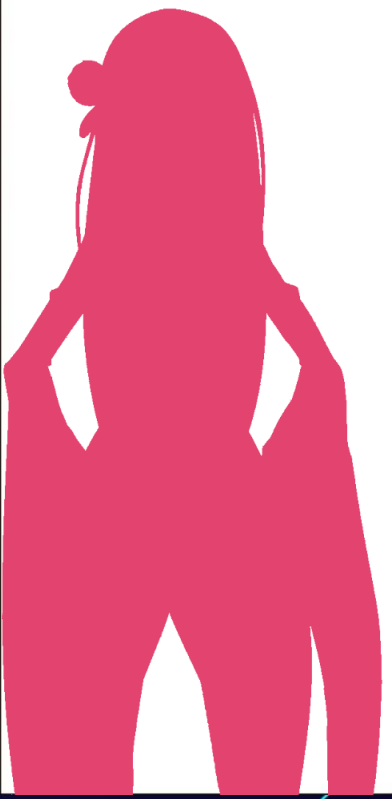
江离数据挖掘俱乐部

训练集按用户平均分成5折，每次取其中一份作为线下测试集，其馀为线下训练集。

我们始终确保我们提交到线上的结果均经过线下测试。本赛题中，线下与线上成绩的变化基本一致。

思路

江离数据挖掘俱乐部



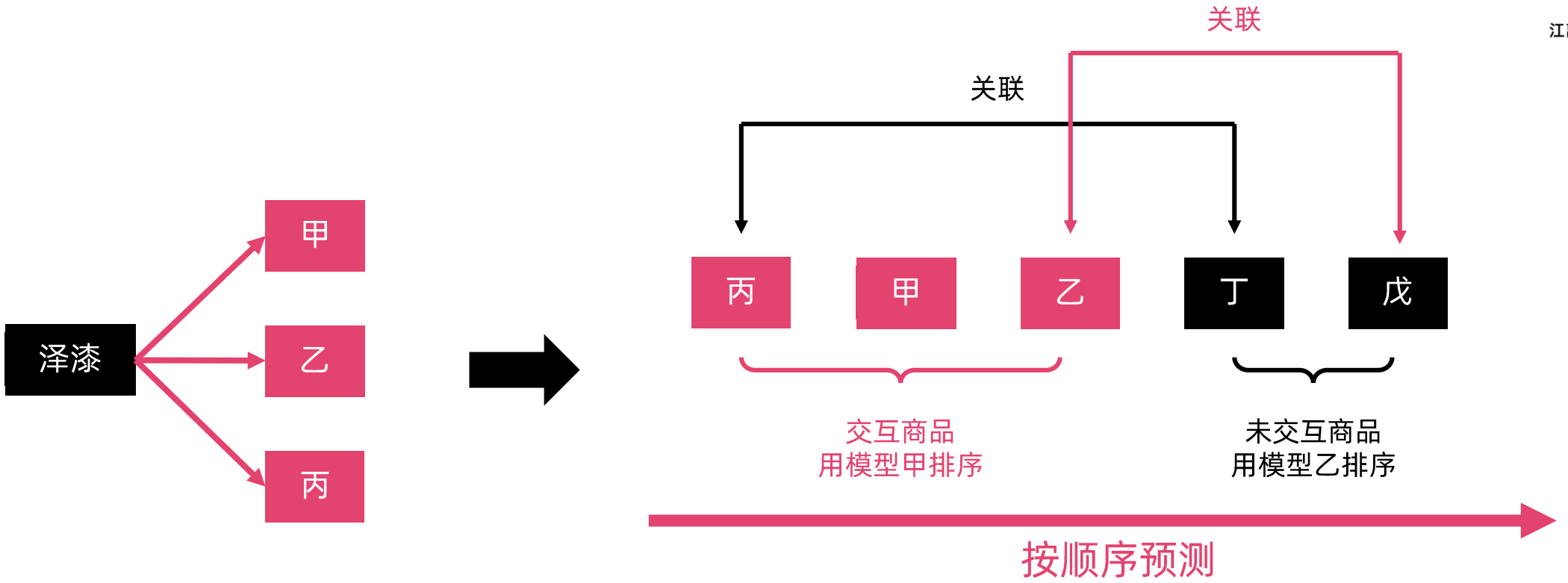
用户会购买什么商品？

- 用户历史看过的商品
- 用户历史看过的商品的关联商品

通过观察数据我们得知，一些商品总是会被同时交互，我们称这些商品是「关联的」。

思路

江离数据挖掘俱乐部



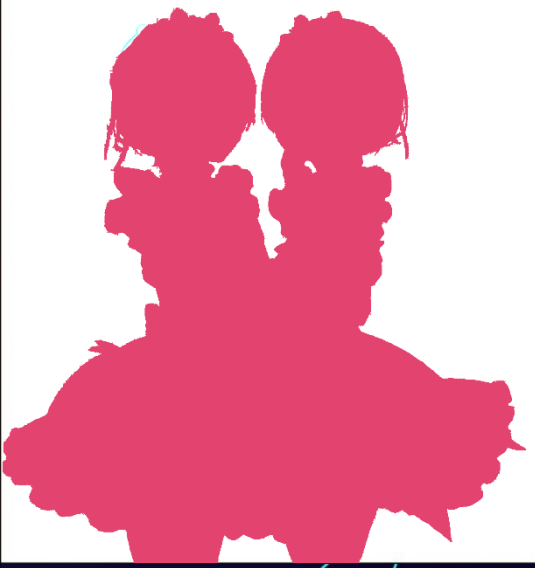
模型甲 交互商品模型

江离数据挖掘俱乐部

构造候选商品
选择用户历史上交互过的商品作为候选

构造特征
从用户历史交互数据中提取出数百维特征

训练模型和预测
训练lightgbm模型并预测



模型甲 买家商品特征

江离数据挖掘俱乐部

买家商品之当日点击数
买家商品之购买日数
买家商品之最小购买日序
买家商品之最大购买日序
买家商品之最小秒序差
买家商品之最大日序差
买家商品之最小单序
买家商品之最大单序
买家商品之单序极差
买家商品之中位单序
买家商品之最大秒序之数
买家商品之最小秒序数大于1单序
买家商品之最小秒序数大于1单序差
买家商品之秒序数大于1数
买家商品之秒序数大于1比例
买家商品之秒序数大于2数
买家商品之秒序数大于2比例
买家商品之当日秒序数大于1数
买家商品之秒序前差分
买家商品之秒序后差分
买家商品之前同类目
买家商品之后同类目
买家商品之总前同类目
买家商品之总后同类目
买家商品之商品单序前差分
买家商品之商品秒序前差分
买家商品之最后秒序前差分
买家商品之最后秒序后差分
买家商品之最后秒序前二差分
买家商品之最后秒序后二差分

模型甲 买家商品排名特征

江离数据挖掘俱乐部

买家商品之点击数排名
买家商品之最后秒点击数排名
买家商品之近一小时点击数排名
买家商品之近两小时点击数排名
买家商品之近三小时点击数排名
买家商品之当日点击数排名
买家商品之购买日数排名
买家商品之最小购买日序排名
买家商品之最大购买日序排名
买家商品之最小秒序差排名
买家商品之最大日序差排名
买家商品之最小单序排名
买家商品之最大单序排名
买家商品之单序极差排名
买家商品之中位单序排名
买家商品之价格排名
买家商品之最大秒序之数排名
买家商品之最小秒序数大于1单序排名
买家商品之最小秒序数大于1单序差排名
买家商品之秒序数大于1数排名
买家商品之秒序数大于1比例排名
买家商品之秒序数大于2数排名
买家商品之秒序数大于2比例排名
买家商品之当日秒序数大于1数排名
买家商品之秒序前差分排名
买家商品之秒序后差分排名
买家商品之前同类目排名
买家商品之后同类目排名
买家商品之商品单序前差分排名
买家商品之商品秒序前差分排名
买家商品之最后秒序前差分排名

模型甲 买家商品排名特征（二）

江离数据挖掘俱乐部

- 买家商品之类目点击数排名
- 买家商品之类目最后秒点击数排名
- 买家商品之类目近一小时点击数排名
- 买家商品之类目近两小时点击数排名
- 买家商品之类目近三小时点击数排名
- 买家商品之类目当日点击数排名
- 买家商品之类目购买日数排名
- 买家商品之类目最小购买日序排名
- 买家商品之类目最大购买日序排名
- 买家商品之类目最小秒序差排名
- 买家商品之类目最大日序差排名
- 买家商品之类目最小单序排名
- 买家商品之类目最大单序排名
- 买家商品之类目单序极差排名
- 买家商品之类目中位单序排名
- 买家商品之最小类目单序排名

模型甲 其它特征

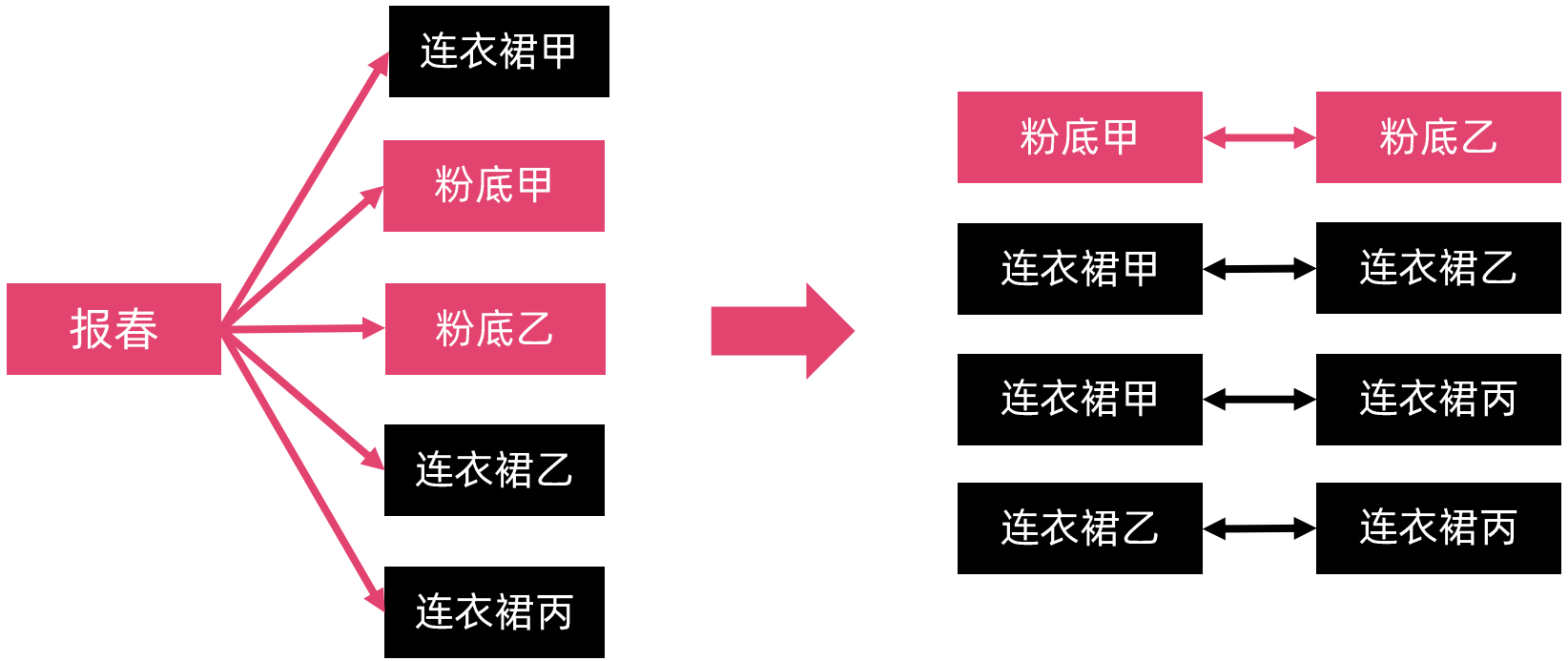
江离数据挖掘俱乐部

- 买家商品之店铺点击数排名
- 买家商品之店铺最后秒点击数排名
- 买家商品之店铺近一小时点击数排名
- 买家商品之店铺近两小时点击数排名
- 买家商品之店铺近三小时点击数排名
- 买家商品之店铺当日点击数排名
- 买家商品之店铺购买日数排名
- 买家商品之店铺最小购买日序排名
- 买家商品之店铺最大购买日序排名
- 买家商品之店铺最小秒序差排名
- 买家商品之店铺最大日序差排名
- 买家商品之店铺最小单序排名
- 买家商品之店铺最大单序排名
- 买家商品之店铺单序极差排名
- 买家商品之店铺中位单序排名
- 买家类目之点击数排名
- 买家类目之最小单序排名
- 买家店铺之点击数排名
- 买家店铺之最小单序排名
- 商品价格

模型乙 关联商品

江离数据挖掘俱乐部

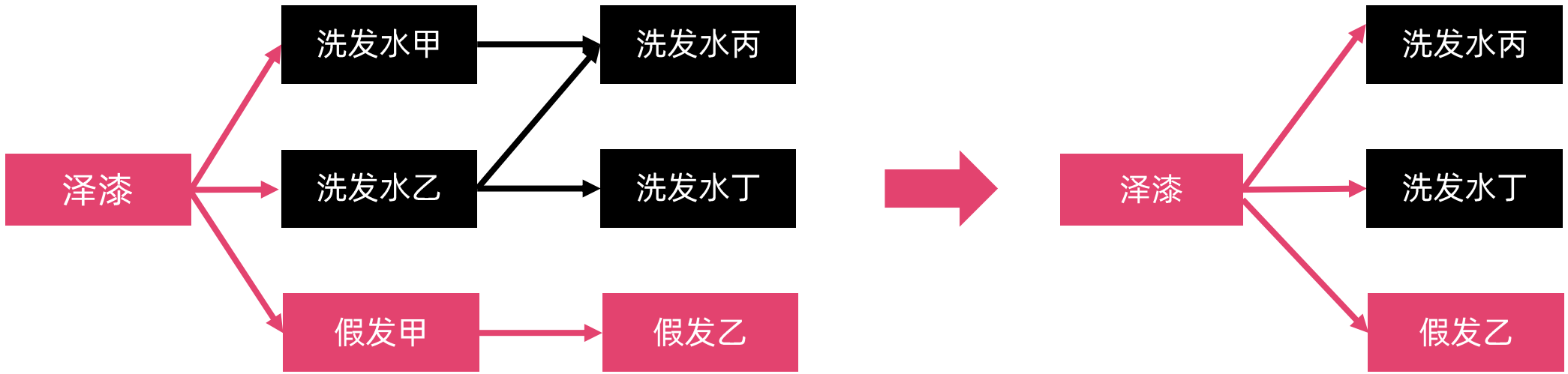
训练用户报春同学的行为记录（非真实数据，仅示意）



模型乙 未交互商品模型

江离数据挖掘俱乐部

测试用户报春同学的行为记录（非真实数据，仅示意）



交互过　　关联　　候选

2.构造特征

商品特征	商品之点击数 商品之点击买家数 商品之购买买家数 商品之点击买家购买比例 商品之购买买家平均点击数 商品之买家购买日数大于1数 商品之买家购买日数大于1比例
买家类目特征	买家类目之点击数 买家类目之最后日点击数 买家类目之商品数 买家类目之最小价格 买家类目之最大价格 买家类目之平均价格 买家类目之最后价格 买家类目之最小单序 买家类目之最小价格与商品价格差 买家类目之最大价格与商品价格差 买家类目之平均价格与商品价格差 买家类目之最后价格与商品价格差
买家店铺特征	买家店铺之点击数 买家类目之最后日点击数 买家店铺之商品数 买家店铺之购买商品数 买家店铺之最后日购买商品数 买家店铺之最小单序
其它	价格 候选打分

江离数据挖掘俱乐部

其它问题

江离数据挖掘俱乐部

- 数据中的「是否购买」含义较为模糊，它可能是指「用户在整个训练和测试时间段是否购买过这件商品」。是否应该改为「当次行为是否是购买行为」？

- 根据赛题，「测试数据中每个用户的最后一条点击数据所对应的商品一定在训练数据中出现过」。是否应该去掉这个规则？

完