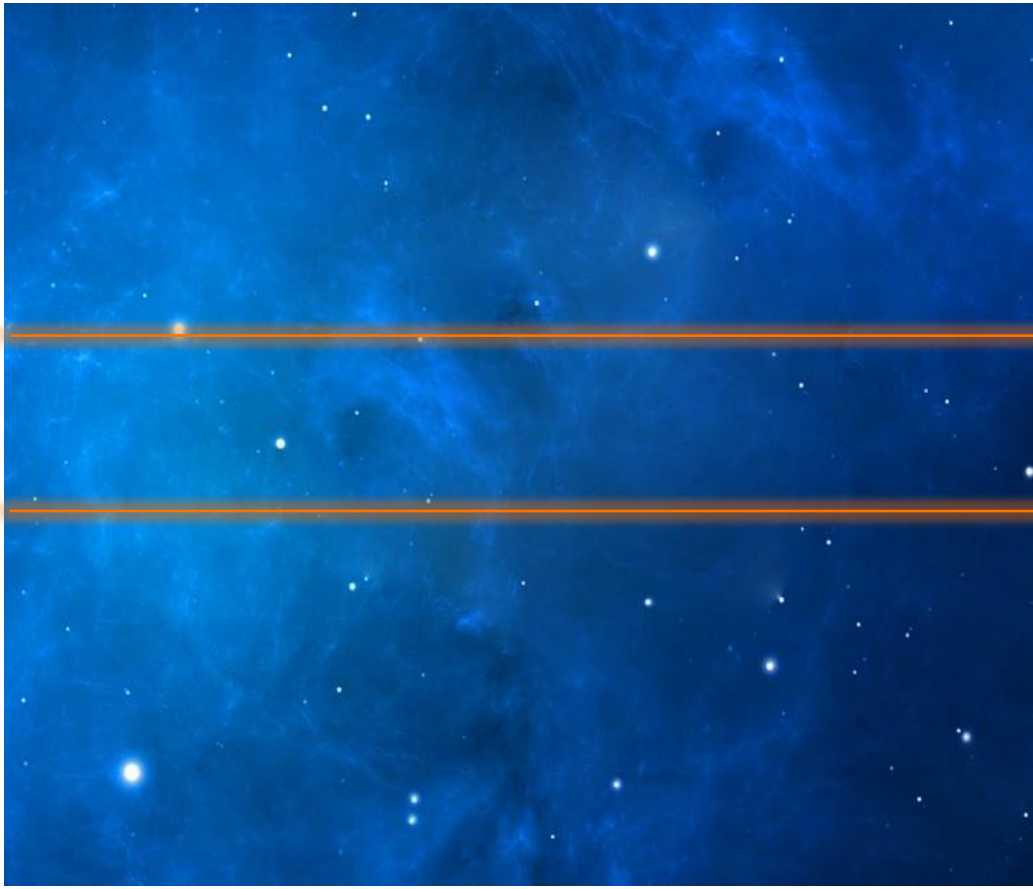
阿里巴巴大数据竞赛

天猫推荐算法大挑战

第二赛季　总决赛

阿里巴巴大数据竞赛分享

目录



目录

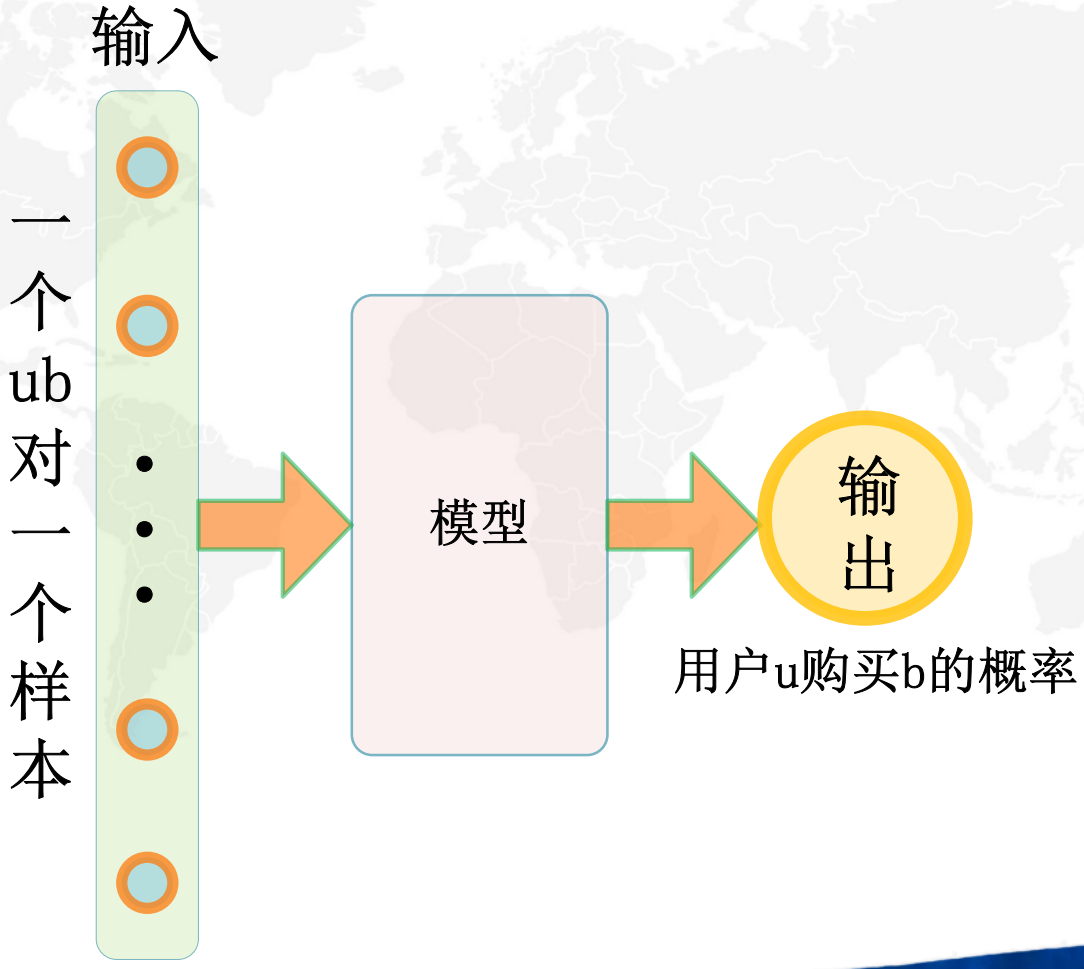
■ 赛题的任务

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会购买哪些品牌。

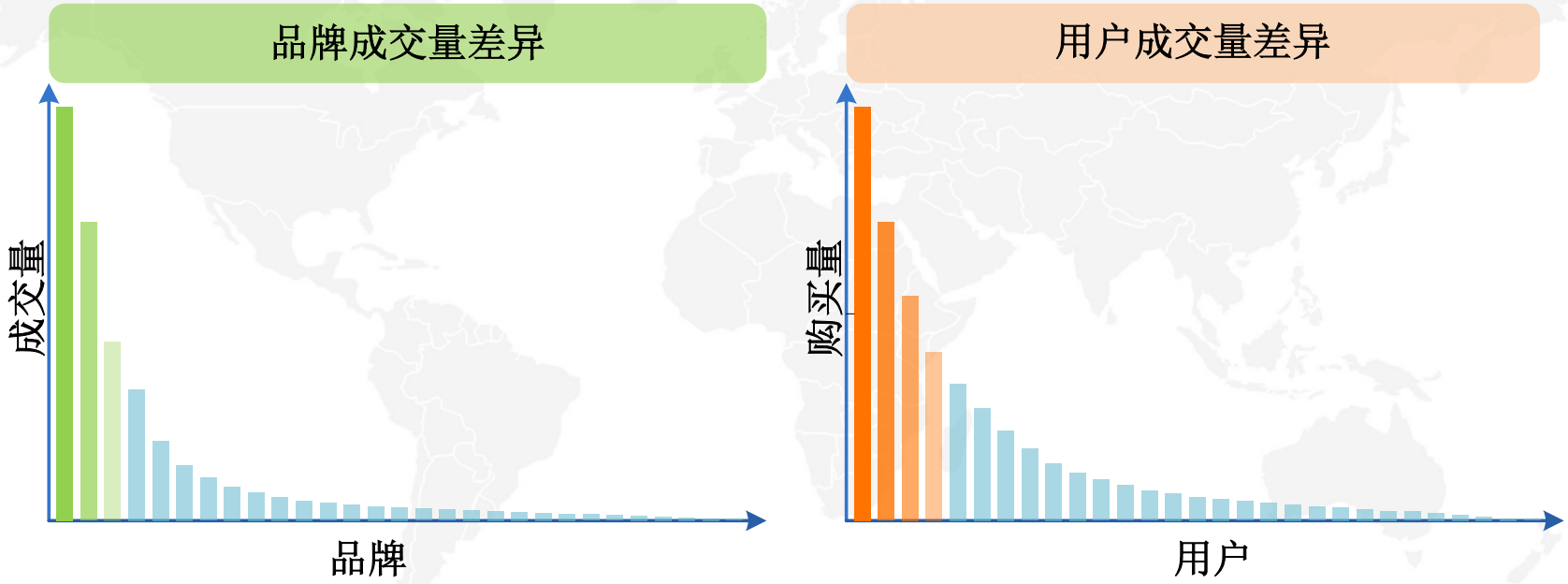
$$\text{Precision}=\frac{\sum_{i}^{N} hitBrands_i}{\sum_{i}^{N} pBrands_i} \qquad \text{Recall}=\frac{\sum_{i}^{M} hitBrands_i}{\sum_{i}^{M} pBrands_i} \qquad F_1=\frac{2*P*R}{P+R}$$

根据用户4个月在天猫的行为日志，预测用户u在将来一个月是否会购买这个品牌b——**分类问题**

■ 解题思路

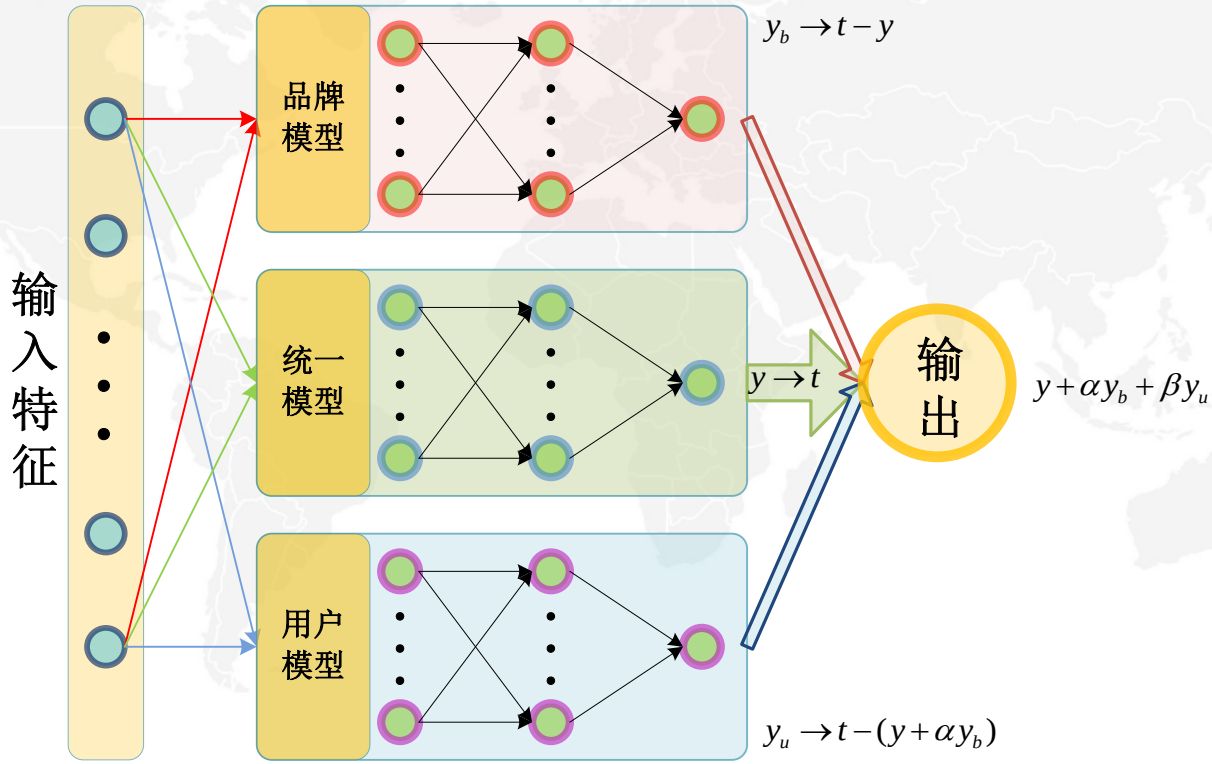


■ 用户和品牌的类内差异



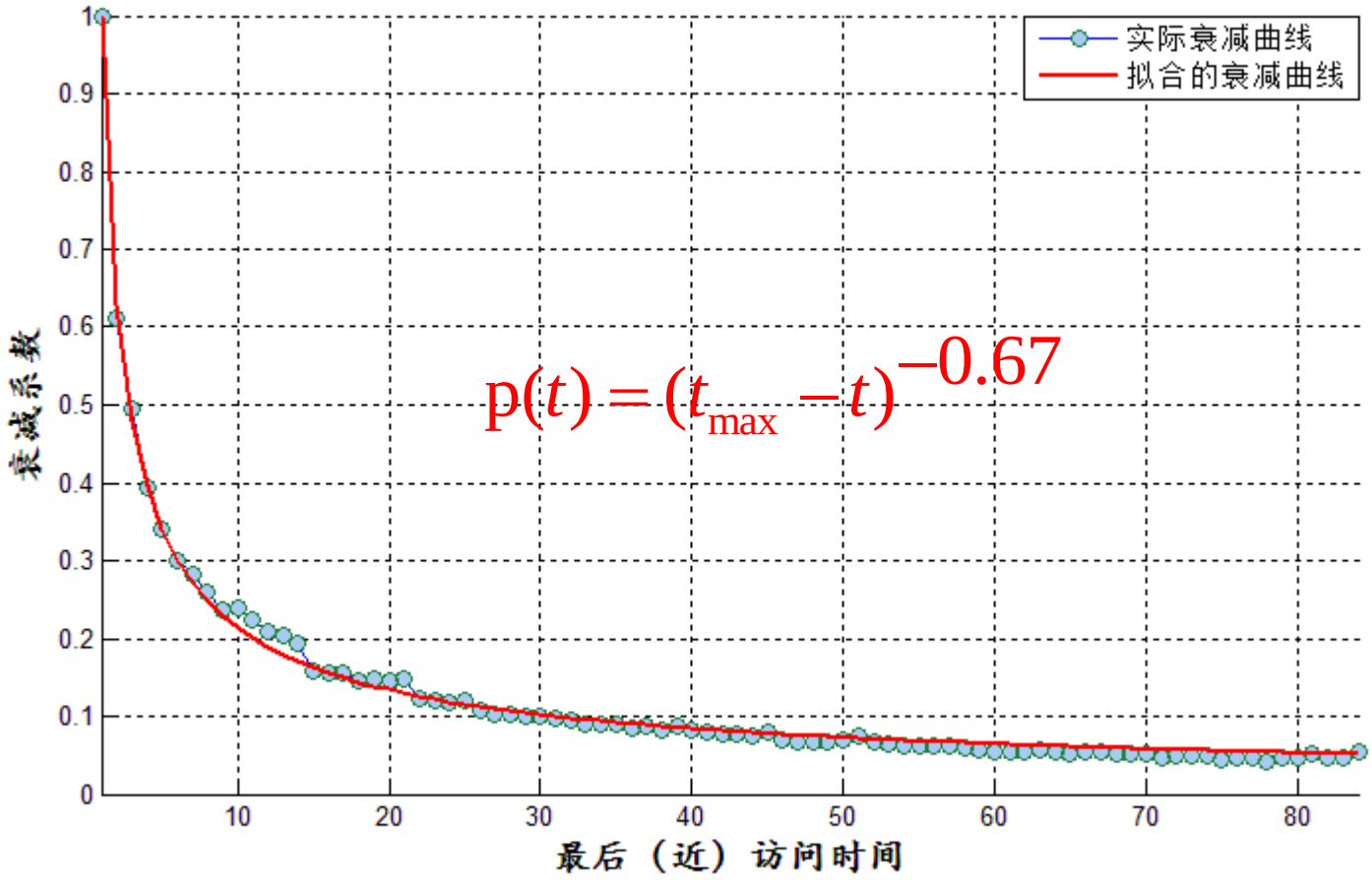
物以类聚，人以群分

■ 三位一体架构

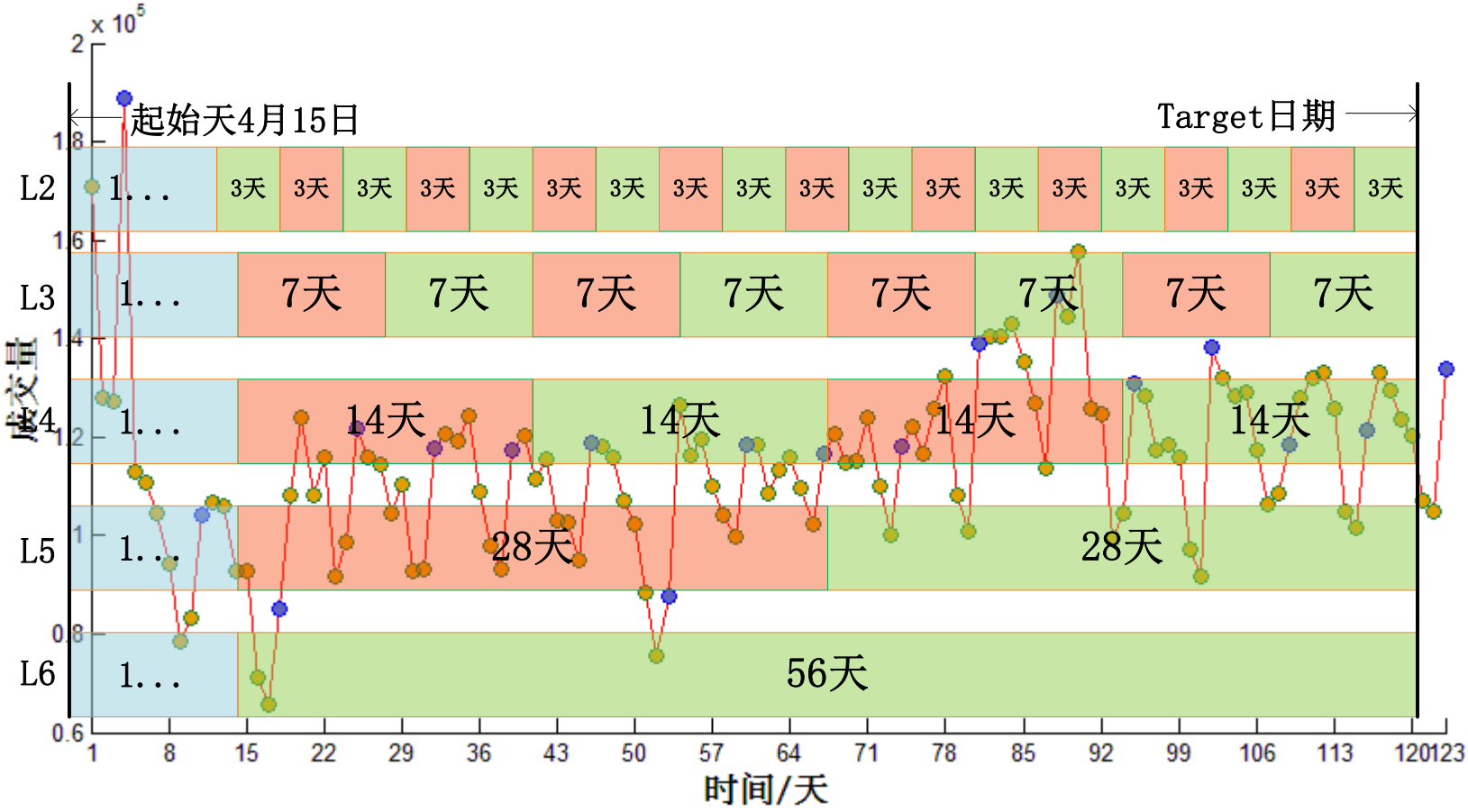

输入特征
品牌模型
统一模型
用户模型
$y_b \to t - y$
$y \to t$
$y_u \to t - (y + \alpha y_b)$
输出
$y + \alpha y_b + \beta y_u$

目录

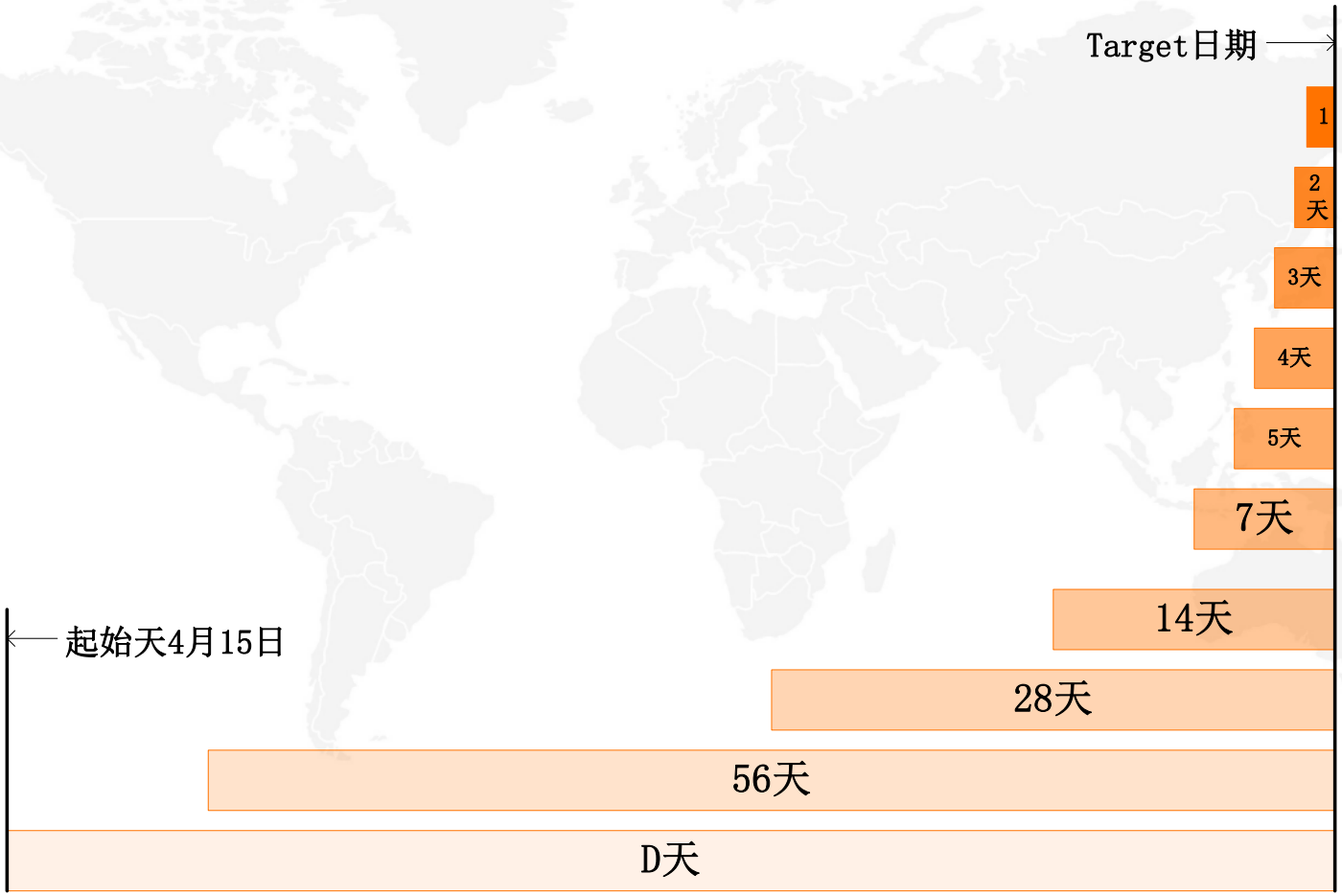
■ 指导思想——源于对数据分析和业务理解



■ 提取方式——层级提取


x 10^5
起始天4月15日
Target日期
L2
L3
L4
L5
L6
1...
3天
7天
14天
28天
56天
成交量
时间/天

提取方式——最近K天0/1提取



■ 特征类型

行为特征——描述用户行为

最后访问时间

$$g(t,\sigma)=\exp(-\frac{(t_{\max}-t)^2}{2\sigma^2})$$

$$\sum_{a=0}^{3} w_a \sum_{t=1}^{D} p(t)\log(x_a(t)+1)$$

平均访问间隔

层级行为次数

层级行为天数

对数特征

$$\sum_{t=1}^{D} p(t)x_a(t)$$

品牌特征——描述品牌属性

层级四种行为次数

重复购买率

最近K天频率特征

转化率

老顾客比例

购买前访问次数

成交量变化趋势

用户特征——描述用户属性

- 四种行为次数
- 购买频率
- 用户行为天数
- 用户转化率
- 用户等级
- 总购买品牌数
- 平均购买前访问天数

交叉特征

行为品牌交叉

访问次数x用户转化率

行为用户交叉

品牌圈特征

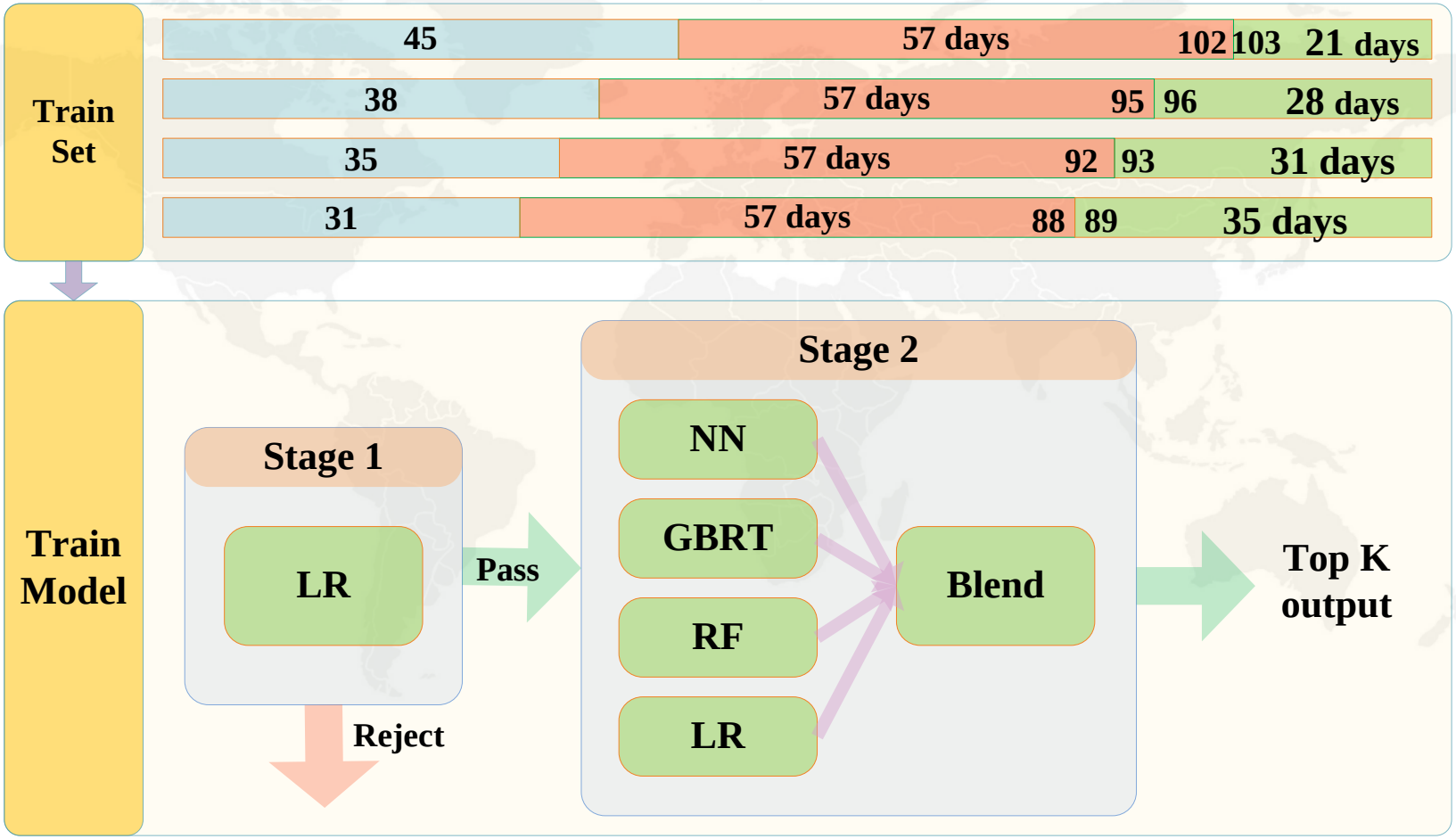
访问次数x品牌转化率

U-B对点击次数/用户总点击次数

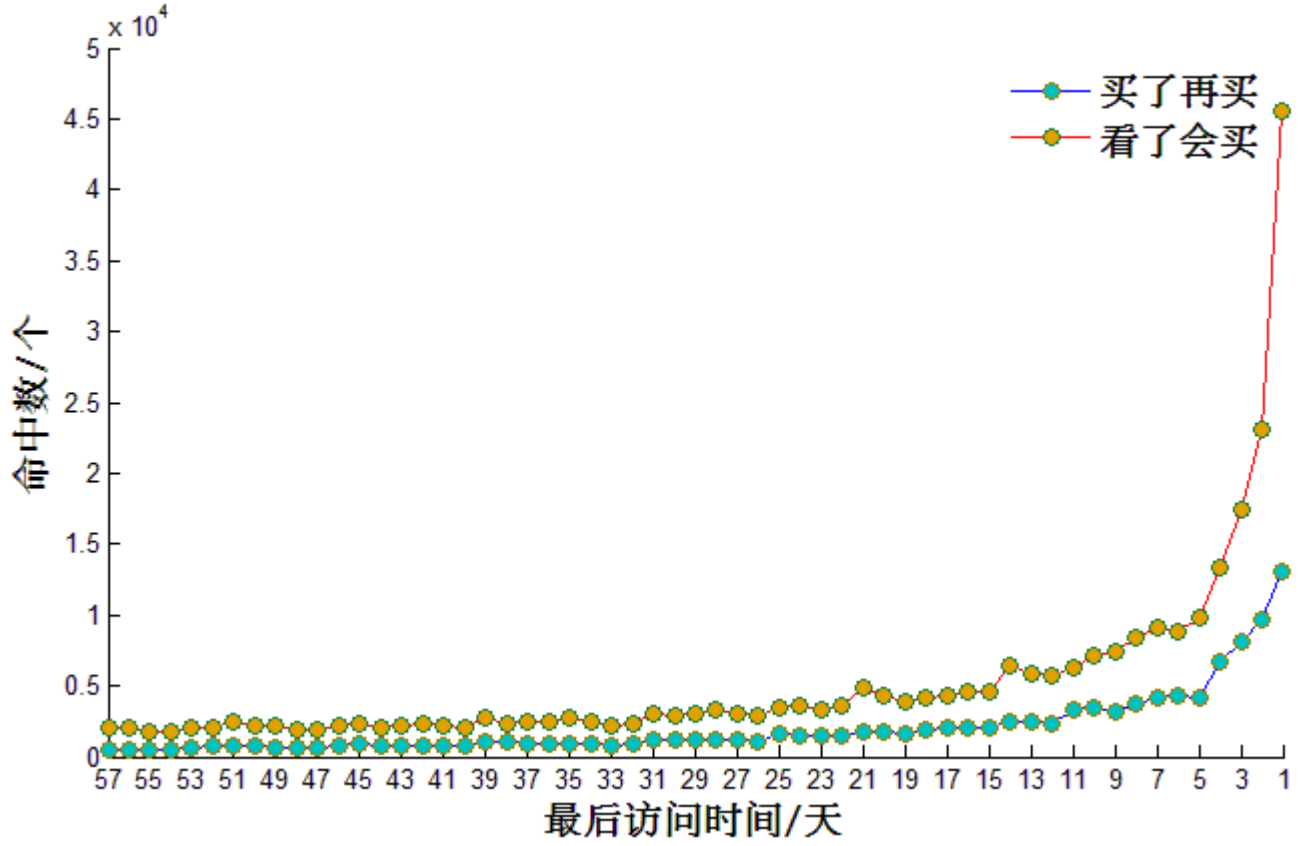
品牌用户交叉

目录

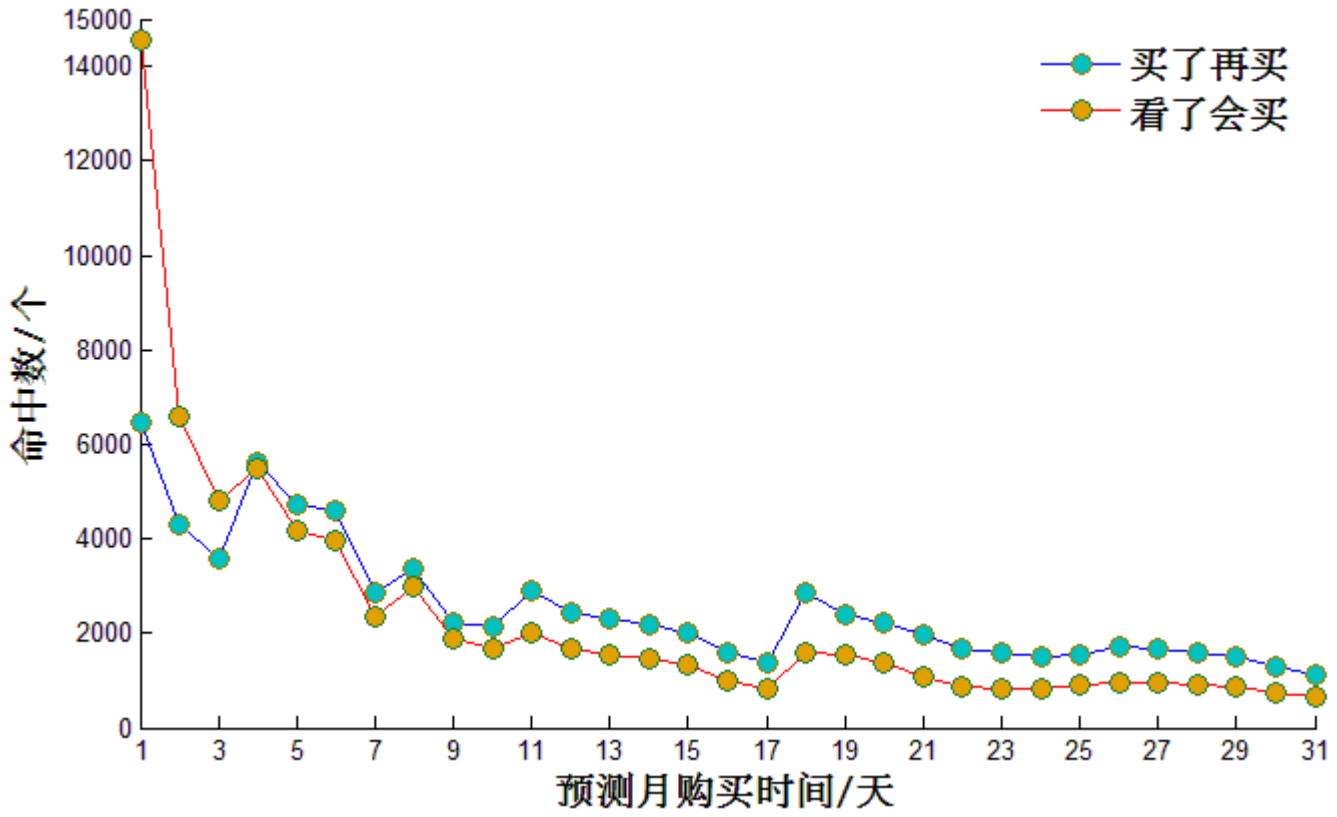
算法整体框架


Train Set
45
57 days
102
103
21 days
38
57 days
95
96
28 days
35
57 days
92
93
31 days
31
57 days
88
89
35 days
Train Model
Stage 1
LR
Pass
Reject
Stage 2
NN
GBRT
RF
LR
Blend
Top K output

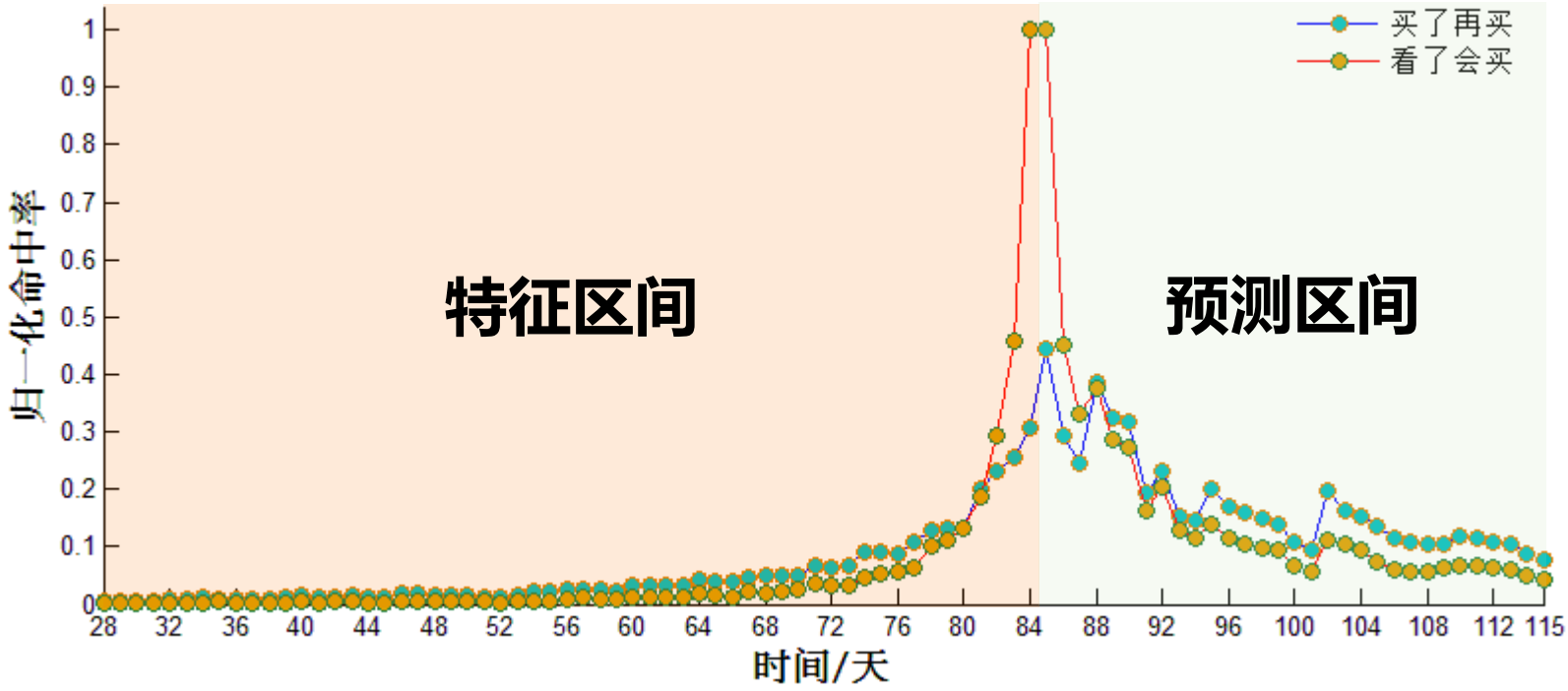
■ 不同行为命中的时间分布



■ 不同行为 模型命中时间分布



■ 模型的预测命中分布



不同行为的差异，算法怎么解决?

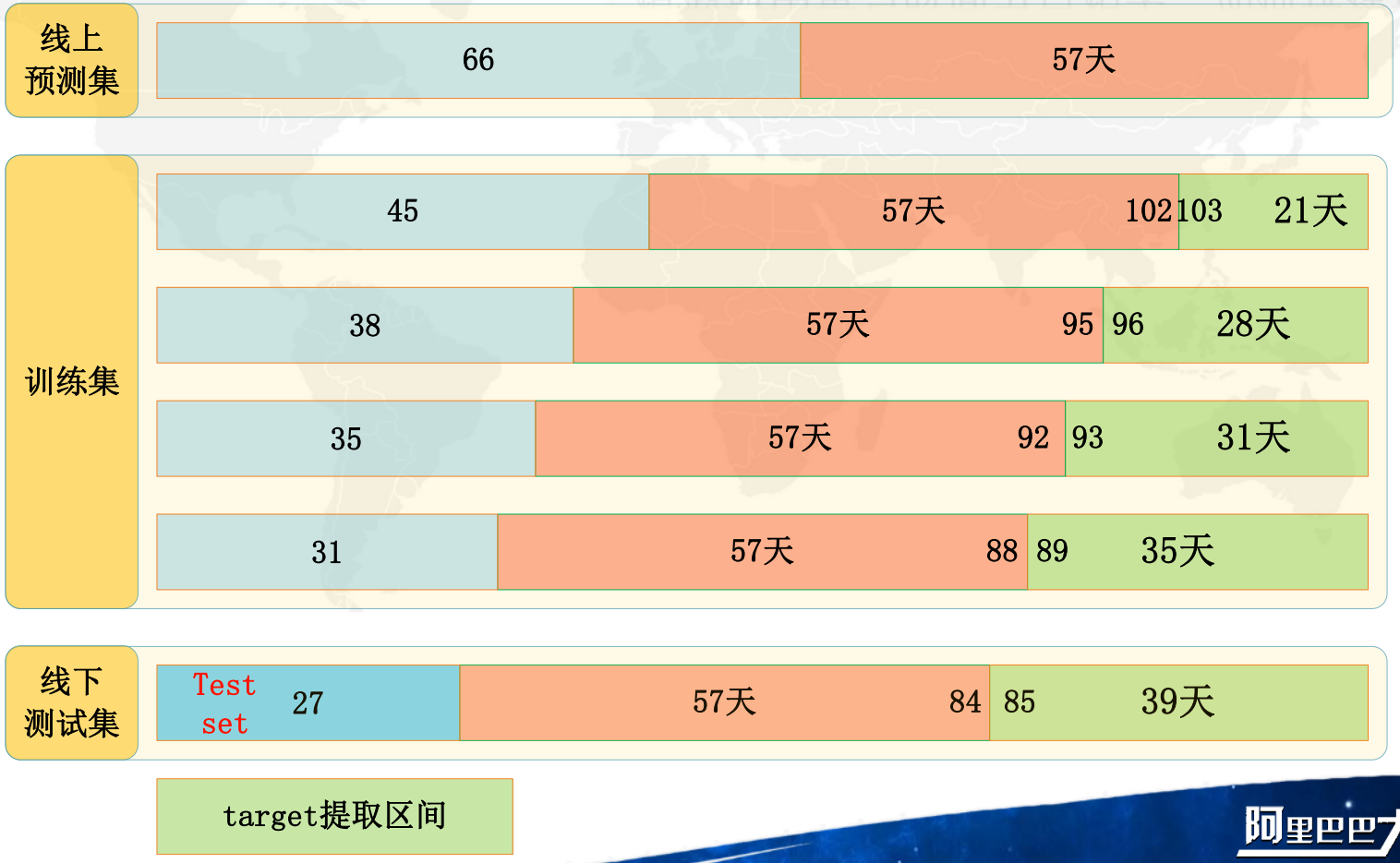
模型命中量与时间节点相关，如何选择最优?

不同行为单独训练

不同行为的差异，算法怎么解决?

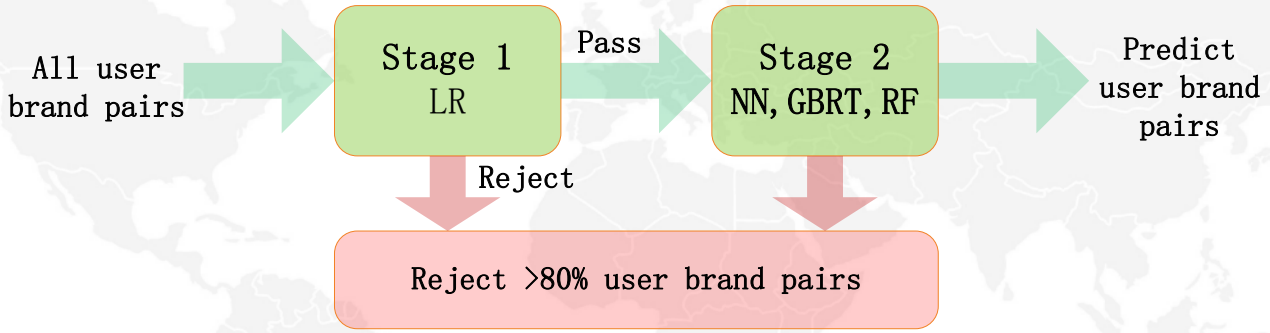
多集合滑窗

模型命中量与时间节点相关，如何选择最优?



■ 如何解决大海捞针问题？

1亿多 vs 40多万



- 通过快速的快速的LR可以滤除>80% 的样本，节约训练预测时间
- 还有助于提高F1分数

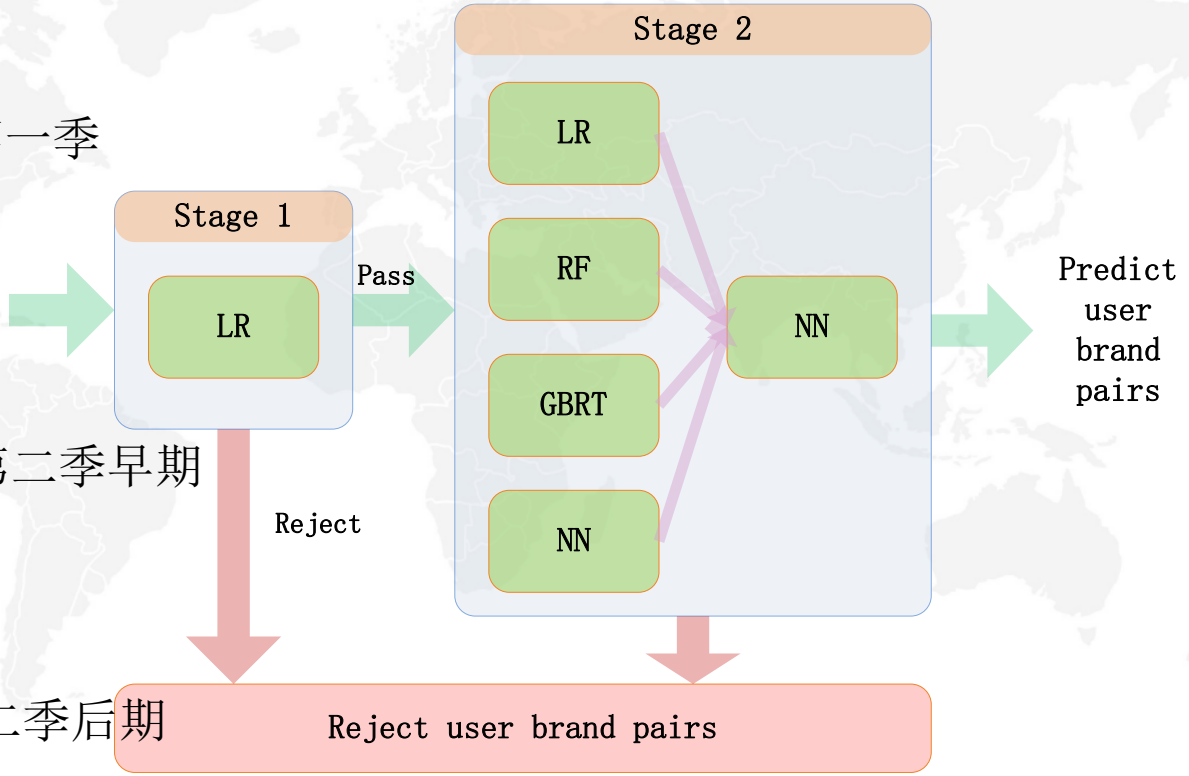
模型融合方法

- Top K 输出合并
- 排序融合
- 加权分数融合
- 线性模型融合
- 非线性模型融合

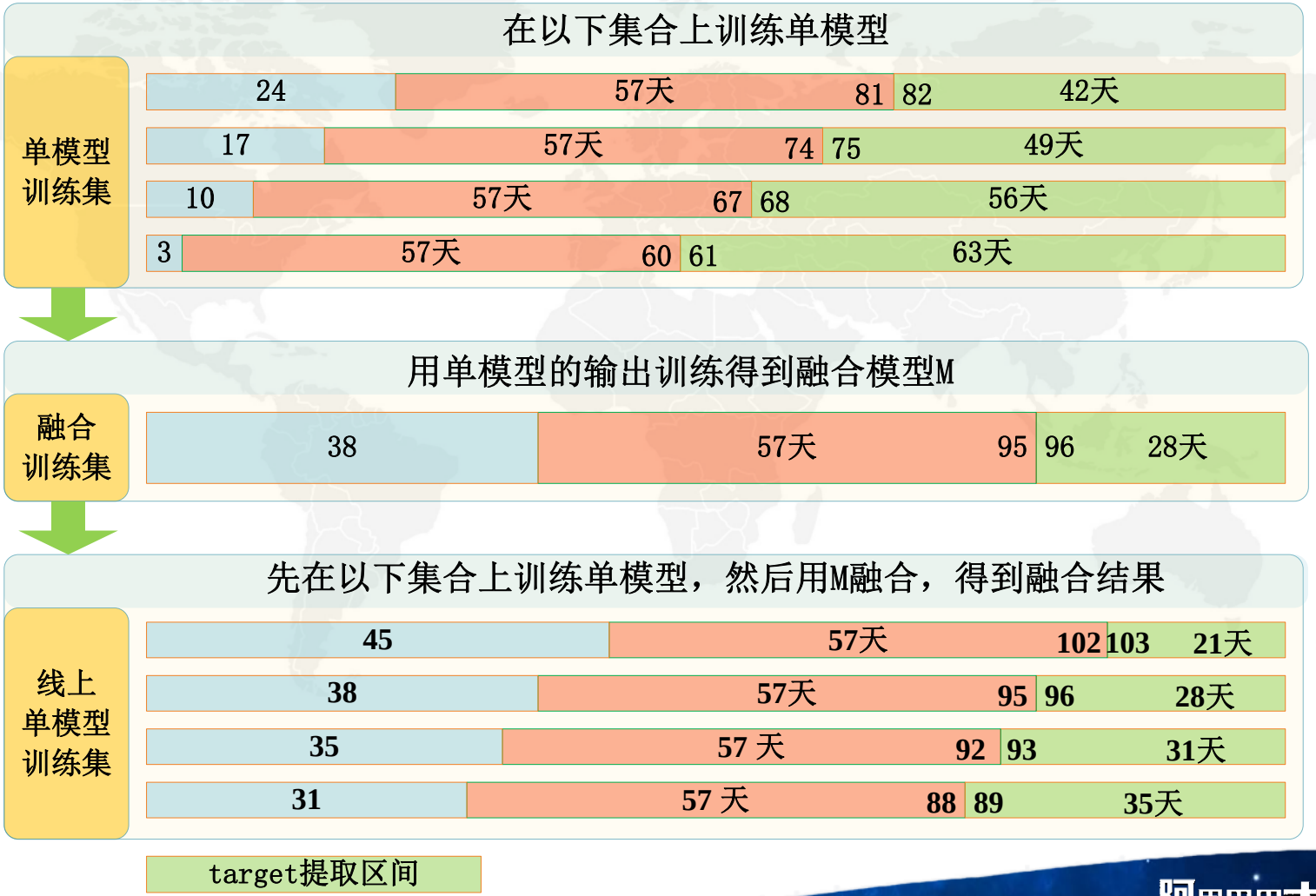
第一季

第二季早期

第二季后期



■ 多集合非线性融合


在以下集合上训练单模型
单模型训练集
24
57天
81
82
42天
17
57天
74
75
49天
10
57天
67
68
56天
3
57天
60
61
63天
用单模型的输出训练得到融合模型M
融合训练集
38
57天
95
96
28天
先在以下集合上训练单模型，然后用M融合，得到融合结果
线上单模型训练集
45
57天
102
103
21天
38
57天
95
96
28天
35
57 天
92
93
31天
31
57 天
88
89
35天
target提取区间

Thank you!

please give some questions