达观个性化推荐算法挑战赛

This_is_test

团队介绍

西安电子科技大学----苏军平

2016 CCF 基于地理位置的精准营销 冠军

销量预测大赛　三等奖

2016 DataCastle“智慧中国杯”全国大数据创新应用大赛（融360“用户贷款风险预测”优秀奖 ）

2016 DataCastle“智慧中国杯”全国大数据创新应用大赛（ “大学生助学金精准资助”17/1964）

1，赛题理解
2，数据探索
3，建模
4，特征工程
5，算法模型
6，总结与展望

赛题理解（1）

1、推荐问题

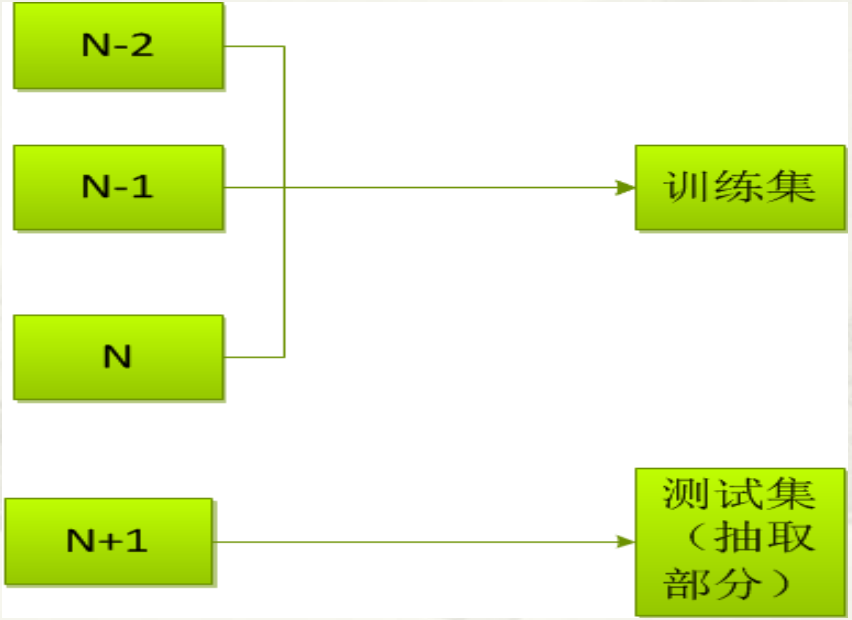
备选集&&二分类排序

2、新闻资讯推荐

新闻资讯可能涉及到热度性质、时效性

3、TOPK推荐

赛题理解（2）



给定三天的数据
预测第四天的资讯情况。

数据探索

Trick1: 用户持续关注性

用户对某新闻资讯具有偏执性，连续几天一直发生行为。

Trick2: 新闻热度性

* 思路来源：kaggle比赛：Outbrain Click Prediction Competition第二名的赛后分享
* 详细细节可以参考kaggle blog

We also hashed all **combinations** of the `document` / `traffic_source` **clicked by a user in** `page_views`. So, if a user came to a document from 'search', it would be treated differently to a user coming to the same document from 'internal'. Any documents occurring less than 80 times in `events.csv` were dropped, because sparse documents tended to just add noise to the model. This information was quite strong in the model as it brought user level preferences in clicks.

数据探索

Trick2:新闻热度性,一些代表意义的统计数据

第16天TOP120　79%

第17天TOP120　81%

第18天TOP120　80%

建模（1）——备选集构建

测试集来源于第19天数据，认为是随机抽取的部分数据，即认为测试集与第19天的数据是同分布的。

利用TRICK2进行构造备选集。即将每天的TOP120热度资讯当作每个用户的候选集。

建模（2）–训练集、验证集构建

模型1：借鉴时间滑窗的思想

训练集：16天当作历史数据，17天为label

验证集：17天当作历史数据，18天为label

预测集：18天当作历史数据，19天进行预测

总结：这种建模方法可以构造验证集，调参方便，不依赖于线上。

建模（3）——训练集、验证集构建

模型2：两天提取特征

训练集：16、17天做历史数据，18天为label

预测集：17、18天做历史数据，19天预测

总结：特征维度上更丰富、更精细，但是无法构造验证集，参数调节需要依赖于线上

建模（4）


TOPN
模型1
训练集16→ 17
验证集17→ 18
测试集18→ 19
训练集16、17→ 18
测试集17、18→ 19
模型2
模型融合

特征工程

1，用户画像刻画

基于用户的历史行为，进行用户画像的刻画。

2，资讯特征

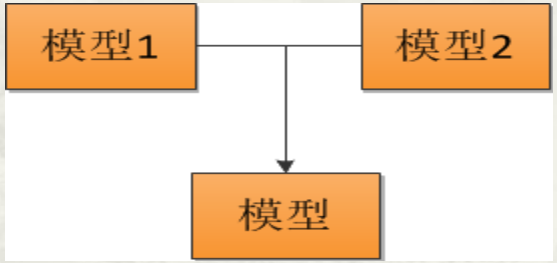
基于该资讯历史数据，进行资讯特征的提取。

算法模型

模型1：0.8*XGBoost + 0.2*Random Forest

模型2: XGBoost

模型融合：



成绩：

5	↓1	this_is_test	0.074838

0.0732325961272314

总结

1，备选集构造

备选集的构造基本是基于资讯热度的，对于个性化体现的不是很明显，没有根据具体用户构造备选集。

2，个性化推荐

特征维度不够丰富，个性化不够明显。

3，算法模型

规则模型、朴素贝叶斯模型。

致谢

感谢“达观”数据与“科赛”平台提供的这次宝贵的比赛&&交流机会。

感谢各位专家与同行的莅临指导。