广东公共交通大数据竞赛
——市民出行公交线路选乘预测

2015天池大数据竞赛

TIANCHI天池

队名：KevinNing

2015年12月22日

提纲

- 个人介绍
- 赛题介绍
- 解决方案
- 总结与思考

个人介绍

- **宁克锋**（正澄）

毕业于清华大学自动化系

从事推荐算法相关工作

赛题介绍

- 以市民出行公交线路选乘预测为赛题
- 基础：广东省部分公交线路的历史公交卡交易数据、气象数据等
- 手段：挖掘固定人群在公共交通中的行为模式，分析推测乘客的出行习惯和偏好
- 目的：建立模型预测人们在未来一周内将会搭乘哪些公交线路
- 终极目标：为广大乘客提供信息对称、安全舒适的出行环境，用数据引领未来城市智慧出行。

- 评测指标：

$$\text{Precision} = \frac{|\cap(PredictionSet,\ ReferenceSet)|}{|PredictionSet|}$$

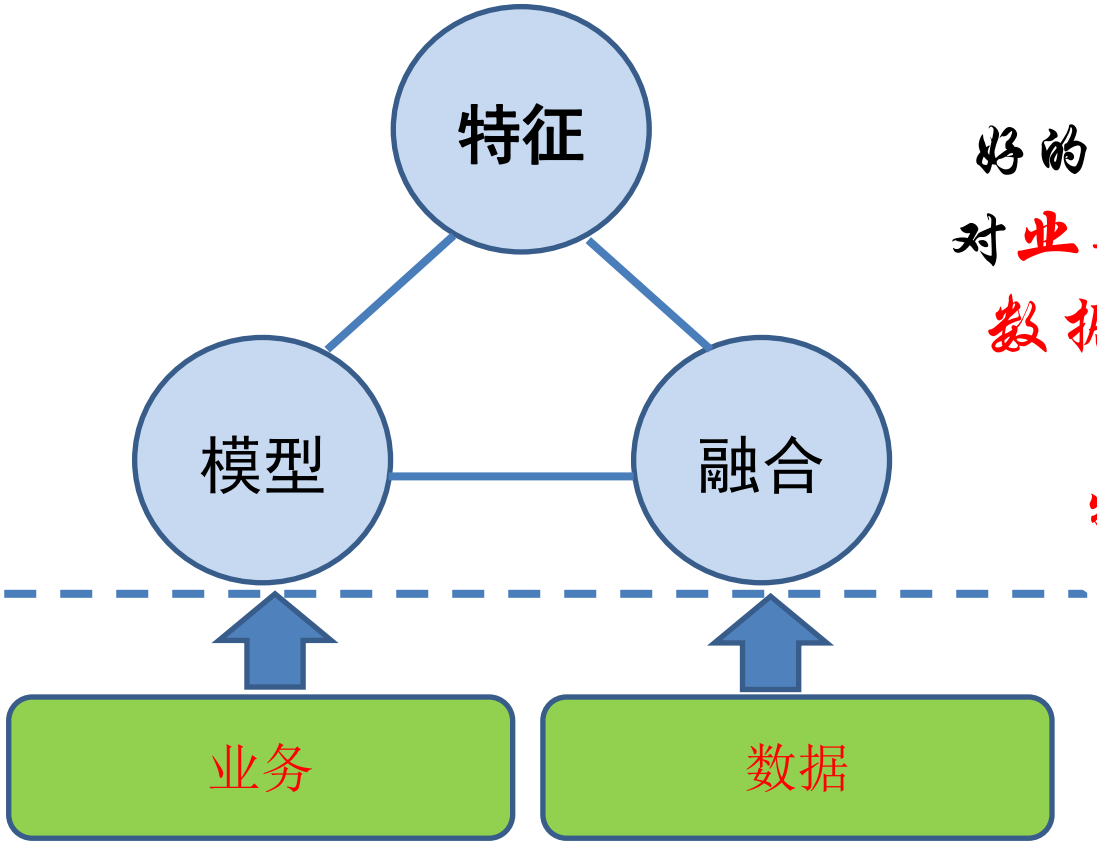
$$\text{Recall} = \frac{|\cap(PredictionSet,\ ReferenceSet)|}{|ReferenceSet|}$$

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

解决方案

铁人三项

两大支撑



好的解决方案源于对业务的深入理解及数据的细致分析！

特征是纽带！

业务为本

- 不同人群，对出行的需求不同

- 我们是上班族：工作日/上班高峰期/下班高峰期等出行刚需

- 我们是老人：出行时间不固定

举例：某上班族部分公交日志如下

A	B	C	D	E	F
1	3	297	579932	5	2014080807
1	3	297	579932	5	2014082007
1	3	297	579932	5	2014082118
1	4	359	579932	5	2014082617
1	6	359	579932	5	2014082617
1	3	297	579932	5	2014082707
1	1	243	579932	5	2014090108
1	2	243	579932	5	2014090108
1	4	243	579932	5	2014090108
1	6	243	579932	5	2014090108
1	7	243	579932	5	2014090108
1	3	297	579932	5	2014090108
1	3	297	579932	5	2014090220
1	1	33	579932	5	2014090508
1	6	33	579932	5	2014090508
1	7	33	579932	5	2014090508
1	3	297	579932	5	2014090508
1	1	20	579932	5	2014090516
1	4	35	579932	5	2014091507
1	3	297	579932	5	2014091917
1	1	142	579932	5	2014092608
1	4	142	579932	5	2014092608
1	3	297	579932	5	2014092608
1	1	17	579932	5	2014092618
1	2	17	579932	5	2014092618

业务为本

- 天气对不同人群出行的影响不同

- 我们是上班族：工作日没办法，下雨也得挤公交，周末天气不好就不出去了

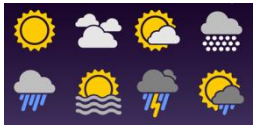
- 我们是老人：不管工作日还是周末，天气好就出去遛遛，不好就呆在家吧

需要构造特征对其进行刻画，使模型能更好的反映实际业务特点

数据为源

- 探索数据边界，知道手中有哪些武器可以使用




线路信息	公交卡信息	乘车信息	气象信息	时间信息
线路类型、站数等	公交卡类型、发卡地等	线路、公交卡、终端id、时间等	雨/阴/晴、温度、风向等	工作日/假日、上下班高峰期

数据为源

- 数据初窥：

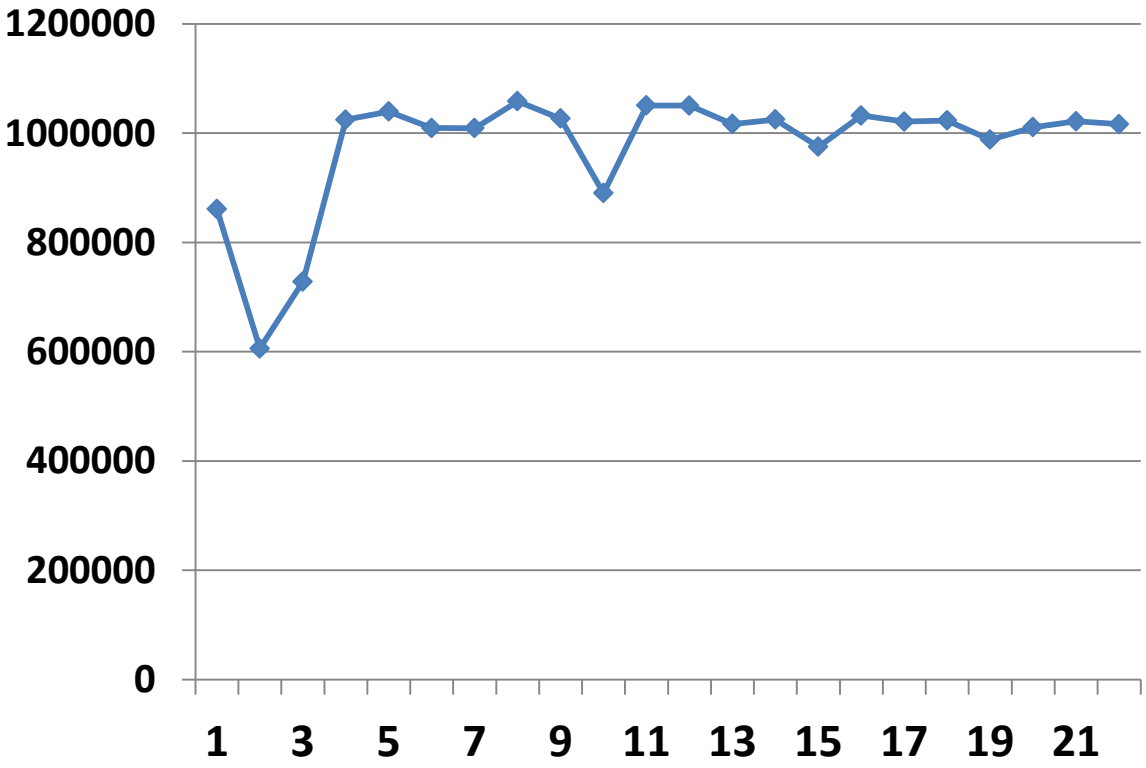
	p1	p2	p1+p2
公交卡数	3448699	4440906	5527605
公交线路数	7	7	14
公交车数	360	439	748
乘车日志数	32493555	26628650	59122205
公交卡类型	7	7	7
公交卡发卡地	20	20	20

- p1公交卡和公交车数量多，日志数反而比p2少？

  - p1为广佛跨区域，p2为广州市区
  - 前者站多，且更多上班族通勤需求

数据为源

- 以7天为周期，card_id、line_name组合数（p1）



- 8月11-15等日期大量日志丢失（噪声多）
- 十一国庆数据偏低（上班族影响）

数据预处理

数据预处理的重要性：garbage in garbage out

数据处理错误

- p1中大概有40%的terminal_id对应两条以上的line_name
  - 对terminal分为两类：对应1条、对应2条以上
  - 生成特征时分开统计

系统记录错误

- 同card_id在同时段同line_name刷卡次数大于2次，有些多达几十次
  - 同card_id同时段同line_name进行排序
  - 对rank大于2的日志进行过滤

数据预处理

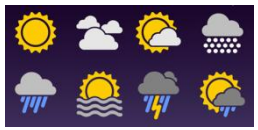
将原始数据处理成模型更易理解的数据

白天温度

- 档位1：10-20度
- 档位2：20-30度
- 档位3：30度+

夜晚温度

- 档位1：0-10度
- 档位2：10-20度
- 档位3：20-30度



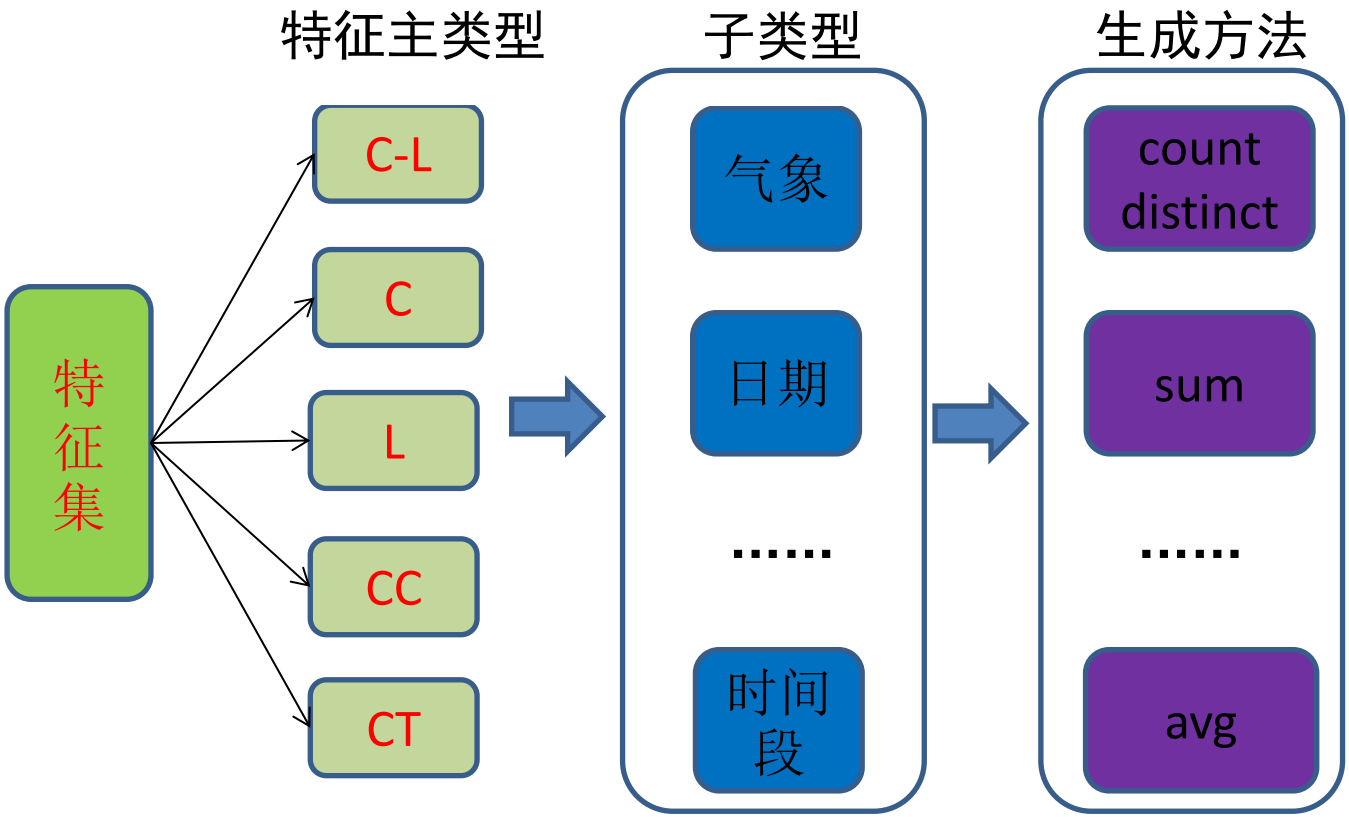
天气

- 类型1：晴、多云
- 类型2：阴、霾
- 类型3：雨

时间

- 工作日
- 周末
- 节假日
- 上/下班高峰期

相对通用的特征工程框架



C:Card_id　　CC:Card_type

L:Line_name　　CT:Create_city

特征设计

Card-Line Feature

- 1/3/7/28/70/126天乘车次数
- 乘车天数/小时数
- 不同终端乘车次数

- 工作日/周末/假日乘车次数
- 上/下班高峰期乘车次数
- 不同气象下乘车次数

- 平均乘车次数/频次
- 最大乘车间隔天数、最早/最近乘车时间

- 不同时间区间内只对应1个终端号的公交卡乘车次数
- ……

特征设计

除与Card-Line Feature相同含义特征外：

□ Line Feature

- 不同Card Type乘车次数
- 不同Create City乘车次数
- 不同Card 乘车次数

□ Card Feature

- 不同线路乘车次数
- Card Type数
- Create City数

□ Card Type Feature

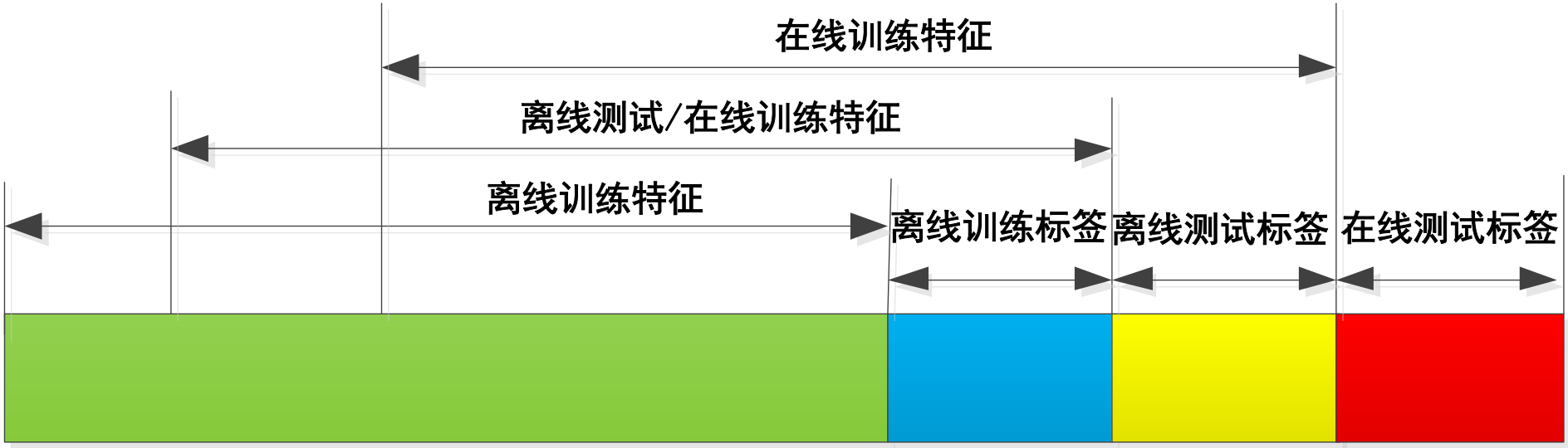
- 不同Card 乘车次数
- 不同线路乘车次数

□ Create City Feature

- 不同Card 乘车次数
- 不同线路乘车次数

离线评测

- 离线评测是重要环节
- 减少对线上提交的过度依赖



✓ 特征不要偷看标签数据

✓ 尽可能模拟在线测试环境

✓ 多个离线评测指标

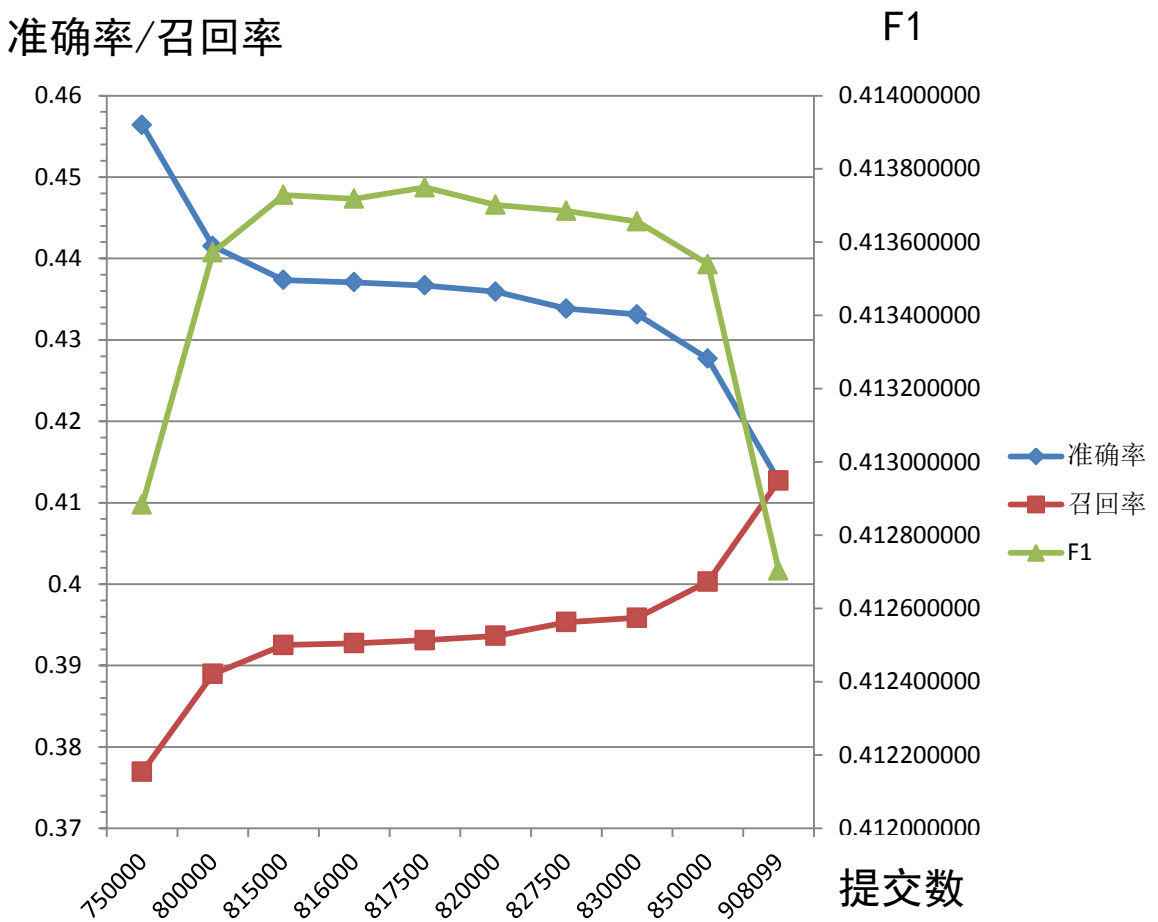
靠谱的离线评测

离线调优、线上成绩胸有成竹

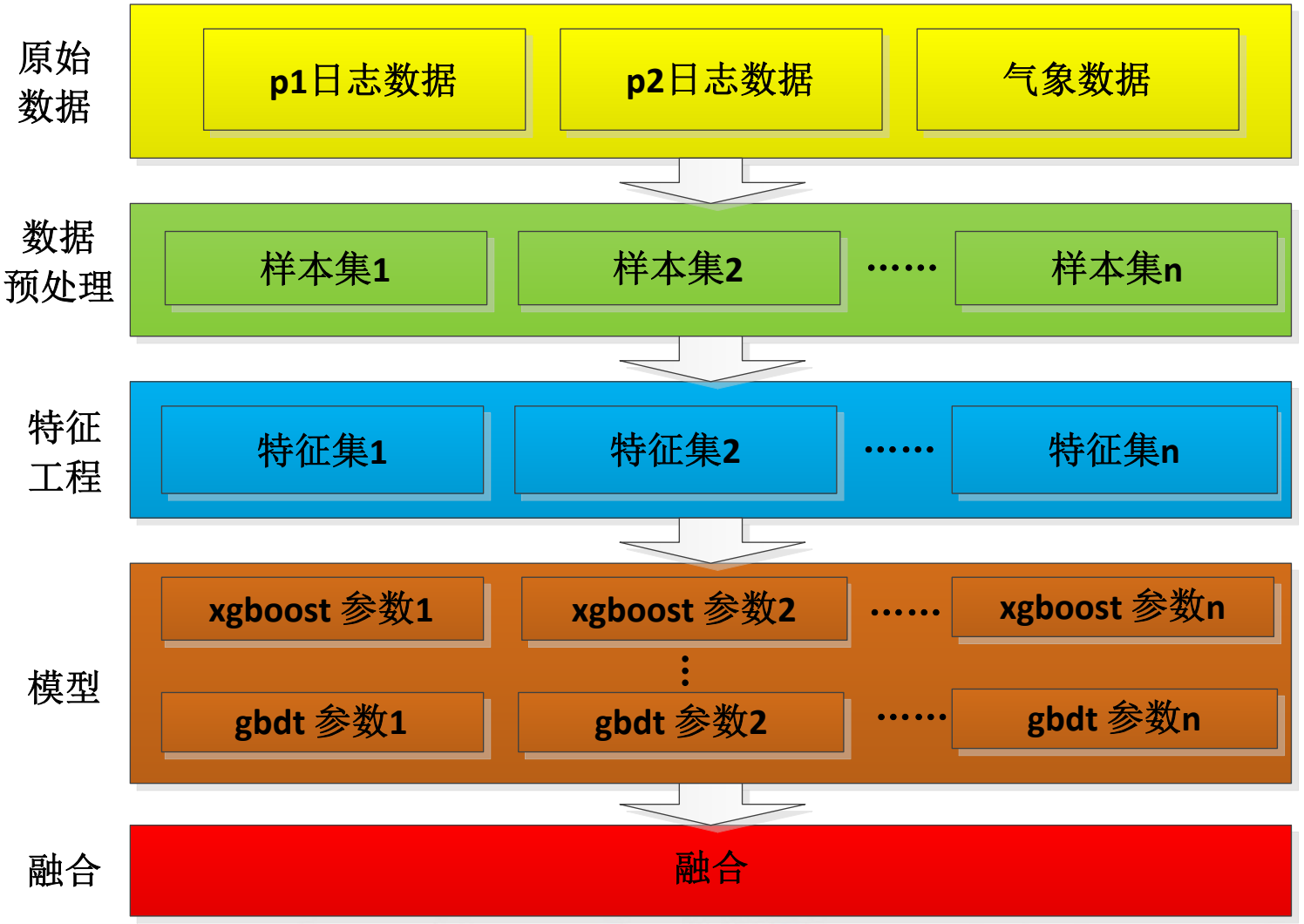
离线评测

提交数离线优化

- 准确率随提交数增加而降低
- 召回率随提交数增加而提升
- F1随提交数增加先提升后降低
- 离线实验结果：F1在提交数约为实际待预测数90%-92%左右取得最优

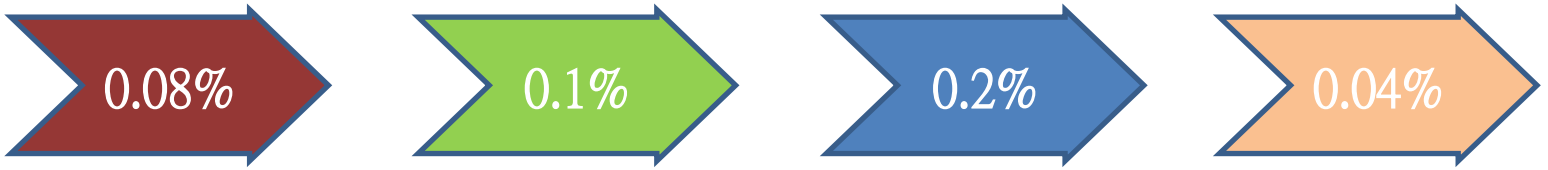


整体解决方案


原始数据
p1日志数据
p2日志数据
气象数据
数据预处理
样本集1
样本集2
……
样本集n
特征工程
特征集1
特征集2
……
特征集n
模型
xgboost 参数1
xgboost 参数2
……
xgboost 参数n
gbdt 参数1
gbdt 参数2
……
gbdt 参数n
融合
融合

回顾

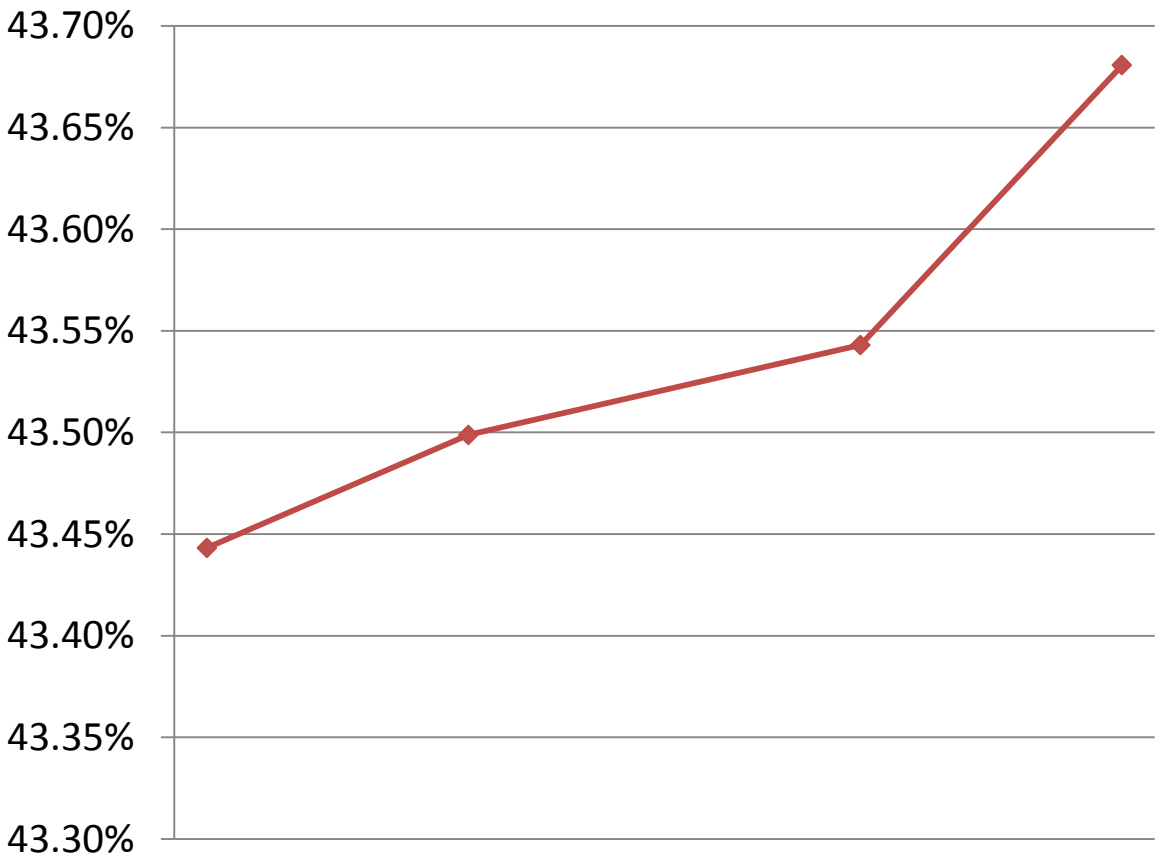
在构造基础特征的基础上，若干提升点：



样本去噪参数调优	利用气象/日期等	同时利用p1/p2数据	模型融合
数据预处理、熟悉模型特点、并调优	对裸数据进行编码，与原日志数据相结合	处理好p1/p2的数据及特征关系	只融合了7个xgboost，还有gbdt等未加入，预期0.1%

回顾

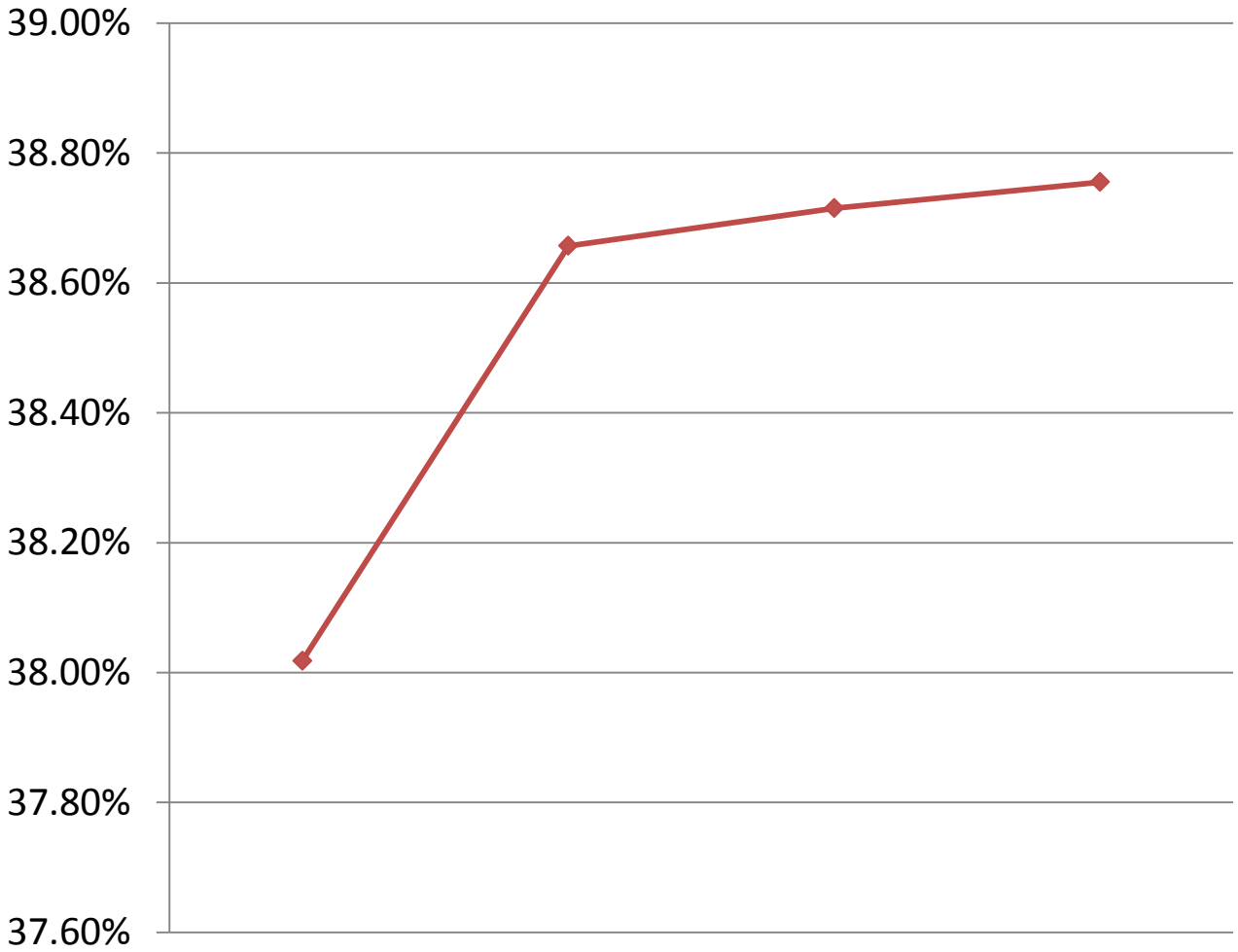
- Part 1
  - ✓ 首次提交模型
    43.44%
  - ✓ 模型调参：
    43.50%
  - ✓ 第二版特征：
    43.55%
  - ✓ 增加气象/日期等信息：
    43.68%



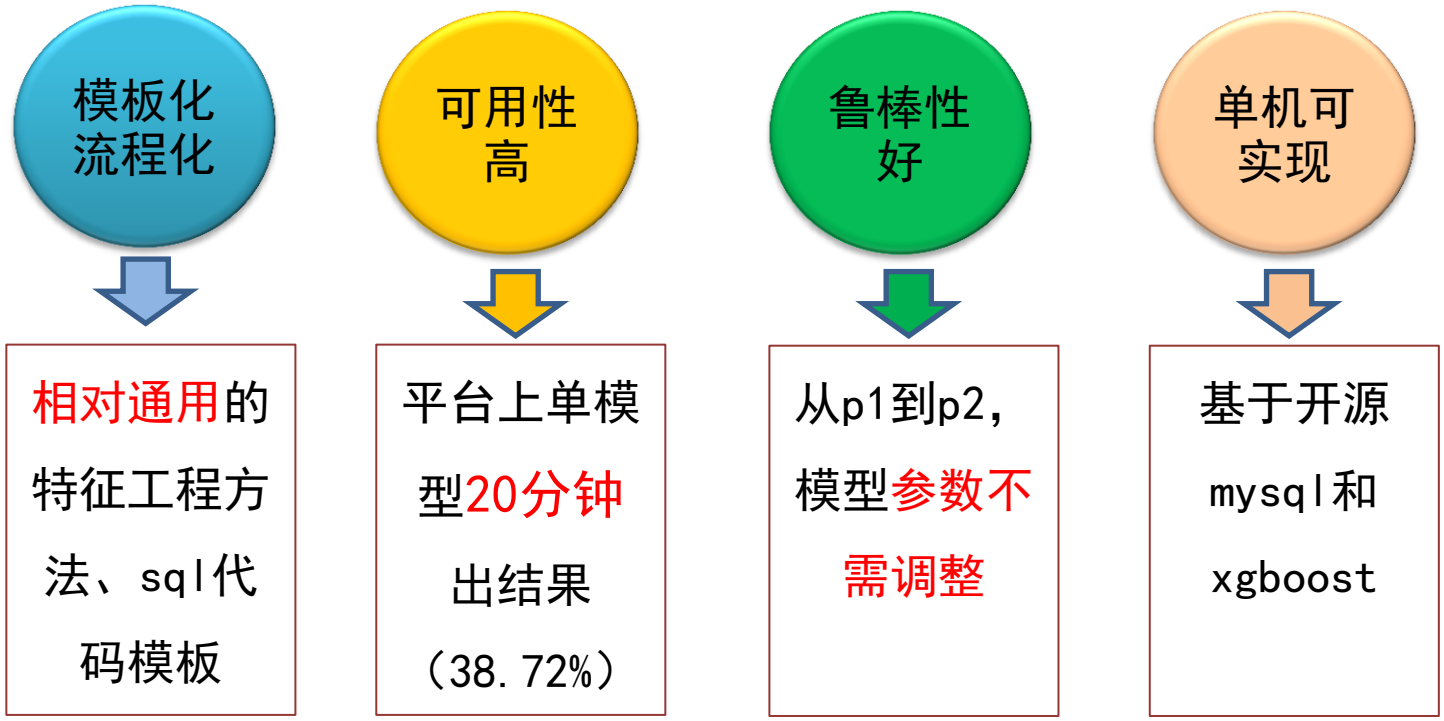
- 周末到了，小心放大招

回顾

- Part 2
  - ✓ 首次提交模型38.02%（修改提交数后实际为**38.52%**）
  - ✓ p1+p2数据：**38.66%**
  - ✓ 第三版特征：**38.72%**
  - ✓ 模型融合：**38.76%**



总结



- 源于对业务的深入理解和数据的细致分析
- 模型融合还有提升空间

总结

- 限制资源的合理性—性价比
- 10点评测对上班狗的好处
- 节约资源的方法：

- 特征工程环节：

  数据预处理（字符串编码等），节约40%左右资源

- 离线评测环节：

  采用资源消耗少的模型、小参数（使用相对指标进行对比）、数据采样

- *三思而后跑job*
- ……

思考

公交数据不仅仅是交通数据，背后隐藏着用户的需求

- 广州站/广州西站：出广州
- 东方金融大厦：上班
- 富力广场：消费
- ……

- 不同人群
- 出行频次/时间
- ……

御膳房建议

- 资源消耗计算的**两点需求：准确+实时**
- sql支持**set功能**：设置mapper、reducer等
- 支持**python UDF**等功能

致谢

以众包的形式，挖掘数据的价值，

解决企业、政府的业务问题，为用户提供更好的服务

- 感谢广东省人民政府、阿里巴巴集团举办如此精彩的大数据竞赛！
- 感谢岭南通公司及天池团队对竞赛的精心组织！
- 感谢一起参赛的小伙伴们！

新浪微博：
kevin_ning_thu