CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

系统认证风险预测

保住绿码团队

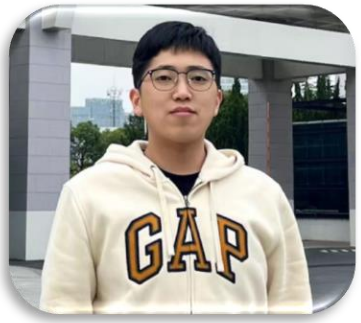
队长：计炜梁
队员：杜嘉伟，李雨晨，
吕思索，谢玲泽

目录

团队介绍

计炜梁
中国工程物理研究院

杜嘉伟
上海工程技术大学



李雨晨
合肥工业大学

吕思索
哈尔滨工业大学（深圳）

谢玲泽
南昌航空大学

AB双榜均为第一，大幅领先第二

A 榜　B 榜

我的成绩

到目前为止，您的最好成绩为 0.53765515 分，第 1 名，在本阶段中，您已超越 370 支队伍。

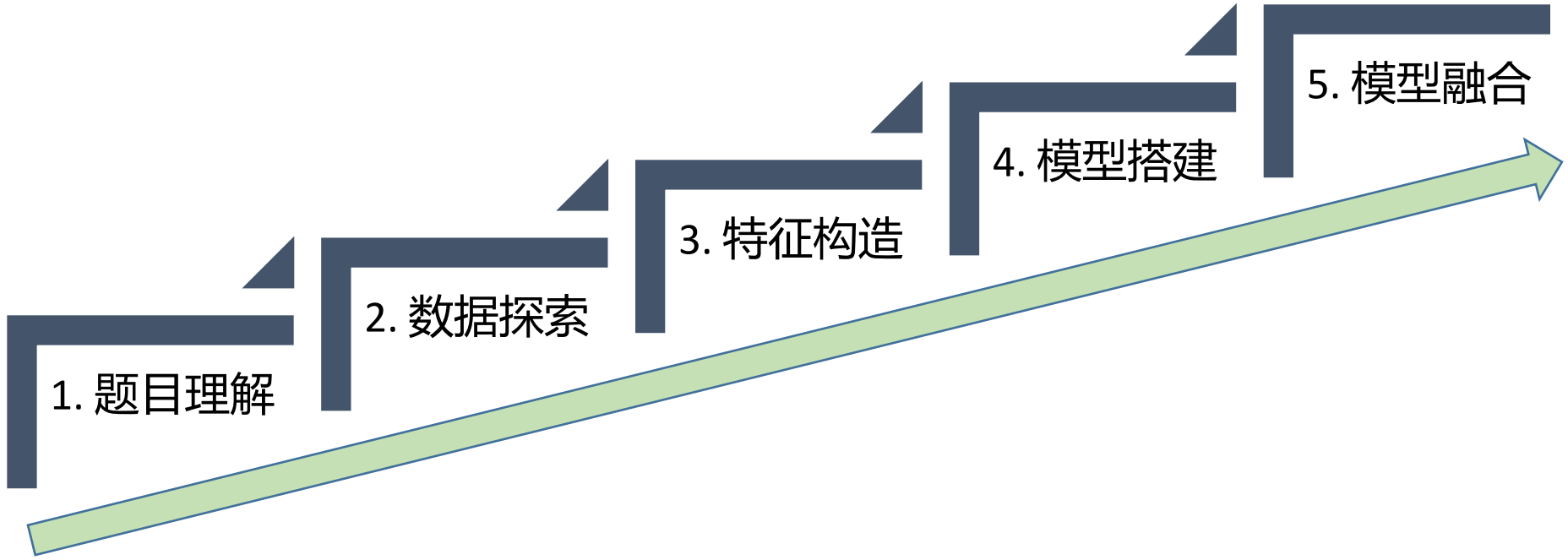
排名	排名变化	队伍名称	有效提交次数	最高分提交时间	最高得分
1	-	保住绿码	1	2021-11-21 23:58	0.53765515
2	-	篮球喜剧人	0	2021-11-21 23:58	0.53224693
3	-	小小笑笑	0	2021-11-21 23:58	0.53192022

团队曾获奖项

- 2021 ATEC科技精英赛科技新星榜冠军
- 2021 讯飞开发者大赛环境质量赛道亚军
- 2021 招商银行FinTech精英训练营冠军
- 2020 链想家计算科技大赛冠军
- 2020 讯飞开发者大赛季军
- 2020 CCF BDCI负载预测赛道三等奖
- 。。。

目录

算法方案解析



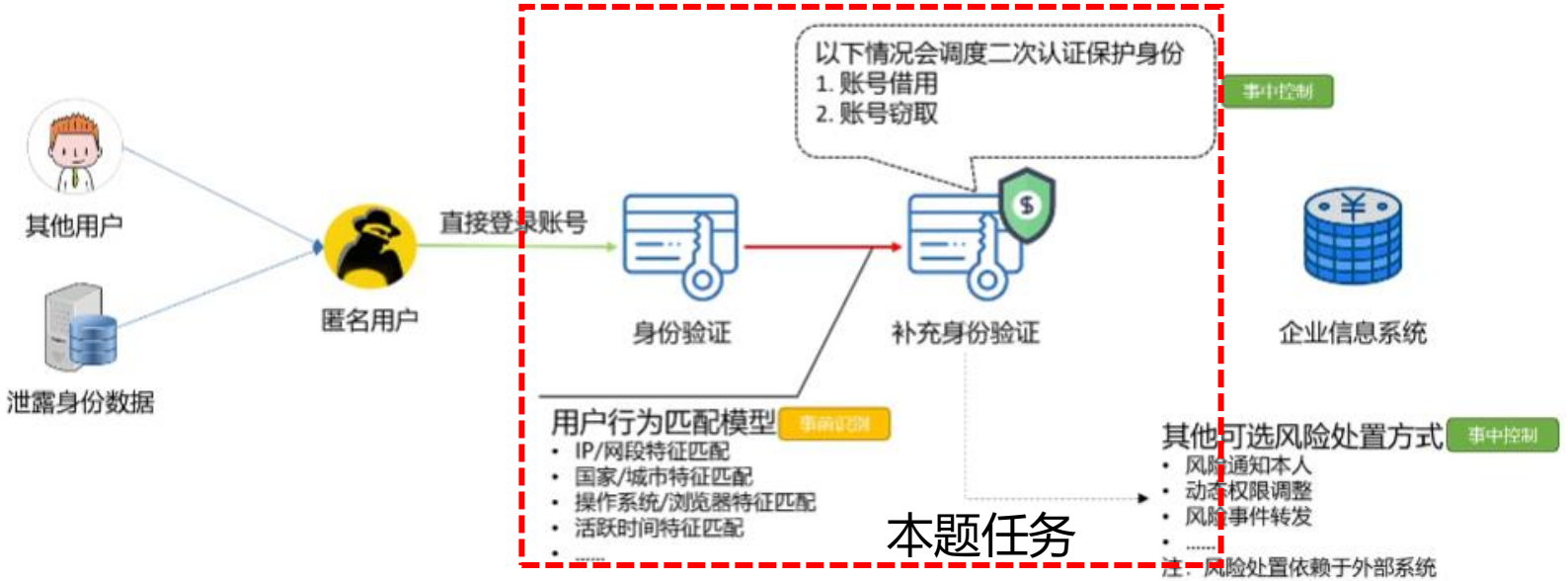
1. 题目理解

基于**用户认证行为数据**及风险异常标记结构，构建用户认证行为特征模型和风险异常评估模型，利用风险评估模型去判断当前用户认证行为**是否存在风险**。

问题类型：二分类

评价指标：AUC $AUC = \frac{\sum_{i \in \text{positiveClass}} \text{rank}_i - \frac{M(1+M)}{2}}{M \times N}$

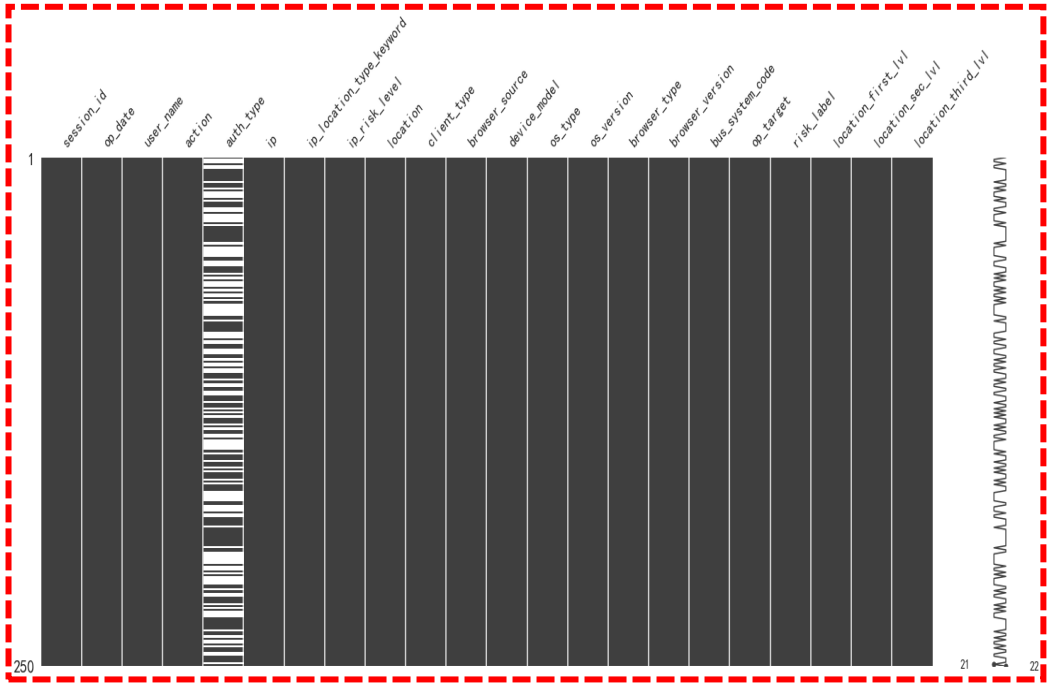
问题特性：异常检测，存在时序关系



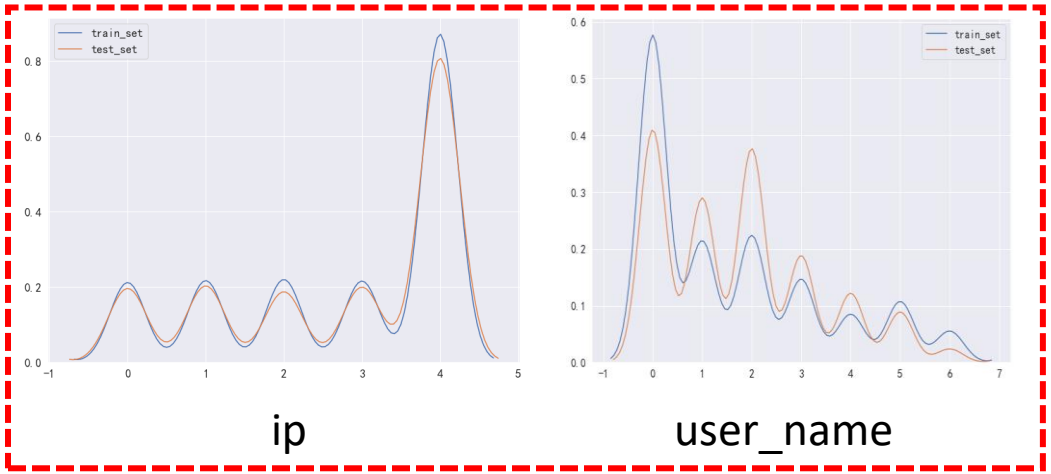
算法方案解析

2. 数据探索

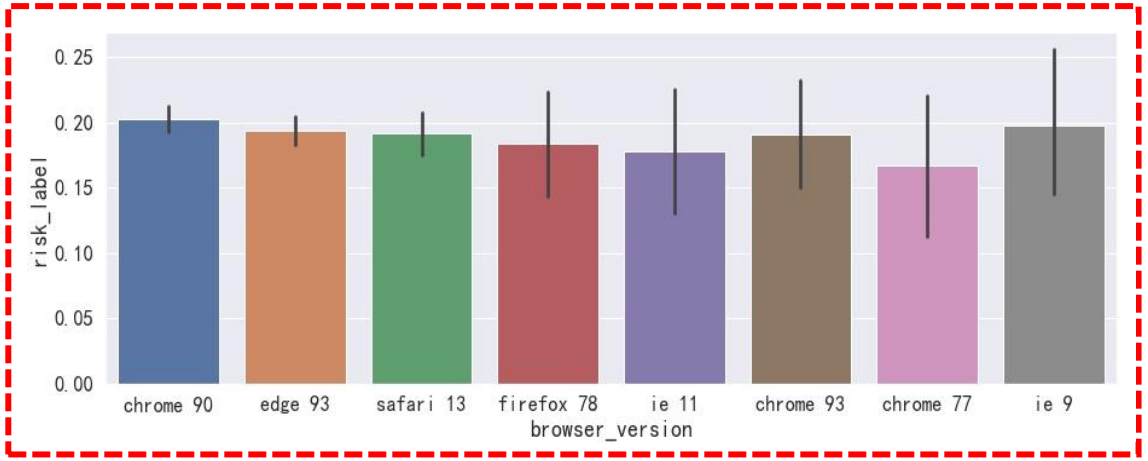
空值分析



分布一致性分析



特征类别与标签列关联性分析



3. 特征构造

- 原始特征
- 时间特征
- 词嵌入特征
- 统计交叉特征

原始特征

在删掉单值特征后，对于给出的原始特征，类别特征占比较大，可以直接使用**标签编码（LabelEncoder）**将其转换为数值特征。

其中地址可以分级提取3级地址，分别再编码。

地点	location	{"first_lvl": "成都分公司", "sec_lvl": "9楼", "third_lvl": "销售部"}; {"first_lvl": "江苏省", "sec_lvl": "苏州市", "third_lvl": "常熟市"}	{"first_lvl": "成都分公司", "sec_lvl": "9楼", "third_lvl": "销售部"} ; {"first_lvl": "江苏省", "sec_lvl": "苏州市", "third_lvl": "常熟市"}

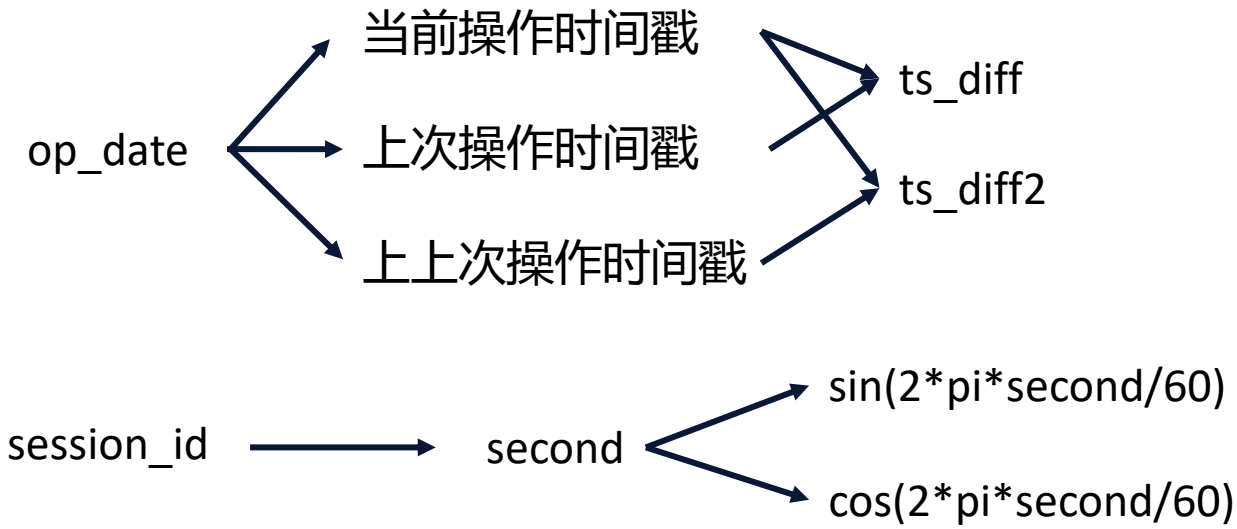
Location = location_first_lvl + location_sec_lvl + location_third_lvl

3. 特征构造-时间特征

1. 利用op_date字段可获取**操作时间戳**，也可以提取操作年、月、日、小时、分钟，但会导致过拟合。

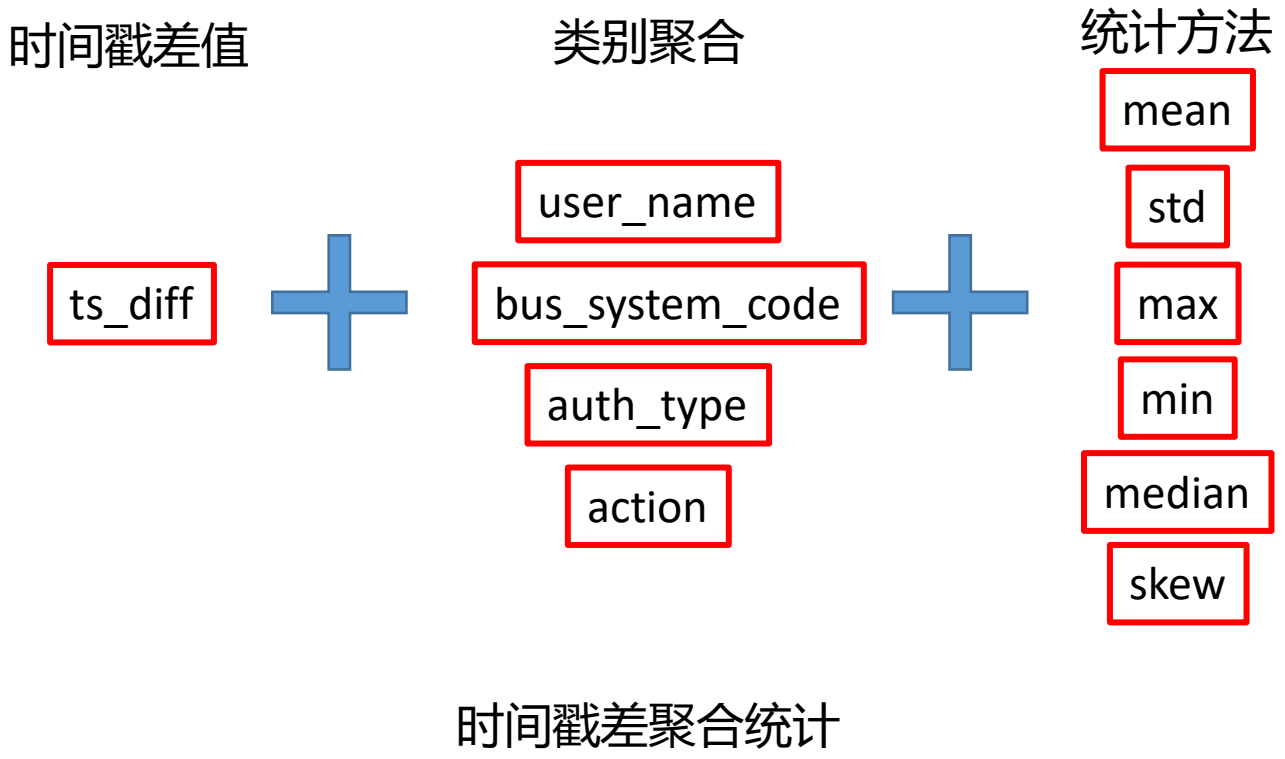
2. 从session_id字段中提取当前**操作秒**，可明显提升模型性能。

3. 利用秒变化的频率进行**正余弦变换**，可以构造出新的时间特征以增强鲁棒性。



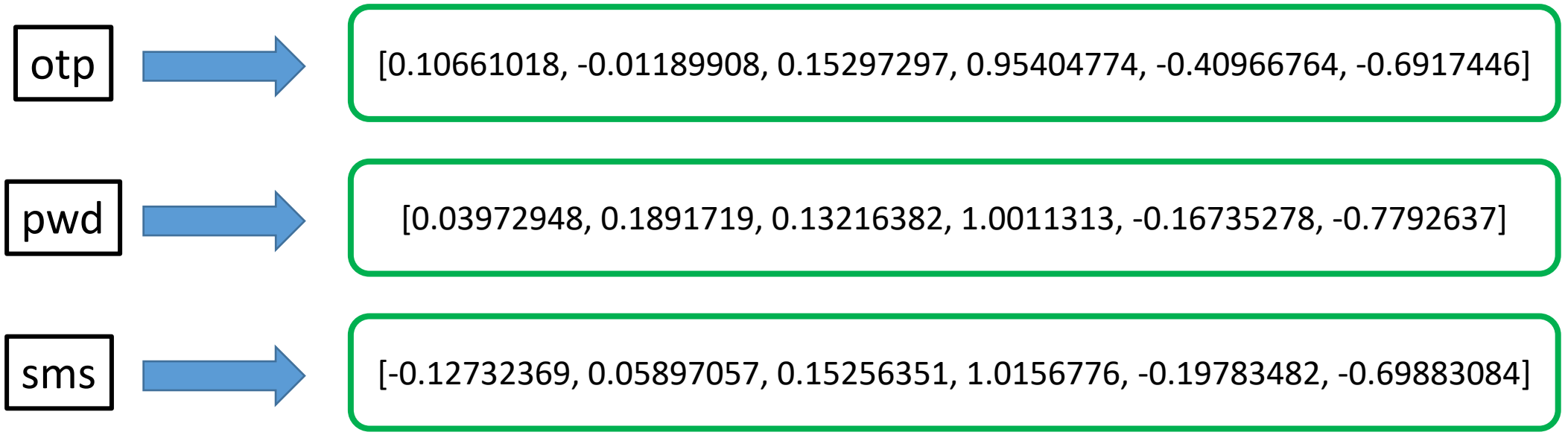
3. 特征构造-统计交叉特征

1. 类别值数量统计
   - 对用户聚合后统计各类别特征值的数量
2. 时间戳差聚合统计
   - 根据类别特征聚合后对时间戳差值进行统计
3. 类别/数值特征交叉
   - 对类别特征进行聚合后对各数值统计，并将统计值与原值做交叉特征



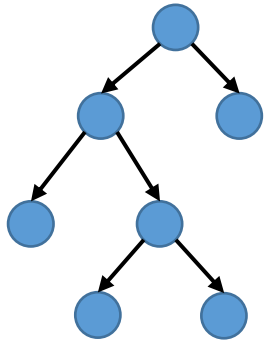
3. 特征构造-词嵌入特征

由于auth_type特征重要度很高，同时同一用户在一天内的行为存在**序列关系**，考虑对每个用户每天的auth_type序列构造句子，并利用Word2Vec进行词嵌入，将每种auth_type表示为一个稠密向量。词向量为6维，滑动窗口大小为12。

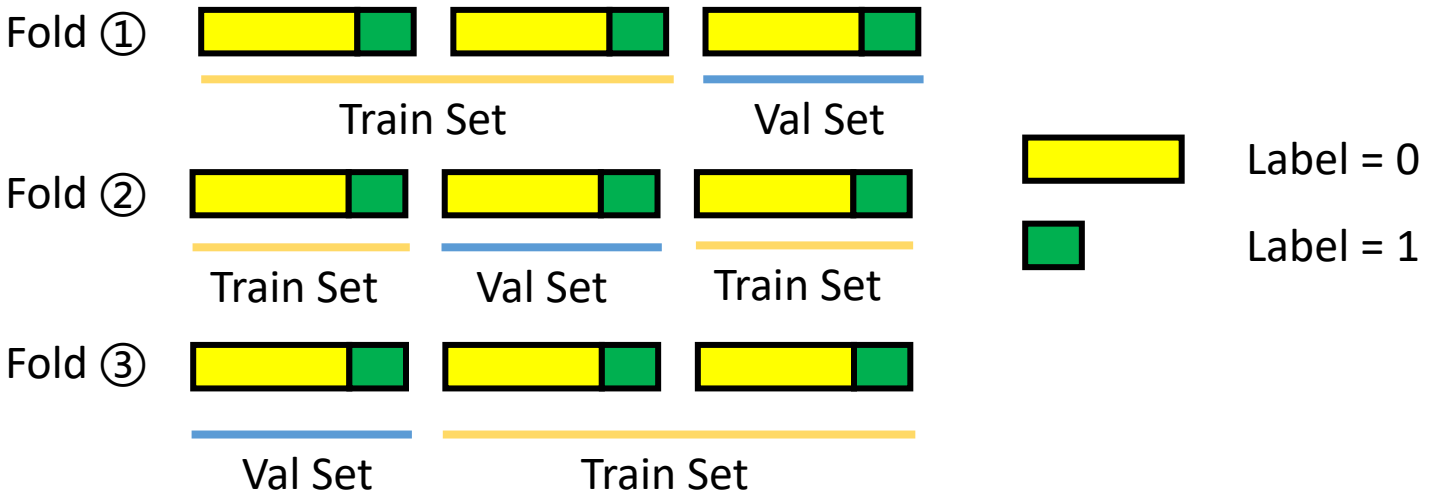


4. 模型搭建

- 基于LightGBM进行模型搭建
- 考虑训练集正负样本比例，采用分层交叉验证划分数据集



LightGBM结构示意图



分层交叉验证

5. 模型融合

由于评分指标为auc，**预测结果的顺序**十分重要，因而考虑进行排序融合

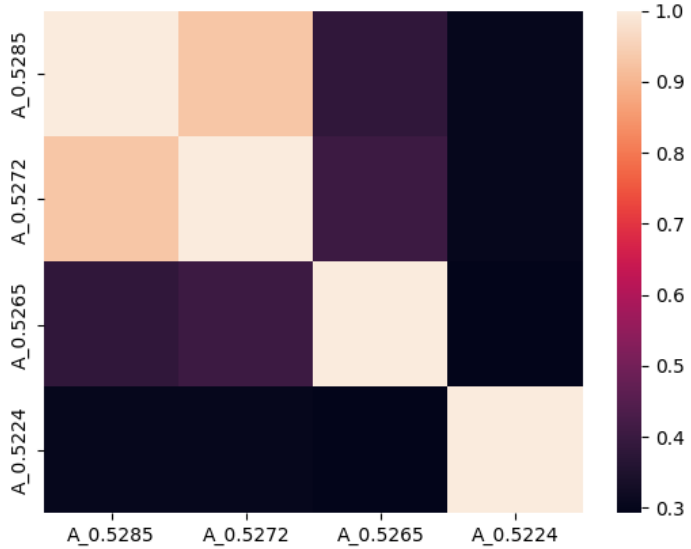
1. 对每个预测结果进行排序，并求加权平均

$$y_0 = \sum_i (\mathrm{rank}(y_i) * w_i)$$

2. 对加权平均结果进行归一化

$$y = \frac{y_0 - \min y_0}{\max y_0 - \min y_0}$$

此外，由于模型预测不稳定，因此融合时不同**模型的差异性**也十分重要



4份结果排序Pearson相似度

A榜单模及融合成绩

6. 上分点及亮点小结

1. 通过挖掘重要**时间特征**，显著提高模型预测性能

2. 利用**统计交叉**和借鉴推荐中常用的对类别特征进行**词嵌入**进一步提升模型性能

3. 通过**分层多折交叉验证**提高模型的泛化性

4. 采用对**排序后融合归一化**及**差异性融合**的方式，增加模型互补能力

目录

后续优化思路

➢ 数据与特征方面

1. **数据预处理**还需要改进，考虑空值的合理填充。
2. 对数据的分析还较为简单，缺少对**用户行为模式**更细致的挖掘。
3. 更多的**关注业务**，考虑引入规则特征

➢ 模型方面

1. 尝试更多类型的模型，目前对**神经网络**尝试较少，可以考虑利用LSTM或CNN提取前后时序关系，考虑进行类别特征的嵌入。
2. 进行更细致的**调参**，后期发现调参提升比预期要大，但机会不够了。

目录

总结

- 模型预测性能波动较大，不同特征在不同模型参数下的表现不同，需要多进行尝试
- 进行模型融合时，模型的差异性十分重要
- 感谢竞赛主办方提供的平台与机会
- 感谢竹云提供的宝贵数据集
- 感谢队友间的相互鼓励与合作

Thanks

Q&A