腾讯2018广告算法大赛
——找寻广告种子用户

小组成员：甘红楠、陶文慧 、元奕超 、刘俊涛 、夏天宇
研究单位：复旦大学软件学院

介绍

基于社交关系的广告（即社交广告）已成为互联网广告行业中发展最为迅速的广告种类之一。如何在复杂的社交场景，多样的广告形态，以及庞大的用户数据等诸多因素干扰下，提供精准高效的广告解决方案成了业界亟待解决的问题。

腾讯社交广告业务推出一种真实的广告产品——相似人群拓展（Lookalike）。该产品的目的是基于广告主提供的目标人群，从海量的人群中找出和目标人群相似的其他人群。

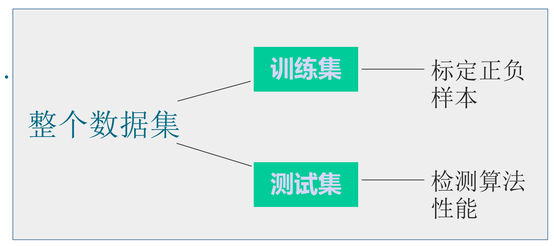
目标 基于广告主提供的一个种子人群（又称为种子包），自动计算出与之相似的人群（称为扩展人群）

数据集

几百个种子人群、海量候选人群对应的用户特征，以及种子人群对应的广告特征（出于业务数据安全保证的考虑，所有数据均为脱敏处理后的数据）。

原数据集共有4个文件，分别为：

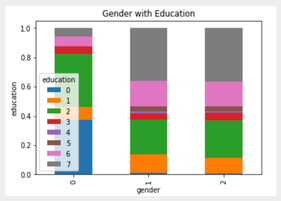
- train.csv：训练集数据文件
- test1.csv：测试集数据文件
- userFeature.data：用户特征文件
- adFeature.csv：广告特征文件



实验和结果

预处理

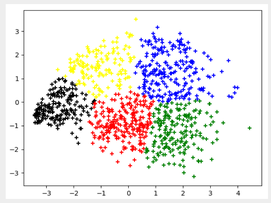
- 预处理：userFeature原始数据存在一些缺失值，需要根据具体的属性情况结合可视化分析进行补全。



性别与受教育程度

预处理
- 可视化分析
- 性别1、2的样本受教育程度类似
- 性别为0的样本教育程度主要集中在低层次阶段或者教育程度未知
- 可能原因、系统性别字段未填写则为0，这部分用户也非活跃用户

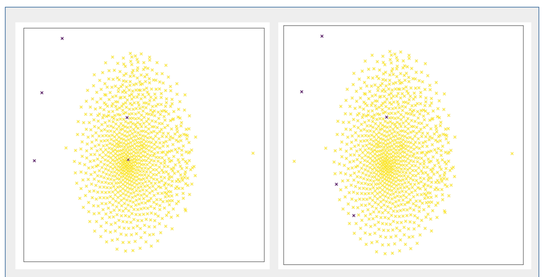
- 聚类分析：通过k-means算法对用户样本进行聚类分析，形成可视化结果



K=5 时，用户样本聚类效果

聚类分析：
- 核心指标：SSE
- 手肘法确定最优k值
- 原始用户数据并无明显类别区分，聚类结果类别间隔较小

- 关联分析：通过FP-Growth算法，计算userFeature数据中Interest1~5属性之间的置信度，从而分析interest之间的关联关系。
- 异常点检测：通过TSNE算法对userFeature降维，然后分别用Isolation Forest和OneClassSVM进行异常点检测分析，并形成可视化结果

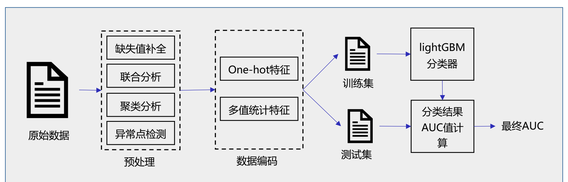


OneClassSVM 异常点检测

模型

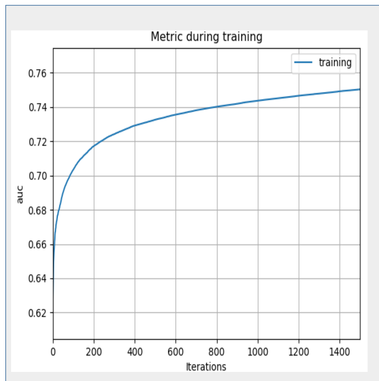
数据编码：
- one-hot编码
  - 针对user_feature 中单个取值的离散特征
  - 针对ad_feature 中单个取值的离散特征
- 多值统计编码
  - 采用类似NLP处理中词向量统计的方法
  - user_feature 中的多取值的离散特征

稀疏处理：
- 编码得到的特征矩阵，包含较多的零值，为稀疏矩阵
- 直接处理稀疏矩阵会有较高的空间复杂度与时间复杂度
- 特征矩阵压缩存储、稀疏化处理



模型框架图

初始模型结果：

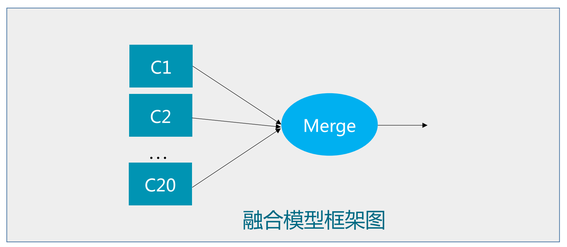


Baseline结果：
- Metrics：AUC
- Baseline模型最终得到的AUC平均值为0.733
- 直接利用原始特征，模型构建简单、方便

训练过程AUC

模型+

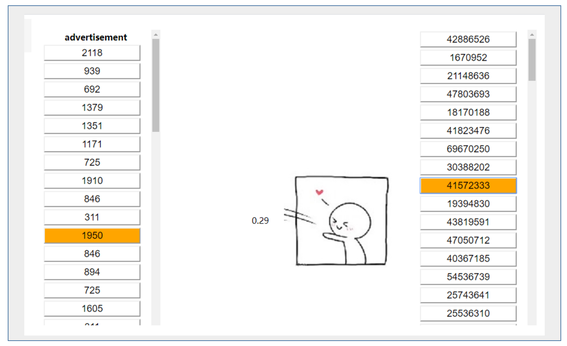
- 特征构建：基于原始数据构建投放量、投放比例、转化率、特殊转化率、多值长度与比例等特征
- 模型融合：将特征划分为5部分、数据划分为四部分，构建20个模型，并将20个模型结果融合
- 试验结果：模型融和结果为0.7559，较baseline提升2个百分点
- 相比于单个模型的最好结果0.7490，提升0.7个百分点



融合模型框架图

可视化展示

计算以后的结果会存入数据库，因为系统是定时计算更新的，所以每次需要进行广告推广的时候，推广人员需要上传相应的推广计划，然后系统会在数据库中调出相应的目标人群，并且可以看到用户与广告的匹配程度。



结论

- 数据预处理不仅仅是简单的异常检测、缺失补全，更可以加深对数据本身的理解
- 特征工程对于点击率、推荐等商业问题模型构建至关重要
- 模型之间的融合可以进一步提升效果