招商银行信用卡中心
金融数据大赛答辩

队伍： X-Driver

名次： A榜冠军 / B榜冠军

成员： 陈瑀　晏世凯　孙万彤

团队曾获荣誉

- **2018搜狐内容识别算法大赛 第7名**

  [CCF、搜狐联合举办，参赛队伍612支]

- **2017中国高校计算机大赛大数据挑战赛 第3名**

  [教育部教指委、清华大学、腾讯联合举办，参赛队伍1222支]

- **IJCAI-17 Customer Flow Forecasts on Koubei.com 第9名**

  [IJCAI、阿里巴巴联合举办，参赛队伍4046支]

- **2016 CCF大数据与计算智能大赛O2O优惠券使用预测 第9名**

  [CCF、阿里巴巴联合举办，参赛队伍1505支]

CONTENTS

赛题分析

数据：

- 掌上生活APP的一个月的操作行为
- 个人属性与信用卡消费数据

目标：

- ✓ 预测未来一周用户是否购买优惠券

分析：

- 单个用户重复一个行为
- 多个用户具有相同的行为

特征工程（常规特征）

主要按照特征群进行提取：基础统计特征，离散特征，时序相关特征

基础统计特征

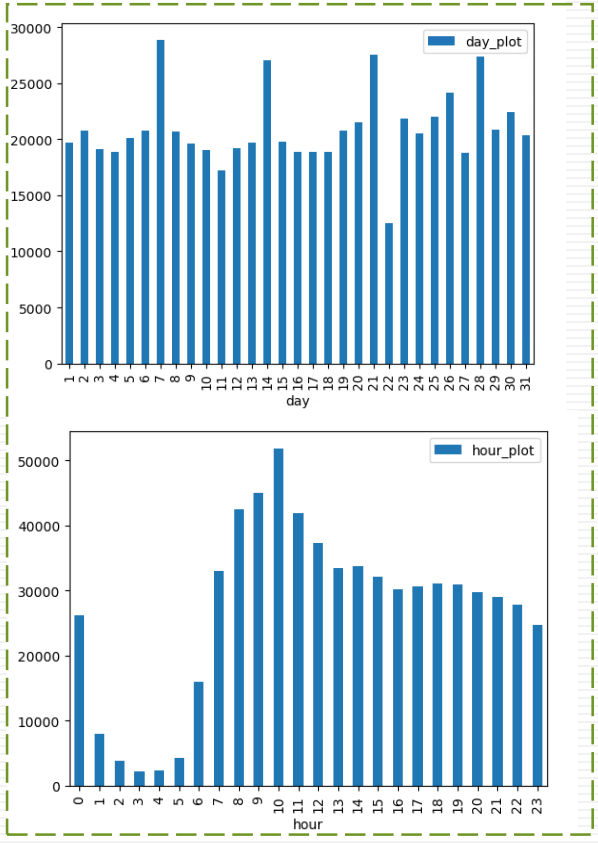
- 用户总的点击次数
- 用户有多少天点击
- 用户每天平均点击量
- 用户各种行为类型次数对总次数的占比
- ……

离散特征

- 用户对于第三级模块分别离散
- 用户在各个小时的点击量
- 用户在各个周几的点击量
- 用户对于各个事件类型的次数
- …….

时序相关特征

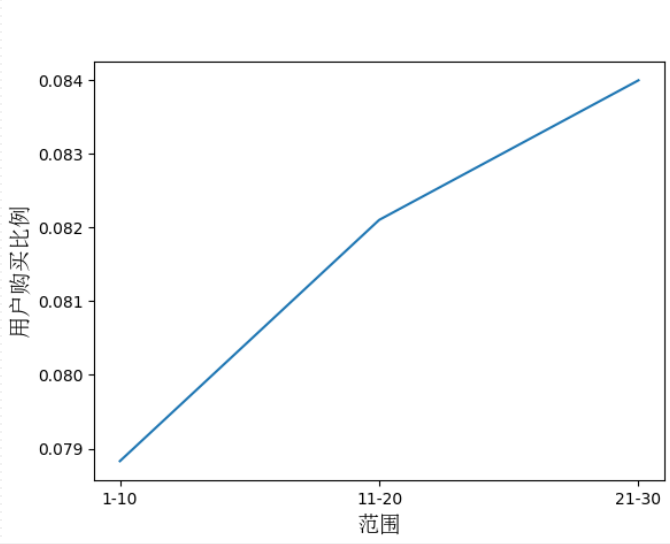
- 用户点击的时间间隔
- 用户最大连续点击天数
- 用户最后一次点击距离最后一天的间隔
- 对于最后一天的统计
- ……



特征工程（亮点一：时序特性）

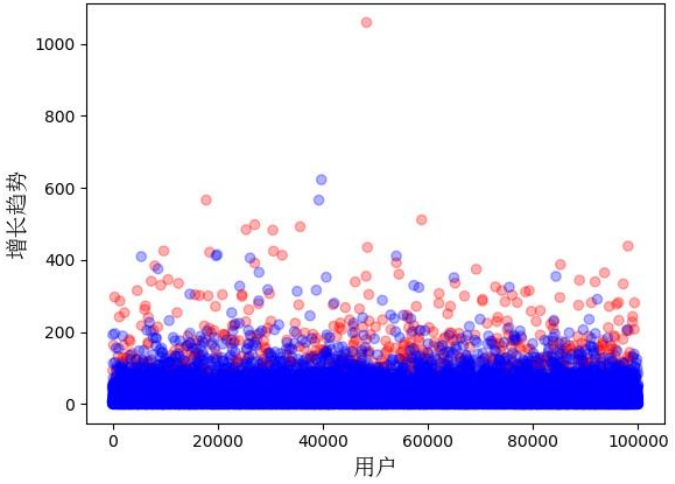
时序特性:

特征权重随时间衰减



不同时间范围内用户购买比例图

用差值反应购买趋势



购买（红）与非购买（蓝）用户的行为增长趋势分布图

特征工程（亮点二：NLP特征）

基本思想：将每个模块看成一个词，一个用户的所有操作就成了一篇文档。

USRID_2： word_1 word_2 word_3 word_4 word_5 word_5 word_5 word_6 word_4

NLP特征：

- Bag of words 特征（未能考虑行为的顺序）
- Word2vec 均值向量特征（行为顺序信息）

可解释性：

- 用户的行为具有一定规律性
- 文字的表达具有一定规律性

USRID	OCC_TIM	EVT_LBL	转换为词语
2	2018/3/13 23:26	38-115-117	word_1
2	2018/3/13 23:26	520-1836-3640	word_2
2	2018/3/13 23:26	0-231-277	word_3
2	2018/3/13 23:30	359-1234-2004	word_4
2	2018/3/13 23:30	326-1041-1678	word_5
2	2018/3/13 23:31	326-1041-1678	word_5
2	2018/3/13 23:33	326-1041-1678	word_5
2	2018/3/13 23:33	359-1233-2003	word_6
2	2018/3/14 19:25	359-1234-2004	word_4

特征工程

特征选择:

- 基于XGB的特征重要性

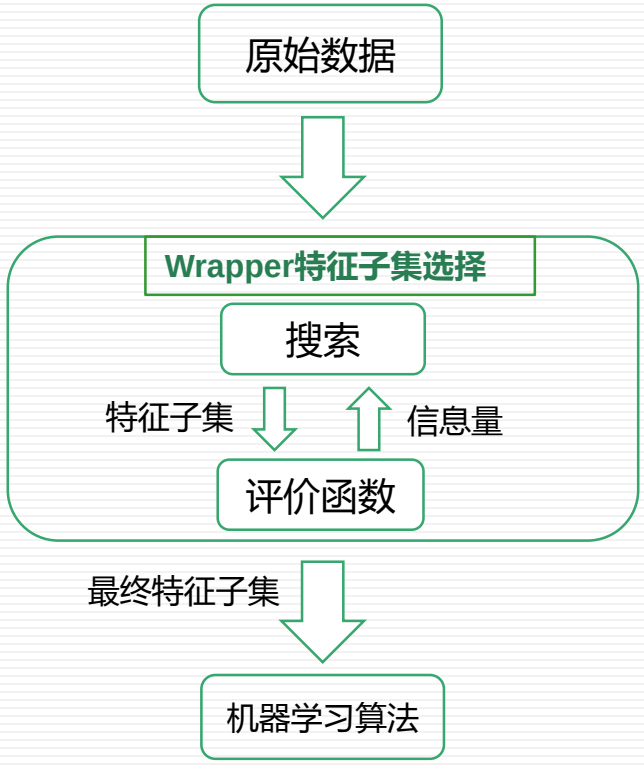
  先训练一个XGBoost模型，输出其特征重要性，然后将重要性为0的特征删除，即完成了特征选择。

- 基于wrapper的方式

  基于贪心算法，寻找最优特征子集，如右图所示。

目的:

- 降维，使模型泛化能力更强，减少过拟合。



模型设计

XGBOOST:

参数	值
booster	gbtree
max_depth	3
eta	0.03
objective	rank:pairwise
min_child_weight	6
scale_pos_weight	3176 / 76824

- 针对指标AUC,使用了rank:pairwise
- 样本不平衡的处理，设置scale_pos_weight为正负样本比例

LIGHTGBM:

参数	值
boosting_type	gbtree
num_leaves	31
learning_rate	0.01
objective	binary
feature_fraction	0.8
bagging_fraction	0.8

- 可以加速训练，节省内存
- 对于类别特征的决策规则

模型融合（基于Rank）

USRID	RST
1391	0.8613208
3679	0.4235010
4176	0.5986956

USRID	RANK
1391	1
3679	3
4176	2

USRID	RST
1391	0.813208
3679	0.9035010
4176	0.6386956

USRID	RANK
1391	2
3679	1
4176	3

$$\sum \text{weight}_i / rank\_i$$

USRID	RANK
1391	1
3679	2
4176	3

比赛评价标准：AUC，其本质为排序优化问题
概率得分线性加权存在问题

方案潜力&应用价值

潜力：

- 特征具有可解释性
- 模型的验证方法具有稳定性
- 随着数据量的增大，Word2vec特征会有更好的效果。
- **自然语言处理领域中成熟的方法可应用到用户行为识别中。**

价值：

- 商家广告精准投放。
- 个性化推荐优惠信息。

参赛总结

不足：

- 对于agg表只进行了简单的二元化和rank处理，没有深度发掘。
- 模型的融合的权重是根据线上成绩，其实可以采用线性回归的方式得到。

收获：

- NLP和普通机器学习的结合
- 坚持的重要性
- 团队协作能力都得到了较大的提升

感谢招商银行信用卡中心给予的宝贵机会

感谢Datafountain平台的工作人员

感 谢 观 看　　请 您 提 问