马上消费金融
www.msxf.com
银监会批准持牌金融机构

《违约用户风险预测》

- 学校　　：　重庆邮电大学
- 团队名称　：　我只吃两个剩下的都给你

让生活更轻松

目录概要

- 1、团队简介
- 2、赛题分析
- 3、数据划分
- 4、特征工程
- 5、模型融合
- 6、总结展望

团队简介

团队数据摇竞赛历史荣誉

竞赛：2016 CCF大数据与计算智能大赛：O2O优惠券使用预测
主办单位：*蚂蚁金服集团，中国计算机协会(CCF)*
排名：9 / 1501

竞赛：IJCAI-17 口碑商家客流量预测
主办单位：*蚂蚁金融服务集团，IJCAI2017*
排名：9 / 4046

竞赛：2017年中国高校计算机大赛—大数据挑战赛
主办单位：*清华大学，深圳市腾讯计算机系统有限公司*
排名：3 / 1222

竞赛：KDD CUP 2017 Traffic Volume Prediction
主办单位：*Alibaba Cloud, AMAP*
排名：13 / 3582

竞赛：IJCAI-18 阿里妈妈搜索广告转化预测第一赛季
主办单位：*阿里妈妈，IJCAI2018*
排名：8 / 4046

赛题分析

- 目前国内违约用户风险预测采取人工审批作业形式，效率低而又面临很大的违约风险，无法进行风险分级管理，影响风险控制的能力及灵活度。

- 本方案使用大数据和人工智能建立违约用户风险预测机器学习模型，对目标客户的基本信息，信用历史记录等特征进行分析，直接预测用户的违约概率，为企业提供稳定可靠的解决方案。

赛题分析

1 银行卡信息

申请贷款唯一编号
银行名称
银行卡号后四位
银行卡类型
银行卡绑定手机号(脱敏)

2 收货地址信息

申请贷款唯一编号
收货人姓名(MD5加密)
收货地址ID
收货地址所在地区
收货人手机号(脱敏)
收货人固定电话号码(脱敏)

3 网购平台信用信息

申请贷款唯一编号
网购平台信用评分
网购平台信用额度
网购平台信用额度使用值

4 个人基本信息

申请贷款唯一编号
性别
出生日期
兴趣爱好
婚姻状况
收入水平
身份证号(脱敏)
学历
所在行业
是否绑定QQ
是否绑定微信
会员级别

5 认证信息

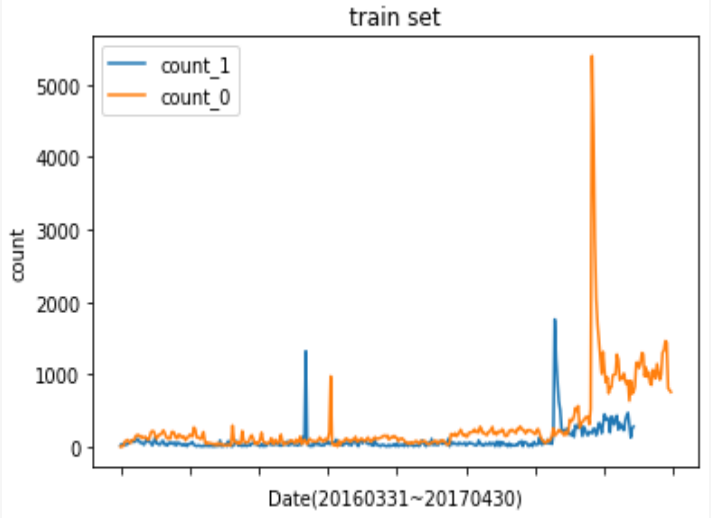
申请贷款唯一编号
身份证号（脱敏）
认证时间
认证电话号码（脱敏）

6 订单信息

申请贷款唯一编号
订单编号MD5加密
收货人姓名MD5加密
订单金额
支付方式
下单时间
订单状态
收货电话（脱敏）
商品编号MD5加密
商品单价

赛题分析—时序关系

是否违约与时序关系



逾期和未逾期趋势有规律性

历史数据中用户逾期值呈小于未逾期的趋势

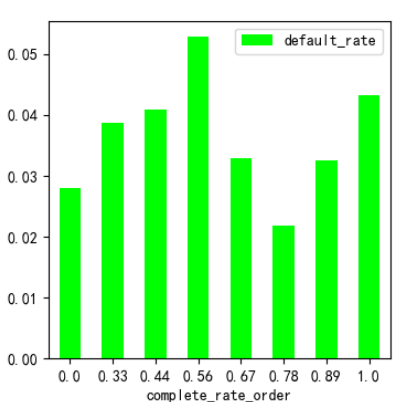
用户数据有波动情况，但大多数稳定中小幅度波动

2017年相比2016年趋势有明显不一致

离预测月越近的历史数据越有代表性

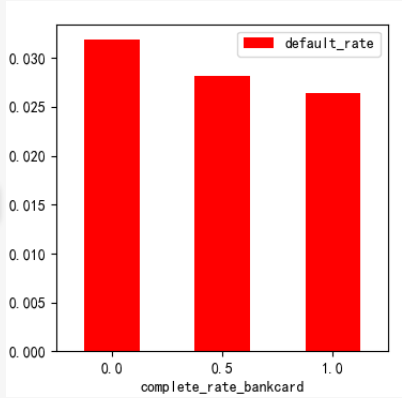
2016年存在异常突出的数据

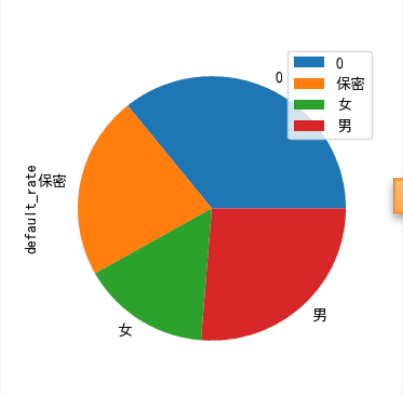
赛题分析—多角度可视化



Order表中用户信息完整度和违约率的关系：完整度为50%的违约率较高。

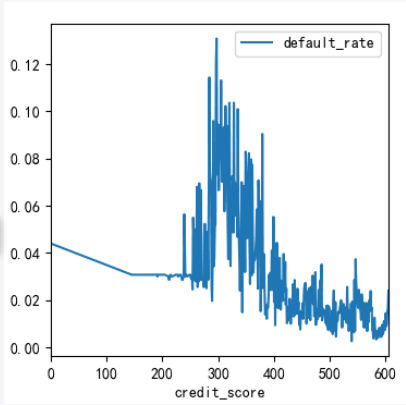
Bankcard表中用户信息完整度和违约率的关系：信息完整度越高的用户违约率越低。



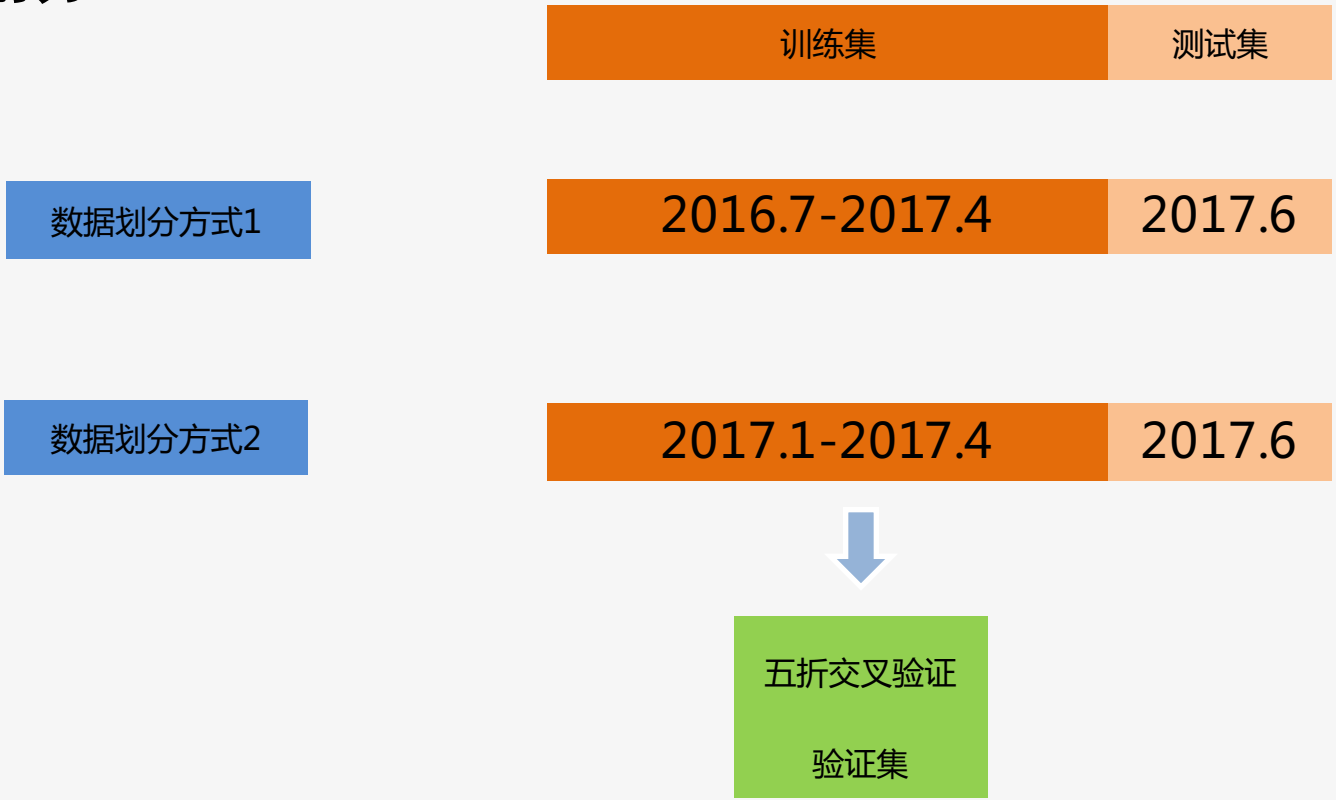


举例分析性别与违约率关系：未知性别和保密性别违约率是最高的，女性逾期还款可能性较低。

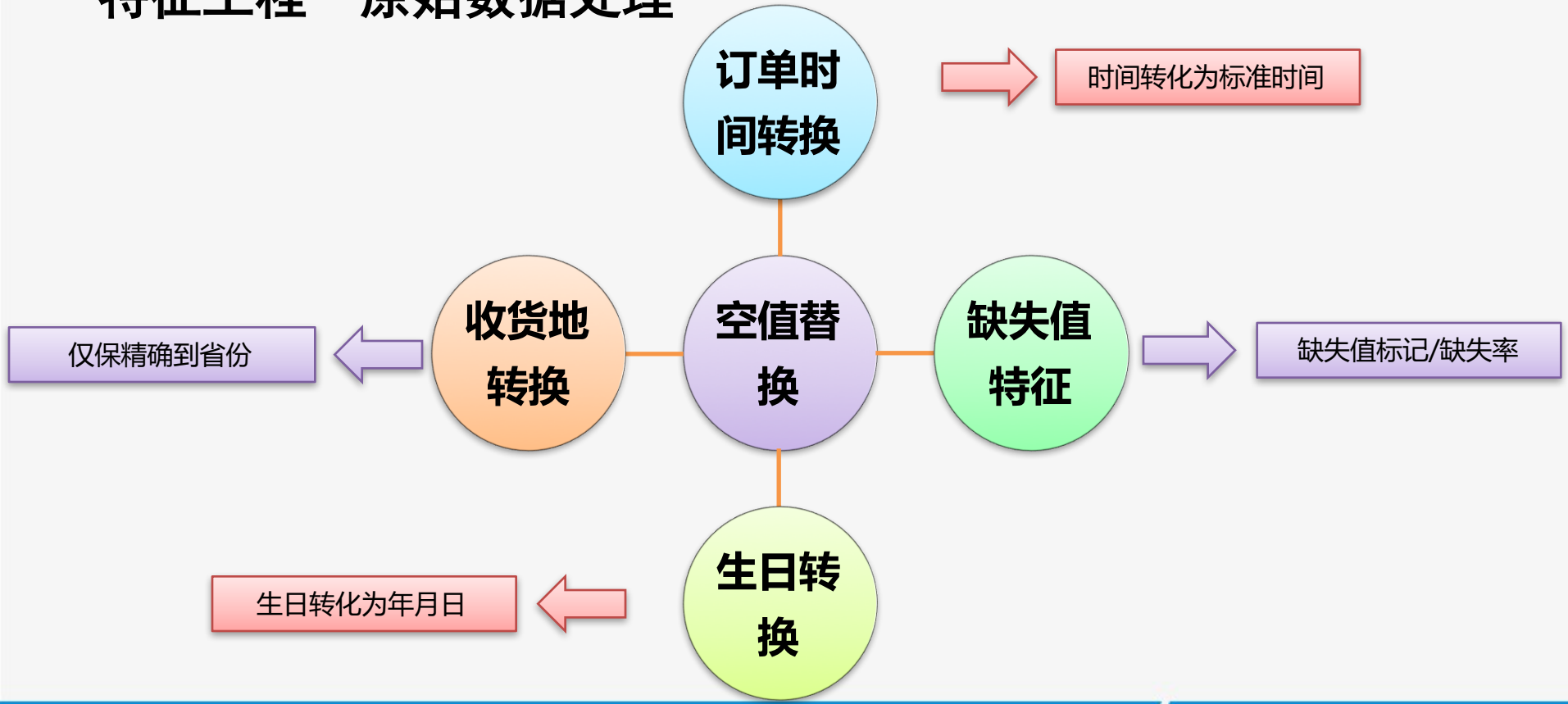
用户网购平台信用评分与违约率的关系：平台信用评分在200-400之间的违约率最高，并非信用评分越低就表征用户越可能逾期还款。



数据划分


训练集
测试集
数据划分方式1
2016.7-2017.4
2017.6
数据划分方式2
2017.1-2017.4
2017.6
五折交叉验证
验证集

特征工程—原始数据处理


订单时间转换
时间转化为标准时间
收货地转换
仅保精确到省份
空值替换
缺失值特征
缺失值标记/缺失率
生日转换
生日转化为年月日

特征工程—XGB构造新特征

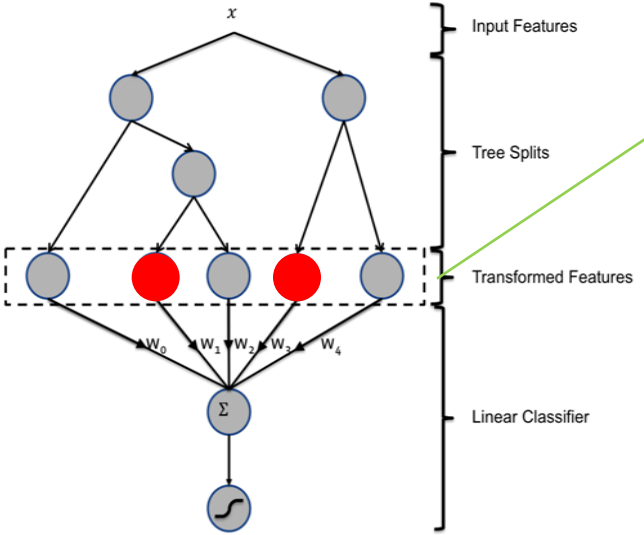


Figure 1: Hybrid model structure. Input features are transformed by means of boosted decision trees. The output of each individual tree is treated as a categorical input feature to a sparse linear classifier. Boosted decision trees prove to be very powerful feature transforms.

新特征向量：[0， 1， 0， 1， 0]

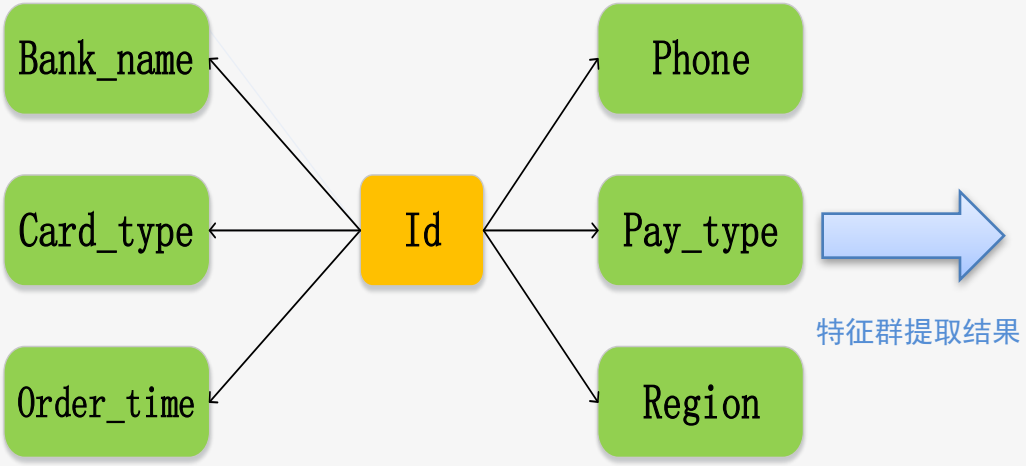
原始特征离散化　　新构造出的特征

Logistic Regression

新模型
（与原模型有较大的差异性）

[参考文献] He X, Pan J, Jin O, et al. Practical Lessons from Predicting Clicks on Ads at Facebook[J]. 2014(12):1-9.

特征工程—特征群提取



- 不同银行卡预留电话（phone）数目
- 不同银行卡类型（card_type）数目
- amt_order 次数、type_pay 在线支付、type_pay 货到付款、sts_order 已完成次数
- 信用额度使用值是否超过信用额度（overdraft是否大于quota）
- 用户贷款提交时间（applsbm_time）和生日（birthday）之差
- 用户贷款提交时间（applsbm_time）和认证时间（auth_time）之差
- One-hot离散user表的sex、merriage、income、degree、qq_bound、wechat_bound、account_grade属性
- 非空的属性数目/总属性数目
- ...

特征工程—借贷时间

以用户的借贷时间为分界点

- 是否认证时间在借贷时间前
- 是否认证时间在借贷时间后
- 认证时间在借贷时间前多少天
- 认证时间在借贷时间后多少天
- 信誉排序

- 借贷时间前有多少次购买
- 借贷时间后有多少次购买
- 借贷时间前有多少次购买最大值
- 借贷时间后有多少次购买最小值
- 银行违约率

特征工程—特征选择

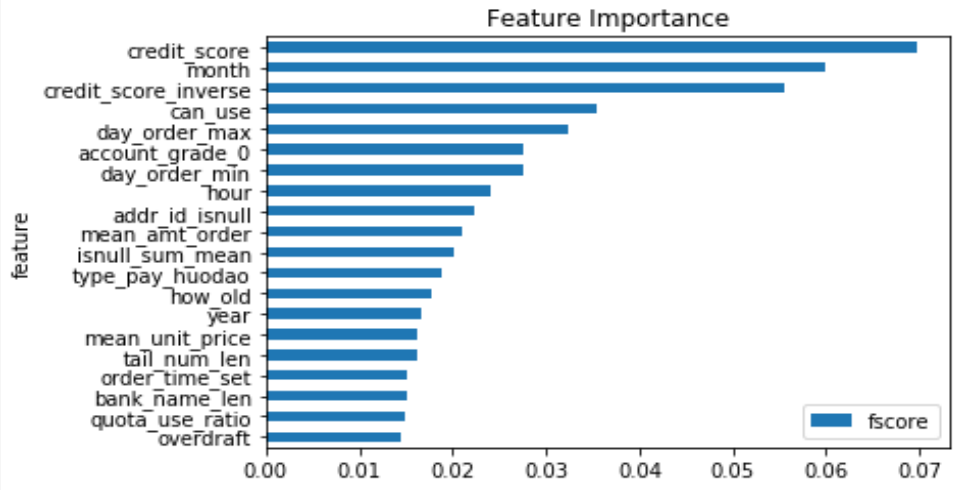
XGBOOST模型输出特征重要性，降序排序，选取top20，可视化如下：

网购平台信用评分

当前样本时间
-
订单最小时间

网购平台信用
评分反序

...



当前样本月份

当前样本时间
-
订单最大时间

订单金额的均值

...

高效的特性选择方案：基于嵌入式的特征排序方法

好处：模型学习的过程和特征选择的过程是同时进行的，基于xgboost 来做特征选择， xgboost 模型训练完成后可以输出特征的重要性，删除重要性趋近于0的特征。

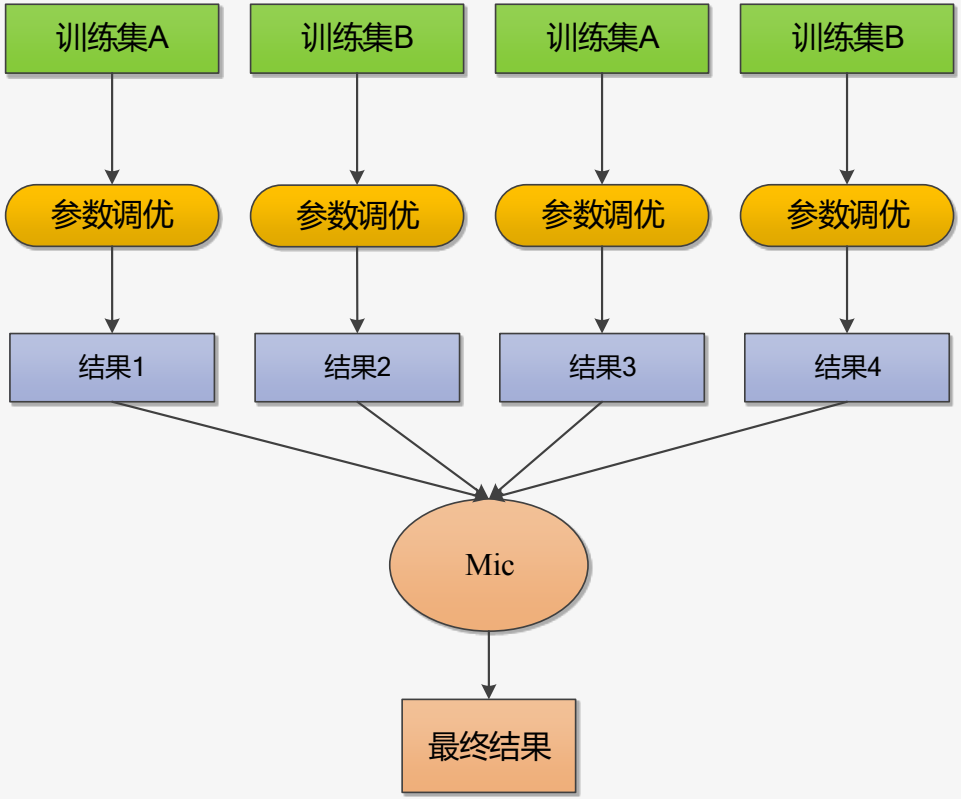
模型融合

模型异构

- 特征工程
- 模型选择

模型融合

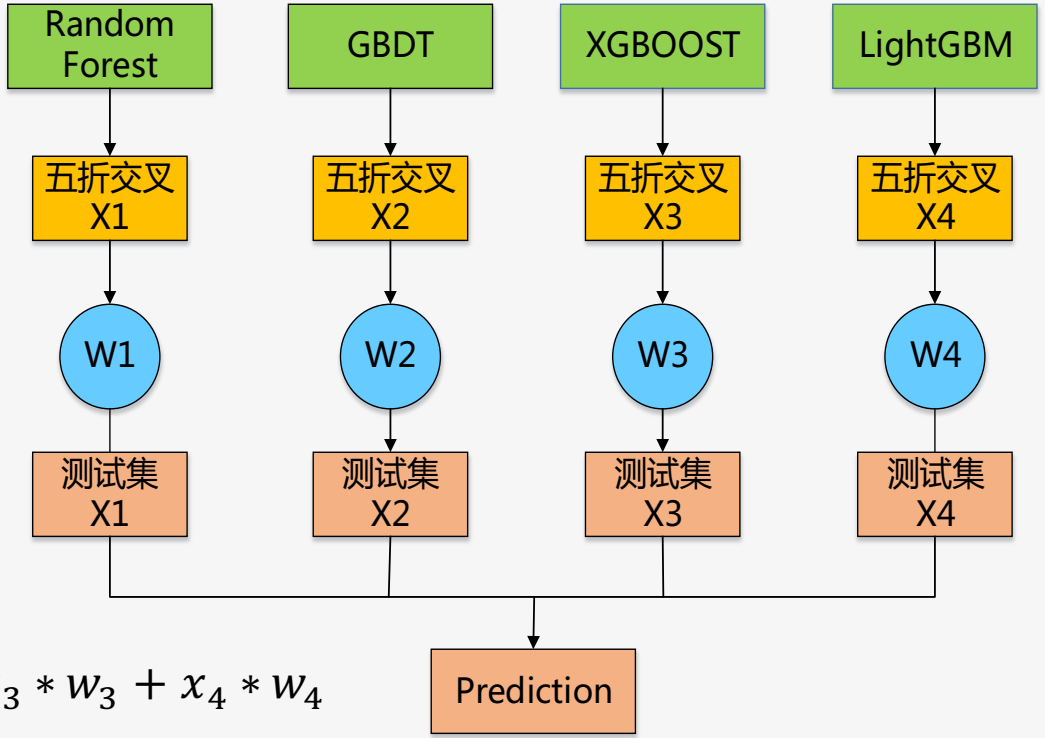
- 算法层面
- 结果层面



模型融合

random_state=None

Stacking模型融合流程



线性回归拟合验证集 $f(Xi) = x_i * w_i$

$$Y(\text{验证集}) = x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + x_4 * w_4$$

$$P(\text{测试集}) = X_1 * w_1 + X_2 * w_2 + X_3 * w_3 + X_4 * w_4$$

总结展望

- 问题分析一定要考虑业务层面。想到可能有用的稍微有一点业务含义的特征就添加，哪怕不太确定，或者觉得和已有特征关联较大!
- 数据的前期处理至关重要！
- 多个异构模型的融合能有效的提高结果精度！
- 有效的线下验证，是提高模型准确性的关键！
- 不到最后一刻坚决不会放弃！
- 初赛排名第2，复赛排名第1，只有稳定的模型才是最好的模型！