

# “智慧中国杯”全国大数据创新应用大赛

## 用户贷款风险预测

团队：SLaughter

# 1

## 问题分析

## ► 问题描述

数据来源：融360

数据内容：用户基本属性以及银行流水、浏览行为、信用卡账单等记录

训练数据集  
(55595)



测试数据集  
(13899)



根据用户历史信用消费行为及基本属性，预测用户是否会发生逾期行为

## ► 问题描述

**应用场景**  
互联网金融贷款逾期预测

**难点**  
数据缺失、验证集选择

**直接预测目标**  
用户是否会发生逾期行为

**本质：二分类问题(分类不平衡)**

**评价指标：KS、AUC**

# 2

## 算法框架

## ► 算法框架



# 3

## 数据探索

► 数据探索

	训练集用户 缺失比例	测试集用户缺 失比例	训练缺失/预测 缺失
用户基本信息	~0%	~0%	——
流水记录	83.3%	94.9%	87.7%
浏览行为	14.7%	13.7%	107.2%
信用卡账单记录	4.4%	1.8%	244.4%

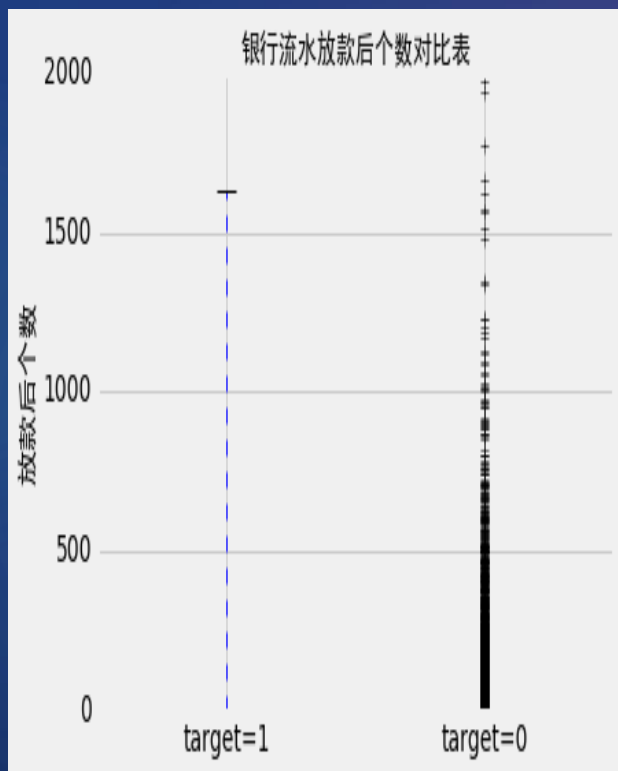
数据缺失严重!!!



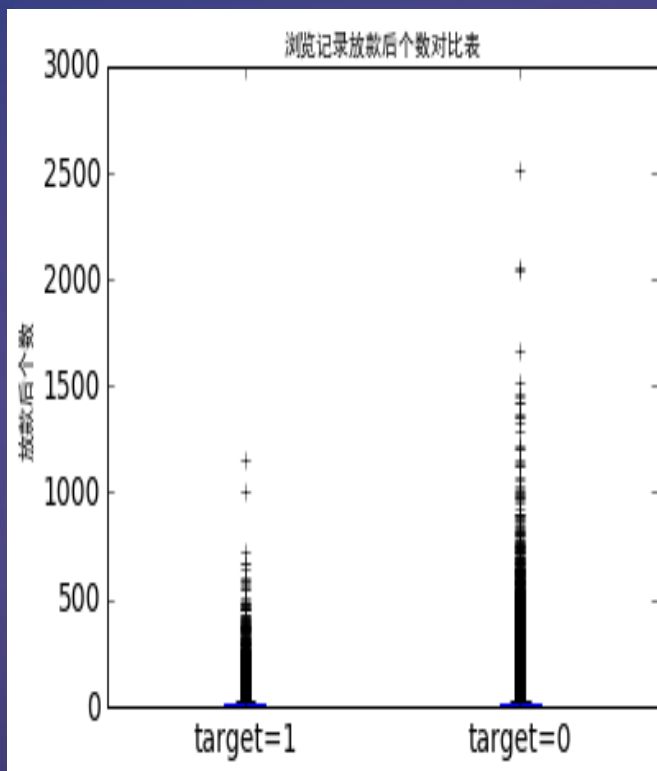
## ► 数据探索

# 放款信息

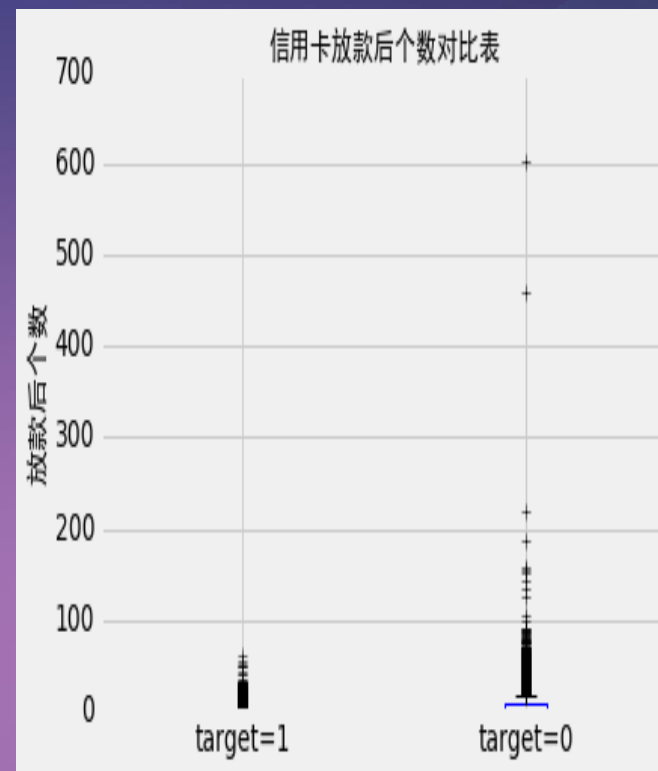
### 银行流失



### 浏览记录



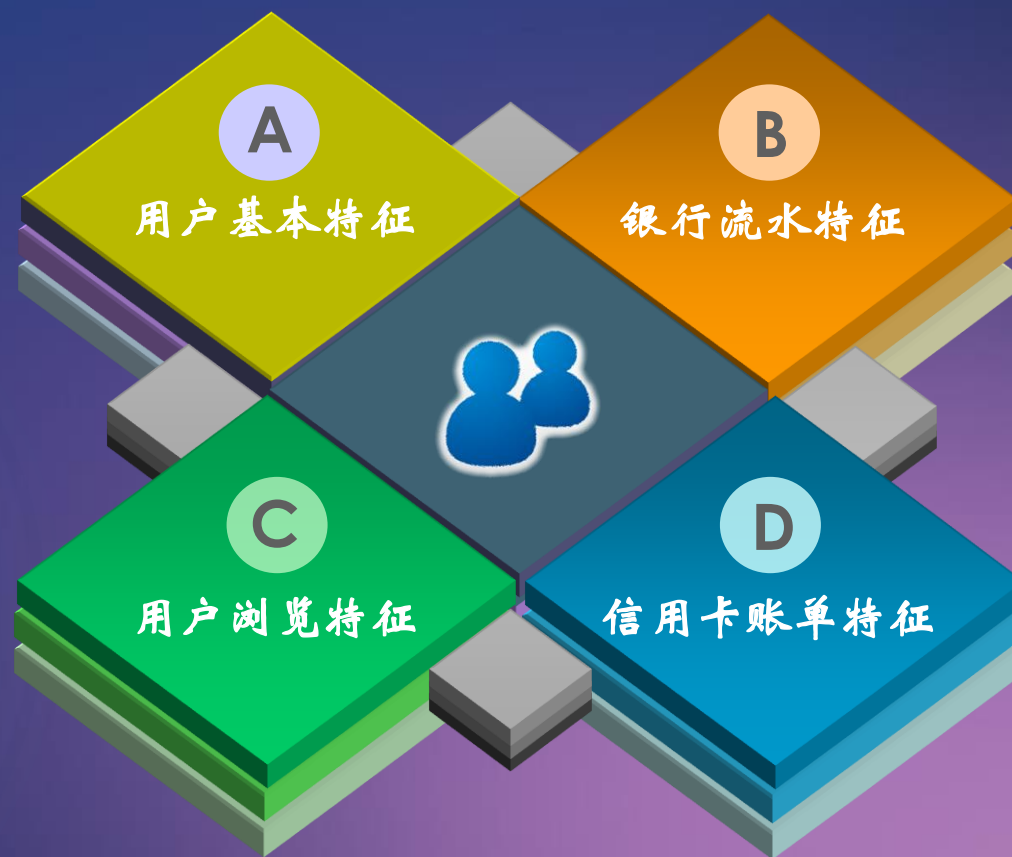
### 信用卡账单



# 4

## 特征工程

## ► 特征工程



## ► 特征工程

### 用户基本特征

#### ◆ 描述用户的基本属性信息

- 性别,职业,教育程度,婚姻状态,户口类型

### 银行流水特征

#### ◆ 描述用户的所有银行消费信息

- 账户流通金额,收入,支出,是否工资收入

——> 交易总量、频繁程度,收入支出比,平均每天收入金额

### 用户浏览行为

#### ◆ 描述用户对本身消费行为的关注情况

- 用户对于自身消费行为的关注程度 ——> 识别用户的责任意识
- 214种主浏览数据,11种子浏览行为
- 274组浏览数据——浏览行为

## ► 特征工程

### 信用卡消费特征 ◆ 描述用户的信用消费状况

#### 统计特征

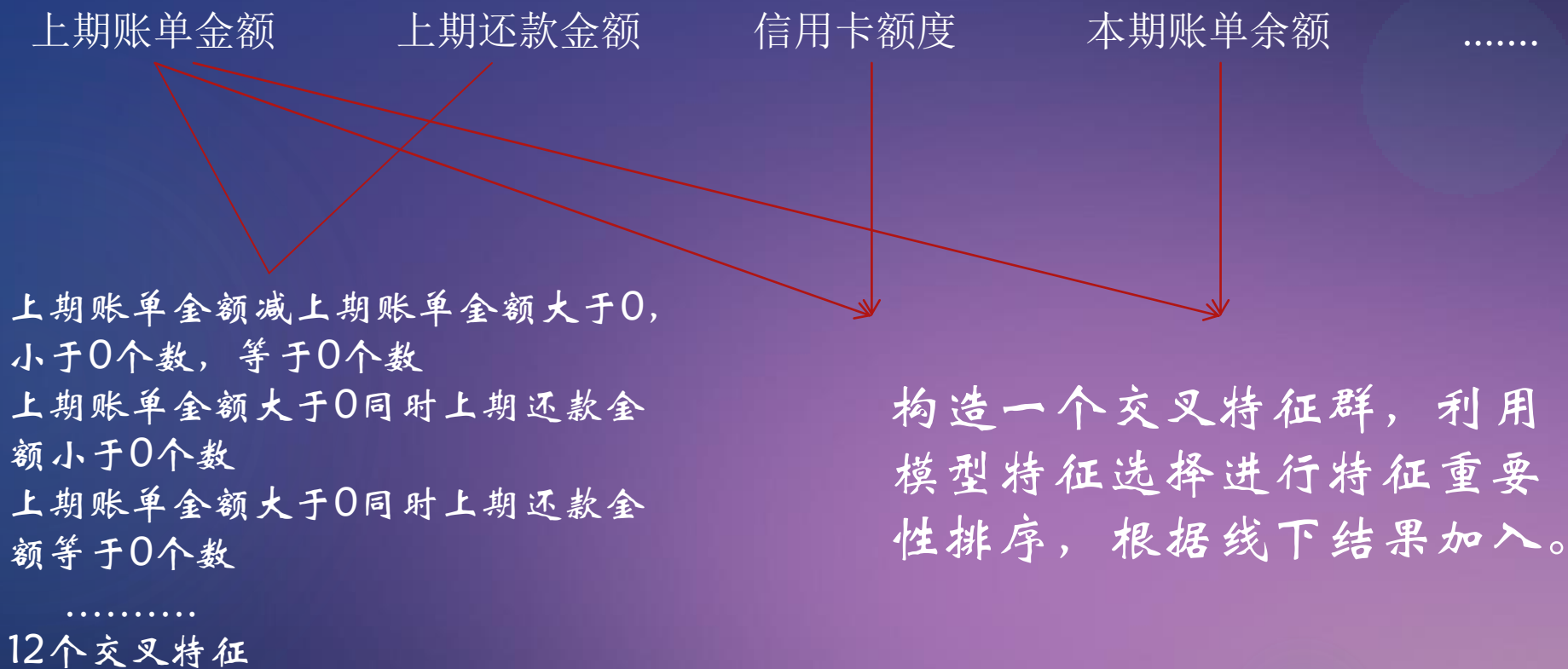
信用卡记录频率  
银行个数  
上期账单金额最大值  
上期账单金额平均值  
上期还款金额最大值  
上期还款金额平均值  
信用额度最小非零值  
.....

#### 业务特征

上期账单金额大于上期还款金额的记录个数  
总体额度上升or下降  
轻度拖欠期数  
轻度拖欠金额  
重度拖欠期数  
重度拖欠金额  
重度拖欠期数占比  
重度拖欠金额占比  
.....

## ► 特征工程

### 交叉特征



## ► 特征工程

### 时间特征

#### ◆ 描述用户的时间信息

##### 统计特征

时间不同的个数  
时间未知记录个数  
记录的时间跨度  
用户最长的卡龄  
用户平均卡龄  
用户最短卡龄  
.....

##### 与放款时间交叉

放款前最近时间戳  
放款后最近时间戳  
时间记录最大值与放款时间差  
.....

## ► 特征工程

### 放款特征

#### ◆ 描述用户放款前后特征对比

##### 银行流水

放款前后记录个数  
放款前后个数比  
.....

##### 浏览记录

放款前后记录个数  
放款前一个月记录个数  
.....

##### 信用卡账单

放款前后记录个数  
放款后还款比例  
放款后信用卡支出金额  
.....





5

模型融合

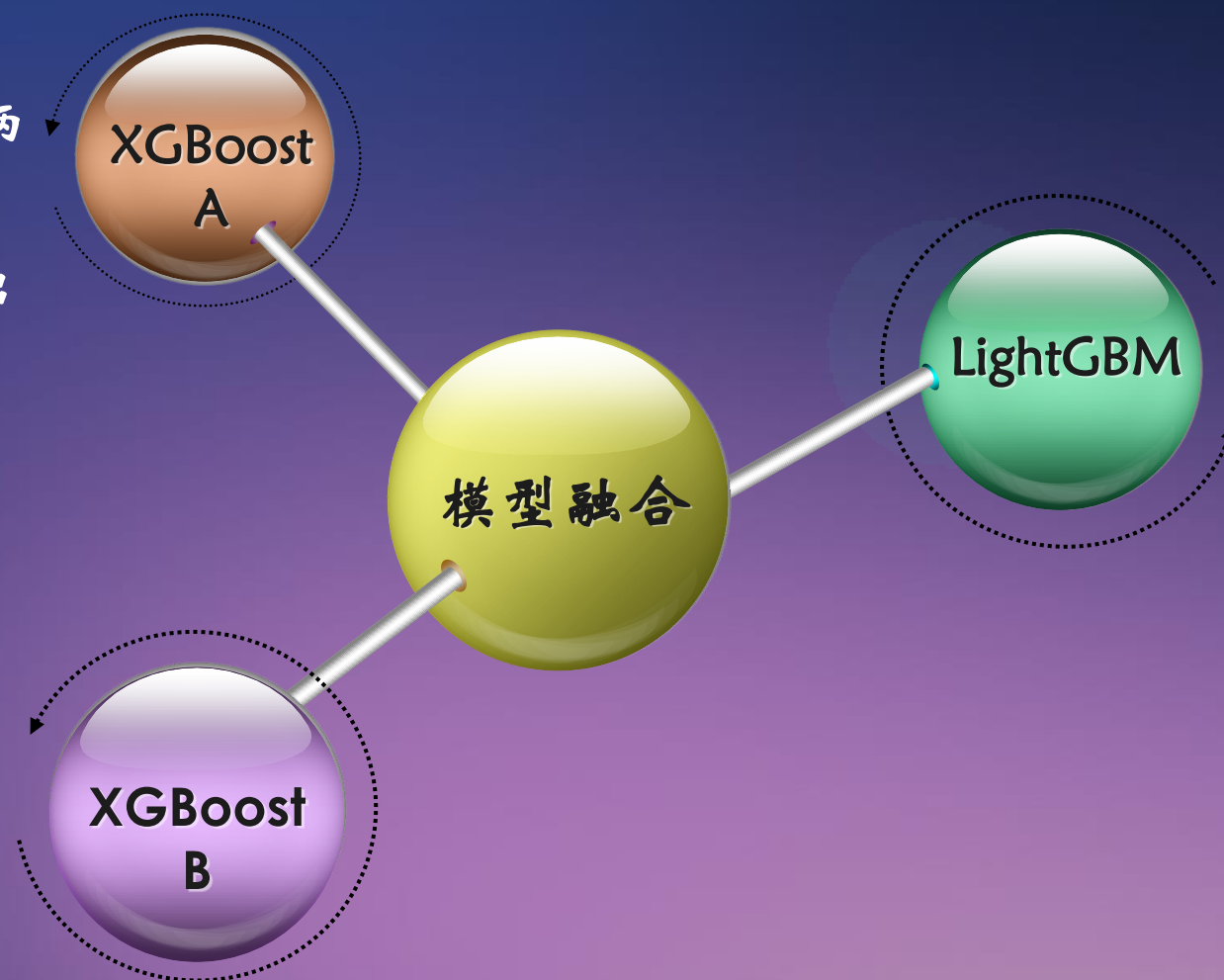
## ► 模型融合

### ◆ 模型获取

- 对于最终数据，使用两种缺失值处理，两种参数得出XGBoost\_A和XGBoost\_B
- 使用lightgbm进行5折Bagging预测，得出LightGBM

### ◆ 直接加权融合

- 最终结果KS值0.46702 =  
 $0.05 * XGB\_A + 0.20 * LGB + 0.75 * XGB\_B$





6

总结

## ► 总结

### 融360用户风险预测

- ◆ 线下验证 → 没有构造稳定的线下验证
- ◆ 模型差异 → 缺少构造多个数据集
- ◆ 模型融合 → 缺少模型融合经验

个人能力的提升与团队协作的锻炼



THANKS

请各位专家批评指导！