四川大学
SICHUAN UNIVERSITY

微额借款用户人品预测

答辩人：朱秋辉（四川大学）

成员：郭柯娜（四川大学）
黄志标（中国科学院）

目录

1、赛题分析

1.1 问题及数据描述

1.2 问题分析

1.1 问题及数据描述

问题描述

利用数据挖掘知识来分析“小额微贷”申请借款用户的信用状况。

数据描述

- 15000带标注样本
  train_x train_y
- 50000无标注样本
  train_unlabeled
- 5000测试样本
  test_x

1.2 问题分析

分类问题（不平衡类）

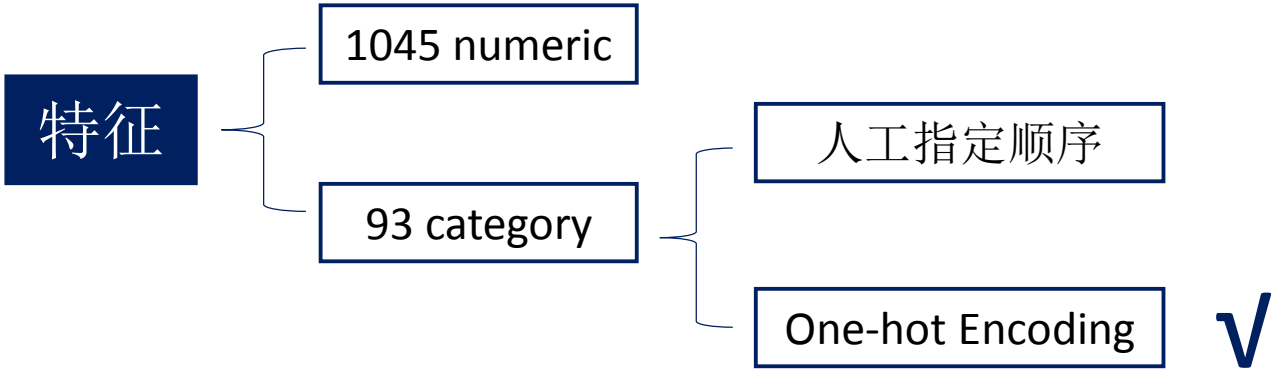
- 特征
  - 经过脱敏处理
  - 1045 numeric 93 category
  - 大量缺失值-1或者-2
  - 特征缺失个数成聚簇类
- 模型
  - 监督模型
  - 半监督模型

2、数据处理

2.1 特征的处理

2.2 新特征的抽取

2.1 特征的处理



改进的One-hot Encoding

特征	取值范围	取值	传统One-hot Encoding	改进One-hot Encoding
X308	[-1, 0, 1]	1	[0, 0, 1]	不变
x308	[-1, 0, 1]	-1	[1, 0, 0]	不变
x415	[1, 2, 3, 4]	3	[0, 0, 1, 0]	[0, 0, 1, 0]

注：-1不能作为特征的取值，选择了36个特征进行OneHotEncoding

2.2 新特征的抽取

uid	train缺失值个数	uid	test缺失值个数
16107	21	1961	24
9791	21	5785	24
14955	21	7007	24
2280	21	10699	24
18201	22	12851	24
13300	22	17957	24
10306	22	3504	25
4889	22	4123	25
7980	22	5686	25
13380	22	7573	25
17847	22	10792	25
11266	22	16398	25
2299	22	18860	25
4694	22	365	26
16661	22	10770	26
15706	23	11569	26
3652	23	18425	26
15938	23	18979	26
15248	23	3081	27
11846	23	5762	27
15837	23	9225	27
11057	23	9818	27
5966	23	11328	27
5858	23	13181	27
19951	23	13242	27
12193	23	16751	27
2435	23	16902	27

根据缺失特征个数，样本成组出现

2.2 新特征的抽取

样本缺失值个数

样本非缺失值个数

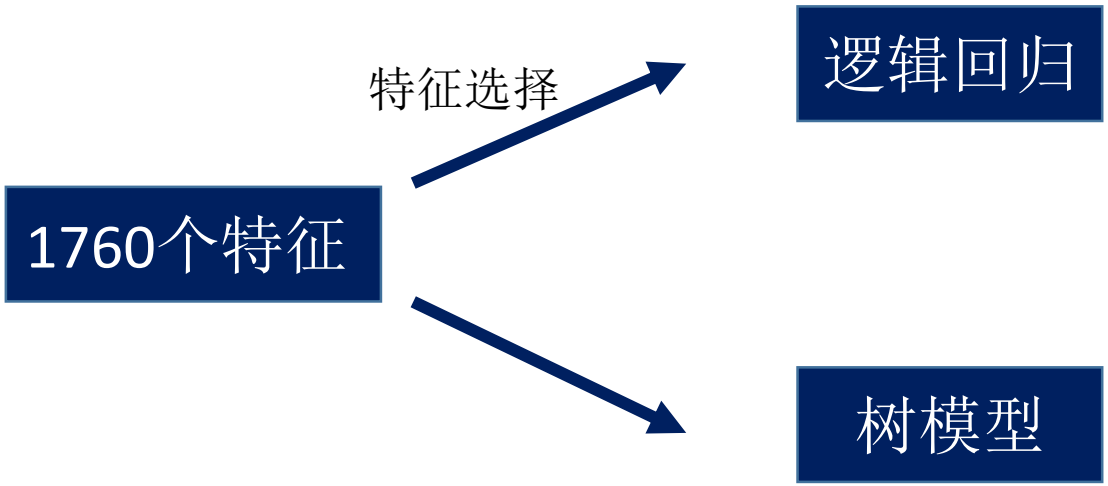
经过One-hot Encoding特征集

1760个特征

3、算法说明

3.1 思路分析
3.2 模型选择
3.3 参数调优
3.4 异常样本分类

3.1 思路分析



3.2 模型选择

模型对比

模型	线下得分	线上得分
逻辑回归[1]	0.756	0.66
随机森林	0.683	0.668
GBDT[2]	0.718	0.722

注：[1] 采用的是启发式搜索（序列前向选择），即每次选择一个特征，使特征函数最优。

[2] 采用的GBDT的多线程版本xgboost。

3.3 参数调优

Xgboost[3]调优过程

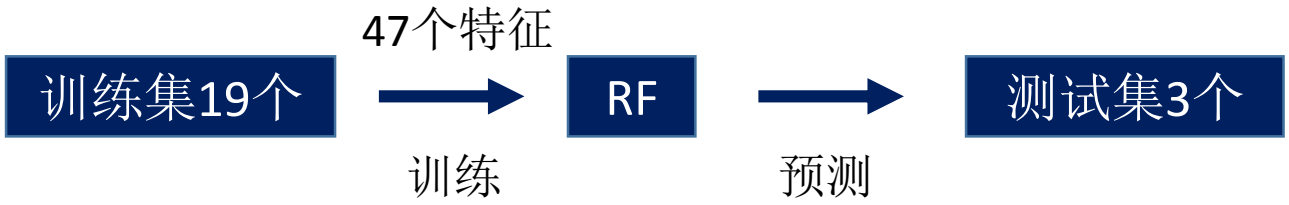
- 随机种子的选择
- 正样本权重
- 最大树深以及树的棵数

注：[3] xgboost 有对缺失值自动处理的功能，如加载数据时可以指定数据的缺失值，dtrain = xgb.Dmatrix(X, label = y, missing=-1)。

3.4 异常样本分类

测试集样本缺失值分布片断

uid	test缺失值个数
5992	1050
10083	1050
19541	1050
10127	583
16422	318



4、参赛收获

4.1 算法未来的改进

算法计划改进

- 对特征进行分组，对每一个特征组提出新特征
- 根据每个样本缺失值个数对样本进行分组，分开预测
- 利用大量无类标的数据

4.2 参赛收获

- 加深了对从数据中发现问题，到解决问题流程的理解。
- 要想达到目标，认清数据的本质才是王道。
- 挑战自己，找一切可能的出路，尝试一切可能的方法。

四川大学
SICHUAN UNIVERSITY

谢谢

致谢：

1、感谢Data Castle 平台

2、感谢主办方的精心组织

3、感谢所有工作人员辛勤的工作

4、感谢所有参赛者，让我学习到很多