用户人品预测大赛--火星人队--竞赛分享

•发布于 2016-03-24 14:55 •作者 DataCastle运营 (/user/5451) •33 次浏览 •来自 微额借款用户人品预测大赛 (/?tab=148)

参赛队队名：火星人

竞赛报告书

一、参赛作品概述

算法概述：本算法分析数据的特点，首先对类别型特征进行了改进的独热编码，并通过分析样本缺失特征数量分析提出了两个具有鉴别力的特征，之后结合全部特征对迭代决策树模型（GBDT）进行训练，初步得出预测结果。最后结合特征缺失异常的样本的预测，得出最终对全部测试集的预测。

关键技术：（1）对特征进行统计分析，利用改进的独热编码技术，对分类型特征进行个别的独热编码。（2）对单个样本缺失值进行统计分析，针对特征缺失个数提取出新特征，并找出异常样本（缺失值个数较多的样本）的预测值。（3）从数据的特点入手，以及对各个模型的分析比较，建模采用迭代决策树进行建立模型，用于对测试集的预测。

二、参赛作品技术路线

算法总思路：

通过特征的预处理，新特征的抽取，采用监督学习的方法建立分类器模型，用于对测试集的预测。

算法原理：

由于分类器往往默认数据是连续的，并且是有序的，而对于 category 型特征，我们通常采用独热编码进行处理，再考虑到数据自身的特点，来改进独热编码，仅处理少量的 category 特征。同时，结合样本缺失值的特点，我们抽取出两个新特征。建模阶段选择能够处理缺失值的迭代决策树来建模。在分析中，根据样本潜在的分组特点，正确预测出了缺失值很多的异常样本。

算法实验结果：

（1） 特征上，利用改进的独热编码，对 36 个 category 型特征进行独热编码，这些特征是：x411,x415~x417,x1107~x1138。编码之后共得到 1758 个特征。

（2） 抽取出每个样本特征缺失的个数与非缺失的个数，总计得到特征 1760 个。

（3） 通过对 GBDT 的分布式版本 xgboost 参数调优后，测试集得到的 auc 值为 0.7229，远高于逻辑回归的 0.66 和随机森林的 0.668.重要参数设置如下：scal_pos_weight:0.13 max_dept:8 eta:0.02 seed:1220 num_boost_round=8000。下表即为正样本权重的选择：

正样本权重	0.11	0.12	0.13	0.14	0.15
score	0.71994	0.72069	0.72105	0.72084	0.72006

答辩PPT

四川大学
SICHUAN UNIVERSITY

微额借款用户人品预测

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

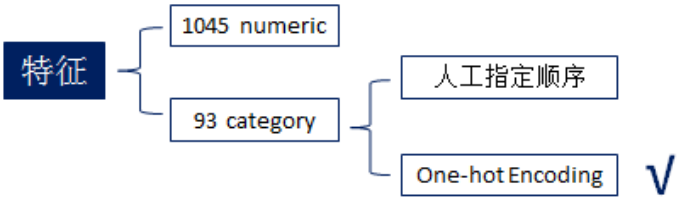
分类问题（不平衡类）

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]

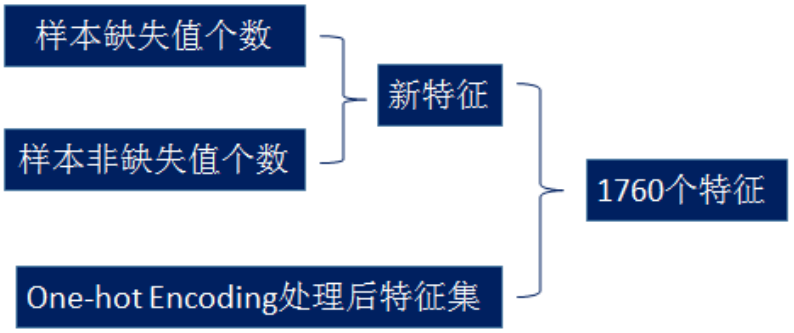
注：-1不能作为特征的取值，选择了36个特征进行OneHotEncoding

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



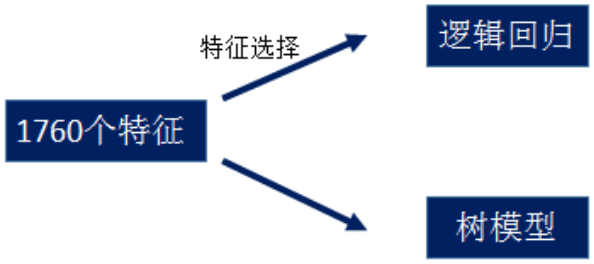
Uid	One-hot Encoding处理后特征	样本缺失值个数	样本缺失值个数
3506	……	26	1732

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

回到顶部

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]参数调优过程

- 正样本权重
- 最大树深
- 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

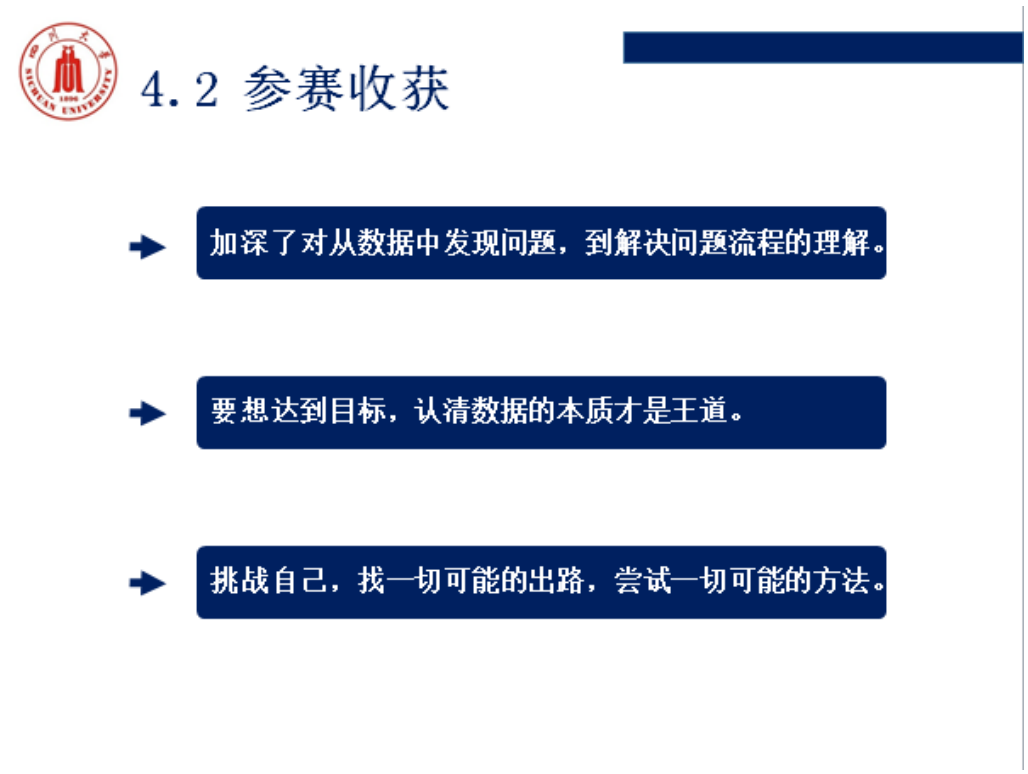
uid	test缺失值个数
5992	1050
10083	1050
19541	1050
10127	583
16422	318

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



0 回复

添加回复 **注:回复会奖励1点DC币，但被管理员删除回复，将扣除作者2DC币;可以使用@符号回复其他人**

回复

作者

(/user/5451) DataCastle运营 (/user/5451)

DC币: 428

无人回复话题

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)

作者其他话题

用户人品预测大赛获奖团队分享 (/topic/58870500b2f84ddb9cbd4f6a45f180df.html)

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)