**微额借款用户人品预测大赛**

**竞赛报告书**

**完成时间： 2016年03月15日**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **参赛题目** | **微额借款用户人品预测大赛** | | | | | |
| **队 名** | **火星人** | | | | **队长电话** | **15881177712** |
| **参评奖项** | **□ 排名奖**  **√ 特别奖** | | | | **线上环节** | **第 12 名（线上排名）**  **0.7254 分（线上得分）** |
| **队 伍**  **基 本**  **信 息**  **(含队长)** | **姓名** | **性别** | **年龄** | **分工** | **在读学校（或工作单位）及身份** | |
| 朱秋辉 | 男 | 24 | 代码编写 | 四川大学硕士研究生 | |
| 郭柯娜 | 女 | 24 | 思路分析 | 四川大学硕士研究生 | |
| 黄志标 | 男 | 24 | 代码编写 | 中国科学院硕士研究生 | |
|  |  |  |  |  | |
|  |  |  |  |  | |
| **作**  **品**  **概**  **述** | 1. **参赛作品概述（算法概述、关键技术，2000字以内）**   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。 | | | | | |
| **技**  **术**  **路**  **线** | 1. **参赛作品技术路线（算法总思路、算法原理、算法实验结果，4000字以内）**   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：      1. 异常样本的识别。训练数据特征缺失个数为1050的样本有19个，测试集3个。利用19个样本进行通过随机森林进行建模，得出结果如下：   Uid 5992 10083 19541  Score 0.935 0.755 0.903  Uid为10083的预测结果与其他两个样本差异性很大，得出uid为5992，10083，19641的类标分别为1，0，1。 | | | | | |
| **作**  **品**  **小**  **结** | 1. **作品总结（算法优势、可能的改进方向，500字以内）**   算法优势：(1)改进的对特征进行One-hot Encoding的方法，相对于传统的One-hot Encoding方法生成的特征集的个数少，使模型训练相对简单。（2）新加入的两个特征相对于原来特征集的特征作用明显。（3）单个模型降低了模型训练的复杂度。  可能改进的方向：（1）仔细分析特征之间的关联性（可根据特征的缺失特性），对特征进行分组，对每个特征组提取出缺失值特征。（2）根据样本特征的缺失个数对样本进行分组，分开预测。（3）利用大量无类标的数据，即半监督学习的方法。 | | | | | |
| **评**  **委**  **意**  **见** | 1. **答辩评审组意见**   **评审组负责人签名：**  **年 月 日** | | | | | |
| **专**  **家**  **意**  **见** | 1. **专家评审组意见：**   **专家组负责人签名：**  **年 月 日** | | | | | |