高校大学生助学金预测文档

运行环境

- python2,ipython notebook
- 第三方库：numpy,pandas,matplotlib,scikit-learn0.18

数据来源

- Datacastle大数据竞赛平台 http://www.pkbigdata.com/common/cmpt/大学生助学金精准资助预测_竞赛信息.html
- 数据分为两组，分别是训练集和测试集，每一组都包含大约1万名学生的信息纪录:
    1. 图书借阅数据
    2. 一卡通数据
    3. 寝室门禁数据
    4. 图书馆门禁数据
    5. 学生成绩数据
    6. 助学金获奖数据
- 训练集与测试集大小都为1G

特征提取

- 对每个表和每一列都进行对应的特征提取，代码详见`feature-extraction-construction`文件

特征连接

- 将处理好后的特征文件，以学生学号为主键连接成一张表，特征总数为500多个特征，相关代码存放在`merge-feature`文件

特征预处理

- 识别特征的异常值并进行相关处理，相关代码存放在`preprocessing`文件

模型

- 以学生获奖数据作为标签，其它数据作为特征，用梯度提升决策树进行建模，相关代码存放在`model`文件。如果想运行`model_fit_glb.ipynb`脚步，需要额外安装Graphlab机器学习框架
- 模型准确率：0.85