



**基于数据挖掘的电力行业客户画像分析**

**The Report on Personas**

我们

指导老师：吴清峰

**二〇一七年 七月二十一日**

目录

[第1章 绪论 1](#_Toc488245234)

[1.1引言 1](#_Toc488245235)

[1.2论文组织结构 1](#_Toc488245236)

[第2章 赛题解析 2](#_Toc488245237)

[2.1学生资助调研 2](#_Toc488245238)

[2.2赛题理解 3](#_Toc488245239)

[2.1.1高校贫困生资助 3](#_Toc488245240)

[2.1.2大数据时代下的高校贫困生资助 3](#_Toc488245241)

[2.3研究思路 4](#_Toc488245242)

[2.3.1数据分析和处理 4](#_Toc488245246)

[2.3.2特征工程 4](#_Toc488245247)

[2.3.3算法调研、建模研究 4](#_Toc488245248)

[第3章 数据分析 5](#_Toc488245250)

[3.1数据表分析 5](#_Toc488245251)

[3.1.1 图书借阅数据borrow 5](#_Toc488245252)

[3.1.2 一卡通数据card 6](#_Toc488245253)

[3.1.3 图书馆门禁数据library 8](#_Toc488245254)

[3.1.4 学生成绩数据score 9](#_Toc488245255)

[3.1.5 助学金数据subsidy 10](#_Toc488245256)

[3.1.6 学生成绩数据score 联合 一卡通数据card 11](#_Toc488245257)

[3.2多表聚合关联分析 12](#_Toc488245258)

[3.2.1 取全学号 12](#_Toc488245259)

[3.2.2 设置默认值 13](#_Toc488245261)

[3.2.3利用Update语句更新特征值 13](#_Toc488245262)

[3.2.4最终表 13](#_Toc488245263)

[第4章 数据处理 13](#_Toc488245265)

[4.1 处理方法 13](#_Toc488245269)

[4.1.1语义分析法 13](#_Toc488245270)

[4.1.2数据不平衡处理 14](#_Toc488245271)

[4.1.3离散化特征 15](#_Toc488245272)

[4.2 特征选择 16](#_Toc488245273)

[特征选择方法 16](#_Toc488245274)

[第5章 算法建模 19](#_Toc488245275)

[4.1. 分析流程图 19](#_Toc488245277)

[4.2. 算法调研 19](#_Toc488245278)

[1.2.1. 各种适于非线性的分类模型的优缺点 20](#_Toc488245282)

[1.2.2. 在本项目下实现结果较好的几种算法 21](#_Toc488245283)

[4.3. 模型融合 22](#_Toc488245284)

[5.3.1. 基于adaboost的单模型融合 22](#_Toc488245293)

[5.3.2. 集成学习VotingClassifier 24](#_Toc488245294)

[4.4. 模型运用 25](#_Toc488245295)

[第6章 总结 25](#_Toc488245297)

[5.1. 总结 25](#_Toc488245299)

[5.2. 展望 26](#_Toc488245300)

[第7章 参考文献 0](#_Toc488245301)

# 绪论

## 1.1引言

随着信息技术的发展，高校以数字化信息和网络为基础，建立集教学、科研、管理、技术服务、生活服务等应用为一体的教育环境。学校的数据库中，保存有学生静态和动态信息。静态信息是学生的相对稳定的信息，包括学生户籍、学号、所属学院等信息；动态信息是随着用户行为不断变化的信息，包括校园卡消费信息、门禁信息、图书馆出入信息等。通过这些信息，可以为学校师生提供更好的服务。

本系统主要目标是将繁杂的学生信息，通过数据分析和数据挖掘技术，把学生中的贫困生抽取出来并分配一定的助学金，取代以往通过人工完成的助学金分配。以往的助学金发放，往往由辅导员等教师，通过班干部讨论来决定是否分配，这种人工方法存在片面性和主观性，很难达到精准的预测；而利用大数据分析和机器学习技术，通过分析学生消费模式，可以实现高精度的贫困生预测。

国内通过学生消费情况进行资助的高校还为数不多，但是中科大的“隐形资助”已经很成熟，前不久还在网络上火了一把。2013年开始，教育部在全国推广中科大的做法，郑州大学、南京理工大学等高校也纷纷效仿。

## 1.2论文组织结构

本文的主要研究工作是利用高校信息化系统采集的学生数据，通过聚类和统计等方法对学生进行贫困预测，最后分配助学金额。学生数据较多，且来自于不同系统，因此需要进行数据规整；分析结果要给行政人员查看，因此结果展示应当简洁、明了、易懂。

本文的论文结构如下：

第一章：绪论。首先是引言，介绍项目的背景、项目目的以及国内高校现状。随后根据前面的分析，简要地叙述本文的研究工作和结构。

第二章：赛题解析。主要从学生资助调研、对赛题的理解、研究思路等三个方面进行介绍。

第三章：数据分析。对比赛提供的几个表进行单表分析和多表聚合分析。

第四章：数据处理。首先介绍有哪些数据处理的方法，而后对数据进行预处理，最后进行特征选择。

第五章：算法建模。给出算法分析流程图，进行算法调研，再对几个较为优秀的算法进行融合，最后是模型应用。

第六章：总结。回顾本文完成的工作，总结在完成项目的过程中应用的技术和走过的弯路，最后对贫困预测的发展进行展望。

# 赛题解析

## 2.1学生资助调研

大数据时代的来临，为创新资助工作方式提供了新的理念和技术支持，也为高校利用大数据推进快速、便捷、高效精准资助工作带来了新的机遇。基于学生每天产生的一卡通实时数据，利用大数据挖掘与分析技术、数学建模理论帮助管理者掌握学生在校期间的真实消费情况、学生经济水平、发现“隐性贫困”与疑似“虚假认定”学生，从而实现精准资助，让每一笔资助经费得到最大价值的发挥与利用，帮助每一个贫困大学生顺利完成学业。因此，基于学生在校期间产生的消费数据运用大数据挖掘与分析技术实现贫困学生的精准挖掘具有重要的应用价值。

本次赛题以贫困预测为主，因此下文将从贫困预测角度来谈论对赛题的理解。。

## 2.2赛题理解

### 2.1.1高校贫困生资助

教育公平是社会公平的重要基础，促进教育公平是国家基本教育政策。党和国家高度重视家庭经济困难学生上学问题，近些年中央有关部门密集出台相关资助政策措施，已建立起覆盖学前教育至研究生教育的学生资助政策体系，从制度上保障了不让一个学生因家庭经济困难而失学。

本专科生教育阶段：国家奖助学金、国家助学贷款、学费补偿贷款代偿、校内奖助学金、勤工助学、困难补助、伙食补贴、学费减免、“绿色通道”等多种方式的混合资助体系。

### 2.1.2大数据时代下的高校贫困生资助

大数据为高校更好地解决贫困生资助工作提供了新的方法。因此，高校必须充分利用各种数据技术来完善贫困生认定机制，建立信息化动态贫困生数据库，提升贫困生的综合素质，达到资助与育人的目的，从而有针对性地做好贫困生资助工作。

利用大数据技术收集分析贫困生在校期间生活、学习消费情况以及不同学生所在地区经济状况的变化情况，结合当地民政部门出具的证明和师生的民主评议，综合判断学生的实际困难程度，进而对提出申请的学生进行全面认定。

图 1.2.2‑1大数据时代用户画像

## 2.3研究思路



### 2.3.1数据分析和处理

此次比赛即通过分析校园卡消费记录、宿舍门禁记录、图书馆门禁记录、图书馆借阅记录、成绩信息等五个方面的信息，得出学生获得助学金信息。

数据存在数据不平衡、数据缺失等问题。

数据不平衡：各类别人数比例不平衡，得助学金人数远少于没获奖人数，属于在大量非目标数据中找少量目标数据的问题。通过过采样，随机增加各个得奖学金类别学生数，使各个类别的学生比例小于4:1,达到数据平衡。

数据缺失：各个表中都缺失一定比例学生数据，如训练集中Student\_id 有10783条数据而Score中仅有9000条数据。在同一个表中，也存在某些字段缺失。我们对数据进行了清洗，去除表中重复条目，并填充了缺失数据。

### 2.3.2特征工程

主要是将数据进行离散化。

离散化的原因其一是众多数据挖掘算法（如决策树、NativeBayes等）只能处理离散化的数据；其二是使模型结果更稳定：把数据分级，而不是原数据可以破除极端值和异常值影响。

具体是将各个特征数据都分成4个级别（1/2/3/4），总共二十余个特征。

### 2.3.3算法调研、建模研究

在具体研究过程中，针对不同的学生行为以及种类，我们可以通过构建模型来分析。可以从如下几方面进行考虑：



**1）选择适当的模型**： 在得到好的属性之后，针对数据的特征，如离散值众多等特点，选用适当的模型进行预测。在比较各个模型的优劣之后，选择较为合适的模型进行细致的调参。这是有效辨别敏感用户的另一种方式。

**2) 多个维度分析建模**： 可以从多个角度入手来分析建立模型，例如抽取用户日用电量进行研究，构建日用电量分析模型，从中分析客户用电是否存在异常行为，尤其是大客户，可以针对电量变化情况特供电量优化方案服务等。

**3)** 不同模型的集成: 不同模型具有不同的偏好，当我们从多个角度用不同的模型对问题进行了预测时就可以得到多个具有不同偏好、对结果具有不同偏差的模型。有效的结合这些模型，让他们互相纠正。这不失为提高模型精度的一种好方式。

# 数据分析

## 3.1数据表分析

### 3.1.1 图书借阅数据borrow

**1）内容构成：**

字段含义：学生id，借阅日期，图书名称，图书编号

9708,2014/2/25,"我的英语日记/ (韩)南银英著 (韩)卢炫廷插图","H315 502"

6956,2013/10/27,"解读联想思维: 联想教父柳传志","K825.38=76 547"

9076,2014/3/28,"公司法 gong si fa = = Corporation law / 范健, 王建文著 eng"

有些图书的编号缺失。字段描述和示例如下（第三条记录缺失图书编号）

对缺失值的处理将在第4章描述

**2）数据表分析：**

**”library\_borrow”：**

该表的一条记录代表了一位学生的借书记录，我们通过主观分析，认为一个学生是否能够获得助学金，以及获得助学金的额度与该学生是否借阅某一些特定的书籍这种具体的数据并没有关联。反之，可能与该学生的总借阅量存在关系。于是，我们通过对该表进行统计分析，生成了一个特征值”library\_borrow”。该特征值代表每位学生的图书借阅总量，将作为训练集的特征之一。

### 3.1.2 一卡通数据card

**1）内容构成：**

学生id，消费类别，消费地点，消费方式，消费时间，消费金额，剩余金额

1006,"POS消费","地点551","淋浴","2013/09/01 00:00:32","0.5","124.9"

1406,"POS消费","地点78","其他","2013/09/01 00:00:40","0.6","373.82"

13554,"POS消费","地点6","淋浴","2013/09/01 00:00:57","0.5","522.37"

**2）数据表分析：**

该表的一条记录代表了一位学生的一次消费记录，通过分析，我们发现消费类别一项的类型包括“charge”,”POS消费”，“交易冲正”，“卡冻结“等19项，我们分析认为对于每一位学生来说，学生卡充值平均金额，即该学生的消费类型为“charge”的消费记录的消费金额之和除以该学生的消费类型为“charge”的消费记录的条数与该学生是否为贫困生有密切关系。于是，我们通过对该表进行统计分析，生成了一个特征值” avg\_charge”。该特征值代表每位学生的每次充钱金额的平均值，将作为训练集的特征之一。此外，分析得出“cost\_amount”（学生历史总消费金额）、“cost\_variance”（学生消费方差）、“cost\_avg\_day\_superMarket”（学生每天超市消费平均值）、“cost\_avg\_day\_laundryroom”（学生每天洗衣房消费平均值）、“cost\_avg\_day\_dinnerHall”（学生每天食堂消费平均值）、“cost\_rate\_supermarket”（学生花费在超市的消费总额占总消费的比例）、“cost\_rate\_laundryroom”（学生花费在洗衣房的消费总额占总消费的比例）、“cost\_rate\_dinnerhall”（学生花费在食堂的消费总额占总消费的比例）、“cost\_times\_day\_supermarket”（学生每天超市消费平均值）、“cost\_times\_day\_dinnerhall”（学生每天食堂消费平均值）、“cost\_times\_day\_laundryroom”（学生每天洗衣房消费平均值）、“cost\_times”（学生消费总次数）、“balance\_rank”（学生卡内余额值在全体学生中的排名）、“card\_days”（学生card活跃天数）、“time6\_7costs”（学生每日6点-7点的消费总额）、“time7\_8costs”（学生每日7点-8点的消费总额）、“totaldinnercosts”（学生该学生日饭堂消费的总额）、“avgdayscosts”（学生的日平均消费）、“consumetimes11\_12”（学生每天 11点 - 12点消费的总次数）、“consumetimes0\_25”（学生单次消费金额在0-2.5元之间的次数）、“countcost0\_10”（学生当日总消费在0-10元范围内的总天数）、“cardrecharge”（学生卡充值总额）、“maxcost7\_8”（学生7点 -8点间的最大单笔消费的消费金额）、“below10\_rank”（学生日消费金额小于10天数占其card活跃天数的比例）、“below2\_5\_rank”（学生单次消费金额小于2.5次数占其总消费次数的比例）、“consume\_rank”（学生的消费排名）、“time7\_8consume\_avg”（学生7点 -8点间的平均消费）都与该学生是不是应该获得助学金有关联，故将他们都作为训练集的特征。

### 3.1.3 图书馆门禁数据library

**1）内容构成：**

字段含义：学生id，门禁编号，具体时间

3684,"5","2013/09/01 08:42:50"

7434,"5","2013/09/01 08:50:08"

8000,"进门2","2014/03/31 18:20:31"

5332,"小门","2014/04/03 20:11:06"

7397,"出门4","2014/09/04 16:50:51"

**2）数据表分析：**

该表的一条记录代表了一位学生的进入或者离开图书馆的记录，图书馆的开放时间为早上7点到晚上22点，门禁编号数据在2014/02/23之前只有“编号”信息，之后引入了“进门、出门”信息，还有些异常信息为null。

**“library\_time\_spand”：**

根据这一特点，我们统计图书馆门禁数据后发现，有很多看似不合理的数据，例如某一学生在某日内有若干次进出图书馆的记录，且时间相隔较短；又如某学生在某天内有进出图书馆的记录，但最后一次是进门记录而不是离开记录。针对这些情况，我们有意忽视“门禁编号”的信息，对一个学生，将其当日记录时间最早的图书馆门禁记录所对应的记录时间作为该学生当日的进馆时间，将其当日记录时间最晚的图书馆门禁记录所对应的记录时间作为该学生当日的出馆时间，来求出该学生当日的图书馆学习时长。通过对所有记录进行统计，得出每一位学生的图书馆学习总时长“library\_time\_spand”，并将其作为训练集的特征之一。

**“library\_times“ ：**

由于在计算特征“library\_time\_spand”时忽略了每天进出图书馆的次数，为了减少这种方式可能产生的副作用，我们又分别统计每位学生进出图书馆的总次数“library\_times”,并将其作为训练集的特征之一。

### 3.1.4 学生成绩数据score

**1）内容构成：**

字段含义： 学生id,学院编号,成绩排名

0,9,1

1,9,2

8,6,1565

9,6,1570

成绩排名的计算方式是将所有成绩按学分加权求和，然后除以学分总和，再按照学生所在学院排序

**2）数据表分析：**

**”score”：**

该表的一条记录代表了一位学生的成绩情况，还包含了他所属的学院，我们通过分析，认为一个学生是否能够获得助学金，以及获得助学金的额度与该学生的成绩存在联系。于是，我们将原数据中的“成绩排名”作为特征“score”并且作为训练集的特征之一。

**“scorerank\_divided\_by\_stunum”:**

由于每个学院的学生总数不同，所以如果单独使用“score”作为特征，可能会导致一些问题。所以我们又在原有数据上统计出各个学院的总人数，并用“scorerank\_divided\_by\_stunum”（学生成绩排名除以所属学院学生总人数）作为特征，来平衡“score”带来的负面影响。

### 3.1.5 助学金数据subsidy

**1）内容构成：**

字段含义：  学生id,助学金金额

10,0

22,1000

28,1000

64,1500

650,2000

**2）数据表分析：**

该表的一条记录表示已知的某一位学生获得的助学金金额，其值可能去4个，分别是0,1000,1500,2000。这是我们最终需要对测试集进行预测后得到的结果。

**“propotion\_of\_1000”、“propotion\_of\_1500”、“propotion\_of\_2000”：**

这三个值分别代表某个学生所在学院获得1000助学金、1500助学金、2000助学金的人数占所有获得助学金学生人数的比例。因为每个学院的情况不一样，学校给每个学院安排的助学金名额也可能存在较大差异，故计算出这三个特征，并将其作为训练集的特征。

### 3.1.6 学生成绩数据score 联合 一卡通数据card

**“score\_rank\*consume\_rank”：**

我们分析认为，一个学生是否属于贫困生，是否应该获得助学金，应该获得多少助学金应该与他的学习成绩，消费情况这两项关系最紧密，为了体现这种关系，我们计算出“score\_rank\*consume\_rank“（学生成绩排名乘以消费排名）”，并将其作为训练集的特征之一。

## 3.2多表聚合关联分析

本系统的数据表众多，一个学生的数据分散在各个不同的表里头，一个表里头有多条相同学生的不同数据记录。为了更加全面的描述学生，我们建立了一个个模型来描述一个学生的所有特征。我们以学生号为连接属性，将上面提到的各种特征进行连接，得到一张总表，在总表中，每个学生的信息只有一行，其主键为学生号，其余字段为上述所有特征的值，在行末再添加一个该学生真实助学金情况。

### 取全学号

由于数据源本身问题，存在部分学生在subsidy表中有记录，而在其他表中无记录的情况，这就给最终表的生成造成的影响。如果直接使用学生号作为连接属性去生成最终表，那么会有很多学生因为在某一张表中没有记录而导致其在最终的表里没有记录。为解决这一问题，我们采取的方法是：首先，将subsidy表（该表中的学号是全的）中的学生号都提取出来；接下来将所有学号插入到最终表中，至此我们就得到了一张每一行只有学号，而其他信息均为默认值的表。



### 3.2.2 设置默认值

上面提及了某些学生可能因为记录的缺失而无法计算一些特征值，对此我们采取的办法是给每一个特征项设置默认值，这样在之后的模型训练中，默认值可以作为特殊的一种取值，不会影响模型的训练。

### 3.2.3利用Update语句更新特征值

在上一部分设置了特征值，之后再以3.2.1中得到的完整学生号作为where语句连接值，一次计算所有学生的每一个特征，如果该学生在该特征上有值，则用Update语句进行更新。

### 3.2.4最终表

经过上面的3个步骤，至此我们就得到了一张最终表，在对该表进行一定转换后就可以用于模型训练。



# 数据处理



## 处理方法

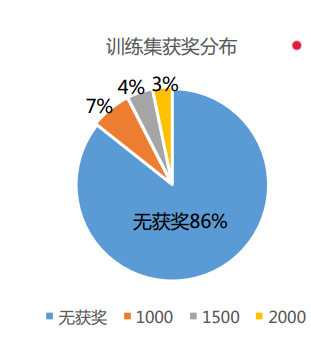
### 4.1.1语义分析法

由于对贫困生的判断除了是否有贫困证明以外，并没有一个统一、标准的判定流程。不能明确的说通过哪几个特征的值就能判断出结果。因此，我们充分挖掘特征实际的语义，挖掘其与贫困程度可能存在的关联，我们通过分析出可能与贫困生判定有关的大量特征，再进一步分析学生画像，从而从学生画像（特征）的角度来判断贫困程度。

### 4.1.2数据不平衡处理

在所给的数据集中，各种类型学生的人数比例相差比较大，此时用标准算法去解决一定会很困难。传统算法往往偏向于多数类，因为他们的损失函数在没考虑数据分布的情况下优化如错误率等量。最坏的情况是，小类别样本会被认为是大类别的异常值而被忽略，学习算法简单的生成一个平凡分类器，将每个样本都分类为大类别。

各类别人数比例不平衡，获得助学金人数远少于没获奖人数，如下图。



我们的目标是找出可以获得助学金的学生的模式，并且还要能区分不同等级助学金之间的差异，对初始训练集来说，属于在大量非目标数据中找少量目标数据的问题。如果不解决数据不平衡问题，就必然会使预测结果偏向不获得助学金这一类型，这违背了我们的目的。

故我们采用过采样的方法，在创建训练数据集时，首先记录下各类型样本的数量，接下来比较各类型样本的数量。若出现任意两种类型的样本数量比超过4:1,则将数量较少的那一类的所有样本复制一遍，再重新比较，直至任意类型样本的数量比都不超过4:1。

### 4.1.3离散化特征

**要求离散化的原因：**

①算法需要：众多数据挖掘算法（如决策树、NativeBayes等）只能处理离散化的数据

②使模型结果更稳定：把数据分级，而不是原数据可以破除极端值和异常值影响

**依次离散化各个特征：**

1）选取一个未转化的特征

2）所有记录按照该特征上的值进行升序排序，得到一个有序的结果集合

3）分别取出该集合的1/4处的记录，1/2处的记录，3/4处的记录，以及集合的最后一条记录

4）用取出的记录在当前选择特征上的取值作为分界的依据，将所有记录的该特征转化为1,2,3,4这4者之一

5）若还有未转化的特征则返回1）

6）结束

**离散化的优势：**

1. 根据上述方法进行特征离散化的处理，可以确保每一条记录在每一个特征上仅有一个唯一的取值，方便后面的模型训练
2. 使得每一特征中各个等级的样本数量大致相同，保证了该特征的可用性

## 特征选择

经过上述处理后可得到一张可以直接用于模型训练的表。该表中包含了许多特征，但有些特征的效果也许不理想，直接使用所有特征必然会降低模型的准确率。故我们要选择某些特征用于下一步的模型训练。

（未筛选特征和筛选之后的准确率的对比，等等看能不能画图）

### 特征选择方法

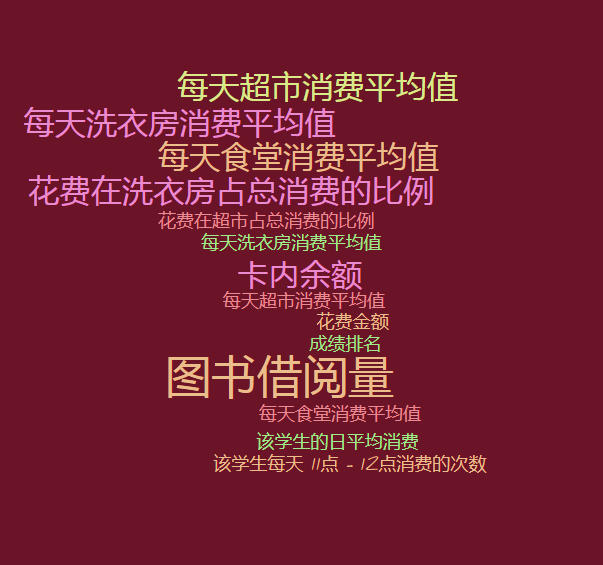
特征重要性评级 (feature importance ranking)

计算输入数据集中每一个属性的重要性，重要性的值可以帮助我们选择出重要的特征。

分类器模型的特征重要度计算值（如上图）所示，我们利用SelectFromModel算法来对上述的特征进行进一步的筛选，最终得到了用于训练的特征，如下表所示：

**最终选择的特征：**

|  |  |
| --- | --- |
| score | 成绩排名 |
| cost\_amount | 花费金额 |
| cost\_avg\_day\_superMarket | 每天超市消费平均值 |
| cost\_avg\_day\_laundryroom | 每天洗衣房消费平均值 |
| cost\_avg\_day\_dinnerHall | 每天食堂消费平均值 |
| cost\_rate\_supermarket | 花费在超市占总消费的比例 |
| cost\_rate\_laundryroom | 花费在洗衣房占总消费的比例 |
| cost\_times\_day\_supermarket | 每天超市消费平均值 |
| cost\_times\_day\_dinnerhall | 每天食堂消费平均值 |
| cost\_times\_day\_laundryroom | 每天洗衣房消费平均值 |
| library\_borrow | 图书借阅量 |
| balance\_rank | 卡内余额 |
| time6\_7costs | 每日6点-7点的消费总额 |
| avgdayscosts | 该学生的日平均消费 |
| consumetimes11\_12 | 该学生每天 11点 - 12点消费的次数 |
| consumetimes0\_25 | 该学生单次消费金额在0-2.5元之间的次数 |
| countcost0\_10 | 该学生当日总消费在0-10元范围的天数 |
| cardrecharge | 卡充值总额 |
| maxcost7\_8 | 7点 -8点间的最大单笔消费 |
| below10\_rank | 日消费金额小于10天数占比 |
| scorerank\_divided\_by\_stunum | 成绩排名除以学院学生人数 |
| avg\_charge | 平均充值金额 |



# 算法建模



## 分析流程图

首先对提供的多张表进行数据清洗、整理，排除掉错误的数据；然后借助上文提到的各种特征数据分析方法对数据进行分析挖掘，挖掘出跟助学金敏感相关的属性特征；接着进行实际的预处理工作，包括数据采样、数据过滤、特征选择、特征生成、缺失值处理、数据变换等；在此基础上，尝试构建不同模型进行训练，主要为适合离散型属性的非线性模型，最后根据定制的模型评估方法（这里使用手动分析）对模型进行评估，需要考虑模型的精度、泛化能力等，目的是得出最理想的模型。



图4.1‑1模型构建流程

经过多次的迭代，选取的特征进行不断的增删，使模型更加精准， 直至达到要求。

## 算法调研

研究各种不同算法的实现原理和它们的适用场景，经过多个模型的相互比较，根据评估指标对比模型在训练集上交叉验证的表现，选择对项目最有利的模型。



### 各种适于非线性的分类模型的优缺点

1. C4.5

以信息增益率为衡量标准实现对数据归纳分类

**优点：**产生的分类规则易于理解，准确率较高

**缺点：**在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效

1. CART

以基于最小距离的尼基指数估计函数为衡量标准对数据进行递归分类

**优点：**抽取规则简便且易于理解；面对存在缺失值、变量数多等问题时非常稳健

**缺点：**要求被选择的属性只能产生两个子节点；类别过多时，错误可能增加的较快

1. Adaboost

针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)

**优点：**高精度，简单无需做特征筛选，不会过度拟合

**缺点：**训练时间过长，执行效果依赖于弱分类器的选择

1. 贝叶斯

通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类

**优点：**算法简单，所需估计的参数很少，对缺失数据不太敏感

**缺点：**属性个数比较多或者属性之间相关性较大时，分类效率下降

1. KNN

如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别

**优点：**简单，无需估计参数，无需训练，适合于多分类问题

**缺点：**计算量较大；可解释性较差，无法给出决策树那样的规则

1. SVM

建立一个最优决策超平面，使得该平面两侧距离平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力

**优点：**更好的泛化能力，解决非线性问题的同时避免维度灾难，可找到全局最优

**缺点：**运算效率低，计算时占用资源过大

1. KMeans

输入聚类个数k，以及包含n个数据对象的数据库，输出满足方差最小标准的k个聚类

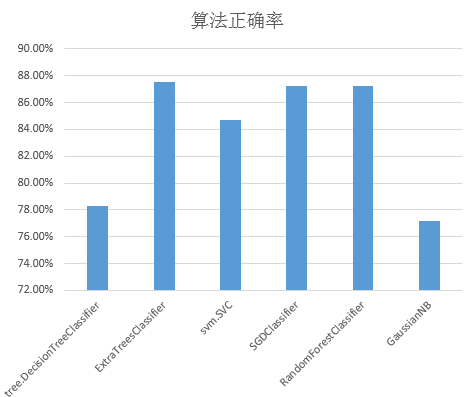
**优点：**运算速度比KNN快

**缺点：**聚类数目k是一个输入参数，不合适的k值可能返回较差的结果

### 在本项目下实现结果较好的几种算法

我们最终发现ExtraTrees、RandomForest和SGD表现相对其他模型更优。这里面主要原因在于大部分属性是离散性的，且维度较高而且彼此间有相互联系，例如学生各类消费比等级、学生消费方差等属性，这对于构建非线性模型是很有帮助。

所有算法单一的的正确率（使用交叉验证，一折作测试集，其余做训练集）如下：



## 模型融合



### 基于adaboost的单模型融合

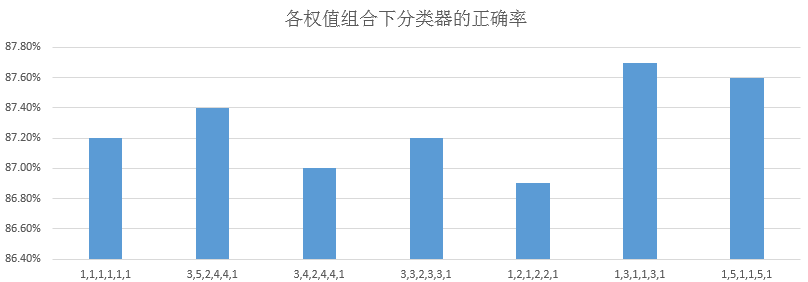
使用上文提到的几种较好的分类算法，通过adaboost方式，将它们训练成较原先更为强大的分类器。

通过迭代不同的次数，使得原先单一的模型在强模型中按不同的权值贡献。其权值为分类后正确率的alpha值。

### 集成学习VotingClassifier

由于单一的分类器在某些方面存在局限性，比如：适于对有高纬度特征的数据进行分类，但是要求各特征间是独立的、和适于对各特征间联系较高的，但对高纬度特征的数据处理困难，这两种算法如果通过某种办法进行优势互补就能得到一个可以处理高纬度且各维度间互有联系的数据的算法模型。

我们使用的是VotingClassifier的软投票机制（即不同分类器对分出的不同类所占的权重比值不一样，相比于硬投票，它是不同分类器占不同比值，对所分的类直接作用）。根据对不同分类器设置分类前的权重来使最后的模型偏向某一正确率最高的分类器(坐标数字为按上述的几个算法顺序赋予的权值组合)。



## 模型运用



# 总结



## 总结

此次项目中，我们团队基于python编程语言，借助Sciki-learn, Pandas,Numpy等开源数据分析模块，采用了多种数据挖掘方法，对学生数据进行深度分析，分别进行了数据预处理，挖掘建模，挖掘结果可视化，并从不同角度进行了统计分析，从而实现了多方位的数据分析挖掘。对所分析的数据，团队通过多次实验、比较分析，寻找并实现了有效的挖掘算法对数据进行建模挖掘。最终，我们得到了如下规则：

* 用电类别上：大工业用电的用户比普通居民用户更敏感

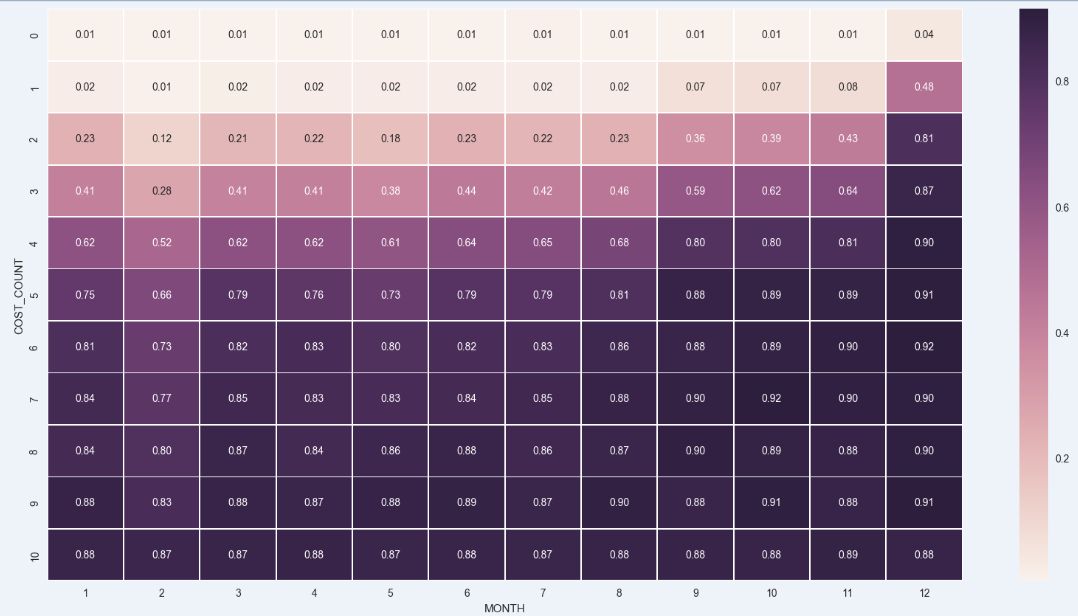


图 4.4.2‑1热力图

## 展望

未来，随着数据量的增长，算法模型的不断优化，希望能够创造出一种自动识别学生画像潜在的模式。不仅仅是助学金情况预测，还可以通过自学习发现诸如学生勤奋度，学生学习有效度等模式，再针对不同的模式，数据挖掘方法能够通过特征分析挖掘出相关的特征因素，建立更加完善的客户画像体系，实现优秀学习模式发现、学生过宅预警、学生失联系预警等目的。

敏感客户画像的建立，有利于我们根据用户的行为特征、个人偏好等等，进行精准化服务，提供适合学生情况的精准化服务。

具体而言，可以有如下几个部分：

**失联预警**

用户可以在首页看到自己管理范围之内的学生失联状况，用户可以自己设置阈值，设置失联多久以后通知用户，超过阈值的学生失联信息会在失联预警模块进行集中展示，并且会通过邮件或者手机短信通知辅导员。

**挂科预警**

同失联预警，学生挂科预警也需要在此处进行显示，学生即将挂科的某门课以及挂科可能性实时显示在首页，供用户辅导学生参考。

**抑郁易感预警**

系统预测学生的在学校内的朋友数量和孤僻指数，然后进行学生的抑郁易感预警，展示在首页，供用户辅导学生参考。

**过宅告警**

系统通过分析学生的宿舍门禁数据，对学生进行过宅告警，告警结果通过过宅级别和平均宿舍逗留时长进行展示和推送给用户。

**失业预警**

系统通过对学生的行为数据进行分析，对学生的失业状况进行预警，及时预测学生的失业，并展示和推送给用户。

# 第7章 参考文献