

2019 年“泰迪杯”数据分析职业技能大赛

B 题

学生校园消费行为分析

一、背景

校园一卡通是集身份认证、金融消费、数据共享等多项功能于一体的信息集成系统。在为师生提供优质、高效信息化服务的同时，系统自身也积累了大量的历史记录，其中蕴含着学生的消费行为以及学校食堂等各部门的运行状况等信息。很多高校基于校园一卡通系统进行“智慧校园”的相关建设，例如《扬子晚报》2016 年 1 月 27 日的报道：《南理工给贫困生“暖心饭卡补助”》。

不用申请，不用审核，饭卡上竟然能悄悄多出几百元……记者昨天从南京理工大学独家了解到，南理工教育基金会正式启动了“暖心饭卡”项目，针对特困生的温饱问题进行“精准援助”。

项目专门针对贫困本科生的“温饱问题”进行援助。在学校一卡通中心，教育基金会的工作人员找来了全校一万六千余名在校本科生 9 月中旬到 11 月中旬的刷卡记录，对所有的记录进行了大数据分析。最终圈定了 500 余名“准援助对象”。

南理工教育基金会将拿出“种子基金”100 万元作为启动资金，根据每位贫困学生的不同情况确定具体的补助金额，然后将这些钱“悄无声息”的打入学生的饭卡中，保证困难学生能够吃饱饭。

——《扬子晚报》2016 年 1 月 27 日：南理工给贫困生“暖心饭卡补助”

本赛题提供国内某高校校园一卡通系统一个月的运行数据，希望参赛者使用数据分析和建模的方法，挖掘数据中所蕴含的信息，分析学生在校园内的学习生活行为，为改进学校服务并为相关部门的决策提供信息支持。

二、目标

1. 分析学生的消费行为和食堂的运营状况，为食堂运营提供建议。
2. 构建学生消费细分模型，为学校判定学生的经济状况提供参考意见。

三、 任务

附件是某学校 2019 年 4 月 1 日至 4 月 30 日的一卡通数据, 请根据附件**自行选择分析工具**完成以下任务。如使用“Python 数据挖掘建模平台”, 使用方式详见附录二。

任务 1 数据导入与预处理

任务 1.1 将附件中的 data1.csv、data2.csv、data3.csv 三份文件加载到分析环境, 对照附录一, 理解字段含义。探查数据质量并进行缺失值和异常值等方面的必要处理。将处理结果保存为“task1_1_X.csv”(如果包含多张数据表, X 可从 1 开始往后编号), 并在报告中描述处理过程。

任务 1.2 将 data1.csv 中的学生个人信息与 data2.csv 中的消费记录建立关联, 处理结果保存为“task1_2_1.csv”; 将 data1.csv 中的学生个人信息与 data3.csv 中的门禁进出记录建立关联, 处理结果保存为“task1_2_2.csv”。

任务 2 食堂就餐行为分析

任务 2.1 绘制各食堂就餐人次的占比饼图, 分析学生早中晚餐的就餐地点是否有显著差别, 并在报告中描述。(提示: 时间间隔非常接近的多次刷卡记录可能为一次就餐行为)

任务 2.2 通过食堂刷卡记录, 分别绘制工作日和非工作日食堂就餐时间曲线图, 分析食堂早中晚餐的就餐峰值, 并在报告中描述。

任务 2.3 根据上述分析的结果, 为食堂的运营提供建议。

任务 3 学生消费行为分析

任务 3.1 根据学生的整体校园消费数据, 计算本月人均刷卡频次和人均消费额, 并选择 3 个专业, 分析不同专业间不同性别学生群体的消费特点。

任务 3.2 根据学生的整体校园消费行为, 选择合适的特征, 构建聚类模型, 分析每一类学生群体的消费特点。

任务 3.3 通过对低消费学生群体的行为进行分析, 探讨是否存在某些特征, 能为学校助学金评定提供参考。

四、 关于竞赛成果提交的说明

1. 登录方式

请使用**队员 1** 的账号登录数睿思网站（www.tipdm.org），进入第二届技能大赛页面。为保证成功提交，**请使用谷歌浏览器**。

2. 作品提交

报告以 PDF 格式提交，文件名为“report.pdf”，要求逻辑清晰、条理分明，内容包括每个任务的完成思路、操作步骤、必要的中间过程、任务的结果及分析。

3. 附件提交

3.1 如使用编程实现，将任务 1、2、3 的源程序分别保存到“program1”，“program2”，“program3”文件夹，然后存放到“program”文件夹中；如使用 Python 数据挖掘建模平台实现，将使用平台建立的工程导出保存到“program”文件夹中。

3.2 将任务 1、2、3 所产生的结果文件，分别保存到“result1”，“result2”，“result3”文件夹，然后存放到“result”文件夹中。

3.3 将“program”、“result”及作品的 word 版本打包成文档“appendix.zip”作为附件提交。

4. 提交界面

4.1 在依次上传完“竞赛承诺书”、“作品”、“附件”后，点击“提交”。

竞赛资讯

赛制介绍

报名缴费

提交A题

提交B题

我的成绩

*对应题目: B

*上传竞赛承诺书(PDF): 选择文件 竞赛承诺书.pdf

注:

1、承诺书模板下载: 竞赛承诺书模板。

2、竞赛承诺书打印后签名, 并扫描生成pdf文件。

*上传作品(报告PDF): 选择文件 report.pdf

*上传附件(zip文件): 选择文件 appendix.zip

注:

1、若选择附件错误, 请按F5刷新页面, 重新选择附件。

2、附件总大小不能超过150MB。

举报: 是否发现违反竞赛章程的现象: ☒ 是 ☐ 否 选择文件 未选择任何文件

注:

若发现有违反竞赛章程的现象, 请以Word描述该现象并收集相关证据(图片, 音频, 视频等)以压缩包的形式提交。

作品论文内容及命名请勿出现学校、学院、队号、队员以及指导老师相关信息, 否则该作品视为无效作品。

提交

4.2 待页面弹出“上传成功”对话框，点击“确定”，在相应位置可以看到“已上传 XXX”字样，表示相关文件提交成功。

竞赛资讯	*对应题目2018-11-11:	B
赛制介绍	*上传竞赛承诺书(PDF):	<div>浏览... 未选择文件。 已上传承诺书!</div> <div>注: 1、承诺书模板下载: 竞赛承诺书模板。 2、竞赛承诺书打印后签名, 并扫描生成pdf文件。</div>
报名缴费	*上传作品(报告PDF):	<div>浏览... 未选择文件。 已上传作品!</div>
提交A题	*上传附件(zip文件):	<div>浏览... 未选择文件。 已上传附件!</div> <div>注: 1、若选择附件错误, 请按F5刷新页面, 重新选择附件。 2、附件总大小不能超过150MB。</div>
提交B题	举报:	<div>是否发现违反竞赛章程的现象: <input type="radio"/> 是 <input checked="" type="radio"/> 否 浏览... 未选择文件。</div> <div>注: 若发现有违反竞赛章程的现象, 请以Word描述该现象并收集相关证据(图片, 音频, 视频等)以压缩包的形式提交。</div>
我的成绩	<div>作品论文内容及命名请勿出现学校、学院、队号、队员以及指导老师相关任何信息, 否则该作品视为无效作品。</div>	

4.3 在比赛当天 20:00 竞赛成果截止提交之前，可多次上传相关文件，系统默认以最后上传的文件为准。

附录一 数据说明

附件包含 3 张数据表，分别为 data1.csv、data2.csv、data3.csv，对应于学生 ID 表、消费记录表和门禁记录表。各字段说明如表 1-3 所示。

表 1 data1.csv 字段说明

字段名	描述
Index	序号
CardNo	校园卡号。每位学生的校园卡号都唯一
Sex	性别。分为“男”和“女”
Major	专业名称
AccessCardNo	门禁卡号。每位学生的门禁卡号都唯一

表 2 data2.csv 字段说明

字段名	描述
Index	流水号。消费的流水号
CardNo	校园卡号。每位学生的校园卡号都唯一
PeoNo	校园卡编号。每位学生的校园卡编号都唯一
Date	消费时间
Money	消费金额。单位：元
FundMoney	存储金额。单位：元
Surplus	余额。单位：元
CardCount	消费次数。累计消费的次数
Type	消费类型
TermNo	消费项目的编码
TermSerNo	消费项目的序列号
conOperNo	消费操作的编码
OperNo	操作编码
Dept	消费地点

表 3 data3.csv 字段说明

字段名	描述
Index	序号
AccessCardNo	门禁卡号。每位学生的门禁卡号都唯一
Date	进出时间
Address	进出地点
Access	是否通过。分为“0”和“1”
Describe	描述。分为“禁止通过-没有权限”和“允许通过”

附录二 Python 数据挖掘建模平台使用说明

Python 数据挖掘建模平台（以下简称平台）是由广东泰迪智能科技股份有限公司提供的一个开源数据挖掘建模工具（官网：<http://python.tipdm.org/>）。基于该平台，参赛者可在没有编程基础的情况下，通过拖拽的方式进行操作，将数据输入输出、数据预处理、挖掘建模等环节通过流程化的方式进行连接，以达到数据分析挖掘的目的。

竞赛专用平台访问网址：<http://103.24.177.103:8081/tipdm-webapp>

平台使用教程：<https://edu.tipdm.org/course/100>

竞赛专用平台访问账号：jn+队伍号（如队号为 2019200001，则平台访问账号为 jn2019200001）

竞赛专用平台初始密码：Jn123456（为保证账户安全，请尽快修改密码。修改密码方式详见“修改平台登录密码.PDF”）

赛题数据已通过共享数据源的方式，分享给所有参赛者，如图 1 所示，参赛者可以直接在共享数据源中查看，并在工程中配置“输入源”使用，无需上传数据。

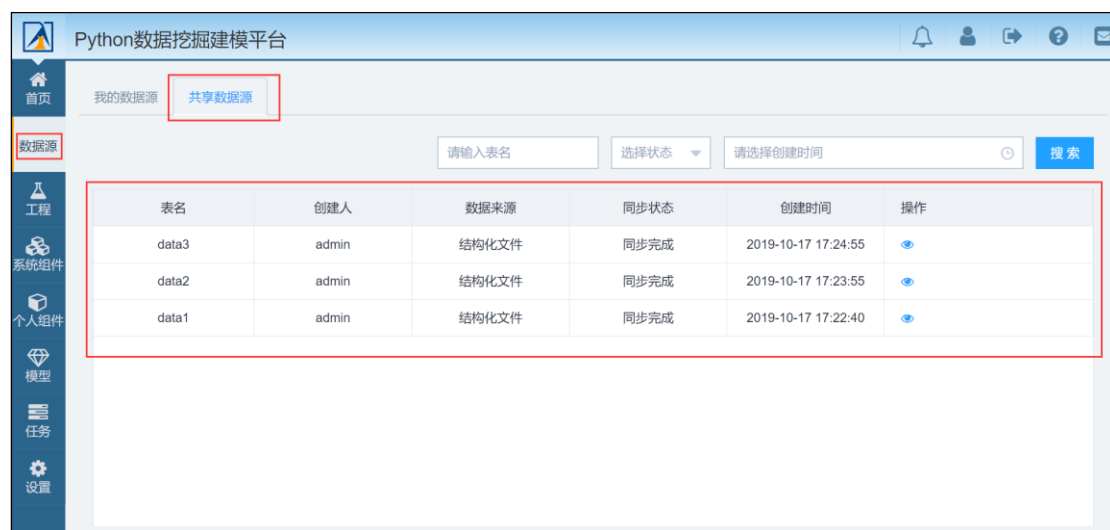


图 1 共享数据源

参赛者可以通过“输出源”组件，将结果保存到本地，并按赛题要求保存到“result”文件夹中提交，步骤如下所示。

（1）拖入“输入/输出”下的“输出源”组件，与想要输出结果的组件相连接，然后单击蓝色刷新按钮，如图 2 所示。

(2) 参数设置中，填入目标表名称，如图 3 所示。

(3) 运行“输出源”组件，运行完成后，右键“输出源”组件，选择“导出 CSV”，如图 4 所示。

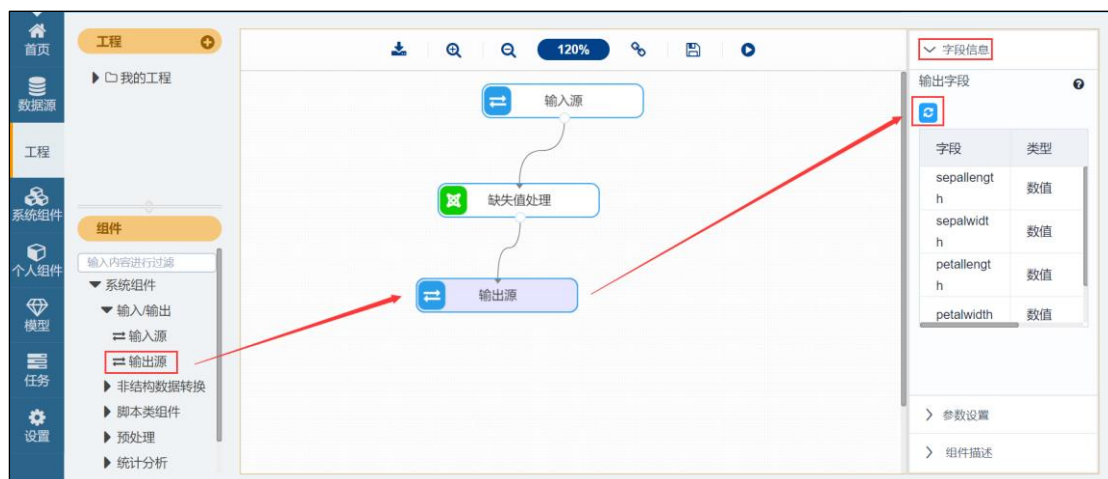


图 2 拖入“输出源”组件

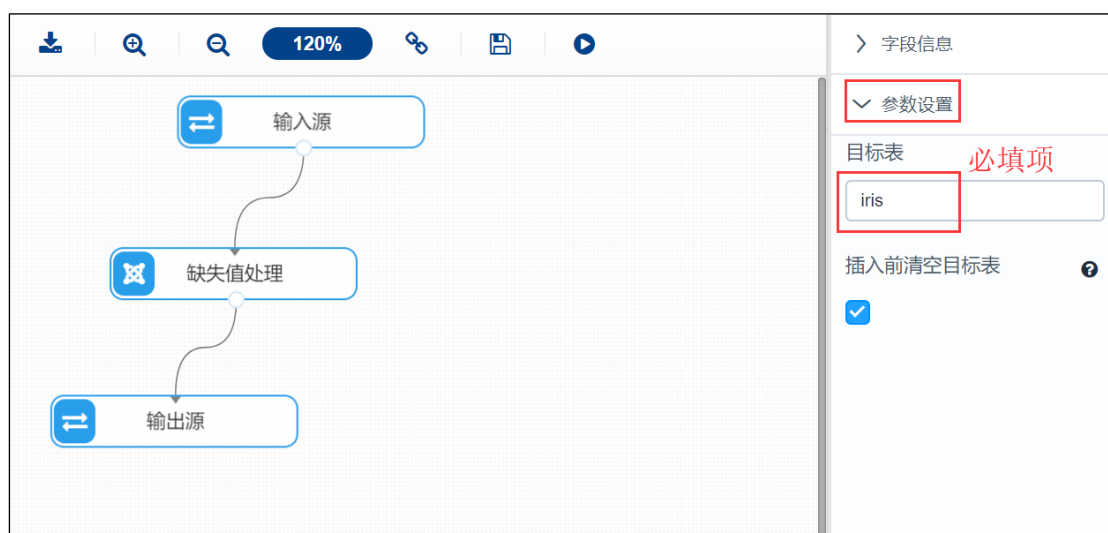


图 3 填入目标表名



图 4 导出 CSV

参赛者使用平台完成赛题后，需使用**工程导出按钮**导出工程文件（如图 5 所示），并保存到 program 文件夹中。



图 5 工程导出按钮