2020 年“泰迪杯”数据分析职业技能大赛

A 题

教育平台的线上课程智能推荐策略

一、 背景

近年来，随着互联网与通信技术的高速发展，学习资源的建设与共享呈现出新的发展趋势，各种网课、慕课、直播课等层出不穷，各种在线教育平台和学习应用纷纷涌现。尤其是 2020 年春季学期，受新冠疫情影响，在教育部“停课不停学”的要求下，网络平台成为“互联网+教育”成果的重要展示阵地。因此，如何根据教育平台的线上用户信息和学习信息，通过数据分析为教育平台和用户提供精准的课程推荐服务就成为线上教育的热点问题。

本赛题提供了某教育平台近两年的运营数据，希望参赛者根据这些数据，为平台制定综合的线上课程推荐策略，以便更好地服务线上用户。

二、 目标

1. 分析平台用户的活跃情况，计算用户的流失率。

2. 分析线上课程的受欢迎程度，构建课程智能推荐模型，为教育平台的线上推荐服务提供策略。

三、 任务

附件是某教育平台 2018 年 9 月至 2020 年 6 月的线上课程运营数据，请根据附件数据，自行选择分析工具完成以下任务，并撰写报告（**报告的要求详见：四、竞赛成果提交说明**）。如使用“TipDM 大数据挖掘建模平台”实现，使用方式详见附录二。

任务 1 数据预处理

任务 1.1 对照附录 1，理解各字段的含义，进行缺失值、重复值等方面的必

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 对用户信息表中 recently_logged 字段的“--”值进行必要的处理，将处理结果保存为“task1_2.csv”，并在报告中描述处理过程。

任务 2 平台用户活跃度分析

任务 2.1 分别绘制各省份与各城市平台登录次数热力地图，并分析用户分布情况。

任务 2.2 分别绘制工作日与非工作日各时段的用户登录次数柱状图，并分析用户活跃的主要时间段。

任务 2.3 记T_{end}为数据观察窗口截止时间（如：赛题数据的采集截止时间为 2020 年 6 月 18 日），T_i为用户 i 的最近访问时间，$\sigma_i = T_{end} - T_i$，若$\sigma_i > 90$天，则称用户 i 为流失用户。根据该定义计算平台用户的流失率。

任务 2.4 根据任务 2.1 至任务 2.3，分析平台用户的活跃度，为该教育平台的线上管理决策提供建议。

任务 3 线上课程推荐

任务 3.1 根据用户参与学习的记录，统计每门课程的参与人数，计算每门课程的受欢迎程度，列出最受欢迎的前 10 门课程，并绘制相应的柱状图。受欢迎程度定义如下：

$$\gamma_i = \frac{Q_i - Q_{min}}{Q_{max} - Q_{min}}$$

其中，γ_i为第 i 门课程的受欢迎程度，Q_i为参与第 i 门课程学习的人数，Q_{max}和Q_{min}分别为所有课程中参与人数最多和最少的课程所对应的人数。

任务 3.2 根据用户选择课程情况，构建用户和课程的关系表（二元矩阵），使用基于物品的协同过滤算法计算课程之间的相似度，并结合用户已选课程的记录，为总学习进度最高的 5 名用户推荐 3 门课程。

任务 3.3 在任务 3.1 和任务 3.2 的基础上，结合用户学习进度数据，分析付费课程和免费课程的差异，给出线上课程的综合推荐策略。

四、 竞赛成果提交说明

1. 登录方式

请使用**队员 1** 的账号登录数睿思，进入第三届技能大赛页面。为保证成功提交，**请使用谷歌浏览器无痕模式**。

2. 报告提交

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。针对各子任务，报告中应包含但不限于如下要点：

（1） **任务 1.1** 应包含每个表中缺失值和重复值的记录数以及有效数据的记录数。

（2） **任务 1.2** 应包含 recently_logged 字段的“--”值的记录数以及数据处理的方法。

（3） **任务 2.1** 应包含各省份与各城市的热力地图以及主要省份和主要城市的数据表格，并进行分析。

（4） **任务 2.2** 应包含工作日与非工作日各时段的柱状图，并进行分析。

（5） **任务 2.3** 应包含对流失率的定义，并给出流失率的结果。

（6） **任务 2.4** 应根据计算结果给出合理的建议。

（7） **任务 3.1** 应包含最受欢迎的前 10 门课程的参与人数、受欢迎程度及柱状图。

（8） **任务 3.2** 应包含相应推荐算法的描述，并给出总学习进度最高的 5 个用户的课程推荐数据。

（9） **任务 3.3** 应包含数据分析的方法、算法描述以及主要结果。

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 TipDM 大数据挖掘建模平台实现，将使用平台建立的工程截图保存到“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”及报告的 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”。

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。

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 1 数据说明

赛题附件包含三张数据表，分别为 users.csv（用户信息表）、study_information.csv（学习详情表）和 login.csv（登录详情表），它们的数据说明分别如表 1、表 2 和表 3 所示。

表 1 users.csv 字段说明

字段名	描述
user_id	用户 id
registration_time	注册时间
recently_logged	最近访问时间
learn_time	学习时长（分）
number_of_classes_join	加入班级数
number_of_classes_out	退出班级数
school	用户所属学校

表 2 study_information.csv 字段说明

字段名	描述
user_id	用户 id
course_id	课程 id
course_join_time	加入课程的时间
learn_process	学习进度
price	课程单价

表 3 login.csv 字段说明

字段名	描述
user_id	用户 id
login_time	登录时间
login_place	登录地址

附录二 TipDM 大数据挖掘建模平台使用说明

TipDM 大数据挖掘建模平台（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（请使用谷歌浏览器无痕模式，该链接仅供竞赛期间使用，竞赛之后链接自动失效）：eb.tipdm.org:10011

竞赛专用平台访问账号：jn+队伍号（如队号为 2020200001，则平台访问账号为 jn2020200001）

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**公共数据集**的方式，分享给所有参赛者，如图 1 所示，参赛者可以直接在公共数据集中查看并复制数据集到我的数据集中，然后在工程中配置“输入源”使用，无需上传数据。