教育平台的线上课程智能推荐策略

摘要

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。

该报告基于用户基本信息、学习详情及用户登录数据，分析平台用户的活跃情况，计算用户的流失率；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。

本文主要工作包括以下三个方面：

1. 基于所给数据集进行数据预处理，包括缺失值的填充与删除，去重，利用数据表之间的关联性处理特殊字段。
2. 利用 pyecharts 绘制各省份与城市的热力地图，利用 Seaborn 绘制工作日与非工作日各时段的用户登录次数柱状图，分析用户分布情况和活跃情况。基于 $\sigma i = T_{end} - T_i$ 计算用户流失率，为该教育平台的线上管理决策提供建议。
3. 根据以上分析，利用基于物品的协同过滤算法，给出线上课程的综合推荐策略。

目录

1.挖掘目标

本次数据挖掘的目标是利用“用户信息表”、“学习详情表”、“登录详情表”三个数据集，利用数据分析与挖掘的方法，数据可视化工具，基于物品的协同过滤算法具体要实现两个目标：

(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。

2.数据预处理

1.1 分析方法与过程

本任务旨在处理数据集的缺失值和重复值。

分别对三个信息表的数据缺失情况进行统计，得到以下结果：

数据集	shape	isNA_1	isNA_2
users	43983	user_id(67)	school(33409)
study_information	194974	price(4238)	\
login	387144	\	\

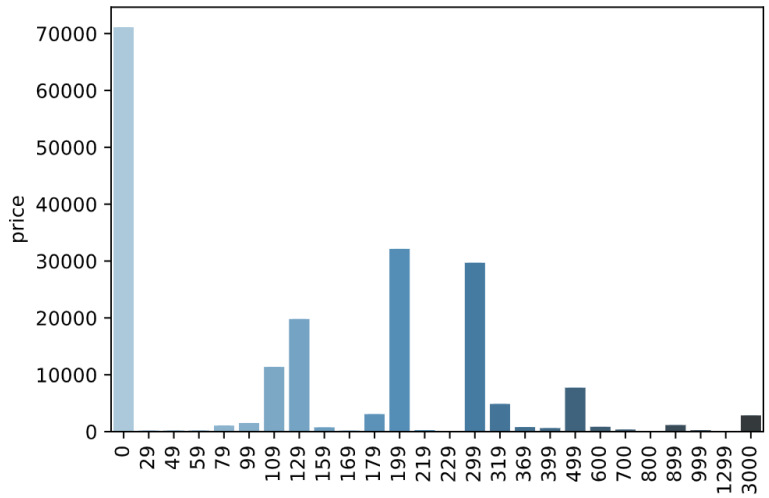
对于 users 表：

user_id：无法从其他数据中获得，且缺失数量不多，直接删除；

school：无法从其他数据中获得，缺失数量较多，用 NaN 填充；

对于 study_information 表：

price：利用 sb.barplot()函数，检测课程单价的分布情况，作出直方图如下：



根据图分析可得 :课程单价不符合正态分布 ,无法用均值填充 ;无法精准构建连续函数，插值法亦不合适。取价格分布最密集（即 price＝0）区间将缺失值填充为 0。

去重：

分别对三个信息表的数据重复情况进行统计 ,仅有 users 表中 user_id 列存在 6 条重复数据，去重即可。

1.2 分析方法与过程

本任务旨在对 users 表 recently_logged 字段的“--”值的进行数据处理。

recently_logged 字段中“--”的数据情况如下：

‘--‘字段数目	使用最后登录时间填充	使用注册时间填充
5375	84	5291

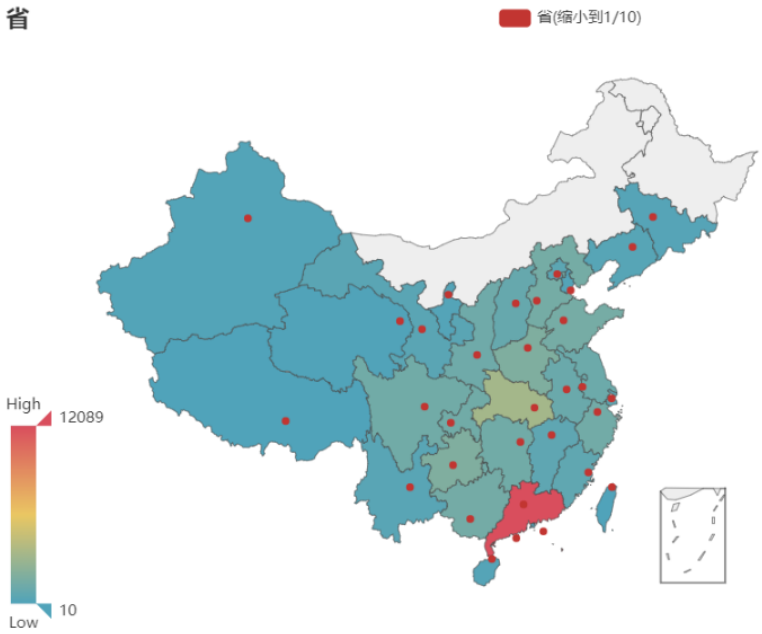
5375 条’--‘数据中，在 login 表“登录时间”列有数据记录的共计 84 条，使用最后登录时间填

充；其余没有记录的数据使用 users 表中注册时间填充。

3.平台用户活跃度分析

2.1 分析方法与过程

本任务旨在绘制各省份与各城市平台登录次数热力地图，并根据热力图分析用户分布情况。



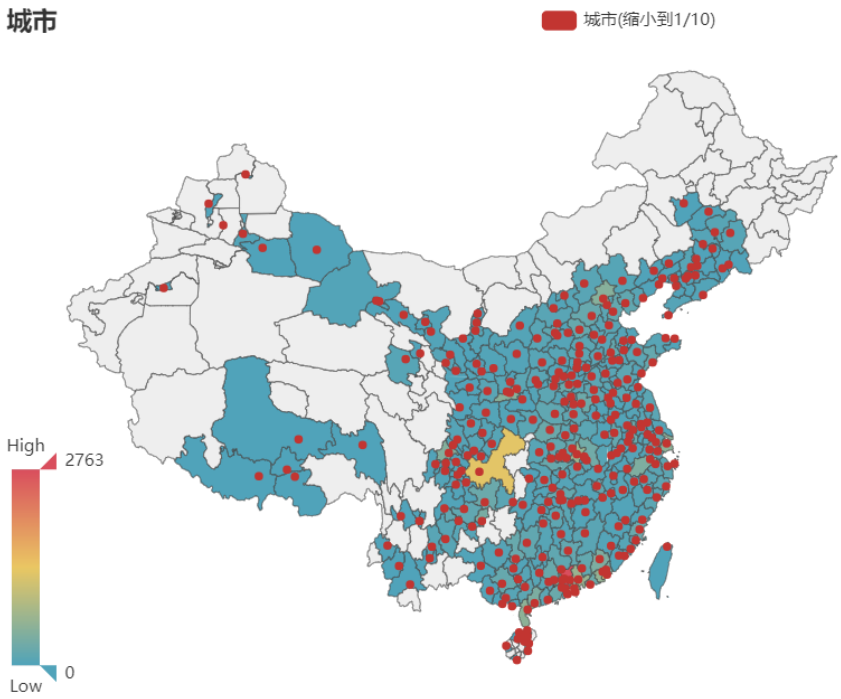
省份平台登录次数热力地图

根据各省份热力图，可以得出以下结论：

1.华南与华中地区用户更为密集，其中广东省占比 31.225%，湖北省占比 8.562%。

2.华西地区用户较为稀疏，其中内蒙古与黑龙江地区占比为 0。

3.根据整体省份密度分析，可以得出各省份之间的密度关系大致为：华南≥华中≥华东≥华北≥华西。



城市平台登录次数热力地图

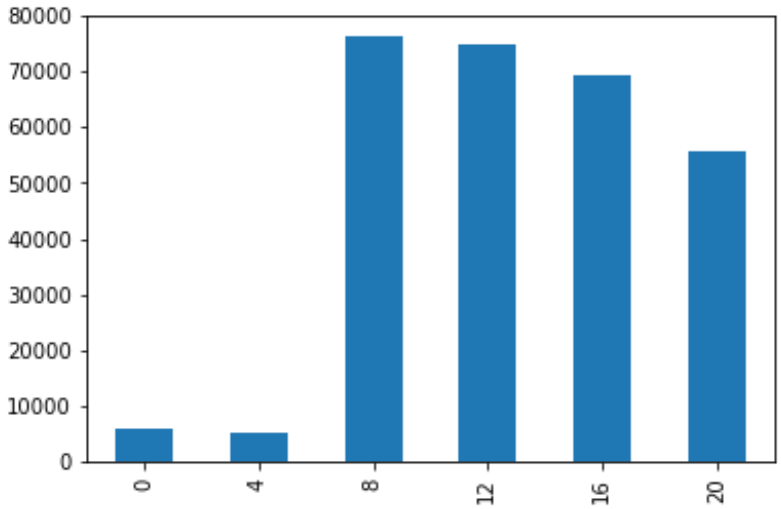
根据各城市热力地图，可以得出以下结论：

1.用户登录次数较频繁的城市为广州市、重庆市、汕头市，分别为 27626 次、13163 次、10146 次，各自占比 7.136%、3.400%、2.621%

2.用户登录次数较多的城市集中分布在地图的东南部，而西北部登录次数稀疏。

2.2 分析方法与过程

本任务旨在绘制工作日与非工作日各时段的用户登录次数柱状图，并根据柱状图分析用户活跃的主要时间段。

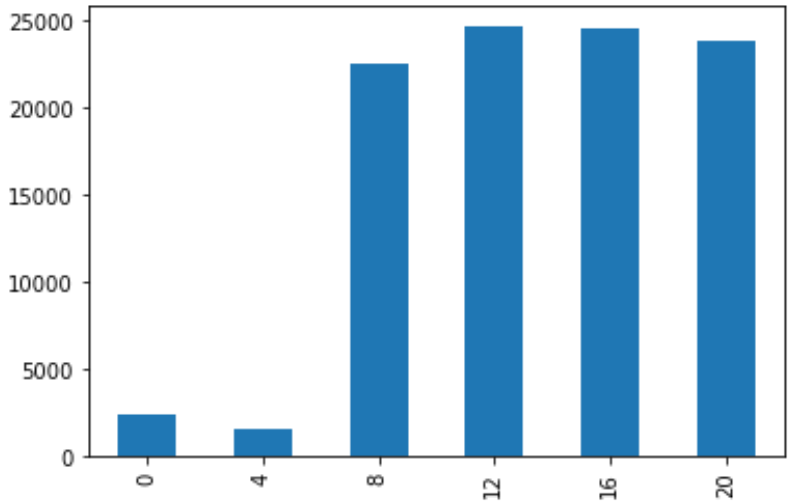


工作日用户登录柱状图

根据工作日各时段用户登录柱状图，分析可得：

1. 用户在 8：00-12：00 最为活跃，在 12：00-24：00 较为活跃，0：00-8：00 不活跃。

2. 活跃情况大致为：上午>下午>晚上。



非工作日用户登录柱状图

根据非工作日各时段用户登录柱状图，分析可得：

1. 用户在 12：00-16：00 最为活跃， 0：00-8：00 不活跃，其他时间较为活跃。

2. 活跃情况大致为：下午>晚上>上午。

2.2 分析方法与过程

本任务旨在计算用户流失率。

(1) 流失用户定义： T_{end}为数据观察窗口截止时间，T_i为用户 i 的最近访问时间，$\sigma i = T_{end} - T_i$ ，若σi>90 天，则称用户 i 为流失用户。

(2) **流失率定义**：流失用户/总用户 *100%

(3) **流失率计算结果**：58.406%，精度 0.001

2.3 分析方法与过程

本任务旨在通过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。

分析：

(1) **根据各省份/城市热力地图，可以得出用户的基本分布情况，用户密集地区与稀疏地区应采取不同的策略来进行管理。**

(2) **根据工作日/非工作日用户登录柱状图，可以适当调整管理维护人员的分配，提高管理效率，减少运营成本。**

(3) 基于$\sigma i = T_{end} - T_i$分别计算用户总流失率、加入班级的流失率以及未加入班级的流失率，可得未加入班级用户的流失率远大于加入班级用户的流失率，因此可以适当调整班级设置。相关计算结果如下图所示：

加入班级流失率	未加入班级流失率	总流失率
20.956%	85.748%	58.406%

建议：

(1) 用户密集地区可以定期监控用户流失数据，查明流失原因，从而改善用户体验并提高用户回头率；用户稀疏地区可以采取某些优惠措施，比如降低课程单价，发放物质奖励等吸引用户，且应提高宣传力度，扩大宣传范围，让教育平台知名度更高。

(2) 在用户活跃时段可以增加老师、管理维护人员的分配，提高互动率；在用户不活跃时段可以减少相关人员分配，以降低运营成本，提高管理效率。

(3) 可以为课程设置行政班级管理，增加课程周期数，提供签到等服务，增大对学生督促作用。

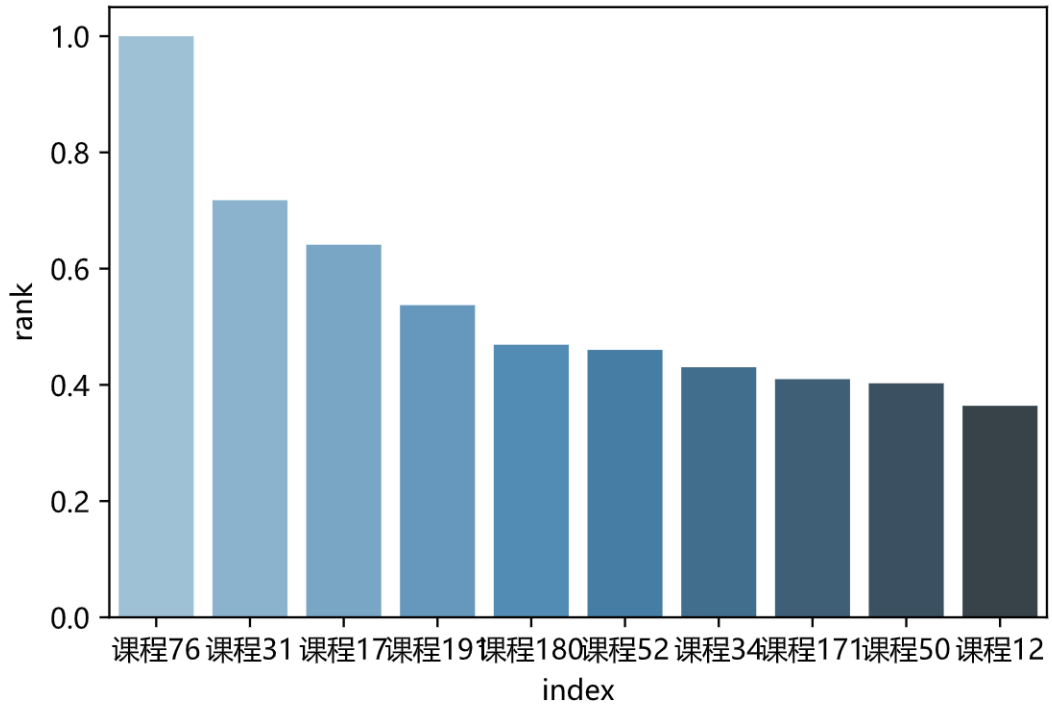
4. 线上课程推荐

3.1 分析方法与过程

该任务旨在通过 $\gamma = \frac{Q_1 - Q_{min}}{Q_{max} - Q_{min}}$ 公式来计算每门课的受欢迎程度，列出最受欢迎的前 10 门课程，并绘制相应的柱状图。

相关计算结果如下：

课程 id	参与人数	受欢迎程度
课程 76	13265	*1.000000*
课程 31	9521	*0.717732*
课程 17	8505	0.641134
课程 191	7126	0.537168
课程 180	6223	0.469089
课程 52	6105	0.460193
课程 34	5709	0.430338
课程 171	5437	0.409831
课程 50	5342	0.402669
课程 12	4829	0.363993



受欢迎程度前 10 的课程柱状图

3.2 分析方法与过程

该任务旨在对相关推荐算法进行描述，并给出总学习进度最高的 5 个用户的课程推荐数据。

ItemCF (基于物品的协同过滤算法)：

(1) 算法核心思想：给用户推荐那些和他们之前喜欢的物品相似的物品。

(2) 算法基本步骤：

i. 计算物品之间的相似度；

ii. 根据物品的相似度和用户的历史行为给用户生成推荐列表；

(3) 算法实现：

$$w_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|}$$

i. 计算物品之间的相似度；使用下面的公式：

ii. 这里，分母$|N(i)|$是喜欢物品 i 的用户数，而分子$|N(i) \cap N(j)|$是同时喜欢物品 i 和物品 j 的用户数。因此，上述公式可以理解为喜欢物品 i 的用户中有多少比例的用户也喜欢物品 j。但是却存在一个问题。如果物品 j 很热门，很多人都喜欢，那么w_{ij}就会很大，接近 1。因此,该公式会造成任何物品都会和热门的物品有很大的相似度,这对于致力于挖掘长尾信息的推荐系统来说显然不是一个好的特性。为了避免推荐出热门的物品，可以用下面的公式:

$$w_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)||N(j)|}}$$

iii. 这里由于还是 0-1 的原因，我们的余弦相似度可以写成上面的形式。但是，是不是每个用户的贡献都相同呢?因此，我们要对数据覆盖领域广而非出于自身兴趣的用户进行一定的惩罚。根据 John S. Breese 在论文 1 中提出的用户活跃度对数的倒数的参数 IUF(Inverse User Frequence)，活跃用户对物品相似度的贡献应该小于不活跃的用户，故应该增加 IUF 参数来修正物品相似度的计算公式:

$$w_{ij} = \frac{\sum_{u \in N(i) \cap N(j)} \frac{1}{\log 1 + |N(u)|}}{\sqrt{|N(i)||N(j)|}}$$

iv. 在得到物品之间的相似度后，ItemCF 通过如下公式计算用户 u 对一个物品 j 的兴趣:

$$p_{uj} = \sum_{i \in N(u) \cap S(j,K)} w_{ji} r_{ui}$$

这里 N(u)是用户喜欢的物品的集合，S(j,K)是和物品 j 最相似的 K 个物品的集合，w_{ji}是物品 j 和 i 的相似度，r_{ui}是用户 u 对物品 i 的兴趣。

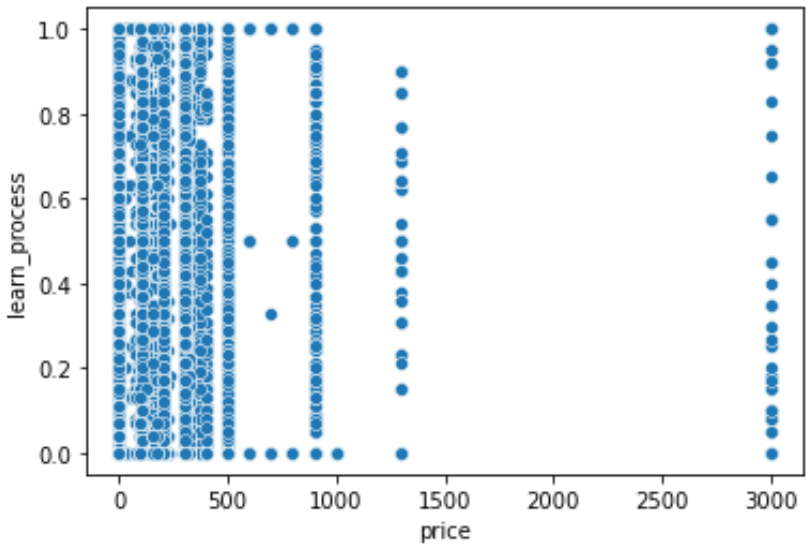
总学习进度最高的 5 个用户的课程推荐数据：

用户 id	课程进度	推荐课程 id	相似度（精度 0.001）
用户 1193	52.38	课程 97	2.871
		课程 99	2.672
		课程 163	2.330
用户 13841	40.42	课程 62	2.898
		课程 231	2.666
		课程 60	2.377
用户 32684	32.91	课程 40	2.674
		课程 62	1.799
		课程 96	1.633
用户 36989	29.60	课程 40	2.706
		课程 180	1.431
		课程 52	1.363
用户 24985	29.51	课程 52	1.597
		课程 180	1.263
		课程 51	0.800

3.3 分析方法与过程

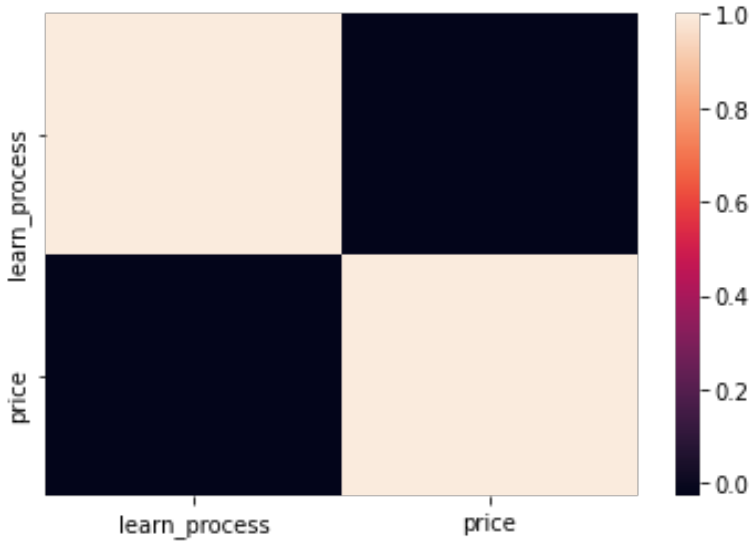
该任务旨在根据任务 3.1、任务 3.2 的结果并结合用户学习进度数据，分析付费课程与免费课程的差异，给出线上综合课程推荐策略。

绘制课程价格和学习进度散点图：



可以得出：付费课程与免费课程的价格分布离散，线性无关。

付费课程与免费课程相关图：



可以得出：课程进度相关系数基本为 0，即用户学习进度数据没有明显差异。

计算得出，付费课程与免费课程学习进度均值如下表：

课程分类	学习进度均值（精度 0.001）
付费课程	15.092

免费课程	13.772

线上综合课程推荐策略：

由于付费课程与免费课程不存在线性相关性，故在构建课程体系时应将两者结合，平均分配二者的课程数目，管理人员等。