2024 年（第 17 届）
中国大学生计算机设计大赛

大数据主题赛作品报告

作品编号：2024007056

作品名称：在线教育综合大数据行为分析系统

填写日期：2024 年 4 月 6 日

摘要

在21世纪的今天，随着5G技术和人工智能技术的蓬勃发展，互联网与通信技术不断取得新的突破，学习资源的共享与构建正在迅猛地向前推进。特别是在2024年，线上教育平台如同雨后春笋般涌现，为广大学习者提供了丰富多样的学习资源和选择。

进入2024年，人们对于线上学习的热情和需求并未减退。特别是在某些特殊时期或情况下，线上教育平台更是成为了教育领域不可或缺的重要力量。因此，如何充分利用这些平台所积累的用户数据，精准地把握用户的课程偏好，并据此进行远程课程的个性化推荐，成为了线上教育领域亟待解决的问题。

在这样的时代背景下，运用先进的数据分析技术，对教育平台的线上信息和用户学习行为进行深入挖掘和研究，变得尤为关键。这不仅有助于提升学习者的学习体验，还能帮助教育平台实现资源的优化配置和运营的高效化。因此，数据分析技术在当前线上教育领域的应用不仅具有深远的意义，而且具备广阔的发展前景。

针对任务一：判断所给数据集是否存在缺失值、异常值、重复值等方面的问题，处理首先利用python对数据进行初步的描述性统计分析，使用pandas中的info函数进行缺失值计数，对数据缺失情况进行检查，然后根据出现的不同情况采取不同的处理方式。而后根据数据分析的需要对处理好的数据进行提取。

针对任务二：首先，根据日期信息和国务院各年份法定节假日文件，对工作日和非工作日进行区分处理，统计24小时各个时段的用

户活跃度、每个星期的用户活跃度、每个月的用户活跃度并进行具体整体分析，绘制相应的柱形图实现对该教育平台用户进行活跃度分析。

针对任务三：根据任务二的活跃度数据，进行进一步分析，通过构建时间序列模型，预测用户活跃度情况。

针对任务四：首先，根据注册时间和最近一次更新时间得出时间差值，并将用户按时间差值分类，得出用户流失的数据，绘制分析图，分析用户在什么时间段流失值最大，然后分析用户留存率，最终结合上述用户活跃度和用户留存率分析，为该教育平台提供线上管理决策建议。

针对任务五：首先，分别计算出用户的学习时间、课程数、时间差值，构建用户行为分类模型，将用户的行为分类，根据分类的结果进行分析，为教育平台对保持活跃度、提高留存率提供策略

我们基于 Echarts 技术制作了可视化大屏。

网站链接：https://nnn45.gitee.io/education-data/

目 录

1. 引言

在线教育作为教育体系的重要补充，其教学质量和用户体验的提升至关重要。然而，随着平台数据的不断积累，如何有效地处理和分析这些数据，挖掘其中的价值，为平台决策提供支持，成为了一个亟待解决的问题。通过使用 Python 作为一种强大的数据分析语言，结合 Web 技术展示，本系统将两者相结合为教育平台提供一个高效、直观的数据展示界面，为后续提供线上管理决策建议打下坚实的基础。

2. 问题分析

1. 分析用户平台活跃度、构建 Prophet 模型预测活跃度趋势
2. 分析用户留存率情况，找到影响用户留存率的原因
3. 构建用户行为分类模型，为教育平台对保持活跃度、提高留存率提供策略

3. 任务一

本项目数据集采用赛事主办方提供的数据集，具体内容如下：

序号	表名	数据量	说明
1	course_chapter	不少于 310000	课程章节表，包含课程 ID 和章节内容等信息
3	classroom_member	不少于 110000	班级成员表, 包含班级 ID 和用户 ID 等信息
4	classroom_courses	不少于 9000	班级课程表, 包含班级 ID 和课程 ID 等信息
5	log	不少于 1000000	日志表，包括日志操作人 ID 和登录日志等信息
6	user_learn_statistics_total	不少于 140000	用户学习统计表，包含用户 ID 和学习相关统计等信息
9	testpaper_result	不少于 340000	试卷结果表，包括试卷名称、用户 ID 和试卷分数等信息

3.1. 判断数据情况

考虑数据量比较多的情况，使用 Python 中的 pandas 对数据进行初步的描述性统计分析。使用 pandas 中的 info 函数进行统计。

如图所示班级课程表、日志表、用户学习统计表均不存在数据缺失的问题，班级成员表中的 lastLearnTime、learnedNum 字段存在缺失情况，考虑值允许为空的情况，留待分析情况中具体对待；试卷结果表中的 teacherSay 字段存在缺失情况，考虑该字段对数据缺失对数据暂无影响，不做处理

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9424 entries, 0 to 9423
Data columns (total 7 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              9424 non-null   int64
 1   classroomId     9424 non-null   int64
 2   courseId        9424 non-null   int64
 3   parentCourseId  9424 non-null   int64
 4   seq             9424 non-null   int64
 5   disabled        9424 non-null   int64
 6   courseSetId     9424 non-null   int64
dtypes: int64(7)
memory usage: 515.5 KB
```

图 3.1 班级课程表数据表信息描述

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1741719 entries, 0 to 1741718
Data columns (total 10 columns):
 #   Column           Dtype
---  ------           -----
 0   id               int64
 1   userId           int64
 2   module           object
 3   action           object
 4   data             object
 5   browser          object
 6   operatingSystem  object
 7   device           object
 8   createdTime      int64
 9   level            object
dtypes: int64(3), object(7)
memory usage: 132.9+ MB
```

图 3.2 日志表数据表信息描述

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142251 entries, 0 to 142250
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   id                  142251 non-null  int64
 1   userId              142251 non-null  int64
 2   joinedClassroomNum  142251 non-null  int64
 3   joinedCourseSetNum  142251 non-null  int64
 4   joinedCourseNum     142251 non-null  int64
 5   exitClassroomNum    142251 non-null  int64
 6   exitCourseSetNum    142251 non-null  int64
 7   exitCourseNum       142251 non-null  int64
 8   learnedSeconds      142251 non-null  int64
 9   finishedTaskNum     142251 non-null  int64
 10  createdTime         142251 non-null  int64
 11  updatedTime         142251 non-null  int64
dtypes: int64(12)
memory usage: 13.0 MB
```

图 3.3 用户学习统计表信息描述

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 111024 entries, 0 to 111023
Data columns (total 16 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   id                111024 non-null  int64
 1   classroomId       111024 non-null  int64
 2   userId            111024 non-null  int64
 3   orderId           111024 non-null  int64
 4   levelId           111024 non-null  int64
 5   noteNum           111024 non-null  int64
 6   threadNum         111024 non-null  int64
 7   locked            111024 non-null  int64
 8   role              111024 non-null  object
 9   createdTime       111024 non-null  int64
 10  lastLearnTime     61984 non-null   float64
 11  learnedNum        61984 non-null   float64
 12  updatedTime       111024 non-null  int64
 13  deadline          111024 non-null  int64
 14  refundDeadline    111024 non-null  int64
 15  deadlineNotified  111024 non-null  int64
dtypes: float64(2), int64(13), object(1)
memory usage: 13.6+ MB
```

图 3.4 班级成员表信息描述

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 349036 entries, 0 to 349035
Data columns (total 23 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   id               349036 non-null  int64
 1   paperName        349036 non-null  object
 2   testId           349036 non-null  int64
 3   userId           349036 non-null  int64
 4   courseId         349036 non-null  int64
 5   lessonId         349036 non-null  int64
 6   score            349036 non-null  float64
 7   objectiveScore   349036 non-null  float64
 8   subjectiveScore  349036 non-null  float64
 9   teacherSay       2070 non-null    object
 10  rightItemCount   349036 non-null  int64
 11  passedStatus     349036 non-null  object
 12  limitedTime      349036 non-null  int64
 13  beginTime        349036 non-null  int64
 14  endTime          349036 non-null  int64
 15  updateTime       349036 non-null  int64
 16  metas            182945 non-null  object
 17  status           349036 non-null  object
 18  checkTeacherId   349036 non-null  int64
 19  checkedTime      349036 non-null  int64
 20  usedTime         349036 non-null  int64
 21  type             349036 non-null  object
 22  courseSetId      349036 non-null  int64
dtypes: float64(3), int64(14), object(6)
memory usage: 61.2+ MB
```

图 3.5 试卷结果表信息描述

4. 任务二

4.1. 用户数据预处理

4.1.1. 日期数据预处理

在对用户的活跃度进行分析前需要对数据进行处理提取日志表的 userId、createdTime、operatingSystem、device 值组成新表，并将 createdTime 时间戳转换为日期格式，而后根据转换出的格式增加新的字段日期、日期编码、月份、时间段编码、登录时间。

基于如上操作得到表如图 4.1 所示

	userId	createdTime	operatingSystem	device	日期	日期编码	月份	时间段编码	登录时间段
0	142708	2023-01-09 02:46:55	Windows NT	computer	2023-01-09	0	1	2	02:46:55
1	142708	2023-01-09 02:46:56	Windows NT	computer	2023-01-09	0	1	2	02:46:56
2	57937	2023-01-09 02:47:14	Windows NT	computer	2023-01-09	0	1	2	02:47:14
3	57937	2023-01-09 02:47:15	Windows NT	computer	2023-01-09	0	1	2	02:47:15
4	9464	2023-01-09 02:47:46	Windows 98	computer	2023-01-09	0	1	2	02:47:46
...	...	...	...	...	...	...	...	...	...
1741714	147723	2023-07-26 16:27:04	Windows NT	computer	2023-07-26	2	7	16	16:27:04
1741715	147723	2023-07-26 16:27:04	Windows NT	computer	2023-07-26	2	7	16	16:27:04
1741716	147723	2023-07-26 16:27:04	Windows NT	computer	2023-07-26	2	7	16	16:27:04
1741717	147723	2023-07-26 16:27:04	Windows NT	computer	2023-07-26	2	7	16	16:27:04
1741718	147723	2023-07-26 16:27:05	Windows NT	computer	2023-07-26	2	7	16	16:27:05

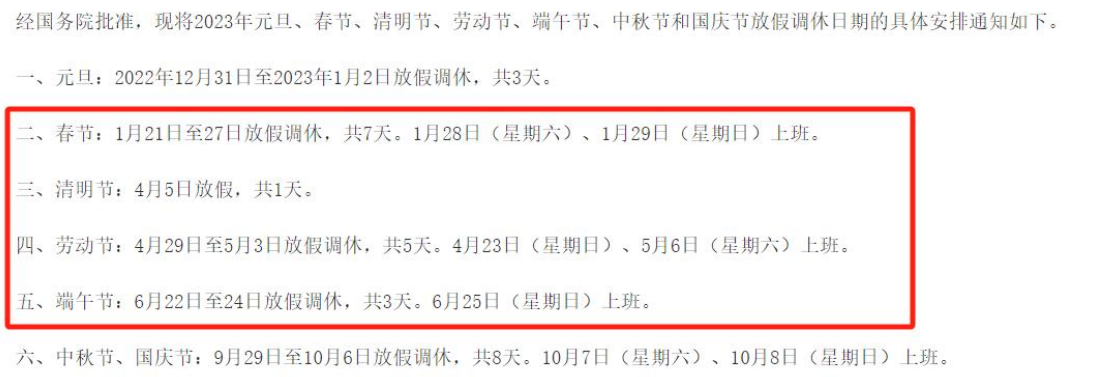
1741719 rows × 9 columns

图 4.1Login_data 表信息描述

4.1.2. 节假日及是否工作日处理

本文将星期一至星期五定义为工作日，星期六和星期天定义为非工作日。考虑到国家法定假日的调休日期，在日期处理过程中根据 2023 年的节假日安排进行调整，以确保工作日日期的准确性（如图 4.2 所示），而后根据分类情况添加新的字段用以区分。

统计工作日与非工作日各个时间段内用户登录总频次记为活跃度情况，同时统计每周、每月的活跃度情况


经国务院批准，现将2023年元旦、春节、清明节、劳动节、端午节、中秋节和国庆节放假调休日期的具体安排通知如下。

一、元旦：2022年12月31日至2023年1月2日放假调休，共3天。

二、春节：1月21日至27日放假调休，共7天。1月28日（星期六）、1月29日（星期日）上班。

三、清明节：4月5日放假，共1天。

四、劳动节：4月29日至5月3日放假调休，共5天。4月23日（星期日）、5月6日（星期六）上班。

五、端午节：6月22日至24日放假调休，共3天。6月25日（星期日）上班。

六、中秋节、国庆节：9月29日至10月6日放假调休，共8天。10月7日（星期六）、10月8日（星期日）上班。

图 4.2 2023 年节假日安排

4.2. 用户活跃度分析

4.2.1. 工作日与非工作日 24 时活跃度分布图分析

从工作日与非工作日 24 时活跃度分布图可以看出用户在工作日活跃度大于同时间段的非工作日活跃度。在工作日期间有三个峰值，0 点的活跃度最高为

100997，0 点至 4 点逐步下降，4 点时活跃度是一个低值为 36532，4 点至 8 点逐步上升，8 点活跃度为另一个峰值为 92607，6 点至 8 点活跃度差值不多，8 点至 12 点先下降而后突然达到又一个峰值为 92973，12 点至 19 点逐步下降，19 点为活跃度最低值为 2689，19 点至 23 点总体呈上升趋势。在非工作日期间总体活跃度不高，峰值位于 7 点为 40153，最低值位于 18 点为 1303。（如图 4.3 所示）

分析原因：用户喜欢在工作日时进行学习，且时间段集中于 0 点、6 点至 8 点、12 点，原因可能是喜欢利用空余时间在平台进行学习：0 点是熬夜学习、6 点至 8 点为晨读学习、12 点处于刚下课吃饭时间的午休学习；15 点至 20 点活跃度普遍低，分析原因用户可能喜欢与这个时间段休息放松、进行娱乐活动

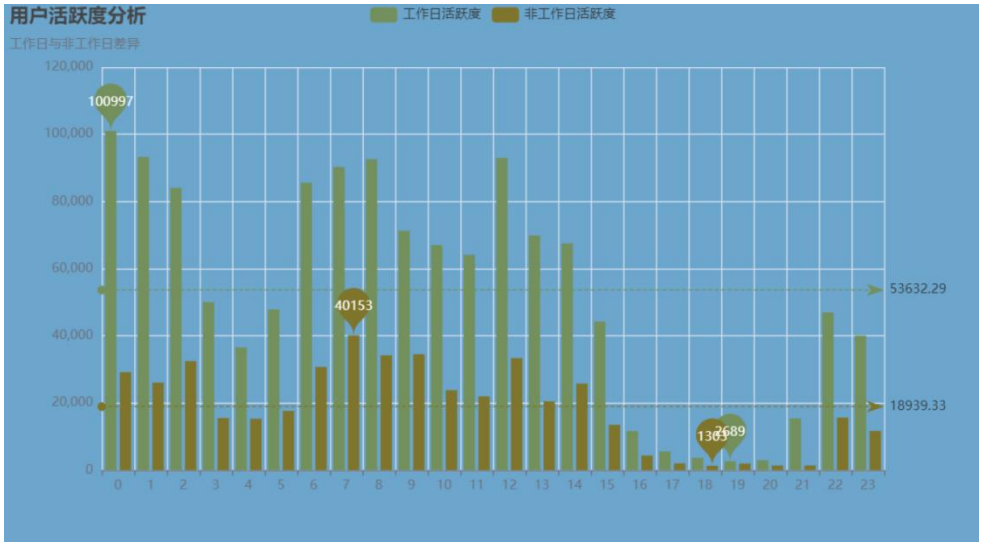


图 4.3 工作日与非工作日 24 时活跃度分布图

4.2.2. 每周活跃度分布图

从一周的活跃度看，总体差距不大，峰值位于周二为 285991，最低值为周天为 225851，

分析原因：说明用户的学习行为与今天为星期几的相关性较小

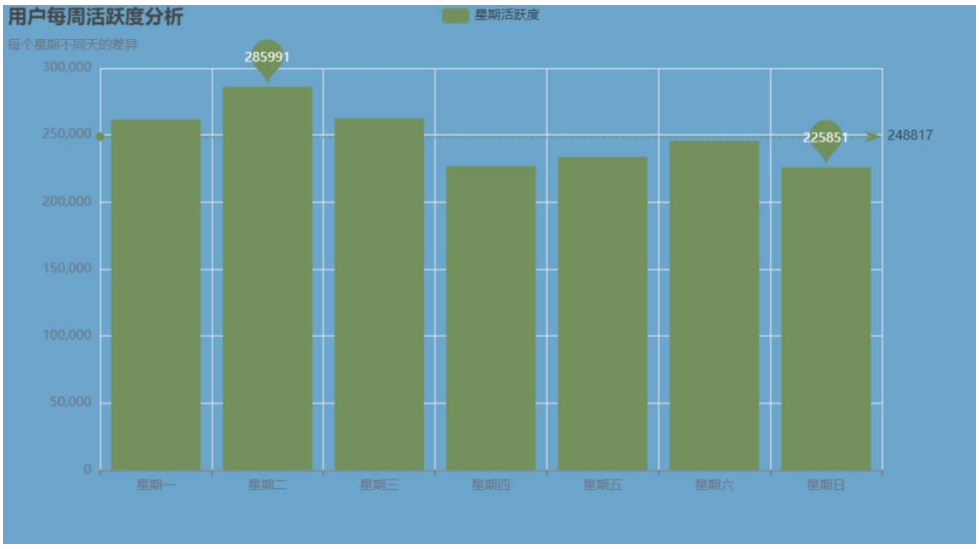


图 4.4 每周活跃度分布图

4.2.3. 每月活跃度分布图

从一月至七月的活跃度分布看，峰值位于四月为 476966，最低值位于二月为 122473，一月与七月活跃度相近，三月、五月、六月活跃度相近。

分析原因：一月、二月、七月基本处于寒暑假期间用户学习兴趣不高，三月至六月处于一个学年的下半学期用户活跃度较高，其中四月值最高，可以是因为教学活动或者比赛在平台学习等原因造成

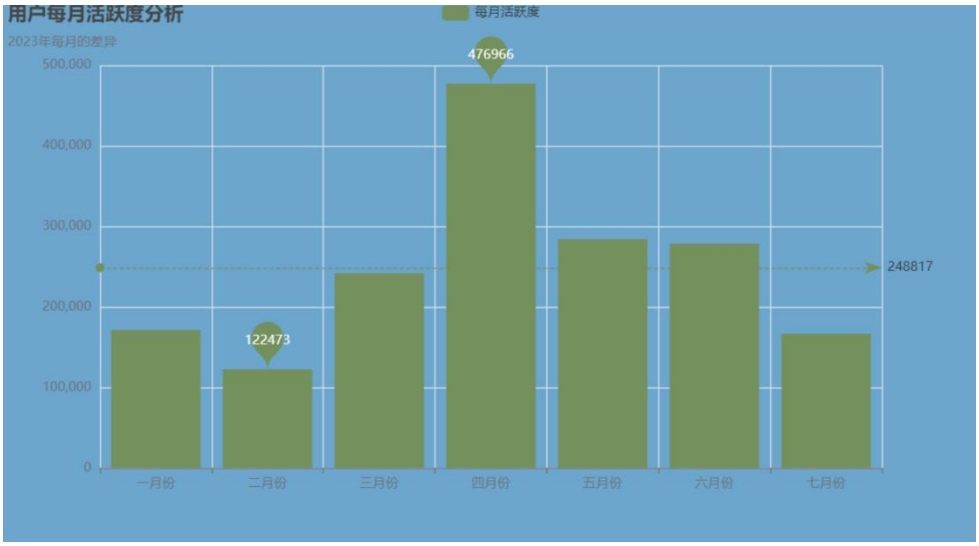


图 4.5 每月活跃度分布图

5. 任务三

5.1. Propht 模型概述

Prophet 模型内部由循环中的分析师与自动化两部分构成一个循环体系。Prophet 的预测过程是根据预测问题建立时间序列模型，对历史数据进行仿真，评估模型的效果，根据出现的问题，进一步进行调整和建模，最终以可视化方式反馈整个预测结果。

Prophet 模型是一个加法回归模型，它由三个核心部分 trend（趋势项）、seasonality（季节项）及 holidays（假期项）构成。

Prophet 模型基本组成公式：y（t）=g（t）+s（t）+h（t）+ ε t

在时间序列分析领域，有一种常见的分析方法叫做时间序列的分解（Decomposition of Time Series），它把时间序列分成几个部分，分别是季节项，趋势项，剩余项。

加法形式：$$y_t = S_t + T_t + R_t.$$

乘法形式：$$y_t = S_t \times T_t \times R_t.$$

Prophet 基于这种方法进行了必要的改进和优化。在实际生活和生产环节中，除了季节项，趋势项，剩余项之外，通常还有节假日的效应。所以，在 Prophet 算法中，同时考虑了以上四项，即：

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t.$$

g(t)表示趋势项，表示时间序列在非周期上面的变化趋势；

s(t)表示周期项，或者称为季节项，一般以周者年为单位；

h(t)表示节假日项，表示在当天是否存在节假日；

ϵ_t表示误差项或者称为剩余项；

Prophet 算法就是通过拟合这几项，最后把它们累加起来就得到时间序列的预测值。

5.2. 用户活跃度预测——基于 Propht 模型

基于任务二的数据处理，将一月到七月份每天的活跃度也统计出来（如图 5.1 所示）

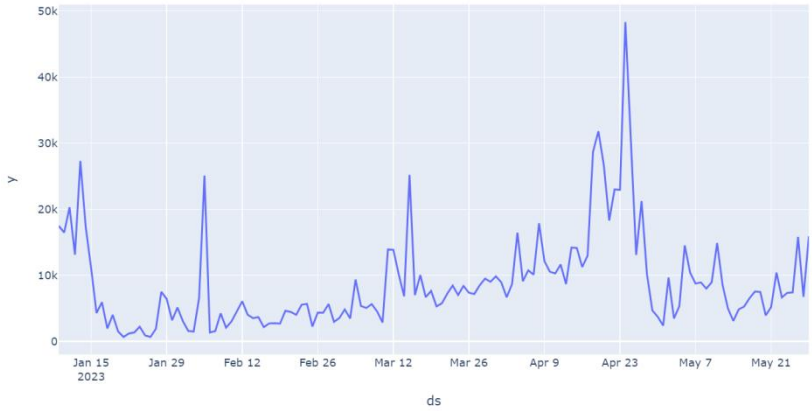


图 5.1 每日活跃度分布图

划分数据，划分为训练集和验证集，将前 140 行的数据作为训练集，后 59 行的数据作为测试集。

实例化 Prophet 对象，并且通过 fit 来训练模型，定义节假日数据，数据的变动会受到季节、周、天的影响，因此我们将这三个参数设置为 True （如图 5.2 所示）

```
# 定义一些节假日数据
holidays = pd.DataFrame({
    'holiday': 'some_holiday',
    'ds': pd.to_datetime(['2023-01-21', '2023-01-22', '2023-01-23', '2023-01-24', '2023-01-25', '2023-01-26'
                          , '2023-01-27', '2023-04-05', '2023-04-29', '2023-04-30', '2023-05-01', '2023-05-02', '2023-05-03'
                          , '2023-06-22', '2023-06-23', '2023-06-24']),
    'lower_window': 0,
    'upper_window': 1,
})

# 数据的变动会受到季节、周、天的影响，因此我们将这三个参数设置为True，并添加自定义的节假日数据
m = Prophet(yearly_seasonality=True, weekly_seasonality=True, daily_seasonality=True, holidays=holidays)
```

图 5.2 节假日数据

prophet 可以计算出 yhat，yhat_lower，yhat_upper，分别表示时间序列的预测值，预测值的下界，预测值的上界（如图 5.3 所示）

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	additive_terms	additive_terms_lower	additive_terms_upper	daily	...	weekly	we
0	2023-01-09	16507.189	11494.243	24771.680	16507.189	16507.189	1925.511	1925.511	1925.511	-3755.098	...	-169.556	
1	2023-01-10	16493.860	11771.830	24618.788	16493.860	16493.860	1490.463	1490.463	1490.463	-3755.098	...	458.444	
2	2023-01-11	16480.532	10615.893	23829.211	16480.532	16480.532	523.447	523.447	523.447	-3755.098	...	579.126	
3	2023-01-12	16467.204	8076.944	20943.576	16467.204	16467.204	-2024.052	-2024.052	-2024.052	-3755.098	...	-873.632	
4	2023-01-13	16453.875	7390.856	20737.306	16453.875	16453.875	-2494.799	-2494.799	-2494.799	-3755.098	...	-259.689	
...	...	...	...	...	...	...	...	...	...	...	...	...	
165	2023-06-23	14308.006	-4680.177	7918.194	14308.006	14308.007	-12445.688	-12445.688	-12445.688	-3755.098	...	-259.689	
166	2023-06-24	14294.678	-4310.650	8556.305	14294.678	14294.678	-12662.928	-12662.928	-12662.928	-3755.098	...	-304.842	
167	2023-06-25	14281.350	2924.467	16109.015	14281.349	14281.350	-4779.232	-4779.232	-4779.232	-3755.098	...	570.149	
168	2023-06-26	14268.021	3142.947	16269.325	14268.021	14268.021	-4698.745	-4698.745	-4698.745	-3755.098	...	-169.556	
169	2023-06-27	14254.693	3489.297	16703.913	14254.692	14254.693	-4103.540	-4103.540	-4103.540	-3755.098	...	458.444	

图 5.3 预测模型数据

对数据进行可视化操作，黑点表示真实数据，蓝线表示预测结果。蓝色区域表示一定置信程度下的预测上限和下限。（如图 5.4 所示）

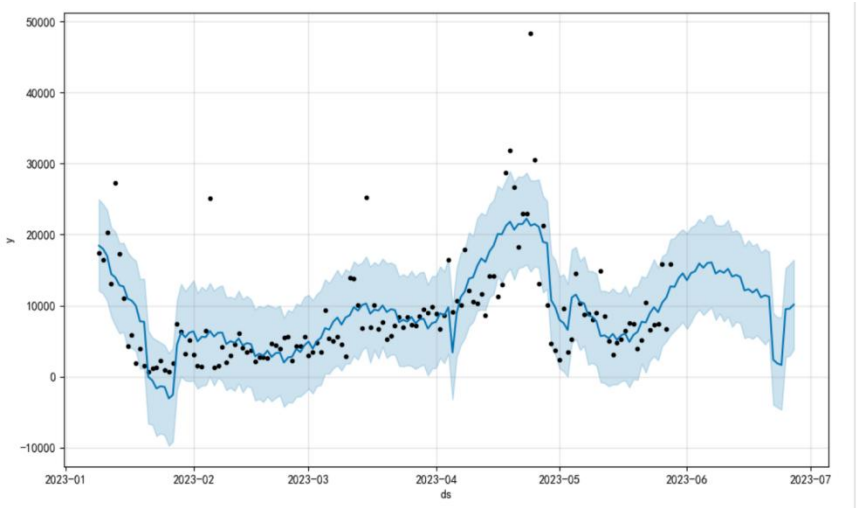


图 5.4 预测模型数据图

将预测的七月份的值和真实七月份的值做对比（如图 5.5 所示）

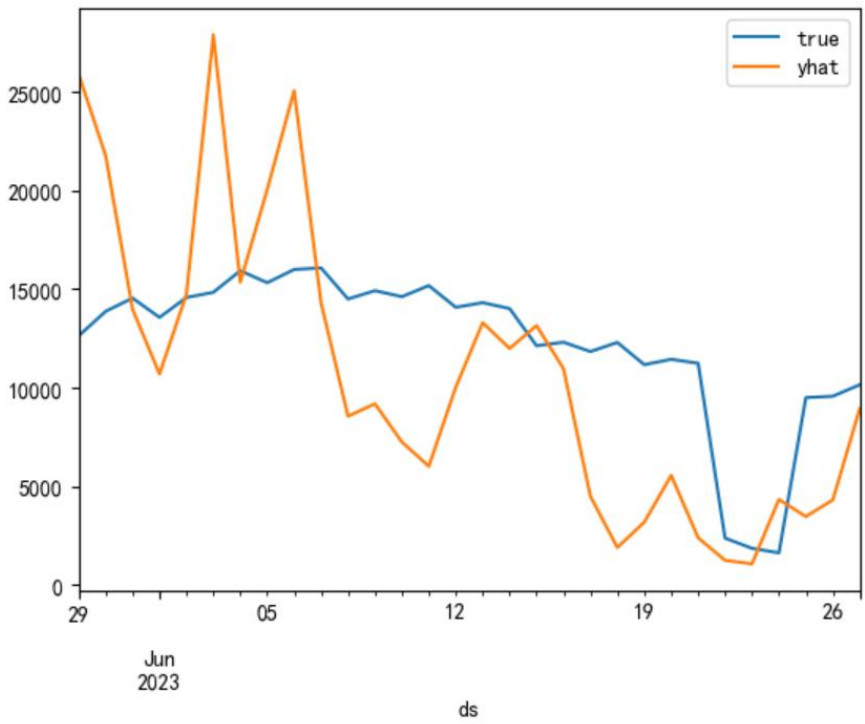


图 5.5 预测模型数据

从模型上看预测的准确率总体趋势符合走向

6. 任务四

6.1. 用户流失数据预处理

6.1.1. 数据处理

计算时间差值提取用户学习统计表中的 userId、createdTime、updatedTime、joinedClassroomNum、exitClassroomNum、learnedSeconds、finishedTaskNum 字段，将 createdTime、updatedTime 的时间戳格式转为日期格式，

6.1.2. 计算时间差值

选取更新时间最大值明确采集时间为 2023-07-26 03:00:05，防止出现异常情况选取采集时间为 2023-8-01 00:00:00，获得用户时间差值的描述信息：时间差值 = 样本采样时间（2023-8-01 00:00:00） - 用户最后一次登录时间，得出

的时间差值作为新增字段加入表中，对时间差值使用 pandas 库中的 describe 函数，统计包括计数、平均值、标准差、最小值、25%、50%（中位数）、75% 分位数和最大值。

根据时间差值将用户分为三类 0-60 为活跃用户、60-90 为潜在用户、90-1999 为流失用户，并将数据绘成直方图。

而后计算流失率：流失率=流失用户/总用户数，并将数据展示出来

6.2. 用户流失数据分析

6.2.1. 用户流失率分析

从用户分类数据得出流失用户为 133814、活跃用户为 6445、潜在用户为 1991（如图 6.1 所示），得出流失率为百分之九十四（如图 6.2 所示），观察时间差值大于 90 天的流失用户图，得出在位于 572 天时流失人数最多为 12869

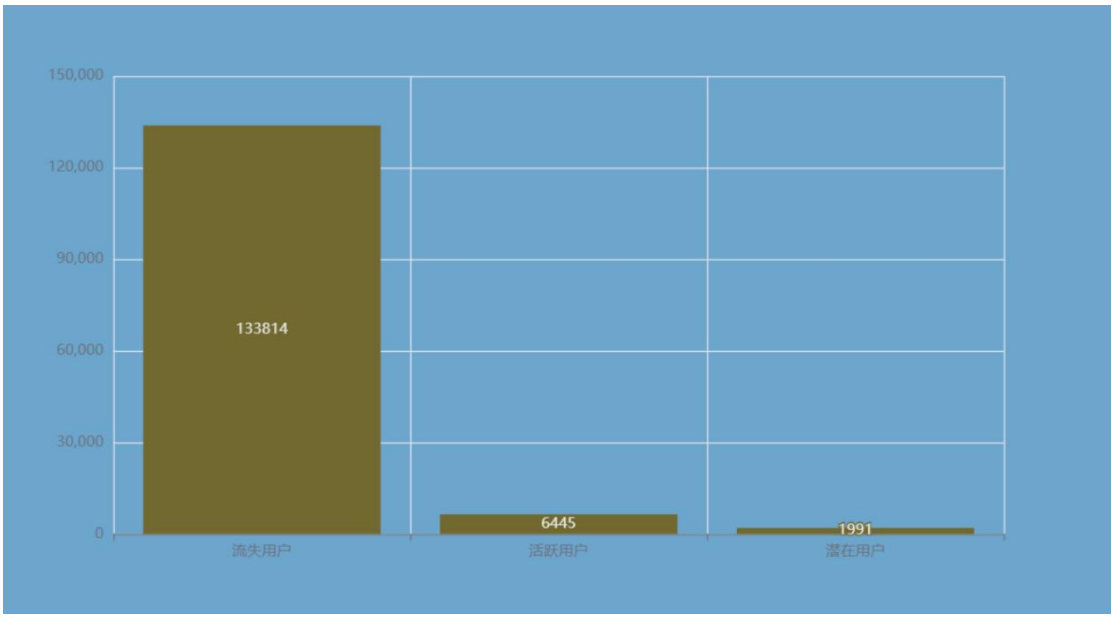


图 6.1 用户类型图

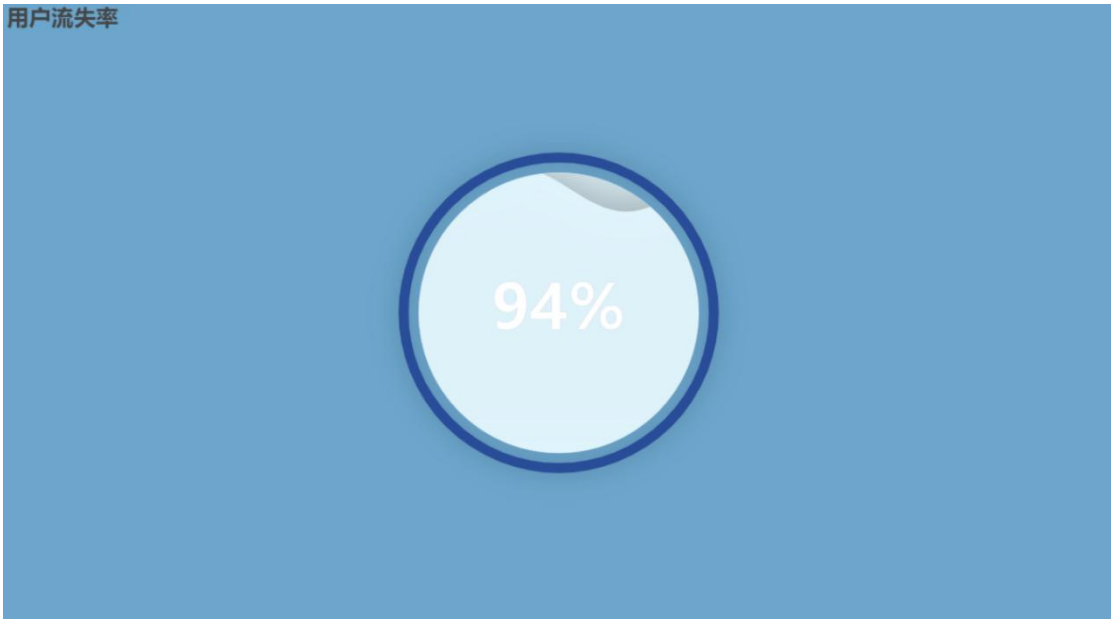


图 6.2 用户流失率图

分析原因：平台的用户流失情况非常严重高达 94%，说明平台对用户的吸引力较低，缺乏能够长时间维持用户登录的产品。将时间差值大于 90 天的人数用图表展示出来（如图 6.3 所示），在 572 天时平台流失人数最大，通过查找得知 2023 年 8 月 1 日前 572 天为 2022 年 1 月 6 日，说明在 2022 年 1 月 6 日就已经达到了 90 天以上没登录，再将时间提前 90 天为 2021 年 10 月份，可能造成原因是当时处于疫情封闭期间，用户在 10 月份前注册平台账户进行学习，因为某种原因导致之后不再登录

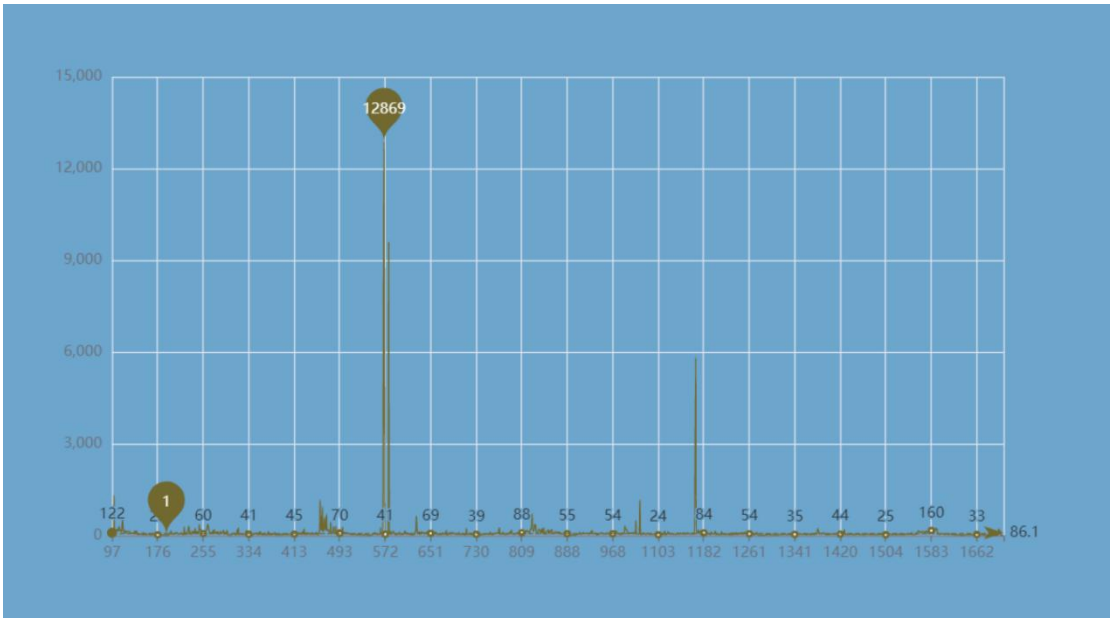


图 6.3 时间差值大于 90 天流失用户展示图

6.2.2. 用户学习情况分析

基于上面的分析再对表中得用户的学习时长进行辅助分析，筛选出时间不为0的用户，再计算出注册天数：注册天数=用最近一次登录时间—创建时间，而后用 learnedSeconds（学习时长）/注册天数，得出平均每天的学习时间，使用pandas 库中的 describe 函数，统计包括计数、平均值、标准差、最小值、25%、50%（中位数）、75% 分位数和最大值，再将学习时间按时长分类 0-200 学习很少、200-1000 学习较少、1000-35000 学习较多、35000-350000 学习很多，绘制直方图（如图 6.4 所示）

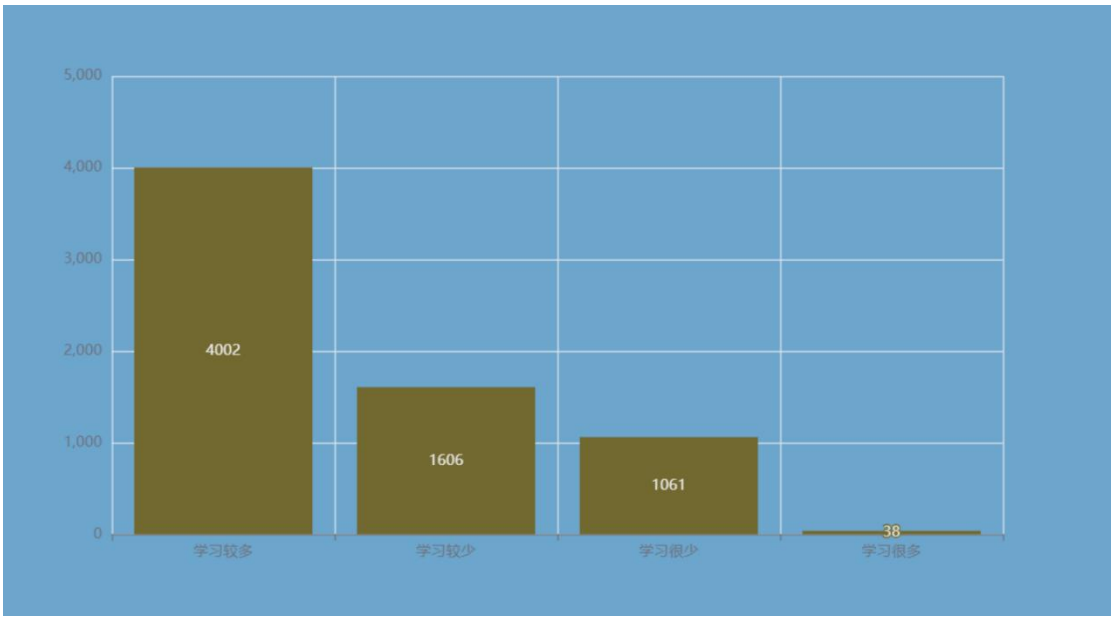


图 6.4 用户平均学习时间图

用户每天的平均学习时间中，学习时间较多的占大部分，但结合高流失率，说明用户是在平台学习到了东西，因为某种原因不再登录。

6.3. 用户留存率分析

6.3.1. 数据预处理

首先以 2023 年 7 月份的登录数据为例，创建新增字段用以表示用户是否在当日，一周，两周，三周，四周内登录过，大于等于 0 为当日、大于等于 1 为第一周、大于等于 8 为第二周、大于等于 15 为第三周、大于等于 22 为第四周，统

计每个时期的用户数，而后计算留存率：留存率=登录用户数/总登录数，所计算数如图 6.5 所示

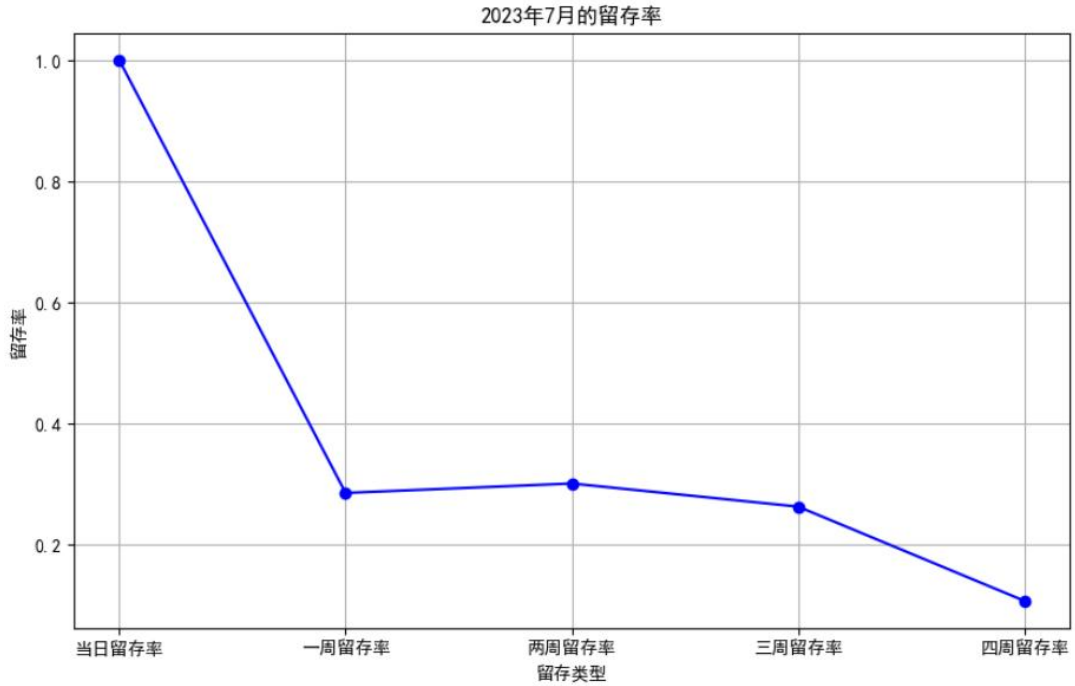


图 6.5 用户 7 月留存率

6.3.2. 留存率分析

从上面的图可以看出用户再刚注册到第一周的留存率下滑的非常厉害，之后二三周则比较稳定，第四周再下跌到最低，结合平均学习时长可以看出，用户不能够长时间学习，基本上学习时间不会超过一周，结合上面的流失率来看，分析用户可能倾向于在短时间内的高强度学习，而后就不再使用平台了，可能是平台对用户的粘性不足，没有足够吸引用户长时间登录的产品。

7. 任务五

7.1. RFM 模型概述

RFM 模型是衡量客户价值和客户创利能力的重要工具和手段。在众多的客户关系管理(CRM)的分析模式中，RFM 模型是被广泛提到的。该机械模型通过一个客户的近期购买行为、购买的总体频率以及花了多少钱 3 项指标来描述该客户的价值状况。

R (Recency)：指用户的最近一次消费时间，简单来说就是用户最后一次下单时间距今天有多长时间了，这个指标与用户流失和复购直接相关。

F(Frequency)：指用户下单频率，简单来说就是用户在固定的时间段内消费了几次。这个指标反映了用户的消费活跃度。

M(Monetary)：指用户消费金额，简单来说就是用户在固定的周期内在平台上花了多少钱，直接反映了用户对公司贡献的价值。

通过 RFM 分析可以把这 3 个指标按价值从低到高排序，并把这 3 个指标作为坐标轴，就可以把空间分为 8 部分，对应 8 类用户：一般保持客户、一般发展客户、一般价值客户、一般挽留客户、重要保持客户、重要发展客户、重要价值客户、重要挽留客户等八个级别。（如图 7.1、7.2 所示）

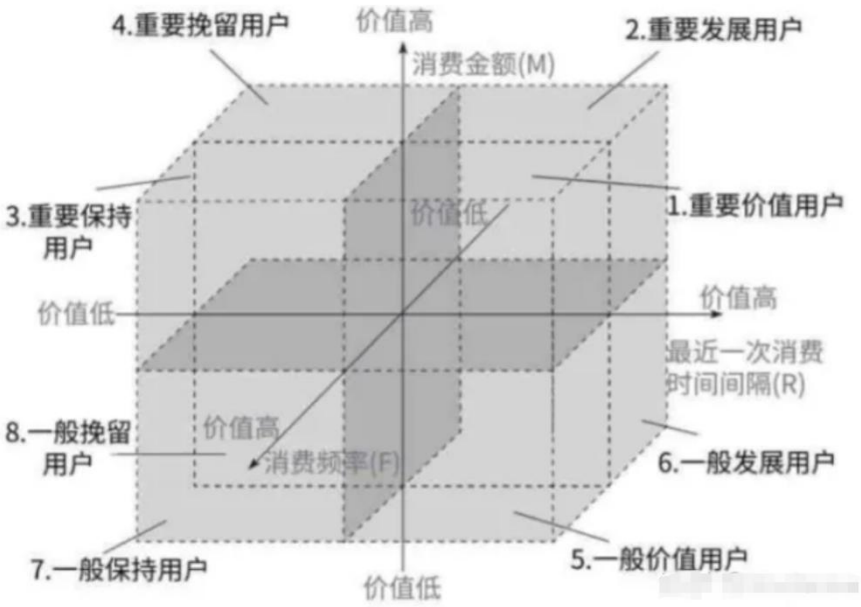


图 7.1 RFM 模型图

类别	R	F	M	运营策略
重要价值客户	高	高	高	保持好现状
重要发展客户	高	低	高	刺激消费频率
重要保持客户	低	高	高	要留住这个客户
重要挽留客户	低	低	高	要留住他并且刺激他的消费频率
一般价值客户	高	高	低	刺激他消费力度
一般发展客户	高	低	低	刺激他消费频率和力度
一般保持客户	低	高	低	要留住他并且刺激他的消费力度
一般挽留客户	低	低	低	要各方面进行刺激

图 7.2 RFM 的用户分类规则

7.2. 学习行为客户分类——基于 RFM 模型

学习行为客户分类将 RFM 中的三个指标进行改变：

R (Recency)：指用户的最近一次消费时间，替换为用户的时间差

F(Frequency)：指用户下单频率，替换为用户的课程数量

M(Monetary)：指用户消费金额，替换为用户在平台的学习时长

将三个数据聚合在一起生成 R、F、M 指标，使用 pandas 库中的 describe 函数，统计包括计数、平均值、标准差、最小值、25%、50%（中位数）、75% 分位数和最大值，确定打分机制，对其评分得（如图 7.3 所示）

	index	userId	R	F	M	R_score	F_score	M_score	R_score_type	F_score_type	M_score_type	RFM_type
0	3	35044	122.000	1.000	679461	1	1	4	低	低	高	低低高
1	4	159756	7.000	1.000	664650	4	1	4	高	低	高	高低高
2	5	159758	152.000	1.000	272079	1	1	4	低	低	高	低低高
3	6	159759	199.000	1.000	508301	1	1	4	低	低	高	低低高
4	7	159760	199.000	1.000	476469	1	1	4	低	低	高	低低高
...	...	...	...	...	...	...	...	...	...	...	...	...
3562	6691	174511	8.000	1.000	17365	4	1	1	高	低	低	高低低
3563	6694	174526	8.000	1.000	2800	4	1	1	高	低	低	高低低
3564	6696	174529	7.000	1.000	29650	4	1	2	高	低	低	高低低
3565	6698	174532	7.000	1.000	28813	4	1	2	高	低	低	高低低
3566	6699	174533	7.000	1.000	55844	4	1	2	高	低	低	高低低

图 7.3 学习行为客户评分数据

带入图 6.2 的用户分类规则得出八张用户数量分布图（如图 7.4 所示）

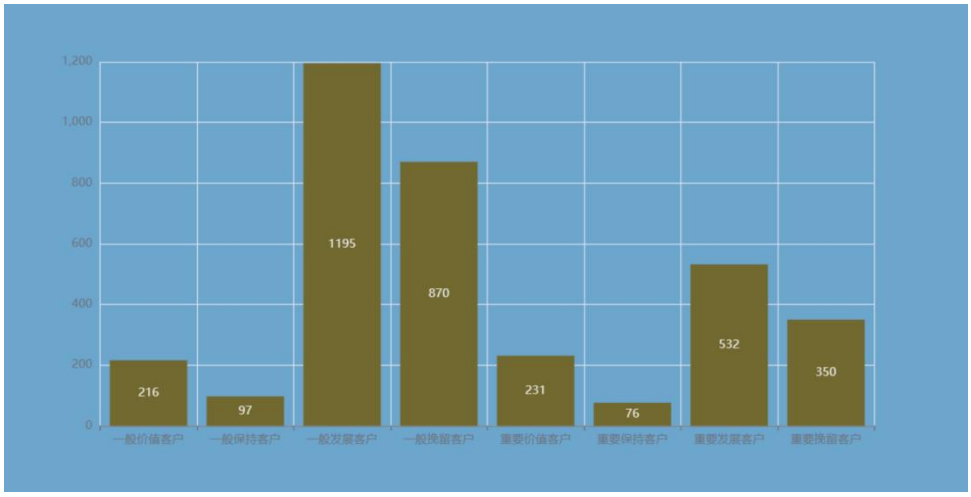


图 7.4 学习行为客户分类图

8. 线上管理决策建议

8.1. 用户活跃度

用户活跃度是教育平台成功的关键因素之一。通过对用户在不同时段、日期的活跃表现进行深入分析，我们可以为平台制定有针对性的运营策略及服务策略。

- 高峰时段维护：在用户活跃度最高时段，如每年三月份至六月份，特别是四月份以及每天的（6：00-8：00、12：00、0：00）等时间段，平台应加强系统维护，确保课程流畅度，避免因系统问题导致用户流失。（如图 8.1 所示）

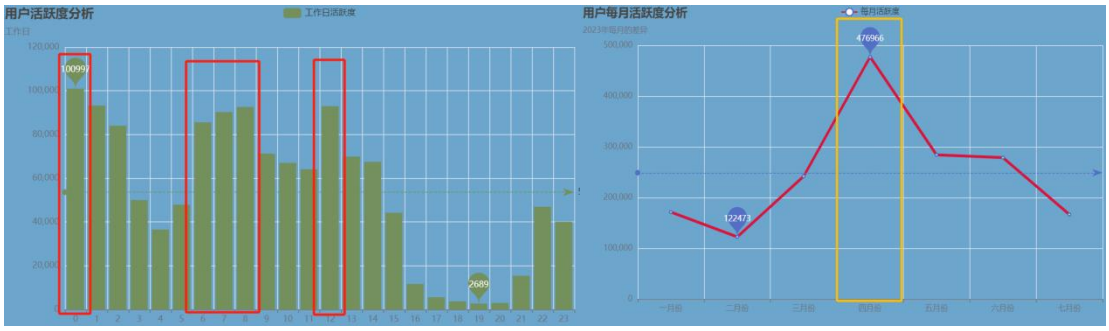


图 8.1 用户活跃度图

- 营销活动推广：在这些高活跃时段，平台可多植入相关课程广告；例如在我的学习—我的课程中下半部分（如图 8.2 所示），根据用户已选的课程推荐相关或者相似的课程，并举办相关课程销售活动，以吸引更多用户，提高课

程的吸引力和销售额。

图 8.2 平台界面图

通过构建的时间序列模型，预测用户未来的活跃度情况。Propht 模型为趋势项、季节性、节假日和误差项组合而成。

季节性趋势 s(t)：使用傅立叶级数来模拟时间序列的周期性：假设 T 表示时间序列的周期，T = 365.25 表示以年为周期，T = 7 表示以周为周期。

傅立叶级数形式是：

$$s(t)=\sum_{n=1}^{N}\left(a_n\cos\left(\frac{2\pi nt}{T}\right)+b_n\sin\left(\frac{2\pi nt}{T}\right)\right)$$

对季节性的拟合需要 2N 个参数，记为向量$\beta=[a_1,b_1,\ldots,a_N,b_N]^T$

这需要对每个历史值、预测值中的时间 t 去构造关于季节性趋势的向量，最终组合成矩阵。用户活跃度预测季节性变化如图 8.3 所示

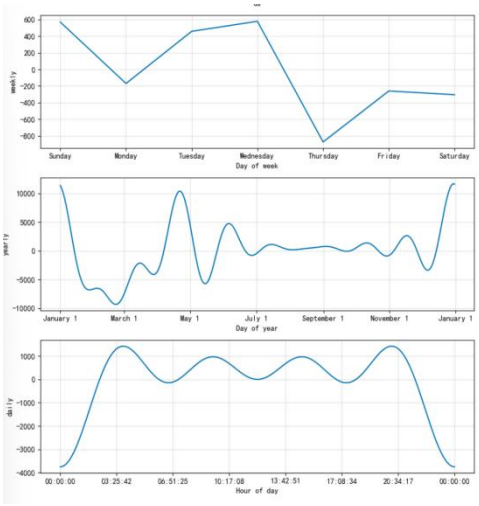


图 8.3 季节性变化趋势图

节假日趋势 h(t)

根据生活经验，节假日、重要事项的影响持续时间不同，同时，假期前后几天的时间都比较重要，假设假期附近时间影响与假期当天一致。于是，需要将不同节假日的影响独立分析，不同的节假日需要设定不同的影响时间范围。

记D_i为第 i 个节假日的前后一段时间，使用参数描述节假日影响范围，满足正态分布，$\kappa \sim Normal(0, v^2)$，默认值是 10，可以调整。其中，v= holidays 为可调整的参数。

假设共有 L 个节假日，得到节假日趋势公式：$h(t) = Z(t)\kappa = \sum_{i=1}^{L} \kappa_i \cdot 1(t \in D_i)$

根据所输入的节假日得到的用户活跃度预测季节性变化如图 8.3 所示

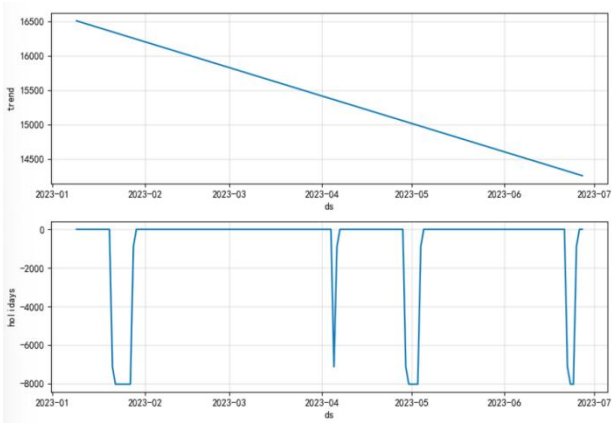


图 8.4 节假日变化趋势图

根据建立的模型平台可以提前预测活跃度变化情况，根据变化情况提前调整课程更新频率、增加或减少广告投放等，以应对可能出现的用户活跃度波动。此外，平台还可以根据预测结果对课程内容进行优化，以满足用户不断变化的学习需求，从而吸引更多用户，提高课程的吸引力和销售额。

8.2. 留存率方面

留存率是衡量平台用户忠诚度和持续使用意愿的重要指标。根据用户数据，我们可以将未流失用户进行划分，并针对不同类型用户实施不同的维护策略。

- 活跃用户：对于活跃用户，平台可提供补贴优惠券、VIP 体验券、VIP 课程套组、免费福利等，以鼓励他们继续使用平台，举办邀请活动通过微信、QQ 将平台分享给他人获取一定的奖励等等。（如图 8.5 所示）

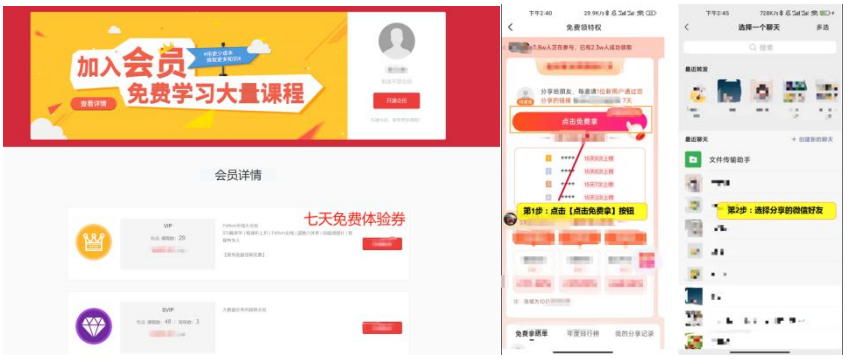


图 8.5 活动样例图

- 潜水用户：对于较少使用的潜水用户，平台可降低价格、提供精准营销等手段，将活动放在登陆界面或者私信等等渠道，重新激发他们的学习兴趣。

图 8.6 折扣课程及推荐图

- 平台优化：平台还应优化用户间交互增加弹幕功能和留言功能，同时提高画面质量和视频稳定性及清晰度，并搭建个性化推荐系统，打造核心课程，以提高用户粘性，让用户能够长时间持续登录平台进行学习。（如图 8.7 所示）

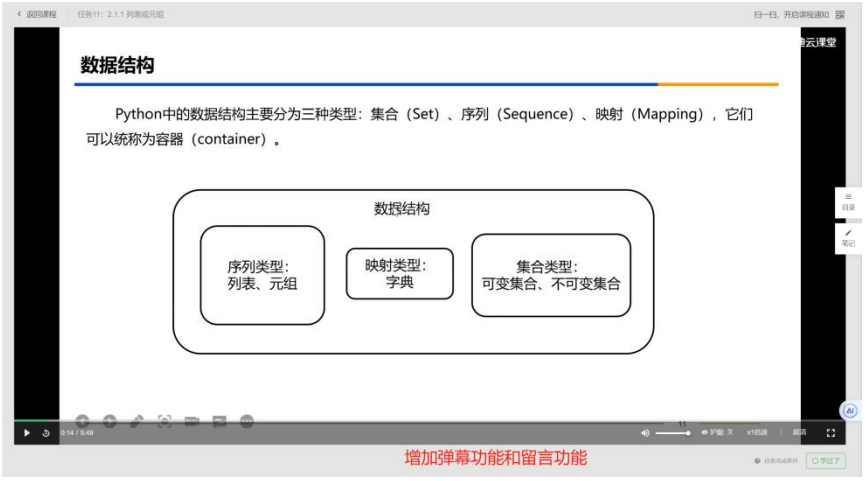


图 8.7 视频界面图

8.3. 学习行为客户方面

通过学习行为客户分类图，针对不同类型的客户，平台应采取不同的策略来提升用户满意度和留存率。

- 重要发展客户、重要挽留客户的策略：两者的占比相对较高，采取的策略就要提供高品质的课程及奖励提高用户的粘性，同时刺激他们的消费并给予更多的优惠。
- 一般发展客户、一般挽留客户的策略：两者的占比很高，采取的策略可以是提供性价比相对较高的课程及一些优惠券、体验券等等，吸引他们在持续在平台学习为主、辅助刺激消费