****

机器学习大作业报告

**Bilibili弹幕视频网用户活跃度分析——基于机器学习模型**

专业： 大数据管理与应用

班级： 大数据201

姓名： 江凌峰

学号： 42007242

目录

[1.引言 2](#_Toc2656)

[1.1研究背景和意义 2](#_Toc1942)

[1.2研究目的 3](#_Toc18200)

[2.文献综述 4](#_Toc16788)

[3.网站背景与数据集介绍 4](#_Toc19402)

[4.数据采集与预处理 5](#_Toc4629)

[5.用户活跃度预测模型 8](#_Toc23197)

[5.1模型总体介绍 8](#_Toc4152)

[5.2描述性统计与特征选取 9](#_Toc23933)

[5.3模型训练与调参 11](#_Toc6235)

[5.4模型评估 15](#_Toc30955)

[5.5总结与分析 17](#_Toc30765)

[6. 参考文献 17](#_Toc23110)

# **1.**引言

## 1.1研究背景和意义

随着互联网信息技术的迅猛发展和社交网络的不断扩张，人们在充斥着各种娱乐内容的环境中，涌现出许多以用户个性化兴趣为核心的社区。在这些社区中，用户之间的信息交互变得越来越频繁和紧密，实时性、社会性和趣味性已经成为用户信息交互的重要特点。[1]

对于视频网站来说，活跃用户是其最宝贵的资产之一，因为用户活跃度直接关系到公司的叙事能力。即使没有盈利，一个拥有数百万日活跃用户的公司仍然能够获得投资人的青睐。如今，产品变现的方式已经相对成熟，通过接入广告，公司可以获得可观的收益。

在用户活跃度价值固定的情况下，用户活跃数越大，收益也相应增加。通过运营手段提升活跃用户数量，公司的收益也会相应提升一个数量级。从这个角度来看，企业通常要求运营部门在保持用户新增的情况下，提高用户的活跃度，以提升收益。

用户活跃度的提升对于产品的生态系统具有重要意义。活跃用户是互联网企业生存与发展的动力所在，拥有活跃的用户意味着更多的选择可能性，用户之间可以相互促进，形成良性循环，从而构建一个有益的商业生态系统。

本研究利用数据挖掘的方法，分析基于Bilibili弹幕视频网站的用户信息，旨在探索用户活跃度与部分用户数据之间的关系。

## 1.2研究目的

使用用户所公开的特征信息评估用户活跃度具有重要的商业和研究意义。通过评估用户活跃度，企业可以更好地了解用户的行为和偏好，从而提供个性化的产品和服务，增强用户留存和用户增长。了解哪些用户更有可能成为忠实用户，并采取相应的措施来提高他们的活跃度，可以帮助企业保持稳定的用户基础和持续的增长。

活跃用户通常对产品更感兴趣并更倾向于参与其中。通过评估用户活跃度，企业可以了解用户对于产品的参与度和互动性，从而改进产品功能、设计更吸引用户的特性，并提供更具吸引力的用户体验。这有助于增加用户的参与度，促进用户之间的互动和社区建设。

通过分析用户的特征信息与活跃度之间的关系，企业可以建立个性化推荐系统和精准营销策略。准确了解用户的兴趣、偏好和行为模式，可以提供定制化的推荐内容和个性化的营销活动，从而提高用户的参与度和忠诚度，并提升商业效益。

用户活跃度与用户满意度和用户体验密切相关。评估用户活跃度可以帮助企业了解用户对产品或服务的满意度，从而发现改进和优化的机会。提高用户活跃度可以增加用户的忠诚度和满意度，提升品牌形象和口碑。

评估用户活跃度对于社会科学研究和用户行为分析也具有重要意义。通过分析用户特征信息与活跃度之间的关系，可以揭示用户行为模式、社交网络结构和信息传播等方面的规律和趋势，为社会科学研究提供实证数据和理论支持。

总而言之，评估用户活跃度通过用户的特征信息在商业和研究领域都具有广泛的应用和意义，可以帮助企业优化运营策略、改善用户体验，提高用户参与度和忠诚度，以及促进社会科学研究和用户行为分析的发展。

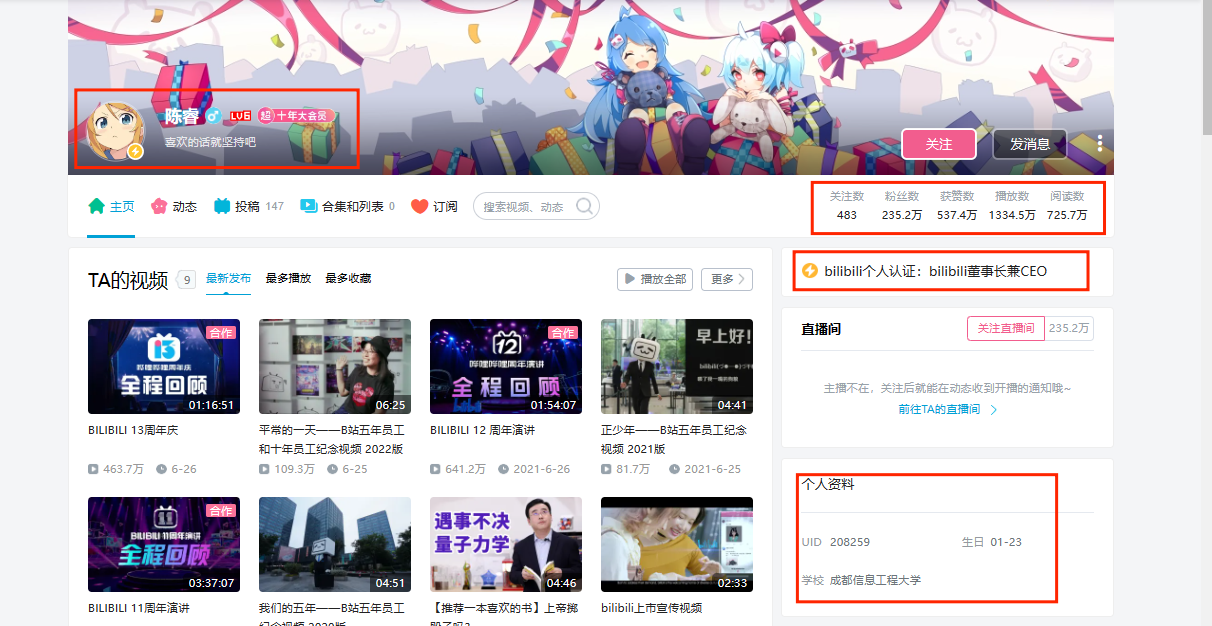
# **2.**文献综述

何明[2]关于视频网站的活跃度研究中，提出了用数据挖掘技术,结合心理学、商学等交叉学科知识,设计建模方法,实现对在线视频弹幕数据的全面理解。包括①对弹幕数据的新特性进行系统分析与量化建模②用于大规模图像分类的深度混合模型③设计门网络用来融合各个基础卷积神经网络的输出以生成最终的分类结果。

李辰宇[3]从多个角度对网络视频业务进行了深入详细的分析,并直接解决了若干现实问题。包括①对视频分发阶段用户与服务器之间的通信流程与通信特点进行分析,并提出了 一个通用的网络视频业务分发服务器检测方法②基于从我国东北某省的移动通信网中采集的超过170亿条网络流量话单数据,从数据消耗、位置移动、业务使用三个方面,对新兴的移动网络环境中视频用户的行为特性进行了深入的分析③基于从网络运营商处采集的超过120亿条大规模流量数据和从视频网站处爬取的连续30天的长时期视频元信息,从用户活跃度、视频属性、用户关系三个方面,对新兴的综合性网络视频业务中两种关键类型用户的喜好特性进行了深入的分析等等。

# **3.**网站背景与数据集介绍

本案例以弹幕视频网站Bilibili为例，通过用户可公开的数据进行查询和爬取整合得到的信息进行挖掘分析。

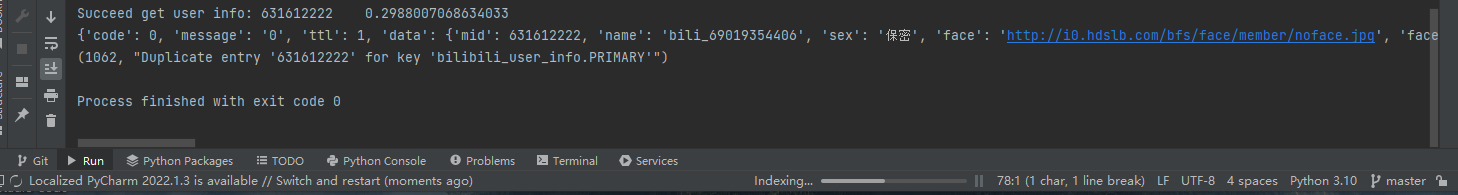


*——用户界面信息*

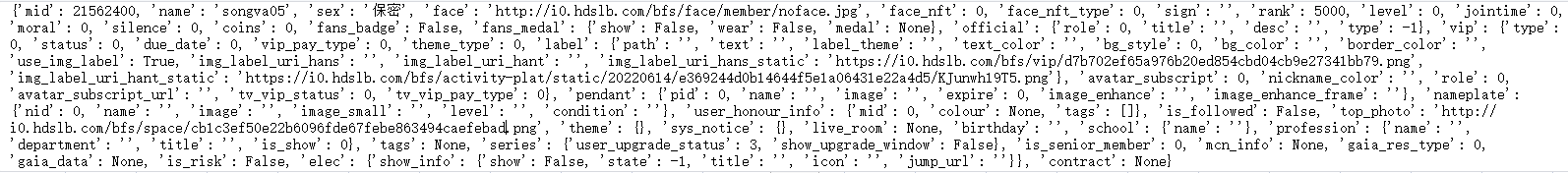
大致所有公开可能有用的信息如上图标注，如昵称、ID号、性别、用户等级、会员状态、关注数、粉丝数、生日、学校等信息，但是，由于用户之间的权限设置不同，可看到的信息也不尽相同，网站源代码中用户性别信息也无法完全获取，因而针对企业内部的分析而言，可以通过更详细的数据，再通过特征选择来构建更合理、符合现实意义的模型。

# **4.**数据采集与预处理

本实验通过爬虫软件，访问用户公开的特征信息界面并收集数据，尝试完全爬取后的一条数据如下：

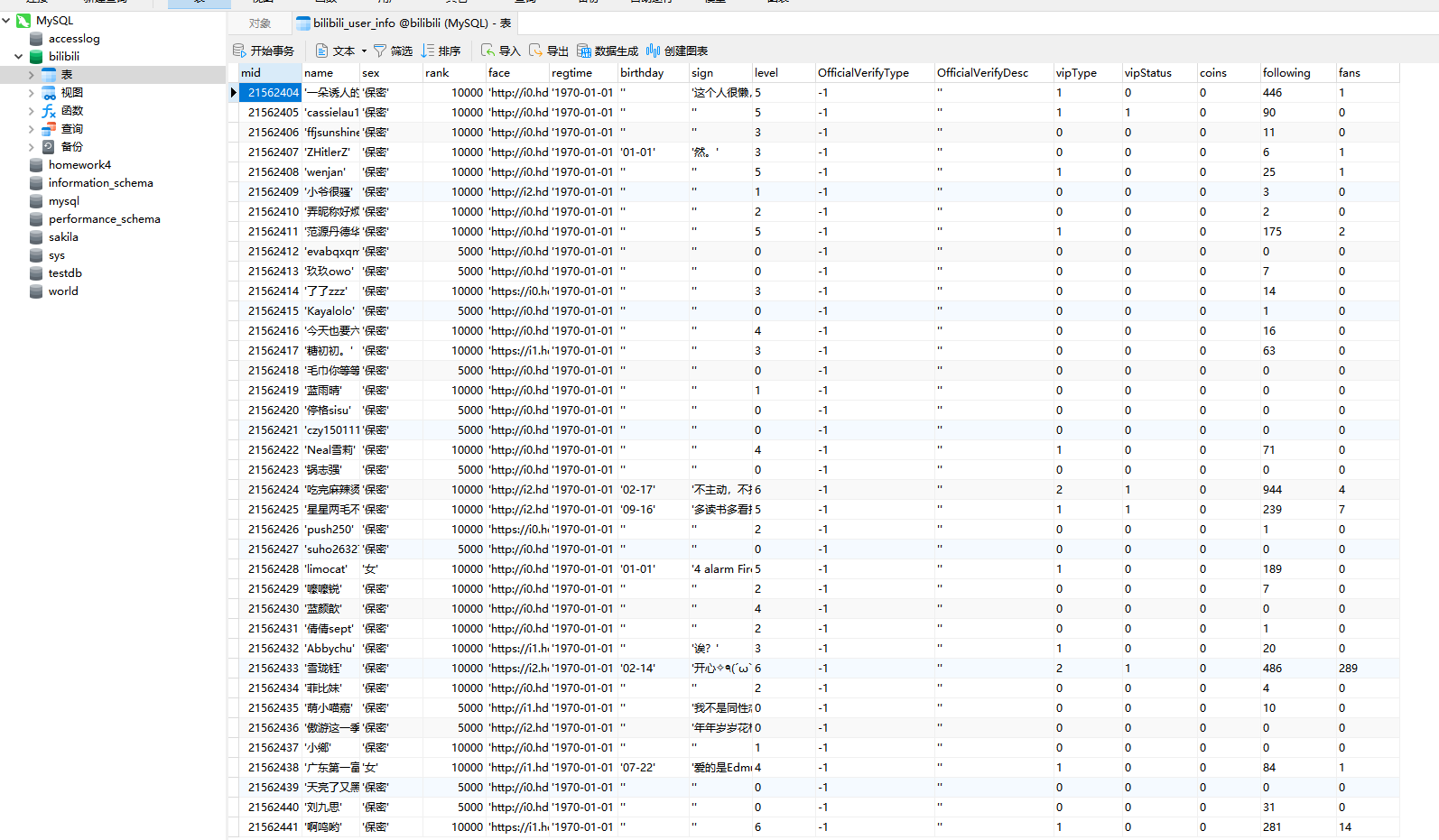


*——程序运行框图*



*——所有可搜集到的信息（字典格式）*

包含了用户的所有信息及设置，因为大部分的用户可能存在默认的情况，或者很共性的选择，仅少数有个性的设置，为了简化特征选取的复杂程度，所以本案例中将大部分默认的特征信息予以排除，筛选出较为有效的属性，得到的数据结构如下：

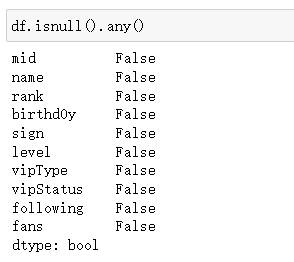


*——MySQL数据库结果*

相关属性解释：

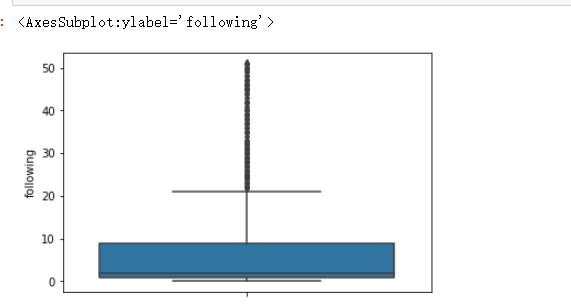
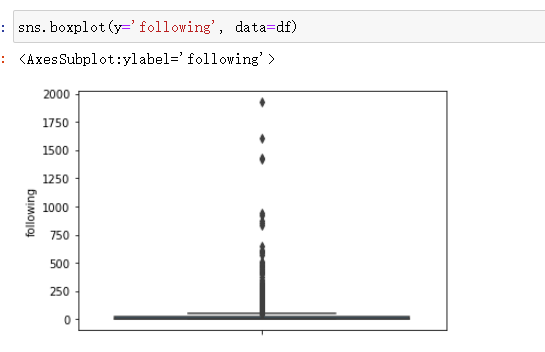
|  |  |
| --- | --- |
| **属性名称** | **内容介绍** |
| Mid | 用户的b站id号，通过此编号检索收集用户特征信息 |
| Name | 用户昵称，标签属性 |
| Sex | 用户性别，类型属性 |
| Rank | 用户等级经验值，数值属性 |
| Birthday | 用户的生日（用户设置了生日信息为1，没有设置为0） |
| Sign | 用户的个性签名（用户设置了生日信息为1，没有设置为0） |
| level | 用户的B站等级，等级是视频网站根据内部经验规则呈现的用户等级，反映了用户的活跃程度（0级一般为注册但未使用，1、2级为非活跃用户，3级为一般活跃用户，4、5、6为活跃用户，所以，在本案例中，3级及以上均可以被认为是活跃用户） |
| Vip Type | 会员种类（分为三级，0是未开通会员，1是普通会员，2是年费会员） |
| VipStatus | 会员状态（指的是用户目前是否处于会员状态，1意味用户还处于会员状态，0意味用户曾经开通会员，但过期后当前时刻还未续费） |
| Following | 关注数，指的是用户在视频平台关注的视频博主账号个数 |
| Fans | 粉丝数，意味着用户的个人粉丝数 |

因为数据获取时，空值由0替代，所以没有空白数据，但对于多个指标同时为0的数据，可以认为是“僵尸号”，做剔除处理，从2476条数据中，筛选和剔除后，留下1098条有效数据。



其中，对于签名和生日资料，将其分为0、1，0意味着没有设置，1意味着设置了，可以一定程度反映用户对于视频网站的黏度。

异常值处理：关注数作为用户的个性选择，差异较大，故观察这方面的数据查看异常值，可以看到关注数的离散程度是较大的。



尽管从箱线图来看，存在一定的异常数据，但由于用户的个性不同，关注数允许有一定量的差异，在本案例中并不影响分析与结论。



*——处理后的有效数据*

# **5.**用户活跃度预测模型

## 5.1模型总体介绍

本文以7：3的比例划分训练集与测试集，分别建立了Logistic回归、决策树、随机森林、SVM模型、K最近邻模型、神经网络模型、朴素贝叶斯模型，并评价了它们的表现。

根据计算结果，我们可以对各个模型的性能进行初步总结：

1. Logistic Regression（逻辑回归）模型表现出较高的训练集和测试集准确率，以及AUC值，说明其在这个问题上具有较好的分类能力和预测性能。

2. Decision Tree（决策树）模型在训练集上表现良好，但在测试集上略有下降，可能存在过拟合的情况。AUC值也相对较低，表示模型在预测上的性能有待改进。

3. Random Forest（随机森林）模型在训练集和测试集上都表现不错，具有较高的准确率和AUC值，这可能是由于其对决策树的集成能够减少过拟合的影响。

4. SVM（支持向量机）模型在训练集和测试集上的准确率和AUC值都相对较高，表现良好，但相较于其他模型，训练集上的准确率较低，可能需要进一步调整模型参数。

5. K-Nearest Neighbors（K最近邻）模型在训练集和测试集上表现相对一致，准确率和AUC值较高，但与其他模型相比，性能略有下降。

6. Neural Network（神经网络）模型在训练集和测试集上表现良好，具有较高的准确率和AUC值，这可能是由于其具有较强的模型拟合能力。

7. Naive Bayes（朴素贝叶斯）模型在训练集和测试集上表现较为一致，准确率和AUC值较低，可能需要进一步改进模型或考虑其他算法。

综上所述，Logistic Regression、Random Forest和Neural Network模型可能是较好的选择。

## 5.2描述性统计与特征选取

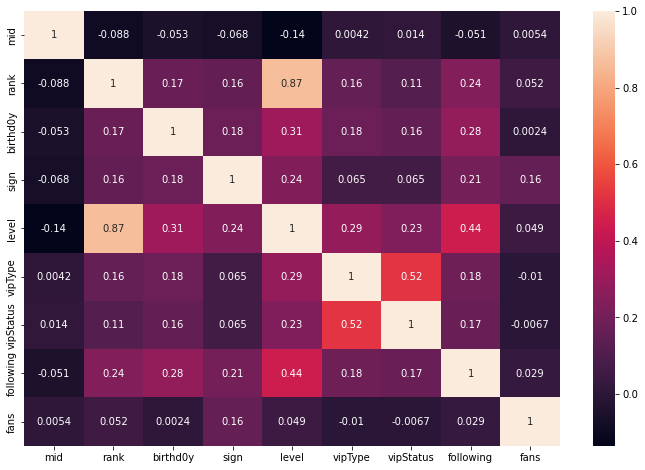
用户活跃度与许多用户个人数据相关，本案例从一些用户的特征信息来判断结果，所以选取的数据主要从数据的可获得性和与活跃度等级相关两个角度来考虑，得到相关性热图与皮尔斯系数结果如下：

通过简单的相关性观察，大多数相关性并不强：

活跃度level与经验值rank相关性0.87 ;

会员等级vipType与会员状态vipStatus相关性0.52 ;

关注数following与活跃度level相关性0.44 .



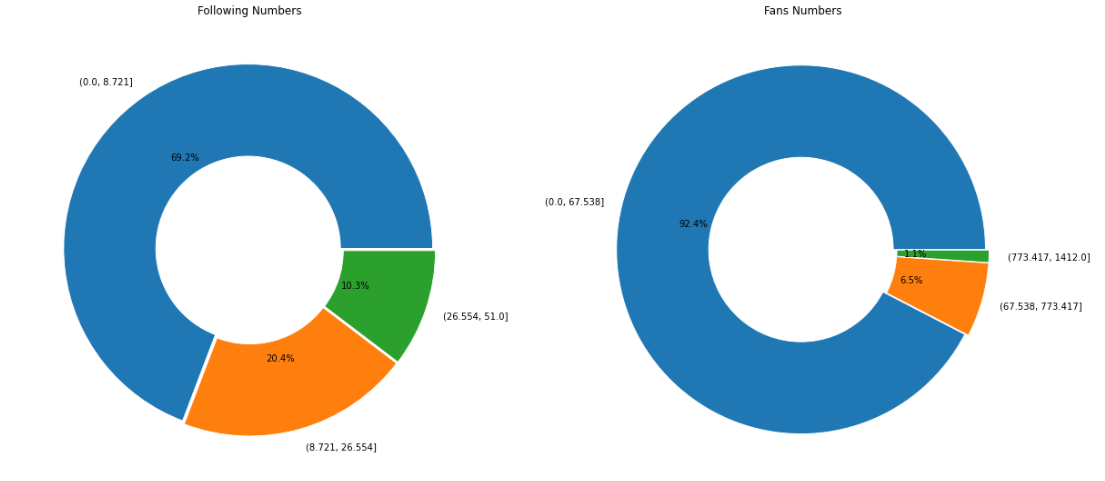
*——相关性热图*

|  |  |
| --- | --- |
| **Feature** | **Correlation** |
| Rank | 0.866592 |
| Following | 0.438446 |
| Birthday | 0.311075 |
| Viptype | 0.289914 |
| Sign | 0.236706 |
| Vipstatus | 0.228864 |
| Fans | 0.049370 |

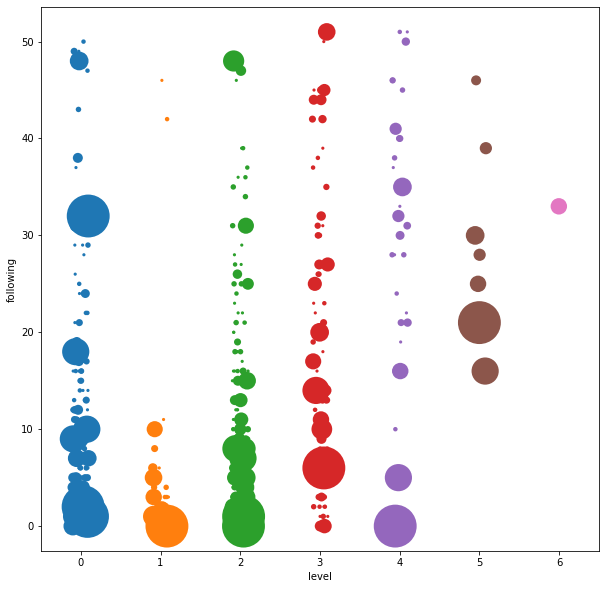
*——皮尔斯相关系数*

结果来说，fans过于不理想，考虑到现实中，绝大多数普通用户使用视频网站的活跃度等级确实与粉丝数相关性并不大，所以可以剔除该变量选择其余属性作为模型使用的数据。同时，我们也可以简单观察一些粉丝数据的一些分布特征，帮助我们理解一些现实问题。

使用K-Means方法将关注数和粉丝数进行聚类离散化，分为了高关注量、中关注量、低关注量和高粉丝数、中粉丝数、低粉丝数三类。



*——三类用户的分布图*



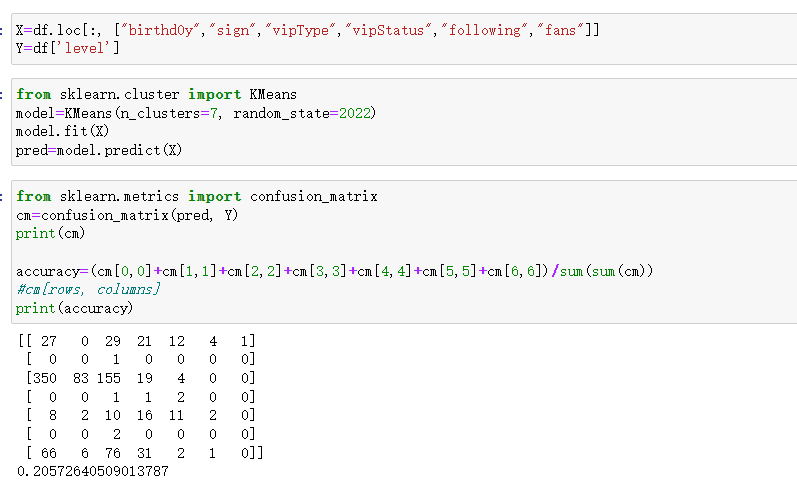
*——活跃度与关注数的计数图*

对于其它参数，可以从图中看出，越是活跃度高的用户，关注数的“气泡”越稍往上靠，意味着关注数多的用户，可能会对网站黏性增强，从而增加活跃度。

从线性回归的角度来看，预测程度一般，生成回归方程为：

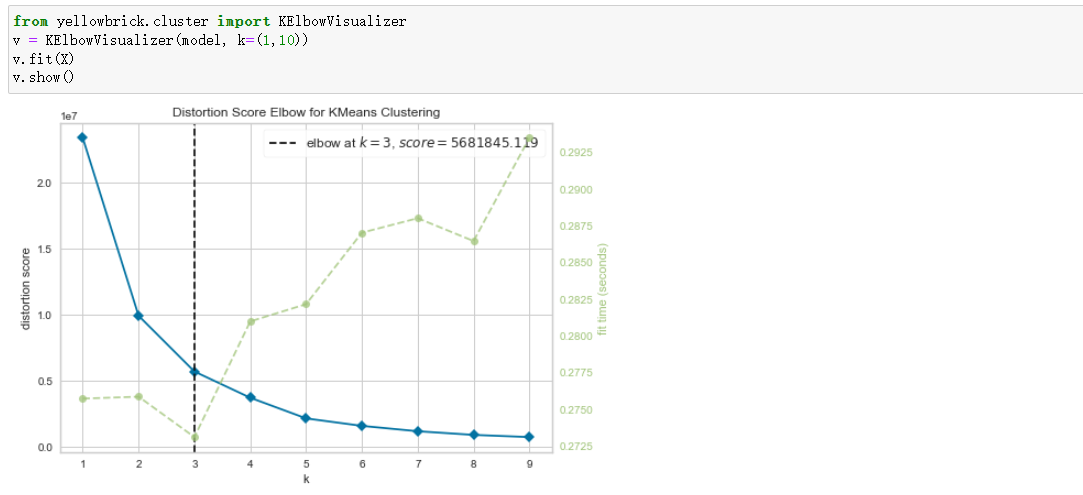
following= [3.95717079] \*level + [3.27438134]

## 5.3模型训练与调参



*——C-均值聚类混淆矩阵图*

通过纯按原来的7种等级进行7簇的K-Means聚类，输出混淆矩阵发现结果并不如意，只有0.2的准确率。原因首先就是，视频平台的经验计算的指标，在我们所能获得的信息中无法完全体现，所以聚类时无法找到不同等级的差异，其次就是算法的不准确性，无法精准定断类型。

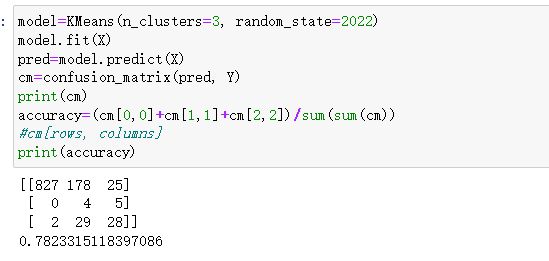


*——“肘型法”匹配图*

Yellowbrick的KelbowVisualizer实现了“肘形”方法，即通过将k-means模型与k的值范围相匹配来选择最佳簇数。如果折线图看起来像一个手臂，那么“肘部”（曲线上的拐点）就是一个很好的提示，表明基础模型在这一点上最适合。

根据测试提示，我们应当分为3簇最为合适。

那么，就我们对数据进行适当的修改，将0，1，2级的用户统一分为0即非活跃用户，而3，4级的用户都归为一般活跃用户，5，6级为重要活跃用户，利用修改的数据再做一次K-Means的聚类分析，查看结果。



*——混淆矩阵*

发现此时准确率大大上升，但依旧错误率较高。大部分集中在活跃度2级上下，且更容易判为非活跃用户，非活跃用户的判断失误仅为2，故非活跃的标准是明显的，而活跃则更为复杂。

所以，我们得到了一个重要的结论：用户活跃度的指标需要时事更新，因为活跃用户变为非活跃用户是动态的过程，特征不够明显，需要实时对比才能发现。

针对决策树，DecisionTreeClassifier 既能用于二分类（其中标签为[-1,1]）也能用于多分类（其中标签为[0,…,k-1]）。,采用输入两个数组：数组X，用[n\_samples, n\_features] 的方式来存放训练样本。整数值数组Y，用 [n\_samples] 来保存训练样本的类标签。

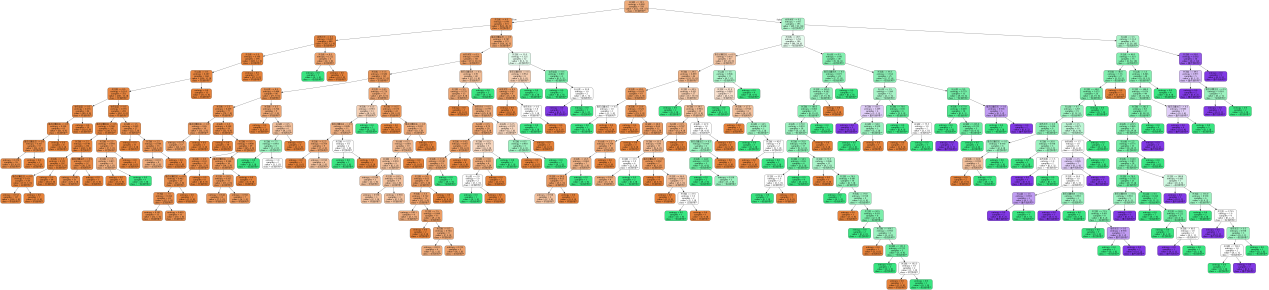
训练后，可以进行预测，clf.predict()，可以预测该样本属于每个类的概率。

优点在于决策树易于理解和实现；易于通过静态测试来对模型进行评测，可以测定模型可信度；如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。

缺点也在于对连续性的字段比较难预测；对有时间顺序的数据，需要很多预处理的工作；当类别太多时，错误可能就会增加的比较快；一般的算法分类的时候，只是根据一个字段来分类。

决策树的构建：特征选择/计算公式，特征选择即决定用哪个特征来划分特征空间，其目的在于选取对训练数据具有分类能力的特征，提高决策树的学习效率。决策树需要找出最佳节点和最佳的分枝方法，而衡量这个“最佳”的指标叫做不纯度。由此还衍生出其他两个常用指标，一个是ID3中信息增益的计算方法可用熵推导，即最为人熟知的信息熵，又叫香农熵，另一个是基尼系数，主要用于CART决策树的纯度判定中。

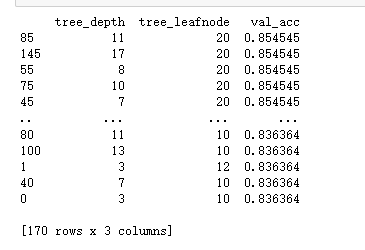
  决策树最终的优化目标是使得叶节点的总不纯度最低，即对应衡量不纯度的指标最低。

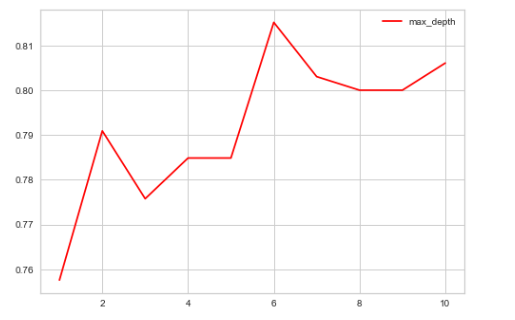


*——决策树结果（较大，详见附件png图片）*

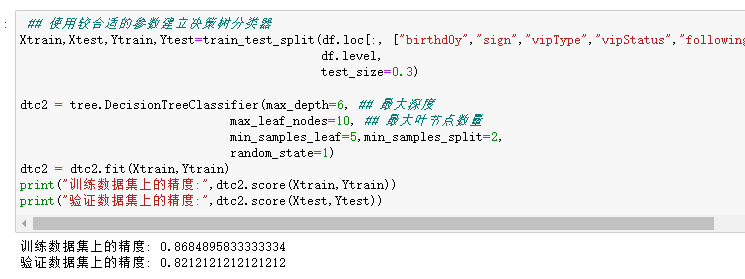
观察所示的模型结构可以发现，该模型是非常复杂的决策树模型，而且决策树的层数远远超过了10层，从而使用该决策树获得的规则会非常的复杂。通过模型的可视化进一步证明了获得的决策树模型具有严重的过拟合问题，需要对模型进行剪枝，精简模型。

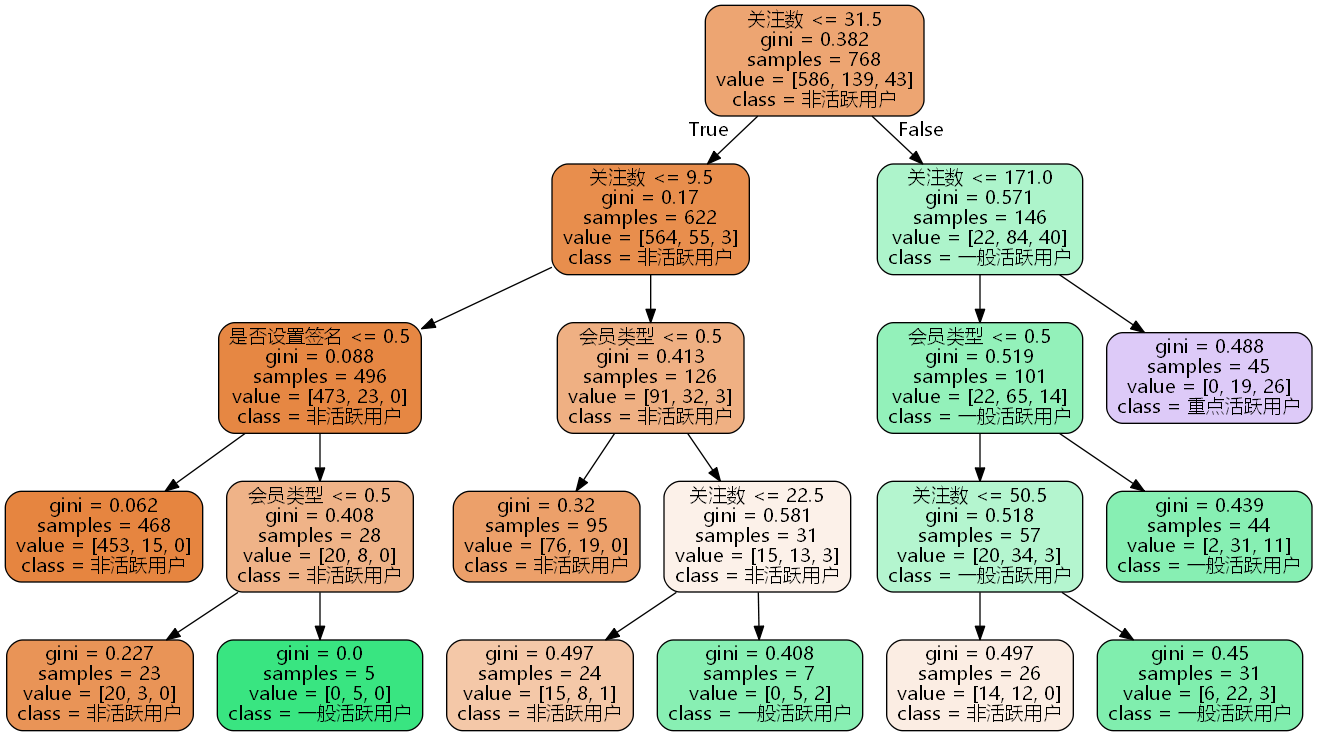
确定最优剪枝参数：





深度为6时，测试集的准确度最高。

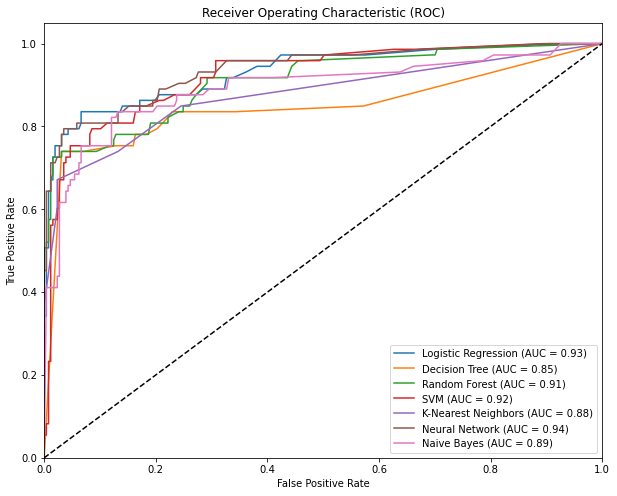




*——剪枝后决策树结果*

剪枝后的模型大大优化，相对于未剪枝时更加直观明了，且模型准确度80%以上，已经较为精确。

为了综合考量不同模型效果，我们下面再将3，4，5，6合并为活跃用户1，0，1，2合并为非活跃用户0。这样模型就都变成了一个二分类问题，便于相互比较。



*——ROC\_AUC曲线*

## 5.4模型评估

在模型评估方面，我们使用了准确率（Accuracy）和AUC值（Area Under the Curve）作为评估指标。这两个指标可以帮助我们了解模型的分类准确性和预测性能。

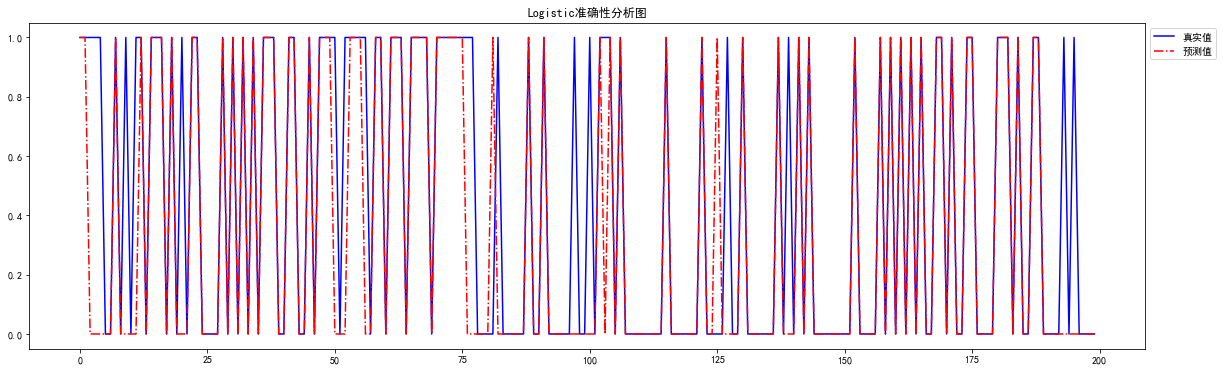
准确率是指模型正确预测的样本数占总样本数的比例，它是最常用的评估指标之一。AUC值是用于评估二分类模型的预测性能的指标。它衡量了模型在不同阈值下的假阳性率（False Positive Rate）和真阳性率（True Positive Rate）之间的平衡。在我们的结果中，各个模型的AUC值如下：

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **训练集准确率** | **测试集准确率** | **AUC值** |
| Logistic Regression | 0.891927 | 0.927273 | 0.931587 |
| Decision Tree | 0.950521 | 0.912121 | 0.838148 |
| Random Forest | 0.950521 | 0.915152 | 0.907254 |
| SVM | 0.876302 | 0.89697 | 0.922472 |
| K-Nearest Neighbors | 0.911458 | 0.909091 | 0.875859 |
| Neural Network | 0.89974 | 0.921212 | 0.928495 |
| Naive Bayes | 0.863281 | 0.893939 | 0.892037 |

综合考虑准确率和AUC值，我们可以看到在我们的实验中，Logistic Regression和Random Forest模型在测试集上表现较好，具有较高的准确率和AUC值，可以作为较优选择。

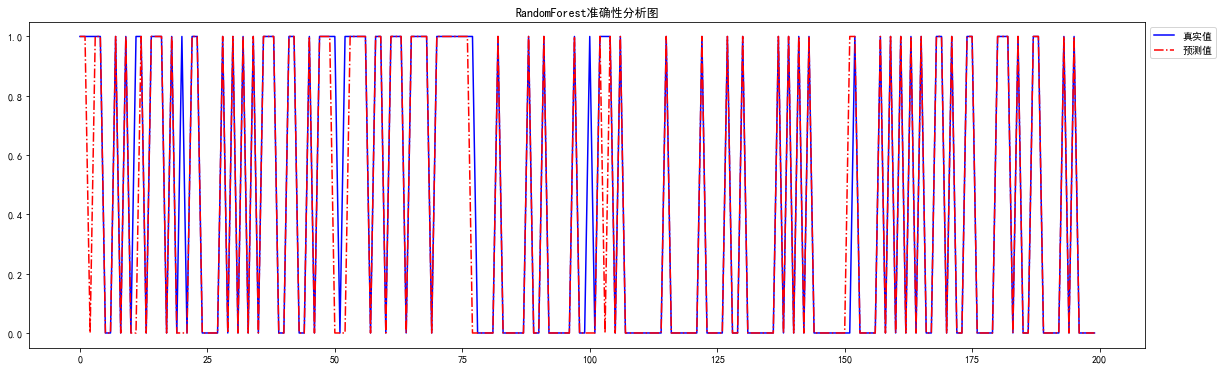
在逻辑回归中，预测效果如下：





在随机森林中，预测效果如下：





## 5.5总结与分析

通过上述所有的分析，我们认为，在视频网站中，用户的关注数、用户的会员状态、用户的会员类型是最主要的三个判别标准，具体体现为，关注数多的用户，对于网站视频的认可度较高，有订阅观看的倾向，其次，会员可以享受网站中特定的视频观看特权，说明用户对于网站内的获得版权的电影、电视剧、动画等认可度高，也具有黏性。

可根据决策树来判断用户类型，再使用自身的大数据算法，例如关联规则，去探寻此类用户共同喜好的视频内容进行推送，做到加强互动，增加用户的依赖。

本实验中，聚类的准确度无法到达90%左右，说明准确度欠佳，问题出现在我们可控的变量较少，即能反映用户活跃度的信息不充分，仅靠部分特征来反映，而这些指标从相关性分析可以看出并非强有效，意味着模型可能有因无法获取而未选取到，但又较为关键的特征变量。

再者，就是实验中对于僵尸号的判断是多个指标同时为空值时认为是僵尸号并删除，而真实环境中，企业应该有更准确的手段来判断，所以，本案例中应该存在弃用账号的影响因素，导致了判断的不准确。

对于互联网企业来说，认真研究自己的用户非常有必要，而本案例受获取权限影响，仅对用户的部分指标进行了数据分析，对于企业而言，利用手中更多的具体数据例如，观看时长、点赞次数、投币次数等一些列信息来进行更精确的判断，同时，也要注意时事更新不同类型用户的各项数据，因为用户的活跃与不活跃是动态的，无法用单一不变的指标来衡量。

# 参考文献

1. 朱钰涵. 在线视频社区中弹幕信息交互群体的用户画像研究[D].南京大学,2019.
2. 何明. 面向在线视频弹幕数据的挖掘方法研究[D].中国科学技术大学,2018.
3. 李辰宇. 基于海量网络数据的视频业务流量分析与应用[D].北京邮电大学,2017.
4. python机器学习：决策树详解[EB/OL].csdn，2022-11-18[2022-11-18].https://blog.csdn.net/oax\_knud/article/details/118826373
5. Python——决策树分类模型剪枝[EB/OL].csdn，2021-09-16[2021-09-16].https://blog.csdn.net/daitulin/article/details/120324542