SHANGHAI UNIVERSITY

课程项目报告

课　　程　数据分析方法

学　　院　悉尼工商学院

专　　业　信管

学　　号　18124399

姓　　名　冯秦萱

基于 B 站番剧信息的点赞量预测

学号：18124399　　姓名：冯秦萱

摘要：B 站作为国内最受年轻人欢迎的番剧平台，其番剧数据为创作者提供了创作目标和方向。本文通过爬取 Bilibili 网站番剧的集数，评分，播放量，投币数等相关信息，利用 python 对这些番剧进行数据挖掘，经过探索性分析和数据可视化直观展现出 B 站番剧的人气特征和发展特点，再结合机器学习算法，进行对比实验，构建了番剧点赞量预测模型并调整优化，最后对实验结果进行了分析。

关键字：数据挖掘；数据分析；Bilibili 弹幕视频网；机器学习；

1 引言

类似点赞互动模式存在极大的商业价值，B 站作为视频点赞的代表平台，在促进视频传播研究的过程中，扮演了至关重要的角色。番剧的点赞量反映了大众对番剧的喜爱程度，一定程度上象征了番剧的人气。B 站上有许多番剧，有的火爆至极，有的却无人问津，作为国内最受年轻人欢迎的番剧平台，其番剧数据为创作者提供了创作目标和方向。本文即是通过对 B 站番剧信息进行数据挖掘，探究影响番剧点赞量的因素，并对番剧点赞量进行预测。

2 相关研究综述

在大数据时代，用户通过观看视频后点赞，除了能够表明自己的喜好，还可以对视频形成口碑效应，所以如果能够对视频的点赞量进行预测，就可以帮助观众在观看视频的选择上提供标准和依据，也可以为视频网站的营销等提供参考[1]。

国内对于网络视频的研究大多是关于主流视频网站。郭英[2]对于视频网站“爱奇艺”的发展及前景进行研究，试图从中找到中国视频网站发展的方向；姜丽秋[3]以搜狐视频为例研究了视频网站传播模式及发展策略，从视频内容、平台、盈利模式多个方面分析视频网站现存问题并提出了相关建议。

目前我国国内有关 B 站数据挖掘的文章大多重在分析其用户和弹幕数据。Jia 等人[4]通过研究 B 站的用户数据，提出播放量与弹幕数仅仅只是反映用户对视频的关注程度，称为“隐性流行”，而点赞和投币量才真正反映观众的喜爱程度，也即“显性流行”；陈华庆等人[5]根据弹幕数据对视频传播的影响，提出了弹幕视频的传播模型。总体来说，国内对于视频的研究较为丰富，但是，对于 B 站番剧的研究较少。

3 数据处理

3.1 数据收集

本文选择将 B 站作为本文研究的数据源。B 站番剧索引页，播放页和详情页中含有关于番剧的大量信息。如剧集数，追番人数，投币数，点赞数，弹幕数，播放数，标签等。本文尽可能全面的收集这三个页面的数据，为后续实验提供支撑。下表为番剧信息字段说明，并对字段作出了相关备注。

表 3-1 番剧信息字段说明

字段名称	含义	备注
title	番剧名	黄金神威 第二季
episode	剧集数	全 12 话
mark	分级	会员专享
score	评分	9.9 分
is_finish	是否完结	1
link	播放链接	https://www.bilibili.com/bangumi/play/ss25679
series_follow	系列追番人数	603655
follows	追番人数	215652
coins	投币数	19693
likes	点赞数	40898
danmus	弹幕数	45297
views	播放数	4037771
season_id	番剧 id	25679
media_id	详情 id	28230074
tags	标签	['漫画改', '搞笑', '冒险', '历史']
comments	评分人数	3963
pub_date	发布日期	2018/10/8
intro	简介	寻求着阿依努的财宝的日俄战争的英雄“不死的杉元”杉元佐一和阿依努族少女阿席莉帕……

本文主要采用 python 网络爬虫对研究所需数据进行收集并存储为 csv 文件。爬虫程序的编写难点在于数据接口的寻找和绕过网站的反爬虫机制。对于这三个页面的信息，并不一定是访问某一个接口就可以获取全部的研究所需数据字段，而是要结合多个数据接口来爬取。最终本文在 2021 年 2 月 28 日爬取了 B 站 100 页，每页 20 个番剧，共 2000 条番剧信息数据。完整的爬虫程序和数据文件见附录。

3.2 数据预处理

从 B 站上获取的原始数据，必然存在一定的问题，在进行数据分析之前，必须对数据进行初步的处理。

3.2.1 缺失值

首先对缺失值进行处理，爬取的两千条数据中存在标签为空和缺失的情况，约四十几条，这部分数据其他信息也不完善，因此直接删除。其他列如集数和分级列，缺失值是有意义的，结合网页的查看，对分级列缺失值填充为“免费”。最后留下 1943 条数据。

3.2.2 数值化处理

本文爬取数据时没有做过多处理导致爬取到的集数，评分等都为文本类型。为了后续的数据分析需要将其数值化。如集数列中“全 12 话”文本类型转换为“12”数字类型。

此外对标签列数据的处理是难点，标签列数据包含了番剧类型，而每部番剧

有多个类型，需要进行独热编码来将其转换为数值方便后续模型使用。爬取时由于使用列表进行存储，导致打开文件后，标签如“['漫画改','搞笑','冒险','历史']”包括符号在内被当作了字符串，在处理这列数据时在网上搜索各种方法，结合使用，编写代码如图 3-1 所示。

```
import re
regStr = ".*?([\u4E00-\u9FA5]+).*?"  #取中文字符的正则化表达
temp_list = [re.findall(regStr, i) for i in data["tags"]]
# 获取分类(去重)
tag_list = np.unique([i for j in temp_list for i in j])
# 增加空白列，用来计数(实际上也是one-hot编码)
tag_df = pd.DataFrame(np.zeros([data.shape[0],tag_list.shape[0]]),columns=tag_list)
for i in range(1943):
    #temp_list[i] ['漫画改', '搞笑', '冒险', '历史']
    #用ix实现复杂切片，类似定位
    tag_df.ix[i,temp_list[i]]=1
```

图 3-1：标签数据处理代码

使用正则表达来提取每行标签字符串中的中文字符，即一个个类别存进列表。通过得到的所有类别创建新的全为 0 值的 dataframe，并用 ix 实现定位，为每个番剧的标签中出现的类别计数，即实现了番剧标签的独热编码。番剧标签类别共有['乙女','偶像','催泪','冒险'……'音乐','魔法']等 42 种，并通过求和得到了这 42 种番剧类别的频次。

3.2.3 异常值处理

去掉集数列中出现的如“2016-3-20 上映”的异常值和发布日期列中出现的“敬请期待”。通过查看数值列的箱线图，发现许多列的箱线图，如图 3-2 出现箱体压扁，有极端偏大的异常值情况。

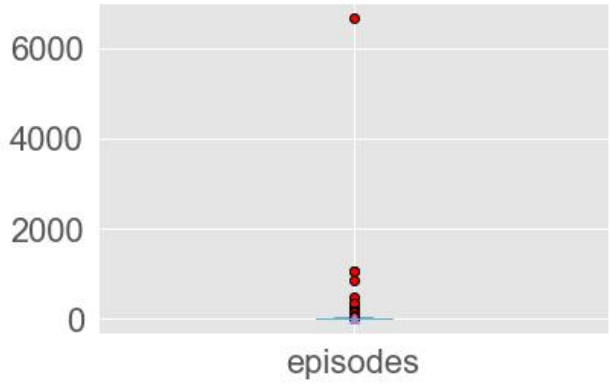

图 3-2：集数列箱线图

因此决定对这些数值列异常值的统一处理，对低于第一个四分位数减去 3 倍四分位数范围和高于第三个四分位数加上 3 倍四分位数范围内的数据进行保留，其他超出范围的值删除，最后留下了 1131 条数据。

3.3 探索性分析

3.3.1 番剧评分总体分布情况

因为爬取的是按照评分排序的前 100 页番剧数据，所以番剧的评分都比较高，统计各个分数番剧的数量如图 3-3 所示。可以从图中看出，番剧评分大多集中在

9.4-9.7 分，并且番剧数量随评分下降而减少，并未呈现正态分布，说明 B 站用户评分普遍偏高。

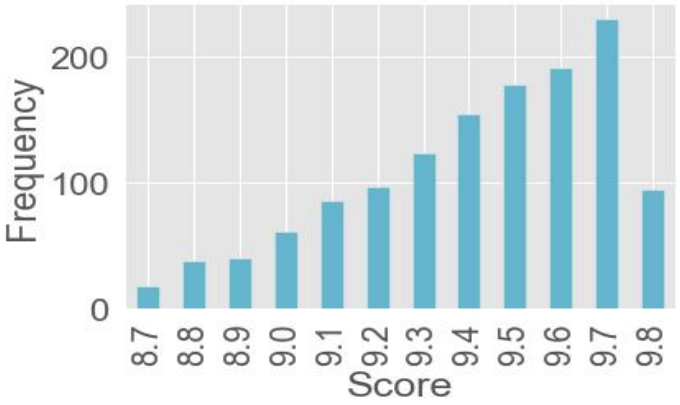

图 3-3：各分数番剧数量图

3.3.2 番剧标签分析

为了更好地了解受众所关注的番剧类型，首先统计了各标签的频率，并选取了排名前 10 的热门标签，如图 3-4 所示。并展示了利用 Python 中的 wordcloud 词云生成库制作番剧类型词云图，如图 3-5 所示。

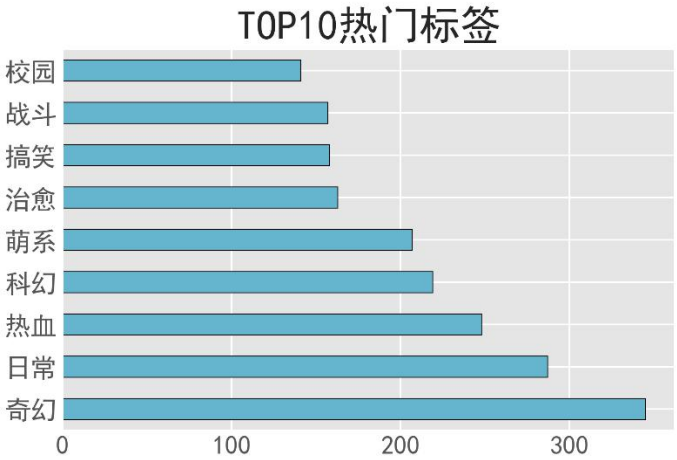

图 3-4：Top10 热门标签统计

图 3-5：番剧标签词云图

从图中可以看出，奇幻、日常、热血、萌系、热血等类别标签是番剧的热门

题材。

3.4 特征构建

通过对数值型数据（除类别标签）的特征重要性及与点赞的相关系数分析，初始选取了投币数，播放量，追番人数，弹幕数，系列追番人数，评论人数，评分，分级 8 个特征。这 8 个特征的相关系数热力图如图 3-6 所示。

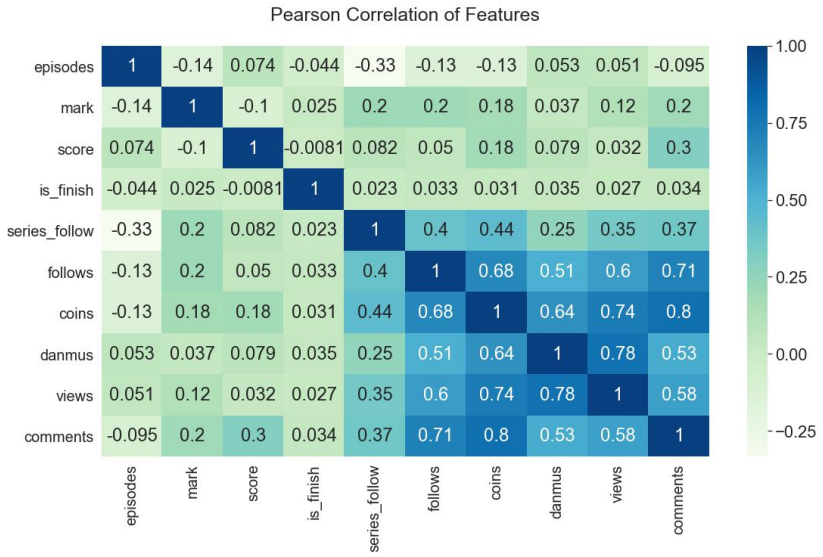

图 3-6：两两特征相关度

通过图 3-6，发现投币数和评论数，弹幕数和播放量，投币数和播放量，评论数和追番人数是高共线特征，因此去掉评论数和播放量两个特征，剩下 6 个特征与 42 个标签特征一起组成共 48 个特征投入模型预测。

4 解决方案

本文选取了逻辑回归、支持向量机、随机森林、Gradient Boosting、K-邻近、AdaBoost 六个算法对番剧处理后的数据进行训练和点赞量预测。通过对这六种算法预测结果的进行评估，选取了梯度提升回归算法对实验数据进行点赞量预测。对选取的模型进行网格搜索寻找最优模型参数，并对比调参前后的效果。最后对模型得到的结果进行解释。

随机森林回归算法是基于决策树分类器所构造的融合算法，是由多棵决策树组成的。该算法的优点是对于变量之间的相关性不敏感，避免了多重共线性的影响[6]。

Gradient Boosting 和 Adaboost 算法。这两个算法都是常用的梯度提升算法。梯度提升回归算法是目前机器学习算法中比较有代表性的算法，可以用于回归或分类问题。

5 实验分析

5.1 模型构建

本文通过 sklearn.model_selection 中的 train_test_split 将特征工程后的 1131 条数据划分训练集和测试集，然后利用 StandardScaler 对预测变量的训练集数据和测试集数据均进行标准化处理，并利用标准化后的数据构建预测模型。本文利用 sklearn 种的各种算法包，将处理好的训练集数据分别使用六种算法来构建线性回归、支持向量机回归、随机森林回归、Gradient Boosting 回归、KNN 回归、Adaboost 回归模型。

5.2 模型评价指标

平均绝对误差(MAE)和均方根误差(RMSE)是关于连续变量的两个最普遍的度量标准。本文选取 MAE 作为评估模型的指标,其公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{1}$$

平均绝对值误差,它表示预测值和观测值之间绝对误差的平均值，反映了预测值与实际值的偏离。

5.3 模型比较

对构建的六个模型利用 predict 方法对测试集预测对应的值，即番剧点赞量。六种模型得到的误差结果如图 5-1 所示。

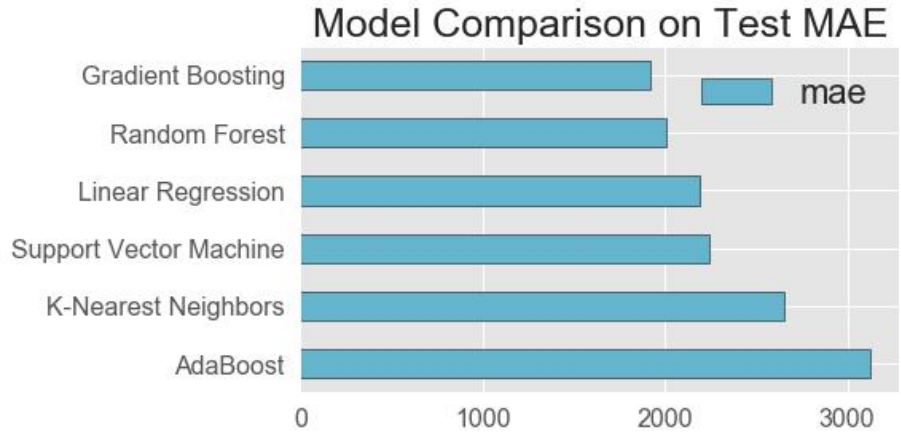

图 5-1：各模型平均绝对误差对比

从图 5-1 中可以看到，Gradient Boosting 回归模型的评估结果是最好的，其平均绝对误差为 1920.2002，优于其他五种模型。所以本文选取 Gradient Boosting 回归模型对实验数据进行点赞量预测。

5.4 模型调参

通过网格搜索对 Gradient Boosting 回归模型的参数进行优化。网格搜索是在所有候选的参数选择中，通过循环遍历，尝试每一种可能性，表现最好的参数就是最终的结果。通过遍历使用的树的数量，损失函数来寻找最优模型。其训练集误差和测试集误差随树的数量的变化如 5-2 所示。

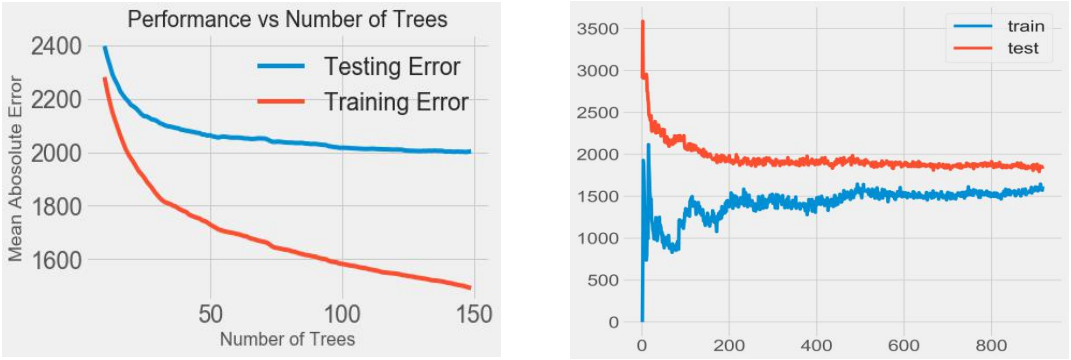

图 5-2：误差曲线

随着模型使用的树的个数增加，训练集误差和测试集误差都会减少。但是，训练集误差比测试集误差下降得快很多。同时也可以直观的推测出来这样的模型存在过拟合现象：它在训练集上表现好，但在测试集上无法达到相同性能。这可能是由于数据量过少的原因。也可以减少树的深度，减少特征数等。

对于最终模型，设定 estimators=147，即交叉验证中最低误差时的超参数值。并测试该模型在测试集上的表现，其调参前后表现如表 5-1 所示。

表 5-1：调参前后模型比较

model	n_estimators	loss	MAE
Base model	Default	Default	1919.7229
Best model	147	lad	1826.7525

可以看出，调参后的模型的 MAE 为 1826.7525，相比较调参前模型有所提升。

5.5 模型评估

将调参后得到的最优模型对测试集进行预测，其表现如图 5-3 所示。

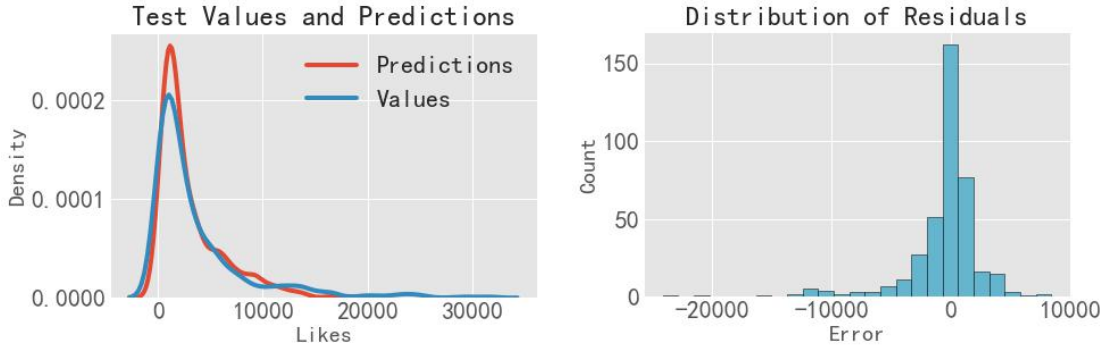

图 5-3：模型预测表现

从左图可以看出，该模型对点赞量的预测曲线基本与真实值相近，从右图可以看出，模型预测的拟合残差基本呈正态分布，但有个别数据预测的残差较大，模型还有改进空间。

5.6 结果解释

将最优模型训练后的特征重要性进行排序，排名前十的重要性如图 5-4 所示。

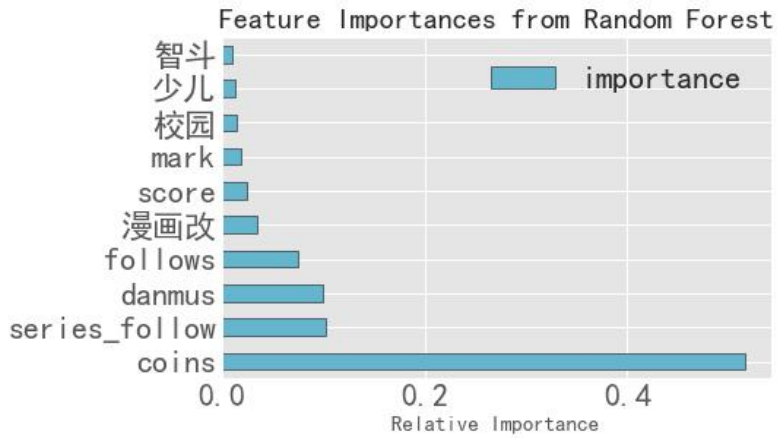

图 5-4：排名前十重要特征

这表明投币数对点赞量的影响程度很大，追番人数和弹幕数次之，类别标签中对点赞数影响较大的有漫画改和校园，这表明受众更喜爱这两个类别的番剧。

6 结论

本文通过爬取 B 站番剧数据，对 B 站番剧信息进行数据挖掘，探究影响番剧点赞量的因素，并对番剧点赞量进行预测。结合六种机器学习算法，进行对比实验，选择 Gradient Boosting 回归模型构建了番剧点赞量预测模型并调参优化。实验结果反映了投币数对点赞量的影响程度很大，漫画改和校园题材的番剧更受欢迎，为番剧创作者提供了创作方向。由于数据量较少，模型训练有过拟合现象。此外在预测点赞量上没有考虑番剧出品方、配音演员、番剧评论、时间等影响因素，需要进一步爬取相关数据进行挖掘分析。

参考文献：

[1] 平澤真大,諏訪博彦，太田敏澄，小川祐樹.对 Niconico 视频网站中社交创新型视频的研究和评测[J]. 電子情報通信学会技術研究報告，2012(03):201-206.

[2] 郭英.视频网站“爱奇艺”发展研究[D]. 曲阜师范大学，2015-06-06.

[3] 姜丽秋. 视频网站传播模式及发展策略研究——以搜狐视频为例[D]. 湖南师 范大学，2015-05.

[4] Jia A. L., Shen S., & Li D. (2018). Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features. *Computer Networks, 140(JUL.20)*, 112-125.

[5] 陈华庆,冼远清,赖建明.网站弹幕视频数据的挖掘与分析[J].福建电脑,2019,35(08):102-103.

[6] 耿娟,郭明欣.豆瓣 Top 250 电影数据挖掘及评分预测[J].河北企业,2021(02):11-13.