SJMMA 2021 D 题

预测奥斯卡

问题背景

在过去的几十年中，我们收集有关消费者偏好数据的能力不断提高。从音乐流媒体到在线购物，推荐系统已成为功能强大的工具，可用于管理用户体验，增加参与度，创造收益。IMDb 和 Rotten Tomatoes 是两个电影评分网站，前者收集观影者对电影的评分，而后者侧重于收集专业影评人的评分。

第 93 届奥斯卡颁奖典礼将于 2021 年 4 月 25 日举行。美国电影学院奖，又称“奥斯卡”，是一年一度的颁奖典礼，旨在表彰过去一年来全球电影人对电影和电影艺术的重大贡献。自成立以来，美国电影学院奖的商业价值和文化影响力在过去 90 年中得到了巨大发展。人们常常认为，赢得这样的奖项对于电影从业者未来的职业发展至关重要，反过来也可以为制片公司本身带来更多的商业收入。

我们将提供三个数据集。您可以选择用一个或几个。

academy_awards.csv

此表包含自 1927 年以来的所有奥斯卡金像奖提名和奖项。10395 行 7 列。

资料来源：https://www.kaggle.com/unanimad/the-oscar-award

imdb.csv

此表包含从 IMDb 抓取的 1986 年至 2016 年的热门电影，提供了电影的行业元数据和用户评分。 6820 行 15 列。

资料来源：https://www.kaggle.com/danielgrijalvas/movies

rotten_tomatoes

此文件夹包含从 Rotten Tomatoes 抓取的专业影评人的电影评分。

在 movies.csv 中，每条记录代表一部在 Rotten Tomatoes 上有记录电影，其 URL，电影标题，描述，流派，时长，导演，演员，用户的评分和影评人的评分。

在 critic_reviews.csv 中，每条记录代表在 Rotten Tomatoes 上发布的影评人评论，其 URL，评论者名称，评论出版物，日期，分数和内容。

资料来源：https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

您的任务

1. 数据清理：这三个数据集具有非常不同的架构，大量重复信息以及少量缺失信息。清理并合并数据集以删除重复项，并且仅保留必要的信息以用于进一步的任务。
2. 用户评分/评论是否可以预测哪些电影/人物将获得特定的奖项？使用数据集中给出的变量构建模型，以预测电影/人获得学院奖提名的可能性。设计度量标准以评估模型的准确性。
3. 设计一种方法来评估电影人在某年的“成功指数”。提示：即使电影制作人在当年不发行任何电影，也可能被认为是成功的电影人。
4. 获得奖项是否对电影人未来的职业发展（成功指数）有所帮助？如果是这样，“获奖效应”能持续多久？
5. 为一家电影制片公司撰写一份非技术报告，他们旨在下一个颁奖季角逐奥斯卡奖。描述您的发现，并提出相应的建议。 （应该迎合观众口味还是影评人口味制作电影？应该制作哪种类型的电影？等等）

提交 你的团队所提交的报告应包含 1 页“总结摘要”、2 页非技术性报告，其正文不可超过 20 页（总页数限于 23 页）。附录和参考文献应置于正文之后，不计入 23 页之限。

*数据: https://pan.baidu.com/s/1GhYoQdacrayBN3Zn4b-wXA 提取码：1s3v

SJMMA2021 Problem D

Predicting the Academy Awards

Background

Over the past few decades, our ability to collect data on consumer preferences has grown exponentially. From music streaming to online shopping, recommender systems have become powerful tools for curating the user experience as well as increasing engagement and revenue. Notably, IMDb and Rotten Tomatoes are two websites that specifically collect consumer preference data on commercial motion pictures, using distinct approaches.

The 93rd Academy Awards is going to be announced on April 25, 2021. The Academy Awards, also known as 'The Oscars', is an annual award show aiming to celebrate major contributions to the art of filmmaking and cinema in the past year, semi-analogous to the Nobel Prize, the Fields medal, or the Turing award, in the field of filmmaking. The commercial value and cultural impact of the Academy Awards has drastically evolved over the past 90 years since its inception. Winning such an award is thought to be essential to the future career development of a filmmaker (director/writer/actor/cinematographer/etc), and could in turn bring more commercial revenue to the production company itself.

Three datasets are provided for you. You do not need to use all of them:

academy_awards.csv

This table lists all Academy Awards nominations and wins since 1927. 10,395 rows & 7 columns.

Source: https://www.kaggle.com/unanimad/the-oscar-award

imdb.csv

This table lists popular movies from 1986 to 2016, scraped from IMDb, providing industry metadata and general user ratings about the films. 6820 rows & 15 columns.

Source: https://www.kaggle.com/danielgrijalvas/movies

rotten_tomatoes

This folder includes movie ratings from critics (in contrast to general viewers in the IMDb dataset), scraped from Rotten Tomatoes.

In the movies dataset each record represents a movie available on Rotten Tomatoes, with the URL used for the scraping, movie title, description, genres, duration, director, actors, users' ratings, and critics' ratings.

In the critics dataset each record represents a critic review published on Rotten Tomatoes, with the URL used for the scraping, critic name, review publication, date, score, and content.

Source: https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset

Your Tasks

1. Data Cleaning: These three datasets have very different schemas, plenty of duplicate information, and traces of missing information. Clean and combine the datasets to remove duplicates and only retain the necessary information for your further tasks.
2. Are user ratings/reviews predictive of which films/people will be nominated for specific academy awards? Construct a model, using the variables given in the datasets, to predict the probability of a film/person to get an academy award nomination. Design a metric to evaluate the accuracy of your model.
3. Design a way to evaluate a filmmaker's 'success index' at any year. Hint: a filmmaker could be considered successful at a time even if they do not release any movies that year. For example, Steven Spielberg. Take this into account.
4. Does receiving an award help the career of a particular filmmaker? If so, how long does the 'award effect' last?
5. Write a non-technical report for a movie production company aiming to contend for the Oscars at the next awards season. Describe your findings, and make recommendations accordingly. (Should you make a movie that appeals to the general viewers, or the critics? Which genre should you produce in? etc.)

Submission Your solution paper should include a 1-page Summary Sheet. The body cannot exceed 20 pages for a maximum of 23 pages with the Summary Sheet AND 2-pages non-technical report inclusive. The appendices and references should appear at the end of the paper and do not count towards the 23 pages limit.

*DATA: https://pan.baidu.com/s/1GhYoQdacrayBN3Zn4b-wXA ,Code：1s3v