AARON BROWN - DSC 630 - FALL 2023

# FINAL PROJECT

Predicting Anime Score Using MyAnimeList

introduction

# objectives

1. To gauge the aspects that lead to Japanese-made animated works' success according to user ratings or "score".

2. Obtain a user-oriented outlook and forecast of “success” using the data.

# introduction

## THE VALUE OF ANIME

Anime series have become the most popular foreign-language TV shows in the US, comprising 30.5% of the market. In 2021, over half of Netflix's subscribers viewed anime.

Anime's worldwide market is estimated to have a worth of more than $27 billion, with projections expecting a growth of nearly 50% by 2028.

Grasping the evaluation of an anime's rating and components that contribute to ratings is advantageous for both fan and investor views.

Recent utilization of streaming platforms such as Netflix, Hulu and Crunchyroll is projected to fuel this sharp rise in popularity, further enhancing the value of studies like these that make use of such information.

# about the data

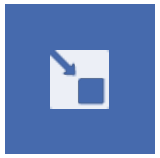
The dataset utilized in the project was taken from Kaggle - it had been harvested from MyAnimeList, a website that houses data from millions of users and tens of thousands of animation productions.

## Features retained from the data:

- title: the name of the animated productions in the data.
- episodes: the number of episodes for the productions.
- type: the classification of the productions (movie, series, etc.).
- favorites: the total number of favorites the titles have received.
- score: (1-10) average rank given by the members or subscribers.
- members: total number of subscribers to the titles.
- studios: the production studio responsible for the titles' adaptation.
- demographic: targeted age/gender group.
- genres: listed genres for each title, can have multiple.
- synopsis: a short overview of the plot for each title.

# data cleaning and preparation

Eliminating around 8K absent values, resulting in over 15K rows.



Converting all strings in the features to lowercase and tokenizing the text.

Substituting the mean for the "Unknown" episode count.

Creating a feature called "tags" which will include information from "type", "studios", “demographic", "genres" and “synopsis”.

Cleansing the "genres" and "synopsis" fields of commas, superfluous whitespaces, as well as other punctuation marks.

Merging and vectorizing the text of the fresh "tags" component to calculate the cosine similarity between their vectors.

Stripping any punctuation from numerical values, which were initially strings.

# METHODOLOGY

## Model Selection

For the purpose of this project, the Linear Regression Model was chosen.

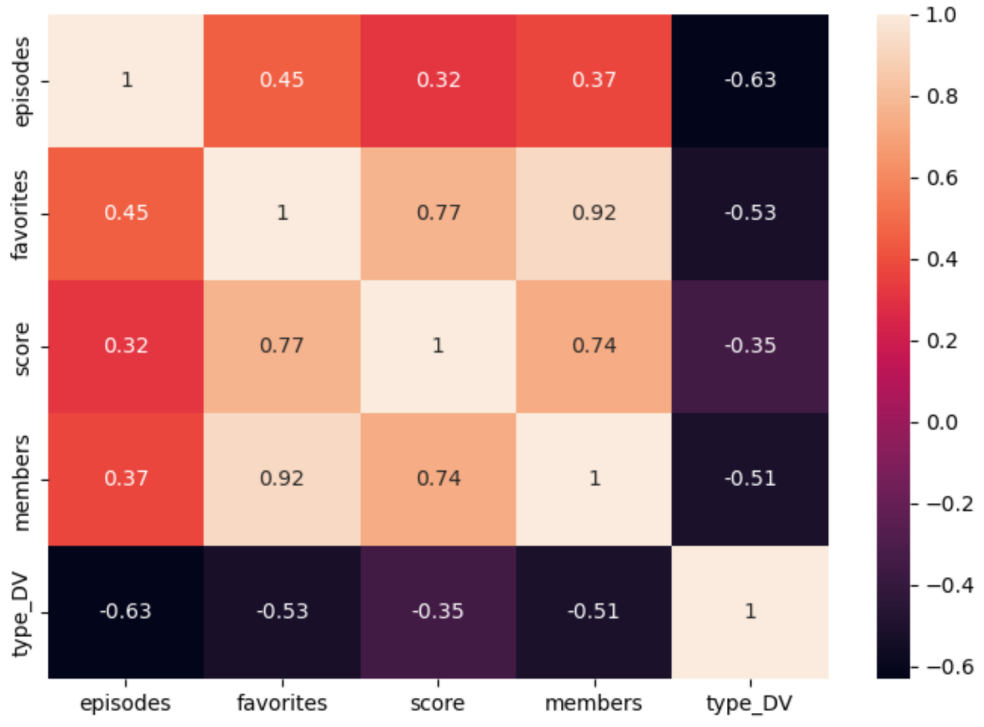
This model was chosen over other alternatives mainly because of its straightforwardness and implementation.

This model, however, has its deficiencies, like overfitting or under-fitting which can be caused by its sensitivity to extreme values.

# METHODOLOGY

**Generating visualizations to examine features and relationships.**

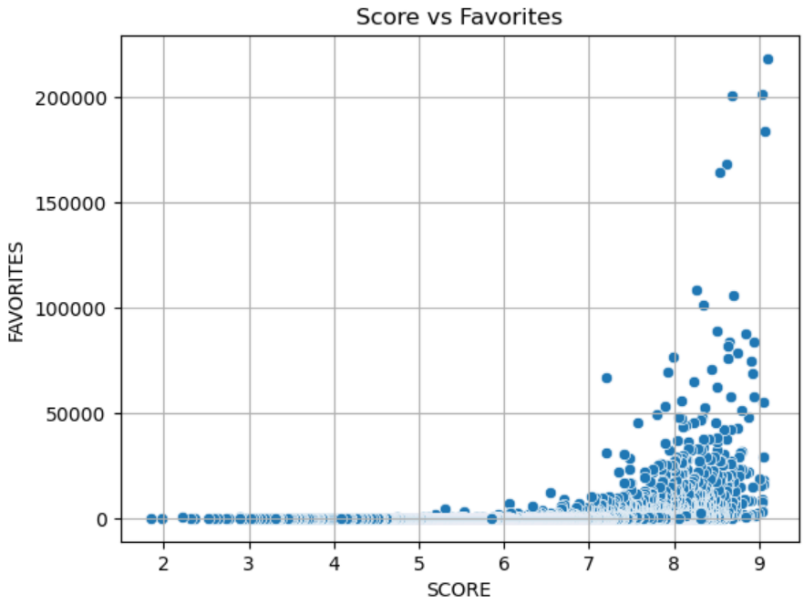
```
corr_matrix = mal_data.corr(method = 'spearman')
fig, ax = plt.subplots(1, 1, figsize=(7,5), tight_layout = True)
sns.heatmap(corr_matrix, annot = True)
plt.show()
```

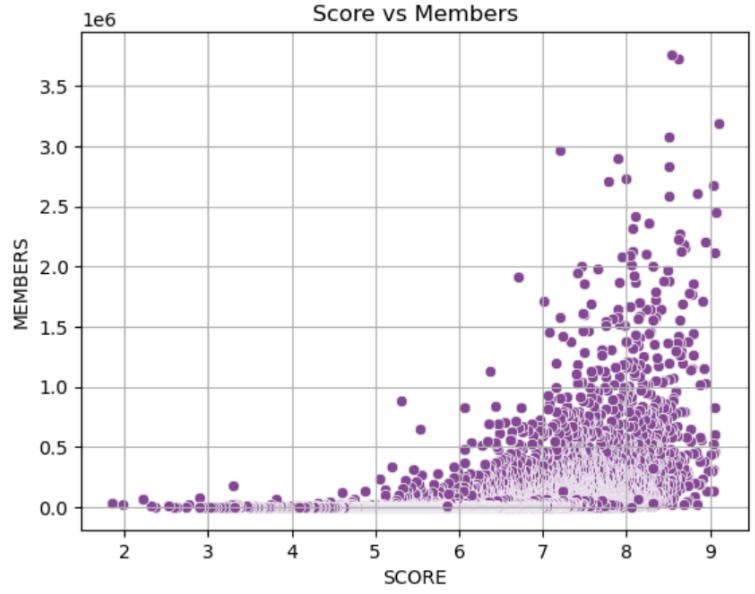


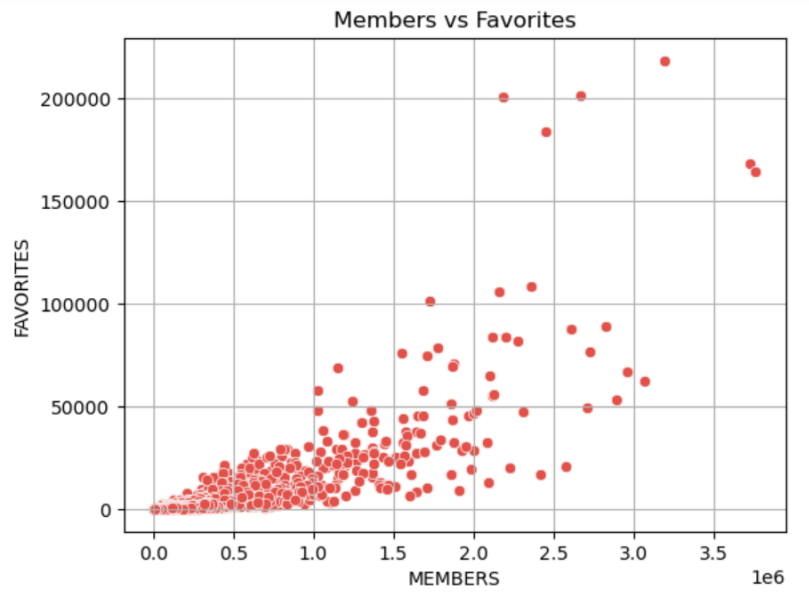

A correlation matrix was used to easily analyze the links between the factors. It is evident that 'members' and 'favorites' correlate with the target variable 'score'. This makes sense, as the more members an anime has, the more favorites it will likely receive. It is plausible that higher scoring anime may have more members too.

Scatter plots depicting the correlations between the target, "members" and "favorites" are shown below.

Scores rise in correlation with member numbers and favorites. Likewise, the more members, the more favorites gained.



# METHODOLOGY

APPLYING AND EVALUATING THE MODEL

The data was divided into training and test sets; the test portion is composed of 25% of the data, using the "score" variable as the target.

The Linear Regression is to be used to forecast "score" values.

Root mean square error was calculated and the outcome was 0.8209.

# CONCLUSION

The model satisfies the objective of this project, and can be used for decision making, and building recommendations for potential titles to explore.

1. The Linear Regression model gave us an RMSE that is very low, suggesting that it fits the data well and provides accurate forecasts.

2. The model's performance is satisfactory, implying that the linear regression model is capable of forecast score accurately.

3. Seeking tactics to drive up member count appears to be the most effective way to improve popularity and rankings for stakeholders.

4. It is conceivable that the Linear Regression Model's low RMSE could be due to overfitting, and thus it is important to consider the implications and assumptions of such a model.

# references


Dataset: Anime_Data_2023
https://www.kaggle.com/datasets/peacelife/anime-data-2023

What is Linear Regression?
https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/

LinearRegression¶
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html