HULU

Data mining and its analytics is helping Hulu in finding the targeted audience choice

Submitted By: Team Blue Khare

Submitted To: Shivanjali

Team Members:

1. Amar Raj Gautam (Lead)
agaut4@unh.newhaven.edu
Student ID: 00763645

3. Sakshi Gowda
sgowd2@unh.newhaven.edu
Student ID: 00761802

2.Rathna Maheswari Nidamanuri
rnida1@unh.newhaven.edu
Student ID: 00794186

Table of contents

Abstract

Hulu is an online movie and TV show streaming platform owned by the Walt Disney Company. Hulu is exclusively in the United States and is not available in other countries. This is a subscription streaming platform launched in October 2007 which is like platforms like Netflix and Amazon Prime. It hosts several video content like movies, TV shows, reality shows, and original Hulu-produced content as well. They are a customer-centric company constantly looking for ways to improve user experience, one of them is a precise content recommendation system that will suggest the most appropriate content to each user based on their previous watch history, and search history, by studying similar audience demography. Our goal in this project is to make use of data mining techniques and data analysis to help find an accurate target audience for various types of content that is available on Hulu. Here, we will be considering the current performance of Hulu, studying the current Hulu audience demography, and finding ways to improve suggesting content that the audience will prefer to watch based on their watch history. The project was a success with few limitations and scope for further improvement.

Introduction

Hulu is one of the leading subscription-based streaming platforms in the US. Currently, Hulu has gained 48.5 million paid subscribers as of 2023. This is due to its huge library and intuitive user interface attracting viewers. Hulu has proven to be among the fiercest competitors in this highly competitive market. The Walt Disney Company, Comcast, and 21st Century Fox jointly own it. Today, Hulu is a globally recognized company owing to its rich collection of up-to-date drama series and award-winning exclusive shows.

In this Big data era, where we have collected a huge amount of data on viewers' preferences and watch history along with the growing catalog of video content, it is becoming a challenge to accurately suggest content to the right target audience since there is so much more data to analyze. This has led to all the platforms like Hulu looking for ways to develop content recommendation systems to be more precise. In this project, we are studying the dataset consisting of metadata about the movies and TV shows such as the title, director, and cast of the shows/movies. Details such as the release year, the rating, duration, etc. We use this data to Understand what content is available in different States, find out if Hulu has more focus on TV Shows than movies in recent years, what the situation of audience engagement such as the imdb voting for now, and what will be the future? Is a feature like runtime useful for predicting future users?

In this project, we are using data mining techniques and data analysis to help find an accurate target audience for various types of content that is available on Hulu. Here, we will be considering the current performance of Hulu, studying the current Hulu audience demography, and finding ways to improve suggesting content that the audience will prefer to watch based on their watch history and the forms of videos that the audience is watching.

Related Work

"Measurement Study of Netflix, Hulu, and a Tale of Three CDNs"- has studied a database of content delivery networks (CDNs) associated with platforms like Hulu and Netflix. This paper proposes new video delivery strategies that can significantly improve the user QoE by effectively utilizing multiple CDNs while still conforming to the business constraints. They study the existing strategies used by these platforms to select CDNs that are best to deliver appropriate content to the end users. Data mining techniques: Clustering is used for grouping CDN performance data or user engagement data into clusters to identify patterns or trends in how different CDNs impact the streaming experience.

"SVOD platform audience. The case of Netflix, Blockbuster, Hulu, and HBO"- It refers to databases of computer and mobile users accessing Netflix and Hulu from July to December 2018. This paper shows how SVOD- the 'subscription to video on demand' model used by platforms like Hulu has affected both the audience and advertising. In recent years, efforts have focused on finding a solution to audience fragmentation, aiming to optimize budget and advertising campaign revenue by studying audiences' preferred platforms to watch content.

"A comparative study of video recommender systems in big data era"- Comparing video recommendation technologies of four famous companies: Netflix, Google, Hulu, and Amazon to understand the basic differences between their recommendation algorithms and investigate the pros and cons. The dataset used: Netflix, Hulu, Amazon Prime, and Google databases on various recommendation systems used and their performance data.

"Hulu video recommendation: from relevance to reasoning" -This paper focuses on improving Hulu, a well-known streaming service's video recommendation system. To improve the entire user experience and satisfaction with the recommended material, it is intended to move beyond conventional relevance-based recommendation systems and add reasoning mechanisms. Dataset used: Hulu content recommendation system and its performance dataset, Hulu subscriber's dataset.

"On the Feasibility of Prefetching and Caching for Online TV Services: A Measurement Study on Hulu" - In this paper we investigate the advantages of having such a prefetching and caching scheme for a free hosting service of professionally created video named "Hulu"

Proposed Methodology:

The proposed methodology for this project is the CRISP-DM framework:

Understanding Business: Understanding the project's objectives and the business problem is the first stage. Here, the objective is to forecast demand and

sales for Walmart stores with accuracy while accounting for events like sales, markdowns, and economic indicators.

Data Understanding: The next stage is to collect and examine the data to learn more about its attributes and caliber. The data in this instance was taken from the Walmart portal on Kaggle and includes sales data for 45 stores along with details on the weather, holidays, CPI, gasoline prices, and unemployment. Roughly 95% of the data are accurate.

Data Preparation: To prepare the data for analysis, this stage entails cleaning and converting it. This could involve resolving data discrepancies, outliers, and missing numbers in addition to developing new features and aggregating data at various granularities.

Modeling: In the modeling process, a machine learning algorithm is chosen and trained to forecast demand and sales using the prepared data. Time series models, regression models, and machine learning models like neural networks or 4 random forests are examples of potential algorithms. Proper performance measures should be used to assess the model, and holdout sets or cross-validation should be employed to validate it.

Evaluation: In the evaluation step, the model's accuracy and reliability are evaluated, its performance is assessed, and whether or not it satisfies the business goals and requirements is determined. Possible tasks include comparing the model's predictions with real sales data and pinpointing areas that need improvement.

Deployment: The model will then be put into production and integrated into the Walmart sales and inventory management system. This may entail creating an API or web interface for users to access the model's predictions, as well as monitoring and updating the model as needed.

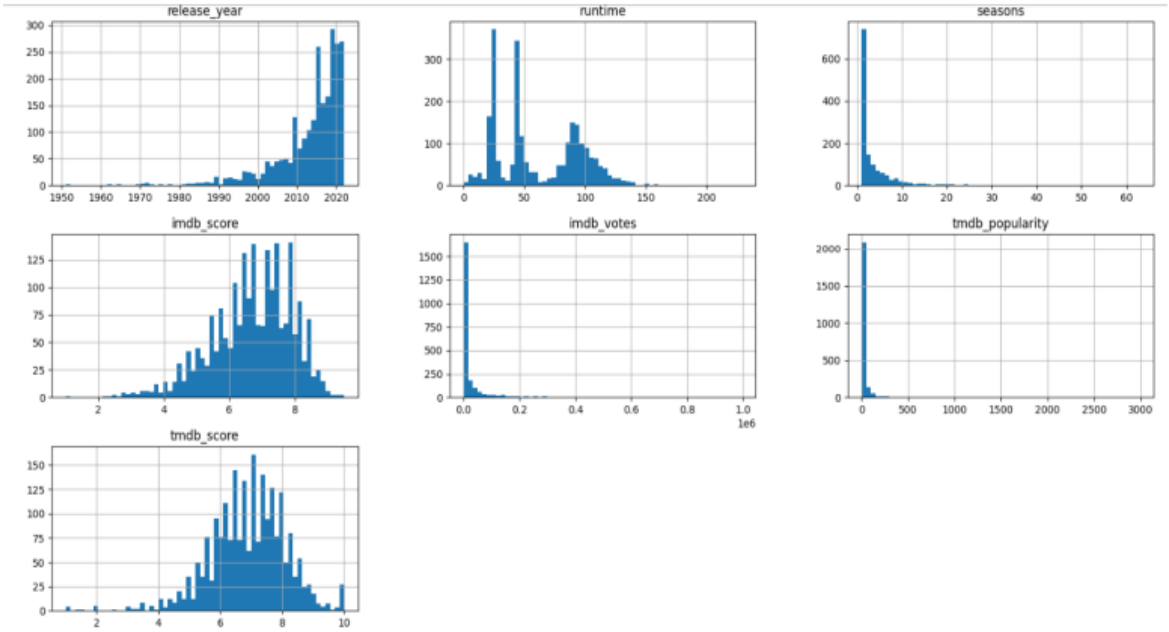
Overall, this methodology follows the CRISP-DM framework and involves iterative cycles of data preparation, modeling, and evaluation to develop an accurate and reliable method to help Hulu find the targeted audience choice.

Data Understanding:
Table of Data:

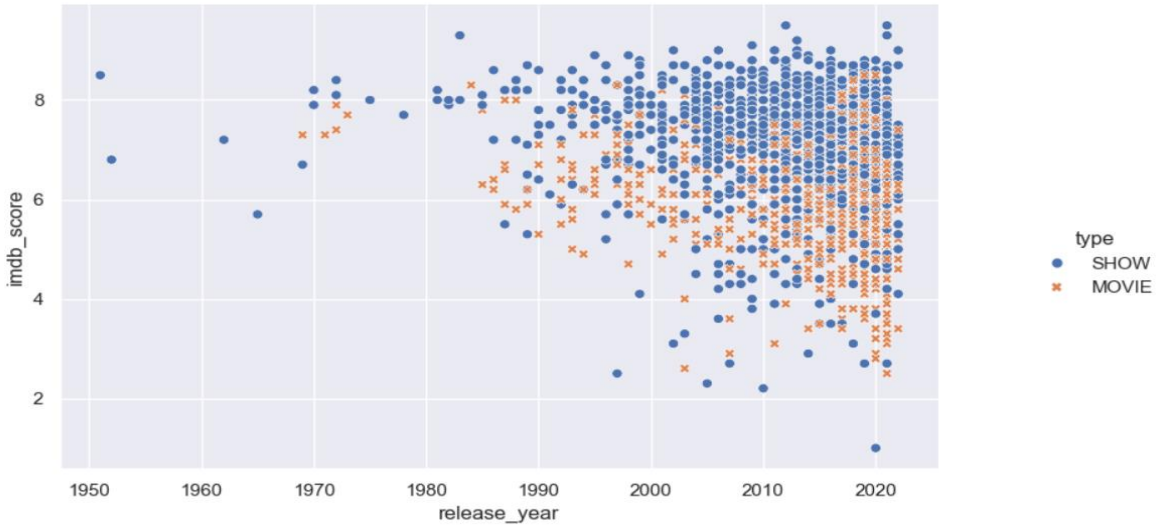
6]:

	release_year	runtime	seasons	imdb_score	imdb_votes	tmdb_popularity	tmdb_score
count	2398.000000	2398.000000	1330.000000	2232.000000	2231.000000	2348.000000	2238.000000
mean	2013.417014	61.518766	3.936090	6.699149	28286.883909	27.870533	6.885657
std	8.483634	35.102738	5.020168	1.208236	78020.103846	91.997761	1.215606
min	1951.000000	0.000000	1.000000	1.000000	5.000000	0.272976	1.000000
25%	2010.000000	25.000000	1.000000	5.900000	621.500000	4.789250	6.100000
50%	2016.000000	48.000000	2.000000	6.800000	3451.000000	10.703000	7.000000
75%	2019.000000	93.000000	5.000000	7.600000	17056.000000	23.910250	7.700000
max	2022.000000	229.000000	63.000000	9.500000	996056.000000	2989.846000	10.000000

Histogram for the possible Integer data type:

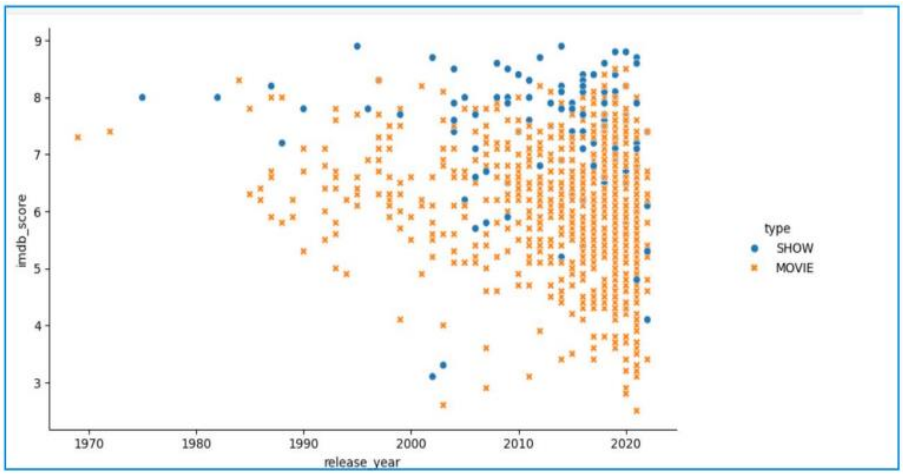


Scatter Graph showing type of shows released over years

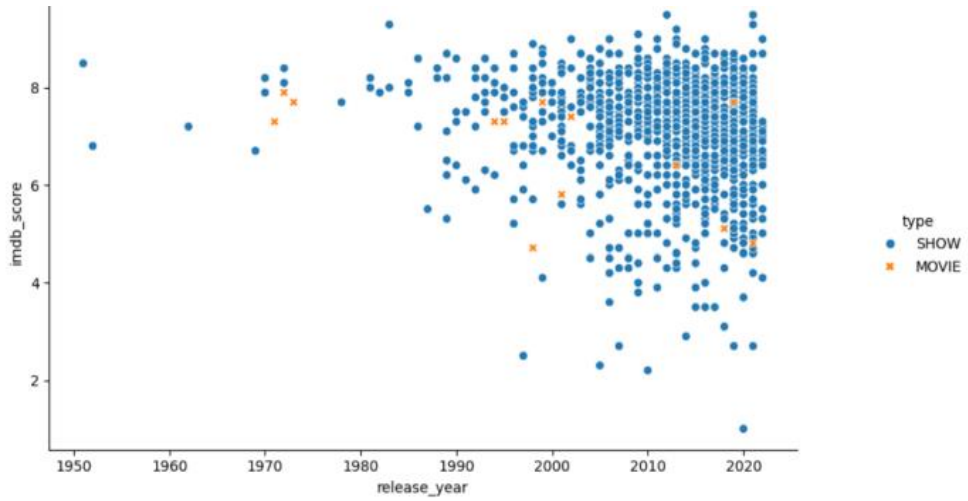


Scatter Graph Based on runtime

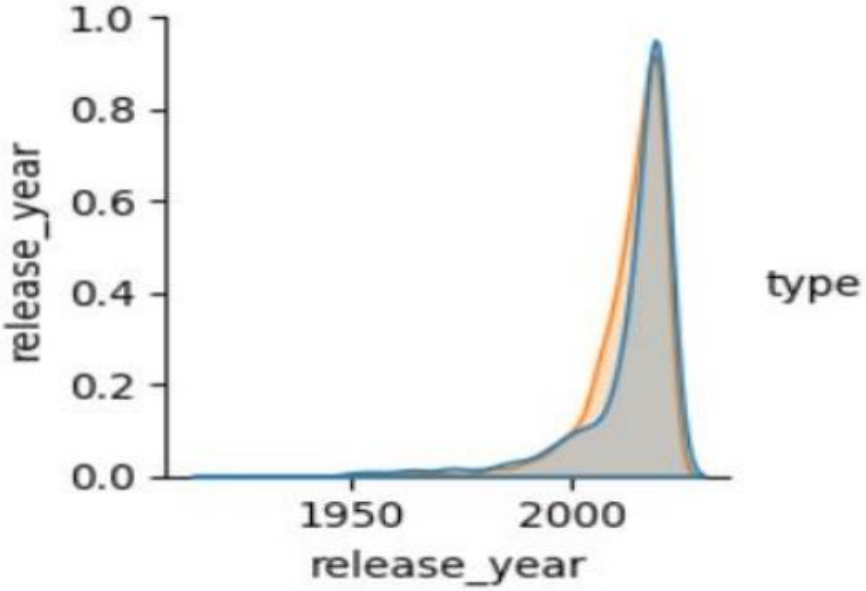
Scatter Graph showing the different types of shows based on run time greater than 50 min



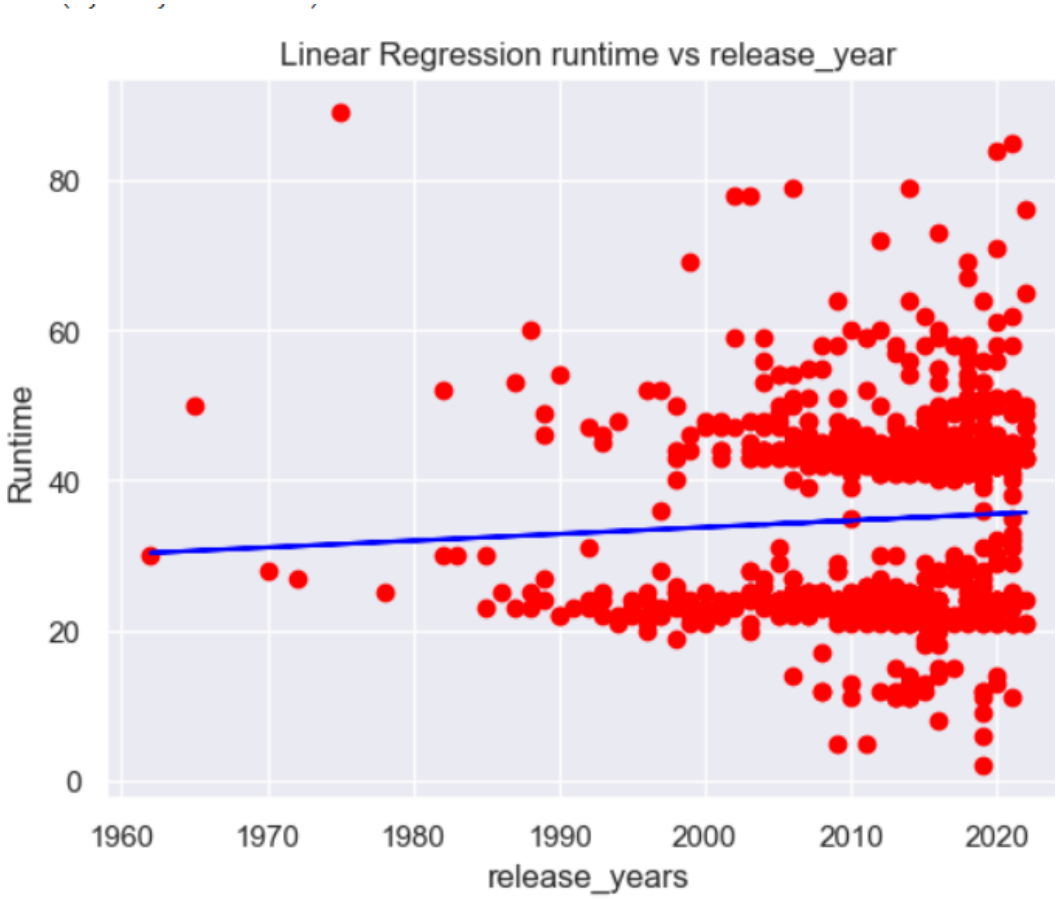
Scatter Graph showing the different types of shows based on run time Less than 50 min



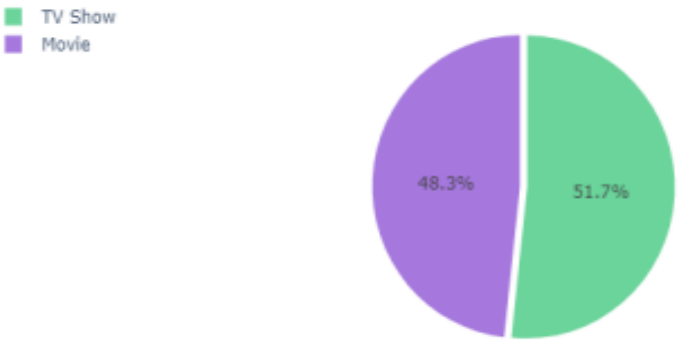
The distribution of films and TV series according to their release years and countries of origin is shown in this scatter plot. The greatest number of films and television series originate from the United States, with a focus on the more recent years, particularly those after 1980. There are also a lot of films and TV series from Japan, the UK, and multi-country projects, mostly from the 1980s onward. A moderate amount of films and TV series are broadcast in Australia, Canada, Israel, Russia, France, Spain, and Ireland. These releases appear to be mostly from after 1980. Less films and TV series are plotted on the graph for nations like Germany, South Korea, New Zealand, Norway, Mexico, Sweden, Venezuela, Italy, Denmark, and Colombia; the majority of releases have occurred in more recent years. All nations had extremely few data points before 1960, and production noticeably increased starting in the 1980s. In conclusion, this graph depicts the globalization of film and television productions during the previous century, with a notable increase in output over the last half of the timeline, particularly in the United States and a few other major nations.



This graph shows the distribution of release years for two different media formats. It is a density plot, also known as a kernel density estimate. The blue curve represents a high density of releases for that media format around the year 2000. The density reduces dramatically before and after this peak, notably beyond 2000. The orange curve shows a more progressive rise, beginning about 1950 and culminating just before 2000. The density falls after this peak, although not as dramatically as the blue curve. The blue curve has a more concentrated peak, suggesting a smaller time frame of high releases, whereas the orange curve has a more spread-out peak, indicating a larger time frame of releases.

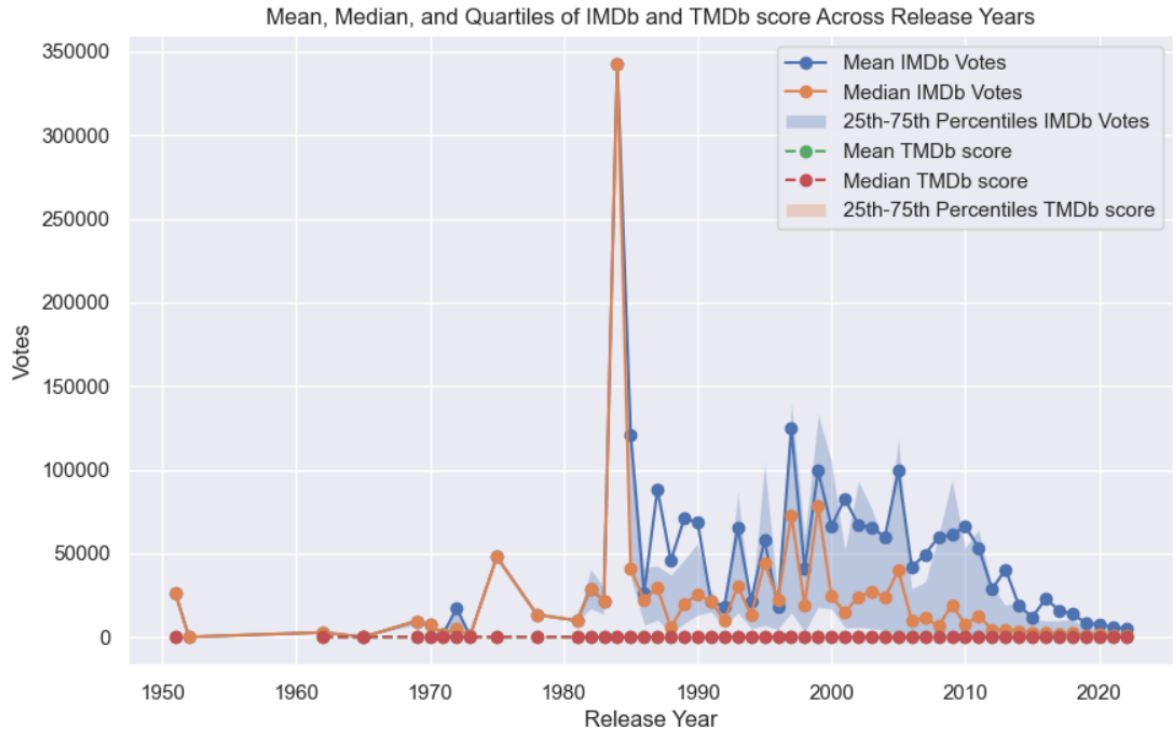


Data Visualization

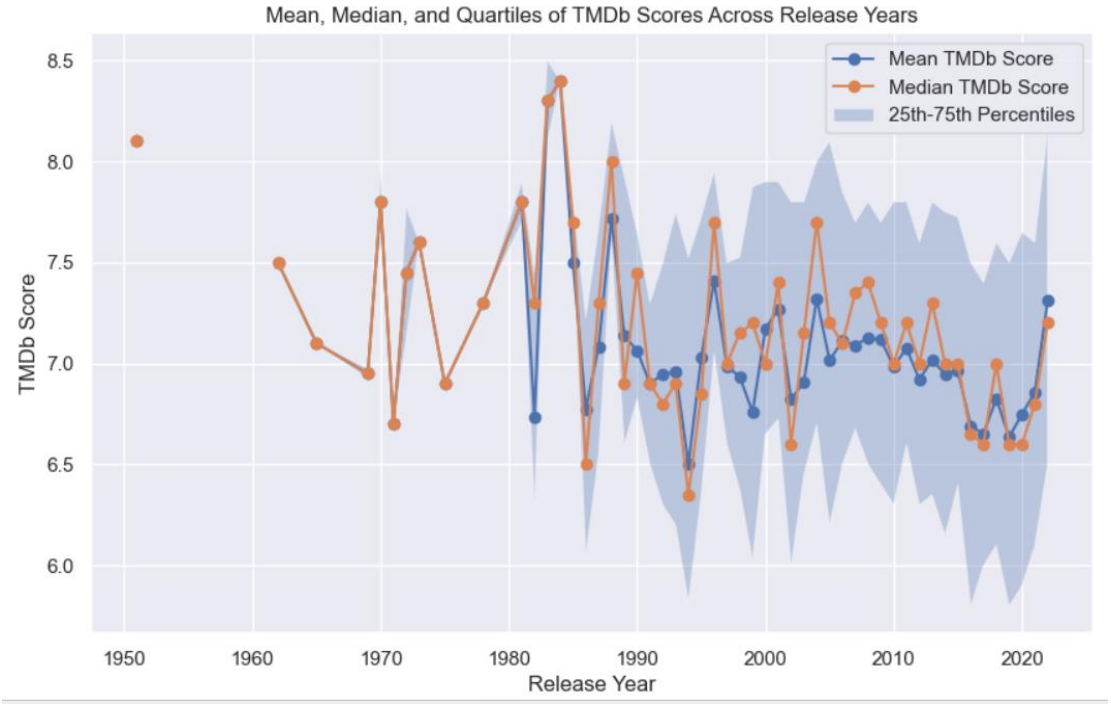


In this below graph its showing the number of votes in the different year. The blue line indicates the mean imdb votes and the orange line indicates the median

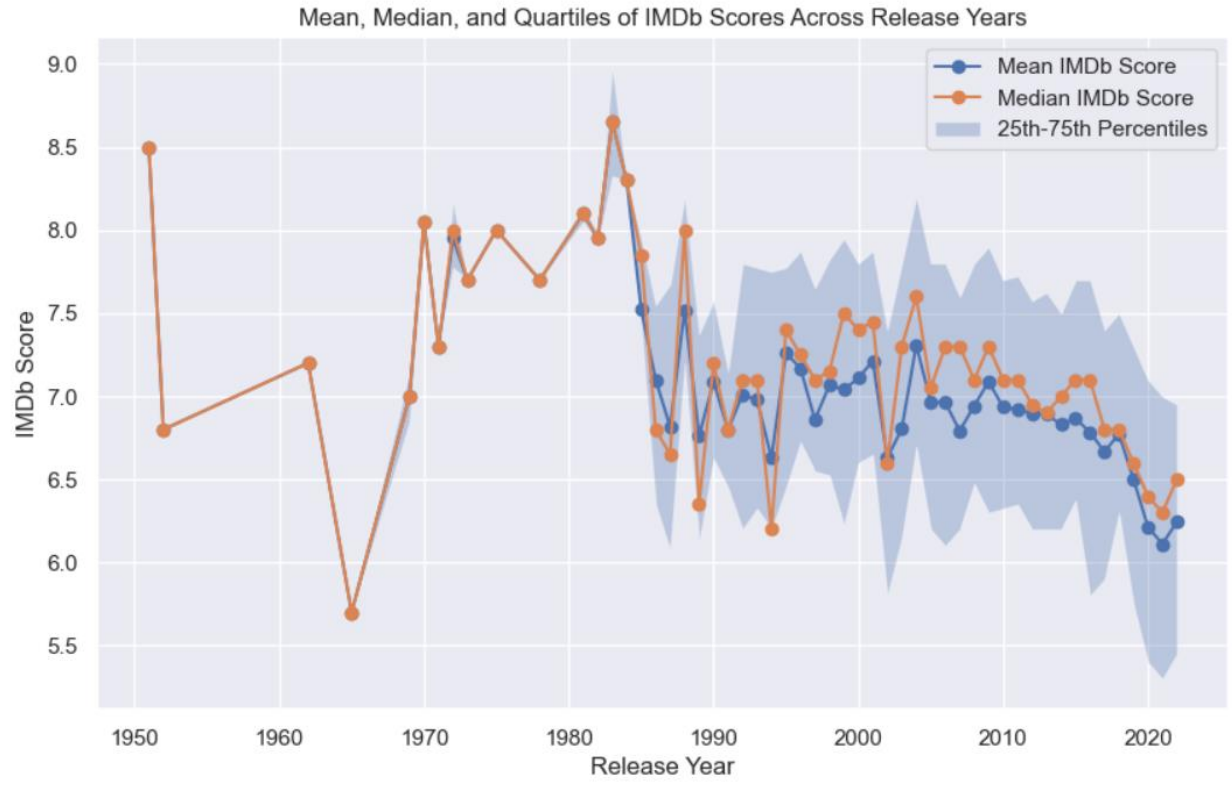
of imdb votes similarly green and red represents the mean and median of the tmdb score.



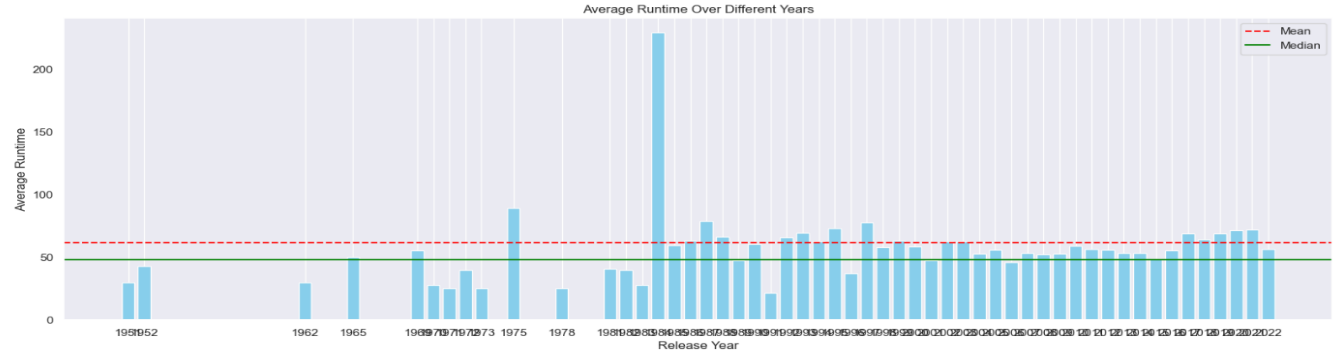
The below graph is only showing for the tmdb score.



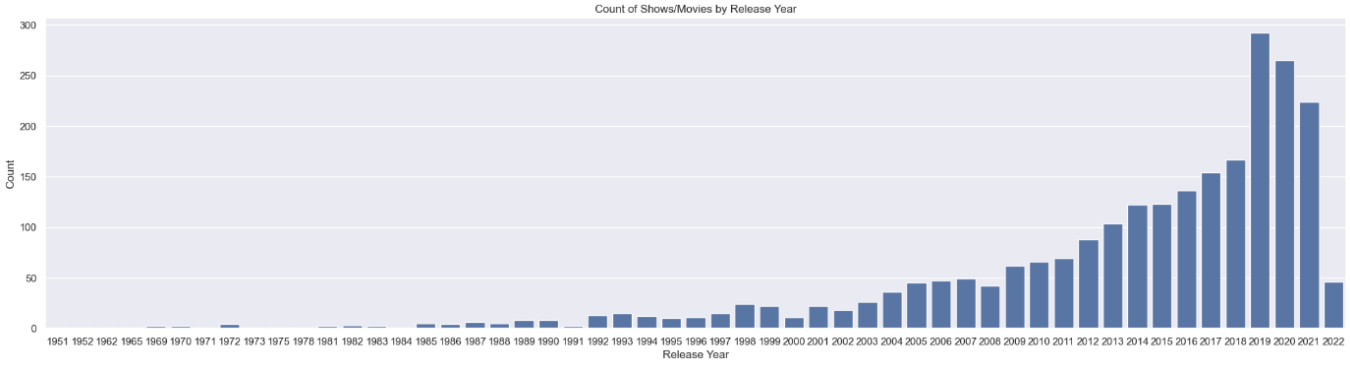
This below graph shows only the imdb score



The below graph shows the mean runtime of the shows



The below graph shows the content released over years



Results

With the implementation of different Algorithms:

Linear regression:

Train R2

0.0030531714921741004

Test R2

-0.00652378515213381

Best hyperparameters for Ridge: {'alpha': 10}

Best hyperparameters for Lasso: {'alpha': 0.1}

*******. Ridge Regression ********

Train r2_score 0.0030531713972477004

Test r2_score -0.006523786136330978

******* Lasso Regression ********

Train r2_score 0.003052415722606394

Test r2_score -0.006524821432379646

Random Forest:

Random Forest Train r2_score 0.0810932649940227

Random Forest Test r2_score -0.049981929288680504

After Hyperparameter Tuning

Best parameters: {'max_depth': 5, 'max_features': 'sqrt', 'n_estimators': 200}

Random Forest Train r2_score 0.07251952938053119

Random Forest Test r2_score -0.03571639158610718

Ada Boost:

Ada Boost Test r2_score -0.003257485143689287

Ada Boost Train r2_score 0.04102009566609177

After Hyperparameter Tuning

Best parameters: {'base_estimator__max_depth': 5, 'learning_rate': 0.01, 'n_esti
mators': 100}

Ada Boost Train r2_score 0.07812928250547013

Ada Boost Test r2_score -0.02891262252723359

Gradient Boost:

Gradient Boost Train r2_score 0.08895633290138161

Gradient Boost Test r2_score -0.04765523311734299

Extra Gradient Boosting Train r2_score 0.08895633210459586

Extra Gradient Boosting Test r2_score -0.06562689814976652

Best Hyperparameters: {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 50}

Extra Gradient Boosting Train r2_score 0.04178473412395467
Extra Gradient Boosting Test r2_score -0.008780014890866328

Results Comparison Table:

Algorithm	Train R2	Test R2
Multi Linear Regressor	0.003	-0.006
Random Forest	0.081	-0.004
Ada Boost	0.041	-0.003
Gradient Boost	0.088	-0.047

After Tuning Results

Algorithm	Train R2	Test R2
Multi Linear Regressor	0.003	-0.006
Random Forest	0.072	-0.035
Ada Boost	0.078	-0.028
Gradient Boost	0.088	-0.065

Result Discussions :

Overall Gradient Boosting initially seemed to perform the best in the training data set but in the test results, it has shown decreased results in the performance.

On the other hand, Ada boost has shown some improvement in the test results but is worse for the training data. So the selection of the best model should not rely solely on the R2 score. It is crucial to consider other metrics and conduct cross-validation.

This kind of situation often calls for revisiting the features used and considering the more complex models or employing different techniques to improve the model performance.

Limitations :

Several limitations and errors may be affecting the results as follows:

The model does not fit the data set, it looks like the model was forcefully implemented.

There might be the underfitting of the data so, the model was not able to capture the underlying patterns or relationships.

Besides these, there might be a possible error in the data set as Hulu was released on March 12, 2008, but the data set had data from 1951.

Conclusion:

We have implemented the machine learning models and we used algorithms like random forest, ada boost, and extra gradient boost and see the results of train and test data.

However, the limitation needs to be addressed that looking into the value of the linear regression model was only able to explain 0.31% of the variance in the training data set and the negative value of R2 in the test indicates that the model performed worse than the model that predicts the mean of the target variable. It suggests that the model's prediction is giving no meaningful improvements over simply using the mean value of the target.

Future Work:

For data accuracy, we can reconfirm the dataset of Hulu as the Hulu was started from 2008 but the dataset was available from 1951. More complex models can be implemented which can catch the pattern from the inconsistent type of data. we could find different findings like is IMDB score and TMDB score interlinked. Based on findings from the connection of different features we would be able to get more accurate results.

Reference:

1. S. Kabiraj et al "Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm, "2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT),2020, pp. 1-4, doi: 10.1109/ICCCNT49239.2020.9225451.

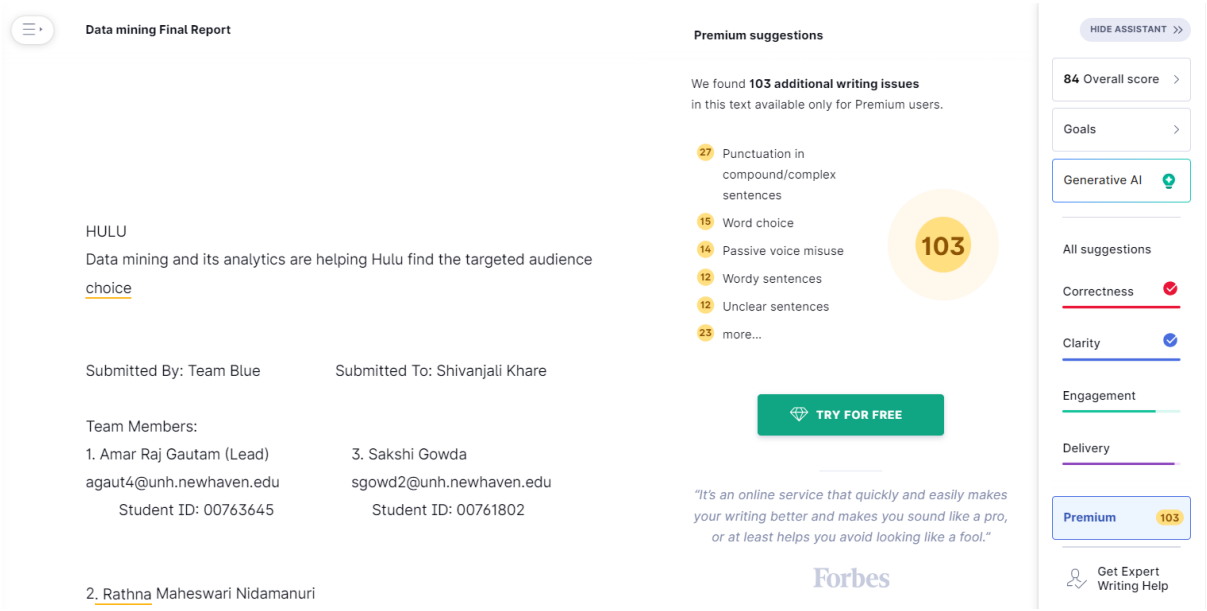
2.Hamid, S., S. Bukhari, D. Ravana, A. A. Norman, and M. T. Ijab. 2016. Role of social media in information-seeking behavior of international students: A systematic literature review. Aslib Journal of Information Management 68 (5):643–66. doi:https://doi.org/10.1108/AJIM-03-2016-0031. [Crossref] [Web of Science ®], [Google Scholar]

3. P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "SemiBoost: Boosting for Semi-Supervised Learning," In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 11, pp. 2000-2014, Nov. 2009, doi: 10.1109/TPAMI.2008.235.

Github

https://github.com/agautam4/Data-mining-Team-Blue/blob/main/Final_datamining.ipynb

Proof Reading –



Data mining Final Report

by Amar Gautam

General metrics

15,554	2,329	157	9 min 18 sec	17 min 54 sec
characters	words	sentences	reading time	speaking time

Score

This text scores better than 84% of all texts checked by Grammarly

Writing Issues

103	✓	103
Issues left	Critical	Advanced

Unique Words

Measures vocabulary diversity by calculating the percentage of words used only once in your document

31%
unique words

Rare Words

Measures depth of vocabulary by identifying words that are not among the 5,000 most common English words.

37%
rare words

grammarly Report: Data mining Final Report

Word Length

Measures average word length

4.9
characters per word

Sentence Length

Measures average sentence length

14.8
words per sentence