Data Mining Project Description

COVID -19 STAY AT HOME: FORECAST MOVIE RATING(SCORE) ON IMDB

Greeshma Vijayakumar | MIS 545 | 07/27/2020

Contents

BACKGROUND

During the COVID-19 stay at home situation, there has been multiple studies on human behavioral aspects including change in human habits and hobbies on a daily basis with most of the population across the globe either working from home or taking online classes. With respect to the recent data released by Comcast for USA it was noticed that people are actually watching so much more video/movie content on websites and TV that weekday and weekend video content viewing habits are now blurring together. This is from the initial research analysis of what the TV watching trends during COVID-19 are showing, including that households are viewing a full workday's worth of content more each week. Data from Comcast also shows that there is a slow decrease in DVR usage, while at the same time there's a ***50% increase in video on demand usage.*** There is also comcast data asserting the customer asks on "what to watch" and "surprise me." This suggests that people have watched more of the shows on their lists and are looking for something new. While they are digging in for something new , the data mining project here will help them make this an easier task, as one of the primary deciding factors for watching or reading content on the internet or TV now is review / rating / scoring.

PROBLEM STATEMENT

This individual project covers a data mining report on predicting how we can help the people watch valuable content on TV or the internet by providing them a rating or review score for each movie they wish to watch. Taking into account, the value of people's time during this pandemic I would like to build a credible movie rating or scoring prediction model using the IMDB data set which is one of the top 10 movie scoring websites trusted by the audience. This will help the audient choose their movie watching content wisely without having spent their valuable time on something which would not satisfy their interests or entertainment needs. While we focus on the video content watchers as our primary consumers of this data, we also have to keep in mind the commercial success of a movie generates a tremendous amount of profit for the video makers as well. So, our secondary consumer of this data model would be the film companies / video content creators.

This is an important problem statement especially during these times because based on the massive movie information available on the internet and TV , it would be super useful to develop an understanding of what are the important factors that make a movie more watchable or interesting than others available in the pool. This saves people's time taken to decide on what to watch, help people watch valuable content based on their interest, help makers create audient favorite data etc.

With that said , I would like to focus my data mining algorithms to what kind of movies are more in demand, in other words, what kind of movies can a get higher IMDB score which will in turn help the audient and movie maker decision making easier.

PROJECT DATA DESCRIPTION

To organize the data and the code I have created a githib repository of the project. You can view the complete data below. The primary dataset is collected Data World and IMDB websites. I have also collected and inserted datasets from TMDB for additional reference and data analysis purposes.

https://github.com/greeshvMIS545Project/PredictRatingMoviesTVShows/commits?author=greeshvMIS545Project

From the initial look of the data, following are my observations: there is metadata for

- ~5000 movies and it contains 28 different variables.
- 19 records do not contain the value for whether the movie is color or black & white. Further data digging will help me understand if we need to delete these records from the dataset as part of data cleaning.
- Spanning across 66 countries in a time range of 100 years.
- 2399 unique director name records and currently uncounted number of actors / actresses.

The identified dependent variable here would be class named ***"imdb_score"*** while the other 27 variables are possible independent variables which are predicting factors for IMDB score or rating.

REFERENCES

- I also referred to the dataset provided in the suggested Project Dataset list. https://www.imdb.com/interfaces/
- https://www.raindance.org/top-10-film-review-websites/
- https://www.mlive.com/coronavirus/2020/05/covid-19-tv-habits-show-a-staggering-change-in-viewing-hours-and-what-were-watching.html
- https://www.comscore.com/Insights/Blog/US-TV-Viewing-Is-Increasing-During-Coronavirus-Pandemic