Movie Recommendation System

Abstract

Due to the importance of entertainment in a time where issues such as high inflation, inflated cost of living and worldwide pandemics are keeping people indoors and confined to their living rooms, people are flocking to have their boredom be rescued by movie streaming services such as Netflix, Disney Plus, Prime Video, Hulu and more.

Movie selection has become ever more important in today's world as watching movies is one of the most popular activities during lock downs and efforts to save money. However, it is not as easy of a task as it seems. Popular streaming services have bombarded individuals with a list of thousands of titles that are hard to choose from.

This study aims to solve the issue of movie selection by building a movie recommendation system using logistic regression analysis and Mean Squared Error (MSE) will be used to measure model performance and accuracy.

1. Description of Applied Problem

Movie selection has always been a problem even before the birth of streaming services, although to a much lesser extent. Going to the movies used to be a big deal as watching a movie in a theatre was uncommon and the choice of movie was important. Choosing a bad movie could be a detrimental waste of time and a waste of hard-earned money.

Similarly, choosing a movie from a streaming service can be daunting. There are many streaming services out there and people usually subscribe to more than one. Choosing from these long lists of titles can be intimidating as individuals typically only have a few hours to unwind, relax and enjoy their movie choice for the evening.

Categories such as new releases, comedy, action, romance and many more are filled with various movie titles people have never heard of or simply aren’t sure if the content matches their interests. This reveals a grave underlying problem of there just simply too many movies available to watch and not enough information about whether a certain individual will enjoy them or not.

Movie choices on streaming services are typically listed with a title, description, genre and actors that were cast in the film. This is sometimes not enough information provided to accurately gauge the quality of the movie, yet alone whether the movie contents align with the user’s interests and their idea of what an entertaining movie is.

These days, people are much more isolated and stuck inside their homes due to the COVID pandemic and rising costs of living. Food, housing and entertainment costs have all skyrocketed, boosting the popularity of simply watching movies indoors and

saving money. Furthermore, some people only have a limited amount of free time and those that choose to spend it watching a movie are having an increasingly difficult time choosing a movie. A bad or incorrect choice means a wasted evening/time, and frustration on the part of the users of the streaming services.

2. Description of Available Data

The data was collected and provided by GroupLens, a research laboratory at the University of Minnesota. They collected millions of movie titles and their reviews, as well as additional information such as their genre, tags, and ratings.

The dataset contains 20000263 ratings, 465564 tag applications from 27278 movies. The dataset was created by 138493 users between January 9th, 1995, and March 31st, 2015. The dataset was generated on October 17th, 2016.

3. Analysis and Visualization Techniques

3.1 Preprocessing phase

For the best and most accurate results, the data set must be cleaned using a technique called data cleaning due to the data being entered manually by humans. This data set as a result is prone to many human errors.

For instance, users may accidentally rate movies twice which causes inconsistencies in the data (multiple accounts or multiple users per account). Their rating may differ than their previous submission and cause anomalies in the result. Secondly, the data itself contains errors as some MovieIDs do not correspond to a movie due to duplicate entries and/or test entries.

The dataset contains 4 features: genome-scores, genome-tags, links, movies, ratings, tags. These 4 features all need to be cross referenced with each other using data integration techniques to provide the result.

In addition, some data may be incomplete or noisy. For example, obscure movies where only a few users have left a rating will add variability and can increase the MSE. These entries should be removed or corrected.

3.2 Analysis phase

The goal of this project is to provide users with a recommendation system for the best movies that adhere to their interests and standards, allowing for the best use of their time in terms of watching movies. The explanatory and response values are supposed to be correlated by using appropriate data mining algorithms which

are also called classifiers. This is a classification problem as the classification of a movie is either watch or do not watch, yes or no, values.

In this study, a method called logistic regression will be used to determine whether a movie is suitable to watch for an individual. This is a large set of data (20 million records of ratings), speed of analysis is not an issue here so a suitable technique for model evaluation will be the cross-validation technique. We are looking for accuracy so avoiding generalization errors as much as possible is key. Reducing bias and variance is important for accuracy in this study.

3.3 Visualization phase

With substantial amounts of data such as this (10,000 movies), visualization can be difficult to achieve as we want to provide the users with a condensed list of movies they will enjoy watching from the total list of movies.

In this project, a simple table will be provided that provides a single column list of movie titles the user will enjoy watching.

Additionally, a confusion matrix will be generated to help with predictive analysis. Confusion matrices can be effective tools for identifying errors in the study.

4. References

a. URL https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset
b. Liu, Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data, Springer Series on Data-Centric Systems and Applications, 200
c. Atanassov, A., & Al-Barznji, K. (2018). *Big Data Recommender System*. LAP LAMBERT Academic.
d. URL https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning#:~:text=Linear%20regression%20is%20used%20to,used%20for%20solving%20Regression%20problem.