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. In addition to logistic regression analysis, a K-Nearest Neighbours analysis was performed as well.

Before the training of any of these models, data preprocessing techniques were performed and resulted in some alterations of the data set. The data set was complete, meaning every single row/entry was filled out (not empty) with correct values. However, additional columns were added which included “Total movie ratings” and “Average movie rating.” These additional columns aided the analysis by determining outliers in the data set which were then removed and corrected. In addition to the new columns, 2 existing columns “zip-code” and “timestamp” were removed as they would not be used in this analysis (they do not provide any value and insight into recommending movies for users).

Based on data preprocessing and visualization techniques, it was determined that the largest factors influencing somebody’s rating of a movie is their age and their occupation. Using this information, a linear regression was performed on the data set to establish whether a user will enjoy a movie based on ratings of others with similar age and occupations. A training/test split of 80/20 was used on the data set, and evaluation of the model was done using Mean Absolute error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

The second model trained was a K-Nearest Neighbours (kNN) model to establish which movies are most like each other. Given a movie title, this model can show the 5 nearest neighbors (in terms of relativity to movie contents) to the given movie. This model uses the movies’ genres to predict the 5 most related movies to a given movie out of the data set and produces the result for the user allowing them to find similar movies to the one they are interested in.

With these 2 models, users can find recommendations using other users' ratings for movies they may enjoy or provide a movie they already enjoyed to the recommendation system and see a list of related movies they could enjoy as well.

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 are not 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 grim 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.

The “movies” data set contains the following features: movie ID, movie title, and genres (a list of genres). The movie ID corresponds to a movie ID in the ratings data set to correlate with.

The “ratings” data set contains the following features: user ID, movie ID, rating, and timestamp. The movie ID corresponds to a movie ID in the movies data set to correlate with.

The “users” data set contains the following features: user ID, gender, age, occupation, and zip-code. The user ID corresponds to a user ID in the ratings data set to correlate with.

The data set was mostly clean, it contained full and complete rows. No incomplete values were entered, and all values were of the correct type. The only data cleaning required would be analysis of outliers, cleaning up unused and unneeded columns as well as adding some additional needed columns.

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 may contain errors if some movie IDs do not correspond to a movie due to duplicate entries and/or test entries.

The additional datasets contain 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 if they are needed in the analysis.

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.

After completing the data preprocessing and analysis phases, it was determined that the data set contains complete and correct data. Meaning, all values that were entered were correct. There were no errors on cross referencing movies ID’s and no one user had rated a movie more than once.

Contained in the “users” and “ratings” datasets were the features “zip-code” and “timestamp”. These features did not add any value in this analysis, so they were removed from the data sets.

To help identify outliers, the “users”, “ratings” and “movies” data sets were merged using the movie IDs and user IDs as the keys. Two additional features were generated to help with the analysis phase and the identification of outliers. The “Average movie rating” and “Total movie ratings” features were added for each movie entry to ensure that a proper number of ratings were given to each movie to determine its overall rating and quality.

Additionally, during this phase it was determined that the genome-scores, genome-tags, links, and tags data sets would not be needed for this analysis and study.

After the preprocessing phase, we now have a data set that is clean of incomplete or empty values, no noisy or outlier values that may skew the results, a data set that contains only useful features, and 2 newly generated features to help with the analysis phase.

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 was 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.

The reason for choosing a logistic regression model is because it is a “statistical analysis method to predict a binary outcome based on prior observations of a data set” (reference #5). In simpler terms, it determines a yes or no outcome based on analysis of data. It is also able to “predict a dependent variable by analyzing the relationship between one or more existing independent variables” (reference #5). In this study, we wanted to analyze information the data set to predict a binary variable of yes or no, true, or false, on whether a user will enjoy a movie.

Before training the regression model, appropriate measures need to be taken to ensure an accurate testing phase. During the preprocessing phase it was discovered that the 2 key features impacting a movie's rating are a user's age and occupation. The focus of the regression should be on those 2 features to provide the most accurate and true result.

To create the logistic regression model, the steps taken were as follows. First, the "genres", "gender" and "title" features were dropped from the combined data sets since these are not numerical values and will not provide any useful data in this regression. Second, the features "Average movie rating" and "Total movie ratings" were dropped as well since they were mostly used for the preprocessing phase and are also not useful in this regression. Third, the data set was split into 2 groups, train/test (an 80/20 split of the overall data). The model was then fitted with the training data and was tested with the testing data in the evaluation phase.

In this study, a 2nd model was created, the K-Nearest Neighbours model. A kNN model was created to provide highly accurate predictions for which movies are related to others of the same genre. The reasoning for creating a kNN model is because it can find the k closest data points to a given data point once the model is trained with a data set. This will allow our system to give suggestions to users based on what they are watching and find related movies to recommend to them.

To create the kNN model, the steps taken were as follows. First, a pivot table was constructed using the title feature as the index, the feature user ID as column, and the rating feature as the value. Second, the kNN model was trained and fitted using this pivot table.

The last step of the analysis phase was the evaluation of the model. The logistic regression model was evaluated with Mean Absolute error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). Using the test data from the split done in the model selection phase, the results of the evaluation of the regression model are:

```
Model 1 (Linear Regression) Evaluation:
    Mean Absolute Error (MAE): 0.9394927445651309
    Mean Square Error (MSE): 1.2776115680903959
    Root Mean Square Error (RMSE): 1.1303148092856237
```

The evaluation of the kNN model was done by selecting a random movie ID from the pivot table. This movie ID was passed to the kNN model, and a list of related movies were returned, along with their respective distances to the original data point. The results differ based on the provided movie ID, but the results look like this:

```
Model 2 (kNN) Evaluation:
    Related movies for: Indian Summer (a.k.a. Alive & Kicking) (1996)

    1: With Honors (1994) (0.768242739612717)
    2: Reality Bites (1994) (0.7738022530889954)
    3: Nothing But Trouble (1991) (0.7753432366587092)
    4: Nine Months (1995) (0.7802646799424787)
    5: Father of the Bride Part II (1995) (0.7834050473328725)
```

As can be seen, the 5 closest data points (movies) are displayed.

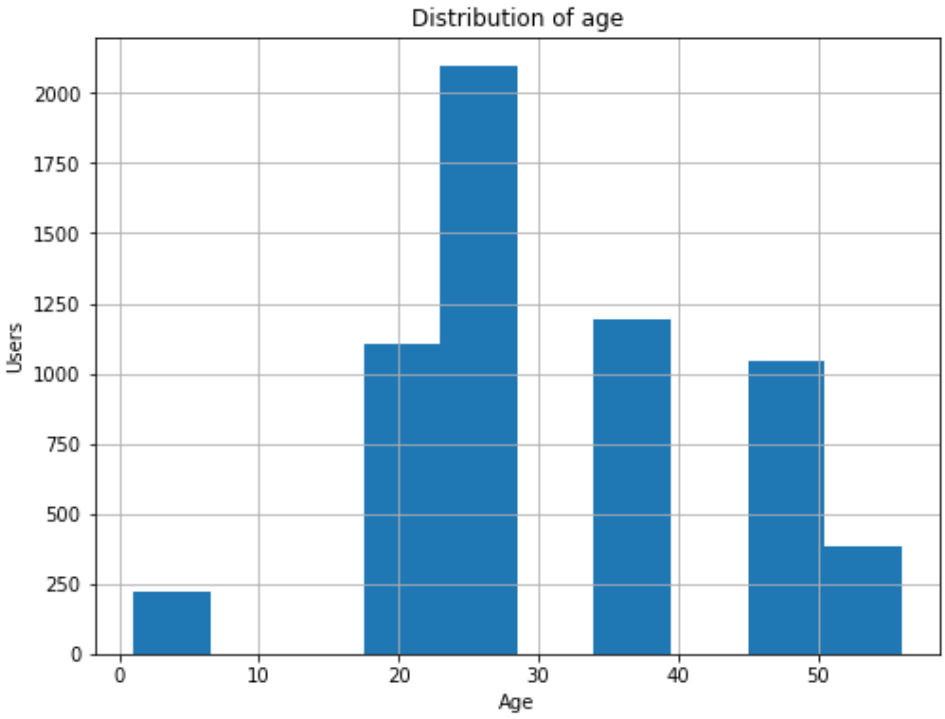
3.3 Visualization phase

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.

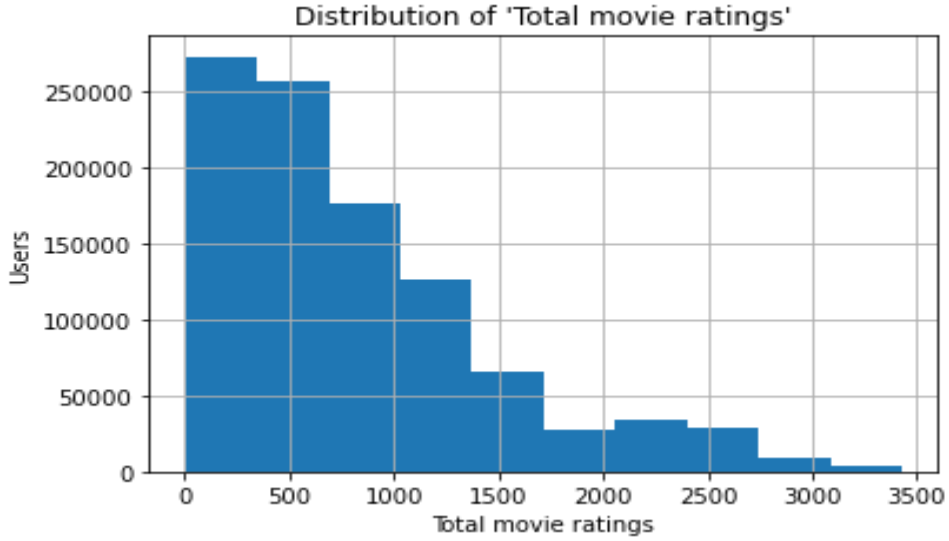
In this project, a simple table will be provided that provides a single column list of movie titles the user will enjoy watching.

During the preprocessing and evaluation phases, a series of visualization techniques were applied and examined.

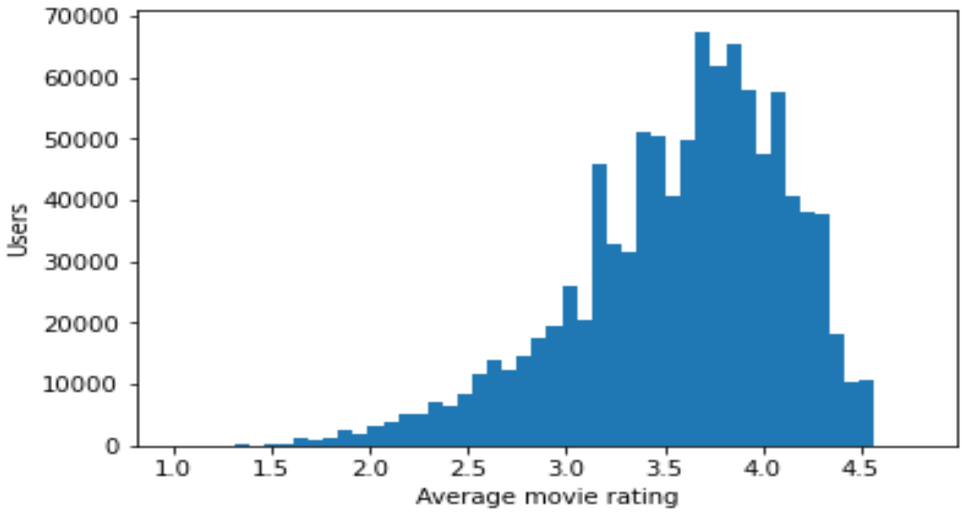
In the preprocessing phase, a histogram was first generated to show the age distribution of the users. This was done to ensure that a uniform distribution of age was in the data set and not skewed, meaning that ratings were from just a few different age groups and that there were multiple users rating the movies. The figure can be seen below:



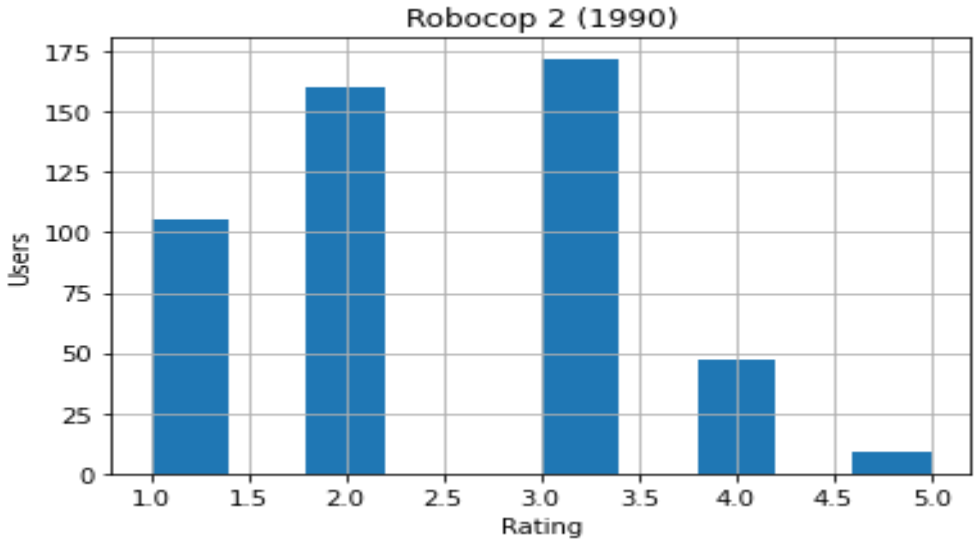
Second, a similar distribution in the form of a histogram was generated. The total movie ratings and the number of users giving out these ratings was plotted to see if a uniform distribution of ratings is given by users. The histogram can be seen below:



Like the above plot, a bar chart was generated to show the average movie rating. Ensuring that the data had an even distribution of movie ratings is key, and that they were not all skewed to be a 5/5 or a 1/5 rating. The plot below shows that the movie ratings are all even and evenly distributed:



The last figure generated in the preprocessing phase was random movie's rating plotted against the number of users who rated that movie that rating. This was done to inspect that movies were rated fairly and honestly, and no movie had a random rating given to it. This was done for a few different movies, but one example of it can be seen in the plot below:



Some data was printed to show which movies had the most/least ratings. This was to ensure that no movies had an inflated number of movie ratings. The output is shown below:

```
American Beauty (1999)                                3428
Star Wars: Episode IV - A New Hope (1977)              2991
Star Wars: Episode V - The Empire Strikes Back (1980)  2990
Star Wars: Episode VI - Return of the Jedi (1983)      2883
Jurassic Park (1993)                                  2672
Saving Private Ryan (1998)                            2653
Terminator 2: Judgment Day (1991)                     2649
Matrix, The (1999)                                    2590
Back to the Future (1985)                             2583
Silence of the Lambs, The (1991)                      2578
Name: Rating, dtype: int64
```

```
Born American (1986)          5
Delta of Venus (1994)         5
Stag (1997)                   5
Cobra (1925)                  5
Enfer, L' (1994)              5
Boys Life 2 (1997)            5
Dancemaker (1998)             5
Feast of July (1995)          5
Frogs for Snakes (1998)       5
In God's Hands (1998)         5
Name: Rating, dtype: int64
```

During the evaluation phase, the output of the Mean Absolute error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) was printed to the screen to show the accuracy of the logistic regression. The output can be seen below:

```
Model 1 (Linear Regression) Evaluation:
    Mean Absolute Error (MAE): 0.9394927445651309
    Mean Square Error (MSE): 1.2776115680903959
    Root Mean Square Error (RMSE): 1.1303148092856237
```

The kNN model was evaluated using a random movie and noting the distance between the selected movie and the generated movies from the kNN model. An example of the output is shown below:

```
Model 2 (kNN) Evaluation:
    Related movies for: Indian Summer (a.k.a. Alive & Kicking) (1996)

    1: With Honors (1994) (0.768242739612717)
    2: Reality Bites (1994) (0.7738022530889954)
    3: Nothing But Trouble (1991) (0.7753432366587092)
    4: Nine Months (1995) (0.7802646799424787)
    5: Father of the Bride Part II (1995) (0.7834050473328725)
```

4. References

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