Movie Recommendation System

Project Report - BAN620 Data mining

A picture containing text, slot machine

Description automatically generated

Under the guidance – BALARAMAN RAJAN



**Table of Contents**

[***Introduction***](#_heading=h.lnxbz9) ***4***

[**Overview**](#_heading=h.35nkun2) **4**

[**Motivation**](#_heading=h.3j2qqm3) **4**

[**Problem Statement**](#_heading=h.1y810tw) **4**

[***Data Mining Phases***](#_heading=h.4i7ojhp) ***4***

[***Data***](#_heading=h.2xcytpi) ***5***

[**Data Source**](#_heading=h.1ci93xb) **5**

[**The dataset is taken from a public platform. The link to the site is given below:**](#_heading=h.3whwml4) **5**

[**Dataset Description**](#_heading=h.2bn6wsx) **5**

[**Dataset Analysis**](#_heading=h.qsh70q) **5**

[***Model***](#_heading=h.3as4poj) ***8***

[**Apriori Algorithm**](#_heading=h.1pxezwc) **9**

[**Collaborative Filtering**](#_heading=h.49x2ik5) **11**

[***Conclusion and Recommendation***](#_heading=h.2p2csry) ***13***

[***Bibliography***](#_heading=h.147n2zr) ***14***

**Table of Figures**

[Figure 1 Data Mining Phases 4](#_heading=h.tyjcwt)

[Figure 2 Average rating per Movie 6](#_heading=h.17dp8vu)

[Figure 3 Customer Rating Values 7](#_heading=h.3rdcrjn)

# **Introduction**

## **Overview**

Netflix initially started as a DVD rental service in 1998. It mostly relied on a third-party postal service to deliver its DVDs to the users. This resulted in heavy losses which they soon mitigated with the introduction of their online streaming service. In order to make this happen, Netflix invested in a lot of algorithms to provide a flawless movie experience to its users. One of such algorithms is the recommendation system that is used by Netflix to provide suggestions to the users. A recommendation system understands the needs of the users and provides suggestions of the various cinematographic products.

## **Motivation**

Movies are an escape route. They distract us from our mundane problems by diverting our attention to the visuals and in the process often releasing emotions or providing relaxation. There by entertaining us. Movies directly connect you to the characters you are seeing on screen and slowly drawing you into their world. After watching a movie whether good or bad your mind is influenced to an extent. Hence people usually try to look at the ratings and then watch the movie.

Looking for the reviews and selecting movies is little time consuming. Hence after brainstorming sessions we came up with an idea to recommend movies that are similar to ones they already like.

## **Problem Statement**

The problem here is which movies should be recommended to the new or existing user on their likings. Our project analysis will recommend those movies to the customer, that they would be interested in, based on the prior customer experience.

**The project goal is to predict and develop personalized recommendation algorithms for better user experience**

# **Data Mining Phases**

This project is carried out on the following phases as shown in the figure below

A picture containing diagram

Description automatically generated

*Figure 1 Data Mining Phases*

# **Data**

## **Data Source**

## The dataset is taken from a public platform. The link to the site is given below:

<https://www.kaggle.com/netflix-inc/netflix-prize-data?select=README>

## **Dataset Description**

The dataset comprises the list of customers who have given the ratings for the movies based on their likings. This dataset was collected between October 1998 and December 2005 and reflects the distribution of all the ratings received during this period.

This dataset involves four text files having the details about the customer and one csv file having details about the movie.

The dataset has of the following variables:

* MovieIDs ranging from 1 to 17770 sequentially.
* The Movie Titles
* Movie ratings from 480 thousand randomly chosen customers. Ratings are on a five-star (integral) scale from 1 to 5.
* Year of release for each Movie ID
* CustomerIDs ranging from 1 to 2649429, with gaps. There are 480189 users.
* Date of each Rating having the format YYYY-MM-DD.

## **Dataset Analysis**

After preprocessing the data,, we selected the first 1000000 records for our analysis. In that we had movie id from 1 to 225 . After selecting the rows , we summarized that data and observed the analysis below.

**Average rating per Movie**

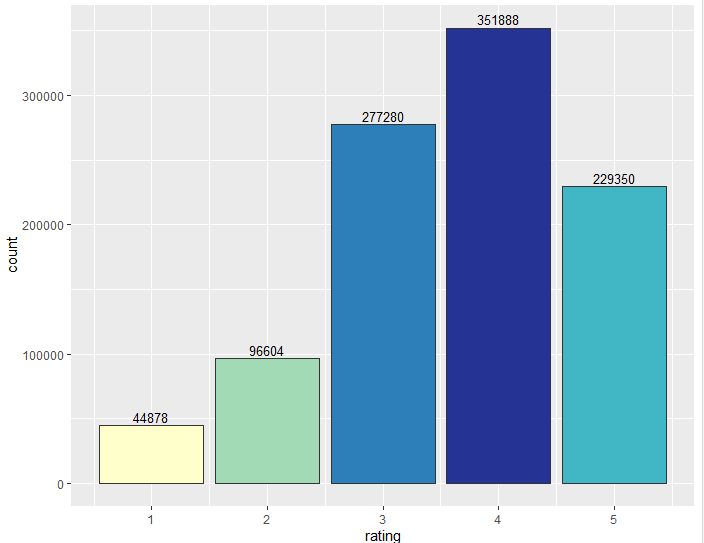
Chart, bar chart

Description automatically generated

*Figure 2: Average Rating Plot of Movies*

From the above plot we can conclude that out of 225 movies , 134 movies got the average rating of 4 while 74 number of movies have an average rating of 3. Likewise, only 13 movies had an average rating of 5.

**Customer Rating Values**



*Figure 3 Customer Rating Values*

This plot shows that there are 351888 number of customers who have given 4 ratings to the movies they watched while 229350 number of customers gave ratings of 5.

Likewise, 277280 number of customers have given rating 3.

From the above two analysis, it is observed that most of the movies classify having rating as 4.

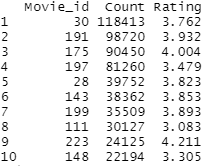
This can be interpreted as

* Either the customer who is giving the rating likes all the movies that he/ she has seen.
* Or the customers who like the movies only give the ratings to those movies and ignore the ones that they don’t like.

After plotting we tried to find Top 10 movies based on popularity and average rating

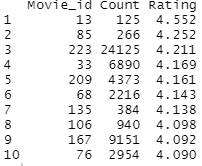
**Top 10 movies based on popularity**

The below chart shows the top 10 movies which are watched many numbers of times



**Top 10 movies based on the highest average rating**

The chart below shows 10 movies having the highest ratings



From the above observation we found out that movies that are highly rated are not the movies which are being watched most. There can be a reason that movies which are being watched most are not the top choice of customers and that's why their average there is not above 4.

# **Model**

After performing basic analysis , we tried to find some seasonality in the data. As in the data there was a column based on date of rating , we thought that there can be some movies which are viewed in particular month eg. holiday movie. For that we grouped the movies based on months in which they were given ratings.

The plot below shows in a month which movie was seen maximum time.

* + The x axis represents the Month
  + The y axis represents the No of views for a movie.

Chart, line chart

Description automatically generated

*Figure 5: Monthly Movie Frequency*

After performing the analysis we found that of the 225 movies considered only few movies showed some seasonality. For example, referring to tag A, we can see that particular movie is watched more during summer likewise tag B is watched more during the start of the year and during the end. Also, tag C is watched more during spring while tag D are evergreen movies which have been watched through the year. From this, we can conclude that very few movies are seasonal.

Further to classify movies based on months we tried to perform the KNN algorithm and classification tree but were not able to conclude anything as there is no seasonality seen in the movies.

## **Apriori Algorithm**

In this project, we applied a data mining algorithm, Apriori, to determine a relationship among films and build a movie recommendation engine. Apriori is a technique in Market Basket Analysis used to discover items that are frequently sold together. Frequently purchased itemset suggests marketing opportunity when customers displayed interest in the subset items. In this case, movies can be viewed as a set of items.

We found that the mining technique can be utilized to uncover an underlying connection within the movies. It can also be used in a movie recommendation, but a number of suggested films can be quite limited, and the quality of such suggestions can be varying.We used variables mainly ratings, movieID and CustomerID to build the rules based on rating received by a movie.

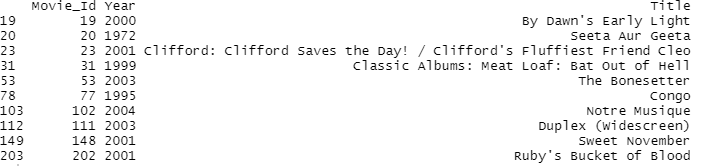
Shown below is an output, after performing Apriori algorithm:

Here 14 rules were constructed. We also got summary statistics of all the rules.In this we can see that rule 5 gives association between movie id 197 and 191 but we can see support is 0.12 which is insignificant but the confidence is 0.44 which is good and its lift is also above 1. So we can say that if a customer has seen movie 197 then there is a chance that he/she will also watch movie 191. As recommending movies is not adding extra risk, so we can put this rule into consideration.

Text, table

Description automatically generated

We also tried Item based collaborative filtering in order to see if we can get the same group of movies as association rules because both the algorithms try to group items based on freq/ transaction . Shown below is the result.



From this we can conclude that association rules work on the popularity of movies and are giving rules based on movies which are frequently watched. We are getting similar results when we find out the top 10 movies based on popularity,on the other hand, Item based is giving different results. So for our use case, we say that association rules are not very good as they are purely based on frequency of viewership and there is not much association between movies.

## **Collaborative Filtering**

Collaborative filtering is a family of algorithms where there are multiple ways to find similar users or items and multiple ways to calculate rating based on ratings of similar users. It can be either User based Collaborative filtering or Item based collaborative filtering. The difference between the two is IBCF tells about “Customers who watched this movie also watched” - If I watched “Home Alone 1”, I might watch “Home Alone 2” but not necessarily “The Wolf of Wallstreet”. And UBCF tells about “Customers similar to you watched” - I may be more inclined to listen to a friend who has similar taste rather than random reviews.

In our dataset, we tried to build a model, based on correlation. In this we have selected a user with customer Id =”16272”. For this we found the users who have watched a similar movie to our user and tried to find the correlation between our current user and all other users. Then based on the highest correlated users and their movie history, we selected the movies which are not seen by our user. After that we tried to find out the mean rating of these movies. Based on the highest average rating , we selected the top 10 movies for our user.

Similarly, we tried to find the top 10 movies for our user using the recommender function of R with user based filtering using pearson correlation. Below are the output of our models.

**Model 1**: Shown below are the movies that are recommended by to the user by using the method described above

Graphical user interface, text, application, email

Description automatically generated

**Method 2 : Using Recommender Package of R With User based Pearson Method**

The following movies were recommended to the same user.

Graphical user interface, text

Description automatically generated

Comparing two outputs we can see the top 10 movies recommended by using the correlation and by Pearson method. We can see that recommendations are almost the same with our basic correlation algorithm. The 10 recommendations are the same except 2 movies.

The advantages of Collaborative Filtering is that it can handle huge datasets. Based on the type of our data, this system can suggest movies watched by similar users, solely depending upon their ratings without understanding the item itself.

**Performance Evaluation of collaborative filtering**

Chart, bar chart

Description automatically generated

We have evaluated the model with the first 50000 records. The distribution of RMSE, MSE and MAE values after conducting cross-fold evaluation are shown in the graph above. From the above graph we can say that user based and item based collaborative filtering is working better than random and popular models in terms of RMSE and MSE but when we look at the MAE model based on popularity it performs best considering our data. The RMSE obtained by our UBCF model is 0.42, considerably lower than other models.

# **Conclusion and Recommendation**

From our analysis we found out that ,for recommendation systems KNN, classification models and apriori recommendation are not the right choice to build models. Although, we can utilize Apriori algorithm findings to construct a new marketing campaign, research customer’s behavior, or make a product suggestion.

User based collaborative filtering is good to give customised recommendation , however if the number of users are too high or the user is new than in that case Item based collaborative filtering is good to give generalized suggestion. For the future , there is scope to develop better models with more features and cloud clusters to have bandwidth.

# **Bibliography**

<https://www.kaggle.com/netflix-inc/netflix-prize-data?select=README>

<https://www.quora.com/Why-do-we-watch-movies>

<https://digitaltransformationpro.com/data-mining-steps/>

<https://towardsdatascience.com/>

<https://www.hackerearth.com/blog/developers/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/>

<https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm/>

<https://www.r-bloggers.com/2020/04/movie-recommendation-with-recommenderlab/>

https://rstudio-pubs-static.s3.amazonaws.com/203264\_7e572fa75b40422d90bd2fd5f3825798.html