Analysis of Movie Recommendation System Data Sets using machine learning techniques

Sohaib Afzal Baig 1* and Hijab Sehar 2++

Received: 01 August 2021; Accepted: 7 Sep 2021; Published: 29 Sep 2021
sohaimirza71@gmail.com  fmscs20012@itu.edu.pk
Information technology university Lahore

Abstract—Multimedia has emerged as one of the top entertainment source due to cheap and uninterrupted availability of high internet speeds. Movie recommendation system have attracted much research interest within the field of recommendation systems. Two widely used techniques, one is collaborative filtering (CF) and second is content-based (CB). However, the accuracy performance of any hybrid system which utilizes more advantage of both systems to better results. Movie recommendation systems has suffered from different problems, such as “Sparsity, Grey sheep problem, Cold start problem, Long-tail problem” etc. Basic Issues can be solved if we take the right choice on what kind of movies to ignore, what movies to suggest. The suggestions generated using approaches such as Linear Regression, Decision Trees, and Bayesian Analysis are examined in this study. Movie-Lens-1M and Movie-Lens-10M are the dataset considered. The results of this experiment suggest that Decision Tree and Linear Regression & Random Forest work well as compared to Bayesian Learning.

Index Terms—Movie Recommendation Systems, Linear Re-aggression, Bayesian Learning, Decision Tree, Random Forest

I. INTRODUCTION
The purpose of recommendation systems is to assist consumers in filtering information according to their preferences. Users' evaluations toward things may be predicted by such algorithms, which may subsequently be used to provide a list of item suggestions. Using methods such as the user's prior behavior, profile, or demographic data, movie recommendation systems give the user with movie options that are more likely to be seen by him. It's tough for a user to locate or choose a movie to watch among hundreds of options, and Movie Recommendation Systems help by suggesting movies. Well-known online multimedia-based sites, such as youtube.com, netflix.com, and movielens.org, have employed movie suggestions.

The algorithm for making suggestions is described as follows:

I.1. Collaborative Filtering technique is the most extensively used method for generating a list of suggestions by learning users' or objects' rating patterns. The similarity in rating patterns between the two groups in this example is significant. The rating prediction is heavily influenced by users or things. For the purpose of producing suggestions, the following findings were obtained. This strategy is effective is based on the pattern of movie ratings provided by consumers. To create, this strategy involves two basic parts. The following are some movie suggestions: (1) a film with a similar plot; (2) a film with.

![Flow Chart of Collaborative Filtering](image)

I.2. Content based by leveraging the similarity of items' content, the content-based method creates a list of suggestions. This strategy, on the other hand, is only useful when, There are objects that exist and will function poorly when used deployed exclusively. In order to put in place a Filtering mechanism based on content. The following steps are shown to implement a content-based filtering system in Fig. 2.
1.3. **Hybrid** method is one that generates suggestions by combining the advantages of the CF and CB methods mentioned above. This method can be implemented using the results of a content-based approach to collaborative filtering (and vice versa) or by linearly aggregating the results of both methods. To create movie suggestions, [1] proposed hybrid technique comprises of four key phases: (1) text Data preparation, (2) term weighting, (3) Collaborative Filtering based approach, and (4) movie clustering. The flowchart for this method is shown in Figure 3.

The following is an overview of the paper's framework: The Literature Review is discussed in Section II, and the Proposed Model and Algorithm is discussed in Section III. The results and conclusion are presented in sections IV and V, respectively.

### 2. Related Work

Over the years, several different types of recommendation systems have been developed. Different ways to constructing recommendation systems are used by different sorts of systems. Content-based approaches, collaborative-based approaches, demographic-based approaches, knowledge-based approaches, and hybrid approaches are only a few examples. Collaborative filtering algorithms are frequently utilized in commercial recommendation systems. Use-based techniques and item-based approaches together are the most studied forms of memory-based collaborative filtering. However, each sort of recommendation system has its own set of advantages and disadvantages.

#### 2.1. Collaborative Filtering based Recommendation Systems

Y. Hu, Y. Yang, C. Li, Y. Wang Wang (2016) [1] to propose the collaborative filtering approach was used by scientific publications. Traditional collaborative filtering and probabilistic topic modeling are combined in this technique. It may build suggestions regarding both current and freshly published Movies and offers an explicable unique system for the customers.

Bei-Bei CUI, 2017 [2] proposed a Movie Recommendation System Based on KNN Collaborative Filtering Algorithm. A recommender system is recommended the design and implements a movie recommendation system prototype combined with the actual needs of movie recommendation through researching of KNN algorithm and collaborative...
filtering algorithm with web-scope data set. Then it give a detailed principle and architecture of JAVAEE system relational database model. Finally, the test results showed that the system has a good recommendation effect. It give a detailed design and development process, and test the stability and high efficiency of experiment system through professional test. This paper has reference significance for the development of personalized recommendation technology.

Nirav Raval, Vijayshri Khedkar (2019) [6] proposed Python library in collaborative method, a movie recommendation system based divided into model-based and memory-based Model. This Methods take action only on a user-item rating matrix and model-based system, like a neural network, generates a model that learns from the information of user-item ratings and recommends new Movies items Different approaches like Item-based filtering, User-based filtering alternating least KNN method, square methods and use movie-lens data set for performance measurement of this system Mean Square method (MSE), Root mean square method (RMSE).

2.2. Content based Recommendation Systems

For item recommendations, the collaborative filtering process is particularly successful. However, because new users have no usage history, it is tough to propose new goods to them or to suggested items for users. The content-based recommendation strategy is appropriate for proposing a new item, thus no can suggest a new things for customers. The content-based recommendation method is suitable for recommending a new item, but it is not effective for recommending an item to a new user.

Elkahky Ali Mamdouh 2015 [3] solved this problem by using web browsing and searching history. They learned preferable topic of user by analyzing search query and clicked URL, and it is used to recommend item to user.

The content-based recommendation strategy is appropriate for proposing a new item, thus no can suggest a new things for customers.

2.3. Deep Learning based Recommendation Systems

Covington, Paul, Jay Adams (2016) [4] to present a deep learning-based technique of selecting films that is appropriate for each user. To produce candidates and rank them, they employed two deep learning networks. Candidates for video are created in the first network from 10 ml of clips. The other one is deep-learning mechanism is then used to score the clips in order to propose the best ones to each viewer.

Nagamanjula R, A.Pethalakshmi (2020), [7] proposed a Movie Recommendation System using SVM classifier. A movie recommendation system based on the combination of opinion mining and user similarity analysis. For this purpose use the SVD KNN algorithm with movie-lens data set. It helps to recommend top-k movies for target user. It is the collect reviews of users for movies and pre-processes data with certain major preprocessing steps. The pre-processed are given for explicit and implicit aspect extraction. The aspect of a word is further classified according to the classes. Finally, top-k movies are recommended for target user. The results suggested by our proposed system are leading and block buster movies and the system is helpful for millions and billions of users around the globe. Here the accuracy of classification is improved using NbSVM classifier and also meet the requirement of the users.

2.4. Hybrid based Recommendation Systems

Noor Ifada (2020) [8] to present a hybrid technique that combines the advantages of the Content-based (CB) approach with the benefits of Collaborative Filtering (CF). As a result, this method needs more movie data and procedures than the Collaborative Filtering method. To produce movie suggestions, the proposed technique consists of four basic stages: The text pre-processing, video clustering, term weighting, and a Collaborative Filtering-based technique are all used.

S. Agarwal (2017) [9] to suggest a Hybrid method to increase the accuracy, quality, and scalability of movie recommendation systems that combines content-based filtering with collaborative filtering, employing a Support Vector Machine as a classifier and gene expression data.

Mahesh Goyani and Neha Chaurasiya (2020) [10] proposed method is hybrid filtering. A recommendation system is to build an efficient recommender system a hybrid combination of different methods of recommendation is must using Nearest N User Neighborhood algorithm and Yahoo Research Web Scope Database and Movielens-10M dataset. It is concluded that by using combination of similarity measure a better user similarity can be generated rather than using single similarity measure and efficiency of the system is also increased. One of the facts that similarity measure like RMSE is evolved by the author and up till now it is only used in movie recommendation. The author also showed that this similarity measure is better than the other in terms of efficiency parameters. Accuracy of this recommender system can be achieved by the hybrid filtering.

It will give progressively explicit outcomes contrasted with different systems that are based on the content-based approach. These systems work on individual users’ ratings, hence limiting your choice to explore more. While our system, which is based on a collaborative approach, computes the connection between different clients and relying upon their ratings, prescribes movies to others who have similar tastes, subsequently allowing users to explore more. It is a web application that allows users to rate movies as well as recommends them appropriate movies based on other’s ratings.

Phonexay Vilakone1, Doo-Soon Park, Khamphaphone Xinchang1 and Fei Hao(2018) [12], proposed improved k-cliques method, in movie recommendation system the output of this method is very effective. A recommendation system use the tow methods which one is Maximal clique method and second one is Collaborative filtering using K nearest neighbor algorithm and movie-lens data set. For performance evaluation, it evaluated the collaborative filtering method using a k nearest neighbor, maximal clique method, k-clique method and improved k-clique methods. The results showed that the improved k-clique method improved the precision of the movie recommendation system more than the other methods used in this. The improved k-clique methods accuracy was good.

3. PROPOSED WORK

A. Datasets

It used two well-known separate data sets from Movie Lens, which are developed via group lens research team to do research in the field of recommender systems, to aid in the evaluation of recommendation systems by developers these are the following: 1M dataset for Movie Lens 2. 10M Movie Lens dataset. The following characteristics are included in the Movie Lens databases [15]

The Movie Lens dataset has the following characteristics:

1. Ratings from 1 to 5 are assigned to user.
   (1 means very bad, 5 means very good)
2. Each user rated at least 20 movies.
3. (Age, gender, occupation, zip)
   Provides simple demographic information for users.

The reason for considering two different data sets of different size Movie lenses is to check the scalability parameter, i.e. when our hybrid approach is applied to two different data sets of different size Movie lenses the system should work well at the same time and provide good performance, if the system.

<table>
<thead>
<tr>
<th>Attributes</th>
<th>MovieLens-1M</th>
<th>MovieLens-10M</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of Genres</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Average no of genres</td>
<td>1.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Number of Users</td>
<td>6,040</td>
<td>71,567</td>
</tr>
<tr>
<td>Number of Ratings</td>
<td>1,000,209</td>
<td>10,000,000</td>
</tr>
<tr>
<td>Number of Movies</td>
<td>3,952</td>
<td>10,681</td>
</tr>
<tr>
<td>Rating Scales</td>
<td>1-5</td>
<td>1-5</td>
</tr>
</tbody>
</table>

TABLE I

Dataset Details

B. Recommendation system quality measures

The effectiveness of the recommender system is measured in numerous ways, including the prediction accuracy and system performance. Because the recommender systems have many different types, there is no one statistic that can assess all aspects of a business system of recommendations. Furthermore, many metrics might be used merged for overall assessment. We will concentrate more on this topic in this part on precision, which is widely regarded as the most significant factor criteria. Statistical accuracy measurements assess how close various types of algorithms’ projected ratings are to real user ratings. One of the unique often used predictive best accuracy metrics in recommender system assessment is Mean Absolute Difference. This is the mean absolute error between the predicted rating and the actual rating provided by the user. The benefits of MAE are self-evident. To begin with, it is straightforward and easily comprehended, making it simple to apply and compelling. Furthermore, because many researchers use it to assess prediction accuracy, comparisons between the different methodologies are simple to make. To be more exact, the formula below demonstrates how to compute MAE.

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{r}_i)
\]

The MAE is a measure of how accurate a forecast is. The lower the MAE, the more accurate the prediction is. Other statistical accuracy metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Correlation. The RMSE Formula is provided below.

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{r}_i)^2
\]
$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (r_i - \hat{r}_i)^2}$$

The decision support accuracy metric focuses on whether the recommendations are useful, that is, whether the user is interested in the system recommendations. Therefore, this type of metric makes it easy to predict binary results, good or bad. Accuracy is a metric that quantifies the number of correct positive predictions made. Precision then calculates the precision of the minority class. It is calculated as the percentage of predicted positives divided correctly by the total number of predicted positives.

$$Precision = \frac{\text{True Positive}}{2 \times \text{True Positive} + \text{False Positive}}$$

Recall is measured in a two-class imbalanced classification problem as the number of true positives divided by the total number of true positives and false negatives. The outcome is a number ranging from 0.0 (no recollection) to 1.0 (perfect recall).

$$Precision = \frac{\text{True Positive}}{2 \times \text{True Positive} + \text{False Positive}}$$

C. Pre-processing Data

In most cases, real-world data is incomplete, noisy, and inconsistent. When it comes to movie ratings, such as MovieLens, respondents may purposefully give inaccurate ratings or data input errors may occur. Good data is required for the best classification results. We did this by preprocessing the rating data, which included data cleansing, data transformation, user selection, and movie selection, among other things. First, it used the Pandas Library to extract data from files. Then, for each movie, we determined the average rating of all the ratings submitted by various users. After that, we constructed a matrix.

Each movie has a features vector in the input data of size M×U (where M is the number of movies and U is the number of genres). Explanation of Algorithm 1

D. Decision Tree

The suggested TRS uses a decision tree as a classifier/model because it offers various advantages to the decision maker, including simplicity and interpretability. Making decisions is a difficult task. Due to its flowchart-like form, it is simple to comprehend. For technological reasons, it is responsible for the TRS’s technical concerns in terms of scalability and sparsity. There are nodes in the decision tree as well as leaves. The root node is the initial node in the chain. The test set’s instances begin to descend to a leaf. Other nodes, often known as internal nodes, are used to test a system. This is where the split — either binary or not — takes place multi-occurrences. The leaf nodes each indicate a class label (i.e., the name of the class) the result of nodes.

E. Linear Regression

We employ Linear Regression to model the relationship between a movie’s feature and user ratings. In each case, we had a feature vector of size 18 in our movie, hence we needed a size 18 parameter vector

$$r_i = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_{16} x_{16}$$

The gradient descent technique is a prominent algorithm for computationally solving linear equations in mathematical optimization theory issues with programming it is incredibly efficient in practice in the calculation of regression functions.

F. Bayesian Learning

In many classification fields, the Simple Bayesian Classifier is one of the most effective machine learning methods. Despite its simplicity, it has been demonstrated to compete with more complicated algorithms in text classification challenges.

Rating was separated into five categories [1, 2, 3, 4, and 5]. The probability of an item belonging to class j rating given its n. The “naive” feature values supposition that features are independent given the rating class.

$$p(r_j | f_1, f_2, \cdots, f_{16}) = p(r_j) \prod_{j=1}^{N} p(f_j | r_i)$$

$p(r_j)$ And $p(f_j | r_i)$ may both be calculated using training data. The likelihood of each rating class is determined, and the example is allocated to the rating class with the highest probability.

Although the assumption that features are independent once we know a movie’s rating class is unrealistic in this domain, the Simple Bayesian Classifier has been proven to be optimum in many circumstances when this assumption is false, and to be competitive with more sophisticated techniques in others. The Simple Bayesian Classifier is also quick since its learning time is proportional to the number of instances in the dataset.
Algorithm-I: Data Preparation

Input: Rating data, Movie Genres data

Process:

1. total movies = n
2. total types of genres = m
3. list of genres = G = [Action, Adventure, Fiction, ...]
4. Output data=[ ] Input data=[ ]

for each movie i:

1. Get the list of ratings of movie i
2. Get the list G movie contain genres of movie i
   (for example movie i is an Action and Fiction Movie)
   \[ G_{\text{movie}} = \{\text{Action, Fiction}\} \]
3. Calculate the mean rating of movie i
   \[ \text{mean_rating} = \frac{\text{sum of all ratings}}{\text{total no of ratings}} \]
4. Append mean rating in Output data
   Output data(i) = mean_rating
5. Multi-Label Binary Encoding of list G with respect to G movie
   encoded_list = [1, 0, 1, ...]
6. Append feature vector of movie i to Input data
   Input data(i,:) = encoded_list

Output: Input data, Output data

4. RESULTS

A set of experiments are given in this section to show these three distinct prediction methodologies.
First, we look at how the data split ratio affects each strategy.
After that, compare them at each stage. Then we’ll look into it
the proclivity of each strategy, and come to the conclusion that

The decision is based on a linear regression model Bayesian Learning & Random Forest is more accurate. The results of the Decision Tree testing are shown in Fig. 5.
The testing results are shown in Fig. 6 for the MovieLens-1m dataset on the MovieLens-10m dataset for Decision Tree
When it comes to presenting

As a result of our findings, we first look into the impact of partitioning on data in both models, then compare accuracy
and tensile strength.

<table>
<thead>
<tr>
<th>Data Split</th>
<th>MAE</th>
<th>MSE</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.504862</td>
<td>0.431601</td>
<td>0.492806</td>
<td>0.492806</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.503378</td>
<td>0.421892</td>
<td>0.493994</td>
<td>0.493994</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.501033</td>
<td>0.429115</td>
<td>0.490556</td>
<td>0.490556</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.513965</td>
<td>0.442344</td>
<td>0.483363</td>
<td>0.483363</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.513177</td>
<td>0.435873</td>
<td>0.484778</td>
<td>0.484778</td>
</tr>
</tbody>
</table>

Figure 5: Decision Tree Results on MovieLens-1m dataset

<table>
<thead>
<tr>
<th>Data Split</th>
<th>MAE</th>
<th>MSE</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.433977</td>
<td>0.291272</td>
<td>0.520000</td>
<td>0.520000</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.444654</td>
<td>0.320811</td>
<td>0.550000</td>
<td>0.550000</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.430691</td>
<td>0.316990</td>
<td>0.542000</td>
<td>0.542000</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.434226</td>
<td>0.317694</td>
<td>0.550000</td>
<td>0.550000</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.466770</td>
<td>0.372625</td>
<td>0.508571</td>
<td>0.508571</td>
</tr>
</tbody>
</table>

Figure 6: Decision Tree Results on MovieLens-10m dataset

As we can see in Fig.7 and Fig. 8, as we increase the split ratio, there is slight decrease in model accuracy, MAE and other evaluation parameters. By comparing the results of Decision Tree model and Linear Regression model, we can conclude that Linear Regression perform well.
Fig. 9 and Fig. 10 display the comparisons between the Bayesian Learning approach on Movielens-1m and Movie lens-10 datasets, correspondingly in terms of MAE, MSE, Precision, Accuracy and Recall. We can notice that the Bayesian Learning based approach always perform poor as compared to other methods.

Fig. 11 and Fig. 12 display the comparisons between the Random Forest approach on Movielens-1m and Movie lens-10 datasets, correspondingly in terms of MAE, MSE, Precision, Accuracy and Recall. We can notice that the Random Forest based approach always perform better as compared to other
methods.

<table>
<thead>
<tr>
<th>Data Split</th>
<th>MAE</th>
<th>MSE</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.504882</td>
<td>0.421601</td>
<td>0.614802</td>
<td>0.614802</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.502378</td>
<td>0.431292</td>
<td>0.613378</td>
<td>0.613378</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.509103</td>
<td>0.423112</td>
<td>0.619122</td>
<td>0.619122</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.508865</td>
<td>0.445805</td>
<td>0.602789</td>
<td>0.602789</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.504177</td>
<td>0.454177</td>
<td>0.609427</td>
<td>0.609427</td>
</tr>
</tbody>
</table>

Figure 14: Random Forest on MovieLens-1m dataset

<table>
<thead>
<tr>
<th>Data Split</th>
<th>MAE</th>
<th>MSE</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.3</td>
<td>0.524023</td>
<td>0.441322</td>
<td>0.644802</td>
<td>0.644802</td>
</tr>
<tr>
<td>1</td>
<td>0.4</td>
<td>0.524122</td>
<td>0.441526</td>
<td>0.643378</td>
<td>0.643378</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.514131</td>
<td>0.432211</td>
<td>0.649122</td>
<td>0.649122</td>
</tr>
<tr>
<td>3</td>
<td>0.6</td>
<td>0.513332</td>
<td>0.453122</td>
<td>0.632789</td>
<td>0.632789</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>0.510091</td>
<td>0.461442</td>
<td>0.639427</td>
<td>0.639427</td>
</tr>
</tbody>
</table>

Figure 15: Random Forest on MovieLens-10m dataset

Figure 16: Random Forest chart

5. CONCLUSION

For both service providers and customers, recommender systems are a valuable tool. This type of site is ideal for an e-commerce business. By promoting technologies, it can help businesses improve revenue things that are most likely to be purchased by consumers. It can also be beneficial to users. Assist them in sorting through the avalanche of information and locating what they’re looking for truly desire in a more refined manner. It put three movie recommendation systems to the test in this research. Decision-making systems based on three Machine Learning algorithms Bayesian Learning, Linear Regression, Random Forest and Tree All Four of them approaches uses rating data and movie genres data to learn. There is a correlation between movie genres and ratings. The performances are evaluated in terms of Precision, Accuracy and Recall metrics.

6. FUTURE WORK

In the future, we could continue this work with the other Machine Learning models like Pre-Trained Model, Transfer Learning, User privacy, and security maintenance using block chain & especially explore the search space by growing the multi-filtering dimensions of the Recommendation System in Machine learning.

REFERENCES


