**Netflix Recommendation System**

**Team Marshmallow**

|  |  |
| --- | --- |
| **Team member Name** | **SRN** |
| Aditya VK | PES2UG19CS017 |
| Aryan Kumar | PES2UG19CS067 |
| Shreesh RD | PES2UG19CS382 |
| Shreesha SK | PES2UG19CS383 |

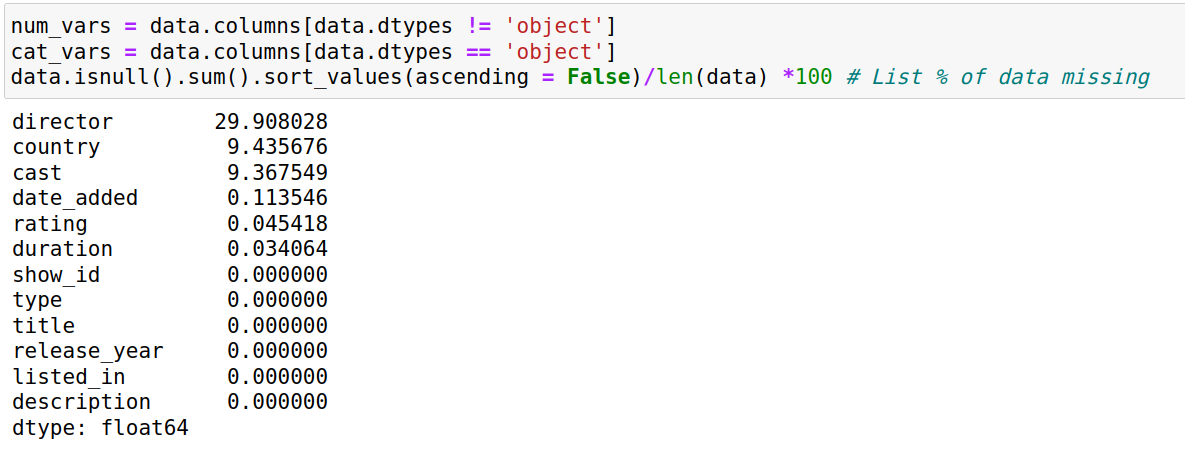
1. **Dataset Name: Netflix Dataset**

Dataset Description: The dataset consists of 12 attributes or columns and 8807 values

1. **Problem statement**: Building a recommendation system for this dataset.
2. **EDA and Visualization**
   * **How many rows and attributes?**

8807 rows and 12 attributes

* + **How many missing data and outliers?**

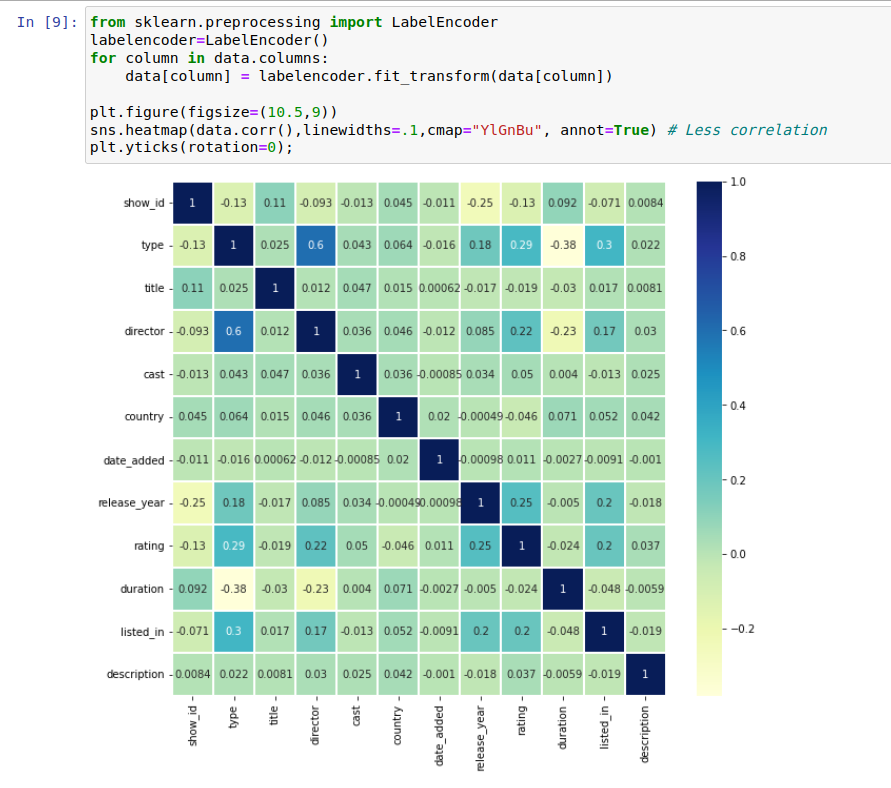


* + **Any inconsistent, incomplete, duplicate or incorrect data?**

No duplicates, incomplete data such as NaN values exist and are considered above

* + **Are the variables correlated to each other?**

Low correlation

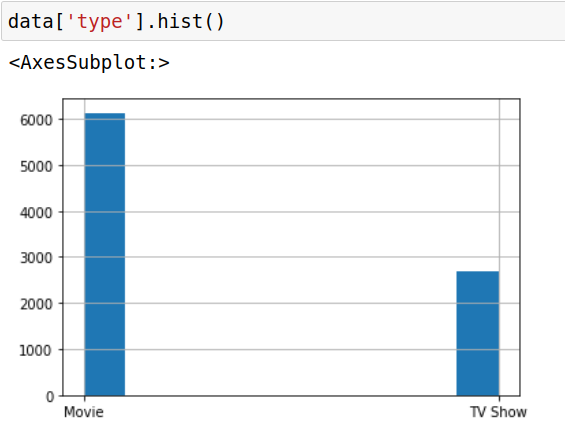


* + **Are any of the preprocessing techniques needed: dimensionality reduction, range transformation, standardization, etc.?**

Alot of the data in the dataset is in text, so the dataset needs preprocessing. The cast and description columns need standardization.

* + **Does PCA help visualize the data? Do we get any insights from histograms/ bar charts/ line plots, etc.?**

Yes some insights can be drawn from the graphs. For example we see that the current dataset has alot more movies than TV shows.



1. **Link for google sheet**

https://docs.google.com/spreadsheets/d/1moIlEb17WK02K6jH-JHkpPS48HXQ0wk-l6OAkZUOQCA/edit#gid=0

1. **Literature Survey ( Summarize)**

**Aditya VK**

**1) Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques**

Research on recommender systems generally focus on designing techniques to improve the accuracy of recommendations. However these tend perform poorly in terms of recommendation diversity. In this paper the focus is on developing algorithmic techniques for improving aggregate diversity of recommendations.

A recommender system with a higher aggregate diversity is more useful to a user as they provide a wider range of items instead of just the bestsellers (which users are capable of discovering by themselves). These systems can also be useful to some business models as they would increase sales of relatively obscure items located in the tail of sales distribution.

Typically the items with highest predicted (as opposed to lowest predicted) ratings are the ones recommended to users. Item-popularity-based ranking is the approach proposed by the authors in order to achieve diversity improvement. It ranks items directly based on their popularity, from lowest to highest, where popularity is represented by the number of known ratings that each item has. This ranking approach increased recommendation diversity from 385 to 1,395; however, recommendation accuracy dropped from 89 to 69 percent. Hence a “ranking threshold” was introduced to allow users to choose a certain level of accuracy.

The proposed recommendation ranking approaches were tested with several movie rating data sets, including MovieLens (grouplens.org), Netflix (netflixprize.com), and Yahoo! Movies (individual ratings from movies.yahoo.com). Three widely popular recommendation techniques were used including two heuristic-based (user-based and item-based CF) and one model-based (matrix factorization CF) techniques. The predicted rating threshold was set as 3.5 (out of 5) to ensure that only relevant items are recommended to users, and ranking threshold was varied from 3.5 to 4.9. The number of top-N recommendations provided by the system can also be varied to control the Accuracy-Diversity Tradeoff.

All the proposed ranking approaches improved the diversity of recommendations by sacrificing the accuracy of recommendations. However, with each ranking approach, as ranking threshold increases, the accuracy loss is significantly minimized. With item-based CF technique on the MovieLens dataset, a 0.1% loss in accuracy allowed a 20% increase in diversity. With a 1% accuracy loss the diversity increased by 81%.

The authors have proposed a number of ranking-based techniques to improve aggregate diversity of recommendations while maintaining adequate accuracy. These ranking techniques also offer flexibility to system designers (i.e., they do not require the designer to use only some specific algorithm). They are also based on scalable sorting-based heuristics and are extremely efficient.

**2) Movie Recommendations Using the Deep Learning Approach**

In this paper, the authors propose a deep learning approach based on

autoencoders to produce a collaborative sltering system which predicts movie ratings for a user based on a large database of ratings from other users. Collaborative filtering is an approach for recommendation systems which relies on the ratings for particular user as well as the ratings of similar users. New movies are recommended to users based on accurate predictions of movie ratings.

The authors proposed a simple neural network model which performs well in terms of root mean squared error for collaborative sltering. Using 90% of the full MovieLens dataset as training it took roughly 4 days to make 30 passes over the entire data before the training loss stabilized around 0.42(RMSE). On test data, the deep learning model-based algorithm outperforms other popular recommendation systems like NPMF, MudRecS and CMF.

A total of 100 participants, who were students at the authors’ university, were used in a study to rate 15 randomly chosen recommendations. In this survey, 71.67% of the time appraisers preferred the recommendation made using our deep learning approach over the recommendation made by the baseline approach.

The experimental results show that the proposed recommendation system outperforms a user-based neighborhood baseline both in terms of root mean squared error on predicted ratings and in a survey in which users judge between recommendations from both systems. In addition, the system was able to handily outperform the neighborhood-based baseline, and was able to provide superior movie recommendations.

**3) Movie Recommender System Based on Percentage of View**

In this paper, an Implicit Opinion Measure (IOM) is proposed to improve the performance of movie recommender systems on implicit feedback. The dataset used is created by Namava, a media service provider.

In this paper the authors propose a percentage of view approach to find relevant movies for customers. To predict this metric, collaborative filtering, content-based filtering and a new method which is called residual method was used.

To this end, the correlation between like probability and percentage of view of movies is shown and it was illustrated that there is a positive correlation between them. For the evaluation of the recommender system, the average of precision of the top 20 recommendations for all of the users was reported as performance of the recommender system.

Prediction results for Content-based Filtering, Collaborative Filtering – Model-based and Collaborative Filtering – Memory-based algorithms are reported.

To improve prediction accuracy, the sign and the value of error was estimated independently at each level. For the comparison, a random recommender system is considered as the baseline and the accuracy of the recommender system will be compared with it.

The recommender system developed using these different prediction methods, works 5 times better than the average random recommendation.

**Shreesha SK**

**4) An Approach for Netflix Recommendation System using Singular Value Decomposition**

The results show an overall improvement in performance for all algorithms and mathematical improvement models as the data size grow. Moreover, the individual increase in performance of each algorithm implies a great difference in their effectiveness, especially between the matrix factorization algorithms SVD and ALS.SVD produced the best ratings among the algorithms with an RMSE of 0.8190 under the 1M data set and a performance increase of 0.09936% from the 100K to 1M data set. This was expected as the winners of the prize used a modified version of the SVD algorithm in their solution. The RMSE indicate how well these algorithms predict ratings for non-rated items. Based on these results, it is suggested to utilize SVD out of the four algorithms for a recommender system in production. This paper focused on comparing four different collaborative filtering algorithms, in which the aim was to find out which one that produced the best prediction rate. The four algorithms were KNN, SVD, ALS and Slope One. This paper also used two mathematical models, Arithmetic mean and weighted arithmetic mean, in order to determine if the mathematical models could produce better prediction ratings than the four algorithms. Out of the four algorithms the SVD had the best prediction rate whereas ALS the worst. However, ALS had the highest performance improvement when the data increased from100K ratings to 1M ratings. AM had a slightly better prediction rate than SVD, and WAM had the overall worst prediction rate.

**5) A New Collaborative Filtering Recommendation Algorithm Based on Dimensionality Reduction and Clustering Techniques**

This paper proposes a new method for recommender systems that benefits from the potentialities provided by the k-means clustering algorithm and SVD technique. Firstly, the k-means clustering algorithm was adopted to cluster users in the same partition according to their preferences, and then the SVD was used in each cluster not only as a dimensionality reduction technique but also as a powerful mechanism, which could efficiently help in finding the most similar users. To evaluate the performance of the proposed method, experimentations were conducted on two real-world datasets for movies recommendation called MovieLens 1M and MovieLens 10M, which contain about 1 million and 10 million ratings made by anonymous users, respectively. In addition, RMSE metric was adopted to evaluate the predictive accuracy of the proposed method in comparison with well-known k-nearest neighbour based recommendation and k-means-based recommendation methods. The experimental results showed that our method improved significantly the performance of the recommendations and remained the lowest values in the RMSE curve in the whole neighbours range.

6) **Personality, Movie Preferences, and Recommendations**

In this work presented results show the relationship between individual’s personalities and their preferences for and use of recommender systems. The first work combines an analysis of personality with actual movie viewing and rating behavior from users. Apersonality test was administered to a pool of 73 subjects and then analyzed their viewing and rating history on Netflix, surveyed them about their attitudes toward recommender systems, and collected ratings of movies that Netflix recommended. Strongest result is a consistent significant correlation between Conscientiousness (a personality trait that reflects careful planners, thoughtfulness, and meticulousness) and a positive attitude about recommender systems, the recommendations themselves, and frequency of use. This has implications for evaluating recommender systems. Users with high Conscientiousness are more inclined to like and use recommender systems, which may be considered when evaluating accuracy with this group. Similarly, lower Conscientiousness users find less usefulness in recommender systems and rate recommended films to be of lower interest. This suggests they are a group where recommendation strategies could be altered to improve the usefulness of the system, and where more attention may be paid to improve recommendation quality. These results highlight personality research and other analysis of user-traits as an interesting and important area of future work for the recommender systems community, with promise for improving algorithms, human decision making, and perception of systems.

**Shreesh RD**

**7) A comparative study of video recommender systems in big data era**

We are present in a big data era and the presence of high-bandwidth access ot the internet has made various kinds of video contents available to everyone. Due to this for video content providers to survive they must be able to provide suitable videos for the user. Therefore recommendation systems are used to suggest video content that the user is most likely to enjoy and hence continue using the service which is beneficial to the video content provider. In this paper popular recommendation systems used by Netflix, Youtube, Hulu and Amazon are analysed and their respective pros and cons are discussed. Each platform has a different requirement for their recommendations systems as each platform has different content. Netflix has a very high recommendation accuracy of 80% and majority of their sales come from the recommendations. Netflix also supports multiple users to use the same account by creating profiles for each user, this helps netflix keep the recommendations separate and the system is not affected by a different user on the same account. The conclusion that can be drawn from the paper after analysing the 4 popular recommendation systems is that there are 2 major steps to the recommendation process. One is to gather data and the second step is to run the algorithm. From the paper it can be deduced that data itself plays more important role than software algorithms in order to improve the recommendation performance. In the big data era, we need an adaptive algorithm which can use the gathered data in real time.

**8) A Comprehensive Survey on Movie Recommendation Systems**

Nowadays users often face the problem of available excessive information. To help combat this recommendation systems are developed and are constantly being refined to enhance the user experience. Recommendation systems are used by digital entertainment platforms like Netflix, prime video, IMDB, and e-commerce portals like Amazon, Flipkart and eBay.

This paper comprehensively reviews several other research papers and discusses the methods used by each of the respective authors. Two traditional methods namely, collaborative filtering (CF) and content-based approaches are used and consist of few limitations individually. Various approaches are shown in the paper such as hard and soft clustering, methods to reduce time complexities, models that review the posters/aesthetic rather than the movie data and also analysing user reviews. From the research, it is found that most movie recommendation systems follow a collaborative filtering approach rather than content based methods. It was also found that hybrid recommendation systems, were preferred by most authors as they tend to utilize the best of both former approaches. However, several major drawbacks remain in recommendation systems despite the use of advanced techniques such as deep learning, neural networks, knowledge discovery over the last few years. There are several fundamental issues faced by movie recommendation systems such as scalability, cold start problem, data sparsity and practical usage feedback and verification based on real implementation are still neglected. Other issues that require significant research attention are accuracy and time complexity problem, which could make RS, a bad candidate for real-world recommendation systems. These latest techniques have resulted in improvements in some areas of recommendation systems but not all, hence there exists an enormous potential for further advancements in the area.

**9)Predictive Self-Learning Content Recommendation System for Multimedia Contents**

The internet is widely used for entertainement and on a very large scale. For instance; one billion hours of YouTube videos are watched every day. One of the key features of such platforms such as the entertainment and shopping platforms is the recommendation system based on past activities of users and the contents of the visited sites to provide related contents to decrease search time and increase the data availability.

The objective of this paper is to create an algorithm that predicts what users search next by using prior collected information and using machine learning to analyze the user behaviors for the future activity. Recommendation Systems face several challenges such as noisy data or bad data because the users don’t want to cooperate. Therefore recommendation systems are general and cannot be changed according to user characteristics after visiting the suggested data. The algorithm predicts what a user searches next by using prior collected information and using machine learning to analyze the user behaviors for the future activity. The Predictive Self-Learnin Recommendation System that uses a Collaborative Filtering Algorithm as well as seven criteria (Popularity, Similarity, Currency, Feedback, Importance, Safety, and Interest) in addition to users’ profiles to make predictive recommendations to users. The system is different from traditional recommendation systems because it allows for more diverse suggestions without decreasing the performance of the system in terms of response time and CPU utilization. Although, the proposed method and related analysis can assists the shopping, entertainment and similar recommendation systems to increase their efficiency by well characterizing users’ behaviors, the more work needs to be done in terms of testing, used dataset, and the learning methodology.

**Aryan K**

**10) A Movie Recommender System: MOVREC**

MOVREC is based on collaborative filtering that makes use of the information provided by users, analyzes them and then recommends the movies that is best suited to the user at that time. Recommended movie list is sorted according to the ratings given to these movies by previous users and it uses K-means algorithm. This system has been developed in PHP using Dreamweaver 6.0 and Apache Server 2.0. Collaborative filtering system recommends items based on similarity measures between users and/or items. The system recommends those items that are preferred by similar kind of users. Content-based filtering is based on the profile of the user’s preference and the item’s description. In CBF to describe items we use keywords apart from user’s profile to indicate user’s preferred liked or dislikes.

In K-means clustering technique we choose K initial centroids, where K is the desired number of clusters. Each point is then assigned to the cluster with nearest mean i.e. the centroid of the cluster. Then we update the centroid of each cluster based on the points that are assigned to the cluster. We repeat the process until there is no change in the cluster center (centroid).

**11) Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications**

Algorithms can help with this task as they help for music, movie, and product recommendations. They have developed an algorithm, and an accompanying Python library, that implements a recommendation system based on the content of articles. They e tested the library on 15K posters from the Society of Neuroscience Conference 2015. They show that the algorithm significantly outperformed suggestions based on keywords. They used Latent semantic analysis which is based on SVD they have also used Rocchio algorithm, The Rocchio algorithm is used to produce recommendations based on relevant and non-relevant documents previously voted by the user [12]. The method can work with term frequency, tf-idf, or any term weighting schemes.They prove that their algorithm works better than other simple recommendation based algorithms and has a much higher accuracy.

**12)A Personalized Electronic Movie Recommendation System Based on Support Vector Machine and Improved Particle Swarm Optimization**

This paper has presented a new rating prediction model for the pre-classification, and later regression, for personalized recommendation. The main advantage of the approach relies on its ability to overcome the limitations of existing collaborative filtering recommendation methods,. In particular, the proposed system starts by establishing a SVM classification model and by identifying a preliminary recommendation list. It then builds a SVM regression model based on the preliminary recommendation list, and predicts items’ ratings. The proposed method is capable of using the items’ content information as well as accounting for the user’s demographic information and behavior information to establish the user-item correlation information matrix and to capture the user’s interests and preferences. To improve the performance of the recommendation system, an improved PSO algorithm with the evolution speed factor and the aggregation degree factor (IPSO) is also proposed to optimize the parameters of the mode.

The experimental results show that the proposed model can provide better recommendation results than the other methods.

1. **Your Plan**

Extract key information from the dataset and use it to develop a model that recommends related shows and movies.

1. **References**

https://www.kaggle.com/shivamb/netflix-shows/tasks?taskId=2447

https://www.kaggle.com/sankha1998/collaborative-book-recommendation-system

https://www.kaggle.com/sankha1998/collaborative-movie-recommendation-system

<https://www.kaggle.com/sankha1998/content-based-movie-reommendation-system>