Gurukul International Multidisciplinary Research Journal (GIMRJ)*with* International Impact Factor 8.249
Peer Reviewed Journal
https://doi.org/10.69758/GIMRJ2406I8V12P030

e-ISSN No. 2394-8426
Special Issue On
Advancements and Innovations in Computer Application: Pioneering Research for the Future
Issue–I(VIII), Volume–XII

Machine Learning Model for Movie Recommendation System

Abhishek Moreshwar Khadse
P G Scholar
Master Of Computer Application,
G H Raisoni University, Amravati,

Abstract— The primary aim of recommendation systems is to recommend applicable objects to a consumer-based totally on ancient data. If a movie is rated excessive by means of a consumer who also watched the movie you are watching now, it's miles possibly to show up inside the recommendations. The films with the highest overall scores are in all likelihood to be enjoyed by way of nearly everyone. The algorithm which does all these features is called CineMatch. For personal users, it also learns from the conduct of the person to higher expect a movie the consumer is anticipated to be fascinated in. Here we have to increase our CineMatch algorithm 10% by using fashionable collaborative filtering techniques.

Keywords—Machine learningmodels,Movies,Ratings,Similarity,matrix,Sparse matrix.

I. INTRODUCTION

A. Motivation and Scope

We are leaving the age of facts and coming into the age of recommendation. Like many device mastering techniques, a recommender system makes a prediction based on users' ancient behaviors. Specifically, it's to expect user choice for a fixed of items based totally on past experience.

B. Need to study

Recommendation systems are getting increasingly important in today's extraordinarily busy world. People are always short on time with the myriad duties they need to accomplish within the restrained 24 hours. Therefore, the recommendation structures are vital as they help them make the right choices, without having to dissipate their cognitive resources. The reason for a recommendation system essentially is to look for content that would bethrilling to an individual. Moreover, it includes a number ofthings to create customized lists of beneficial and exciting content unique to every user/individual. Recommendation structures are Artificial Intelligence primarily based algorithms that skim thru all possible alternatives and create a customized listing of objects which might be thrilling and relevant to an individual.

C. Literature Survey/Review of Literature

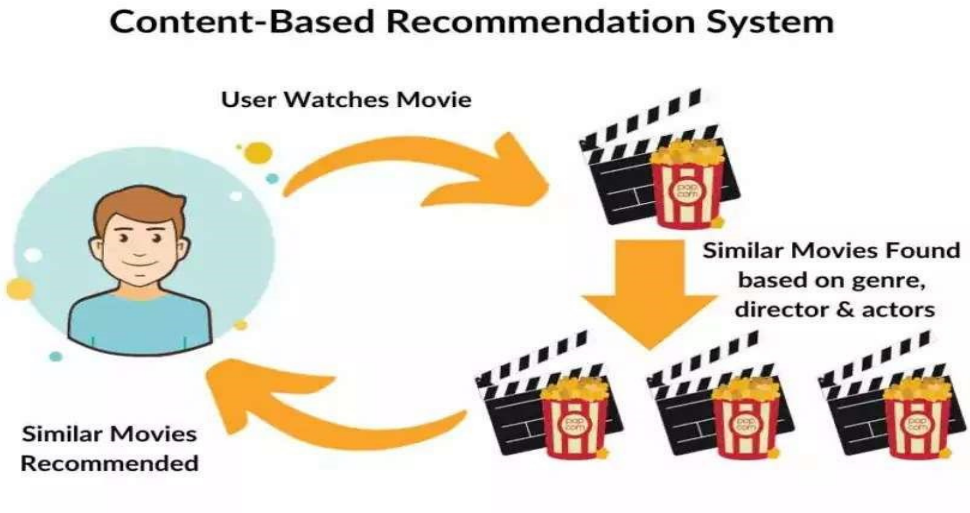
- The two principal tasks addressed by way of collaborative filtering techniques are rating prediction and rating. In contrast, ranking fashions leverage implicit feedback (e.g. Clicks) so that you can offerthe user with a customized ranked listing of encouraged items .
- With the increasing need for retaining confidential statistics at the same time as supplying tips, privacy- maintaining Collaborative filtering has been receiving increasing attention. To make statistics proprietors experience more comfortable even as imparting predictions, various schemes were proposed to estimate pointers without deeply jeopardizing privacy. Such methods dispose of or reduce statistics proprietors' privacy, financial, and legal concerns by means of employing exceptional privacy-retaining techniques].
- In the spread of information, the way to quickly locate one's favorite film in a massive variety of movies end up a very essential issue. Personalized recommendation machines can play a crucial role in particular whilst the person has no clean target movie..

- In this paper, we design and implement a movie recommendation machine prototype blended with the actual wishes of movie recommendation thru gaining knowledge of KNN algorithm and collaborativefiltering algorithm.
- In this study, we examine a privacy-retaining collaborative filtering method for binary facts referred to as a randomized reaction technique. We develop a method focused on the second thing of privacy to find out faux binary rankings the usage of auxiliary and public information .
- If privacy measures are provided, they may decideto grow to be worried about prediction generation processes. We advocate privacy-maintaining schemes getting rid of e-commerce sites' privaten ess concerns for imparting predictions on allotted data .
- With the improvement of the Internet and e- commerce, the recommendation machine has been widely used. In this paper, the electronic commerce recommendation system has a similar look at and makes a specialty of the collaborative filtering algorithm in the utility of personalized film recommendation system

II. RESEARCH GAP

The data set provided quite a few rating information, and a prediction accuracy bar this is 10% better than what Cinematch algorithm can do on the equal training data set. (Accuracy is a measurement of the way closely predicted scores of films in shape subsequent actual rankings).And we have to Predict the score that .

RESEARCH METHODOLOGY



User-Item Sparse Matrix

In the User-Item matrix, each row represents a person and every column represents an object and every cell represents rating given with the id of a user to an item.

A. User-User Similarity Matrices

Here, two customers could be similar to the premise of the comparable ratings given with the id of each of them. If any two users are similar then it means both of them have given very comparable scores to the items due to the fact here the consumer vector is nothing however the row of a matrix which in flip contains rankings given through user to the items. Now considering cosine similarity can variety from '0' to '1' and '1' means the highest similarity, so consequently, all the diagonal elements could be '1' because the similarity of the consumer with him/herself is the highest. But there's one

hassle with user-user similarity. User alternatives and tastes change over time. If any consumer favored some item one year in the past then it isn't important that he/she will like the identical object eventoday.

B. Item-Item Similarity Matrix

Here, two items can be comparable to the idea of the comparable rankings given to each of the items via all of the users. If any two gadgets are comparable then it means both of them had been given very comparable ratings by means of all of the users due to the fact here the item vectoris nothing however the column of the matrix which in flip contains scores given with the aid of consumer to the objects. Now due to the fact cosine similarity can variety from '0' to '1' and '1' means the highest similarity, so consequently, all of the diagonal elements might be '1' due to the fact the similarity of an item with the identical itemis the highest.

C. Cold Start Problem

The cold start problem concerns the personalized guidelines for users without a few past histories (new users). Providing suggestions to users with small beyond history turns into tough trouble for CF models due to the fact their studying and predictive ability is limited.

III. SURPRISE LIBRARY MODELS

A. XGBoost

However, with regards to small-to-medium structured/tabular data, choice tree primarily based algorithms are taken into consideration best-in-class proper now.. XGBoost and Gradient Boosting Machines (GBMs) are each ensemble tree techniques that follow the precept of boosting susceptible learners using the gradient descent architecture.

B. Surprise Baseline

This Algorithm predicting a random totallyon the data.

Predicted rating: (baseline prediction)

$$\hat{r}_{ui} = b_{ui} = \mu + b_u + b_i$$

μ : Average of all trainings in training data.
bu : User bias.
bi : Item bias (movie biases)

C. Suprise KNNBaseline Predictor

It is a number one collaborative algorithmconsidering a filtering baseline rating.

Predicted Rating: (based on User-User similarity)

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{v \in N_i^k(u)} \text{sim}(u,v) \cdot (r_{vi} - b_{vi})}{\sum_{v \in N_i^k(u)} \text{sim}(u,v)}$$

This is exactly same as our hand-crafted features 'SUR'- 'Similar User Rating'. Means here we have taken 'k' such similar users 'v' with user 'u' who also rated movie 'i'. $rvi$is the rating which user 'v' gives on item 'i'. bvi is the predicted baseline model rating of user 'v' on item 'i'.Generally, it will be cosine similarity or Pearson correlation coefficient.

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in N_u^k(i)} \text{sim}(i,j) \cdot (r_{uj} - b_{uj})}{\sum_{j \in N_u^k(j)} \text{sim}(i,j)}$$

Predicted rating (based on Item Item similarity):

A. Matrix Factorization SVD

The Singular-Value Decomposition, or SVD for short, is a matrix decomposition technique for decreasing a matrix to its constituent elements in order to ensure the next matrix calculations simpler. The SVD is used broadly both within the calculation of different matrix operations, including matrix inverse, but also as a statistics reduction approach in machine learning.

Predicted Rating:

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

qi — Representation of item(movie) in latent factor space.
pu — Representation of user in new latent factor space.

D. Matrix Factorization SVDpp

Here, an implicit rating describes the fact that a consumer u rated an item j, regardless of the rating value

yi is an object vector. For every object j, there is an object vector yj that is an implicit remarks. Implicit feedback in a roundabout way displays opinion by looking at consumer behavior including purchase history, surfing history, seek patterns, or even mouse movements. Implicit comments commonly denotes the presence or absence of an event

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T \left(p_u + |I_u|^{-\frac{1}{2}} \sum_{j \in I_u} y_j \right)$$

Iu — the set of all items rated by user u.

yj— implicit ratings.

For example, there's a film 10 in which a person has just checked the info of the film and spend some time there, which will contribute to implicit rating. Now, since here our records set has now not provided us the details that for how long a person has hung out on the movie, so right here we are considering the fact that despite the fact that a user has rated some film then it means that he has spent some time on that film which contributes to implicit rating. If person u is unknown, then the bias bu and the elements pu are assumed to be zero. The equal applies for item i with bi, qi, and yi

IV. MPLEMENTATION

Reading and Storing Data

The dataset I am working with is downloaded from Kaggle .https://www.kaggle.com/Netflixinc/Netflix-prize-data.

It consists of four .txt files and we have to convert the four .txt files to .csv file. And the .csv file consists of the following attributes.

Then we need to do away with duplicates, Duplicates are the values which befell extra than once inside the given information. Here we should find the duplicates and dispose of it by way of duplicate characteristic

A. Performing Exploratory Data Analysis on Data

In statistics, exploratory data analysis isn't the same as initial data analysis (IDA), which focuses extra narrowly on checking assumptions required for version becoming and hypothesis trying out, and coping with lacking valuesand making transformations of variables as needed making transformations of variables as needed

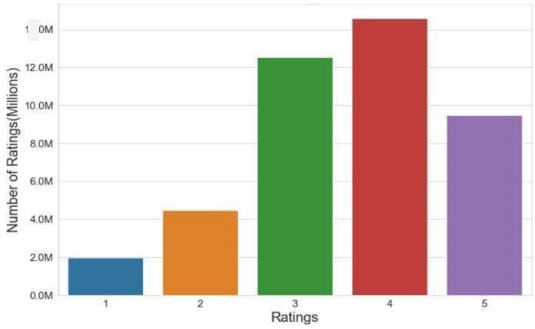


Fig. 1. Distribution of Ratings in data

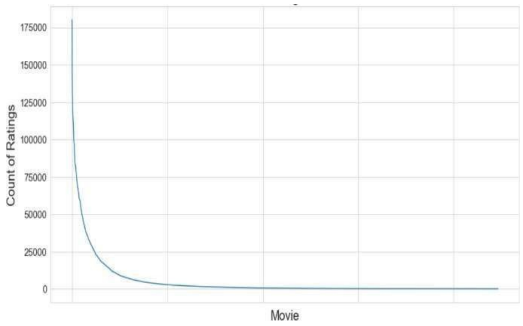


Fig. 2. Analysis of Ratings per movie.

The above graph shows the distribution of ratings from the data set. For example it implies that there are 2millions of ratings with a rating of 1.And similarly for the reaming ratings also.

It clearly shows that there are some movies which are very popular and were rated by many users as compared to other movies.

B. Creating User-Item sparse matrix for the data

Once the data preprocessing was completed then we have to create a user-Item sparse matrix for the data. Shape of sparse matrix depends on highest value of User ID and highest value of Movie ID.Then we have to find the global average of all movie ratings, average rating per user and average rating per movie. And next we have to compute the similarity matrices, there are mainly two similarity matrices such as user-user and item-item and we have to compute both matrices with our data set.And there is a csv file which consists of movie names for the movie id's which are present in our data set.

Let's check does movie-movie similarity works. Pick a random movie and check its top 10 most similar movies.

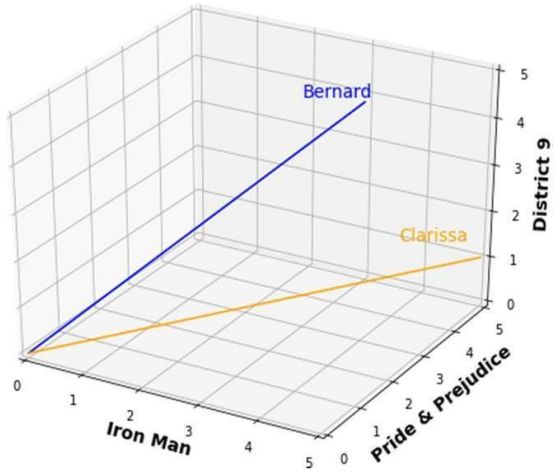
Suppose pick a movieid with number 17767. The number with particular movieid is picked from the movie titles and will show the name of the movie. Then by using the movie-movie similarity matrix we can find the total number of ratings given to the particular movie and it will also show the similar movies.

For example the movie with movieid 17767 is American experience. The top ten similar movies for American experience are as follows

D. Applying Machine Learning Models

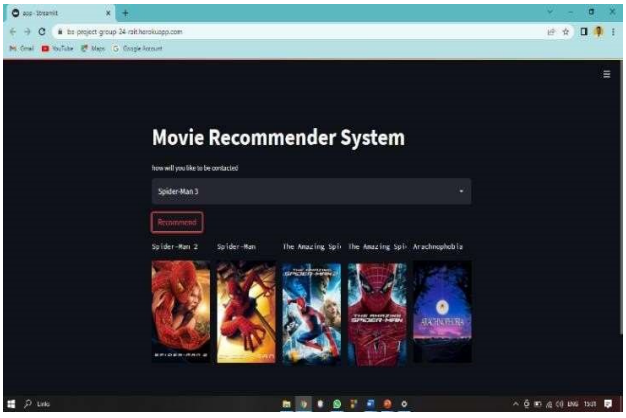
Before us applying the models we have to featurize data for the regression problem. Once it was completed we have transform data to surprise models. We can't giveraw data (movie, user, and rating) to train the model in Surprise library. Following are the models which we are applying for the data.

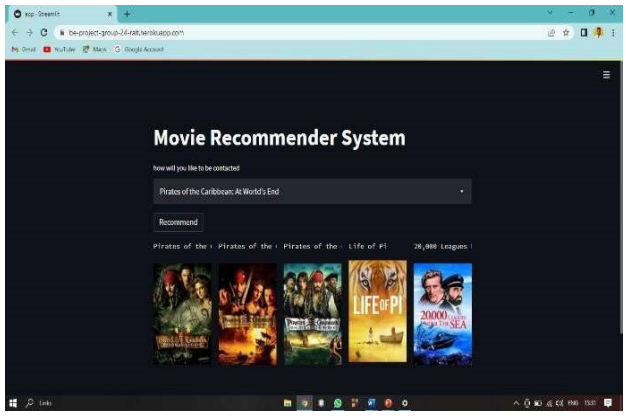
1. XGBoost was the first model which we are applying for the featurize data. When we run the model we get the RMSE and MAPE for the train and test data

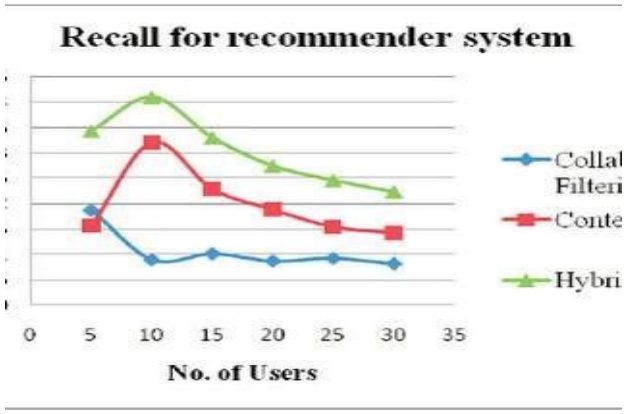


V. RESULTS AND DISCUSSION

With the model trained, it must be tested to check if it'd operate well in planet things. A comparative analysis is essential for evaluating the performance of your movie recommendation system in relation to existing methods or alternative approaches. Here's how you can structure this analysis in your presentation:







VI. CONCLUSION

So, far our best model is SVDpp with Test RMSE of 1.0675.Here we are not much worried about our

RMSE because we haven't trained it on the whole data . Our main intention here is to learn more about Recommendation Systems .If we taken whole data we would definitely get better RMSE .

VII. FUTURE ENHANCEMENT-

Tune hyper parameters of all the Xgboost models aboveto improve the RMSE Here we used 10K users and 1K movies to train the above models due to my pc ram issues In the future, I am going to run on the entire information set using cloud resources

Acknowledgment

We would like to express our sincere gratitude to everyone who contributed to the successful development and deployment of our movie recommendation system.

Firstly, we extend our deepest appreciation to our project supervisors and academic mentors, whose invaluable guidance and insightful feedback have been instrumental throughout this journey. Their expertise and encouragement have significantly shaped the direction and quality of this work

We are profoundly thankful to our peers and colleagues for their continuous support and collaboration. Their constructive criticism and shared knowledge have greatly enriched our project.

Special thanks to the data providers, including the online movie databases and streaming services, for granting access to the extensive datasets that were crucial for training and validating our model. Without their contributions, this project would not have been possible.

We also acknowledge the open-source community for providing the robust libraries and frameworks that facilitated the development of our machine learning algorithms. Tools such as TensorFlow, PyTorch, Scikitlearn, and others have been indispensable in our work.

Our heartfelt thanks go to our family and friends for their unwavering support and understanding during the countless hours dedicated to this project.

Finally, we would like to thank the users of our movie recommendation system for their feedback and engagement. Their input has been essential in refining our model and enhancing its performance. This project has been a collective effort, and we are truly grateful to everyone who contributed in various ways to its success.

REFERENCES

1) Davidsson C, Moritz S. Utilizing implicit feedback and context to recommend mobile applications from first use. DOI: 10.1051/,04008 (2017) 712012ITA 2017 ITM Web of Conferences itmconf/201 40084 In: Proc. of the Ca RR 2011. New York: ACM Press, 2011. 19
http://dl.acm.org/citation.cfm?id=1961639 [doi:10.1145/1961 634.1961
2) Bilge, A., Kaleli, C., Yakut, I., Gunes, I., Polat, H.: A survey of privacy-preserving collaborative filtering schemes. Int. J. Softw Eng. Knowl. Eng. 23(08), 1085–1108 (2013)CrossRefGoogle Scholar
3) Calandrino, J.A., Kilzer, A., Narayanan, A., Felten, E.W., Shmatikov, V.: You might also like: : privacy risks of collaborative filtering. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 231–246, Oakland, CA.
4) Okkalioglu, M., Koc, M., Polat, H.: On the discovery of fake binary ratings. In: Proceedings of the 30th Annual ACM Symposium on Applied Computing, SAC 2015, pp. 901–907. ACM, USA(2015).
5) Kaleli, C., Polat, H.: Privacy–preserving naïve bayesian classifier based recommendations on distributed data.
6) Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30,
7) Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237
8) Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19
9) Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", *International Journal of Scientific Research in Science and Technology (IJSRST)*, 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682
10) Usha Kosarkar, Prachi Sasankar(2021), " A study for Face Recognition using techniques PCA and KNN", Journal of Computer Engineering (IOSR-JCE), 2278-0661,PP 2-5,
11) Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w
12) Usha Kosarkar, Dipali Bhende, " Employing Artificial Intelligence Techniques in Mental Health Diagnostic Expert System", International Journal of Computer Engineering (IOSR-JCE),2278-0661, PP-40-45, https://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/9.%2040-45.pdf?id=7557