Application Of Machine Learning in Recommendation System: A Movie Recommender

Rehmanali Dauva[1],K Amuthabala[2], Sahil Raja Mohammed[3], Sagar Kumar Rajput[4], Salahuddin Khan[5]

[1,3,4,5]Students,School of CSE, REVA UNIVERSITY

[2]Assosicate Professor,School of CSE,REVA UNIVERSITY

Abstract.

The use of internet has brought vast numbers of users online through different platforms. Unlike old days internet is not just limited to email surfing, there's so much in the internet or let's say everything is on internet. In modern world it feels like not just internet but the whole world is revolving around the content. Different kind of users have different taste, and they demand different kind of content. This scenario has made in rise of video sharing platform like YouTube and various OTT platforms like Netflix or Amazon prime videos and many more. When the internet is revolving around content, one of the major contents are movies. Recommendation systems in such cases are used to recommend user movies based on various factors like genre, rating etc. Recommendation systems use various algorithms to suggest best suitable movie based in various factors. We have implemented final datasets following algorithms: a. Singular Value Decomposition (SVD) b. K-Nearest Neighbor (KNN) Algorithm. KNN outperforms SVD in terms of Mean Absolute Error (MAE) and Root Mean Square Error(RMSE). The recommended movie in KNN is more precise than SVD and in each value of K in cross-validation, it is clearly seen that KNN is much better. Hence, KNN is a better model in movie recommendation than SVD.

Keywords. K-mean Algorithm, SVD, RSs- Recommendation Systems.

1. INTRODUCTION

A Recommendation System or simply a recommender would be a model or system where user would get suggestions for various thing in real life, it could be while shopping online or watching videos online. Recommender Systems (RSs) assist individuals in discovering new or recommended things. It helps consumers find new items and services, such as movies, videos, podcasts, series and sometimes even people, example Facebook Suggests friends in You May Know list. These mechanisms are also important in making decisions, assisting humans in getting optimum results or lowering the negative impacts or hazards. Many information-based organizations, like Google, Twitter, LinkedIn, and Netflix now use RSs to better serve their user base. The field of RSs is thought to have started almost in the 1990s when team of researchers and developed launched first RSs known as Tapestry. Scholars & Researchers have been studying the application of Machine Learning (ML) algorithms which is subset or an artificial intelligence field, as the RS field has been evolving (AI). The study or research of Machine Learning was started in late 1950s when

the domain of AI-first started growing. Presently, there are huge number of ML algorithms (Such as SVD, k-nearest neighbor or Bayes network are few of them) which can or are used in applications ranging from small application like vacuum cleaner robots to complex life problems like providing help for differently-abled people or maybe in pattern identifying in images and self-driving cars etc.

ML algorithms provides personalized recommendations in various domains especially for Recommendation systems, as previously stated. However, due to the large number of techniques and tweaks presented in the literature, the ML field needs a comprehensive classification approach for its algorithms. As a result, when developing a RS, choosing an ML approach that meets one's needs becomes difficult and perplexing.

An approach for assisting practitioners and researchers in deciding the Machine Learning algorithm to use in a system for recommendation and discovering the areas that may be progressed in the development of RSs is to examine the RS and ML areas. Implementing the ML algorithms can help highlight patterns of growth and paves the way for further researches. As a result of this systematic study, researchers and practitioners should be able to understand more about the Recommendation application area and make well-informed research and implementation decisions.

A. Existing System

Over the past couple of decades, many recommendation systems are proposed that use various filtering methods. Different Big Data and Machine learning approaches are taken into consideration while developing the systems. Researchers and Scholars previously developed a collaborative filtering-based recommender system that leverages user ratings to offer suggestions.

B. Problem Definition

- To identify false Rating
- To identify Ranking Prediction

C. Objectives

- To identify best ML algorithm suitable for RSs
- To recommend the user movies based on rating and other factors
- To eliminate false rating

2. METHODOLOGY

A. System Design and Architecture:

People who enjoy viewing movies will benefit from the system. Different people have different tastes in movies. The figure below shows the system diagram of the movie recommendation system.

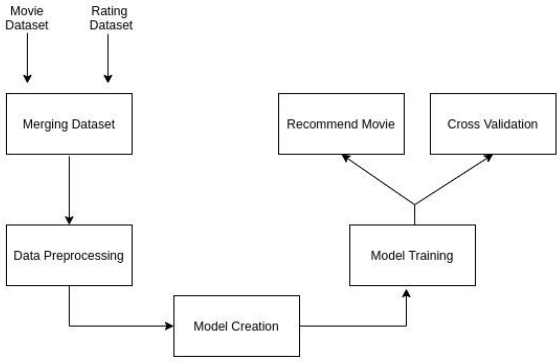

Figure-1:System Archiecture

A. *Data Collection* :
We gathered information from numerous movie rating websites like IMDB, MovieLens etc. Finally, collected data are in CSV format.

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

Figure-2:Movie Dataset

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931

Figure-3:Rating Dataset

userId	1	2	3	4	5	6	7	8	9	10	...	601	602	603	604	605	606	607	608	609	610
movieId																					
1	4.0	NaN	NaN	NaN	4.0	NaN	4.5	NaN	NaN	NaN	...	4.0	NaN	4.0	3.0	4.0	2.5	4.0	2.5	3.0	5.0
2	NaN	NaN	NaN	NaN	NaN	4.0	NaN	4.0	NaN	NaN	...	NaN	4.0	NaN	5.0	3.5	NaN	NaN	2.0	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	NaN	5.0	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	3.0	NaN	NaN	NaN	NaN	NaN	NaN

Figure 4: Combined Dataset

3. DATA PREPROCESSING

After the data is collected, it must be examined to make sure that it is in the correct format. Only mandatory fields should be chosen. To clean the files, MySQL is used. Some manual work has been done with Excel. The Following Processes are used for data pre-processing:

A. Imputing NaN with 0:

userId	1	2	3	4	5	6	7	8	9	10	...	601	602	603	604	605	606	607	608	609	610
movieId																					
1	4.0	0.0	0.0	0.0	4.0	0.0	4.5	0.0	0.0	0.0	...	4.0	0.0	4.0	3.0	4.0	2.5	4.0	2.5	3.0	5.0
2	0.0	0.0	0.0	0.0	0.0	4.0	0.0	4.0	0.0	0.0	...	0.0	4.0	0.0	5.0	3.5	0.0	0.0	2.0	0.0	0.0
3	4.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure-5: 0 imputed Dataset

B. Removing Noise from the data:

In the actual world, ratings are sparse, hence very famous movies are used to acquire the points for data and movie passionate people are considered .No one would prefer to watch a movie with only a few ratings because are considered unreliable. The consideration of individuals on the other hand is solely based on the number of movies one has rated. Considering all of the above and doing some error testing and trial testing, we'll use various filters to reduce the noise inside the final dataset.

- A minimum of ten users must vote for a film to be eligible.
- To be eligible, a user should have voted on at least 50 films.

Visualize how these filters look like:

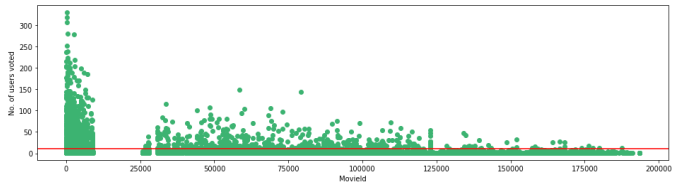

Figure-6: Representation of the number of visitors who voted using our 10-point criteria.

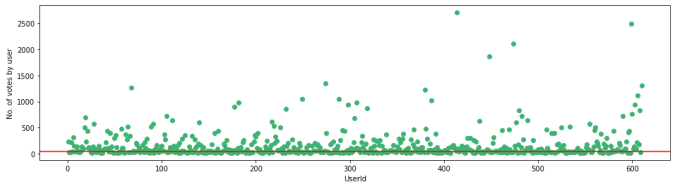

Figure-7: Representation of each user's voting results based on our 50-vote threshold

C. Removing Sparsity:

The enormous number of redundant zeros contained in the matrix structure makes sparse matrices computationally expensive. The difficulty of a huge scale greatly increases the complexity of space, making it difficult to solve these issues.

```
(0, 2)        3
(1, 0)        4
(1, 4)        2
(2, 4)        1
```

Figure-8: CSR Sample

Sparse values cannot be seen in the above CSR samples. Row and column indexes are assigned to the values. The number 3 occurs in the 0th row and 2nd column.

D. Model Creation:

We have implemented a movie dataset on two machine learning algorithms one of them is K-Nearest Neighbor (KNN) and other Singular Value Decomposition (SVD).

1) **Singular Value Decomposition:** A matrix's SVD is a factorization of that matrix into three vectors in mathematical concepts. It has several fascinating algebraic characteristics and offers crucial geometrical and theoretical insights regarding linear transformations.

```
U, S, V = randomized_svd(csr_data,
                         n_components=15,
                         n_iter=5,
                         random_state=42)
```

Figure-9:SVD

2) **K- Nearest Neighbour (KNN):** The full form of KNN stands for "K-Nearest Neighbour ". The approach could be used to solve both kind of problem i.e. classification as well as regression. The number of nearest neighbours to a new unknown variable that must be predicted or categorised is represented by the sign 'K' sign. It's a data classification approach that evaluates the likelihood of a data point belonging to one of group based on which measured values are adjacent to it.

```
knn = NearestNeighbors(metric='cosine', algorithm='brute', n_neighbors=20, n_jobs=-1)
knn.fit(csr_data)
```

```
NearestNeighbors(algorithm='brute', metric='cosine', n_jobs=-1, n_neighbors=20)
```

Figure-10:KNN

E. Model Training and Evaluation:

For Model Training, we have used CPU for the KNN algorithm and Colab GPU for the SVD algorithm.

1) **Single Value Decomposition:** We have trained and evaluated SVD models. For evaluation, we have used the same movie for recommendation and cross-validation approaches.
   a. **Movie Recommendation:** SVD was unable to recommend the movie accurately. Besides using different parameters tuning SVD fails to provide accurate results.

```
Recommendations for ['GoldenEye (1995)']:

GoldenEye (1995)
It Takes Two (1995)
White Balloon, The (Badkonake sefid) (1995)
Drop Zone (1994)
Hunted, The (1995)
Once Were Warriors (1994)
Hoop Dreams (1994)
Pie in the Sky (1996)
Unzipped (1995)
Walk in the Clouds, A (1995)
```

Figure-11: SVD Movie Result

   b. **Cross Validation:** Cross-validation, also referred as out-of-sample test or also called as rotation estimation, is a model validation methodology for examining how well statistical analysis results will generalize to a new data set. On subsequent iterations, cross-

validation is a resampling methodology that checks and trains a model using diverse chunks of data. While evaluating the results using SVD we have used 5 iterations. The Following values of MAE (Mean Absolute Error) and Root Mean Square Error (RMSE) e obtained.

Table-1:RMSE and MAE for SVD

Test	0.87	0.873	0.878	0.875	0.870
Test MAE	0.669364 43	0.672 15072	0.676 03402	0.671 06658	0.671 31548

2) **K-means Neighbor (KNN):** We have trained and evaluated KNN models. For evaluation, we have used the same movie for recommendation and cross-validation approaches.

a. **Movie Recommendation:** KNN was able to recommend the movie accurately. We have trained using different parameter tuning and accurate results were obtained.

	Title	Distance
1	Jurassic Park (1993)	0.412374
2	Stargate (1994)	0.409341
3	Batman (1989)	0.406579
4	Batman Forever (1995)	0.405572
5	Terminator 2: Judgment Day (1991)	0.404621
6	Mission: Impossible (1996)	0.397863
7	Speed (1994)	0.397093
8	Clear and Present Danger (1994)	0.390702
9	True Lies (1994)	0.387416
10	Die Hard: With a Vengeance (1995)	0.363683

Figure-12: KNN Movie Result

b. **Cross Validation:** We utilized five iterations to evaluate the findings using KNN. The following values of MAE (Mean Absolute Error) and RMS (Root Mean Square Error) are obtained (RMSE).

Table-2:RMSE and MAE for KNN

Test	0.9390	0.942	0.939	0.930	0.942
Test MAE	0.7176 5578	0.717 6315	0.720 06309	0.711 21797	0.720 1747

c. Content Based Filtering:

The underlying premise is that if you enjoy something, you'll like something "alike." It usually works effectively when determining the context/properties of each object is simple. The data user provides is used by a content-based recommender, such as explicit movie ratings from the MovieLens collection. Based on the data obtained, a user profile is developed, which ultimately gets used to provide suggestions to the user.The more the number of inputs obtained the accuracy of the machine gradually increases and can recommend far more accurately.

1050	Aladdin and the King of Thieves (1996)
2072	American Tail, An (1986)
2073	American Tail: Fievel Goes West, An (1991)
2285	Rugrats Movie, The (1998)
2286	Bug's Life, A (1998)
3045	Toy Story 2 (1999)
3542	Saludos Amigos (1943)
3682	Chicken Run (2000)
3685	Adventures of Rocky and Bullwinkle, The (2000)
236	Goofy Movie, A (1995)
12	Balto (1995)
241	Gumby: The Movie (1995)
310	Swan Princess, The (1994)
592	Pinocchio (1940)
612	Aristocats, The (1970)
700	Oliver & Company (1988)
876	Land Before Time III: The Time of the Great Gi...
1010	Winnie the Pooh and the Blustery Day (1968)
1012	Sword in the Stone, The (1963)
1020	Fox and the Hound, The (1981)

Figure-13: Filtered Movies

F. Comparison between SVD and KNN:

We have plotted the comparison multi-bar chart of both KNN and SVD algorithms and the result was compared between RMSE and MAE values.

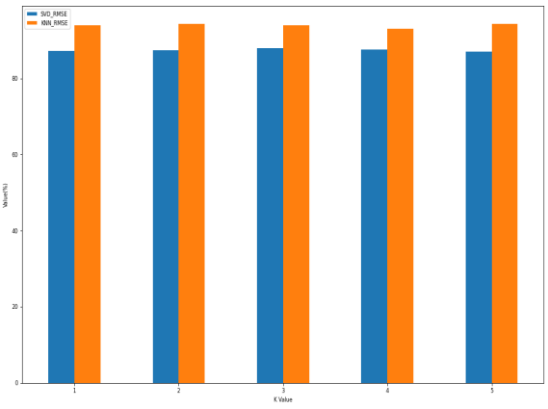

Figure-14 : Comparison between KNN and SVD of RMSE metric.

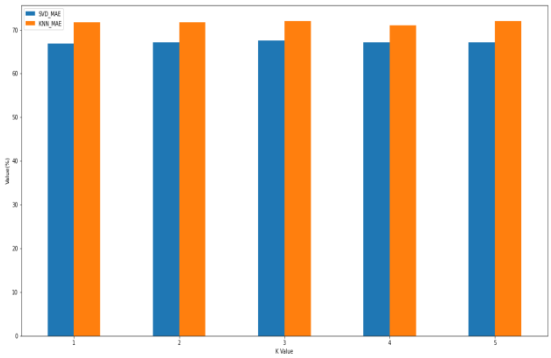

Figure-15: Comparison between KNN and SVD using MAE metric.

4. RESULTS

The Movie Recommender ML app proposes or suggests a movie to the user that is devoid of false ratings and has a very user-friendly interface. Out of KNN and SVD, KNN also proves to be the best algorithm to employ. The predictions were compared, and KNN was found to be the superior algorithm that is employed in the app and can be employed in future Recommendation Systems.

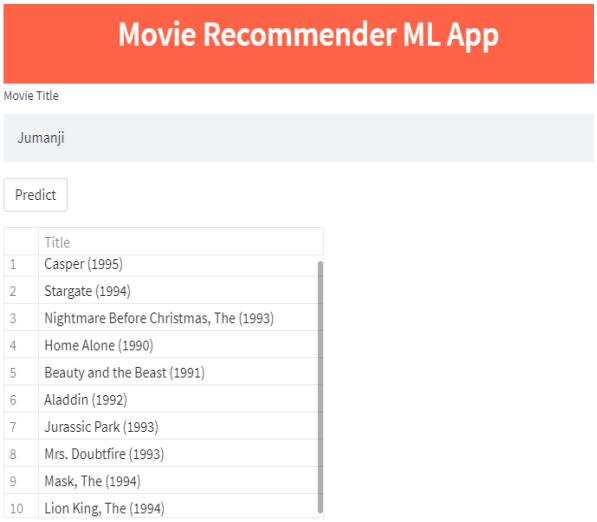

Figure-16: Movie Prediction Result

5. CONCLUSION

In AI based assistant, online selling platforms, social networking sites, and many other formal and informal sectors, recommender systems (RS) are now frequently used.

Since its inception in the mid-1990s, RSs research has advanced. Machine learning (ML) techniques, which allow machines to understand from user data and tailor suggestions much more, are a significant advancement in the history of RS. Machine learning is one of subset of Artificial Intelligence (AI) that includes algorithms that aim to foresee the outcome of data to be processed. In the disciplines of image identification, search engines, and security, machine learning has achieved significant advances. However, there are other algorithms in the ML field that have been reported in the literature, each with its own set of properties. There is no classification scheme for algorithms in the literature that shows which environments they are best suited for. As a result, selecting an ML algorithm for use in RSs is tricky. Furthermore, researchers in RSs lack a comprehensive vision of trends in Machine Learning Algorithms adoption, making it difficult to identify where to focus their field of study called research efforts.

Scientists and Developers on the other hand, need to determine which SE areas lack of enough assets or tools for RSs development. This paper then suggests a comprehensive review of the ML algorithms used in RSs, as well as what SE domains can help with the creation of such RSs.

We looked at the domains where RSs using an ML algorithm were authenticated during the systematic review. Due of the ease with which test data may be retrieved, movies, articles, and product reviews are the three popular domains.In the movie realm,MovieLens is one and IMDb the other are multiple online collections of movie ratings. As a result, we used the SVD and KNN Algorithms to build this dataset. In both RMSE and MAE , KNN has a higher value than SVD. The recommended movie in KNN is more precise than SVD and in each value of K in cross-validation, it is seen that KNN is much better. Hence, KNN is a better model in movie recommendation than SVD.

6. REFERENCES

[1] Rekha KB, Gowda NC, "A framework for sentiment analysis in customer product reviews using machine learning", International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 267-271, Oct 2022.

[2] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6), 734-749.

[3] 3. Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In Recommender systems handbook (pp. 217-253). Springer US.

[4] 4. Ahmed, A., Kanagal, B., Pandey, S., Josifovski, V., Pueyo, L. G., & Yuan, J. (2013, February). Latent factor models with additive and hierarchically-smoothed user preferences. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 385-394). ACM.

[5] P Amuthabala, Dr. R Santosh,"Robust analysis and optimization of a novel efficient quality assurance model in data warehousing", Computers and Electrical Engineering. 2019, 74, 233–244.

[6] Basha, S. M., Poluru, R. K., & Ahmed, S. T. (2022, April). A Comprehensive Study on Learning Strategies of Optimization Algorithms and its Applications. In *2022 8th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-4). IEEE.

[7] P Amuthabala, Dr. M.Mohanapriya ,"Pattern Based Technique in Evaluation of Data Quality on Complex Data", Journal of Advance research in Dynamical and control systems, 2017, Issue-15,36-41.

[8] Apte, C. (2010). The role of machine learning in business optimization. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 1-2).

[9] P Amuthabala, M Mohanapriya," Cost Effective Framework for Complex and Heterogeneous Data Integration in Warehouse", CSOC 2016, Vol 2, Software Engineering Perspectives and Application in Intelligent Systems, ISBN: 978-3-319-33622-0, pp. 93-104, 2016. © Springer International Publishing Switzerland 2016.