**MOVIE RATING PREDICTION USING MACHINE LEARNING TECHNIQUES**

Guggilam Amarnath1

[**guggilam.amarnath2020@vitstudent.ac.in**](mailto:guggilam.amarnath2020@vitstudent.ac.in)

Perubhotla Srinivasa Aditya Manish2

[**srinivasa.adityamanish2020@vitstudent.ac.in**](mailto:srinivasa.adityamanish2020@vitstudent.ac.in)

Naveen Krishna Makeena3

[**makkenanaveen.krishna2020@vitstudent.ac.in**](mailto:makkenanaveen.krishna2020@vitstudent.ac.in)

School of Computer Science and Engineering,

Vellore Institute of Technology (VIT), Chennai, Tamil Nadu, India.

**ABSTRACT**

To produce precise suggestions, recommendation systems employ ratings that users have provided to things. Companies that sell a large number of items to a large number of consumers and allow those customers to evaluate their products, such as Amazon, can accumulate vast datasets that can be used to forecast what rating a certain user will give to a specific item. Items with a high expected rating for a certain user are then recommended to that person. The same might be said of other products, such as movies in our situation. One of the most often used models in machine learning algorithms is recommendation systems. Netflix's success is claimed to be due to its powerful recommendation algorithm. In reality, the Netflix reward (an open competition for the best collaborative filtering algorithm to forecast user ratings for films based on prior ratings without any other information about the users or films) is representative of this algorithm for products recommendation system. For our movie rating prediction system we are going to use several machine learning algorithms such as logistic regression, Navies Bayes, Support Vector Machine, KNN, Decision Tree, Random Forest, Perceptron, Stochastic Gradient Descent.

1. **INTRODUCTION**

Movie industry has been expanded worldwide and now people, wherever on earth, have chance to watch a movie in the day it is released. There is a huge sector behind the preparation phases of each film and lots of directors and movie stars have burst. Every year, hundreds of films are produced. These movies have different genres, varying from comedy to romance or war to science fiction. To keep track of every movie produced, an online platform was needed. We have Taken a dataset from movielens, which has the data movies of more than 1000 and the ratings of the movies given by the users. In this project our plan is to develop a model that can predict how a user will rate a specific movie, similar to a movie recommendation system. Our model will make predictions based on user ratings of other movies and the average rating of the specific movie.

1. **LITERATURE SURVEY**

[1]Arash Oshnoudi, Behzad Soleimani Neysiani\*, Zahra Aminoroaya andNaser Nematbakhsh in the paper titled, “ Improving Recommender Systems Performances Using User Dimension Expansion by Movies’ Genres and Voting-Based Ensemble Machine Learning Technique” proposed a novel approach for prediction the movies and rating them based on there geners.[2] Processing of System start with the data collection and According to the data mining process as a conceptual model, The experiments performed on the MovieLens dataset show that the proposed method is more successful than other previous methods in predicting user clusters with 93.81% accuracy, 94.45% precision, and 92.81% recall.[3]

[4]O. Bora Fikir, lker O. Yaz, Tansel Özyer in the paper titled, “ A Movie Rating Prediction Algorithm with Collaborative Filtering” proposed a novel approach for prediction the movies and rating them based on there geners. This Method performs a novel collaborative filtering method on the entire missing values. [5]Iteratively, predicts ratings in random order. As missing values are predicted they are used for latter missing values. We have proposed an algorithm for predicting all missing values and used QR factorization method for predicting each entry.

[6]Xiaoyue Li, Zhezhou Yu, Zhuo Wang and Haonan Zhao in the paper titled, “Research on Movie Rating Prediction Algorithms” proposed a novel approach for predicting and rating movies using the RF that uses the users' activity and rating to select suitable experimental data and proved this method can efficiently reduce RMSE and MAE of various recommendation algorithms.[6] The MCBF-SVD uses weighting factors to discuss the impact of movie categories on predicting future rating behavior of users, and also improves the filtering method based on movie categories.

[7]Farshad B. Moghaddam, Mehdi Elahi, Reza Hosseini, Christoph Trattner, Marko Tkalciˇ cˇ in the paper(2019) titled, “Predicting Movie Popularity and Ratings with Visual Features”. The proposed method uses a large dataset, containing 13053 movie trailers that had their titles available in the Movielens dataset. Prior work has reported large values of similarity, based on visual features, extracted from movie trailers and their corresponding full-length movies. [8]For each movie, we also collected the meta-data such as #ratings, average rating, genre, and the year of production from IMDB. Every movie can have one or multiple genre label(s) out of 30 possible genres (e.g., drama, comedy, romance, etc.). The following list represents the entire methodology (Splitting movies, Identifying Key-frames, Extracting key features, Aggregating features, Training & Predicting.[9]

[9]Kavya Pradeep, Tintu Rosmin C R, Sherly Susana Durom, G S Anisha in the paper(2020) titled, “Decision Tree Algorithms for Accurate Prediction of Movie Rating”. This model assists with discovering the rating of the upcoming motion picture through qualities or attributes of that movie. In this proposed system, there are three different algorithms. By comparing these algorithms we identify the best algorithm that shows the highest accuracy and with the help of this algorithm we can predict the success of upcoming movies. [10]For finding the accurate algorithm, we downloaded dataset from Kaggle.com namely “Bollywood movies”. It needed a high level cleaning. We included rating, budget and net-gross for each movie manually with the help of IMDB and Box-Office India websites. We then processed this raw data in order to classify those using decision algorithms. 400 movies from the considered movie set had taken for training set. This training set is pre-processed and loaded in the first stage. After that using this remaining data is classified considering the classified training set[11]

[12]Rijul Dhir, Anand Raj in the paper(2018) titled, “Movie Success Prediction using Machine Learning Algorithms and their Comparison”. [13]It proposes a way to predict how successful a movie will be prior to its arrival at the box office instead of listening to critics and others on whether a movie will be successful or not.[18] The first step is to identify a dataset of movie data that’s representative and suitable for analysis. Relevant attributes of such data must include general pre-production information regarding film productions such as genre, language and information about the actors and directors involved. Likewise, the data must also include some measure of success, such as user originated movie ratings. Secondly, the relevant dataset has to be prepared and structured in such a way that the data used is representative of the movie scene at large, as well as viable for analysis by the relevant machine learning techniques and algorithms. [14][15]Lastly, the prediction performance of the relevant machine learning algorithms has to be evaluated based on the specified dataset. This means that a set of suitable tools has to be acquired, as well as configured for evaluating both algorithms in comparison to each other based on the data, whilst still ensuring equivalence between in measurements.

[16]Taeryong Jeon,, Soojin Lee , Gyeongdong Baek , Sungshin Kim and Jaewoo Cho in the paper titled, “ A Movie Rating Prediction System of User Propensity Analysis based on Collaborative Filtering and Fuzzy System” proposed a novel approach for the prediction the movies and rating them based on there geners to solve the collaborative filtering problems by that we just got the results from the samples with 99% of reliability.

[17]May Saikiran Gogineni, and Anjusha Pimpalshendeank in the paper titled, “ Predicting IMDB Movie Rating Using Deep Learning” proposed a novel approach for the prediction the movies and rating on the imbd using deep learning .[19] The work done will be useful to understand the potential of classical machine learning algorithms and the latest deep learning architectures in evaluating emotion from the raw text.

[20]Warda Ruheen Bristi, Zakia Zaman and Nishat Sultana in the paper titled, “Predicting IMDb Rating of Movies by Machine Learning Techniques” proposed a novel approach for the movies rating using machine learning .[17] Then machine learning classification algorithms are applied of the data set. Lastly an efficient model is developed to predict a movie’s IMDb rating. The model gives good classification measures with the data set.

1. **METHODOLOGY**
   1. **RESEARCH APPROACH**

We have collected a movielens website dataset. The film industry is one of the biggest contributors to the entertainment industry's unpredictability in success and failure (Raj & Aditya, 2017). Because of quick digitization and the rise of internet-based life the film business is developing significantly as the average number of movies produced per year is greater than 1000, therefore to make the movie profitable, it becomes a matter of concern that the movie succeeds (Bhave et al., 2015). The success rate is the fraction or percentage of success among several attempts, and also, the average task success rate can be calculated either per participant or per task that users complete correctly.The purpose of this project is to create a model that will predict how a user will rate a specific movie, similar to a movie recommendation system.Our model will make predictions based on user ratings of other movies and the average rating of the specific movie.We are going to make a deep insights from the dataset by doing analysis and build a model .

* 1. **METHOD APPROACH**

We have used Naive Bayes,SVM and KNN algorithms where Naive Bayes is a straightforward supervised machine learning technique that obtains results by applying the Bayes theorem to a set of features under the strict assumption of independence; this algorithm just takes the independence of each input variable for granted and Support Vector Machine Both classification and regression problems can be solved by the Support Vector Machine algorithm. The SVM algorithm objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. The hyperplane is the name given to this optimal decision boundary and Algorithm of K Nearest Neighbors. The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point.

We will also test three different regression models to predict each rating in the training set Then, I will select the best model and apply it to the test set .

**Model 1 :** Predicted Rating = Global Average Rating + Movie Effect

**Model 2 :** Predicted Rating = Global Average Rating + User Effect

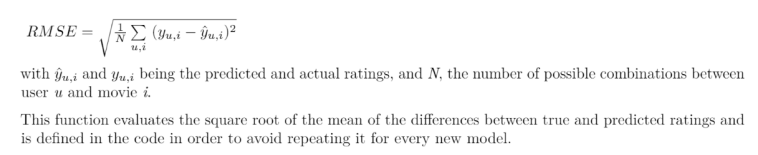
**Model 3 :** Predicted Rating = Global Average Rating + Movie Effect + User Effect

🡪The **global average rating** is the average rating across all entries in the dataset.

🡪The **movie effect** is the difference between the average rating for the specific movie and the global average rating.

🡪Similarly, the **user effect** is the difference between the average rating for the specific user and the global average rating. To evaluate three models we will use RMSE(Root Mean Square Error).

RMSE is used measure of the differences between values predicted by a model and the values observed. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular dataset, a lower RMSE is better than a higher one. The effect of each error on RMSE is proportional to the size of the squared error; thus larger errors have a disproportionately large effect on RMSE. 14 Three models that will be developed will be compared using their resulting RMSE in order to assess their quality. The function that computes the RMSE for vectors of ratings and their corresponding predictors will be the following

****

**3.3Regression Algorithm:**

The algorithm aimed to predict movie rating , dislikes, and the view count of a trailer, release date , star ranking, and so on. Multiple Linear Regression Algorithm was used for the prediction of earnings of the movie. Once the movie is released, we use social media and the opinion shared by people in respective platforms. The goal was to define a relationship between the prediction value and the features by solving for the linear coefficients, θ that best map the features to the prediction value. Where the ratings have been collected in a vector Y. Y is a (m x 1) vector (where m=50000). The movie set was to be pruned to select a set of features that have been found to make a major impact on the success or failure of a film. After the identification, all the producers, directors, actors, andactresses were rated based on their past performance at the Box Office.

**3.4Logistic Regression**

This statistical analysis method, logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable.

**3.5 Naive Bayes**

Naive Bayes is a straightforward supervised machine learning technique that obtains results by applying the Bayes theorem to a set of features under the strict assumption of independence; this algorithm just takes the independence of each input variable for granted.

**3.6 Support Vector Machine**

Both classification and regression problems can be solved by the Support Vector Machine algorithm. The SVM algorithm objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. The hyperplane is the name given to this optimal decision boundary.

**3.7 KNN**

Algorithm of K Nearest Neighbors. The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point.

**3.8 Decision Tree**

Decision Tree is a supervised learning technique that may be applied to classification and regression problems, however it works best when dealing with classification problems. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the result.

**3.9 Random Forest**

A supervised learning technique called Random Forest Regression leverages the ensemble learning approach for regression. The ensemble learning method combines predictions from various machine learning algorithms to provide predictions that are more accurate than those from a single model.

**3.10 Perceptron:**

A Perceptron is a neural network unit that does certain computations to detect features or business intelligence in the input data. This algorithm enables neurons to learn elements and processes them one by one during preparation.

Binary classifiers decide whether an input , usually represented by a series of vectors

**3.11 Stochastic Gradient Descent**

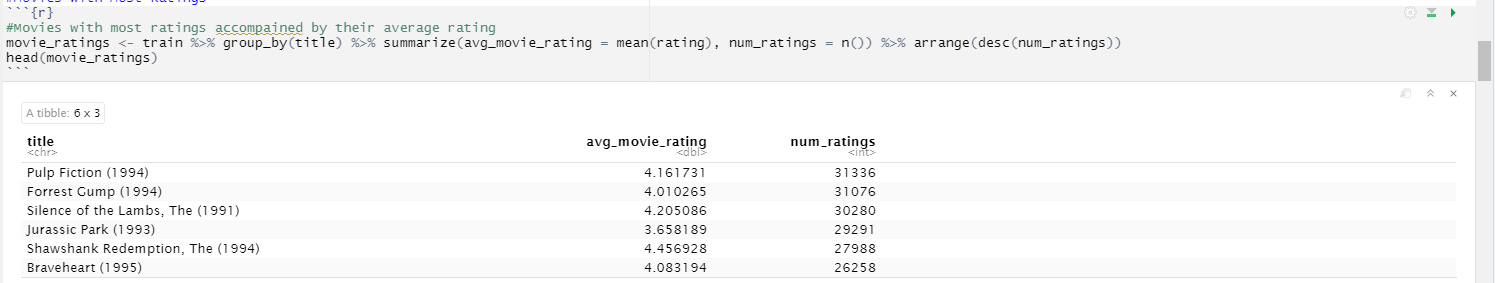
**Stochastic gradient descent** (**SGD**) is an iterative method for optimizing an [objective function](https://en.wikipedia.org/wiki/Objective_function) with suitable [smoothness](https://en.wikipedia.org/wiki/Smoothness) properties(e.g. [differentiable](https://en.wikipedia.org/wiki/Differentiable_function) or [subdifferentiable](https://en.wikipedia.org/wiki/Subgradient_method)). It can be regarded as a [stochastic approximation](https://en.wikipedia.org/wiki/Stochastic_approximation) of [gradient descent](https://en.wikipedia.org/wiki/Gradient_descent) optimization, since it replaces the actual gradient (calculated from the entire [data set](https://en.wikipedia.org/wiki/Data_set)) by an estimate thereof (calculated from a randomly selected subset of the data).

**3.12 Data Visualiztion:**

****

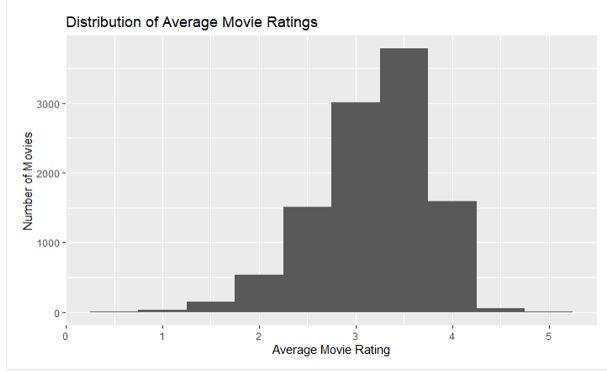
**Fig1.Distribution of Ratings**

From Fig1 we can see the ratings appear to be left-skewed since there are few ratings b/w 0 to 2 stars and many ratings b/w 3 to 5 stars .



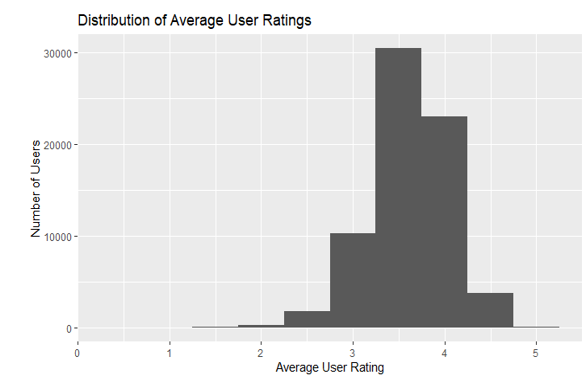
**Fig2.Movies By Ratings**

From Fig2, Pulp Fiction, Forrest Gump, The Silence of the Lambs, Jurassic Park, Shawshank Redemption, and Braveheart have the most ratings (about 30,000 each) with an average rating ranging between 3.66 and 4.46.



**Fig3:Distribution Of Movie Ratings**

Based on the histogram from Fig3, most movies appear to have an average rating between 2.5 and 4. In addition, there are only be a few movies with an average rating of 0.5 stars (worst possible rating) and 5 stars (perfect rating).

****

**Fig4.Distribution of Average User Ratings**

From Fig4, Based on the histogram, most users give an average rating between 3 and 4.5. In addition, only a few users have a very high or very low average rating (i.e. 0 to 2 stars; 5 stars).

1. **RESULT ANALYSIS**

The result of the research from Table1 concludes that the Random Forest Classifier algorithm shows more accurate results than other algorithms. In our dataset, all the attributes are having categorical values. The Stochastic Gradient Descent is a probability-based approach that has very low accurate results.

|  |  |  |
| --- | --- | --- |
| MODEL | ACCURACY | |
| Random Forest | 98.09 | |
| Decision Tree | 98.09 | |
| KNN | 52.71 | |
| Perceptron | 28.38 | |
| Support vector Machines | 28.08 | |
| Logistic Regression | 25.46 | |
| Naïve Bayes | 24.52 | |
| Linear Svc | | 18.16 |
| Stochastic Gradient Descent | | 15.63 |

Table1. Observations of each Machine Learning Algorithm.

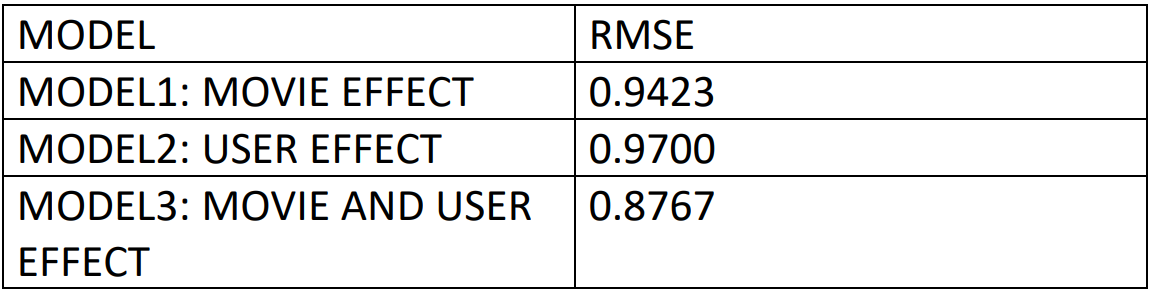


Table2:Observation for 3 different regression models

From Table2, Model 3 (Movie & User Effect) has the lowest RMSE. Thus, I will deploy this model to the validation dataset.

1. **CONCLUSION**

We are finally concluding with our model to help every one to save their money and time with this Foundations of Data Analysis project.After finding all the accuracies for the algorithms we found ,Random Forest ,Regression, Decision tree algorithm as the best suited algorithm as it has got more accuracy.We have decided to extend the modules into different supervised and Unsupervised algorithms.We will collect various data sources based on more survey results based on extensions.We enhance the model with better filtered test data and train data to get better results.

**REFERENCES**

1. S. Gaenssle, O. Budzinski, and D. Astakhova, “Conquering the box office: Factors influencing success of international movies in russia,” 2018.
2. M. Saraee, S. White, J. Eccleston *et al.*, “A data mining approach to analysis and prediction of movie ratings,” *Transactions of the Wessex Institute*, pp. 343–352, 2004.
3. M. H. Latif and H. Afzal, “Prediction of movies popularity using ma- chine learning techniques,” *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 8, p. 127, 2016.
4. S. Pramod, A. Joshi, and A. Mary, “Prediction of movie success for real world movie dataset,” *Int. J. of Advance Res., Ideas and Innovations in Technol*, vol. 3, no. 3, 2017.
5. R. Parimi and D. Caragea, “Pre-release box-office success prediction for motion pictures,” in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2013, pp. 571–585.
6. J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, “Movie success prediction using data mining,” in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017, pp. 1–4.
7. R. P. Ruhrla¨nder, M. Boissier, and M. Uflacker, “Improving box of- fice result predictions for movies using consumer-centric models,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 655–664.
8. R. Khandelwal and H. Virwani, “Comparative analysis for prediction of success of bollywood movie,” *Available at SSRN 3350907*, 2019.
9. B. C¸ izmeci and S¸ . G. O¨ gu¨du¨cu¨, “Predicting imdb ratings of pre- release movies with factorization machines using social media,” in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE, 2018, pp. 173–178.
10. S. Mundra, A. Dhingra, A. Kapur, and D. Joshi, “Prediction of a movies success using data mining techniques,” in *Information and Communication Technology for Intelligent Systems*. Springer, 2019, pp. 219–227
11. Burke, R. "Recommender Systems: Survey and Experiments", In User Modeling and UserAdapted Interaction. Vol 12. Kluwer Academic Publishers, 2002, pp. 2- 40.
12. Ho A T, Menezes I L L, Tagmouti Y. E-mrs: Emotion-based movie recommender system[C]//Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell. 2006: 1-8.
13. Ricci F R. L. and Shapira, B. 2011[J]. Introduction to recommender system handbook.
14. A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. a. Mikic-Fonte, and A. Peleteiro, "A hybrid contentbased and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition," Inf. Sci. (Ny)., vol. 180, no. 22, pp. 4290–4311, Nov. 2010.
15. Paterek A. Improving regularized singular value decomposition for collaborative filtering[C]//Proceedings of KDD cup and workshop. 2007, 2007: 5-8.
16. Barragáns-Martínez A B, Costa-Montenegro E, Burguillo J C, et al. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition[J]. Information Sciences, 2010, 180(22): 4290-4311.
17. Vozalis M G, Margaritis K G. Applying SVD on item-based filtering[C]//5th International Conference on Intelligent Systems Design and Applications (ISDA'05). IEEE, 2005: 464-469.
18. D. Billsus and M. J. Pazzani. Learning collaborative information filters. In 15th International Conference on Machine Learning, Madison, WI, 1998
19. Reihanian A, Minaei-Bidgoli B, Alizadeh H. Topic-oriented community detection of rating-based social networks[J]. Journal of King Saud University-Computer and Information Sciences, 2016, 28(3): 303-310.
20. Çoba L, Zanker M. Replication and reproduction in recommender systems research-evidence from a case-study with the rrecsys library[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 305-314.