****

**School of Computer Science and Engineering**

**J Component report**

**Programme : B.Tech**

**Course Title : Foundations of Data Analytics**

**Course Code : CSE3505**

**Slot : F2**

**Title: << MOVIELENS RATING PREDICTION >>**

**Team Members: <<Guggilam Amarnath | 20BCE1543>>**

**<<Perubhotla Srinivasa Aditya Manish | 20BCE1932>>**

**<<Naveen Krishna Makeena| 20BRS1232>>**

**Faculty: Dr. Trilok Nath Pandey**   **Sign:**

**Date:**

*November 2022*

*­­­*

i

**DECLARATION**

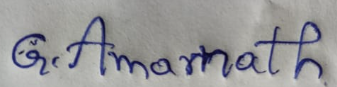
I hereby declare that the report titled “**MOVIELENS RATING PREDICTION”** submitted by me to VIT Chennai is a record of bona-fide work undertaken by me under the supervision of **Dr.Trilok Nath Pandey**, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai.

Signature of the Candidate

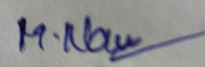
****

***PERUBHOTLA SRINIVASA ADTIYA MANISH***

*(20BCE1932)*

**

*GUGGILAM AMARNATH(20BCE1543)*

**

*NAVEEN KRISHNA MAKEENA (20BRS1232)*

ii

**CERTIFICATE**

<Times New Roman, Font 14, Underlined, Bold, CAPS>

Certified that this project report entitled “**MOVIELENS RATING PREDICTION”** is a bonafide work of **GUGGILAM AMARNATH(20BCE1543), PERUBHOTLA SRINIVASA ADITYA MANISH(20BCE1932),NAVEEN KRISHNA MAKEENA(20BRS1232)** and they carried out the Project work under my supervision and guidance for CSE3505 – Foundations of Data Analytics.

**DR.TRILOK NATH PANDEY**

**SCOPE, VIT Chennai**

**ACKNOWLEDGEMENT**

First and foremost, praises and thanks to God, the Almighty, for His showers of blessings throughout our internship and its successful completion. I would also like to press our deep and sincere gratitude to Dr. Trilok Nath Pandey, our course faculty B.Tech Computer Science and Engineering.SCSE, VIT Chennai for providing us with invaluable supervision, and support. and tutelage during the course of our project work. I would also like to thank Dr. Ganesan R, Dean of the School of Computer Science & Engineering, VIT Chennai, and Dr. Parvathi R, Associate Dean (Academics) of the School of Computer Science & Engineering, VIT Chennai for their empathy, patience, and knowledge that he imparts unto us.

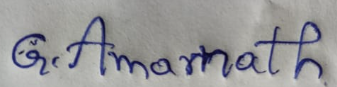
It was a great privilege and honor to work and study under of guidance Dr. Trilok Nath Pandey and greatly indebted to him for his full coordination and support in the completion of the project .

Signature of the Candidate

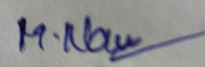
****

***PERUBHOTLA SRINIVASA ADTIYA MANISH***

*(20BCE1932)*

**

*GUGGILAM AMARNATH(20BCE1543)*

**

*NAVEEN KRISHNA MAKEENA (20BRS1232)*

**ABSTRACT**

Entertainment is no longer just anything that we enjoy occasionally, with over two million spectators a day, the amount generated by the movie industry is huge. The movie sector is one of the biggest contributors to the entertainment industry’s unpredictability in success and failure. The aim of this research work to design an efficient movie recommendation algorithm that will increase prediction accuracy, the Movie Review Rating Prediction (MR2P) was achieved through a systematic review of the existing movie success algorithm. This research work will enable movie stakeholders (producers, directors, crew, cast already in the movie industry or aspirants) to know the kind of movie to invest in which will, in turn, be beneficial in terms of higher profit.

Loss of revenue by stakeholders is the main concern in the film industry. The number of available viewing logs and friendship networks is too limited to design effective recommendation algorithms for movies, thereby leading to a largely inefficient

algorithm. Due to the inefficiency of some existing algorithms, there could be a loss in revenue by stakeholders, and inaccurate movie prediction mechanism. Hence, this research work is aimed at building a web application that will not only make

new movies more intriguing to the general public but present an accurate recommendation algorithm for movie prediction.



**CONTENTS**

Declaration i

Certificate ii

Acknowledgement iii

Abstract iv

1 Introduction 1

1.1 Objective and goal of the project . . . . . . . . . . . . . . . . . …….7

1.2 Problem Statement. …. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

1.3 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………...8

1.4 Challenges . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………....8

2 Literature Survey 9

3 Algorithms 10

Requirements Specification 11

3.1 Hardware Requirements . . . . . . . . . . . . . . . . . . . . . . . . . … 11

3.2 Software Requirements . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4 System Design 12

5 Implementation of System 13

6 Data Collection 14

7 Data Visualization 15

8 Model Development 18

9 Accuracy for algorithms used 19

10 Results & Discussion 21

11 Conclusion and Future Work 22

12 References 22

v

1. **Introduction:**

Recommendation systems use ratings that users have given to items to make specific recommendations. Companies that sell many products to many customers and permit these customers to rate their products, like Amazon, are able to collect massive datasets that can be used to predict what rating a particular user will give to a specific item. Items for which a high rating is predicted for a given user are then recommended to that user.

The same could be done for other items, as movies for instance in our case.

**1.1Objective and goal of the project**

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 .

**1.2Problem Statement**

In general, the more a movie is rated by users, the greater its average rating. However, this relationship is relatively weak.

Now, we will perform a deeper analysis of ratings by user. Similar to our analysis of ratings by movie, We will compare the average rating of users and determine whether the number of ratings they have given in total impact their average rating.

**1.3Motivation**

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 (Nielsen, 2006).

**1.4Challenges**

We could generally not face much potential risk with our strategy as it is based on the people's choice but it may harm us in some cases if the members of any particular movie try to take action on us if they are not convinced with our review.It may harm us physically or financially.

Different people may have different views and it varies by perspectives with respect to a particular movie review and some people may also try to either overhype or demolish the result in the Data collection strategy.

**2.Literature Survey:**

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. 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.

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. 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.

Rijul Dhir, Anand Raj in the paper(2018) titled, “Movie Success Prediction using Machine Learning Algorithms and their Comparison”. 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. 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. 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.

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 . 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**.**

**ALGORITHMS USED:**

**Regression 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.

**Logistic 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.

**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.

**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.

**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.

**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.

**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.

**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

**Stochastic Gradient Decent**

**Stochastic gradient descent** (often abbreviated **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** **Requirements Specification**

3.1 **Hardware Requirements**

There are four steps for preparing a machine-learning model:

Preprocessing input data

Training the model

Storing the trained model

Deployment of the model

**Base software:** ENVI 5.6.3 and the ENVI Deep Learning 2.0 module

**Operating systems:**

Windows 10 and 11 (Intel/AMD 64-bit)

Linux (Intel/AMD 64-bit, kernel 3.10.0 or higher, glibc 2.17 or higher)

**Hardware:**

NVIDIA graphics card with Capability version 3.5 to 8.6. A minimum of 8 GB of GPU memory is recommended for optimal performance, particularly when training deep learning models.

NVIDIA GPU driver version: Windows 461.33 or higher, Linux 460.32.03 or higher.

3.2 **Software Requirements**

We have developed a web application that can detect Whether a website is phishing or legitimate based on the URLs. For this web application, we are using FLASK a python framework for the backend implementation.

To train the models and comparative study Jupiter notebook is required.

Some of the packages or modules required for the machine learning training and testing are:

**NumPy** — For working with arrays, attributes are stored in the arrays.

**Pandas** — For reading/writing data.

**Scikt-learn** – For metrics evaluation and machine learning models.

Before importing the libraries first need to install all the packages listed in the above requirements.

**R:** For data visualization and data analysis

**4.System Design:**

* First, we will run an exploratory data analysis to get a general overview of the data and explore possible predictor variables.
* Then, we will include the most important features into numerous models i..e We are planning find the accuracy of the algorithms that we have used are logistic regression, random forest, decision tree, KNN, perceptron, support vector machines , naïve bayes, Linear SVC , stochastic Gradient Descent .Then, we will select the best model and apply it to the test data set (validation).

**MODULES:**

**DATA COLLECTION:**

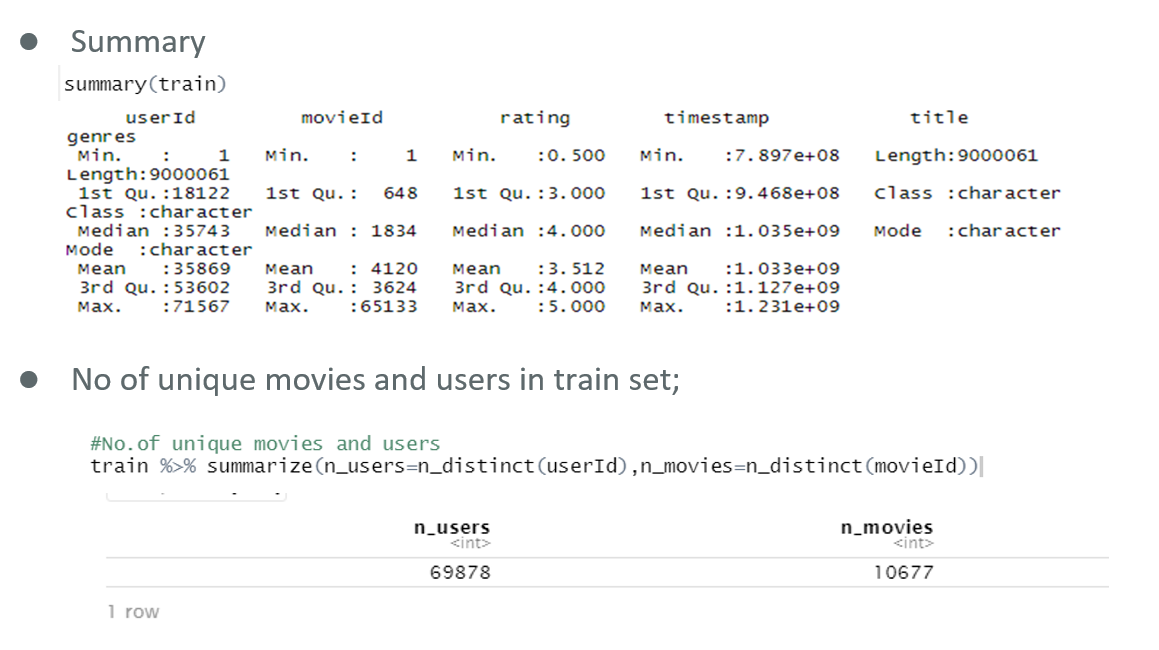
For the project,we used the 10M version of the MovieLens dataset which we collected from

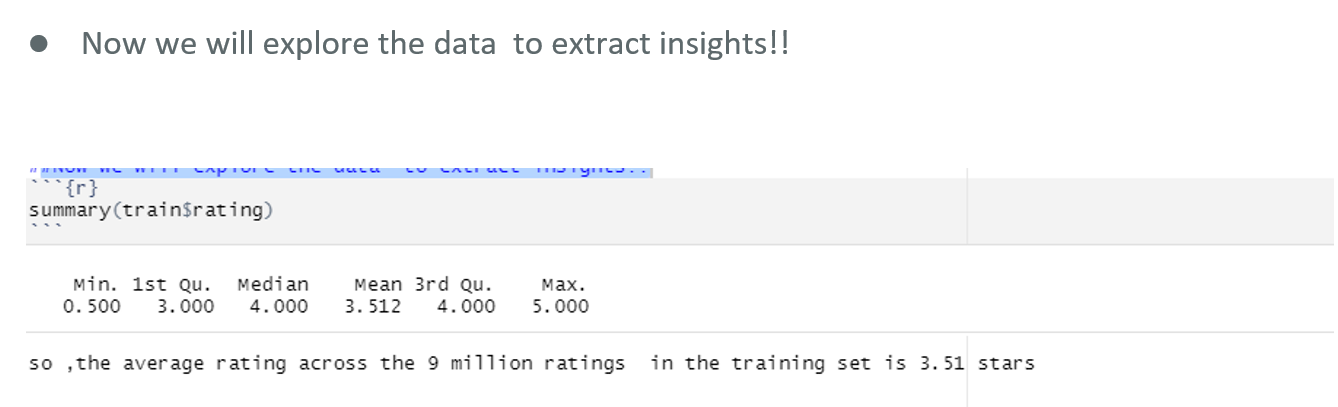
“https:// grouplens.org/datasets/movielens/10m/”.

The dataset presents information about 10 million movie ratings including user id, movie id, user rating of the movie (between 0.5 to 5 stars), timestamp of the rating (seconds since midnight Coordinated Universal Time of January 1, 1970), title of the movie, and movie genre(s): Action, Adventure, Animation, Children’s, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, IMAX, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and/or Western.

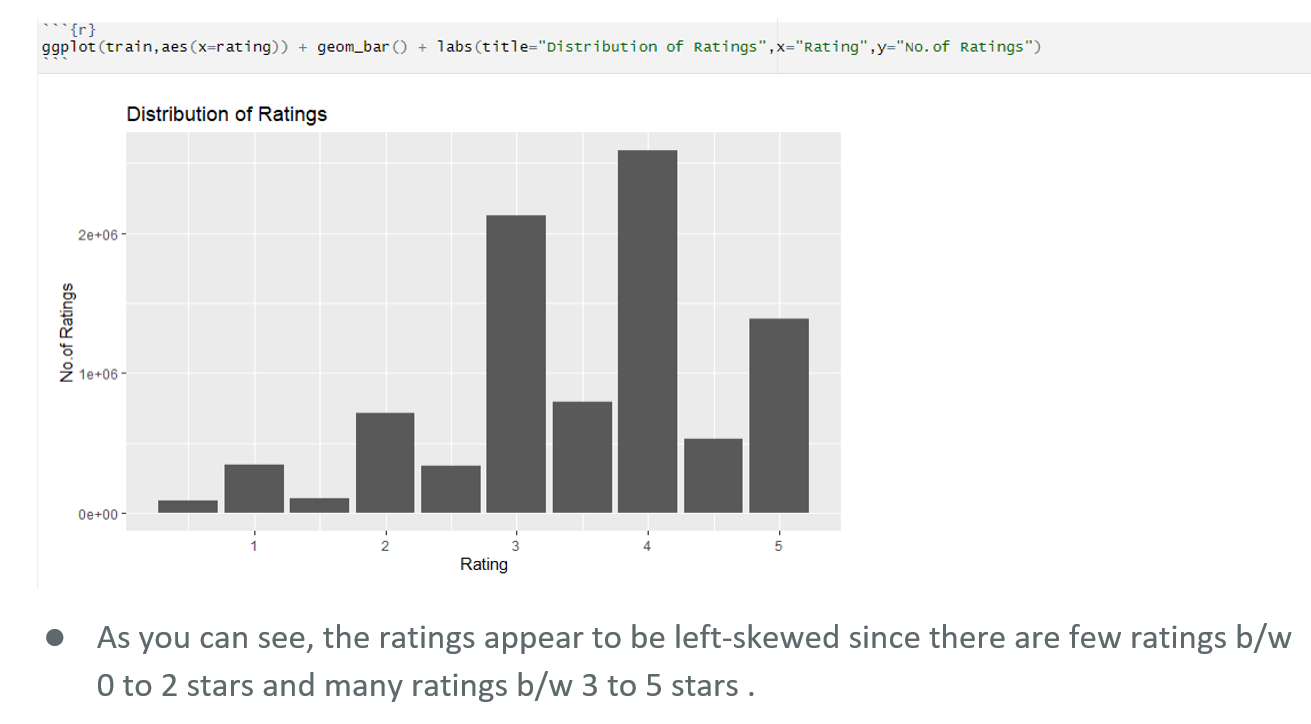


Exploring train set:

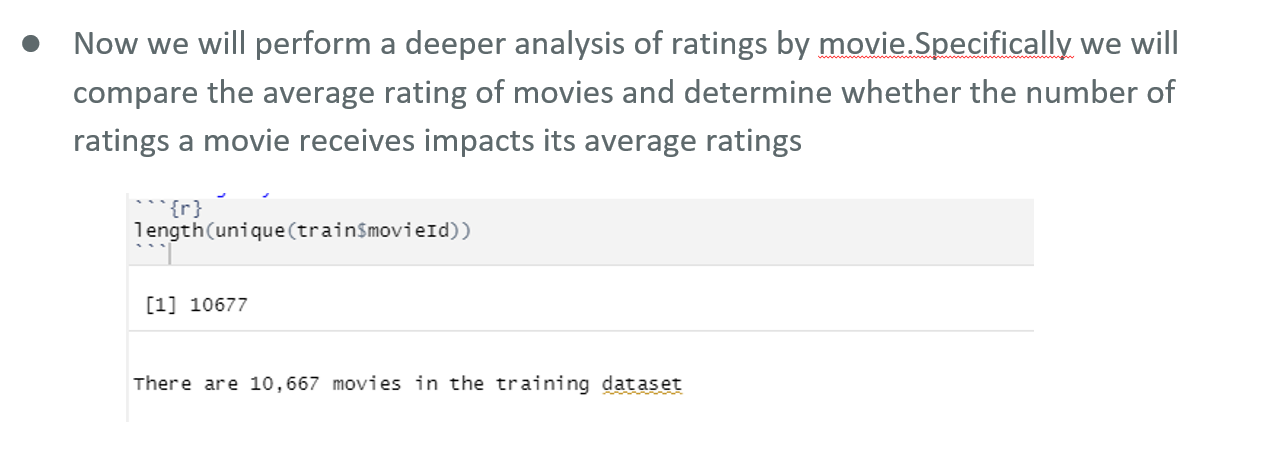
****

****

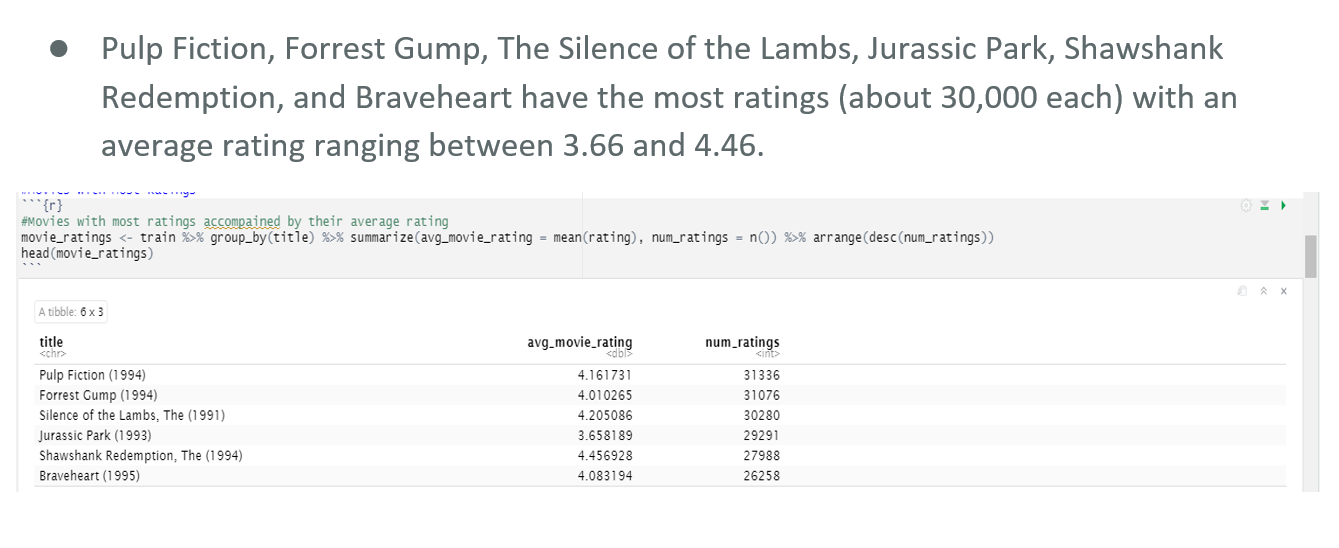
Data Visualization:

****

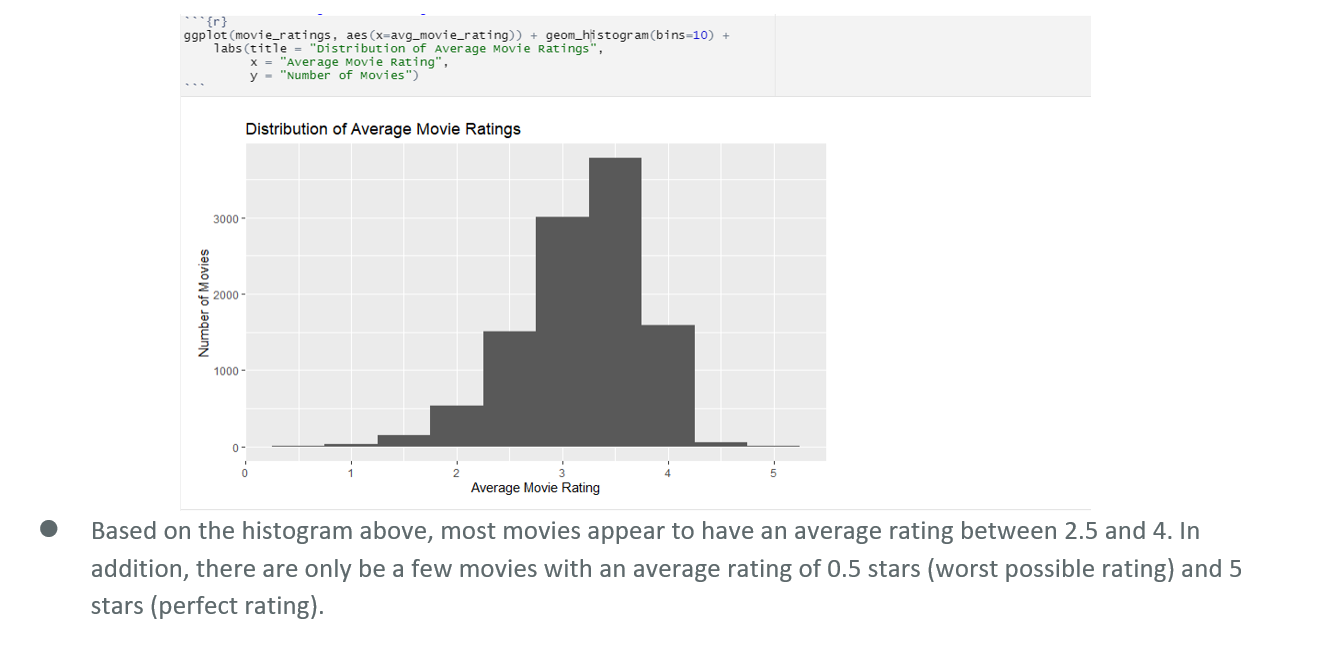
Ratings by Movie:

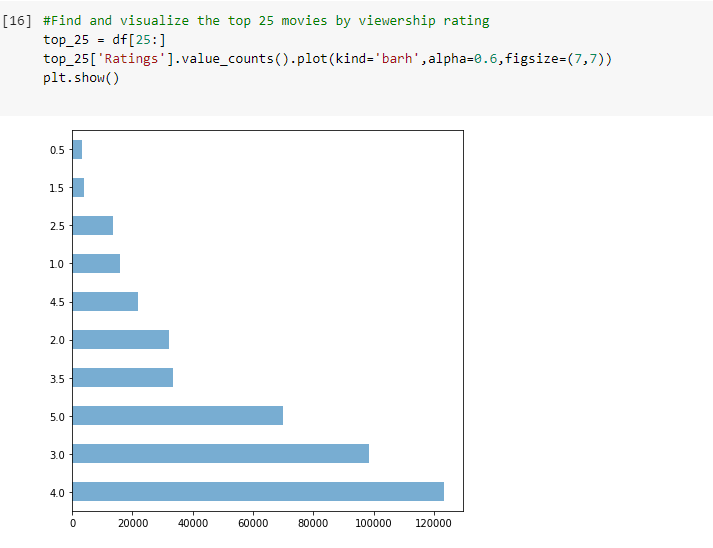


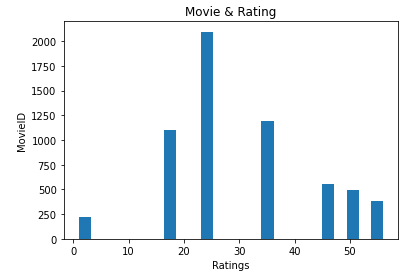
Movies with Most Ratings:



Most Common Average Movie Rating:







**MODEL DEVELOPMENT**

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

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

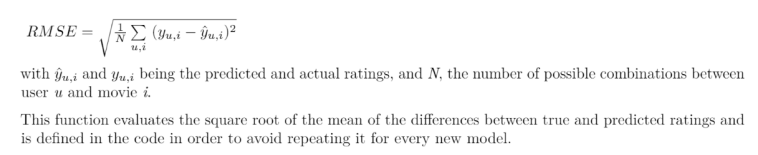
Model 2: Predicted Rating = Global Average Rating + User Effect

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

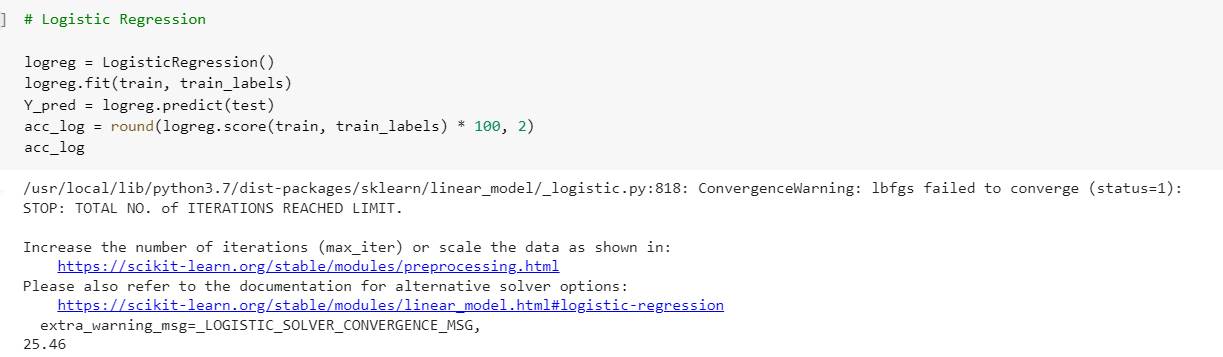
The global average rating is the average rating across all entries in the dataset. The movie effect is the difference between the average rating for the specific movie and the global average rating. Similarly, the user effect is the difference between the average rating for the specific user and the global average rating. To evaluate the three models, we will use RMSE (Root Mean Square Error). Ultimately, we will select the model with the lowest RMSE.

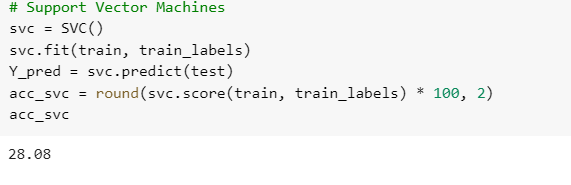
**RMSE:**

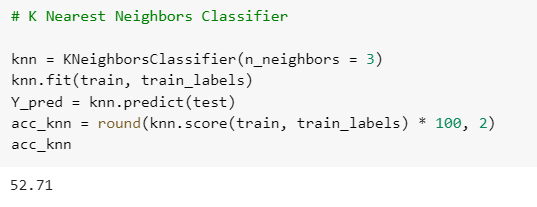
RMSE is used measure of the differences between values predicted by a model and the values observed. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular dataset, a lower RMSE is better than a higher one. The effect of each error on RMSE is proportional to 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.

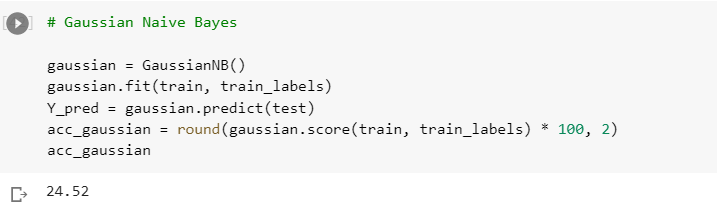
****

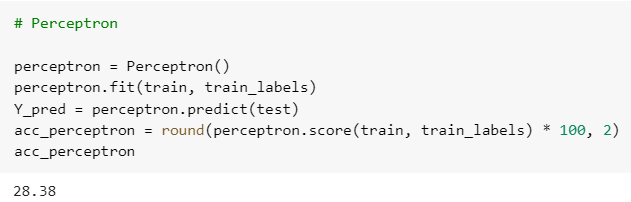
**Accuracy for the algorithms used:**

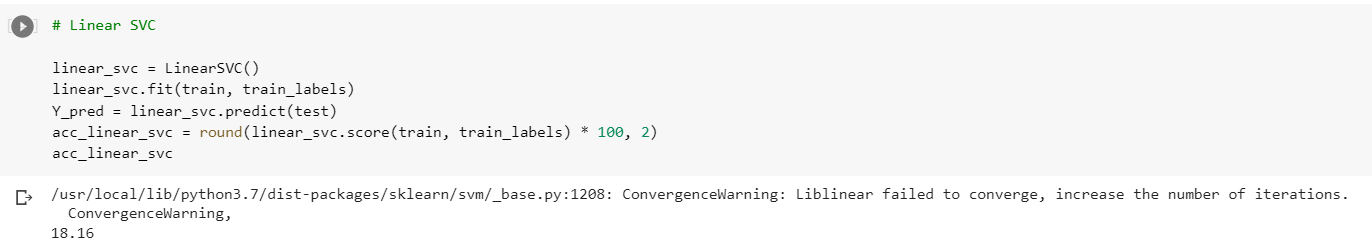
****

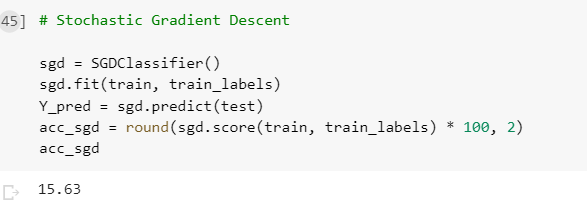
****

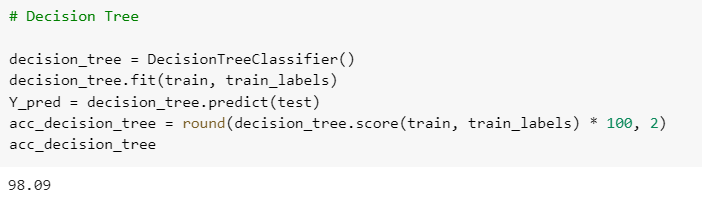
****

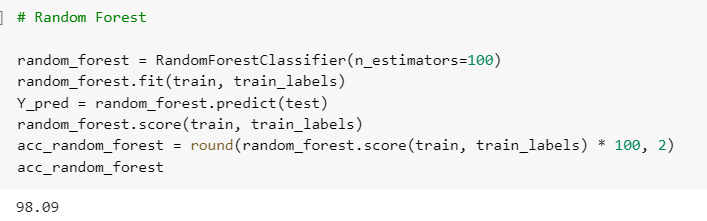
****

****

****

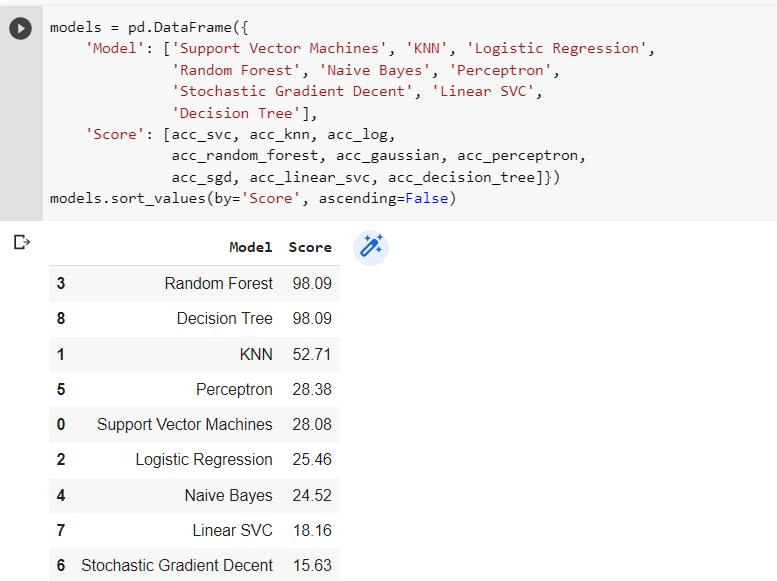
****

****

****

**RESULTS AND DISCUSSION:**

The result of the research 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 has very low accurate results.

****

**7.Conclusion and Future Work:**

* 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.

**8.REFERENCES**

<https://files.grouplens.org/datasets/movielens/ml-10m.zip>

* <https://mickteaching.wordpress.com/2016/04/19/data-need-to-be-normally-distributed-and-other-myths-of-linear-regression/>
* <https://stats.stackexchange.com/questions/148803/how-does-linear-regression-use-the-normal-distribution>
* <https://iovs.arvojournals.org/article.aspx?articleid=2128171>
* <https://ieeexplore.ieee.org/Xplore/home.jsp>
* https://ieeexplore.ieee.org/document/5277415
* <https://ieeexplore.ieee.org/document/9137994>
* https://ieeexplore.ieee.org/document/8944604

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. International Journal of Forecasting. 28, 161.

Marovic, M., Mihokovic, M., Miksa, M., Pribil, S., & Tus, A. (2011). Automatic movie ratings prediction using machine learning. Convention on Information & Communication Technology Electronics & Microelectronics.

Meinhausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research. 7, 983-999.

Meyer, D., Leisch, F., & Honk, K. (2003). The support vector machine under test. Neurocomputing. 55, 169-186.

Olson, D.L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems. 52, 464-473.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems. 14, 1-37.

Yu, H., & Kim, S. (2012). SVM Tutorial — Classification, regression and ranking.