International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Big Five Factor Based Movie Rate Prediction Using Machine Learning Techniques

Shetty Bhuvaneshwari[1*], Preethi[2], Anisha P. Rodrigues[3], Roshan Fernandes[4]
[1]*Lecturer, Department of Computer Science and Engineering, Government Polytechnic for Women, Mangalore, India*
[2]*Student, Department of Computer Science and Engineering, N.M.A.M. Institute of Technology, Nitte, India*
[3]*Assistant Professor, Department of Computer Science and Engineering, N.M.A.M. Institute of Technology, Nitte, India*
[4]*Associate Professor, Department of Computer Science and Engineering, N.M.A.M. Institute of Technology, Nitte, India*

***Abstract*: Predicting ratings make up a big part of our cultural environment, but little research has been done on what such films indicate about our personalities. Using the Big Five-Factor model of personality as a guide, we set out to see if there were any links between film and movie interests and individual personalities. Ratings-based recommender systems may fail to provide ideal levels of diversity, popularity, and serendipity for their users because the type of movie one watches is tied to one's personality. Individual user's preferences in recommendation lists for diversity, popularity, and serendipity cannot be inferred just from their ratings. When we incorporate user's personality qualities into the process of creating recommendations, we can boost user satisfaction. The proposed model can be used to recommend movies to users. In this paper, the dataset having 1834 users with 12 different movie ratings. Then the data set is used to analyze the movie rating based on the individual user personality. The Big Five Factor model algorithm determines an individual's personality. The experimental result is then analyzed, and it shows that the proposed methodology calculates movie ratings more accurately. Our trials on predicting user enjoyment of movie lists were conducted using a variety of machine learning methods, with LGBM providing the lowest Mean Absolute Error.**

***Keywords*: Big five-factor model, personality prediction, movie rating, machine learning.**

1. Introduction

Since the turn of the twentieth century, movies have been a significant part of world culture, and with the advent of the digital age, sharing and consumption of cinema has never been easier. Everyone has a favorite movie, but what do those choices reveal about us? The concept makes intuitive sense. Even if we don't know anything else about someone, we will undoubtedly create a set of conclusions about them based on their favorite movie and favorite character. Are these types of assumptions unfair? Despite the movies' cultural and economic influence, there is a scarcity of research on the psychological effects of our movies and media preferences. The ability to predict someone's personality based on their favorite films may be unsettling to some, but it brings us closer to solving the age-old question of whether we are what we consume.

A modest amount of research has attempted to determine the nature of the relationship between users and their film or media preferences. The Big Five personality model, sometimes known as the five-factor personality model, is a potentially helpful framework for approaching this relationship. The five-factor model consists of extraversion, agreeableness, conscientiousness, neuroticism (emotional stability), and openness to experience. Energy, excitement, talkativeness, and a desire to be in the company of others characterize extraversion. High agreeableness scores are associated with compassion, cooperation, and the ability to help, rather than suspicion and distrust of others.

Conscientious people are more organized, trustworthy, and thoughtful than spontaneous or unorganized people. Neuroticism is a personality trait that measures a person's emotional consistency. People who score high in this category are impulsive and regularly feel negative feelings like anger, anxiety, and depression. Individuals with high openness scores are more likely to like art, adventure, and a wide range of experiences, which is the final factor to examine. Individuals that are inventive, educated, and self-reliant are more likely to be open. A composite model of a person's personality emerges when all of these qualities are acknowledged.

The present study takes a more in-depth look at the elements that influence film preferences. We can investigate the preferred movie by concerning all five dimensions of the five-factor model of film preferences and personality. When given a list of films, each individual displayed different choices, even though the movie rating was identical. If people differ in their Big Five traits in addition to their film interests, it's reasonable to infer that they differ in how those qualities and tastes interact. Using these characteristics instead of filling out the questionnaire, it is possible to evaluate someone's personality by looking at their preferences. Because we regularly use and share our film and media choices, which are publicly accessible on social networks, such media-based personality models have become beneficial in the digital era. The media we consume may impact our personality and value development, making a model that combines the two particularly robust. The current study's purpose is to delve deeper into this strategy and build a model based solely on one sort of media: movies. While the goal of this study is to investigate whether individual

*Corresponding author: bhuvana.shetty@gmail.com

personality factors will be connected to film genre preferences. Furthermore, we predict that each personality attribute score will be unique. A movie can belong to any of the genres based on motion picture categories, such as action, comedy, science fiction, biographies, and horror. Different people enjoy different kinds of movies: some like films that make them think, while others choose films that make them laugh. We observed that reviews and ratings may not be enough to predict a person's movie preference because they mainly reflect the person's observations about the context and narrative of the films. As a result, these evaluations don't pay attention to users' every- day choices, psychological states, or affective and cognitive mechanisms. For our study, we created a dataset of 1834 users. Aside from their personality, each user's choices for a group of 12 movies are evaluated using serendipity, popularity, diversity value, or none of the above (default option). Based on the specified measure and condition, the users provided a list of movies to watch (high, medium, low). For example, if serendipity is the assigned metric and the condition is high, the films on the list are firmly serendipitous.

The Big Five Factors are measured as follows:

Openness: It is a rating scale (from 1 to 7) that measures a user's willingness to try new things. 1 indicates that the user does not prefer new experiences, whereas 7 indicates that the user does.

Agreeableness: It is a rating scale (from 1 to 7) that assesses a user's proclivity to be empathetic and cooperative toward others rather than distrustful and combative. 1 indicates that the user does not have compassion or the nature of cooperation. 7 shows that the user is sympathetic and cooperative.

Neuroticism: It is an assessment score (from 1 to 7) assessing a user's tendency to have psychological stress. 1 means the user has the tendency to have psychological stress, and 7 means the user has the tendency not to have psychological stress.

Conscientiousness: It is a rating scale (from 1 to 7) that assesses a user's ability to be orderly, dependable, and self-disciplined. 1 indicates that the user has no such propensity, whereas 7 depicts that the user does.

Extroversion: It is a scale (from 1 to 7) that measures a person's predisposition to be outgoing. 1 indicates that the user has no such propensity, whereas 7 indicates that the user does.

2. Related Work

Lukito, Louis Christy, et al. [1] have suggested that using the fame of social media it has become possible to predict a person's behaviors and personality traits. He has explored Twitter as an open vocabulary prediction data source in Indonesia. A Naive Bayes classifier is used. A statistical model has been made to classify the personality of an individual with his Twitter username and gender as an input. Pre-processing data includes steps such as removal of retweets, deletion of repeated letters, and emotional symbols.

S. V. Kedar [2] proposed that the automatic personality evaluation should consist of the automatic classification of data of a person such as a video, speech, or text. The latest state-of-the-art for assessment of personality. Despite enormous work in this area, the current state-of-the-art is difficult to establish. A psychological personality evaluation explains an inside person. If you or someone else needs to describe yourself, you are using lots of adjectives and descriptions. Models based on characteristics map a single adjective that can be described as behavior.

Michael M. Tadesse [3] presents a framework for designing and implementing SNA as well as linguistic characteristics like Linguistic Inquiry and Word Count (LIWC) and Structured Programming for Linguistic Cue Extraction (SPLICE). The effects of a user's social interaction behavior on personality were investigated. This study uses Facebook profile data to calculate personality scores. For example, the data suggest that agreeableness is favorably connected to the number of tags, whereas neuroticism is negatively connected to the number of friends. During the pre-processing tokenization is done using open NLP to separate the last word of each sentence with punctuations. Names, spaces, URLs, symbols, etc., are removed. LIWC dictionary is widely used in psychological studies to extract certain linguistic features. SPLICE is a newer dictionary with updated processes. SNA technique analyses the socializing behavior which originates from the relationships among members.

K. Maheshwari [4] proposes a method to predict customer's behaviors in online shopping using SVM Classifier. The character may have motivation, quality, perceptions, employment and revenue, attitude, culture, and social beliefs. Some customers are attached to a product based on a friend's recommendations. Few people browse the data of the product but refuse to buy. Younger generations will shop more online. So, age-based categories are also possible. Data mining can be used to investigate a customer's activities on shopping using various methods and algorithms. Every activity of a person is stored as a byte of data in a database which is used to collect information such as how often they buy a product, or how much they spend their valuable time in buying decisions. Items bought and their quantity is also considered. Here customers are categorized based on their behavioral characteristics on shopping activities. The classification by using a multidimensional hyperplane. Frequent customers are identified. Which product is sold more on which date can also be achieved? Here dataset is pre-processed using data mining techniques.

Manish Mishra [5] describes the characteristics and applications of Artificial Neural networks or ANN. In the development of smart systems based on biological neural networks, many developments have been made. It's more like a web of interrelated neurons. A neuron is a cell that processes data into a different neuron. It consists of an input layer, an output layer, and a hidden layer or layers. Many ANN systems do not describe how they solve problems since the hidden layer of the black box implementation. Each system must be smart to solve problems based on inputs in the world of today. Applications in airline security, data validation, target marketing, customer research, etc. may be performed. These applications may develop the ability such as backpropagation, perception learning, etc.

Roopa M., [6] describes ANN using a backpropagation

algorithm that is distributed MANETs. A MANET is a network with many dynamic nodes and less infrastructure. The theory proposed is that multi-layer neural perceptron networks can use a clustering technique. Clustering is a virtual process in which nodes with similar properties are grouped. To initialize the input parameters, they should include the speed of the node and the number of clusters. Mean Square Error (MSE) and epoch should be initialized to know the number of iterations. Weights and a bias should be initialized too. Some sigmoid functions should be chosen as an activation functions. Once the output value is calculated it is compared with the target values. Hence the mobility can be calculated and the weights of each node. MSE is obtained and epoch is calculated. Weights are then modified along with the bias. The networks are trained with data gathered at the hidden layer. The backpropagation algorithm helps the network to minimize the error rate.

Di Xue et al. [8] proposed AttRCNN-CNN to recognize the Big Five personality traits from each online user's text post in 2018. They used an AttRCNN structure to collect deep semantic information from sentences, as well as a CNN-based inspection structure to extract document vectors. The Authors tested the idea of combining deep semantic features with statistical linguistic data extracted directly from text posts, then feed them into standard regression algorithms to predict real-valued Big Five personality ratings.

Authors [9] Cold start problem, long-tail problem, sparsity, shared account problem, grey sheep problem, scalability, and other issues are common in recommendation systems. The “collaborative filtering” technique was applied, and the “Pearson correlation coefficient” was chosen as the similarity measure. Movie-Lens-100k is the dataset under consideration. The results of this experiment reveal that low-rated films have little bearing on movie predictions. As a result, it's best to ignore them while making movie predictions.

Authors [10] combined feature extraction and classification in CNN as a cooperative job based on this fundamental. According to the study, scientists used the CNN algorithm to classify personality traits in 250 users out of 9917 Facebook status updates by applying deep semantic features and LWIC-derived features. The openness characteristic has the highest categorization accuracy of 0.76. This study found that combining CNN with classical language features produces excellent results.

There is very little research work carried out on the Big Five Factors-based movie rate predictions. Hence the proposed work implemented the various machine learning algorithms, namely, Support Vector Machine, Random Forest, LGBM Classifier, Naive Bayes, Passive-Aggressive Classifier, Logistic Regression, XGB Regressor, and Linear Regression to predict the movie ratings based on the Big Five Factors. The results are analyzed and compared for each classifier by considering the Mean Absolute Error (MAE).

3. Methodology

As seen in figure 1, enjoyed movie list prediction is performed by knowing the user’s personality and type of list. In this work, the main stage is to train the various classifiers and find mean absolute error values.

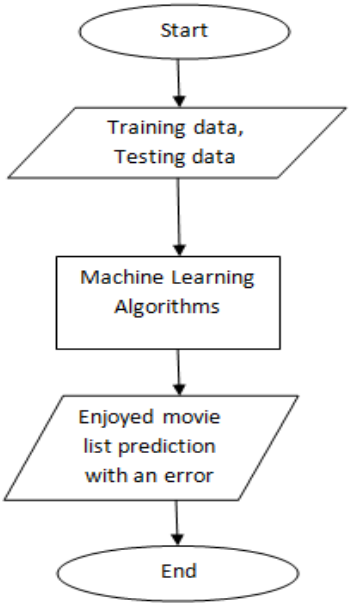

Fig. 1. Prediction of enjoyed movie list flow chart

A. Data Collection

This section gives an overview of the data collection and classification techniques which are we used for developing a predictive model to recommend the movies to the users. In our project we used 1834 rows with 34 columns is about users. This dataset is formed by measurements of five personality traits of the users. There are 12 movies on the list (where x is a number between 1 and 12). These fields include the ids of the twelve movies in the list. Predicted rating x (x is a number between 1 and 12) is the user's estimated rating for the accompanying movie x. Enjoy watching is the user's response to the question, "This list covers movies I believe I enjoyed seeing." Users were asked to respond on a 5-point Likert scale. (Strongly Disagree: 1, Strongly Agree: 5)

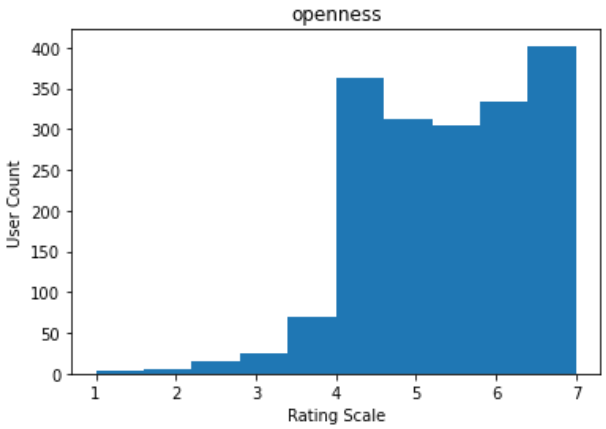

Fig. 2. Distribution of openness users

Fig. 3. Distribution of agreeableness users

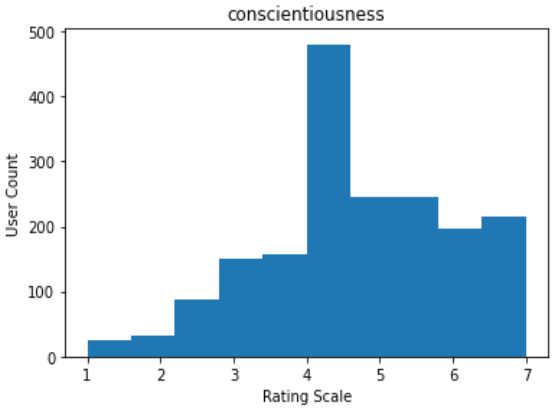

Fig. 4. Distribution of conscientiousness users

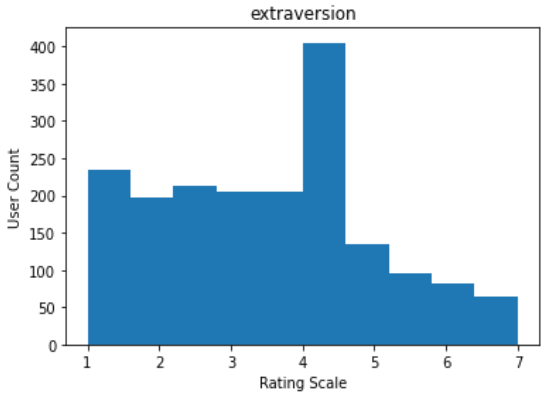

Fig. 5. Distribution of extraversion users

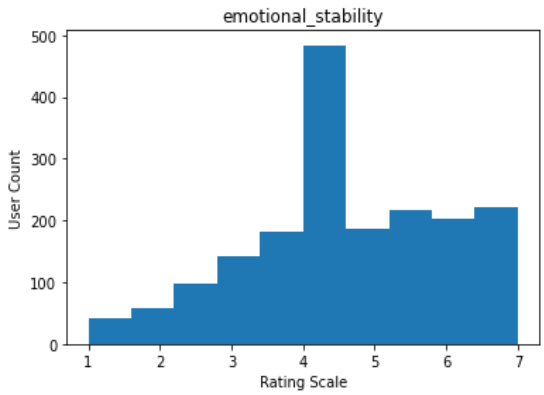

Fig. 6. Distribution of emotional stability users

Figure 2 through 6 shows the distribution of Openness Users, Conscientiousness Users, Extraversion Users, and Emotional Stability Users in the data set used for experimental purposes.

The analysis shows that most of the distributions look similar to normal distribution but the openness is more negative unbalanced, which represents the people who score high in openness are more tend to like arts. They most prefer to watch movies which are in the form of art.

B. Machine Learning Algorithms

On the dataset, various machine learning algorithms were utilized, and the error rate was computed using the models. The lesser the error in technique shows how much the user enjoyed the movie list.

1) Support Vector Classifier

It is an administered AI model, for two gathering characterization issues that utilize grouping calculations. At the point when the marked preparing information for every classification has been given to the SVC model, the model can order the new content. This order calculation functions admirably in restricted measures of information, this is additionally considered one of the quick and reliable characterization calculations. SVC arrangement calculation gives better precision on our dataset contrasted with different calculations.

2) Random Forest Classifier

It is an administered learning calculation. The "forest" assembles a gathering of choice trees is normally prepared with packing technique. The blend of learning models utilizing packing techniques builds the general execution result. The irregular woodland is utilized for both characterization and relapse issues is one of the enormous benefits, that shapes most of current AI frameworks. In Irregular backwoods, the extra haphazardness is added to the model, while developing the tree. This calculation consistently looks for the best element among an arbitrary subset of highlights. On this present reality dataset, these calculations give a better blunder rate contrasted with different calculations.

3) LGBM Classifier

Light Angle Boosting Machine classifier depends on the choice tree, it builds the proficiency of the model and diminishes the utilization of the memory. Essentially it utilizes the two procedures Angle based one-side examining and elite component packaging which finishes the impediments of histogram-based calculation. This calculation parts the tree leaf-wise whereas other boosting calculations develop tree-level insight. Normally leaf shrewd calculation has lower misfortune contrasted with the level astute calculation yet it expands the intricacy of the model and here and there overfitting in little datasets. On account of the more modest size of the dataset, LGBM shows more mistakes.

4) Naive Bayes Algorithms

The Naïve Bayes algorithms are quite simple to design and provides significant benefits in many complex real-world situations. The Naive Bayes algorithms are in different flavors such as Multinomial, Bernoulli, and Gaussian Naive Bayes classifiers. In our project, we used two Naive Bayes Classifiers

those are namely Multinomial naive Bayes classifier and Gaussian naive Bayes classifier. The Multinomial naive Bayes works well on our datasets, we got a lesser mean absolute error than the Gaussian naive Bayes.

5) Logistic Regression and Linear Regression

The Logistic Regression and Linear Regression use an equation as the representation. Logistic regression is usually used to solve classification problems and to predict the categorical whereas linear is used to solve the regression problems, for predicting the continuous dependent variable using the set of independent features. In our project, Logistic Regression provides a better result than Linear Regression.

6) Passive-Aggressive Classifier

This is one of the online learning algorithms, generally used for large-scale learning. When this algorithm is used to make the model for a given dataset, this will respond as passive or aggressive. Passive is when the prediction is correct and Aggressive is when the prediction is incorrect. In our dataset, the Passive-Aggressive Classifier provides more MAE.

7) XGB Regressor

The Extreme Gradient Boost classifier is a decision tree-based ensemble algorithm of machine learning. This algorithm uses the texts and images as input, which is of unstructured data format. In our work, we used this classifier on texts. As per the result, using this classifier we got a medium MAE value compared to others.

4. Results and Discussion

Finding a correlation allows us to figure out which personality type is more likely to provide higher ratings to recommended lists. Table 1 gives the correlation between Personality and Enjoy Watching attributes. The values show that these two attributes have a very low correlation value.

Table 1
Correlation between personality and enjoy watching

Personality	Enjoy watching
agreeableness	0.0368
openness	0.064
extraversion	0.027
emotional stability	-0.0022
conscientiousness	-0.044

Table 2
MAE obtained for various machine learning models

Classifier	MAE
LGBM	0.7864
Support Vector Classifier (SVC)	0.7886
Multinomial Naïve Bayes (MNB)	0.7886
Logistic Regression (LR)	0.7886
Gaussian Naïve Bayes (GNB)	0.7952
Random Forest Classifier (RF)	0.8562
Linear Regression (LR)	0.8600
XGB Regressor (XGBR)	0.8719
Passive Aggressive (PA)	0.9520

We estimated the mean absolute error to investigate the correlation between user enjoyment of the movie list and personality.

The sum of the absolute discrepancies between each prediction and its corresponding rating, divided by the number of ratings, is the Mean Absolute Error (MAE). We repeated the procedure for each method for each user in the test group to estimate how much they loved the movie list with an error.

In the proposed work, we used multiple classifiers, to see the performance on the personality prediction dataset to recommend the movie on the ratings given by the users. Table 2 shows the predicted values from the various classifier, by seeing this we can predict how much they enjoyed the movie list with an error for each classifier. The LGBM classifier gives less error while predicting on the personality dataset. Different types of a classifier are used to determine which algorithm works well in predicting. As already discussed in the previous section Passive Aggressive Classifier shows a greater error rate that means the movie list enjoyed by the user is less.

.

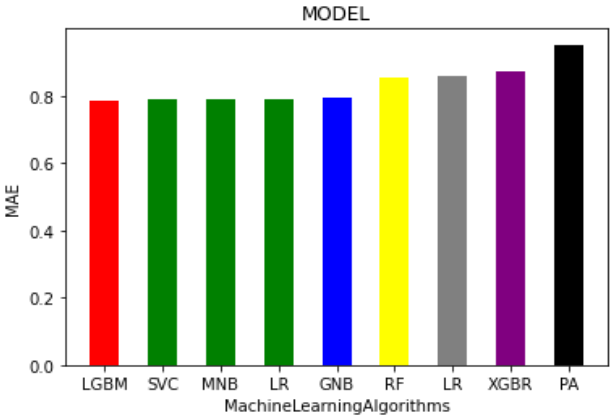

Fig. 6. Classifiers with error rate

5. Conclusion and Future Work

Our current research stumbled onto several forms of machine learning models that were capable of offering the error rate on how much the user enjoyed the movie, and from these ratings can successfully propose movies based on their preferences. The Support Vector Classifier, Multinomial Nave Bayes, and Logistic Regression methods we utilized to create the model have a lower error rate than the others, hence it is considered one of the best-suited classifiers for predicting personality on movie lists. As a result, we can readily classify the type of personality a person has, as well as the types of movies they prefer. In the future, more amount of data set values may be analyzed using deep learning models.

References

[1] Louis Christy Lukito, "Social Media User Personality Classification using Computational Linguistic", 2016.
[2] S.V. Kedar, D. S. Bormane, "Automatic Personality Assessment: A Systematic Review", 2015.
[3] Michael M. Tadesse, Honfei Lin, Bo Xu, and Liang Yang, "Personality Predictions Based on User Behaviour on the Facebook Social Media Platform", 2016.
[4] K. Maheswari, "Predicting Customer Behavior in Online Shopping Using SVM Classifier", 2017.
[5] Manish Mishra, "A View of Artificial Neural Network", 2014.
[6] Roopa M, S. Selva Kumar Raja, "Artificial Neural Network Using Back Propagation Algorithm Using Distributed MANETs", 2016.
[7] Reem Alnanih, Nadia Bahatheg, Melad Alamri, Rana Algizani, "Mobile-D Approach-based Persona for Designing User Interface", 2019.

[8] D. Xue, L Wu, Z Hong, S. Guo, L Gao, Z. Wu, X. Zhong, and J. Sun, "Deep learning-based personality recognition from text posts of online social networks." Applied Intelligence 48, no. 11 (2018): 4232-4246.

[9] M. M. Reddy, R. Sujithra, B. Surendiran, "Analysis of Movie Recommendation Systems; with and without considering the low rated movies," 2020 International Conference on Emerging Trends in Information Technology and Engineering, 2020.

[10] C. Yuan, J. Wu, H. Li, and L. Wang, "Personality Recognition Based on User Generated Content," in 2018 15th International Conference on Service Systems and Service Management (ICSSSM), pp. 1-6, 2018.