CineRater: A Machine Learning Strategy for Predicting Movie Ratings

Submitted by:

Kamalhasan B   a.narexxxxx@gmail.com
Narendra A   kamxxxx@gmail.com

1   Introduction and Project background

In today's world, where movies are a major source of entertainment for people, it has become increasingly important to develop accurate methods for predicting movie ratings. Selecting which movies to watch can be a challenging task for movie enthusiasts, given the large number of movies released every week. Although there are many resources available to help make informed decisions, such as movie trailers, reviews, and recommendations, these sources can be subjective and not always reliable. Movie ratings are an important factor to consider, but they can also be highly variable and dependent on individual preferences.

The aim of this project is to address this issue by using machine learning to develop a movie rating prediction system that is both accurate and trustworthy. By analyzing a comprehensive dataset of movie ratings and applying a range of machine learning algorithms, a model will be created that can anticipate the ratings of upcoming movies.

The goal of this project is to provide a model for movie enthusiasts that will help them make informed decisions about which movies to watch. With an accurate movie rating prediction system, viewers can save time and money by only watching movies that are likely to meet their expectations. This approach has the potential to transform the way people choose which movies to watch and could have a significant impact on the film industry.

2   Problem statement

The motive of this project is to develop a machine learning-based rating prediction system for movies that is both accurate and reliable. Leveraging a comprehensive dataset of movie ratings and applying a range of machine learning algorithms to construct a model that can anticipate the ratings of upcoming movies. By developing an accurate and trustworthy movie rating prediction system, this work aims to make informed decisions for movie enthusiasts. This approach has the potential to revolutionize the way people choose which movies to watch and could have a significant impact on the film industry.

The project addresses the problem of subjective and unreliable movie ratings that make it difficult for movie enthusiasts to decide. By developing an accurate and reliable movie rating prediction system, it helps movie enthusiasts to save time and money by only watching movies that are likely to meet their expectations.

3   Literature review

In the paper [3] author proposed a MCGF-SVD model to perform the task of movie rating prediction. Initially, the MCBF-SVD enhances the filtering approach based on movie categories and discusses the impact of movie categories on forecasting users future rating behavior using weighting factors. Lastly, the SVD method, which performs well in Colborative Filtering, is merged with MCBF. This significantly improved the accuracy of rating prediction compared to numerous other current algorithms. The RMSE and MAE of the model[3]were recorded at 0.651 and 0.5767 respectively.

In the study, [4] a customized prediction model DeepFM is developed for the prediction. This model is developed to reflect better the low and high-order feature interaction of the data. The model in this paper[4] uses a combination of machine learning and deep learning techniques such as Word Embedding, Text vectorization, Long-Short-Term-Memory Networks, Max Pooling layers, Dense

Layers, and a few other processing techniques. Finally, Root Mean Squared Error of the model is noted at 0.9782.

The authors in the paper[2] investigated the accuracy of machine learning and deep learning models in the context of movie rating prediction. The author also investigated the use of different word embedding methods such as CBOW and skip-gram to finally compare the model accuracy. Machine learning and deep learning models such as Naive Bayes, Recurrent Neural Network, Long-Short-Term-Network, and Gated Recurrent Unit techniques were employed and investigated. According to the authors[2], the Naive Bayes and LSTM models were concluded as better prediction models in the case.

In summary, the literature on movie rating prediction has demonstrated the potential for machine learning algorithms to accurately predict movie ratings. However, the variability and subjectivity of movie ratings suggest that a comprehensive approach that considers a range of factors is necessary to develop an accurate and reliable movie rating prediction system.

4   Objective

The objective of this study is to develop a reliable and accurate movie rating prediction system. This will involve collecting and analyzing a comprehensive dataset, applying a range of machine learning algorithms to develop a prediction model, evaluating the model's accuracy and usability, and comparing it against existing systems to assess its potential impact on the film industry and movie-watching experience. The study aims to provide movie enthusiasts with a reliable model for making informed decisions about which movies to watch.

5   Research design and Methodology

The research design for this study is a mixed-methods approach that combines quantitative data analysis with a user study. Exploratory data analysis is performed on the dataset to provide a better intuition of the data. The quantitative data analysis involves collecting a comprehensive dataset of movie ratings and related information, including movie genres, runtime, budget, release date, and critical reception. The dataset will be sourced from publicly available data repositories IMDb through web scraping techniques based on the requirement. The dataset will be cleaned and pre-processed to ensure that the data is accurate and consistent. Data analysis methods such as missing values imputation, outlier detection, and clustering are performed during the project implementation.

A machine learning technique will be used to create a prediction model for movie ratings. Scikit-learn, a Python-based machine learning library, was used to create the models. The performance of the developed model will be evaluated against a set of validation data to determine its accuracy and reliability. To prevent the model from being overfitting to the training data, the validation data will be kept apart from the training data. Proper research will be carried out to use the methods of Natual language processing, and sentiment analysis of the movie summaries based on the requirement to increase the model performance.

6   Conceptual Framework

The conceptual framework for this study is based on the principles of machine learning and predictive analytics. The study will leverage existing research on machine learning algorithms and apply them to the domain of movie rating prediction. The developed rating prediction system will be grounded in the relevant theory and literature on movie ratings and related factors, including movie genres, runtime, budget, release date, and critical reception. The framework will also incorporate user-centered design principles to ensure that the developed system is intuitive and user-friendly.

We will use a range of machine learning algorithms to develop the prediction model, including regression models, decision trees, and ensemble methods. Regression models will be used to predict movie ratings based on a range of factors, including movie genres, runtime, budget, release date, and critical reception. Decision trees will be employed to pinpoint the key elements that affect movie ratings. Ensemble methods, such as random forests and gradient boosting, will be used to combine the strengths of multiple algorithms to improve the accuracy of the prediction model.

7   Dataset

For this study, the IMDb dataset[1] will be used, which is a comprehensive collection of movie data sourced from IMDb.com. The dataset includes a wide range of information about movies, such as movie titles, genres, release dates, production companies, cast and crew members, ratings, and reviews.

The dataset will be preprocessed to ensure that the data is accurate and consistent, as well as to extract the necessary features for our machine learning model. This work includes removing any duplicates or irrelevant data points, and also convert categorical data into numerical data using appropriate encoding techniques.

The IMDb dataset is a valuable resource for our study, as it provides a comprehensive collection of movie data that can be used to train and evaluate our machine learning models. It's worth noting that the IMDb dataset is a publicly available dataset that is often used in research and industry applications. As such, we will make sure to adhere to ethical data usage practices and follow the terms of service for IMDb.com.

8   Software requirements

This project will be implemented using python programming in the google colab environment and uses the following scientific stack of modules Tensorflow, Keras, Opencv, Numpy, Matplotlib and few other modules based on the implementation requirement.

9   Conclusion

In conclusion, this project aims to develop a reliable and accurate movie rating prediction system using machine learning algorithms. The study will collect and analyze a comprehensive dataset of movie ratings and related information, apply a range of machine learning algorithms to develop a prediction model, evaluate the model's accuracy and usability, and compare it against existing systems to assess its potential impact on the film industry and movie-watching experience. The study has the potential to revolutionize the way people choose which movies to watch and could have a significant impact on the film industry.

The developed rating prediction system could provide movie enthusiasts with a reliable tool for making informed decisions about which movies to watch and enable movie producers to better understand audience preferences and create more successful movies. The study's findings could also contribute to the broader field of predictive analytics and machine learning, demonstrating the applicability of these techniques to a range of domains beyond traditional business applications.

References

[1] Imdb dataset. https://www.imdb.com/interfaces/.

[2] Saikiran Gogineni and Anjusha Pimpalshende. Predicting imdb movie rating using deep learning. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1139–1144, 2020.

[3] Xiaoyue Li, Haonan Zhao, Zhuo Wang, and Zhezhou Yu. Research on movie rating prediction algorithms. In *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, pages 121–125, 2020.

[4] Dong Uk Won and Hwa Sung Kim. A prediction scheme for movie preference rating based on deepfm model. In *2022 International Conference on Information Networking (ICOIN)*, pages 385–390, 2022.