Performance Prediction for IMDB Movies

Muhammad Sudais (sudaismsm@gmail.com)
FAST - National University of Computer and Emerging Sciences https://orcid.org/0000-0002-2209-8104

Mohammad Hasan Khan
FAST - National University of Computer and Emerging Sciences

Abdul Jabbar Tabani
FAST - National University of Computer and Emerging Sciences

Method Article

Keywords: Machine Learning, Data Science, Predictions, Movies

Posted Date: January 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1243202/v1

Performance Prediction for IMDB Movies

Muhammad Sudais
National University of Computer and Emerging Science
Karachi, Pakistan
sudaismsm@gmail.com

Mohammad Hasan Khan
National University of Computer and Emerging Science
Karachi, Pakistan

Abdul Jabbar Tabani
National University of Computer and Emerging Science
Karachi, Pakistan

Abstract

Filmmakers and other associated people in the fraternity are very much concerned about the performance of their movies on box office. They pay a lot of hard work and invest a big fat amount on their babies to present them in theatre. In return they want reviews from the users, houseful theaters, healthy nominations and award wins and good evaluation from critiques. We decided to predict the performance of the movies which can help producers and filmmakers in making their movies and invest son right place.

Introduction

Predicting a movie's performance can help filmmaker in many ways and can help them assess different aspects of filmmaking. For generating these predictions, we need to follow the well-defined and widely used Data Science Predictions Model Methodology which has multiple steps. Starting from identifying the research goals followed by collecting and preprocessing data. An exploratory data analysis is done then before creating and evaluating a prediction model. Different models of Machine Learning are used here like K Nearest Neighbors, Decision Tree, Neural Networks, Support Vector Machine. Each of these models are used in an Ensemble with majority voting. Results of these models are then presented. Following are the details of the steps are carried out for predicting the performance of movie using Data Science toolkit and Machine Learning Basics.

Research Goal

The primary goal is to predict the performance of the movie in terms of Box Office Collection and critical reviews. A good prediction of the performance is helpful for future filmmakers to focus on the things which they lack while making a movie. One way to achieve this objective is by discovering knowledge for prediction regarding ratings of a movie, number of photos, videos, stories, reviews, rating counts and other attributes that affect the performance of the movie on Box Office. This process will be done by using Machine Learning and Data Science tools and techniques. The intention behind the model is identification and extraction of potentially valuable knowledge for foreseeing how accurate the system is in prediction of movies performance.

The main objective of the research is to predict the future performance of a movie using genre, number of award wins, number of nomination, duration, number of news article, number of user reviews, IMDB ratings and rating counts as to evaluate the success criteria of the movie.

Data Collection and Understanding

The data required is collected from Kaggle – IMDB movie rating. The data provided of the movies is from 1988 to 2017. The data provided has 14,762 rows and 44 columns. Each row has the detail of a movie that is released between these years. It contains details of the genre, ratings and user reviews. The attributes of the data collected are the following:

1. Performance: How well the movie performed.
2. tid: The identification of movie Title.
3. Title: The name of the particular movie.
4. wordsInTitle: The words that are present in a movie title.
5. url: imdb link to the movie.
6. imdbRating.: Ratings from the IMDB that are given to a particular movie from 10.

7. RatingCount: Number of people that rated the movie.
8. duration: duration of movie in seconds.
9. year: The year in which the movie is released.
10. type: file type in which the movie is present.
11. nrofWins: number of award that a movie has won.
12. nrofNominations: number of times a movie is nominated for award.
13. nrofPhotos: number of pictures that are posted of each movie.
14. nrofNews: Number of times a movie appeared on News.
15. nrofUserRiviews: The number of users that have given reviews about the movie.
16. nrofGenre: Number of genre a movie belongs to.

The remaining attributes are Boolean variables which indicates that whether movie belongs to a given genre or not:

17. Action
18. Adult
19. Adventure
20. Animation
21. Biography
22. Comedy
23. Crime
24. Documentary
25. Drama
26. Family
27. Fantasy
28. Film Noir
29. GameShow
30. History
31. Horror
32. Music
33. Musical
34. Mystery
35. News
36. RealityTV
37. Romance
38. SciFi
39. Short
40. Sport
41. TalkShow
42. Thriller
43. War
44. Western

In the data columns, there is a column called Performance. It is the column we need to work on. It has following four classifications:

1. Worst
2. Below Average
3. Average
4. Blockbuster

Each movie lies in one of these classes.

Data Preparation

As discussed above, the data acquired contains redundancy and useless information that is not needed for the predictions and modelling, we need to clean the data. Furthermore, the structure of the data is also needed to change because it is not in the form that can be entered in the model to train it and make predictions from it.

3.2 Data Transformation

For data transformation, we have removed the unwanted columns. We have the columns; title, url and so on. So, we have deleted these columns.
After this transformation, we have only the required columns needed for predictions.

3.3 Data Cleaning

The data we now have after transformation has many missing values and some other errors.
Given below is the checklist of the whole cleaning process step by step.

3.3.1 Missing Values

The data collected has many missing values. Those missing values are filled using Microsoft Excel.
In columns like rating, duration, number of genre, we used mean value of columns and for remaining columns with missing values the observations were removed because changing the missing values with mean was not suitable there. The variance of the value is very high and putting the mean value would end up in inappropriate modelling.

3.3.1 Invalid Data Format

Some of the rows in the dataset collected were misaligned, Excel was used to align them manually in their appropriate column.

4. Exploratory Data Analysis

During exploratory data analysis, you analyze the data with full concentrate and in full depth. Pictures, graphs and plots are used to grasp the information from the data. The phase is about exploring data so keeping your mind open and eyes peeled is essential. The goal is not to clean the data but its common that you will still discover anomalies you missed before. So, this is about stepping back and fix.
Graphs, Plots and Charts are used to explore data and are combined so that they can provide even more insights.
For understanding the relation between courses, we used scatter plots, Frequency Histograms, Bar charts and graphs, some of them are discussed below.

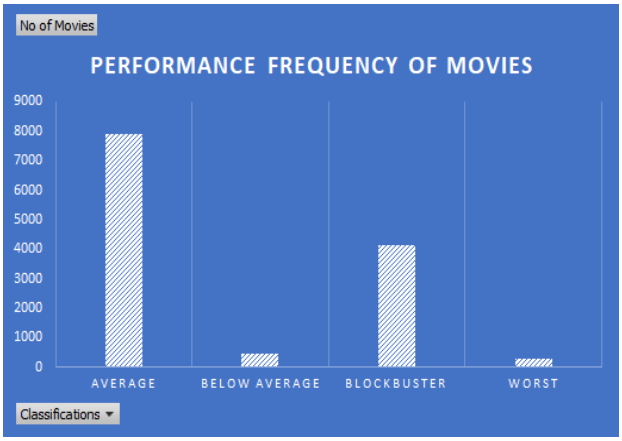

Figure 1: Frequency Histogram for Movies and Performance Classifications.

This graph shows how many movies lie in a particular class. According to which:

- Worst : 293
- Below Average: 443
- Average: 7918
- Blockbuster: 4125

There is a class imbalance problem in the dataset. As we can see that we have 7918 observations for Average and only 293 observations for Worst movies. To solve this issue, we removed rows from the Average class to see how much it affects the results but it has not affected the results significantly, so, the issue of the class imbalance is neglected to observe more data without deleting the rows. The figure below shows how the IMDB rating of a movie is related to the performance of the movie on box office. This is the most important attribute of the problem. Because the relation between these two is very high. This is also the reason behind the high accuracy of the Decision Tree Algorithm because the classifier has selected IMDB rating column for taking the first decision.

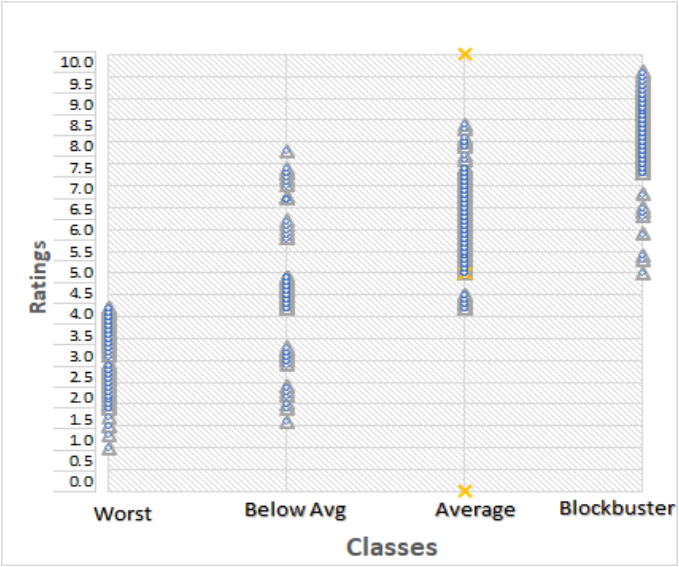

Figure 2: Plot for Movie Classification and IMDB Ratings.

The scatterplot below shows how many times the movie has come in award nominations, where ratings show the user review of the movie, the nominations and award wins show the reviews of the critiques and other specialist. This helps us in considering both the reviews for evaluation of a movie.

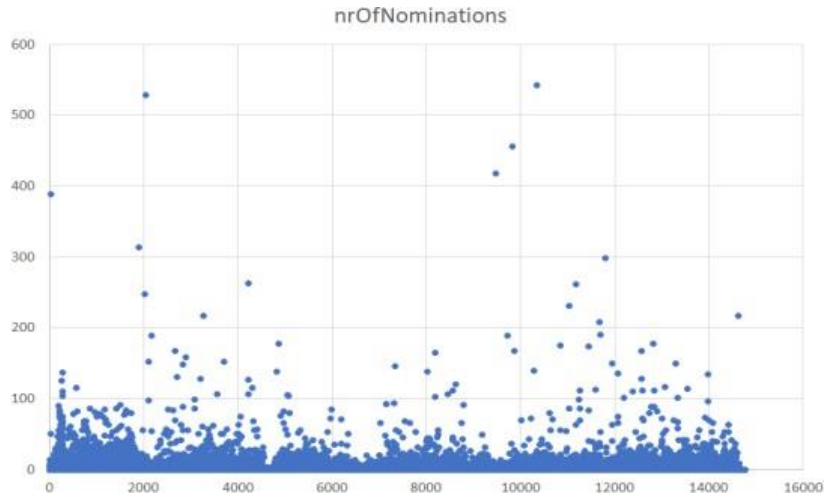

Figure 3: Scatterplot for the number of nominations for each movie.

Building Model and Evaluation

The variables selected are the ones obtained after transformation. 70-30 Cross Validation is used. 70% of the observations are used for training the models and the remaining 30% are used as testing data.

The classifiers that have been used to make model and to evaluate its accuracy are:

- K-nearest Neighbor Classifier
- Support Vector Machine
- Neural Networks
- Decision Tree.

Decision Tree is the best performer for the problem because the classification is highly dependent on ratings and decision tree selects it as the first decision making attribute.

The purpose for using four kinds of classifiers is to get the highest possible accuracy from each of the four classifiers and select the best possibilities. For this we have used ensemble classifier which basically selects the best possible accuracy from the all the classifiers.

The ensembling technique used is Majority Voting which takes the statistical mode of predictions from each classifier for each test. Each of the classifier give its own performance on different instances. If one classifier gives less performance on some movie then there are three other classifiers in support to give the best possible result. One classifier cannot give best results on each and every instance, so four different combinations of classifiers are used to work on diverse instances.

Following are the accuracy results from different classifiers:

	Accuracy	Recall	Precision	F1 Score
KNN	0.68	0.43	0.35	0.36
SVM	0.93	0.66	0.65	0.62
Decision Tree	0.97	0.91	0.91	0.91
Neural Networks	0.86	0.68	0.61	0.62
Ensemble	0.94	0.79	0.69	0.71

Table 1: Accuracy, Precision, Recall and F1 Score for classifiers.

Ensemble Prediction = Mode (KNN Prediction, SVM Prediction, DT Prediction, NN Prediction)

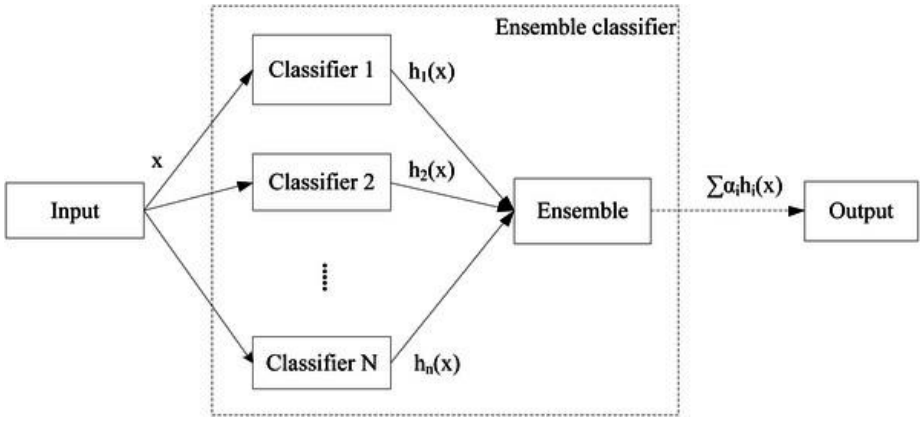

Figure 4: Working of the Ensemble Classifier.

The diagram above shows how the ensembling technique work. The data is fed to multiple classifiers, the output of these classifiers is then fed as the input of the ensemble classifier which uses them for making the new predictions by using majority voting, mean or any other ensembling technique.

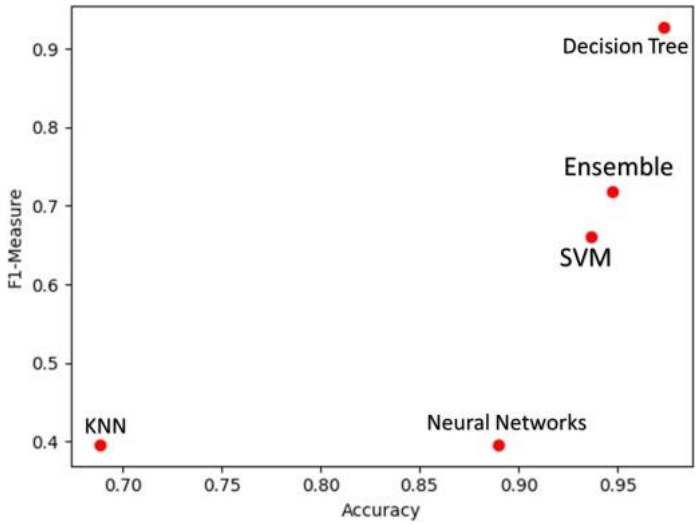

Figure 5: Plot for Accuracy and F1 Measure for different classifications.

Presentation of the findings and Application Development

The findings, accuracy and other scores are obtained on python console. An application for these predictions will work for the filmmakers and help them in predicting the future of the movies but it is not designed by us in this project because it is out of our course scope. Furthermore, we can also automate this for common usage.

Conclusion

Movies are the primary source of entertainment. It also is important part of a country's economy. Predicting the performance of movies is helpful to predict the business of a movie and help assess in casting and other aspects of the filmmaking.

References

Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing user modeling on twitter for personalized news recommendations. In International Conference on User Modeling, Adaptation, and Personalization, pages 1–12. Springer, 2011.

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

Ehsan Aslanian, Mohammadreza Radmanesh, and Mahdi Jalili. Hybrid recommender systems based on content feature relationship. IEEE Transactions on Industrial Informatics, 2016.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In Proceedings of the Fourth Conference on Recommender Systems, pages 39–46. ACM, 2010. ISBN 978-1- 60558-906-0.

Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. ACM Transactions on Information Systems, 22(1):143–177, 2004. ISSN 1046-8188.

Rahul Katarya and Om Prakash Verma. An effective collaborative movie recommender system

with cuckoo search. Egyptian Informatics Journal, 18(2): 105–112, 2017.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In Recommender Systems Handbook, pages 73–105. Springer, 2011.

Mohammad Soleymani, Guillaume Chanel, Joep JM Kierkels, and Thierry Pun. Affective ranking of movie scenes using physiological signals and content analysis. In Second Workshop on Multimedia Semantics, pages 32–39, 2008.

Conflicts of Interests

There were no conflicts while working on the project but we had some arguments and discussion over the selection of the dataset to work on. Also, there was an argument on what to choose for solving problem between Classification on Performance and Regression on ratings.