Big Data Continual Assessment

Movie Prediction Analysis

Neil Sankar

T00170945

B.Sc. (Honours) in Computing with Mobile App Development

Abstract

The Internet Movie Database (IMDb) has been a go to source for everything movie related. It features the information of thousands upon thousands of movies and allows users the chance to rate these movies based on their own feedback and through a scoring system. IMDb uses a scoring system of 0.0 to 10.0. This project is designed to predict the IMDb score using machine learning algorithms.

*Keywords:* Movies; IMDb; Dataset; Prediction; Ratings; Machine Learning; Big Data; Analysis

1. Introduction

Can movie ratings be predicted? If so, how accurately can they be predicted? The film industry is one that can be extremely rewarding, extremely unrewarding, or somewhere in the middle, in terms of success both financially and reputationally. Thousands of movies are released every year to cinema, and the success of these movie can depend on many factors, such as the cast, the director, the budget, the genre, the plot etc. In this project, a dataset of movies provided from the IMDb movie database is analyzed in order to find patterns and interesting correlations. It is then cleaned up in a way to maintain data integrity. Once the dataset is deemed to be satisfactory in terms of integrity and featuring only required columns of data, it is then transformed in order to streamline and increase the potential success of the machine learning algorithms that are used to predict the movie rating. This document covers what was learned from the assignment and acts as a guide for the reader in understanding the work carried out.

* 1. **Resources**

Dataset - IMDb dataset of 5000+ movies. Available from Kaggle

Notebook – T00170945\_Big\_Data\_Neil\_Sankar

Report – T00170945\_Big\_Data\_Neil\_Sankar\_Report

* 1. **Literature Overview**

There have been many different examples of research carried out in the field of movie prediction. Examples such as the research by Akshay Sadarangani and Ashton Vaz has been done with the aim of predicting the movie rating by analyzing different features of information in the IMDb dataset, and predicting the IMDb rating using machine learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Decision Trees. (Sadarangani & Vaz, n.d.). Research has also been carried out to predict the IMDb rating by comparing the correlations of the rating vs different columns in the dataset, as done by Chuan Sun. (Sun, 2016). The work of Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke aimed to predict the IMDb rating by analysing social media platforms such as Twitter and YouTube to gather public feedback on different movies and did so successfully. (Oghina, et al., 2012). This gave an indication that this was a feasible project and similar findings and results could be achieved.

1. Methodology/Methods

The dataset was imported into the notebook and initial observations were made on the data contained. In a project of this nature, a lot of the early analysis was to see what the highest grossing and rated movies were, what the lowest rated and financially disastrous movies were. Who the most popular director was etc.? These findings are for the most part displayed on bar charts.

The next method was to separate the dataset into 2 categories, numerical data and worded data. From doing this, it became apparent that there was a lot of missing data in the dataset. This needed to be corrected so for numerical data, the mean value of all values in its respective column substituted for null missing values, and as for written data, the most popular category for each column replaced missing values i.e. USA and English replaced missing values in the country and language columns respectively.

The next step carried out was done to assist the machine learning algorithms which were to be employed later in the project. Data, both numerical and worded was banded together into categories represented a single integer based on which category this data fell into. This was done until every single column in the dataset was reduced down to numerical values. This was a long drawn out task but enabled very quick machine learning times. The concept of this was observed from viewing Jeremy Hummel’s analysis on IMDb Predictive Analysis. (Hummel, 2017)

Before the machine learning algorithms were started, the data set had to be changed. Any missing values that may have appeared after the categorical banding were removed altogether. The data was then split into 80% training data and 20% test data.

The machine learning algorithms all ran successfully, and the findings can be seen below.

* 1. **Findings & Tables**

In this section, Tables are featured displaying some of the various findings from the analysis of this project.

Table 1. Breakdown of worded and numerical data in the dataset.

|  |  |
| --- | --- |
| Type | Features |
| String | color, director\_name, actor\_2\_name, genres, actor\_1\_name, movie\_title, actor\_3\_name, plot\_keywords, movie\_imdb\_link, language, country |
| Numerical | imdb\_score , movie\_facebook\_likes, title\_year, duration, aspect\_ratio, num\_critic\_for\_reviews, director\_facebook\_likes, gross, actor\_1\_facebook\_likes, actor\_3\_facebook\_likes, actor\_2\_facebook\_likes, facenumber\_in\_poster, num\_user\_for\_reviews, cast\_total\_facebook\_likes, num\_voted\_users, content\_rating, budget |

Table 2. Model training accuracy and predicted IMDb scores by each model.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Prediction |
|  |  | From 1009 movies |
| Logistic Regression | 75.98 | 759 / 1009 |
| Decision Tree | 99.88 | 720 / 1009 |
| k-Nearest Neighbor | 85.32 | 722 / 1009 |
| Random Forest | 99.88 | 794 / 1009 |
| Support Vector Machines | 83.74 | 784 / 1009 |
| And another entry | 73.00 | 1. 1009 |

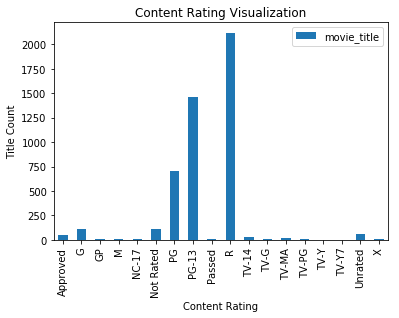


Figure 1. Distribution of movies based on content rating

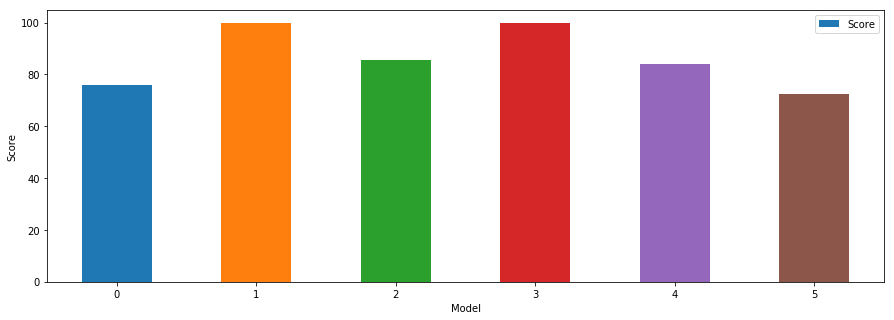


Figure 2. Training accuracy by each model

0 – Logistic Regression, 1 – Decision Tree, 2 – k-Nearest Neighbor, 3 – Random Forest, 4 – Support Vector Machines, 5 – Gaussian Naïve Bayes

1. Discussion

The task of this project from the very start was to make predictions on movie ratings. By using the IMDb dataset available from Kaggle, I was able to do exactly what I set out to do, which was to analyze movie data, find meaningful correlations, and ultimately use Machine Learning algorithms to predict movie ratings. I believe my findings echo what some of the previous researchers (outlined in literature overview and references) obtained during their own analysis. Can you use machine learning algorithms to predict movie ratings? Yes.

1. Conclusion

I am satisfied with the results I obtained during this analysis of the IMDb Dataset. It was a very interesting project, and one that I can use as a basis for further analytical projects in the future.

References

Hummel, J., 2017. *IMDb Predictive Analysis.* [Online]   
Available at: https://github.com/jmhummel/IMDb-predictive-analytics

Oghina, A., Breuss, M., Tsagkias, M. & de Rijke, M., 2012. [Online]   
Available at: https://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/oghina-predicting-2012.pdf

Sadarangani, A. & Vaz, A., n.d. [Online]   
Available at: https://github.com/aksh4y/IMDb-Rating-Prediction/blob/master/Predicting%20Movie%20Ratings%20Using%20IMDb%20Dataset.pdf  
[Accessed n.d].

Sun, C., 2016. *Predict Movie Rating.* [Online]   
Available at: https://nycdatascience.com/blog/student-works/web-scraping/movie-rating-prediction/