**Predictive Analysis of Movies on OTT platforms Using Machine learning Approach**

Devesh Shetty

*Department of Information Technology, Shree LR Tiwari College of Engineering, Mumbai, India*

deveshetty.2201@gmail.com

Abstract— **Movies are considered to be an important art forms, a worldwide source of entertainment, and a powerful medium for educating or indoctrinating citizens. As far as the current pandemic situation is concerned, OTT platforms act as one of the most entertaining factors and a significant stress reliever for people around the globe. This project aims to explore all the movies in popular OTT platforms, in order to gain interesting insights. This is carried out with the aid of a Kaggle dataset, collected from Netflix, Prime Video, Hulu and Disney+ API. Dataset contains the complete information of all the movies, their ratings and the corresponding OTT platforms in which they are available. It provides detailed information such as Year of release, Genre, IMDb rating, Director and the Language of each movie. Here in this project, we are using Multiple linear regression and Random Forest to analyze our data and to get meaning full insights from the data collected from different OTT platforms collected from Netflix, Prime Video, Hulu and Disney+. Furthermore, the result obtained from each of these algorithms are compared to understand their respective suitability under varied conditions*.***

Keywords— **R, dplyr, ggplot, Netflix, Amazon Prime, Hulu, Disney+**

1. Introduction

Traditionally TV has been the source of entertainment along with recording it within CDs, after the boom of technology and internet there have been many OTT (Over the Top) apps available in the market which has now a days almost replaced the traditional ways of consuming content, OTT apps like Netflix, Prime Video, Disney Plus, Hulu have emerged and are preferred more due to their ease and self-paced content consumption. Here the data is analysed and some visualization and manipulation are carried to get a more precise and a graphical picture of the entire dataset and this will help to shed views of a particular product or topic. If people find topics relevant or interesting, then they would desire to share their opinion about the topic. The topic could be a product or any other object. Understanding this can help us decide and OTT platforms the type of movies that are popular among the people rather it be depending on reviews, rating or age group. .[3] In this project we are also using algorithms like Multiple linear regression and Random Forest to analyse our data and to get meaning full insights from the data collected from different OTT platforms collected from Netflix, Prime Video, Hulu and Disney+. Furthermore, the result obtained from each of these algorithms are compared to understand their respective suitability under varied conditions.

1. Related Work

An [1] In this paper Movies Reviews Sentiment Analysis and Classification, the authors main focus is how sentimental is performed here goal of this work is to address SA by constructing an approach that can classify movie reviews and then compare the results in an inclusive study of eight well-known classifiers. To evaluate the proposed model, IMDB reviews real dataset was utilized. Tokenization was applied on the dataset to transfer strings into word vector, then stemming was used to extract the root of the words, afterwards gain ratio was applied on the dataset as an attribute selection algorithm. Then, the data was split into training and testing datasets using the percentages 66%, 34% respectively. In order to compare the eight different classifiers, five different evaluation metrics are utilized. The results show that Random Forest outperforms the other classifiers. Furthermore, Ripper Rule Learning performed the worst on the dataset according to the results attained from the evaluation metrics.

[2] Here we followed paper based on The Performance Comparison of Multiple Linear Regression, Random Forest the paper gave us an detailed classification about the processes . For comparison of there are several data mining techniques, the power production data from a Photovoltaic Module was used in the research. In this study, the model was constituted from seven variables. The highest correlation coefficient was obtained in Artificial Neural Network architecture (R = 0.997). The study by author also showed the importance of data mining method. If this study had been evaluated by MLR then the findings of the study would have been obtained biased and non-robust. So, a study must be evaluated by robust statistical methods in order to estimate a model in a high accuracy rate. This study showed that the MLP-ANN architecture has the best performance when compared with MLR and RF.

1. Proposed system

### A. Data Pre-processing

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

### Steps Involved in Data Pre-processing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

### 2. Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

Removing NA value: This approach is suitable only

when the dataset we have is quite

large and multiple values are missing within a table.

### 3. Filling or Skipping missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value. Or we can skip such values

Helps us to perform data cleaning operations on the data

dp111 = complete.cases(dp11)

dp11\_11= dp11[dp111, ]

dp1= dp11\_11[!duplicated(dp11\_11),]

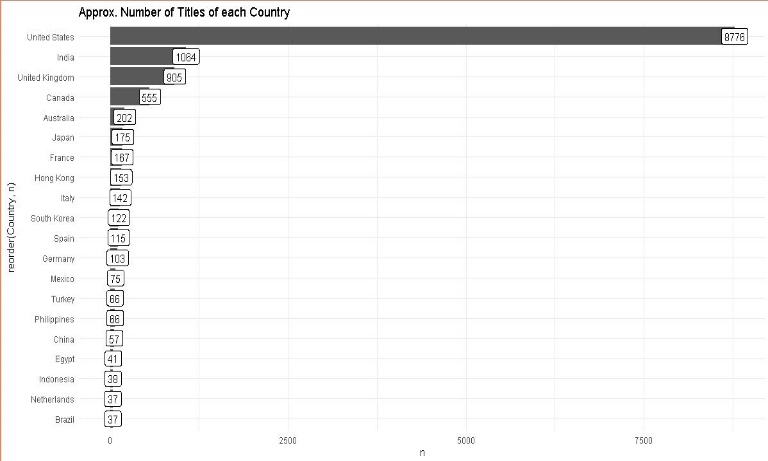
*code example*

B. Data Visualization and manipulation of data:

Our dataset is quite large so let’s visualize through graph

* *Visualizing Number of Titles by each country:*

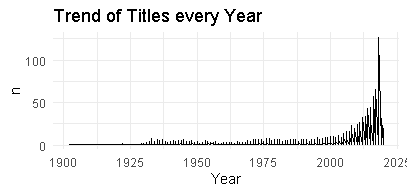
This graph shows number of titles of each country within different OTT platform. As we can see major titles is of USA, India and United Kingdom

****

**Fig. B1. Number of titles of each country**

* *IMDB Rating trends by year:*

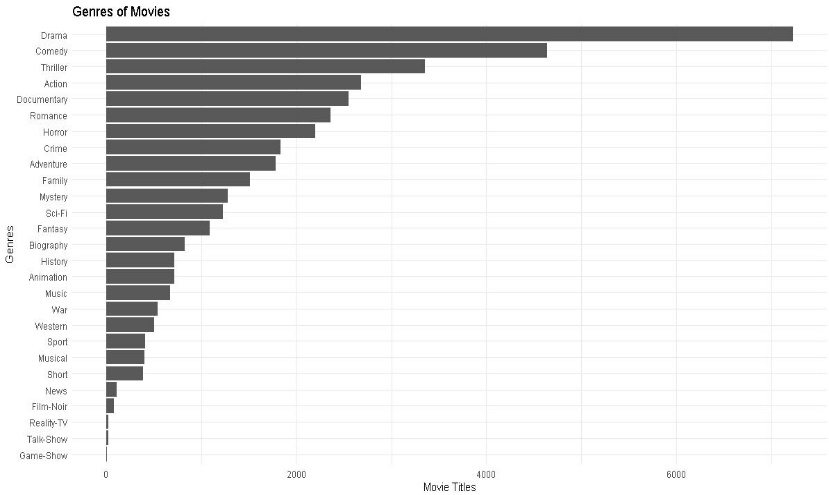
As we see the profit yearly so plot the graph as per year of titles. Below is graph of different titles trending yearly. In these modern times as, the internet is growing so is the availability of various movies and series. With the increasing time the watching among people is also increasing.

****

**Fig. B2.Trends each year**

* Genre of Titles:

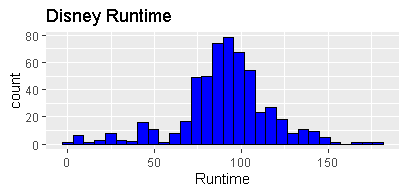
Next graph is trend by genres of movies across different OTT platform. Majority of people like drama and comedy because they made them fresh. So, we can see the highest trend is among drama, comedy and thriller.

****

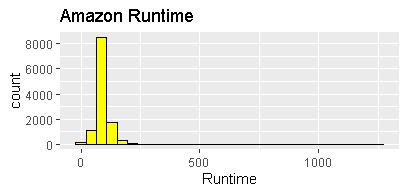
**Fig. B3.Trends by genres**

* Runtime Analysis of platforms:

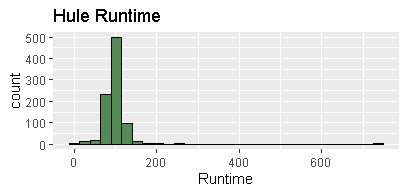
The runtime analysis of the platforms gives us a brief idea about hours of content present across different OTT platforms. Below we can see the runtime of Disney, Amazon Prime, Netflix and Hulu.



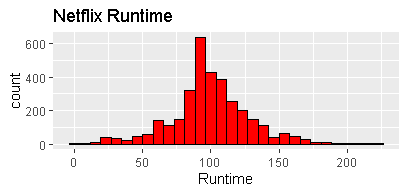
**Fig. B4. Disney Plus runtime analysis**



**Fig. B5. Amazon Prime runtime analysis**



**Fig. B6. Hulu runtime analysis**



**Fig. B7. Netflix runtime analysis**

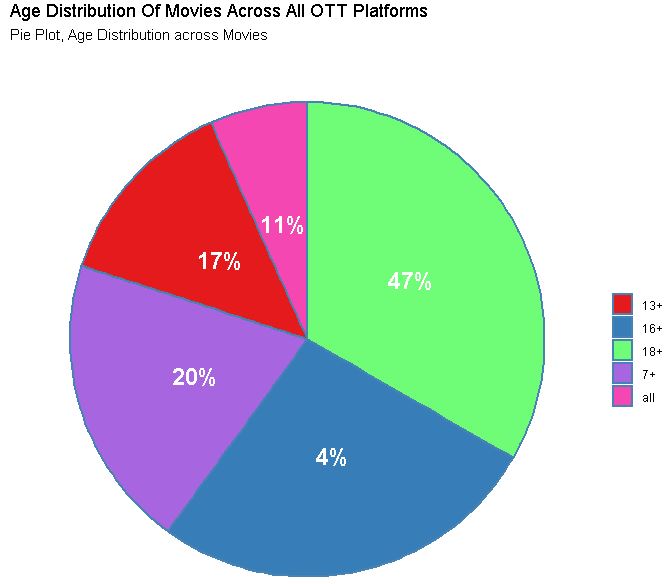
The above figures dhow a detailed runtime analysis of each movies on various OTT platforms such as Netflix, Amazon Prime, Hulu, Disney Plus and Netflix.

These analysis helps us to get detailed analysis of the entire system and how the trends follow up.

These runtime gives us an average time and how divers are the movies on these OTT platforms.

* Age Distribution of Movies:

Age is the major factor of what movies or shows to watch below is the graph of age distribution. From the graph we can see majority of adult people is watching movies and shows.

****

**Fig. B8. Age Distribution**

C. The Training and Testing Phase:

* Data preprocessing before training:

As there are various OTT platform, we have taken a large amount of dataset so for getting a higher accuracy we need to cleanse data. So, in preprocessing step we are cleaning the data by first removing all the N.A values in the respective column age, year and IMDB. After we are selecting particular column for our processing.

dp3\_ml = dp1 %>%

select (Runtime ,Genres , Year,IMDb,Age)

%>%

mutate(Age\_i = parse\_number(Age) )%>%

drop\_na()

dp3\_ml %>% filter(is.na(Runtime)==T

& is.na(Age\_i)== T & Age\_i!= 'all'

& is.na(Year)==T & is.na(IMDb)==T)

%>% select(Runtime ,Genres , Year,IMDb,Age\_i )

**code example**

* Splitting data for training and testing training and testing set.

The next step would be dividing the data into training and testing set. We divided dataset into 70% for training set and 30% for testing set

set.seed(88)

split <- sample.split(dp2\_ml$IMDb, SplitRatio = 0.70)

**code example**

* Multiple linear regression:

Here we are using multiple linear regression which often known as multiple regression, is a statistical technique that predicts the result of a response variable by combining numerous explanatory variables. MLR aims to

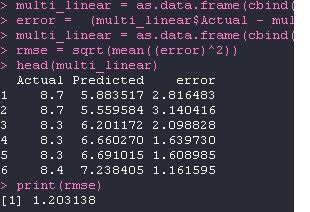
represent the linear relationship between the explanatory (independent) variables and the response (dependent) variable.

model <- lm(IMDb ~., data=train)

predicted\_value <- predict(model, newdata = test)

multi\_linear = as.data.frame(cbind(Actual = test$IMDb, Predicted = predicted\_value))

**code example**



**Fig. C9. Multiple Linear Regression**

* Random Forest:

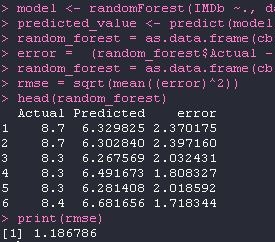
Random forest is a learning algorithm that is supervised. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method is based on the premise that combining different learning models improves the final output. Random forest has the advantage of being able to be utilize for both classification and regression issues, which make up the majority of existing systems.

model <- randomForest(IMDb ~., data=train)

predicted\_value <- predict(model, newdata = test)

random\_forest = as.data.frame(cbind(Actual = test$IMDb , Predicted predicted\_value))

**code example**



**Fig. C10. Random Forest**

1. comparative Study

### The comparison of both algorithms through results can give us a better understanding about their accuracy and usefulness within the project. The Fig C9 which is for Multiple Linear regression gives us the actual value along with predicted and error value. The root mean square error is 1.203138 here. In Fig C10 which is for Random Forest, it gives us the actual value along with predicted and error value. The root mean square error is 1.186786 here. Hence from here from the root mean square error we can judge that the Random Forest algorithm is much more accurate and efficient as compared to the multiple linear regression.

1. conclusion

Here in, after cleaning and removing duplicate values from data analyzed the data with respect to its IMDb ratings which as predicted using many other independent parameters. The aim of this experiment was movies on OTT analysis wherein we predict the IMDb ratings based on various other parameters using Multiple Linear Regression Algorithm and Random Forest Algorithm. For the data we used Kaggle which provided us with the data the data was around 16,744 which was later processed and manipulated to get meaning full insights. From results It is clear that the accuracy of Random Forest Algorithm with root mean square error = 1.186786 is a bit higher as compared to Multiple Linear Regression Algorithm with root mean square error = 1.203138 but as the data will increase the accuracy of the algorithms will also change. Hence the addition of more data will help us in better prediction for actual values.

1. Future Scope

The Existing Database is of around 16k data which is not sufficient to get meaningful insights from the data that accurately. Also, this type of database is limited for processing of structured data and has a limitation when dealing with a large amount of data. So, the use of Big Data technologies like Hadoop can be used to achieve better results. We could further improve our analysis by using various different algorithms and see which gives us better accuracy.

References

1. Yasen, M., & Tedmori, S. (2019). *Movies Reviews Sentiment Analysis and Classification. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT).* doi:10.1109/jeeit.2019.8717422.

[2] Kayri, M., Kayri, I., & Gencoglu, M. T. (2017). The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data. 2017 14th International Conference on Engineering of Modern Electric Systems (EMES). doi:10.1109/emes.2017.7980368

[3] Prof Ria Patnaik, Prof Reema Shah, Prof Upendra More, Rise of OTT Platforms: Effect of the C-19 Pandemic, -- Palarch’s Journal Of Archaeology Of Egypt/Egyptology 18(7). ISSN 1567-214x

[4] Song, Minzheong A Comparative Study on Over-The-Tops, Netflix & Amazon Prime Video: Based on the Success Factors of Innovation -International Journal of Advanced Smart Convergence Vol.10 No.1 62-74 (2021)