Coursework-1
ECS784U/ECS784P - Data Analytics - 2022/23

Shrey Mahar (ID No.- 220307635)

School of Electronic Engineering and Computer Science

Queen Mary University of London

s.mahar@se22.qmul.ac.uk

IMDB Top 250 Movies

1.Introduction:-

IMDB is a famous platform for movies, TV shows, and all entertainment related topics. IMDB has a list of all the movies and entertainment related topics in which we can see the details like genre, rating, budget and everything. In this project we got a dataset of top 250 movies and it consists of 250 rows, in which each represents a movie and 15 columns, which contains details of each movie like rating, director, genre, etc. The dataset contains a variety of films from different genres, era and nations, old films like “The Godfather" as well as recent ones like “Avengers”.

The IMDB Top 250 Movies dataset's high data quality is one of its most important benefits. Users of IMDB, a popular and dependable online source for movie information, take ratings seriously. The dataset offers accurate information about the directors, actors, and plot summaries of the films as a result, making it a trustworthy source of data for analysis. This dataset can be used for a variety of uses, like by examining the top film’s release year, checking how many hits and flops there were, or which genre is most liked by the audience.

2.Literature Review:-

Film industry is an excellent contributing factor to the economy of a country with a good amount of revenue generated every year. In recent years, ML and data mining have been used to analyse the success of movies.This literature review aims to explore some of the existing research on predicting movie success in the U.S. market.

In their paper, **"Predicting Movie Success in the U.S. market,"** Darin Im, Minh Thao, and Dang Nguyen from Stanford University explored the use of multiple regression and correlation analysis to predict movie success. The study found that a movie's budget, genre, and release date significantly impact its success. The paper **"Predicting Movie Rating based on Text Reviews"** by Sagar V. Mehta, Rose Marie Philip, and Aju Talappillil Scaria, used a machine learning based approach to predict ratings using text review. The authors collected reviews from the IMDb database and used natural language processing techniques to extract relevant features. The paper **"Using Social Networks to improve Movie Ratings**

predictions" by Suhaas Prasad proposes a method to improve movie rating predictions by leveraging social networks. The author uses data from Facebook to determine the social connections between users who rated the same movies.
W. Zhang and S. Skiena, **"Improving movie gross prediction through news analysis"**, IEEE/WIC/ACM International Conference on Web In-telligence and Intelligent Agent Technology, Milan, 2009, the authors collect news articles and extract relevant features to train machine learning models that outperform traditional models in terms of accuracy.

3.Data Processing:-

The data we obtained needed to be cleaned for further analysis, so we need to remove inconsistency and make it simple to better use it for our analysis. Once the data is suitably cleaned and integrated, it goes through selection and transformation activities, to translate the textual information (where necessary) into numerical information, which is better analyzed by data mining processes. This also discards irrelevant data, and selects a subset of the data to be mined, which is better suited to perform the analysis of our choice.
Below is a description of all the steps that were performed to process and clean the data :

3.1)Load Data from Kaggle:-

We took the dataset from Kaggle and we can see rank, name, year, rating and all the details of the top movies.
df = pd.read_csv("IMDB Top 250 Movies.csv")

3.2)Removing and converting values:-

->Budget section:-
In the budget section we have to convert the string into float so that we can use it, like there was a dollar sign in the string so we've removed the sign and made it into a simple float value. Next the values were in different currencies, so we have converted the budget of the movie into a single currency i.e. U.S. dollar according to the conversion rate, and converted into simple float.
->Box-office section:-
In this section 'estimated' was written in some places, so we removed them all and converted them into float value like we did in the budget section.
->Run-Time section:-
We've converted the hours value into minutes, like some movies are of one and half hour so it'll be converted into minutes float value.
->After that we've splitted the genre and made a list and same did with the director, writers and cast so that they can become a list of strings and we can separate them out later.

4.Exploratory Data Analytics:-

Now our data is processed and cleaned up, we'll perform EDA on the dataset by using data visualisation techniques to get some graphs or patterns and information regarding the movies.

Data Visualization may be viewed as the process of extracting and visualising the data in a very clear and understandable way without any form of reading or writing by displaying the results in the form of pie charts, bar graphs, statistical representation and through graphical forms as well.

The following section describes the results obtained after performing Exploratory Data Analysis on the cleaned dataset :-

4.1)<u>Genre Analysis</u>:-

First of all we've observed the top 10 popular genres and plotted a graph as shown in Fig.1.

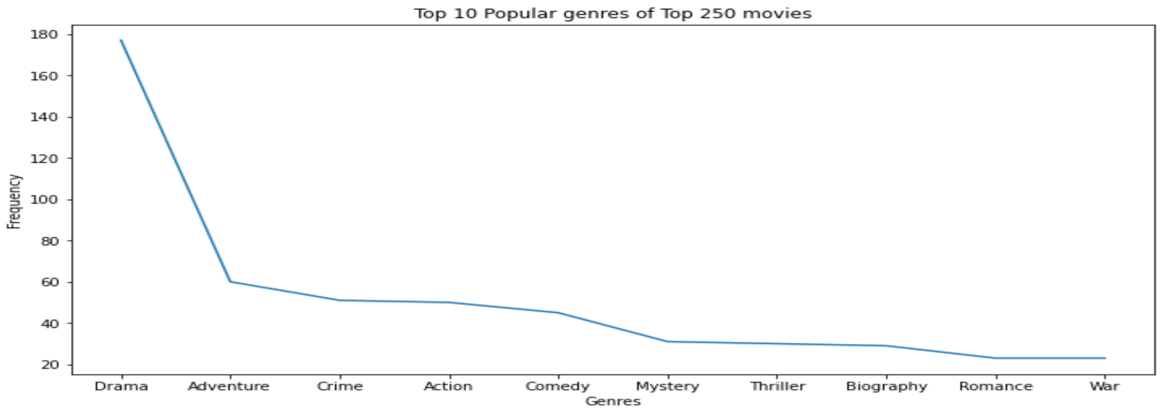

Fig.1

Now as we can see in Fig.2, the rating of the movie is varying across the genre, here's the box graph. Box graph is telling us the first quarter, fliers, IQR(interquartile range), median and mean(blue colour dots).

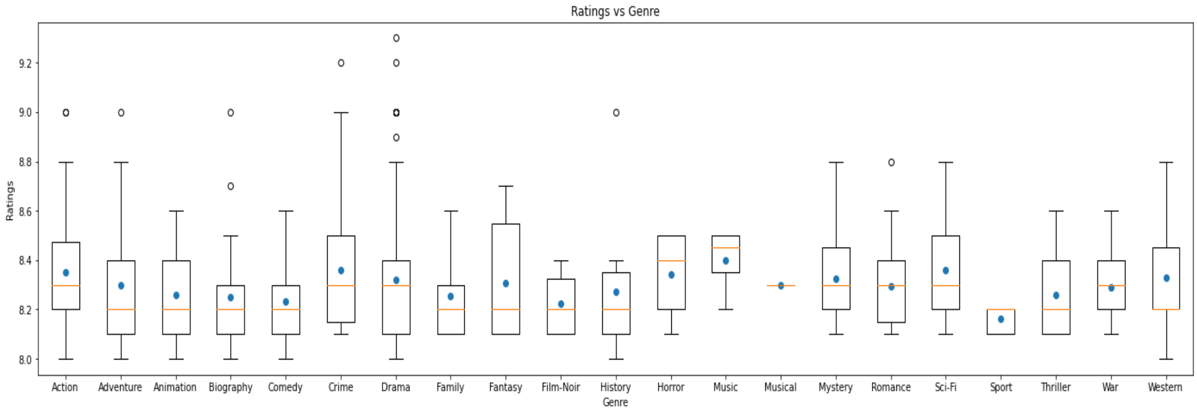

Fig.2

Q1-1.5IQR Q1 median mean Q3 Q3+1.5IQR

```
             |---------:-----------|

  o     |--------|        :    o    |--------|   o  o

             |---------:-----------|

flier        <--------------------->          fliers

                    IQR
```

4.2)<u>Category Analysis</u>:-

There are several box office categories but there is no specific formula or threshold for determining the categories. Instead, the categories are based on industry standards and expectations, as well as the production budget and marketing costs of the movie. For the sake of simplicity, we're using the following formula in comparison to their budget.

All time Blockbuster: Above 300%
Blockbuster: 200%-300%
Super Hit: 175%-200%
Hit: 125%-175%
Average: 100%-125%
Flop: 75%-100%
Super Flop: 50%-75%
Disaster: Below 50%

Now we’ve calculated the Return of investment(ROI), and plotted it as shown in Fig.3.

ROI=(Box office collection)/(Budget)

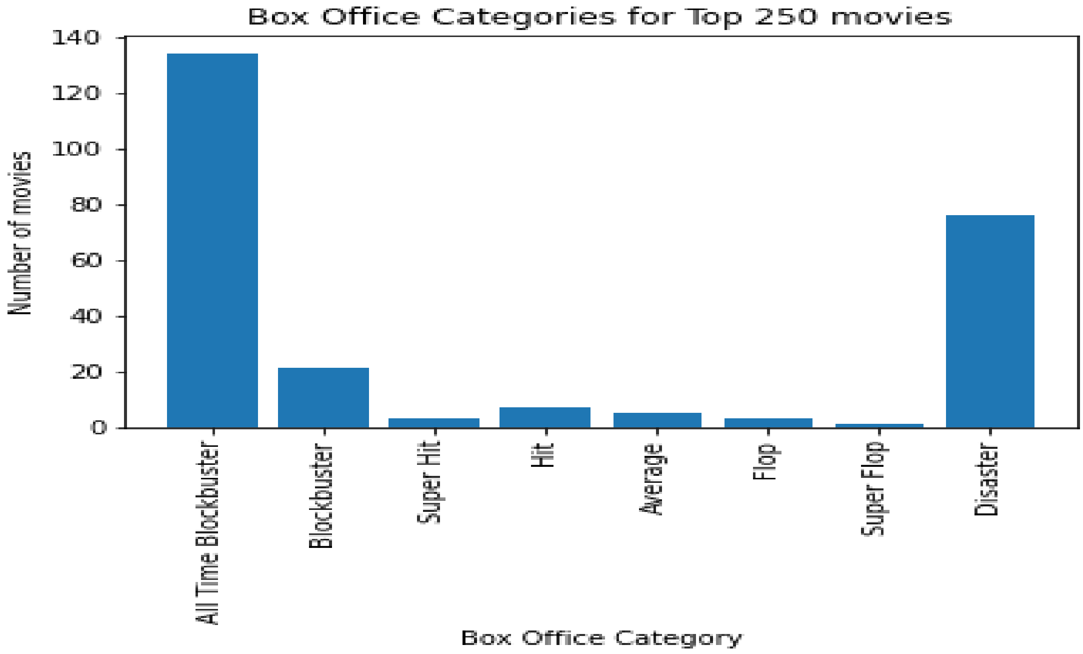

Fig.3

4.3)Distribution according to certificates:-

In this the data is distributed according to the certificate, whether it is R-rated, PG, not rated, etc. as shown in Fig.4.

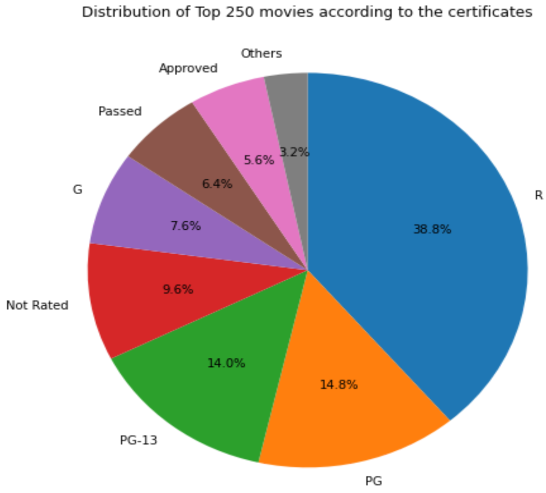

Fig.4

4.4)Director Analysis:-

Plotted list of top 10 best directors whose movies are in top 250 as shown in Fig.5.

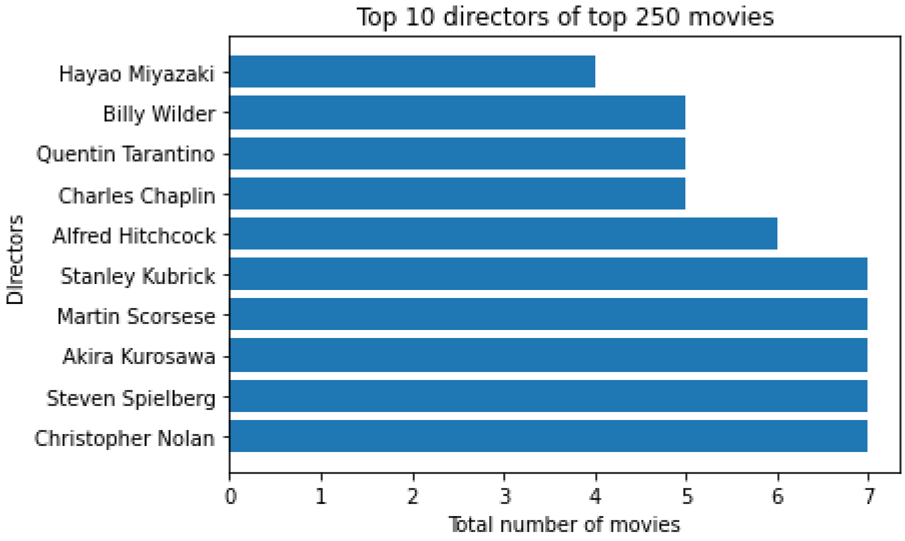

Fig.5

5.Model Selection:-

For regression we've calculated the probability of the values, like how many times the name of director came and its probability, made different factors for each, converting all the values(genre, director, etc.) into numbers, so that we can represent the number.

The following section describes the results of the various models fitted over the data and the inferences that were derived out of them. Linear regression and SVM can both be used for predicting the rank of movies in the used dataset,

5.1)<u>Linear Regression</u>:- Linear regression is a way to predict the rank of a movie in this dataset because the target variable, which is rating, is a continuous variable. Linear regression is a statistical method that models the relationship between a dependent variable (in this case, the rating) and one or more independent variables (the features of the movie) as a linear equation. The model tries to minimise the difference between the predicted rating and the actual rating for each movie in the dataset. In this we are splitting the data in training and testing set. The accuracy score, which is **<u>0.6578947368421053</u>** in this case, helps to observe how well our model is working and predicting the rating of movies in the dataset.

5.2)<u>Support Vector Regression</u>:- SVR is a SVM algo used for regression analysis, which makes it perfect to predict the rank of movies in the above dataset. In regression, the SVM algorithm attempts to find the hyperplane that best fits the data by minimizing the distance between the predicted values and the actual values. SVR can be useful for a non-linear relationship. In short SVR is a good way to predict rank of movie in this dataset as it is an algo that capture non-linear relations between target variable and features, making it useful for predicting continuous variable like movie rating. Accuracy we got in SVR is **<u>0.7631578947368421</u>**.

-> Comparing above mentioned models, we can see that SVR is giving better accuracy so it’s better to use SVR as linear regression can be effective in handling complex datasets, but linear regression is simpler and faster whereas SVM is more complex, but can handle more complex dataset and give higher accuracy as we can see in above case.

6.<u>Conclusion:</u>

IMDB Top 250 Movies is a very helpful dataset for learning about the greatest movies of all time. In the result we can find the ranking of the movies based on their genre, director and some other factor and it provides details and exact information about the actor, genre, director, collection and all the other aspects of movies and the user can search it according to them.

6.1)<u>Advantages:-</u>

-> It is a reliable indicator of the general consensus because it provides information on the highest-rated movies based on user ratings or public opinions.
->It provides excellent information on public opinion, a huge selection of movies, the dataset's range of films from different genres, ages, and nations allows for the analysis of historical and cultural patterns in the cinema industry. ->Another advantage is that the dataset can be used to identify famous personalities in the film industry, such as directors and actors.

6.2)<u>Limitations:-</u>

->Firstly the dataset only contains top 250 movies, which may not be completely representative of the film industry. So many movies that are not included in the dataset may be important to figure out current movie trends or the main players.

->Whenever the user gives rating to a movie, the dataset may be biassed towards a particular genre preference. This suggests that the dataset may not be representative of the entire film industry, as certain genres may be overrepresented or underrepresented.

->The dataset is static and does not take into account any revisions or changes to the top-rated movies over time.This means that some films that were previously popular may have been replaced by newer films, and the dataset may not accurately reflect changes in public opinion over time.

6.3)<u>Possible Future Improvements or Directions:-</u>

->We can expand this analysis to include some more datasets such as box office data, to identify trends and patterns that may help predict which movies will be successful at the box office.

->It will also be helpful to include sentimental analysis of reviews to check which aspects of a movie are most commonly appreciated or criticised.

References:-

1.) https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset

2.) Darin Im, Minh Thao, Dang Nguyen, Predicting Movie Success in the U.S. market, Dept.Elect.Eng, Stanford Univ., California, December, 2011

3.) Sagar V. Mehta, Rose Marie Philip, Aju Talappillil Scaria, PredictingMovie Rating based on Text Reviews, Dept.Elect.Eng, Stanford Univ.,California, December, 2011

4.) Suhaas Prasad, Using Social Networks to improve Movie Ratingspredictions, Dept.Elect.Eng, Stanford Univ., California, 2010

5.) Cohen, J., Cohen P., West, S.G., & Aiken, L.S. (2003). Applied multiple regression correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates

6.) W. Zhang and S. Skiena, Improving movie gross prediction through news analysis, IEEE/WIC/ACM International Conference on Web In-telligence and Intelligent Agent Technology, Milan, 2009