

1. INTRODUCTION

In this new world of Data Science, we wanted to create something special and unique. That's when we came up with an idea of predicting how much a movie would make using data from the IMDb website. We scraped information from the top 1000 movies on the IMDb website, which includes attributes like Directors, Stars, Meta score, Gross, etc. We thought this data could help us understand what makes a movie successful.

A. BUSINESS UNDERSTANDING

We started building a prediction model which makes educated guess on a movie's earnings. This project isn't about numbers; it's about decoding the magic behind top movies. We are not using Crystal balls, but magic of math and Machine Learning with a spice of storytelling.

Join us as we explore the world of movies and data science, where we turn raw scraped data into a useful dataset with a prediction model. This is the story of predicting revenue, one at a time.

B. BUSINESS OBJECTIVE

Predictive Accuracy: Create a strong predictive model that can be used to forecast box office receipts for films with precision, assisting producers in making the best decisions.

Key Factor Identification: Assist production businesses in customising their strategy, identify the crucial elements that greatly impact a film's box office performance.

Enrichment of the Audience: Give film buffs insightful information about the business to help them develop a greater appreciation and understanding of the film industry.

User-Friendly Interface: Provide a user-friendly interface that allows both experts and amateurs to access and interact with the prediction model.

Continuous Improvement: Add additional data sources for increased accuracy and update and improve the model on a regular basis to accommodate the changing needs of the film business.

Our goal is to provide a complete, approachable solution that connects the two worlds.

C. PROBLEM STATEMENT

With the aim to get the most out of their promotional and distributional approaches movie making organizations aim to predict a movie's future box office revenue . They look for important elements that aim for that have a big impact on films box office performance.

The issue is that , abundance of data, box office prediction remains a mystery . We need to establish a connection between the delight of movies and statistics.

Making the art of cinema into science was our greatest hurdle. For the purpose of figuring out the potential revenue of a film , we began to come up with a model .

Considering that we are here we have the opportunity to look into and notify other spectators what we have managed to learn. We intend to use the task at hand as a change to learn the inner working of successful movie makers.



Fig 1.C.1 IMDb top 1000 movies

2. DATASET

A. DATA ACQUISITION

Data was obtained by web scraping the website of IMDb's top-rated movie listings using one of the most famous library beautiful soup .

B. INSPECT THE DATA

	Name_of_movie	Year_of_release	Watchtime	Movie_Rating	Metascore	Votes	Grosses
0	The Shawshank Redemption	1994	142	9.3	82	2,811,733	\$28.34M
1	The Godfather	1972	175	9.2	100	1,959,532	\$134.97M
2	The Dark Knight	2008	152	9.0	84	2,793,415	\$534.86M
3	Schindler's List	1993	195	9.0	95	1,413,454	\$96.90M
4	The Lord of the Rings: The Return of the King	2003	201	9.0	94	1,925,087	\$377.85M

Description	Director	Stars	Genres
Over the course of several years, two convicts...	Frank Darabont	['Tim Robbins', 'Morgan Freeman', 'Bob Gunton']...	Drama
Don Vito Corleone, head of a mafia family, dec...	Francis Ford Coppola	['Marlon Brando', 'Al Pacino', 'James Caan', '...]	Crime, Drama
When the menace known as the Joker wreaks havo...	Christopher Nolan	['Christian Bale', 'Heath Ledger', 'Aaron Eckh...]	Action, Crime, Drama
In German-occupied Poland during World War II,...	Steven Spielberg	['Liam Neeson', 'Ralph Fiennes', 'Ben Kingsley']...	Biography, Drama, History
Gandalf and Aragorn lead the World of Men agai...	Peter Jackson	['Elijah Wood', 'Viggo Mortensen', 'Ian McKell...]	Action, Adventure, Drama

Fig 2.B.1 First 5 rows of out dataset(df.head())

C. DATASET DESCRIPTION

This dataset comprises of 1000 observations and 11 attributes, providing a rich mix of both numeric and text data. Notably, this dataset was web scraped from IMDb's top-rated movie listings

(https://www.imdb.com/search/title/?count=100&groups=top_1000&sort=user_

rating) on 2nd September , 2023, by Sneh Patel, Shubh Agarwal and Shreyasee Shinde.

Attributes are as follows:

1. Name_of_movie
2. Year_of_release
3. Watchtime
4. Movie_Rating
5. Metascore
6. Votes
7. Grosses
8. Description
9. Director
10. Stars
11. Genres

3. DATA CLEANING AND MANIPULATION(DPL)

Below we can see that are some columns like year_of_release , Metascore , Votes, Grosses which were extracted as object datatype during scrapping due to which (‘’) was considered as data so there we no null Values . So we need to convert the required columns to their respective datatype.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Name_of_movie         1000 non-null   object
1   Year_of_release       1000 non-null   object
2   Watchtime             1000 non-null   int64
3   Movie_Rating          1000 non-null   float64
4   Metascore             1000 non-null   object
5   Votes                 1000 non-null   object
6   Grosses               1000 non-null   object
7   Description           1000 non-null   object
8   Director              1000 non-null   object
9   Stars                 1000 non-null   object
10  Genres                 1000 non-null   object
dtypes: float64(1), int64(1), object(9)
memory usage: 86.1+ KB
```

Fig 3.1 Information of all the attributes(df.info())

A. DATA TYPE CONVERSION & REPLACING EMPTY STRINGS WITH NaN

Now we have to convert numerical columns to their specified datatype which were considered as object datatype.

```
] import re

# Clean and extract the numeric part from 'Year_of_release' using regular expressions
df['Year_of_release'] = df['Year_of_release'].str.extract('(\d+)').astype(float).astype('Int64')

# This code will extract the numeric part and handle missing values with 'Int64' data type

# Replace non-numeric and empty string values with NaN
df['Metascore'] = pd.to_numeric(df['Metascore'], errors='coerce')

# Convert the column to an integer data type, handling NaN values
df['Metascore'] = df['Metascore'].astype(pd.Int64Dtype())

# Clean the 'Votes' column by removing non-numeric characters and leading/trailing spaces
df['Votes'] = df['Votes'].str.replace(r'^\0-9', '', regex=True).str.strip()

# Convert the cleaned column to integer
df['Votes'] = df['Votes'].astype(int)

# Remove non-numeric characters (e.g., currency symbols, commas, and other non-numeric characters)
df['Grosses'] = df['Grosses'].str.replace(r'^\0-9.', '', regex=True)

# Replace empty strings with NaN
df['Grosses'] = df['Grosses'].replace('', np.nan)
```

Fig 3.A.1 Code Snap for Data Type Conversion and NaN replacement

B. CHECKING FOR NULL VALUES and DATA TYPE



Fig 3.B.1 Matrix Plot

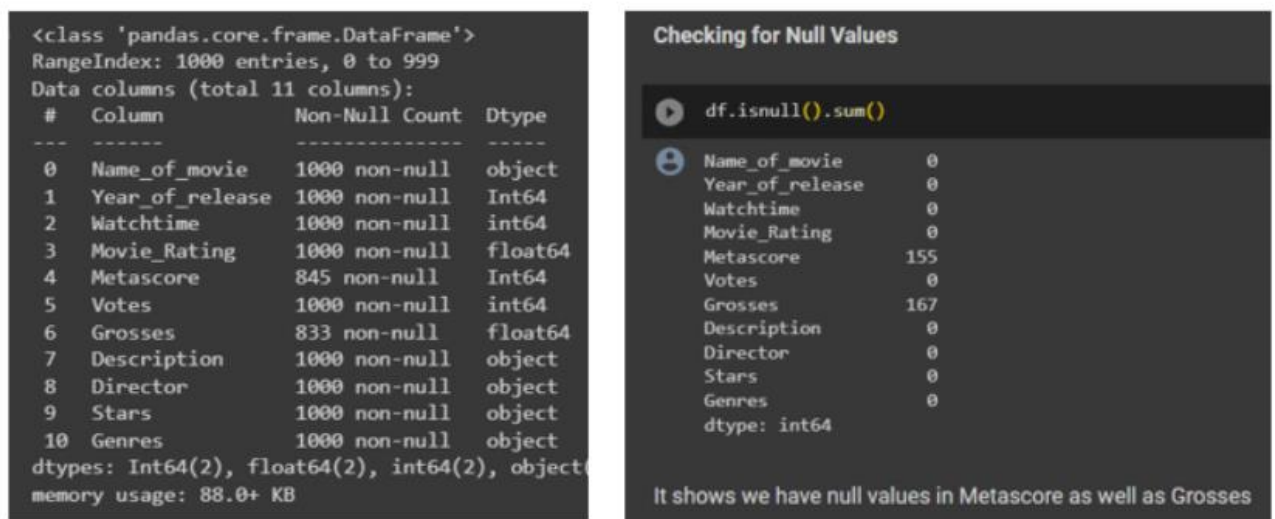


Fig 3.B.2 Total null values and Data types for respective columns

C. OUTLIER ANALYSIS

We are going to fill the values of both 'Grosses' and 'Metascore' on the basis of 'Votes' and 'Movie_Rating' respectively. So in order to decide the imputation

technique whether we should use mean / median , first we have to check the no. of outliers present in Votes.

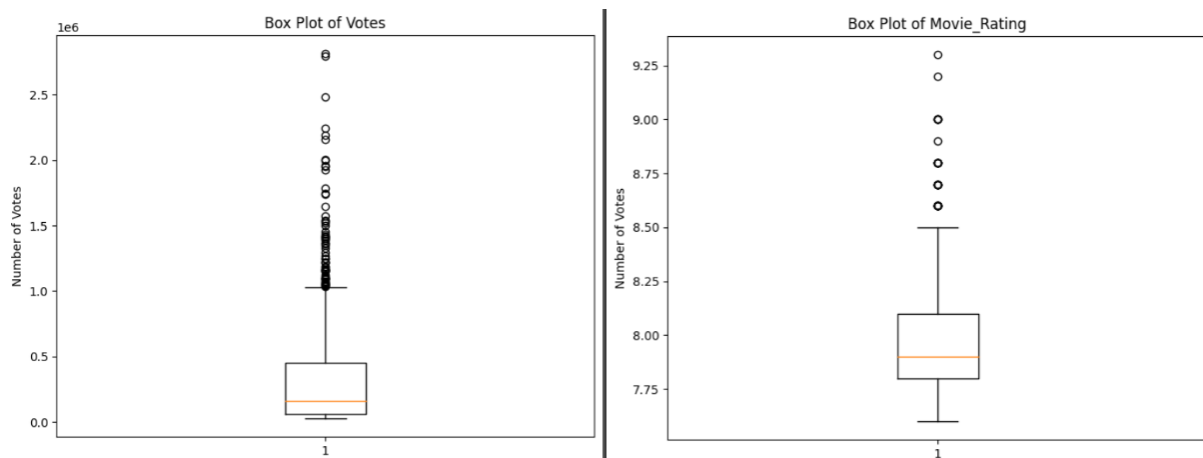


Fig 3.C.1 Box plot

D. NULL VALUE IMPUTATION

As there are lot of outliers present in votes so we are going to use median as an imputation as it is robust and does not get affected by no. of Outliers.

And there are less no. of Outliers in Movie_Rating so we can use the technique of mean for filling up the Null Values.

```
# Replace NaN values in 'Grosses' with the median of the 'Votes' column
median_Votes = df['Votes'].median()
df['Grosses'].fillna(median_Votes, inplace=True)

# Ensure 'Metascore' is of a floating-point data type
df['Metascore'] = df['Metascore'].astype(float)

# Replace NaN values in 'Metascore' with the mean of the 'Movie_Rating' column
mean_rating = df['Movie_Rating'].mean()
df['Metascore'].fillna(mean_rating, inplace=True)
```

Fig 3.D.1 Null value Imputation

E. ANALYZING & DELETING UNNECESSARY COLUMNS IN OUR DATASET

We will remove the Watch time , Description as it has no use for our Gross prediction.

```
df = df.drop(['Watchtime', 'Description'], axis=1)
```

Fig 3.E.1 df.drop

F. STANDARDIZE VALUES

Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

$$z = (x - \mu) / \sigma$$

Now we are going to initialize a `StandardScaler` to standardize the training data (`X_train`) and then applies the same transformation to the testing data (`X_test`). Standardization scales the data to have a mean of 0 and a standard deviation of 1, making it suitable for many machine learning algorithms.

```
# Create and fit a StandardScaler to standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Fig 3.F.1 Standard Scaler

4. EXPLORATORY DATA ANALYSIS(EDA)

Exploratory Data Analysis (EDA) is crucial for movie box office prediction as it offers insights into the dataset's characteristics, helping identify trends, outliers, and relationships among variables. EDA aids in data preprocessing, feature selection, and the creation of informative features. It allows for the visualization of patterns and helps choose appropriate models. EDA helps ensure that data is properly prepared, enhancing model performance, and enables better understanding of the factors influencing box office earnings, improving the accuracy of predictions. This process is fundamental for making informed decisions and building reliable predictive models in the movie industry.

A. HEAT MAP

A heatmap is essential for visualizing the relationships between multiple variables in a dataset. It is particularly valuable when dealing with a large dataset with numerous features.

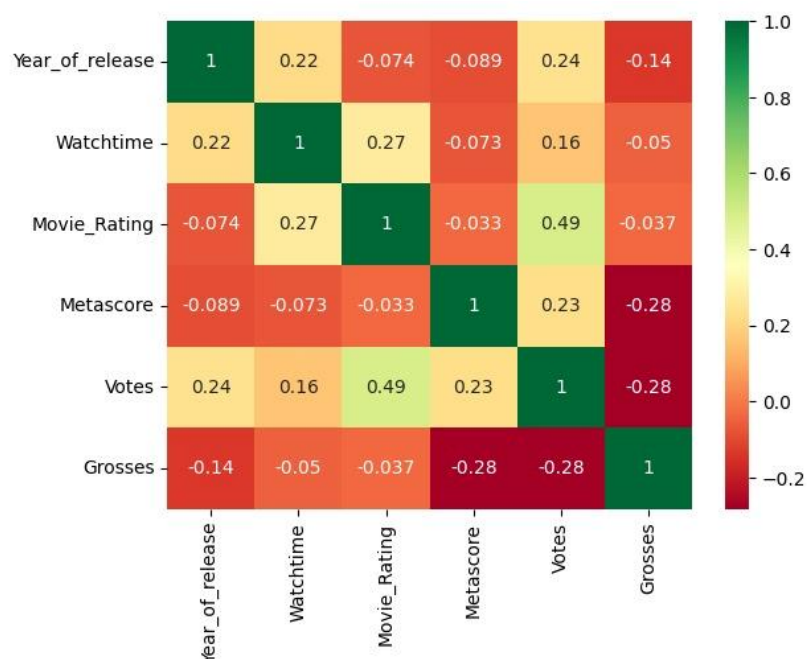


Fig 4.A.1 Heat Map for correlation

B. BOX PLOT

Our one of the target variables is 'Movie Rating' to we need to analyze it carefully . First we need to know about is statistical measures like no. of outliers , median , Interquartile Ranges from that we can also came to know about the name of the movies with highest Movie_Rating as well as lowest Rating this can be done using BOX PLOT .

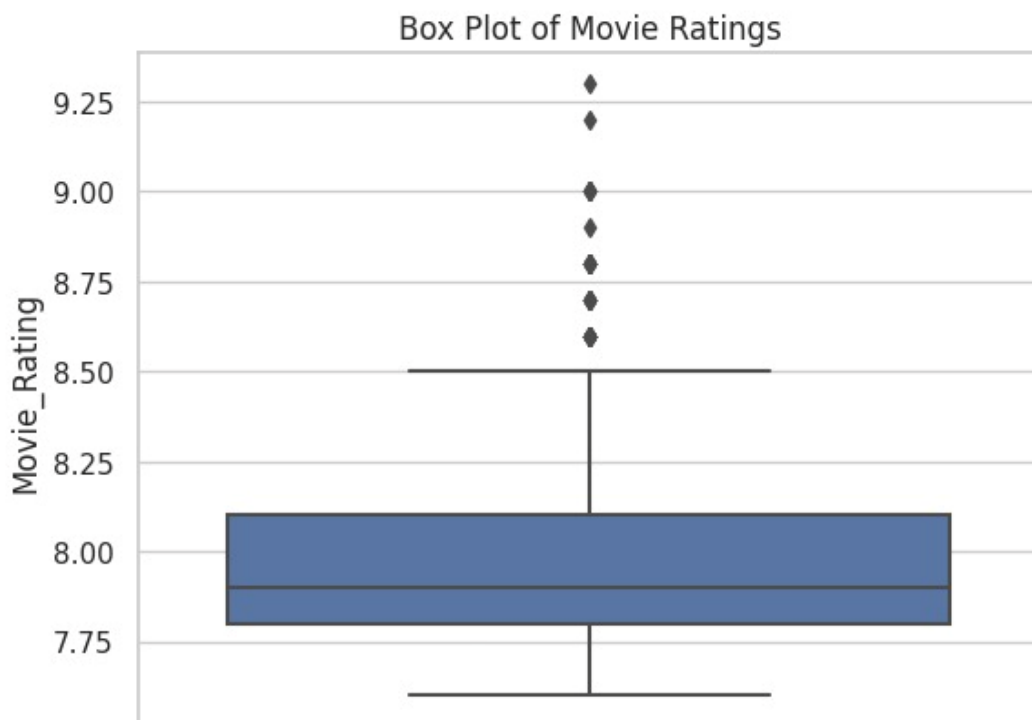


Fig 4.B.1 Box plot for Movie Ratings

Inference: We can see from the box plot for Movie_Rating:

- i) There are total 7 Outliers
- ii) Highest Movie Rating is 9.3
- iii) $Q1 = 7.8$
- iv) $Q2(\text{Median}) = 7.9$
- v) $Q3 = 8.1$
- vi) Movie rating > 8.54 and < 7.35 are considered as Outliers

C. SCATTER PLOT

Bivariate analysis using a scatter plot allows us to examine the relationship between two variables, in this case, Metascore and Votes. Metascore, assigned before a movie's release, reflects critical reception, while Votes represent viewer responses. By plotting Metascore against Votes, we can visually assess if there's a correlation. If there's a positive correlation, it implies that well-received movies by critics tend to attract more viewers. Conversely, a negative correlation suggests that critical acclaim may not always translate to popularity.

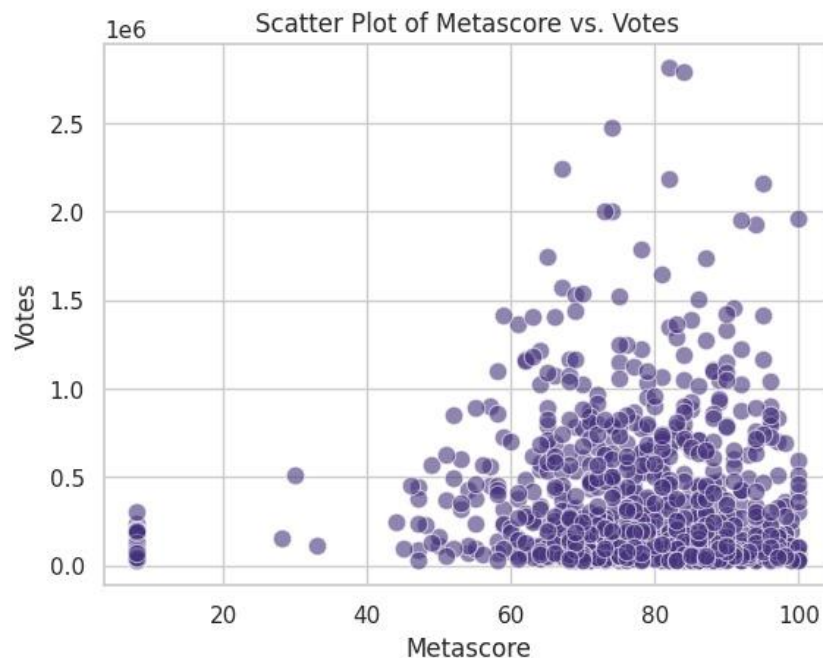


Fig 4.C.1 Scatter Plot plot for Metascore vs Votes

Inference: So it shows that More the Meta score more will be the Votes are less positively correlated which also justifies our heat map where are correlation was 0.23 i.e they are less positively correlated there are some movies where votes increases with Metascore while majority of movies have more Metascore but less Votes.

D. HISTOGRAM

Analyzing the distribution of movie ratings is essential to gain insights into the dataset's overall movie quality. This examination provides a clear overview of the most common rating ranges and any potential outliers. It aids filmmakers, production companies, and distributors in understanding the preferences of audiences and can be used to tailor marketing and content strategies.

Additionally, it offers a foundation for identifying patterns in audience reception, which can influence future film projects and their success in the industry.

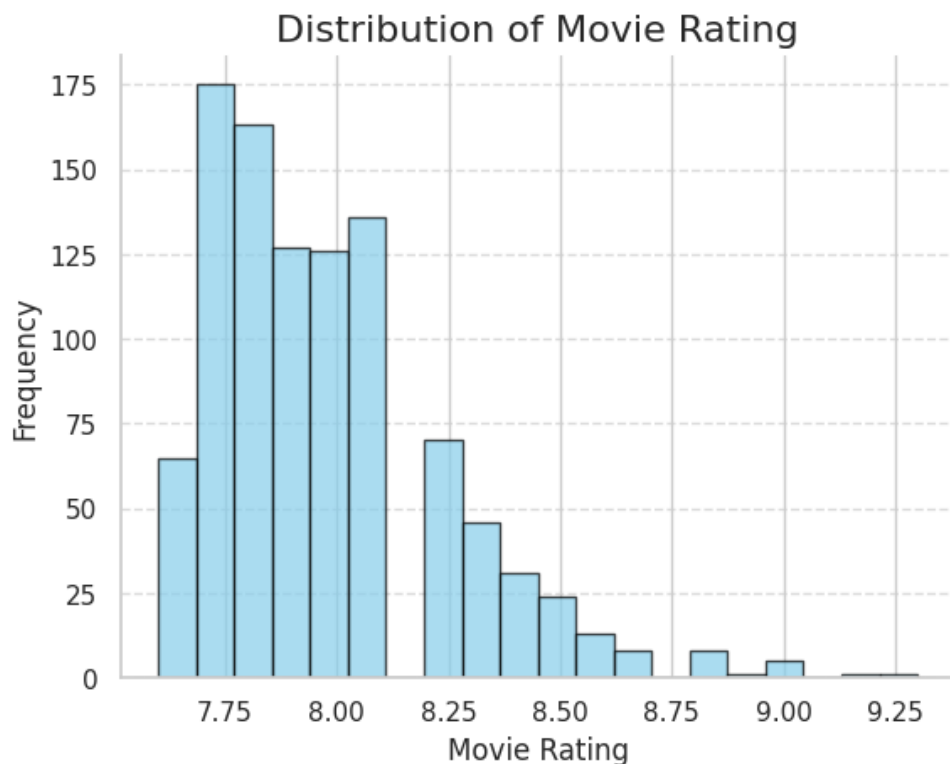


Fig 4.D.1 Histogram for Movie Ratings

Inference: So we can see that movies with Movie Rating are greater than 5 . So it justifies that we have scraped top 1000 IMDB movies and our dataset is more concentrated towards good Movies . This is also positively Skewed i.e most of the data is concentrated on the left tail of the graph.

5. METHADODOLOGY OF IMPLEMENTATION

As our project mainly focuses on EDA and DPL so detailed analysis has been included however a Machine Learning has also been created to predict Movie Box office Predication which included the following step:

- i) Data Acquisition using Web scrapping of IMBD website.
- ii) Data Preparation was done by cleaning and manipulating the data.
- iii) Exploratory Data Analysis was done in order to analyze the data and concluding important inferences from that.
- iv) Model building was done by first splitting the data , standardizing it . We used Random Forest Classifier for prediction .
- v) Model Evaluation was Done by calculating Mean Absolute Error , Mean Squared Error , Root Mean Squared Error and R-squared on test set.

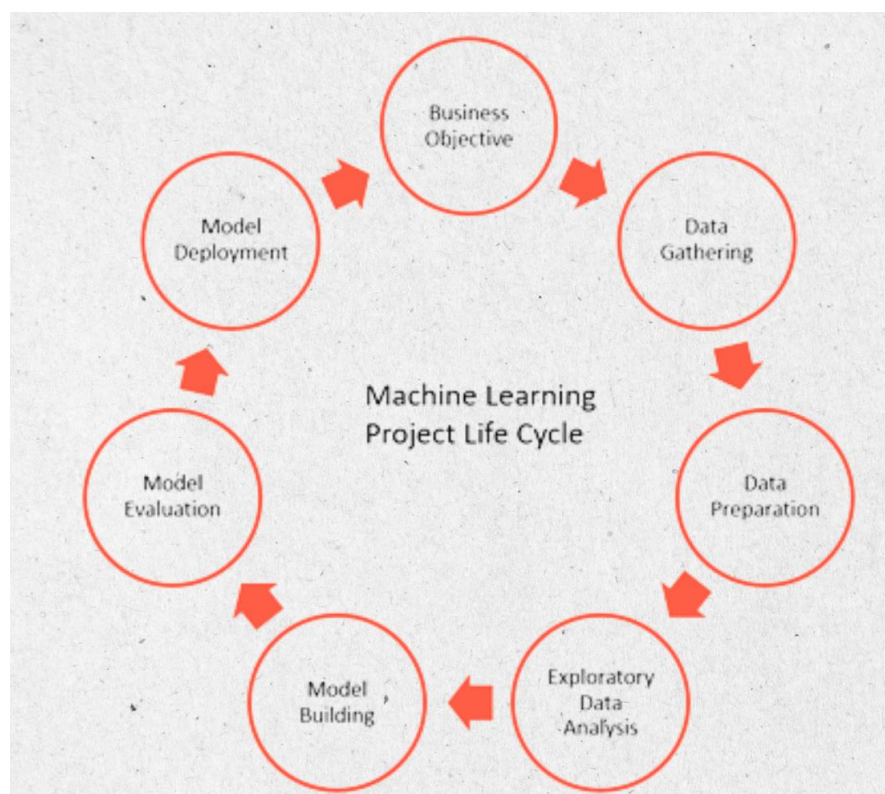
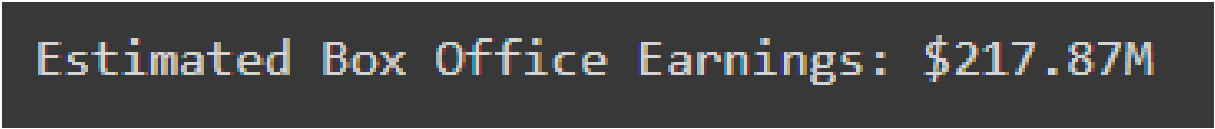


Fig 5.1 Machine Learning Life Cycle



Estimated Box Office Earnings: \$217.87M

Fig 5.2 Final Output of the model

6. JUSTIFICATION

In a world where decisions are driven by data, our project bridges the gap between filmmaking and analytics of data science. It's not just about predicting box office earnings; its also about understanding the art of cinema.

In order to sum our model for predicting movie grosses is the result of our most significant efforts to utilize information to commemorate the love of films . We handed IMDb data into an approachable tool that provides filmmakers and movie lovers more power. With vigilant model engineering, efficiency optimizing and data planning, our project is now prepared to provide services to the film industry. It symbolizes our contribution to ensure that detailed data - driven insights are available to everybody and it is beyond a prediction model .

CONCLUSION

We will keep growing and enhancing and also improving this project, open to new data sources , leading advancements in field of data science and the film industry . it is lot more than jut a prediction model - its a data which demonstrates what drives us for movies.