**MOVIE SUCCESS PREDICTION**

# Project Report

# Bachelor of Technology

(Computer Science & Engineering)

SUBMITTED BY

**V. SAI SUMANTH(210257)**

**K. SRI TEJ VISHNU(210237)**

**K.SHREYAS KUMAR(210340)**

UNDER THE SUPERVISION OF

**DR. KIRAN SHARMA**

School of Engineering and Technology

Logo, company name

Description automatically generated

# BML MUNJALUNIVERSITY Gurugram, Haryana - 122413

# Dec 2022

**CANDIDATE’S DECLARATION**

We hereby declare that we have undergone six months Project 1 at BML Munjal University and worked on project entitled, “**Movie Success Prediction**”, in partial fulfillment of requirements for the award of Degree of **Bachelor of Technology** in name of the department at **BML Munjal University**, having University Roll No.1232434, is an authentic record of my own work carried out during a period from July, 2022 to December, 2022 under the supervision of Dr. Kiran Sharma.

## V. Sai Sumanth

**K.Sri Tej Vishnu**

**K.Shreyas Kumar**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

## Dr. Kiran Sharma

Person from Organization with designation

The movie industry is no longer just an industry or a center of entertainment; it is also a center of global business. Everyone in the world is now excited about a movie's box office success, popularity, and so on. There is a wealth of information available online about the success and popularity of these films. The film industry is an important sector in the global market.

As a result, it is important to increase profits by predicting movie success before its release. The number of movies released all around market is growing and the success rate of each film is important because huge amounts of money are invested in its production. In such cases, awareness about a movie's success or failure and what factors influence a movie's success will benefit production houses because these predictions will give them a promising idea of how to handle marketing, which is an expensive affair in and of itself. As a result, predicting the success of a film is critical to the film industry.

As a result, predicting the success of a film is critical to the film industry. We proposed using methods of machine learning and algorithms to build a model for trying to predict if a movie will be a failure or long before it is released.

We am highly grateful to **Dr. Kiran Sharma,** BML Munjal University, Gurugram, for providing supervision to carry out the project-1 from July-December 2022.

**Dr. Kiran Sharma** has provided great help in carrying out my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the training in this manner.

We would like to express thanks profusely to thank **Dr. Kiran Sharma**, for stimulating me time to time. We would also like to thank entire team of BML Munjal University. We would also thank my friends who devoted their valuable time and helped me in all possible ways towards successful completion.

## **V. Sai Sumanth**

**K. Sri Tej Vishnu**

**K. Shreyas Kumar**

**Figure No. Figure Description Page No.**

Fig(4.2.1) Information of columns data description . . . . . . . . . . 6

Fig(4.3.1) Data format representation . . . . . . . . . . . . . . . . . . . . . 7

Fig(4.3.2) Bar plot of type of movie released in each year . . . . . 8

Fig(4.3.3) Line plot of average budget in each year . . . . . . . . . . 8

Fig(4.3.4) Line plot of average gross in each year. . . . . . . . . . . . . 9

Fig(4.3.5) Bar plot between budget and gross revenue

in each year . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

Fig(4.3.6) Bar plot of total no. of movies in each genre. . . . . . . . 11

Fig(4.3.7) Bar plot of total no. of movies for each main lead. . . . . 11

**Abbreviation Full Form**

**EDA** Exploratory Data Analysis

**ML** Machine Learning

**SVM** Support Vector Machine

**NLP** Natural Processing Language

**IMDb** Internet Movie Database

**OS** Operating System

## Contents PageNo.

***[Candidate’s Declaration](#_bookmark0)* i**

[Abstract](#_bookmark1) ii

[Acknowledgement](#_bookmark2) iii

[List of Figures](#_bookmark3) iv

[List of Abbreviations](#_bookmark5) v

1. [Introduction to Project](#_bookmark7) 7
2. [Literature Review](#_bookmark12) 8
   1. Comparison . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
   2. Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

3.3 [Objectives of Project](#_bookmark14) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

1. [Exploratory Data Analysis](#_bookmark15) 10
   1. [Flow Chart](#_bookmark24) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
   2. [Description of Dataset](#_bookmark16). . . . . . . . . . . . . . . . . . . . . . . . . 10

4.3 [Exploratory Data Analysis and Visualizations](#_bookmark17) . . . . . . . . . . . . . . . 11

1. [Methodology](#_bookmark19) 13
   1. [Introduction to Languages . . . . . . . . . . . . . . . . . . . . . . .](#_bookmark20) . . . . . . . . . . . 14
   2. [Any other Supporting Languages/ packages](#_bookmark21) . . . . . . . . . . . . . . . . 14

[5.3 Flow Chart](#_bookmark24) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

[5.4 L](#_bookmark29)ibraries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .16

[5.5 Implementation](#_bookmark30)  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

1. [Results](#_bookmark31) 17
2. [Conclusion and Future Scope](#_bookmark32) 19

[7.1 Conclusion](#_bookmark33) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

[7.2 Future Scope](#_bookmark34) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

**8 Bibliography 20**

# 

# Introduction

Movies have the potential to control lifestyle on an international and domestic level. Movies are no longer just a source of entertainment and major source of global business and marketing. Movies create a new trend among people, especially among young people. Not only are film makers and movie box - office executives worried about success of their movies, but so are the public in general. People used to talk these on social media.

Movies can be found online. People can share their movie reviews on websites such as IMDb, Rotten Tomatoes, Metacritic, and others. In­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­any part of the world, movies continue to stay a major source of entertainment. However, if the movie fails to do well at the box office, this industry could face many flops. Our project will attempt to predict the movie's success rate by conducting prediction on the movie's various features. People are increasingly using these platforms because they provide honest feedback. As a result, a wealth of data about movie ratings and reviews is available online. As a result, predicting a film's success is critical in the film industry.

This prediction can help actors, producers, and directors decide more effectively. They will have the opportunity to make a choice even before movie is released. This proposed effort aims to build a model using data mining techniques that can predict the success of the movie in beforehand, thus improving outcomes. IMDb is a tremendous resource for finding specific info regarding all of movie ever produced. A massive quantity of data that gives a variety of useful information about general movie trends. Data mining techniques enable us to dig up data that confirms or dismisses big movie misconceptions, as well as predict the success of a future movie based on minimal awareness about the movie prior to its theatrical release

# Literature Review

Much research has been done on predicting movie success with machine learning and algorithms using many languages like Java Script, R programming, Python, and all using meta data like social media data, cast and crew whose focus is whether the model working is helpful for the prediction and all.

# 3.1 Comparison

**R. Dhir and A. Raj -** The research paper “Movie Success Prediction Using Machine Learning Algorithms and their Comparison,” is discussed based on some released features, SVM, NLP, and neural about how they have been used and how the prediction is calculated and how it is based on the variety criterion such as Rotten Tomatoes, IMDb vote count, number of displays, expense, and box office.

**Jeffrey Ericson and Jesse Grodman -** The report “A Predictor for Movie Success” discussed about which are the main attributes need to be known for predicting the success, measures need to be taken and the input values which can be influenced by producer before it’s launch and all. Using Java Script and python as languages building a model and changes in trends over the years is discussed in the report.

**Muhammad Hassan Latif and Hammad Afzal -** The research paper “Prediction of Movies popularity Using Machine Learning Techniques” discussed about it’s plans and processes to predict movies popularity by using ML as main method. Having a goal to check the reasons behind the failures of many prediction models and classifying the reasons behind it’s failure and ways to improve the model is all included in it. The data for this article was taken from social media. The article claims that there is a correlation between social media content and box office receipts. So, they predicted the box office using the Linear Regression and Support Vector Regression techniques. Additionally, they employed both linear and nonlinear regression, depending on how much popularity a given video attained based on user comments and postings.

# 3.2 Problem Statement

In the movie industry, accurate analysis of movie success is required, that benefits various people working in the industry, particularly investors. Many people would prefer to invest their time and money in knowing about movies, while others would prefer to invest in movies (distributors, satellite, and digital supporters) to avoid losses in the future.

Catching the genres which attract audience having good success , knowing the cast and crew with high success by analyzing past data and finding audience interest by analyzing recent movies with high success are the major problems for need to be solved in proper way which helps in making better movies.

# 3.3 Objectives of Project

Our objective is to develop a mathematical model that tells the probability of the success of movie being a Flop or Hit and a model which predicts box office before a movie is released using machine learning techniques and algorithm

# 4.Exploratory Data Analysis

# 4.1 Flow chart

Raw Data Collection

Data Processing

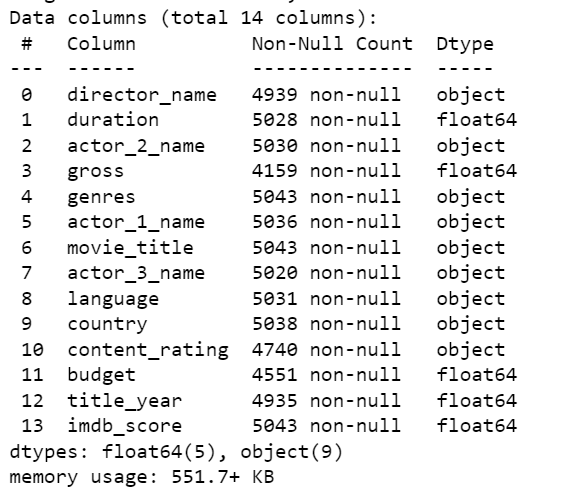
Data Cleaning

EDA

Modelling

# 4.2 Dataset Description

Movie Meta dataset contains Movie Name, Year, duration, genre, rating, votes including cast and crew director, actor 1, actor 2, actor 3 and many other elements. Analyzing the meta data for better view of data and for easier understanding so that in the future we can predict the outcome using the analysis we did here. We analyzed the dataset using anaconda jupyter notebook.



Fig(4.2.1):Information of columns data description

# 4.3Exploratory Data Analysis and Visualizations

# With the CSV data file which was gathered from Kaggle the data has been brought into jupyter notebook after importing necessary libraries, it undergoes some processing before being cleaned and then EDA is then performed. As the EDA is an iterative process we can re process and re clean the required data any time during EDA and use the cleaned data set and knowledge from EDA to perform modelling.

# 

# Therefore, the objectives for doing the EDA are as such:

# To check the quality of data for further processing and cleaning if necessary.

# To get better view about the metadata inside it which further help in modelling.

# We are using movies\_data.csv from Kaggle to illustrate the concept of ratings of Indian movies from many years. Focusing our project objectives and problem statement is to forecast the different type of meta data influencing the ratings of movies ,our objectives are therefore:

# To check the changes in meta data in the timeline; and

# To check the impact of meta data, cast and crew on the movie.

# Our code template shall perform the following steps:

# Preview data:

# The necessary dependencies like python libraries and start off with generating a simple preview and statistics of ratings data set which is known as Preliminary Data Processing.

 Table

Description automatically generated

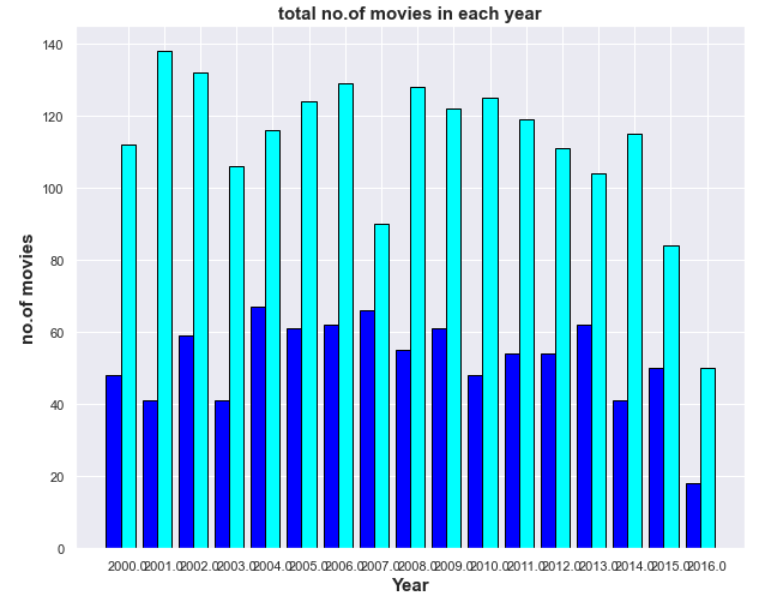
Fig(4.3.1):Data format representation

# Check any null values:

# As in preview data we found out that there are any null values so there is no need to check us any further for finding any null values again and for risking in future we tried to drop null values using Boolean if there are any null values it shows true and drop them. So as pre-processing and cleaning of data is done, we can start doing graphical EDA.

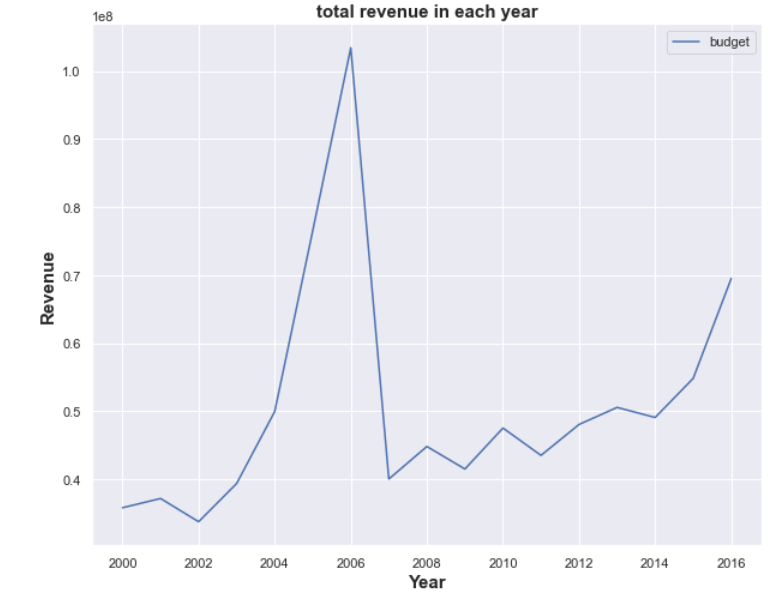
# Time series EDA:

# We planned to represent the total number of movies released in each year but dividing into two parts number of hits ( movies with IMDb score more than 7) and number of flops (movies with IMDb score less than 7). X- axis of the graph shows year and Y-axis shows the number of movies count and the blue bar in the graph shows the hit movies and aqua bar shows the flop and average movies according to IMDb score.

Fig(4.3.2):Bar plot representing type of movies released in each year

# The money rotation in each year can be viewed using line plot so we decided to plot average budget in each year and average gross in each year. Here the X- axis is taken as year in both graphs and Y-axis values are according to their data

# Coming to average Budget representation in each movie the Y-axis values are reached to max limit so it is shown with values between 0 to 1 which would be multiplied with 1e8 ( 1 multiplied by 108)



Fig(4.3.3):Line plot of average budget in each year

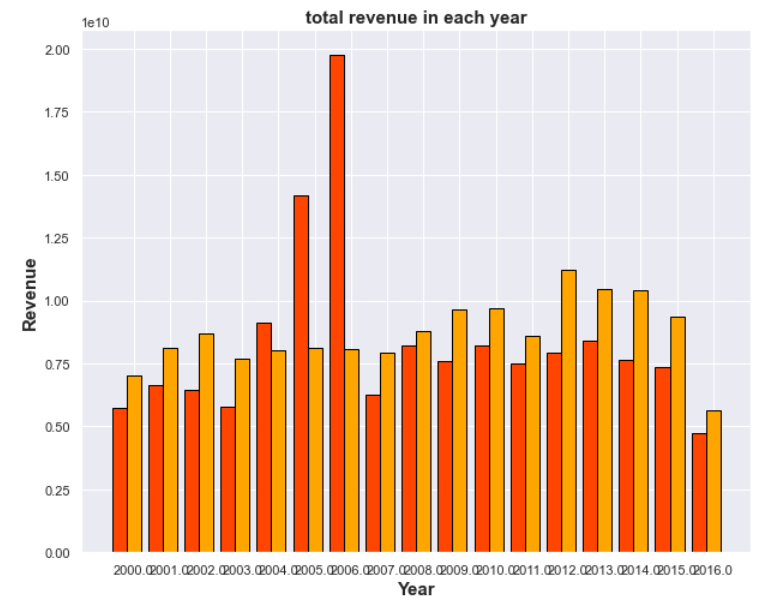
# Coming to average gross representation in each movie the Y-axis values are reached to max limit so it is shown with values between 0 to 1 which would be multiplied with 1e7 ( 1 multiplied by 107)

Chart, line chart

Description automatically generated

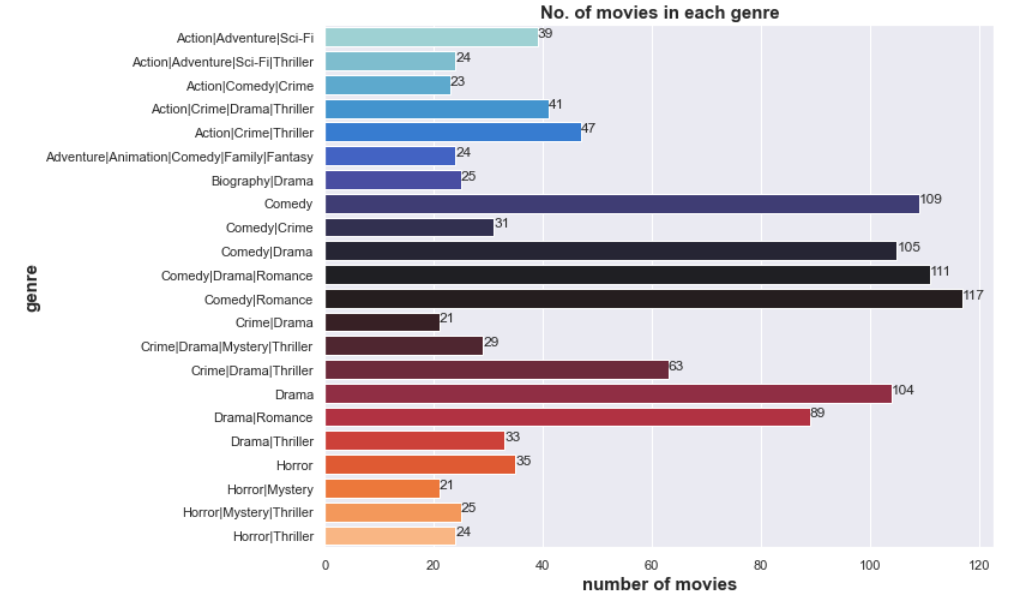
Fig(4.3.4):Line plot of average gross in each year

# The total budget and gross comparison are being to be compared using bar plot which would help us to understand the difference between their values using graph. So, we did the sum of each movie gross and budget in each year and used orange to represent gross and orange red to represent budget for easier understanding. The x-axis is the year and Y-axis is the total revenue in which the values are in range of between 0 to 2 which are multiplied with 1e10 (1 multiplied by 1010 )



Fig(4.3.5):Bar plot between budget and gross revenue in each year

Showing the relation between numerical value and categorical value on the plane is helping us to convey the data in better format is the sole reason to bar plot.We decided to count number of movies released and also number of hits and flop in each genre in each meta data like genre, director, actor 1,actor 2,actor 3 so here we shown genre graphs which we had done as example:



Fig(4.3.6):Bar plot of total number of movies in each genre

Chart, bar chart

Description automatically generated

Fig(4.3.7):Bar plot of total number of movies for each main lead

# Movie Rating Model

# INPUT

After analyzing and doing further processes we designed code so that it takes input one- on- one in this way

# 

# 

# 

# Table Description automatically generated

# OUTPUT

With the given input into variable, it checks if the given values are in data and according to that it takes the success percent and calculate to give the required output.

# 

# Box Office Model:

# INPUT

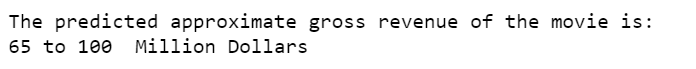
We give the crew and cast along with meta data of the movie in the input so that it analyse the data.

# 

# OUTPUT

After encoding the input, the algorithm applies and predict the results to give the output.





# 5.Methodology

The proposed methodology addresses the project's various stages, which include

data-collection, data preprocessing, generating training and testing datasets, model generation, prediction, and outcomes.

* **Data Analysis:**

In data analysis, all selected attributes are analyzed on the basis of different factor that help us to gather most accurate outcome for further stages. Selected features for analysis are some of the attributes for greater view on the meta data set.

* **Data Collection:**

As per project requirements IMDb movie data set and revenue data set of movies in recent years are required in a proper format for analyzing and preprocessing it.

* **Data Preprocessing:**
* In this stage dataset is prepared for applying data mining technique.
* Before applying data mining technique, pre-processing methods like cleaning, variable transformation and data partitioning and other techniques which are required as per data is done for further processes.
* **Generating Training and Test Dataset:**
* Training dataset is a set of attributes used to fit the parameters of the model.

# 5.1 Introduction to Languages

Python language is used to build this model and whole model, analyzing part done using anaconda jupyter notebook using python 3 version.

Python is an object-oriented, interpreted programming language. Modules, exceptions, dynamic typing, very high-level dynamic data types, and classes are all included. It supports a variety of programming paradigms other than object-oriented programming, including procedural and functional programming.

# 5.2 Any other Supporting Languages/ packages

Common libraries like pandas, NumPy and we used SKLEARN and using SKLEARN we imported many libraries package like Count Vectorizer linear model and many more for the data mining the data and building the model .

We imported the tree from SKLEARN for representation of the analyzed structure and build a statistical of success model using importing math package. Using SKLEARN we imported many packages which further explained

# 

# 5.3 Flow Chart

## 

**START**

**DATA COLLECTION**

**IMDb DATASET**

**DATA PREPROCESSING**

**GENERATING TRAINING DATASET**

**DATA ANALYSIS**

**MODEL GENERATION**

**MODEL GENERATION**

**PREDICTION**

**PROBABITITY USING STATS**

## 

**RESULTS/ OUTCOME**

# 5.4 Libraries

## SKLEARN:

SKLEARN is a Python-based machine learning library that is free to use. It includes support-vector machines, random forests, gradient boosting, k-means, and regression, and clustering algorithms, and it is designed to work with the Python statistical and systematic libraries NumPy and SciPy.

## LABEL ENCODER:

Label encoding is the process of converting labels into a numerical form so that they can be read by machines. Machine learning algorithms then can make better choices regarding how those labels should be used. In supervised methods, it is an important pre-processing step for the formalized dataset.

## MATH LIBRARY:

Math is a basic Python 3 library module that contains standard mathematical exponent and operations. The math component allows you to do a variety of mathematical calculations, which include numeric, exponentials, log scale, and infinite calculations

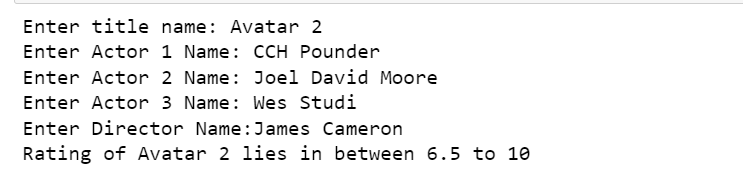
* 1. **Implementation**

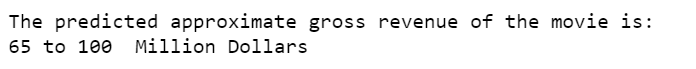
Two Datasets Stage2lower.csv and WithoutGrossLower.csv are taken to build model as the values in each column is stored in variable which in future when we enter the values it checks if the input value is in variable if it’s true it calls the values and implement the method which is given into model.

The statistical model which we built also follows in this process which gives the needed output like the chances of success and chances of collection if the possible rating is entered to model. With the above main libraries, it works in this process

# 6 Results

After entering the meta data and data of cast and crew as input it gives the output after analysing the data given to it earlier so we get the possibility of the ratings which would come to the movie.Coming to box-office model it calculate the prediction which takes the meta data and other values for good accuracy and perfect outcome.





# 7 Conclusion and Future Scope

# 7.1 Conclusion

Two different system models using the static factors i.e the trating model and the box office model which are the used for the prediction of success of the movie as described.The main component like the meta data and the revenue of the movie are taken from kaggle.

We like to conclude that this project has given us great insight and experience on exploring and analysing the data. There many types of libraries which are used in building a prediction system. Using the earlier analysis and libraries we tried to build a prediction system to get a desired outcome of the success of the movie.

# 7.2 Future Scope

* Trying other ML algorithms like Random Forest , Naive Bayes and many more and checking which have more accurate results.
* Designing front end by building a website for giving input instead of entering input in terminal.

# [1] Muhammad qHassan iLatifz andxHammadaAfzal,x“Prediction of Movies popularity fUsing sMachine aLearning qTechniques*” tInternational rJournale of Computer uScience andwNetwork Security,w*VOL.16 No.8,/August 2016.

[2]JeffreyrEricsonsandwJesse4Grodman,5“A6Predictor0foriMovieqSuccess”sCS229, StanfordyUniversity,gUSA,eSeptembert2015.

# [3][RijulgDhir](https://ieeexplore.ieee.org/author/37086827236)5and1[Anand2Raj](https://ieeexplore.ieee.org/author/37086827496), “Movie Success Prediction using Machine Learning4Algorithms2and1their2Comparison*” International0Conference9on8Secure Cyber5Computing7and8Communication,*u15-17jDecemberk2018.

[4] N. Quader,zM. O.eGani, D.wChaki and M. H. Ali, "A machine learning approach to predict movie box-office success," *20th International Conference of Computer and Information Technology*, Dhaka, 2017.

[5] Kanitkar, "Bollywood Movie Success Prediction Using Machine Learning Algorithms," *3rd International Conference on Circuits, Control, Communication and Computing*, Bengaluru (Bangalore), India, 2018.

