**A**
**Project Report**
**on**

**Big Data Analytics**

# MOVIE DATA ANALYSIS AND PREDICTION

**Submitted by-**


***Kosana Sai Venkata Pavan Kumar***

***1800230C203***

***CSE – 2018***

***Goli Abhilash***

***1800218C203***

***CSE – 2018***


**Under the guidance of**

**Dr. Yogesh Gupta**
Associate Professor



**Department of Computer Science and Engineering**
**SCHOOL OF ENGINEERING AND TECHNOLOGY**
**BML MUNJAL UNIVERSITY GURGAON-122413, INDIA**


***May, 2021***

# Acknowledgement

On the accommodation of my project report, I want to expand my sincere thanks and my earnest gratitude to my professor Dr. Yogesh Gupta sir, Department of Computer Science and engineering, for providing me chance to do this project under his guidance and giving significant direction all through this exploration. He has demonstrated the techniques to make the exploration and to present this work as clearly as possible. It was an implausible advantage and regard to work under his direction. I am amazingly thankful for what he has explained me in the classes and offered me. Finally, I am grateful to my parents and friends for their motivation and support throughout my career.

Thanking you
Goli Abhilash & Kosana Sai Venkata Pavan Kumar
CSE – 2018 Batch

# INDEX OF THE REPORT

# Abstract

Big Data in Movie Analysis is very useful which allows us to analyze and predict the revenue of the model with greater accuracy, and eliminating the guesswork often involved. The main objective of this project is to analyze and build a machine learning model that predicts the revenue of the movie. We have taken the dataset from Kaggle which contains details of 3000 movies with the data about title, actors, budget etc.

In this project, the dataset is analyzed, visually represented, and trained with two different classification algorithms. They are Ridge regression and Random Forest. In Evaluation, RMSE scores are calculated for each algorithm and chosen the best one according to these scores. Finally, we have predicted the revenue of films in the dataset which has was not attached with the revenue.

# Motivation

Movie production, distribution is also a kind of business. Many people involved in these businesses are getting huge losses because of their incapability of prediction of the movie revenue. So, by considering the movie cast, crew, its prequel, popularity etc, the revenue of the movie is predicted which helps in better estimation to the producer and the distributor to be in the part of venture or not. The main aim of this project is to build a machine learning model to predict the revenue of the film.

## 1) INTRODUCTION

Big Data is a collection of data that is huge in volume, yet developing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. It is also data but with huge size. It is also used in predictive modeling, and other advanced analytics applications. Big data analytics give entertainment business decision-makers the data-driven insights they need in order to help their film companies compete in a highly competitive market. This data allows them to identify realistic goals and shows them how to meet them. .


*Fig 1: Big Data on Movies*

In this project, a machine learning model is being developed by using Ridge regression and Random forest algorithm. The main aim is to create a machine-learning model that can forecast the sales of a new film based on budget, release dates, genres, production firms, and production countries.We have taken the dataset from Kaggle which contains details of 3000 movies(between 1960 to 2017) consisting of all the information regarding each and every movie(like cast, crew, budget, popularity, date, Genre etc..). It helps in decreasing the loss to producers who is gona make a film.

## 2) PROBLEM STATEMENT

As we see in our daily life many people are getting losses by investing huge amounts in movies. They are getting huge losses because of their incapability of prediction of the movie revenue. So, by considering the movie cast, crew, its prequel, popularity etc, the revenue of the movie is predicted which helps in better estimation to the producer and the distributor to be in the part of venture or not.

## 3) LITERATURE REVIEW

We have done a lot of literature review on the similar movie revenue prediction projects. We have got some of the inters existing projects.

### [1]. Early Prediction of movie Box office success Based on Wikipedia Activity Big Data:

They presented the results of developing a simple statistical model for movie financial performance based on internet users' cumulative activity data.

By calculating and evaluating the activity level of editors and viewers of the corresponding entry to the movie in Wikipedia, the well-known online encyclopedia, they demonstrated that the success of a movie can be predicted much before its publication.

### [2]. Predicting Movie Revenue

Cast, budget, film critic reviews, MPAA ranking, release year, and other considerations all influence movie sales. Because of many of these variables, there is no analytical method for estimating how much money a film would make. However, by studying the revenue created by previous films, they can create a model that can assist us in predicting a film's projected revenue.

| S.No | Existing State of Art | Drawbacks in Existing State of Art | Overcome |
|---|---|---|---|
| 1 | The popularity of a movie can be predicted much before its release by measuring and analyzing the activity in encyclopedia | The only considered data is the popularity among encyclopedia users. | Few other important parameters are also considered for movie prediction like cast, crew, genre, budget etc. |
| 2 | The first week-end sales, budget, and number of theatres on opening week were used as parameters in a linear regression | They used K-means clustering with two clusters to isolate the results, with only the number | Using only the number of theatres as a metric equivalent to popularity is not |

| | | | |
|---|---|---|---|
| | algorithm. The global box office income is the linear algorithm's output. | of theatres as their dataset for the algorithm. | better we used the dataset containing in dataset. |

*Table 1: Existing state-of-art*

## 4) METHODOLOGY

This project comprises three steps. They are Pre-processing, and Model implementation, and Evaluation. Internally, each step contains a different process. Each process is explained clearly in this project. They are:

### 1) Pre-processing:

Data Preprocessing is a basic step in the process of building a model. It consists of several steps like Data collection, Data cleaning, Data analysis, Handling categorical values. A good data provides better performance in the model implementation and evaluation. The steps involve in this stage are:

#### i. Data collection:

The dataset is collected from the Kaggle website which contains details of 3000 movies ranging from 1960 to 2017. It consists of information like movie cast, crew, its prequel, popularity, budget, Genre etc.. Below we can see the image of the dataset taken.

```
[ ] data = pd.read_csv("/content/drive/MyDrive/BDA project/train.csv")
    print(data.shape)
    data.head(n=2)
```

(3000, 23)

| | id | belongs_to_collection | budget | genres | homepage | imdb_id | original_language | original_title | overview | popularity | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | [{'id': 313576, 'name': 'Hot Tub Time Machine ... | 14000000 | [{'id': 35, 'name': 'Comedy'}] | NaN | tt2637294 | en | Hot Tub Time Machine 2 | When Lou, who has become the "father of the In... | 6.575393 | /tQtWuw |
| 1 | 2 | [{'id': 107674, 'name': 'The Princess Diaries ... | 40000000 | [{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam... | NaN | tt0368933 | en | The Princess Diaries 2: Royal Engagement | Mia Thermopolis is now a college graduate and ... | 8.248895 | /w9Z7A0Gi |

*Fig 1: Dataset of 3000 movies*

### ii. Data Cleaning:

The dataset that we have obtained was raw data. It is just a collection of details like the movie cast, crew, popularity, Genre etc... Data cleaning is a very important step which cleans the data and converts the raw data into the data needed for model training. Here, we are removing the null values present in the dataset. The below Fig-3 shows the column that contains null values in the data set. If we look at the fig-3 homepage column it contains 2034 null values.

```
data_explore.isna().sum()
```

```
id                          0
belongs_to_collection    2396
budget                      0
genres                      7
homepage                 2054
imdb_id                     0
original_language           0
original_title              0
overview                    8
popularity                  0
poster_path                 1
production_companies      156
production_countries       55
release_date                0
runtime                     2
spoken_languages           20
status                      0
tagline                   597
title                       0
Keywords                  276
cast                       13
crew                       16
revenue                     0
dtype: int64
```

*Fig 3: Checking Null values*

### iii. Data Analysis:

Data Analysis is third step in this process. Understanding the data is very important to implement the further steps. Data Visualization involves in this step. Converting the numerical or categorical data into graphical representation makes us clearer and may help to find the next operation that should be performed.

- The below Fig-4 consists of information regarding Top 20 most popular movies with X-axis as popularity and Y-axis as name of the movies. The top most popularity movie is Wonder Women with 294 popularity.

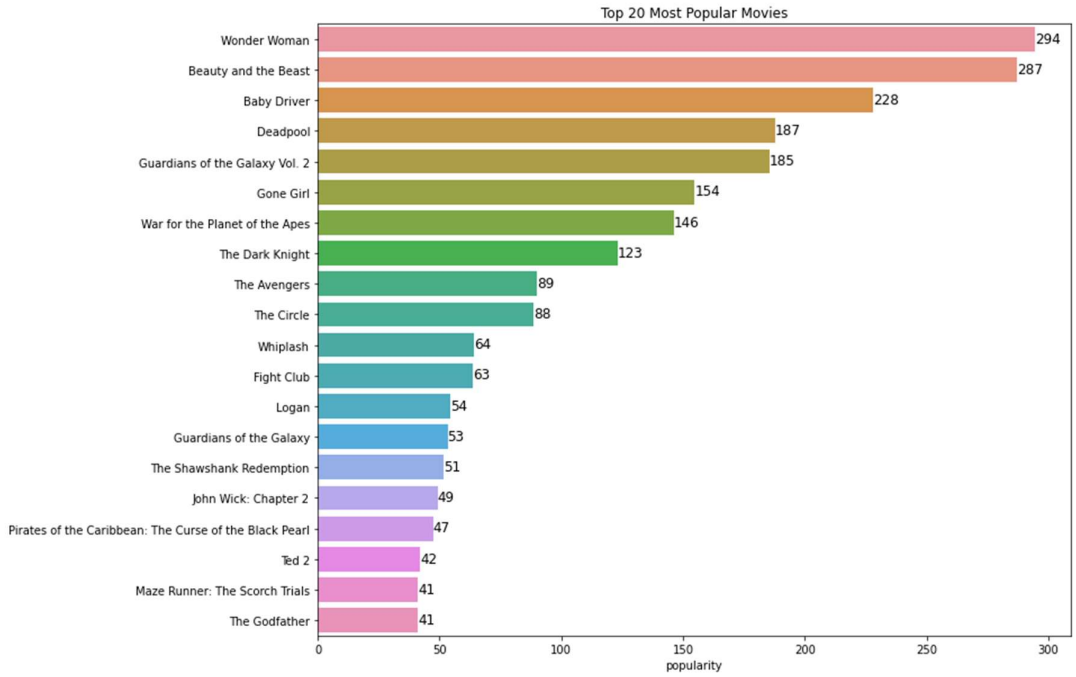


*Fig 4: Graphical representation of TOP 20 popularity movies*

- The below Fig-5 consists of information regarding Top 20 high revenue movies with X-axis as revenue(as million) and Y-axis as name of the movies. The top most revenue movie is The Avenger with 1519 million.

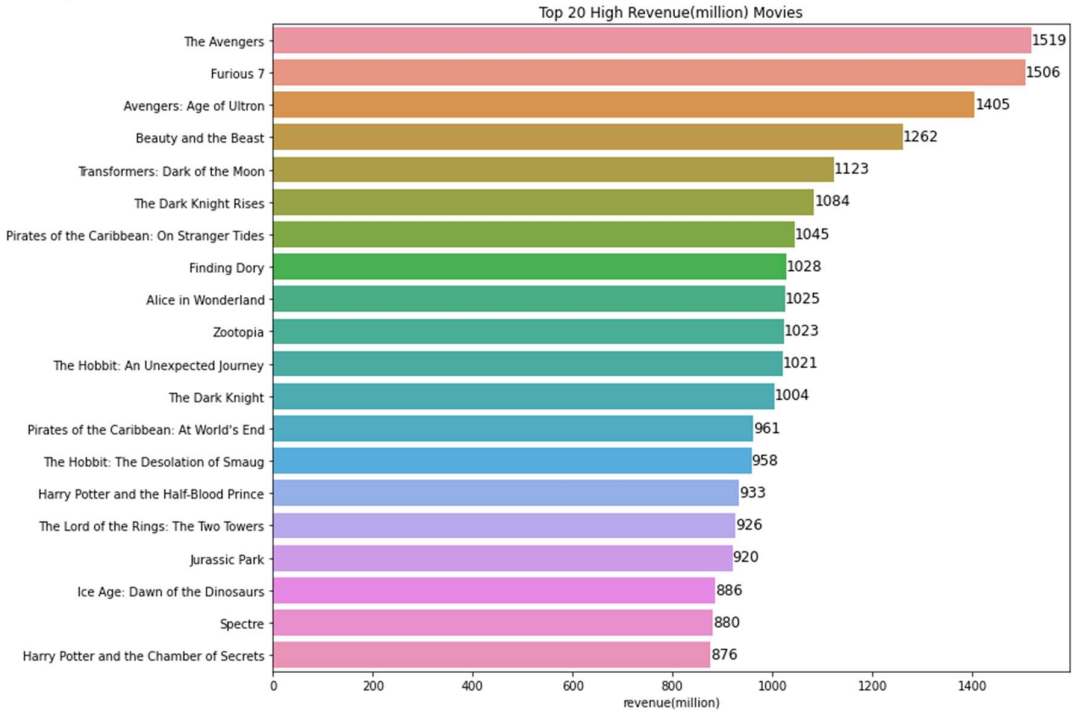


*Fig 5: Graphical representation of TOP 20 high revenue movies.*

- The below Fig-6 consists of information regarding Top 20 high budget(million) movies with X-axis as Profit(million) and Y-axis as name of the movies. The top most budget(million) movie is Pirates of the Caribbean- On stranger tides with 380 million.

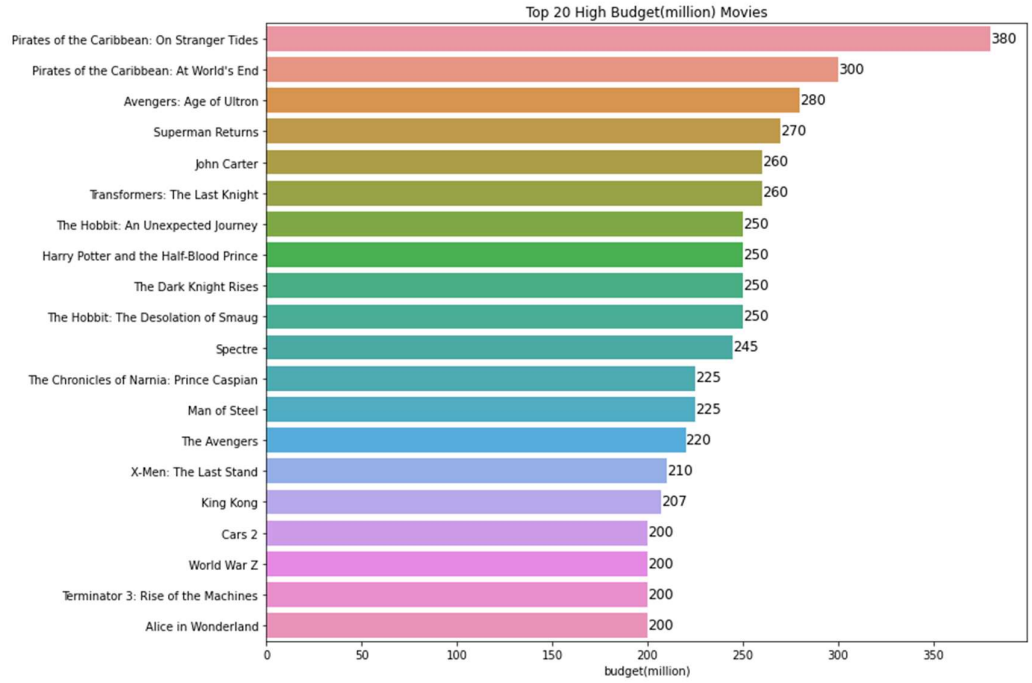


*Fig 6: Graphical representation of TOP 20 high budget(million) movies*

- The below Fig-7 consists of information regarding Top 20 highest Grossing movies with X-axis as Budget(as million) and Y-axis as name of the movies. The top most Profit(million) movie is Furious-7 with 1316 million.

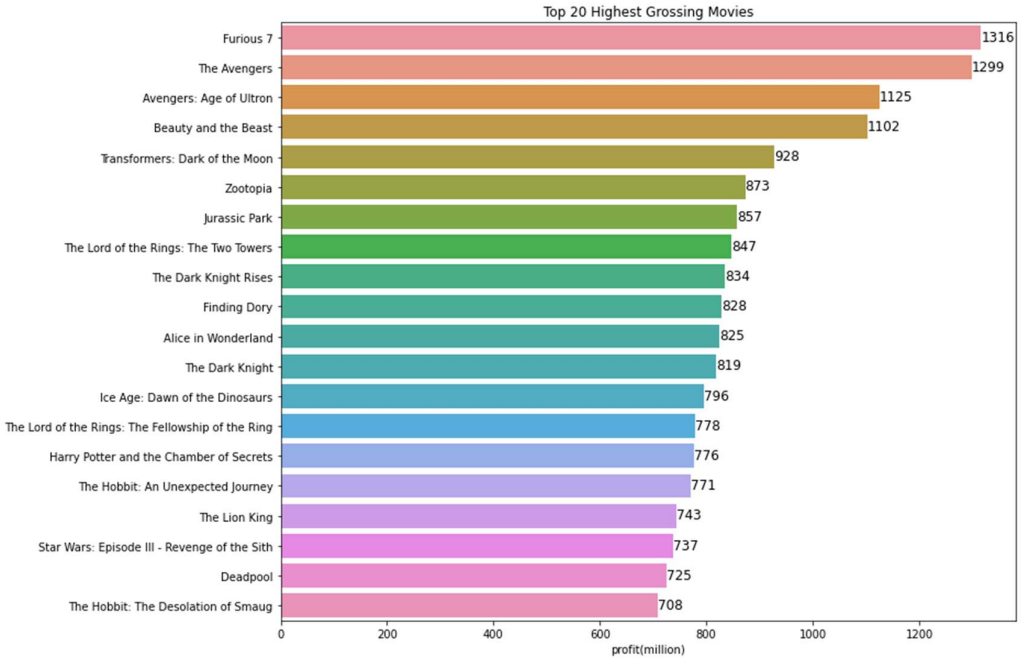


*Fig 7: Graphical representation of TOP 20 highest Grossing(million)*

- The below Fig-8 shows the movie count of different types of genres. X-axis contains the types of Genres and Y-axis represents the count of movies. By below graph we can conclude that the most movies released are based on Drama which is 1531 movies.

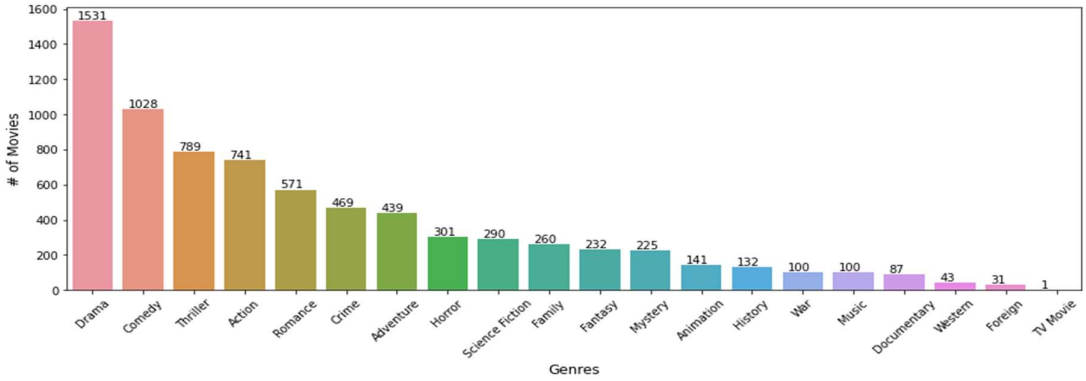


*Fig 8: Graphical representation of Genres*

- The below Fig-9 shows the relationship between Genres and Median popularity. In X-axis we took types of Genres and in Y-axis it shows popularity of that particular Genre.

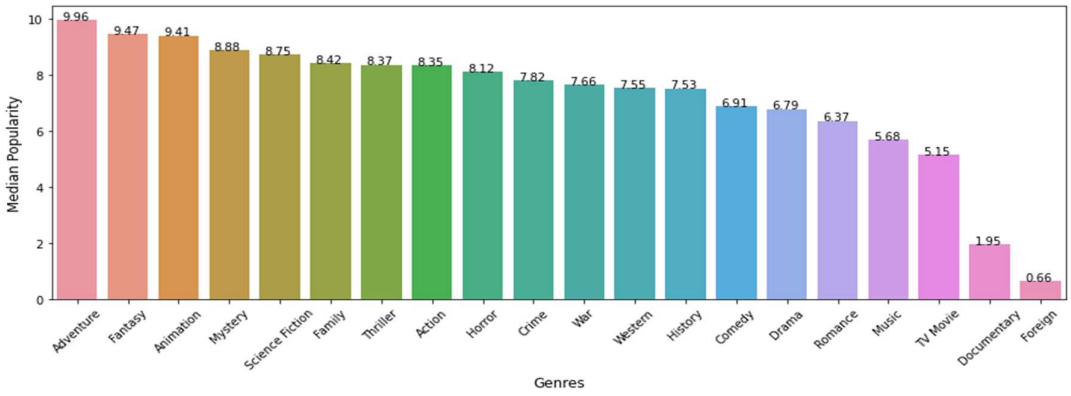


*Fig 9: Genres versus Median popularity*

- The below Fig-10 tells us the total budget kept on particular genre and revenue earned for that. In X-axis we took types of Genres and in Y-axis it shows sum value(in million).

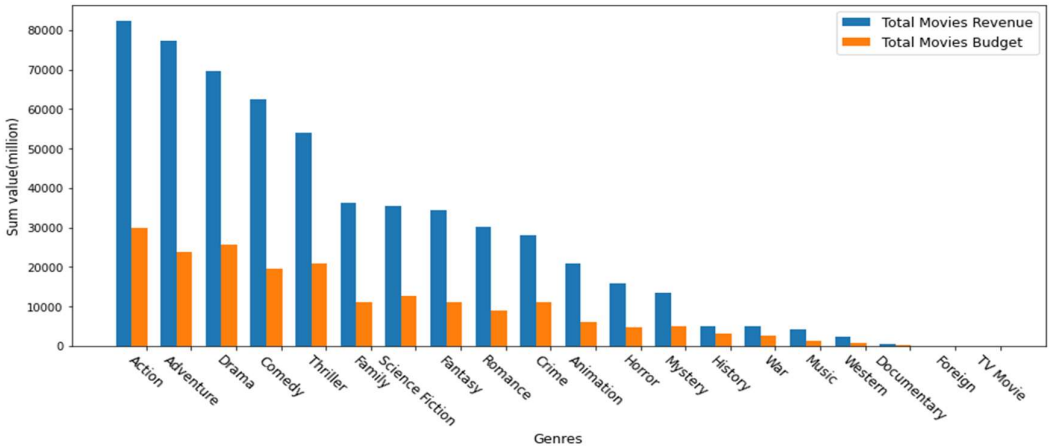


*Fig 10: Budget and revenue of particular Genre*

- The below Fig-11 shows the relationship between Revenue versus Budget. In X-axis we took Budget and in Y-axis we took Revenue.

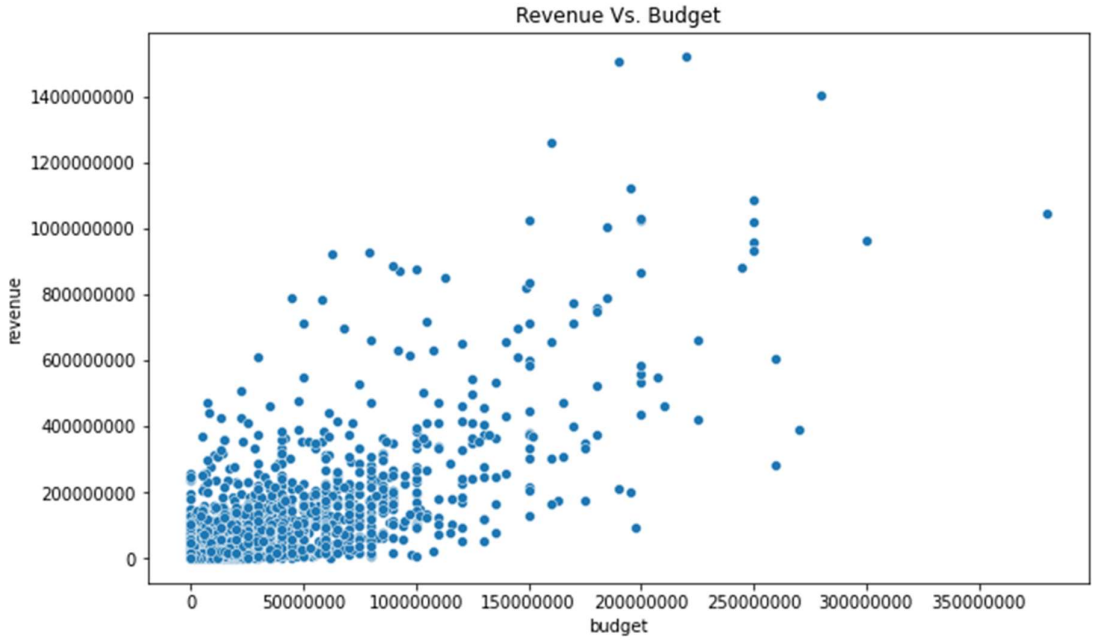


*Fig 11: Scatter plot representation of Budget vs Revenue.*

- The below Fig-12 shows the relationship between Revenue versus Popularity. In X-axis we took popularity and in Y-axis we took Revenue.

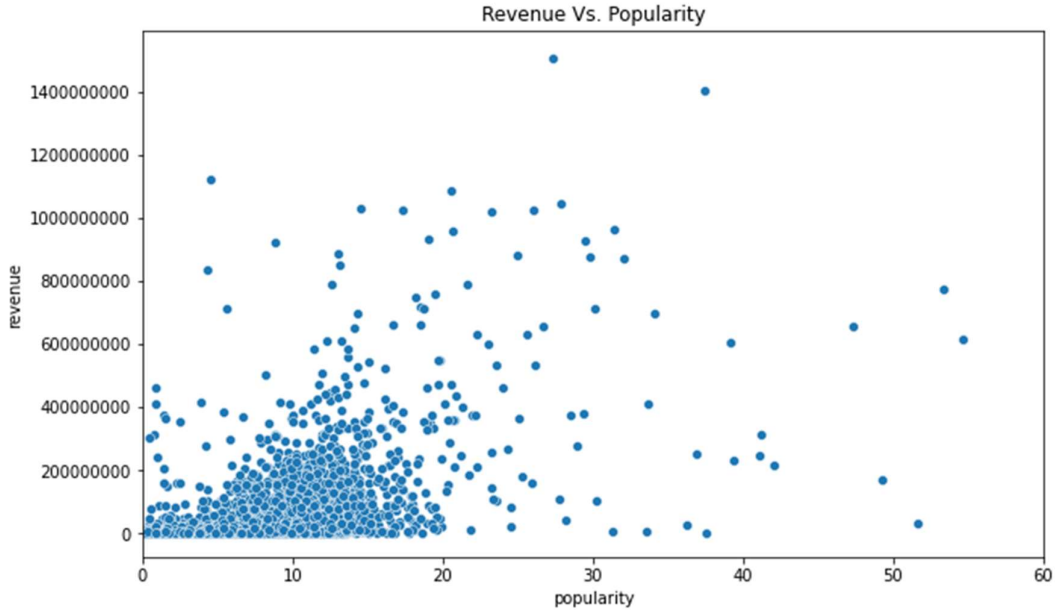


*Fig 12: Scatter plot representation of Revenue vs Popularity*

- The below Fig-13 shows the relationship between Revenue versus Movie Runtime. In X-axis we took Runtime and in Y-axis we took Revenue.

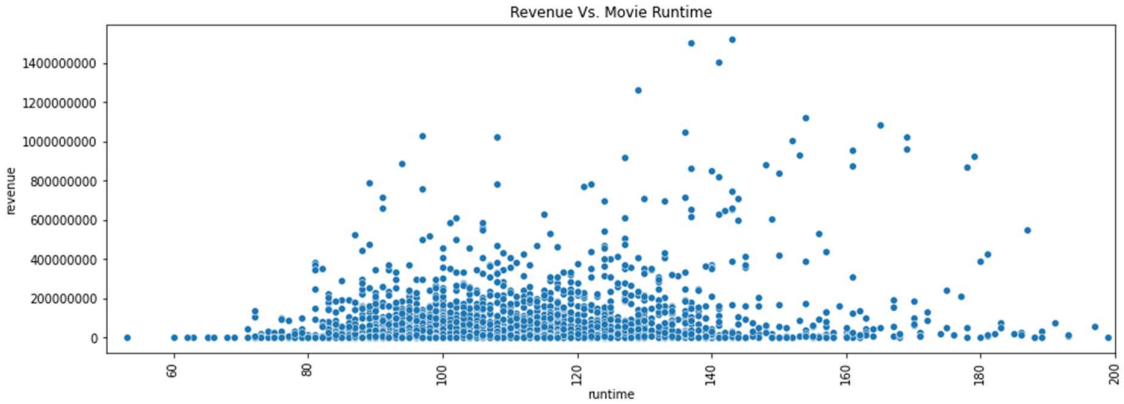


*Fig 13: Scatter plot representation of Revenue vs Movie Runtime.*

- Revenue across the release years of the movies are shown in the below figure.

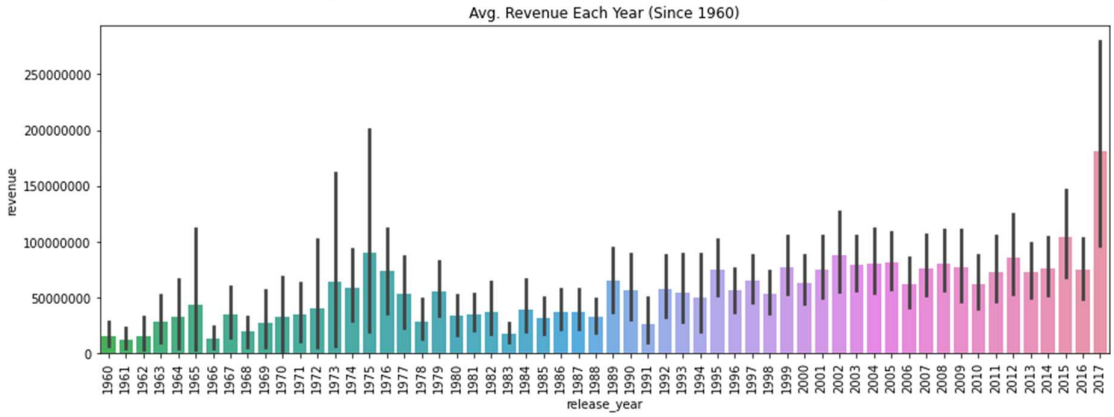


*Fig 14: Scatter plot representation of Revenue vs Movie release year.*

- Revenue across each months of release of movies are depicted in the below figure.

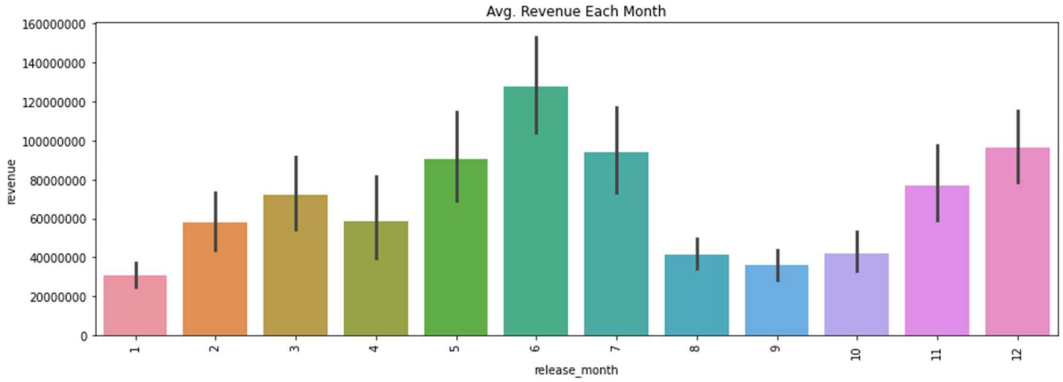


*Fig 15: Scatter plot representation of Revenue vs Movie Release month.*

- Revenue across language is shown below and the revenue is highest for English movies.

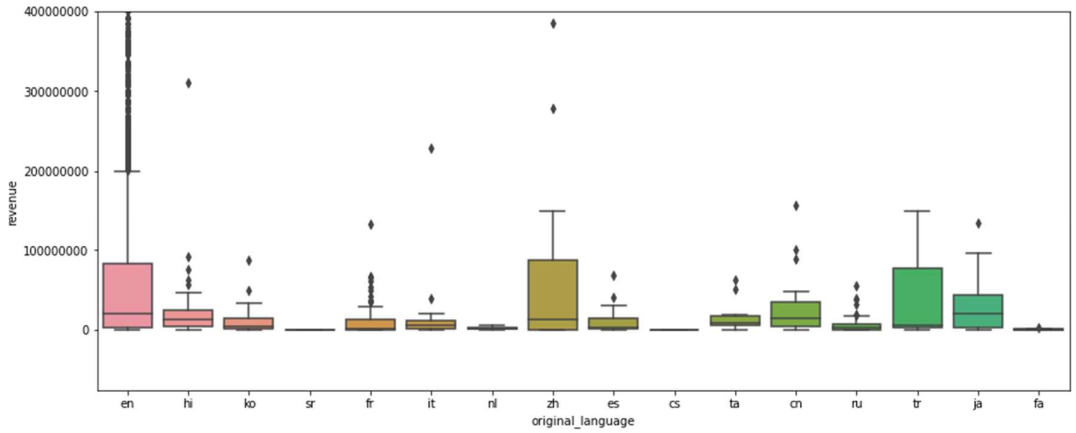


*Fig 16: Scatter plot representation of Revenue vs Movie Original language.*

- We have also found the correlation between each feature and also represented it graphically. By seeing that, we came to know that some of them are positively correlated and some are negatively correlated.

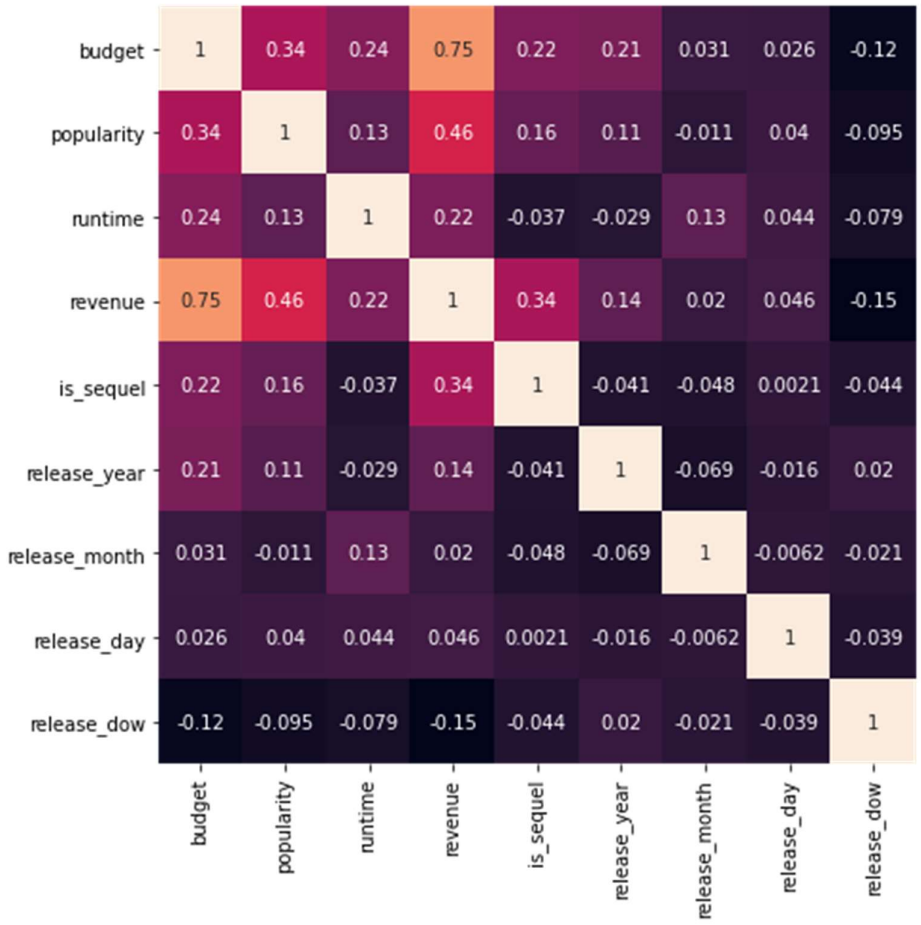


*Fig 17: Correlation of each feature of dataset*

### iv. Handling Categorical Values:

Machine learning models cannot handle NULL values and categorical values. So, it is necessary to handle the categorical values before implementing the model. Handling is nothing but converting those values to numerical values. It helps to convert the categorical values to numerical.

| Comedy | Drama | Family | Romance | Thriller | Action | Animation | Adventure | Horror |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Fig 18: Dataset and datatypes after applying one hot encoding*

**Model Implementation:**

The data is ready for the model implementation. Now, we have split the dependent and independent variables into train and test set by using the inbuilt method train_test_split() which is also provided by sklearn. The We know that it is a multiple classification problem. We Initially performed feature engineering and found the best features.

i. **Feature engineering:**

Imputation is used to get the missing values. Handling outliers to have bad impacts on the model. Binning is used to make model robust and prevent from overfitting. Log transformation is used in order to have normalize the effect of change of magnitude over a variable i.e., marks scored in two consecutive assessments by two students are 45,50 and 85,90. The change of 5 marks value is different in both the cases.

ii. **K-Fold Cross Validation:**

Cross-validation is a resampling technique for evaluating machine learning models on a small sample of data.The procedure has only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the technique is often referred to as k-fold cross-validation. It is a technique used in ML to estimate a machine learning model's ability on unknown data. That is, to use a small sample to estimate how the model would do in general when it's used to make predictions on data that was not used during model's training.

The overall method is as per the following:

- Mix the dataset.
- Split the dataset into k gatherings
- For every interesting gathering:
    - Accept the gathering as a hold out or test data set
    - Accept the excess gatherings as a training data set
    - Fit a model on the training set and assess it on the test set
    - Hold the assessment score and dispose of the model
- Sum up the skill of the model utilizing the example of model assessment scores

### iii. Ridge Regression:

Ridge regression is a model tuning technique that can be used to analyse data with multicollinearity. L2 regularisation is achieved using this approach. Where there is a problem with multicollinearity, least-squares are unbiased, and variances are high, the expected values are far from the actual values.

Ridge regression can be used as a classifier, just code the response labels as -1 and +1 and fit the regression model as normal. The Accuracy of the ridge regression model for the revenue analysis was 0.7123. this is comparatively lower than accuracy of Random Forest Algorithm used in later parts of the project.

### iv. Random Forest:

Random Forest is the collection of decision trees, each and every tree will be trained independently by selecting best attribute to be next by using indicators like IG, gini index, gain ratio etc. Also for each tree the given random samples are independent. Here each tree vote and most voted class will be considered and the test data will be assigned to that class. The accuracy Of the model in Random Forest is around 90%.

| | Models | Accuracy |
|---|---|---|
| 0 | Ridge Regression | 0.716500 |
| 1 | Random Forest | 0.899700 |

*Fig 19: Accuracy Comparison of both the algorithms.*

The accuracy of Ridge regression is 71% while the accuracy of the Random Forest algorithm is ~90%. This is shown in the figure 19. As the accuracy is compared, the Random Forest algorithm is clear dominant. So for the further prediction, random forest regressor is used.

| | Model | Train RMSE | Test RMSE |
|---|---|---|---|
| 0 | Ridge | 2.280826 | 2.264044 |
| 1 | RandomForestRegressor | 2.149805 | 2.063491 |

*Fig 20: RMSE Scores for Ridge and random Forest Regressor.*

When the two models are compared together, their corresponding RMS Scores are found that the error for random forest regressor is less compared to the error in Ridge regressor. So Random forest regressor is used in all the other cases.

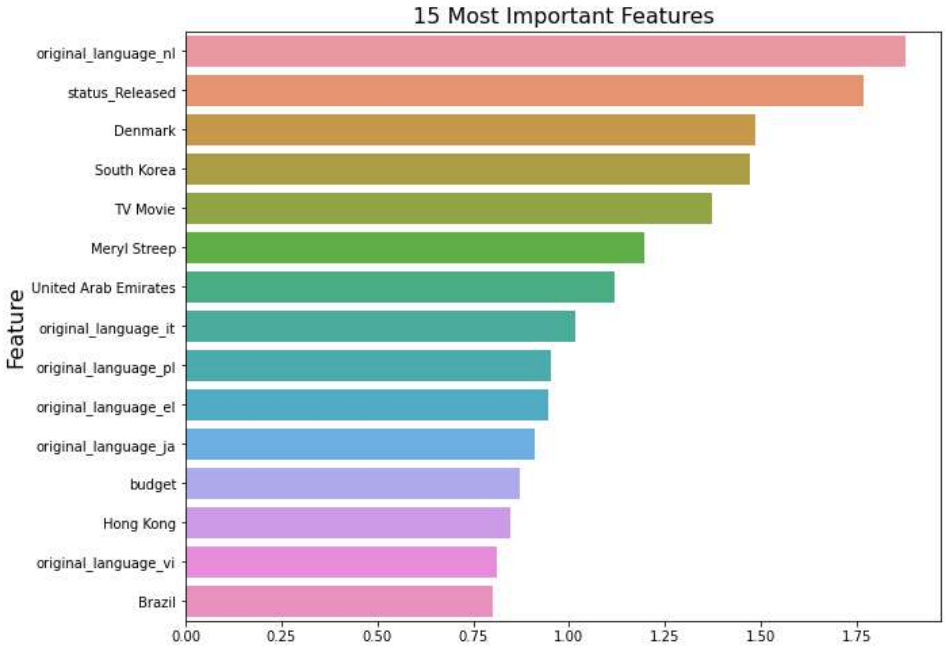


*Fig 21: Top 20 features extracted from Ridge Regression*

The top 20 features are extracted based on the coefficients of the variables(columns). The highest 20 coefficients are considered, which shows that these 20 features are more important compared to other coefficients. In Ridge regression, the important features are shown in the y-axis of the Fig 20. Which are original languages, status of release, etc. we felt these are also not satisfactory for the Ridge Regression. So, we moved to Random Forest Regression.

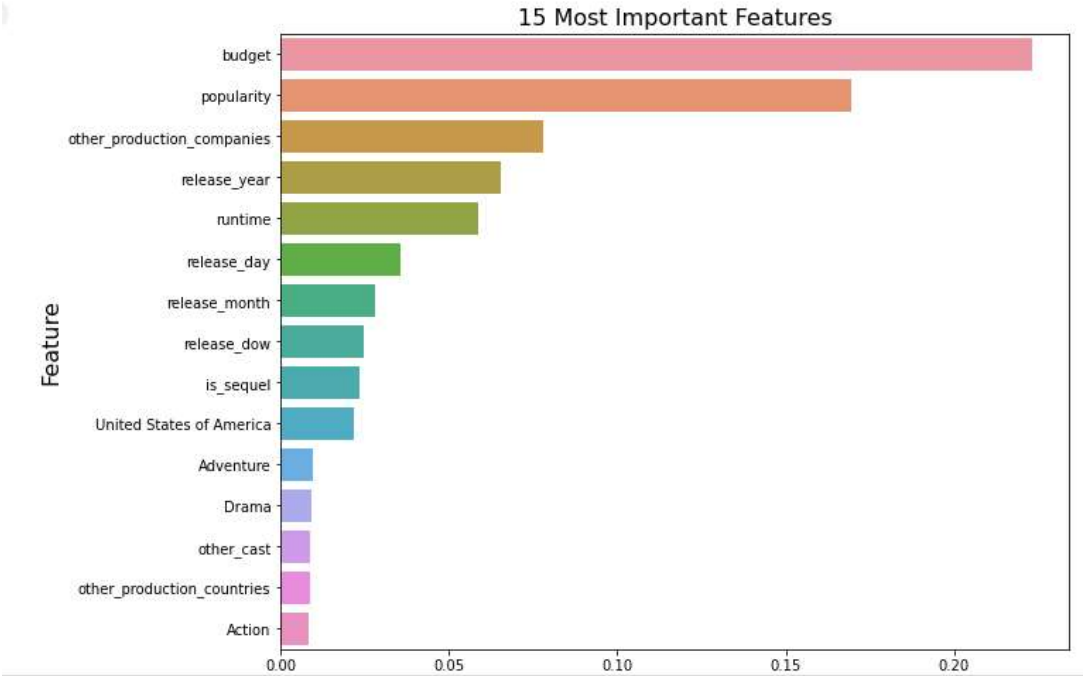


*Fig 22: Top 20 features extracted from Forest Regression*

### 4.3) Block Diagram of the project:

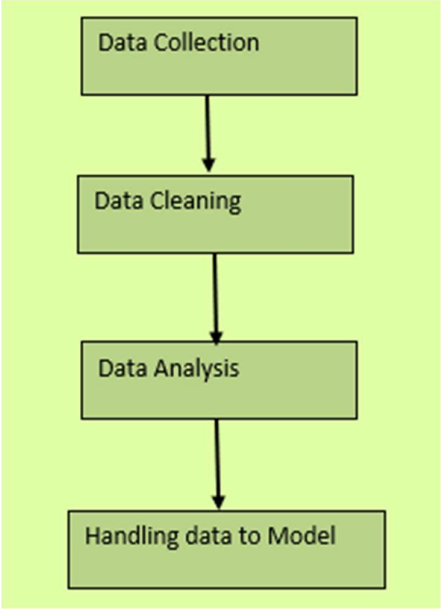


*Fig 11: Block diagram of Data cleaning and Analysis in this project*

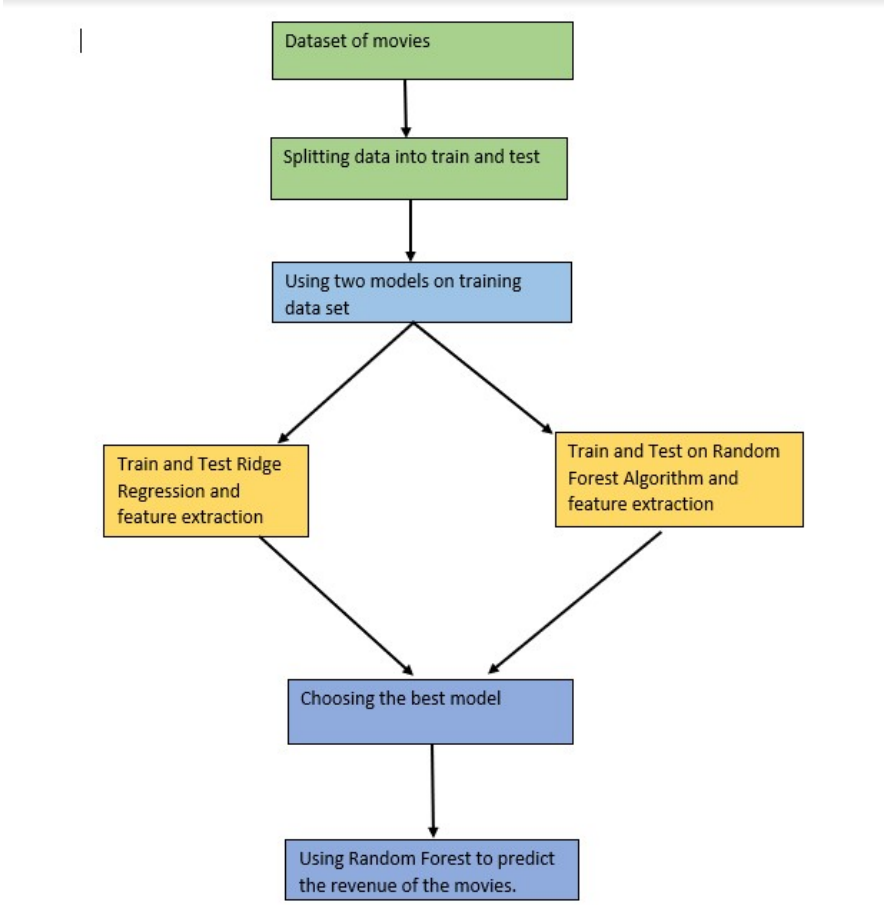


*Fig 11: Block diagram of Model in this project*

### 4.4) Hardware and Software Components used in our project:

We have implemented this project in Google Collab notebook in python language. It allows anybody to write and execute arbitrary python code through the browser and is especially well suited to machine learning and data analysis. Some of the python open-source libraries like NumPy, pandas for data analysis and Matplotlib, Seaborn for data visualization. And another open-source package like sklearn is used for Model training and implementation.

We implemented the entire project on Google Colaboratory which is called as Google Colab. This Colab files will be helpful in prototyping ML models on powerful hardware devices like TPU's and GPU's. We implemented the entire code in Python language and used the libraries like Pandas, matplotlib, seaborn, scikitplot, sklearn, StandardScalar and classification_report.

**pandas:** Used to create dataframe, do manipulations and changes accordingly to the dataframe.

**matplotlib:** Used to plot the graph for finding the optimal K value in doing K-NN.

**seaborn:** Used in data analysis part and this library plays a key role in doing data analysis for our project.

**scikitplot:** Used for building the confusion matrix which is used for evaluation purpose.

**sklearn:** It is a ML library which has various clustering, classification and regression algorithms in it and in our case we are using classification algorithms only. We import the above built 6 models for this library.

**StandardScalar:** For the dataset (where every variable is numeric) it removes the mean and scales every variable/attribute to unit variance.

### 4.5) Features of our project that are new and distinguish them over the closest technology:

Every project has its own methodology to obtain best results. Generally, in every project one algorithm will be applied and one accuracy score will be calculated. But In our project, we have checked with two classification algorithms and selected the best model.

The closest technology i found to my project is Early Prediction of movie Box office success Based on Wikipedia Activity Big Data they predicted the movie through the popularity. But it fails many times as it cannot decide whether a particular movie is good or bad. The thing our project make a difference is that we took many parameters to predict which can make a huge difference in both.

### 4.6) Alternative ways of implementing our project:

There are many alternative ways of implementing this project.

- When coming to model implementation part, we have used two Machine learning classification algorithms (Ridge Regression and Random Forest Regressor). There are other classification algorithms like Support vector machine, Naive Bayes, Random forest,Stochastic Gradient Descent. We can also apply these algorithms on data set and choose the best one according to their accuracy and performance.
- We can also implement this project using deep learning and neural networks.

### 4.7) Status of our Project:

Our project is successfully implemented. We have analyzed the dataset and trained with two classification algorithms and also calculated respective errors. In the comparison of both the models, it is found that the error is less in Random forest algorithm and it was used to predict the revenue of few films.

Thus, we are confident that our project is completed on time and with the best accuracy possible. But there is chance of extending it further with good optimizations.

## 5) RESULT AND DISCUSSION

After the successful training of the model, few movies whose revenue is unknow was given to the model to predict the revenue. Our model was successful enough in calculating the revenue to all the 4000 movies given along with all the data (with few missing values) and same parameters like genre, budget, popularity etc. The revenue is predicted for the movies like Pokemon: The Rise Of Darkrai, Attack of the 50 Foot Woman, addited to Love,Incedies and Inside Deep Throat. we have predicted the revenue of 4000 films in the dataset which has was not attached with the revenue and are saved in ‘output.csv’.

| | title | revenue |
|---|---|---|
| 0 | Pokémon: The Rise of Darkrai | 4.312409e+06 |
| 1 | Attack of the 50 Foot Woman | 1.574562e+06 |
| 2 | Addicted to Love | 6.327415e+06 |
| 3 | Incendies | 1.014175e+06 |
| 4 | Inside Deep Throat | 6.030557e+05 |

## v. CONCLUSION AND FUTURE WORK

Finally, Big Data in Movie Analysis is very useful which allows us to analyze and predict the revenue of the model with greater accuracy, and eliminating the guesswork often involved. The main objective of this project is to analyze and build a machine learning model that predicts the revenue of the movie. We have taken the dataset from Kaggle which contains details of 3000 movies with the data about title, actors, budget etc.

In this project, the dataset is analyzed, visually represented, and trained with two different classification algorithms. They are Ridge regression and Random Forest. In Evaluation, RMSE scores are calculated for each algorithm and chosen the best one according to these scores. Finally, we have predicted the revenue of 4000 films in the dataset which has was not attached with the revenue

## REFERENCES


[1] Mestyán, Márton, Taha Yasseri, and János Kertész. "Early prediction of movie box office success based on Wikipedia activity big data." *PloS one* 8.8 (2013): e71226.

[2] Apte, Nikhil, Mats Forssell, and Anahita Sidhwa. "Predicting movie revenue." *CS229, Stanford University* (2011).

[3] Latif, Muhammad Hassan, and Hammad Afzal. "Prediction of movies popularity using machine learning techniques." *International Journal of Computer Science and Network Security (IJCSNS)* 16.8 (2016): 127.