***A Mini Project Report***

**ON**

**REVENUE ANALYSIS AND VISUALIZATION OF**

**BOX OFFICE**

***Submitted In partial fulfillment for the Degree of B.Tech***

***In***

**Artificial Intelligence**

***by***

*Sayani Aryan Raj (19911A3544)*

*Velpula Sai Kiran Yadav(19911A3557)*

***Under the guidance of***

Dr. I. S. Hephzi Punithavathi

Associate Professor

Department of Artificial Intelligence



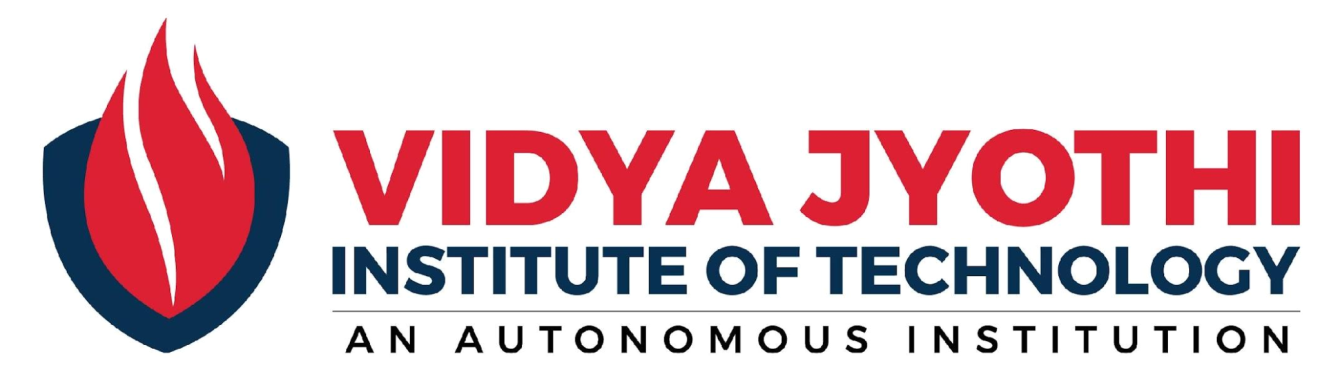
**DEPARTMENT OF ARTIFICIAL INTELLIGENCE**

**VIDYA JYOTHI INSTITUTE OF TECHNOLOGY**

(An Autonomous Institution)

Accredited by NAAC & NBA, Approved by AICTE New Delhi & Permanently Affiliated to JNTUH

**2022 – 2023**



[Accredited by NAAC & NBA, Approved by AICTE New Delhi & Permanently Affiliated to JNTUH]

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE**

***CERTIFICATE***

This is to certify that the project report entitled **Revenue Analysis and Visualization of Box office** submitted by **Sayani Aryan Raj(19911A3544)** and **Velpula Sai Kiran Yadav(19911A3557)** to Vidya Jyothi Institute of Technology, Hyderabad, in partial fulfillment for the award of the degree of **B.Tech in Artificial Intelligence** a bonafiderecord of project work carried out by us under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institution or University for the award of any degree.

**Dr. I. S. Hephzi Punithavathi Dr. Avuku Obulesh**

**Supervisor Head of Department**

**Department of Artificial Intelligence Department of Artificial Intelligence**

ⅰ

***DECLARATION***

We declare that this mini project report titled **Revenue Analysis and Visualization of Box office** submitted in partial fulfillment of the degree of **B.Tech in Artificial Intelligence** is a record of original work carried out under the supervision of **Dr. I. S. Hephzi Punithavathi** and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice of reporting scientific information, due acknowledgments have been made wherever the findings of others have been cited.

Sayani Aryan Raj (19911A3544)

Velpula Sai Kiran Yadav (19911A3557)

ⅱ

***ACKNOWLEDGEMENTS***

This acknowledgment transcends the reality of formality when we would express deep gratitude and respect to all those people who helped, supported, guided and inspired us throughout the completion of this mini project.

We would like to express our gratitude to our guide **Dr.I. S. Hephzi Punithavathi** Associate Professor, Department of Artificial Intelligence, for her valuable guidance, suggestion, and encouragement in completing the mini project successfully.

We would like to express our sincere thanks to **Dr. Avuku Obulesh**, Associate Professor, Project Review Committee of the Department of Artificial Intelligence for their valuable guidance, and suggestions and for supervising us at every instance in completing the mini project successfully.

We would like to convey our sincere thanks to **Dr. A. Padmaja**, Principal of Vidya Jyothi Institute of Technology, for her valuable support.

We would like to convey our sincere thanks to **Dr. E. Saibaba Reddy**, Director of Vidya Jyothi Institute of Technology, for his valuable support.

We would like to express our sincere thanks to our friend **G. Hemanth Sai**, for encouraging us, helping us with hurdles, and rectifying our mistakes.

ⅲ

***ABSTRACT***

In this project we explore different attributes affecting box office revenue and find out how they impact a particular film's revenue. We use the data taken from Kaggle called TMDB dataset, we mainly aim to analyze and visualize data that shows a better interpretation of data, and we try to check how the movie has performed by utilizing the various visualization libraries like matplotlib, and seaborn. Prediction of movie revenue with the application of regression algorithm, which aims to predict revenue based on popularity and budget. Prediction of revenue with root mean squared logarithmic error which gives us a ratio of true value and predicted value in performing we get a marginal value which says predicted values are less than true values analyzing movie descriptions like which words are better to use in a movie description. The vectorization process combines linguistic techniques from nltk with machine learning techniques. Transformation of text data into numbers and finding respective weights of words and indicating in the scale of positive and negative impacts on revenue.

ⅳ

**LIST OF FIGURES**

| **FIGURE NUMBER** | **TITLE** | **PAGE NUMBER** |
| --- | --- | --- |
| Figure 3.1 | Regression Line | 19 |
| Figure 3.2 | Structure of random forest | 21 |
| Figure 3.3 | Bar Chart | 23 |
| Figure 3.4 | Histogram | 23 |
| Figure 3.5 | Scatter Plot | 24 |
| Figure 3.6 | Count Plot | 25 |
| Figure 3.7 | Top words in titles | 26 |
| Figure 4.1 | Revenue v/s Budget | 27 |
| Figure 4.2 | Top ten movies highest revenue | 28 |
| Figure 4.3 | Top ten high budget films | 28 |
| Figure 4.4 | Top ten popular films | 29 |
| Figure 4.5 | Movie release in a week | 30 |
| Figure 4.6 | Top words used in film overview | 30 |
| Figure 4.7 | Language revenue collection | 31 |
| Figure 4.8 | Feature importance of film description | 32 |

ⅴ

**LIST OF TABLES**

| **TABLE NUMBER** | **TITLE** | **PAGE NUMBER** |
| --- | --- | --- |
|  |  |  |
| Table 3.1 | Dataset attributes | 10 |
| Table 3.2 | Example Stop words | 14 |
| Table 3.3 | Ngrams | 14 |
| Table 3.4 | Examples of N-gram types | 15 |
| Table 4.1 | Accuracy of models used | 33 |

ⅵ

**ABBREVIATIONS**

NLTK - Natural Language Toolkit

OTT - Over The Top

WOD - Word of Mouth

eWOD - electronic Word of Mouth

IMDB - Internet Movie DataBase

ASBA - Aspect Based Sentiment Analysis

TF-IDF - Term Frequency - Inverse Document Frequency

ⅶ

**TABLE OF CONTENTS**

| **DESCRIPTIONS** | **PAGE NUMBERS** |
| --- | --- |
| CERTIFICATE | ⅰ |
| DECLARATION | ⅱ |
| ACKNOWLEDGEMENTS | ⅲ |
| ABSTRACT | ⅳ |
| LIST OF FIGURES | ⅴ |
| LIST OF TABLES | ⅵ |
| ABBREVIATIONS | ⅶ |
| CHAPTER 1 INTRODUCTION | 1 |
| 1.1 Introduction | 1 |
| 1.2 Background Information | 1 |
| 1.3 Objective | 2 |
| 1.4 Problem Statement | 2 |
| 1.5 Organization of Project | 2 |
| CHAPTER 2 LITERATURE SURVEY | 4 |
| 2.1 Existing Research Works | 4 |
| 2.2 Proposed Systems | 6 |
| CHAPTER 3 METHODOLOGY | 8 |
| 3.1 Data Exploration | 8 |
| 3.2 Data Preprocessing | 9 |
| 3.3 Libraries and Packages | 10 |
| 3.3.1 ELI5 | 11 |
| 3.3.2 NLTK | 12 |
| 3.3.2.1 Stop Words | 13 |
| 3.3.2.2 N-grams | 14 |
| 3.3.3 Scikit Learn | 15 |
| 3.3.3.1 Count Vectorizer | 15 |
| 3.3.3.2 TF-IDF Vectorizer | 16 |
| 3.4 Regression Algorithms | 17 |
| 3.4.1 Linear Regression | 18 |
| 3.4.1.1 Simple Linear Regression | 18 |
| 3.4.1.2 Multiple Linear Regression | 20 |
| 3.4.2 Random Forest Regression | 20 |
| 3.5 Data Visualization | 22 |
| 3.5.1 Matplot Library | 22 |
| 3.5.1.1 Bar Chart | 22 |
| 3.5.1.2 Histogram | 23 |
| 3.5.1.3 Scatter Plot | 23 |
| 3.5.2 Seaborn | 24 |
| 3.5.3 Word Cloud | 25 |
| CHAPTER 4 RESULTS AND DISCUSSIONS | 27 |
| 4.1 Revenue v/s Budget | 27 |
| 4.2 Top Ten High Revenue Films | 27 |
| 4.3 Top Ten High Budget Films | 28 |
| 4.4 Top Ten Popular Films | 29 |
| 4.5 Number of movies released on respective day | 29 |
| 4.6 Top Words used in film descriptions | 30 |
| 4.7 Revenue and Languages | 31 |
| 4.8 Analysis of film descriptions | 32 |
| 4.9 Revenue Prediction | 33 |
| CHAPTER 5 CONCLUSIONS | 34 |
| CHAPTER 6 REFERENCES | 35 |

**CHAPTER 1**

**INTRODUCTION**

**1.1 Introduction**

Films are part of the entertainment industry which provides the country with huge growth in the economic sector. Films are such a great deal in building a stable economic balance. The film, and television industries (together referred to as “creative industries” in this report) in India represent a combined revenue of about INR 930 000 million. Against the backdrop of a sluggish economy, these industries on aggregate have been resilient and performed well.

The projection and analysis of these earnings are greatly important for the creative industries and often a source of interest to fans. These figures can always be an interesting topic while discussing with your friends and family. The success or Failure of a movie can be determined with the help of box office statistics.

The film industry has grown immensely over the past few decades generating billions of dollars of revenue for the stakeholders. Now people can watch movies online and offline on a variety of mobile devices during leisure or travel through Netflix, Youtube, and downloads. A prediction system to assess the box office success of new movies can help movie producers and directors make informed decisions when making the movie in order to increase the chance of profitability and box office success.

**1.2 Background Information**

IMDB is the world’s most popular and authoritative source for movies. An online database of information related to films and television series including ratings, and fan and critical reviews. This site not only gives reviews, and ratings but also suggests the best movies with the highest ratings, the top 50 movies in particular in month and year. And also provides the following information about a movie: cast and crew names, character names, summaries, ratings, and language. The rating scale is from 1 to 10.

The other is Rotten Tomatoes, an American review-aggregation website for film and television series. Showing the best movies to watch or more likely recommending. The Tomatometer score is based on people's reviews and ratings. When at least 60% of reviews for a movie or TV show are positive, a red tomato is displayed to indicate its Fresh status. When less than 60% of reviews for a movie or TV show are positive, a green splat is displayed to indicate its Rotten status.

Soundtracks, visually striking scenes, fights, sets, locations, pre, and post-film works, and costume designs make the film more captivating, so these can't be ignored since they have a better impact on the result of the film. Therefore spending a fair amount of money on these things is also an important part of making a film successful. The amount of money spent as a whole in making a film is considered as the budget, it will always be a cautious figure in the minds of film producers.

Language is the most important connection between films and the audience. It inspires an individual in a certain way to connect completely to the story. In which different languages the film is released, which regional audience will like the film, and in which language the film has collected more revenue, all of these are important for a film to make good revenue.

The income of the film industry comes from screening movies in the theater, which is called “Box-Office”. The film industry is a highly competitive industry. Many new movies queue up to be released each week, so a theater owner has to decide on which movie to be shown, based mainly on revenue.

With the improvement in people’s living standards and film production level, more and more people enrich their lives by watching movies. For the success of a movie, box office is an important evaluation index, and how to reasonably and accurately predict The box office through data has become the focus of researchers. An effective box office prediction model can provide business decision support and guidance for film production and distribution companies, which is very important to the sustainable development of the film industry.

**1.3 Objective**

The objective of our project is to analyze the box office revenue. In this project, we found which movies have gained more revenue, and which movies have a bigger budget, and analyzed the revenue collected by the movie. This project provides deeper insight into the whole financial perspective of the film Industry, thus it is very useful for the filmmaking department saving lots of money and time.

**1.4 Problem Statement**

In this project, we will investigate the different attributes from the TMDB dataset and their signaling power for box-office revenue. By using multiple regression analysis, we will examine the relationship between box-office revenues.

**1.5 Organization of project report**

This project report is divided into 6 chapters. Chapter 1 consists of the Introduction and Objectives of the project. Chapter 2 consists of a Literature Survey. Chapter 3 consists of the methodologies used, this covers the implementation part in detail. Chapter 4 consists of the results and outputs obtained. Chapter 5 consists of the conclusions and Chapter 6 consists of References used in the project.

**CHAPTER 2**

**LITERATURE SURVEY**

In this section we will discuss different works done on the analysis of box office revenue. We will be discussing expert works and methodologies more specifically we will review the results of previous studies and also their influence on movie box office revenue. These reviews are given by professionals who worked on the analysis of box office revenue. We will discuss different challenges they faced while working and the conclusions of their works.

**2.1 Existing Research Works**

Works that are done on the box office revenue analysis so far were based on the factors affecting movie revenue, electronic Word of Mouth, study on ratings, and reviews of films.

According to Skyler and Victor's study, they proposed the following, using multiple regression analyses they found there will be a significant positive effect of the amount ratings from users as well as experts on box office revenue. Furthermore, they found, within the multiple regression model, there was a negative effect between the user and expert ratings on box office revenue, with this negative effect being caused by a mediation effect from several ratings. Using simple linear regression they showed, the influence on box office revenue from the quality and quantity of user rating is larger than that of expert rating. They used the following multiple regression,

Box office revenue = b1 + b2.Quality of user rating + b3.Quantity of user rating + b4. Quality of expert rating + b5.Quantity of expert rating + ∑ all control variables + 𝜺

The following conclusions were obtained, user rating is slightly higher than expert rating, range of rating is larger with experts than with users when looking at the minimum and maximum rating is given. The amount of “paid professionals” who give reviews and rate movies is limited, thus it is obvious that the number of expert ratings is lower than the number of user ratings.

Sangjae et al., proposed decision trees, k-nearest-neighbors(k-NN), and linear regression using ensemble methods and the prediction performance of decision trees based on bagging and boosting using a sample of 1439 movies. Using ensemble methods they combined several models to produce better performance rather than utilizing a single model. Ensemble methods combine multiple supervised models into a “supermodel”. The purpose of using an ensemble is to improve the weak power of individual models for the best performance of the combined model to achieve improved predictive accuracy. It consists of several learners known as base learners, which can be decision trees, neural networks, or other kinds of algorithms. Bagging and boosting are two relatively new but popular methods for producing ensembles with the same algorithm. These ensemble methods have been used in various applications of bioinformatics problems, bankruptcy prediction, and time series prediction.

The following results have observed that ensemble methods based on decision trees outperform ensemble methods based on k-NN in predicting box office. Decision trees using ensemble methods provide better prediction performance than ensemble methods based on linear regression analysis of the box office. They also compared the predicted performance of ensemble methods as greater than that of non-ensemble methods.Finally, they concluded that decision trees using ensemble methods provide better application effectiveness of ensemble methods than k-NN and linear regression analysis.

(Yuan Ni et al., 2021), proposed a model for predicting the profitability class of a movie namely “profit” and “loss” based on the data about movies released between the years 2010 and 2015. Their methodology considers both historical data as well as data extracted from social media. Data is normalized and then given a weight using standard normalization techniques. The cleaned and normalized dataset is then used to train a back-propagation cross-entropy validated neural network. They defined an artificial neural network model that is trained using pre-processed dataset. The model is then validated and tested to predict the success or failure of a movie given a set of input data items; the neural network model classifies movies into “success” or “failure”.

A multilayer backpropagation neural network from Matlab's toolbox specifically the neural pattern recognition tool to design, train and validate neural network models for classifying movie datasets into two classes namely success and failure. Used conjugate gradient backpropagation algorithm. Performance is evaluated using a cross-entropy method in which the training stops when the percentage of error in the output cannot be improved. They presented the following results given by their model showing the training confusion matrix was 90.9% accurate in the training phase. and the test confusion matrix is 78.6% accurate. These are the following research work we studied and observed the different methodologies and performances of different models giving a good and acceptable accuracy.

Electronic Word of Mouth, (eWOM) impacts box office performance. About 70% of customers would reference eWOM before making purchase decisions, implying that eWOM has become a reliable source of information for movie consumers. We split up the user-generated content on IMDB.com into quality and quantity. When we consider the quality, we consider the rating given. The literature reviews provide us with several results. Positive star ratings on amazon.com on books showed to have a significant positive effect on sales. So the valence of word of mouth positively influences sales. When we focus on the results of research done on User generated data and its effect on box-office revenue, we can draw a different conclusion. When looking at the quality of user ratings, the research found there was no significant effect on box-office revenue. The numerical value of the rating did not yield explanatory power.

Aspect-based Sentiment Analysis (ABSA) is a method that defines the terms related to the aspects and identifies the sentiment associated with each aspect. Customers could express their attitudes toward aspects that were the attributes of the product. Different aspects can produce different sentiment responses.By utilizing ABSA, there is the possibility to capture detailed information about objects of interest.

**2.2 Proposed Systems**

We propose to predict the revenue with given inputs such as budget, popularity, and runtime. Analyzing revenue with a certain budget and how much popularity it has gained and with respective runtime, we used two different regression algorithms: linear regression and random forest regressor comparing the accuracy of the two models. We will discuss methodologies that we will be using in our work which mainly includes regression algorithms.

A film budget can make certain expectations of the possible range of revenue that a movie can make, so with this hypothetical statement, we can certainly use the budget as one of the influencing parameters of revenue. The popularity of the film is a significant variable in real-time instances and also in expected revenue. The popularity of a film is directly proportional to the revenue it has collected. We can say that movies with more popularity can collect more money than movies with low popularity. So popularity is a significant parameter that influences revenue collection.

The runtime of a movie or film duration both are the same, the runtime is a pertinent parameter in films. The runtime of the movie can be of different time intervals based on the story and what the director wants to show and how much can be shown in his film. Film duration should be in an acceptable time period so that film is enjoyable and entertaining. Long-duration of films can create disinterest in movies and bad talk comes out so bad for revenue. Film Descriptions, a good description of a film can cause a meager level of interest in a movie to watch that film yet if film descriptions aren’t good then it may have an effect. In this project analyze the film descriptions of which words have a good influence on the movie. Feature extraction of text with Tfidvectorizer which converts raw documents into Term Frequency - Inverse Document Frequency(TF-IDF) features. TF-IDF is a measure that can quantify the importance or relevance of string representations like words and phrases in a document amongst a collection of documents or corpus.

Mostly this project gives a detailed analysis of movie revenue and other important factors which have influenced a movie so that it can make good collections. We used data visualization techniques to interpret the analyses of data using two different data visualization libraries: matplotlib, word cloud, and seaborn.

**CHAPTER 3**

**METHODOLOGY**

In this chapter we will be discussing the different methodologies, algorithms, and machine learning techniques we followed. We collected data from The Movie Database. The data set consists of a fine set of attributes about film and how they are affecting any movie. The main aim of our project is to give a good analysis with visualization about the box office and also prediction of revenue based on popularity, budget and runtime. In the following section 3.1 we will be discussing data exploration, and in section 3.2 will be discussing data preprocessing and various techniques.

**3.1 Data Exploration**

The dataset is chosen from The Movie Database from Kaggle which mainly aims to find an analysis of box office revenue from the given dataset. Utilizing better visualization techniques to easily interpret the data so others can understand data easily and analysis of data about box office revenue retrieving some conclusions about overall revenue collection around the world which contributes to the different country economies and how much is it affecting.

In the data set we have set finite variables which are used for analyzing data like budget, the amount of money invested in making a film, and revenue the amount of money which movie has gained, popularity, how popular the movie is, and the rating it has gained. In the process we explore the dataset using functions like head() which gives the first five rows of the dataset, tail() gives the last five rows of the dataset and describe() descriptive statistics including those that summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN values. Analyzes both numeric and object series, as well as DataFrame column sets of mixed data types. The output will vary depending on what is provided. Description of data in the form of Data Frame as given below,

count: Number of Non-empty values

mean: Average value.

std: The Standard Deviation.

**3.2 Data Preprocessing**

In this section we will be discussing the data preprocessing of data like filling the null values with any of the methods like median, mean, and forward fill which are used in different cases with respective requirements of data. And dropping some insignificant columns. Data preprocessing can refer to the manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process. After the preprocessing, the raw data is transformed into useful and efficient data. We can refer to data preprocessing as the manipulation of data or dropping of data before it is used in order to ensure or enhance performance.

Initially, we need to find out which columns are having null values and also find the number of null they are containing in that column which can be achieved by using the function is null().sum(). By identifying such columns, we easily find the columns we want to drop which aren’t relevant to the analysis and prediction process. Using the forward fill method which is one of the methods to fill null values by values from the previous row or previous columns based on the values assigned to the axis attribute. And will propagate the last valid observation forward.

And creating some new columns or deriving some data from release date columns like release weekday, release day, release month, and release year by using the date time format and assigning respective new columns with corresponding parts of date. By creating such new columns using release date we can perform analysis in different ways which can be insightful and data interpretation can be done in different ways.Columns with a higher number of null cannot be used for any type of analysis which can cause more errors during the visualization which allows us to display wrong information to others so in order to eliminate such kinds of errors we aren't including those columns which can be erroneous. In this project we mainly focus on revenue and how other attributes affect the revenue and which way, like how the budget is affecting the revenue higher the budget can cause the film with huge popularity so everyone will know about film and make them curious.

Variables with certain significance will be used in algorithms as independent variables and dependent variables. Independent variables which can affect the dependent variables. Let’s say variables like budget, runtime, and popularity are independent variables, and revenue as the dependent variable or target attribute. In the process, we find how the dependent and independent variables are correlated to each other and whether they can have certain relations among them. If we can find such a correlation we can use those variables in predicting target attributes with independent variables.

In the dataset we have the following set attributes, datatype, no-null count.

**Table 3.1**: Dataset attributes.

| **S.No** | **Column** | **No –Null count** | **Data type** |
| --- | --- | --- | --- |
| 1 | Id | 3000 | int |
| 2 | Belongs\_to\_collection | 604 | Object |
| 3 | Budget | 3000 | Int |
| 4 | Genres | 2993 | Object |
| 5 | Homepage | 946 | Object |
| 6 | Imdb\_id | 3000 | Object |
| 7 | Original\_language | 3000 | Object |
| 8 | Original\_title | 3000 | Object |
| 9 | Overview | 2992 | Object |
| 10 | Popularity | 3000 | Float |
| 11 | Poster\_path | 2999 | Object |
| 12 | Production\_companies | 2844 | Object |
| 13 | Production\_countries | 2945 | Object |
| 14 | Release\_datae | 3000 | Object |
| 15 | Runtime | 2998 | Float |
| 16 | Spoken\_languages | 2980 | Object |
| 17 | Status | 3000 | Object |
| 18 | Tagline | 2403 | Object |
| 19 | Title | 3000 | Object |
| 20 | Keywords | 2724 | Object |
| 21 | Cast | 2987 | Object |
| 22 | Crew | 2984 | Object |
| 23 | revenue | 3000 | int |

**3.3 Libraries and Packages**

We have used a few algorithms in the project along with some better visualization techniques like ELI5. We also used NLTK and feature extraction techniques from the scikit learn module. All the algorithms and techniques used in the project will be explained in detail in this section. We are using linear regression to visualize the importance of the features from the movie overview or the description. This means that what are the words or the features that have a better impact on the revenue of the movie are being figured out so to extract the impact of the features we have used linear regression.

**3.3.1 ELI5**

ELI5 is a python package that helps to debug machine learning classifiers and explain their predictions. It provides support for machine learning packages and frameworks. Currently, ELI5 allows us to explain weights and predictions of scikit-learn linear classifiers and regressors, print decision trees as text or as SVG, show feature importances, and explain predictions of decision trees and tree-based ensembles. The eli5 has support for a list of libraries below it can explain that we'll be primarily concentrating on sklearn as a part of this project.

1. Scikit-learn

2.XGBoost

3.CatBoost

4. Keras

5. lightning

6.LightGBM

7. sklearn-crfsuite

At high-level, the eli5 provides two ways to understand ML models and their predictions:

1. Global Features Importances (Model Level), It lets us analyze model weights to understand the global performance of the model.

2. Local Features Importances (Individual Example Level), It lets us analyze individual data examples' predictions to understand the local performance of the model. This can help us drill down why the particular prediction was made and which data features played what role in that prediction. "Eli5" provides many different formatting functions through "elif5.formatters" that let us format features importance present in the Explanation object in different ways. This can be useful for individuals who want to format importance in different ways.Some of the important ones are format\_as\_html(), format\_html\_styles(), format\_as\_text(), format\_as\_dict(), format\_as\_dataframe(), format\_as\_dataframes(), format\_as\_image().

1)explain\_weights()

It takes any ML Model as input and returns an explanation object which has global features importances. This object can be formatted in different ways to display features importances.

Based on the input model type it calls an underlying method to handle that model.

"explain\_weights\_sklearn()"

"explain\_linear\_regressor\_weights()"

"explain\_linear\_classifier\_weights()"

"explain\_decision\_tree()"

"explain\_rf\_feature\_importance()"

"explain\_permutation\_importance()"

"explain\_weights\_lightgbm()"

"explain\_prediction\_xgboost()"

"explain\_weights\_catboost()"

"explain\_weights\_lightning()"

"explain\_weights\_sklearn\_crfsuite()"

2)explain\_prediction()

It takes any ML Model and single data example as input and returns an Explanation object which has local features importances explaining contributions of features for that individual example prediction. We can format this explanation object as per our requirements using the method mentioned in the previous point.

Based on the input model type it calls an underlying method to handle that model.

"explain\_prediction\_sklearn()"

"explain\_prediction\_linear\_regressor()"

"explain\_prediction\_linear\_classifier()"

"explain\_prediction\_tree\_classifier()"

"explain\_prediction\_tree\_regressor()"

"explain\_prediction\_lightning()"

"explain\_prediction\_lightgbm()"

"explain\_prediction\_xgboost()"

3)show\_weights()

This method works exactly like "explain\_weights()" with the only difference being that it returns an IPython.display.HTML object instead. We can display this object in a Jupyter notebook to show global weights.The usage of this method is the same as calling explain\_weights() to generate an Explanation object and then calling format\_as\_html() with an explanation object to display it. Therefore we have used this model to show the global feature importance in the project.

4)show\_prediction()

This method works exactly like "explain\_prediction()" with the only difference being that it returns an IPython.display.HTML object instead. We can display this object in the Jupyter notebook to show the local features' importance. The usage of this method is the same as calling explain\_prediction() to generate an Explanation object and then calling format\_as\_html() with an explanation object to display it. Hence we have used show\_prediction instead of explain\_prediction to show local feature importance.

**3.3.2 Natural Language Toolkit**

We use the Natural Language Toolkit(NLTK) to convert this text into features. It provides lots of text processing libraries to work on the text formats in machine learning.

Natural language toolkit(NLTK) has to be installed in your system to work on it. The NLTK module is a massive toolkit, aimed at helping you with the entire Natural Language Processing (NLP) methodology.

In order to install NLTK run the following commands in your terminal,

1. sudo pip install nltk

2. Then, enter the python shell in your terminal by simply typing python

3. Type import nltk

4. nltk.download(‘all’)

In this project we used NLTK for text tokenizing. So basically, tokenizing involves splitting sentences and words from the body of the text.

**3.3.2.1 Stop Words**

Stop word is a commonly used word (such as “the”, “a”, “an”, or “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or take up the valuable processing time. For this, we can remove them easily, by storing a list of words that you consider stop words. We have imported stop words from NLTK because this will be used in converting text into tokens most of the time we remove the stop words from the text, hence we use NLTK to identify the stop words and extract exact tokens. stopwords example table,

**Table 3.2**: Example for Stop Words.

| **Sample text with stop words** | **without stop words** |
| --- | --- |
| GeeksforGeeks - A computer science portal for Geeks. | GeeksforGeeks Computer Science, Portal, Geeks. |
| Can listening be exhausting? | Listening, Exhausting |
| I like reading, so I read | Like, Reading, Read |

**3.3.2.2 N-grams**

We used n-grams from nltk.util. These are used in grouping the words together by a specific number. N-gram can be defined as the contiguous sequence of n items from a given sample of text or speech. The items can be letters, words, or base pairs according to the application. The table shows N-grams,

**Table 3.3**: N-grams.

| **n** | **Term** |
| --- | --- |
| 1 | Unigram |
| 2 | Bigram |
| 3 | Trigram |
| n | n-gram |

As clearly depicted in the table above, when n=1, it is said to be a unigram. When n=2, it is said to be a bigram, and so on. An example of n-grams. Let’s understand n-grams practically with the help of the following sentence: “I reside in Bengaluru”;

**Table 3.4** : Examples of N-gram types.

| **S.No.** | **Type of n-gram** | **Generated n-grams** |
| --- | --- | --- |
| 1 | Unigram | [“I”,”reside”,”in”,  “Bengaluru”] |
| 2 | Bigram | [“I reside”,”reside in”,”in  Bengaluru”] |
| 3 | Trigram | [“I reside in”, “reside in  Bengaluru”] |

From the table above, it’s clear that unigram means taking only one word at a time, bigram means taking two words at a time and trigram means taking three words at a time.

**3.3.3 Scikit-learn**

The sklearn.feature\_extraction module can be used to extract features in a format supported by machine learning algorithms from datasets consisting of formats such as text and image. The sklearn.feature\_extraction.text submodule gathers utilities to build feature vectors from text documents. The sklearn.feature\_selection module implements feature selection algorithms. It currently includes univariate filter selection methods and the recursive feature elimination algorithm.

**3.3.3.1 CountVectorizer**

In order to use textual data for predictive modeling, the text must be parsed to remove certain words, this process is called tokenization. These words need to then be encoded as integers, or floating-point values, for use as inputs in machine learning algorithms. This process is called feature extraction (or vectorization). CountVectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. CountVectorizer creates a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample.

**3.3.3.2 TF-IDF vectorizer**

We have used TfidfVectorizer to extract features from the text it is available in the scikit learn library. Text vectorization algorithms, namely TF-IDF vectorizer, which is a very popular approach for traditional machine learning algorithms, can help in transforming text into vectors.

In this scheme, features and samples are defined as follows. Each individual token occurrence frequency (normalized or not) is treated as a feature. The vector of all the token frequencies for a given document is considered a multivariate sample. A corpus of documents can thus be represented by a matrix with one row per document and one column per token (e.g. word) occurring in the corpus. We call vectorization the general process of turning a collection of text documents into numerical feature vectors. Parameters are defined in detail below,

1) input {‘filename’, ‘file’, ‘content’}, default=’content’, If 'filename', the sequence passed as an argument to fit is expected to be a list of filenames that need reading to fetch the raw content to analyze. If 'file', the sequence items must have a ‘read’ method (file-like object) that is called to fetch the bytes in memory. If 'content', the input is expected to be a sequence of items that can be of type string or byte.

2) sublinear\_tf , apply sublinear tf scaling, i.e., replace tf with 1 + log(tf). Sublinear tf-scaling is a modification of term frequency. It keeps sublinear\_tf= True in this project.

3) analyzer {‘word’, ‘char’, ‘char\_wb’} or callable, default=’word’, Whether the feature should be made of word or character n-grams. Option ‘char\_wb’ creates character n-grams only from text inside word boundaries; n-grams at the edges of words are padded with space. In our project, we used words, since we need to make the feature out of words.

4) token\_pattern: str, default=r ”\b\w\w+\b”, Regular expression denoting what constitutes a “token”, only used if analyzer == 'word'. The default regex selects tokens of 2 or more alphanumeric characters (punctuation is completely ignored and always treated as a token separator). In our code we used the regular expression token\_pattern=r'\w{1,}' this means that it accepts the word characters. \w means a word character.

5) ngram\_range, An n-gram is just a string of n words in a row. E.g., the sentence 'I am Groot' contains the 2-grams 'I am' and 'am Groot'. The sentence is itself a 3-gram. Set the parameter ngram\_range=(a,b) where a is the minimum and b is the maximum size of n-grams you want to include in your features. In this project, we used ngram\_range=(1, 2) which means it accepts unigrams and bigrams.

6) min\_df, The other parameter we used is min\_df. min\_df is used for removing terms that appear too infrequently. For example, min\_df = 0.01 means "ignore terms that appear in less than 1% of the documents". Min\_df means minimum document frequency. We have used min\_df=5 in this project.

**3.4 Regression Algorithms**

In this particular section we have tried to predict revenue based on a few attributes provided in the data set like popularity, budget, revenue, language, etc. Basically in our data set the problem is we are trying to predict the revenue based on few values available, hence it is a regression problem. To do this we have used 2 algorithms that are linear regression and random forest regressor. We will try to learn how we have implemented these algorithms in detail in the below sections. Regression searches for relationships among variables. For example, you can observe several employees of some companies and try to understand how their salaries depend on their features, such as experience, education level, role, city of employment, and so on. This is a regression problem where data related to each employee represents one observation. The presumption is that experience, education, role, and city are independent features, while the salary depends on them. In the same way, as we observe our data we have runtime, popularity and budget are independent features, while revenue depends on them.

Generally, in regression analysis, you consider some phenomenon of interest and have a number of observations. Each observation has two or more features. Following the assumption that at least one of the features depends on the others, you try to establish a relation among them. The dependent features are called the dependent variables, outputs, or responses. The independent features are called the independent variables, inputs, regressors, or predictors

It’s a common practice to denote the outputs with 𝑦 and the inputs with 𝑥. If there are two or more independent variables, then they can be represented as the vector 𝐱 = (𝑥₁, …, 𝑥ᵣ), where 𝑟 is the number of inputs In our project, the independent variables are also represented by a vector,

x=Data[‘popularity’,’ runtime’,’ budget’]

y=Data[‘revenue’]

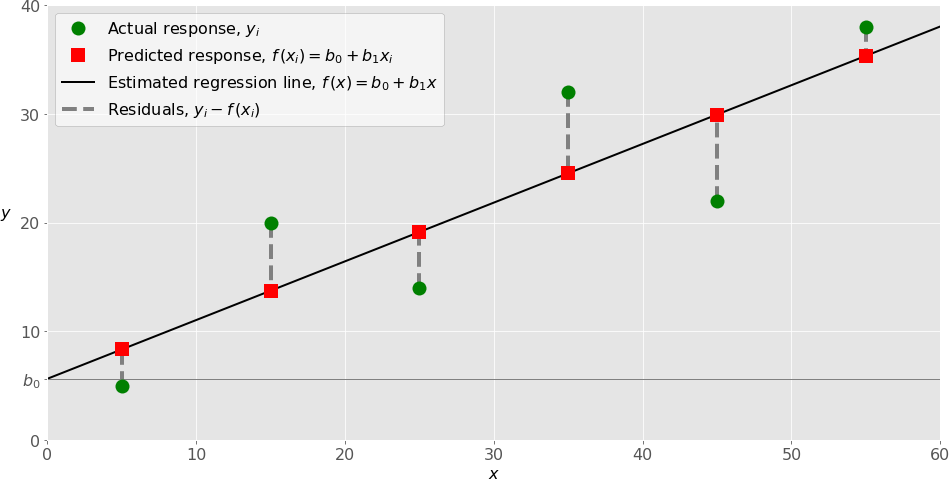
After categorizing data into independent and dependent features that are our x and y vectors we need to use train-test split. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset. The objective is to estimate the performance of the machine learning model on new data: data not used to train the model. This is how we expect to use the model in practice. Namely, to fit it on available data with known inputs and outputs, then make predictions on new examples in the future where we do not have the expected output or target values.

**3.4.1 Linear Regression**

We have implemented two regression algorithms in this project. The first one is linear regression. Linear regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results. When implementing linear regression of some dependent variable 𝑦 on the set of independent variables 𝐱 = (𝑥₁, …, 𝑥ᵣ), where 𝑟 is the number of predictors, you assume a linear relationship between 𝑦 and 𝐱: 𝑦 = 𝛽₀ + 𝛽₁𝑥₁ + ⋯ + 𝛽ᵣ𝑥ᵣ + 𝜀. This equation is the regression equation. 𝛽₀, 𝛽₁, …, 𝛽ᵣ are the regression coefficients, and 𝜀 is the random error. According to our problem, we have independent variables such as {popularity, runtime, budget}, and we are implementing linear regression on revenue. It is a dependent variable.

**3.4.1.1 Simple Linear Regression**

Simple or a single-variate linear regression is the simplest case of linear regression, as it has a single independent variable, x = 𝑥.The following figure illustrates simple linear regression. When implementing simple linear regression, you typically start with a given set of input-output (𝑥-𝑦) pairs. These pairs are your observations, shown as green circles in the figure. For example, the leftmost observation has the input 𝑥 = 5 and the actual output, or response, 𝑦 = 5. The next one has 𝑥 = 15 and 𝑦 = 20.



**Figure 3.1** : Regression line.

The estimated regression function, represented by the black line, has the equation 𝑓(𝑥) = 𝑏₀ + 𝑏₁𝑥. Your goal is to calculate the optimal values of the predicted weights 𝑏₀ and 𝑏₁ that minimize SSR and determine the estimated regression function. The value of 𝑏₀, also called the intercept, shows the point where the estimated regression line crosses the 𝑦 axis. It’s the value of the estimated response 𝑓(𝑥) for 𝑥 = 0. The value of 𝑏₁ determines the slope of the estimated regression line. The predicted responses, shown as red squares, are the points on the regression line that correspond to the input values. For example, for the input 𝑥 = 5, the predicted response is 𝑓(5) = 8.33, which the leftmost red square represents. The vertical dashed gray lines represent the residuals, which can be calculated as 𝑦ᵢ - 𝑓(xᵢ) = 𝑦ᵢ - 𝑏₀ - 𝑏₁𝑥ᵢ for 𝑖 = 1, …, 𝑛. They’re the distances between the green circles and red squares. When you implement linear regression, you’re actually trying to minimize these distances and make the red squares as close to the predefined green circles as possible. This is the example for simple linear regression but in our project, we have more than one independent variable so it is considered as multiple linear regression.

**3.4.1.2 Multiple Linear Regression**

Multiple or multivariate linear regression is a case of linear regression with two or more independent variables. If there are just two independent variables, then the estimated regression function is 𝑓(𝑥₁, 𝑥₂) = 𝑏₀ + 𝑏₁𝑥₁ + 𝑏₂𝑥₂. It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights 𝑏₀, 𝑏₁, and 𝑏₂ such that this plane is as close as possible to the actual responses, while yielding the minimal SSR.

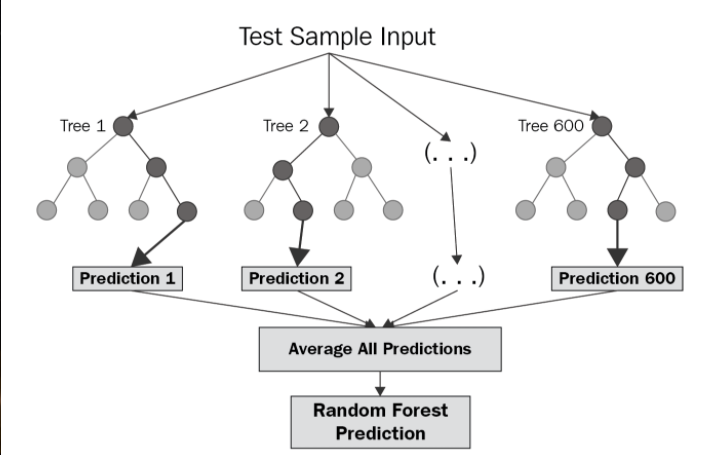
The case of more than two independent variables is similar but more general. The estimated regression function is 𝑓(𝑥₁, …, 𝑥ᵣ) = 𝑏₀ + 𝑏₁𝑥₁ + ⋯ +𝑏ᵣ𝑥ᵣ, and there are 𝑟 + 1 weights to be determined when the number of inputs is 𝑟. Implementing linear regression starts with the simplest case, which is simple linear regression. There are five basic steps when implementing linear regression:

1. Import the packages and classes that you need.
2. Provide data to work with, and eventually do appropriate transformations.
3. Create a regression model and fit it with existing data.
4. Check the results of model fitting to know whether the model is satisfactory.
5. Apply the model for predictions.

These steps are more or less general for most of the regression approaches and implementations.

**3.4.2 Random Forest Regression**

The other algorithm we have used in the project for revenue prediction is random forest regression. Random Forest Regression is a supervised learning algorithm that uses an ensemble learning method for regression. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, and we must choose the number of trees to include in the model. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



**Figure 3.2** : Structure of random forest.

The diagram above shows the structure of a Random Forest.The trees run in parallel with no interaction among them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. Implementing random forest regression. To get a better understanding of the Random Forest algorithm, let’s walk through the steps:

1. Importing the libraries.
2. Importing the data set.
3. Splitting the dataset into the Training set and Test set.
4. Training the Random Forest Regression model on the training set.
5. Predicting the Results.

We followed those steps for the importing of the random forest regressor model for our data set.

**3.5 Data Visualization**

In this section we will be discussing data visualization, Data Visualization is the graphical representation of information and data. By using usual elements like charts, graphs, and maps. These visual displays of information communicate complex data relationships and data-driven insights in ways that are easy to understand. In this project we used two different data visualization libraries, they are matplotlib and seaborn.

**3.5.1 Matplot library**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in python. matplotlib makes easy things easy and hard things possible.

With matplotlib, we can,

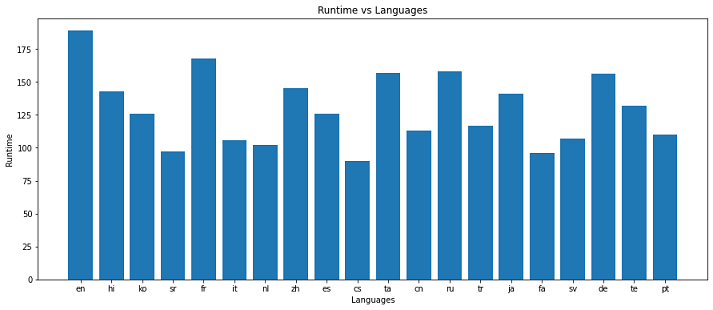
* create a publication of quality plots.
* Make interactive figures that can zoom.
* Customize to many file formats.
* Embedded in Graphical User Interface.

Various matplotlib plotting techniques used in this project are the following

1. Bar Chart
2. Histogram
3. Scatterplot

**3.5.1.1 Bar Chart**

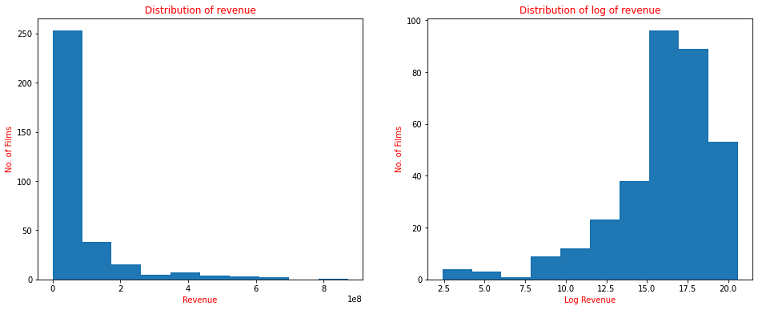
Bar Charts are divided into sections that represent parts of the whole. They provide a simple way to organize data and compare the size of each component to one another. In this project, we used a bar chart in different analyses like, comparison of different languages and their runtime, used for plotting runtime and languages of films, from that graph we could analyze languages like English, French, and Russian films have more runtime than compared to other language films as shown below figure ,



**Figure 3.3** : Bar Chart.

**3.5.1.2 Histogram**

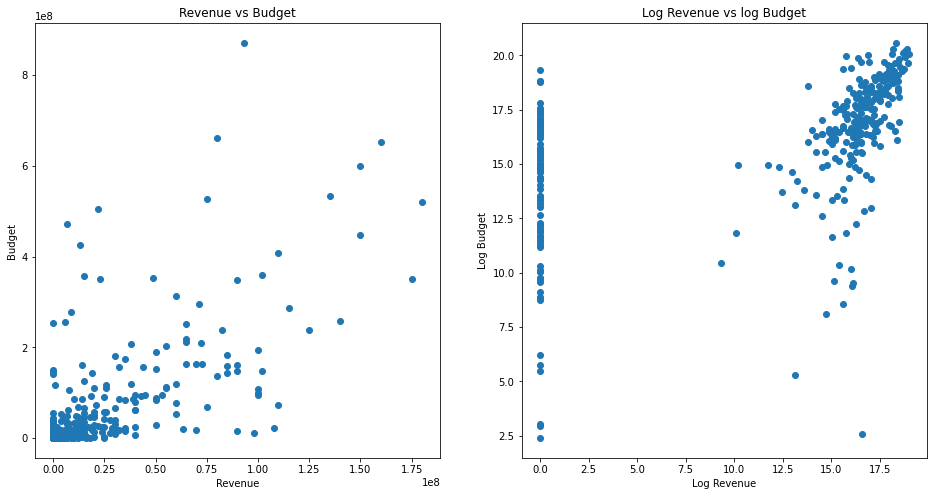
Histogram plots a distribution of numbers using a bar chart with no spaces between bars, representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.like the distribution of overall revenue from the dataset so that we can understand revenue gained by many films, as shown below in figure,



**Figure 3.4** : Histogram.

**3.5.1.3 Scatterplot**

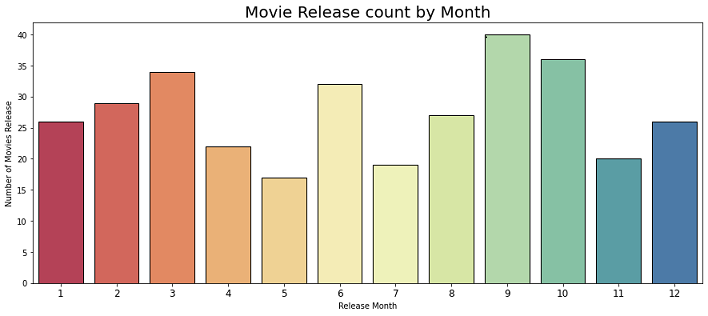
Scatterplot these visuals are beneficial in revealing the relationship between two variables, and they are commonly used with regression data analysis. We used the Scatterplot in particular scenarios like finding the correlation between the budget and revenue for the given data set as shown below in figure,



**Figure 3.5** : Scatter plot

**3.5.2 Seaborn**

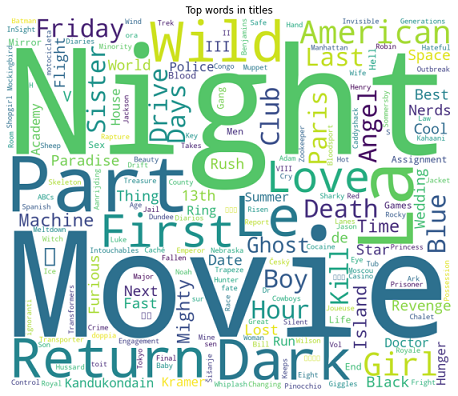
Seaborn is a python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. And it helps to understand the data, however complex the significance of data by summarizing and presenting a huge amount of data in a simple and easy-to-understand format and helps communicate information clearly and effectively. Seaborn provides various plots, some plots used in this work are countplot, which show the counts of observations in each categorical bin using bars, a countplot can be thought of as a histogram across a categorical instead of a quantitative variable. In our work, we used countplot for the number of movies released in each month and the number of movies released in each corresponding year, as shown below in figure,



**Figure 3.6** : Count plot.

**3.5.3 Word Cloud**

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using the word cloud. This library is widely used for analyzing textual data from social media, the internet, and Digital forums so that we can see different trends across the internet. In our project, we used this library to find frequently used words in film titles and film descriptions, which shows an output of words with varying sizes showing their importance or frequency as shown below in figure,



**Figure 3.7** : Top words used in titles.

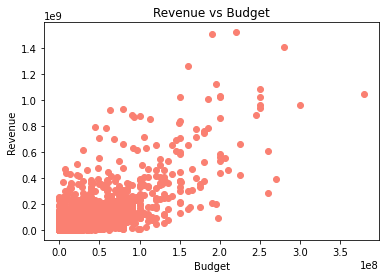
**CHAPTER 4**

**RESULTS AND DISCUSSIONS**

In this section we will be discussing the results, and output we got, and also we will be discussing a few metrics like accuracy, performances of some models, and a few data visualizations to interpret the results.

**4.1 Revenue v/s Budget**

Below scatter plot is plotted between revenue and budget that gives a better interpretation of the relation of the budget with revenue. Movies with very high budgets gained proportionate revenue, and only very few movies with the highest budget were produced, from the plot we can find the budget of the respective film to revenue. It is shown in the below figure,



**Figure 4.1** : Revenue vs Budget

**4.2 Top Ten Highest Revenue Films**

According to the Performed analysis we got “The Avengers”, as a film which has gained the highest revenue of 151.9 crores USD with a budget of 22 crores USD. Given the whole dataset, The Avengers has the highest revenue and the second highest collection made by a film is Furious 7, its revenue is 150.6 crores USD. Also, the top ten films with the highest revenue are shown in the output below,



**Figure 4.2** : Top Ten Movies With Highest Revenue.

**4.3 Top Ten Highest Budget Films**

The top ten films with the highest budget are Pirates of the Caribbean: On Stranger Tides with the highest budget of 38 crores USD which has collected revenue of 104.5 crores USD, the second highest budget film is Pirates of the Caribbean: At World’s End, its budget is 30 crores USD and the third highest budget film was Avengers: Age of Ultron, its budget is 28 crores USD.



**Figure 4.3** : Top Ten Highest Budget Films.

**4.4 Top ten popular films with their revenue**

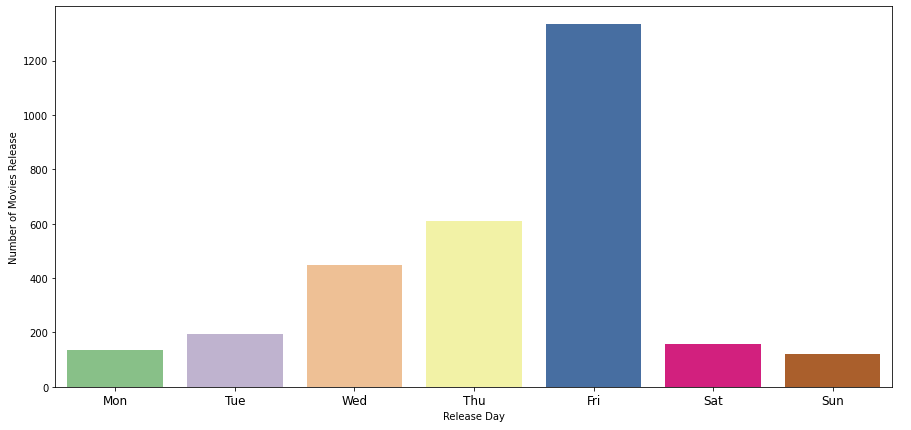
Wonder Woman was the most popular movie with a popularity value of 294.34 and it has collected 82 crores USD. In our analysis, we have sorted the top ten films according to their popularity with their respective revenue gained. We have also visualized it using a background gradient which is also an effective analytical tool.



**Figure 4.4** : Top Ten Popular Films

**4.5 Number Of Movies Released On The Respective Day**

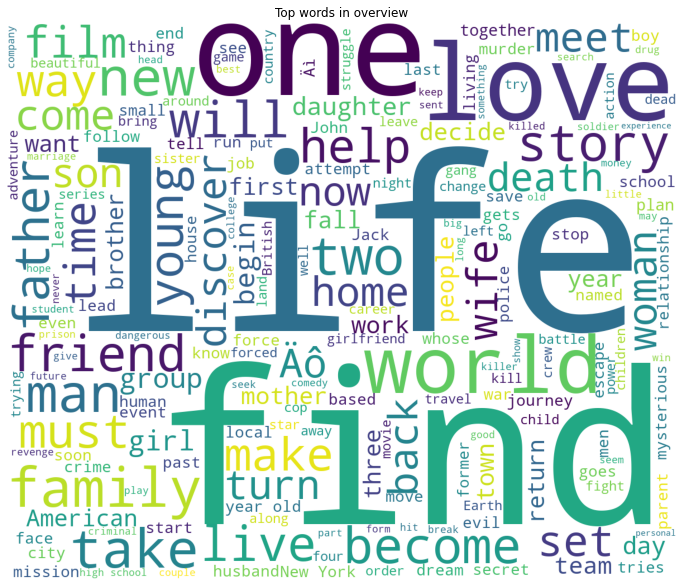
Most of the movies are released on Fridays so that people can watch them and give reviews about them. If it's good then people will watch them on Saturday and Sunday as it is a weekend. The number of movies released on each day of the week, more than six hundred movies were released on Thursday which is the second highest number of movie releases. On Sunday fewer movies are released, also on Monday's number of movie releases is low.



**Figure 4.5** : Movie releases in a week.

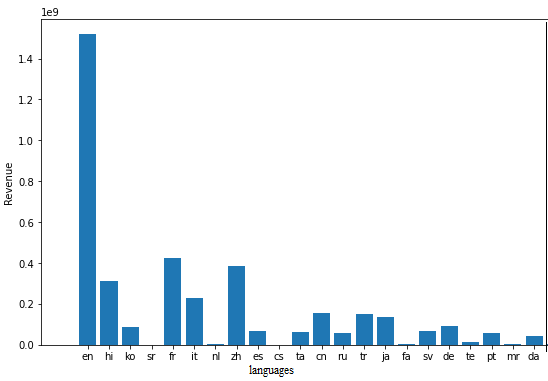
**4.6 Top words used in film descriptions**

Most words used in film descriptions are indicated by the size of the word. As we can see in the figure below we can observe that words are scattered with different sizes. If we consider the word life it has the largest size comparatively, so it is the most frequently used word in the film descriptions, words like drug, evil, and killed are very small in size so they are used very less in the film descriptions.



**Figure 4.6** . Top words used in Film Overview.

**4.7 Revenue and Languages**

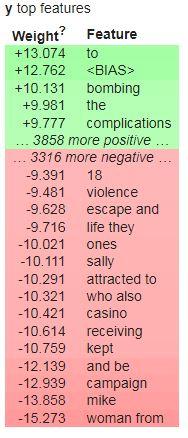


**Figure 4.7** : Languages Revenue Collection.

From the above plot we can clearly conclude that, highest grossing language is English and revenue collection is greater than 140 crores USD. English movie revenue collections are much higher when compared with other language movies: French, Hindi, Chinese and Korean. French movies have second highest revenue which is 40 crores USD and Chinese movies have third highest revenue collection. English movies are the only movies in the 100 crores USD club, but other language movies are having average collections of 35 crores USD. So, Hollywood is making more revenue in the film industry.

**4.8 Analysis Of The Film Descriptions**

Visualizing feature Importance of a film description, if we observe text from the description of the movie L’Armee des Ombres we can see the words in two color gradients red and green, the word betrayed is green which means it has a better effect on the revenue, and the word occupied is a dark red that means it has a poor effect on the revenue. In this way, we locally analyze the feature importance of the movie overview on the revenue.Here the green color indicates a strong feature and the red color indicates poor features.



**Figure 4.8** : Feature Importance of all descriptions.

Output film description of the film L’Armee des Ombres. Words highlighted in green color are considered good features, as the green color intensity value decreases then its feature importance also reduces.‘betrayed’ word acts as a good feature and has a positive impact on creating a good film description, so it has a positive effect on the sale of the movie tickets, eventually rise of collections of the movie. While words highlighted with red color have less importance, also when the intensity of red decreases then words importance increases. ‘occupied’ in the film description given below has very low importance.

Betrayed by an informant, Philippe Gerbier (Lino Ventura) finds himself trapped in a torturous Nazi prison camp. Though Gerbier escapes to rejoin the Resistance in occupied Marseilles, France, and exacts his revenge on the informant, he must continue a quiet, seemingly endless battle against the Nazis in an atmosphere of tension, paranoia, and distrust.

**4.9 Revenue Prediction**

During the revenue prediction we have used the following algorithms. The algorithms we used and their respective accuracies are in the table below. Using the Linear regression model we achieved an accuracy of 62.6%, and by using Random Forest Regressor model we achieved an accuracy of 83%. The Random Forest Regressor model could nearly predict the actual output values of revenue with inputs given, whereas using the Linear Regression model predicted values of revenue were approximate compared to the actual values.

**Table 4.1** : Accuracy of the models used.

| S.No. | Model | Accuracy |
| --- | --- | --- |
| 1 | Linear Regression | 62.6 % |
| 2 | RandomForestRegressor | 83 % |

**CHAPTER 5**

**CONCLUSIONS**

In this project we have performed an analysis of all the important features of the box office revenue of a movie. Using machine learning techniques and statistical methods for analyzing the features affecting the income of the movie. We have followed all the steps of machine learning from data collecting to preprocessing step by step and carefully completed the model based on the data set we used. Some of the results we have concluded from this project are how the revenues of the movies are weak and how they can be strong. We have tried to answer such questions.

According to the analysis we performed, we found attributes like runtime, budget, and popularity are affecting the movie's box office. In revenue and budget analysis, we have compared strong factors affecting the movie’s results like revenue vs budget using graph techniques available in machine learning. From the analysis, we could conclude that movies with large budgets had good revenue compared to movies with low budgets.

In the graphs, we have visualized features, runtime, and revenue, from the graphs we got to know that the movies with a runtime of 2 to 2.5 hours, did very well compared to the movies with large or very low runtimes in this way we have used graphs for visualization of the features. Another conclusion seems to be that the popularity of films also plays an important role in bringing the audience to the theaters. According to the sorting we have done on the movies based on their popularity we can conclude that movies with high popularity had better income than those with low popularity.

The most important analysis done in this project is extracting the features to be used in writing the proper film overview or description. To conclude this issue we have found the features which have a better impact on the revenue and also we extracted the features which have a bad impact on the revenue. An avenue for further research is to see whether these similar approaches can be used in real-time and present-day data.

**CHAPTER 6**

**REFERENCES**

[1] Breuss, M., de Rijke, M., Oghina, A., & Tsagkias, M., 2012.“Predicting IMDB movie ratings with social media”. In proceedings of the 34th European Conference on IR Research: Advances in Information Retrieval, ECIR, Barcelona, Spain, Eds. R. Baeza-Yates & A.P. de Vries & H. Zaragoza & B.B. Cambazoglu & V. Murdock & R. Lempel & F. Silvestri, LNCS vol 7224, pp. 503-507, Springer.

[2] Chang, W., & Lee, K.J., 2009. “Bayesian belief network for box-office performance: A case study on Korean movies”. Expert Systems with Applications, Elsevier, vol. 36(1), pp. 280–291.

[3] Delen, D. & Sharda, R., 2006. “Predicting box office success of motion pictures with neural network”. Expert Systems with Applications, Elsevier, vol. 30 (2), pp. 243–254.

[4] Kaur, A. & Nidhi, A.P., 2013. “Predicting movie success using Neural Network”. International Journal of Science and Search, India, online, vol. 2(9),pp. 69-71.

[5] Rotten Tomatoes, 2022. Retrieved from http://www.rottentomatoes.com/.

[6] Zhang, W., and Skiena, S., 2009. “Improving Movie Gross Prediction through News Analysis”. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent

Agent Technology, vol. 1, pp. 301-304, IEEE Computer Society, Washington, DC, USA.

[7] Li, Z., Jianhua, L., & Suying, Y., 2009. “Forecasting box office revenues of movies with BP neural network”. Expert Systems with Applications, Elsevier, vol. 36(3), Part 2, pp. 6580–6587.

[8] Internet Movie Database,2022. Retrieved from http://www.imdb.com/.

[9] The Movie Database, 2022. Retrieved from https://www.themoviedb.org

[10] Skyler, Frank, and Victor 2012. “A study on box-office revenue

How user and expert ratings determine movie success” . Elsevier, vol. 36(1), pp. 280–291.

[11] Bae, Giwoong, Kim, Hye-jin, 2019. The impact of movie titles on box office success. J. Bus. Res. 103, 100–109.

[12] Du, J.; Xu, H.; Huang, X. Box office prediction based on microblog. Expert Syst. Appl. 2014, 41, 1680–1689.

[13]Basuroy, S.; Ravid, S.C.A. How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets.J. Mark. 2003, 67, 103–117.

[14]Gaenssle, S.; Budzinski, O.; Astakhova, D. Conquering the box office: Factors influencing success of international movies in Russia. Rev. Netw. Econ. 2018, 17, 245–266.