**MOVIE BOX OFFICE REVENUE PREDICTION**

**Table of content:**

1. [Title Of Project 3](#_Toc96341549)
2. [Executive Summary 3](#_Toc96341550)
3. [Abstract 4](#_Toc96341554)
4. Introduction………………………..5
5. [Methodology 6](#_Toc96341555)

5.1.Data Extraction

5.2.Data Preprocessing

5.3.Feature Selection

5.4.Classification

1. [Results Section 8](#_Toc96341556)
2. [Discussion 10](#_Toc96341557)

7.1.Features

7.2.Impact of features on revenue

7.3.limitations and challenges

7.4.Comparision with existing research

1. [Conclusion 11](#_Toc96341558)
2. [Contributions/References 12](#_Toc96341559)

**1. TITLE OF PROJECT:**

**MOVIE BOX OFFICE REVENUE PREDICTION**

**2. EXECUTIVE SUMMARY:**

The Movie Box Office Revenue Prediction project aims to create a predictive model for estimating pre-release box office revenue, utilizing diverse data sources including movie metadata, casting information, and social media metrics. The methodology involves comprehensive data preprocessing, exploratory analysis, and experimentation with machine learning models. The project aims to deliver a trained model, an insights report highlighting key findings, and an optional user interface for user-friendly predictions. The potential impact of the project is to empower filmmakers and studios with data-driven insights, facilitating informed decision-making for marketing strategies, release timing, and resource allocation in the highly competitive film industry.

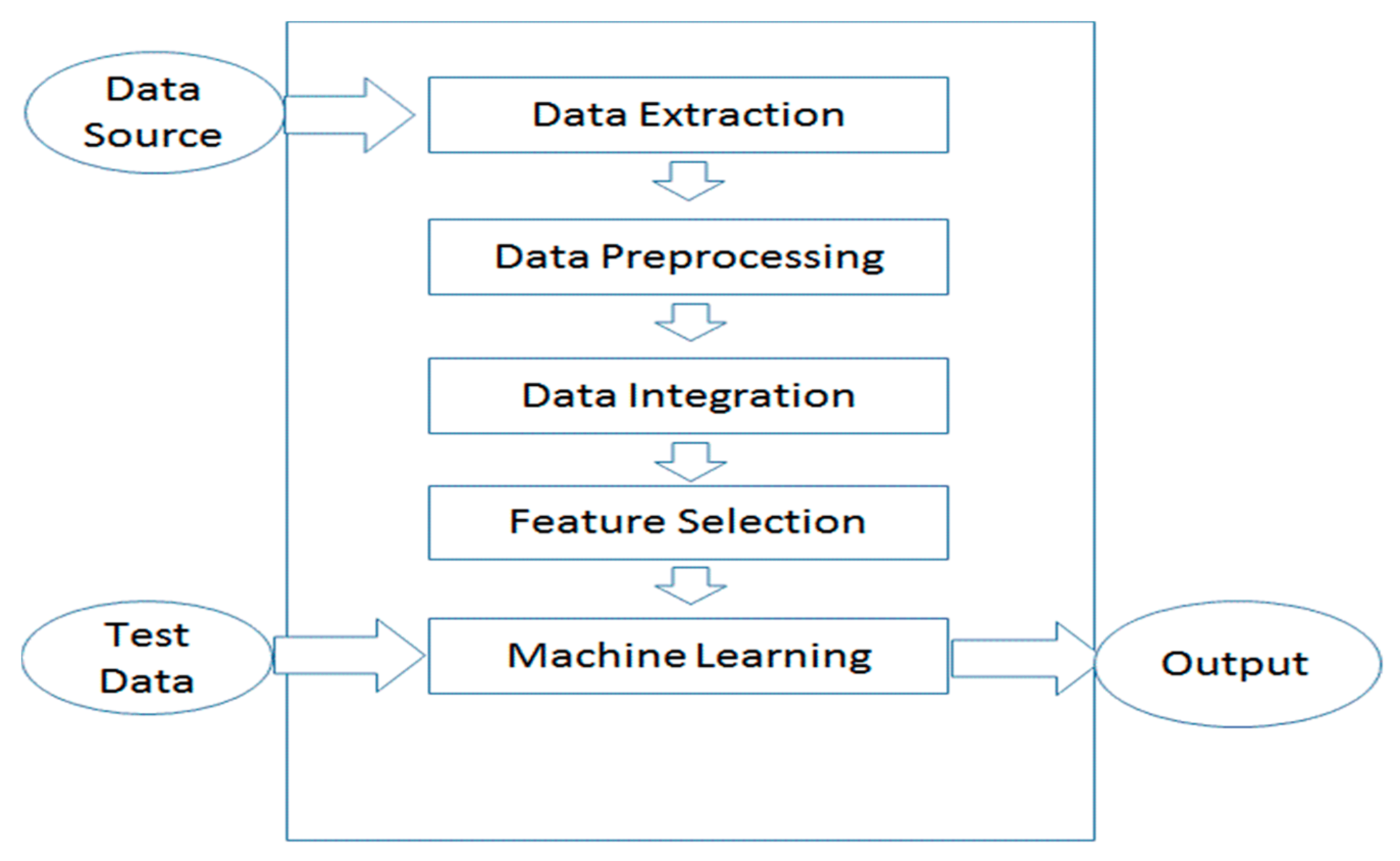
**3. ABSTRACT:**

This project focuses on the development of a predictive model for estimating box office revenue of movies prior to their release, employing a comprehensive dataset comprising movie metadata, casting details, and social media metrics. Through meticulous data preprocessing, exploratory data analysis, and the application of diverse machine learning models, the project aims to deliver a robust tool for filmmakers and studios. The predictive model, evaluated on its performance metrics, provides insights into the factors influencing box office success. An optional user interface facilitates user-friendly predictions. The project's potential impact lies in empowering industry stakeholders with data-driven decision-making, optimizing marketing strategies, release timing, and resource allocation in the dynamic landscape of the film industry.

**4. INTRODUCTION:**

The film industry is a dynamic and competitive landscape where the success of a movie is influenced by numerous factors. Understanding and predicting box office revenue before a movie's release is a critical aspect of strategic decision-making for filmmakers and studios. This project aims to address this challenge by developing a predictive model that leverages a comprehensive dataset, encompassing movie metadata, casting details, and social media metrics. The goal is to empower industry stakeholders with a tool that provides actionable insights, enabling them to optimize marketing strategies, choose optimal release timing, and allocate resources judiciously. Through the integration of advanced data science techniques and machine learning, this project seeks to contribute to a more informed and efficient approach to navigating the complexities of the film industry..

**5. METHODOLOGY:**



## The methodology is divided into 5 steps, the first one being data selection and extraction, The data is then preprocessed The preprocessed data set used with different algorithms to determine and select which features to use in the final data set used in the classification model

**5.1. DATA EXTRACTION:**

Kaggle is used as source for the data used in this study. Kaggle an open API. Explore the different datasets available in Kaggle and choose the appropriate dataset for the project. Here we used the following dataset for this project [(https://www.kaggle.com/c/tmdb-box-office-prediction/data).](file:///C:\Users\91913\Downloads\tmdb-box-office-prediction.zip)The movies are sorted in descending order by number of rating and contains movies that originate from the US as movies from other countries have their budget represented in their native currency. Multiple currencies would require further pre-processing to normalise the data which was avoided in this case

**5.2. DATA PREPROCESSING**:

Duplicate values are removed, cells without a value were interpreted as NaN instead of removing the entire row. Movies with less than 10000 ratings where taken out of the data set as these movies tended to be quite volatile and not representative of the majority of viewers. Movies with very few ratings also tend to have inflated ratings. Movies that made less than 1 million USD in gross box office revenue after taking inflation into account where also removed from the data set.

**5.3. FEATURE SELECION:**

The data set includes features that are categorical such as genre, which means that numerical transformations are inappropriate, thus feature selection is main technique of reducing the number of dimensions. Feature selection reduces the number of predictor variables by selecting a subset of predictor variables to create a prediction model.

**5.4. CLASSIFICATION:**

The following algorithms are used in the project

**5.4.1Random Forest:**

Random forests or random decision forests is an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time. For classification tasks, the output of the random forest is the class selected by most trees.

**5.4.2.Linear Regression:**

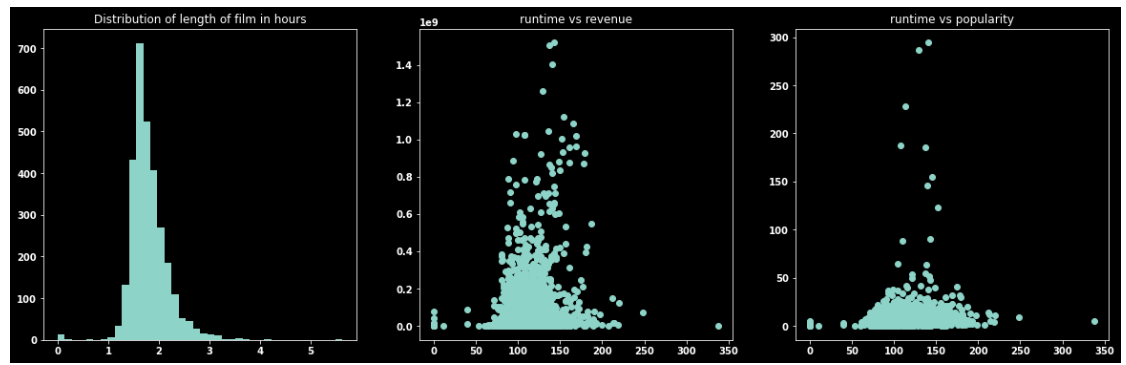
Linear Regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features **.**

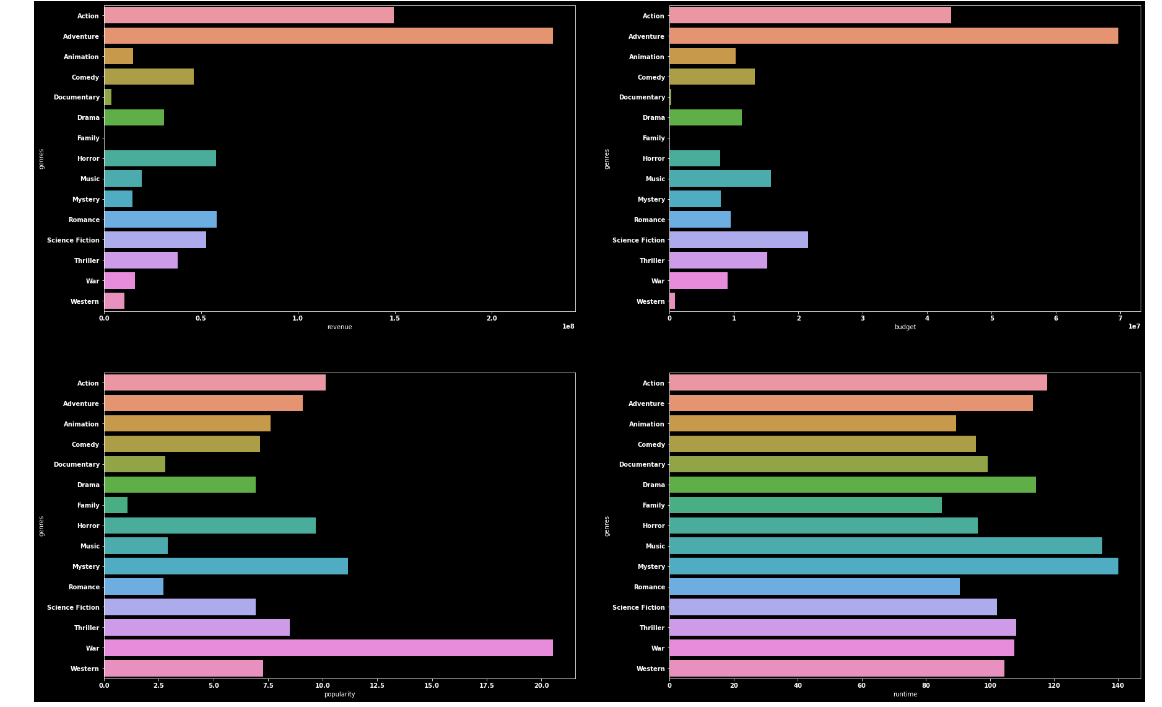
**5.4.3.Support Vector Regression:**

Support Vector Regression (SVR) is a regression algorithm that extends the concepts of Support Vector Machines (SVM) to solve regression problems. Similar to SVM for classification, SVR aims to find a hyperplane that best fits the data, while considering a margin of tolerance**.**

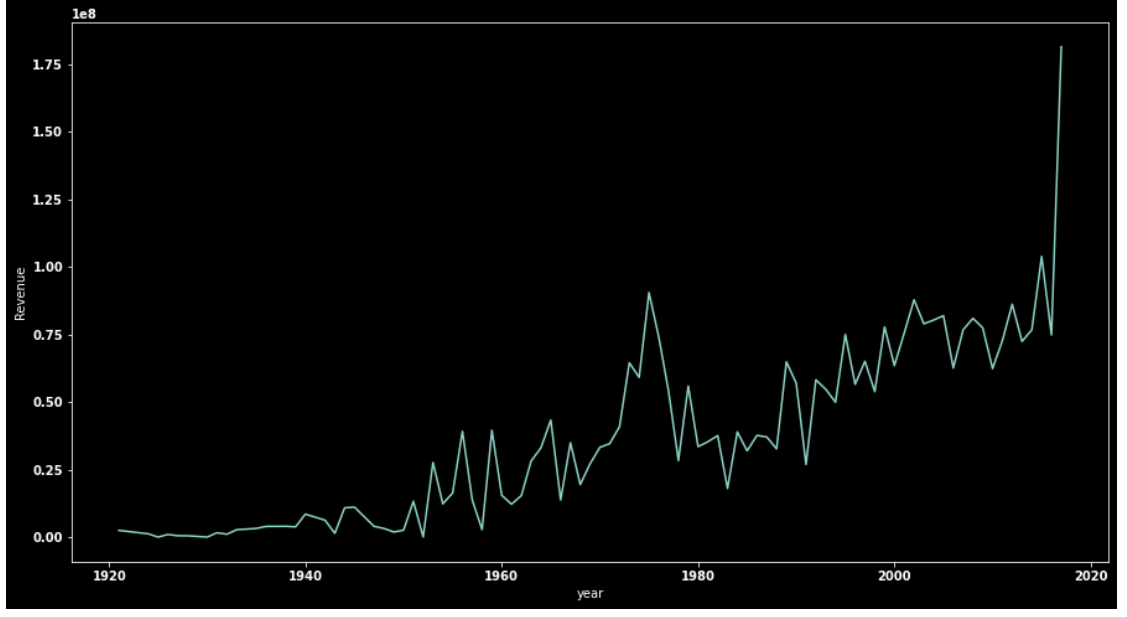
**6. RESULT:**

Sample output images of plots of a feature against other feature

****

****

Bar plot against the features runtime, revenue, budget and popularity .It clearly shows that runtime has a huge impact in different genres of movies**.**

****

Plot for revenue generated in different years. The revenue has been increased drastically in recent years.

**7. DISCUSSION:**

**7.1. FEATURES**

The results of this study is arguably on par with the other studies mentioned when predicting the rating. Since the aim of this study was to take the techniques used by previous studies and improve them adding more features and a larger amount of data, it is hard claim that to be a success. In part due to our achieved rate of successfully predicting the rating was equal or lower than all the studies used as theoretical basis for this study.

**7.2. Impact of Features on Revenue:**

Exploring the influence of specific features such as genre, casting choices, production budget, release timing, and social media engagement on box office performance, we observed a notable trends and patterns that contribute to a movie’s success**.**

**7.3. Limitations and Challenges**

New trends, technologies, or shifts in audience preferences may not be adequately captured in historical data. The limited historical data might make it difficult to predict the performance of innovative or unprecedented movies.

Access to comprehensive and up-to-date data, including marketing budgets, promotional strategies, and social media metrics, may be restricted. Incomplete or outdated data may impact the model's accuracy**.**

**7.4. Comparison with Existing Research**

Many existing models heavily rely on traditional features such as cast, director, and production budget. This approach might oversimplify the complex dynamics influencing box office performance and fail to capture emerging trends or unconventional successes.

**8. CONCLUSION:**

The primary aim of this study, focused on developing a method to predict movie success using historical data, did not yield direct success. The findings underscore the necessity of incorporating additional features for accurate predictions. In the context of rating prediction, the model achieved its highest success rate of approximately 65% with six groups and 83% with four groups. While these results may be considered promising, a comparison with similar studies reveals a comparable level of predictability. However, when extending the prediction task to box office revenue, the model's success rate reached only 15%, indicating the difficulty in achieving accurate predictions using the methods employed in this study. In contrast to prior research, it appears that our data preprocessing methods may not be optimal. Notably, a binary decision on whether the revenue would exceed the budget resulted in a success rate of approximately 60%. In conclusion, while using historical data to model and predict movie ratings shows promise and warrants further evaluation, the same approach does not hold promise for accurately forecasting box office revenue. Previous studies have demonstrated the potential for accurate predictions, but the methods employed in this study did not surpass or improve upon them.

**9.GITHUB LINK:**

https://github.com/Harshisampath/Team13

**10. REFERENCES:**

[1] Statista Inc. Global box office revenue from 2016 to 2020, 2016. URL https://www. statista.com/statistics/259987/global-box-office-revenue/.

[2] MathWorks. Classification learner, 2017. URL https://se.mathworks.com/ help/stats/classificationlearner-app.html.

[3] Roc curves and area under the curve explained, 2017. URL http://www. dataschool.io/roc-curves-and-auc-explained/.

[4] Jeff Schneider. Cross validation, 1997. URL http://www.cs.cmu.edu/ ~schneide/tut5/node42.html.

[5] Imdb weighted mean rating, 2017. URL http://www.imdb.com/help/show\_ leaf?votes.

[6] Nikhil Apte, Mats Forssell, and Anahita Sidhwa. Predicting movie revenue. CS229, Stanford University, 2011.

[7] Muhammad Hassan Latif and Hammad Afzal. Prediction of movies popularity using machine learning techniques, 2016. http://paper.ijcsns.org/07\_book/ 201608/20160820.pdf.

[8] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, volume 1, pages 492–499. IEEE, 2010.