## Data Science Project Report

# Movie Revenue Predictor

CT-21013: Data Science

Group Members
Prathamesh Wani - 112003152
Isha Surve - 112003142
Siddhi Shinde - 112003137

Project Guide
Dr.Y.V.Haribhakta
Department of Computer Engineering,
COEP Technological University,
Shivaji Nagar, Pune

# CONTENTS

## Introduction:

The Movie Revenue predictor is a data science project aimed at building a model that can predict the revenue of a movie based on various features. The project utilizes machine learning techniques to train the model and uses a dataset of movie features and their corresponding revenues to evaluate the model's accuracy.

## Pipeline:

The pipeline of the Movie Revenue predictor project can be broadly divided into five stages, namely Data Collection, Data Preprocessing, Exploratory Data Analysis, Model Training, and Model Evaluation.

## Data Collection:

The first stage of the pipeline is Data Collection. In this stage, we tried to gather data from various sources, such as IMDb, Rotten Tomatoes, and Wikipedia. We collect data on various features of movies such as budget, genre, director, actors, actresses, release date, runtime, and ratings. This data was usually available in the form of APIs, or web scraping.

## Data Preprocessing:

The second stage of the pipeline is Data Preprocessing. In this stage, we clean and transform the collected data to prepare it for model training.

- We remove missing values, handled outliers, encode categorical variables, and normalize the numerical variables.
- We adjusted the data of box office collection and the budget of the movies from the 1930s to 2023 according to the inflation rates.
- The Release date data was encoded into holiday, non-holiday, and near-holiday dates.
- The genres were given values for the model to understand them.
- The Actor and actresses’ ratings were calculated and saved for every instance.

## Exploratory Data Analysis:

EDA was done on the cleaned preprocessed data to find any hidden trends which can be helpful for the model. Such as trends in box-office reports of overall movies, Optimum genre Quantity, Release date impact, Runtime effects, Excessive budget Quotations, etc.

## Model Training:

The Fourth stage of the pipeline is Model Training. In this stage, we train various machine learning models such as Linear Regression, Random Forest, and Decision tree Regressor. We use the cleaned and transformed data to train these models and evaluate their performance using various metrics such as Mean Squared Error, R-squared, and Mean Absolute Error. We then choose the best-performing model based on these metrics.

## Model Evaluation:

The fifth and final stage of the pipeline is Model Evaluation. In this stage, we use the chosen model to predict the revenue of a new movie based on its features. We also evaluate the model's accuracy on a test dataset to ensure that it generalizes well and is not overfitting to the training dataset.

## Conclusion:

The Movie Revenue Predictor project is a data science project aimed at building a model that can predict the revenue of a movie based on various features to help reduce the high risks involved in making a movie. The project utilizes machine learning techniques to train the model and uses a dataset of movie features and their corresponding revenues to evaluate the model's accuracy. The results satisfyingly entail the film makers about their combination of elements for making a movie are good or not.