Movie Revenue Prediction
SML Project

Vikranth Udandarao
Computer Science & Engineering Dept.
IIIT-Delhi, India
vikranth22570@iiitd.ac.in

Pratyush Gupta
Computer Science & Engineering Dept.
IIIT-Delhi, India
pratyush22375@iiitd.ac.in

***Abstract*—In the contemporary film industry, accurately predicting a movie's earnings is paramount for maximizing profitability. This project aims to develop a machine learning model for predicting movie earnings based on input features like budget, director, genre, leading crew, and IMDb ratings. Through a structured methodology involving data collection, preprocessing, analysis, model selection, evaluation, and improvement, a robust predictive model is constructed. Linear Regression, Decision Trees and Random Forest Regression have been tested . Model improvement strategies include hyperparameter tuning and feature selection. The resulting model offers promising accuracy and generalization, facilitating informed decision-making in the film industry to maximize profits.**

I. Introduction

A. Motivation

Imagine you are a filmmaker or head of a movie production house and you have a big question: what makes a movie a blockbuster hit or a flop?

You might think it depends on the star power of the actors, the vision of the director, the budget of the production, or the genre of the story.

Or you might think it is simply the quality of the storytelling that captivates the audience and earns high ratings. But the answer is not straightforward or easy.

There are many factors that influence the earnings of a movie, and the true combination of these factors has not been mastered yet. That's why we have developed a machine learning model that reveals the most important factors for box office success by analyzing real data from a wide variety of movies produced around the world. With our model, filmmakers can make more informed decisions and optimize their movie production for maximum profit and popularity.

B. Rationale

We hypothesize that certain parameters hold more significance in predicting movie revenue than others. Specifically, we conjecture that the director's track record and the genre of the film carry substantial weight in this prediction model.

Our observations suggest that despite lower IMDb ratings, action-oriented films often demonstrate strong performance at the box office. Conversely, genres such as comedy or emotional dramas, despite potentially higher IMDb ratings, may not achieve comparable revenue outcomes to their action counterparts.

These insights underscore the complex interplay between film attributes and audience preferences, prompting us to assign greater importance to factors like directorial history and genre classification within our predictive framework.

C. Overview

In this project, we follow a structured methodology to build and evaluate our predictive model.

We first collect a large dataset of movies and their features from various sources and custom tailor the datasets to suit our needs.

We then pre-process the data to handle missing values, outliers, and categorical variables. We perform data analysis to explore the data and understand its characteristics and relationships.

We use descriptive statistics, inferential statistics, and data visualization techniques to gain insights into the data like using a graph to compare the accuracy of our model's performance of training and test data.

We then select several machine learning algorithms that are suitable for regression tasks, such as decision trees and random forests. We train and test our models using cross-validation and compare their performance using metrics such as R-squared mean error.

We also apply model improvement strategies such as hyperparameter tuning, feature selection and regularization to enhance the accuracy and generalization of our models.

The resulting model offers promising results and can be used to predict the revenue of any movie based on its features.

II. Data Analysis

In the data analysis phase, we will thoroughly examine the collected movie dataset to gain insights into its structure, distribution, and relationships between features and the target variable (movie earnings). We dropped the unnecessary extra parameters and have come down to 5 parameters, which are the director's name, actor 1's name, actor 2's name, genre and budget.

References

[1] scikit-learn Models
   1) https://scikit-learn.org/stable/

[2] Kaggle Datasets

1) Movie Industry Dataset,
https://www.kaggle.com/datasets/danielgrijalvas/movies

2) IMDB 5000 Movies Multiple Genres Dataset,
https://www.kaggle.com/datasets/rakkesharv/imdb-5000-movies-multiple-genres-dataset
https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

3) Top 500 Movies Budget,
https://www.kaggle.com/datasets/mitchellharrison/top-500-movies-budget