Movie Revenue Prediction System

SML Project

IMDb

Done by:

Vikranth Udandarao
2022570

Pratyush Gupta
2022375

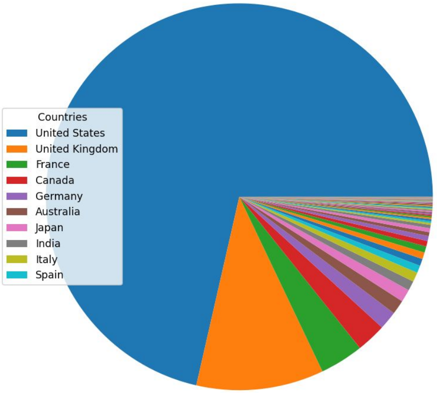
Introduction

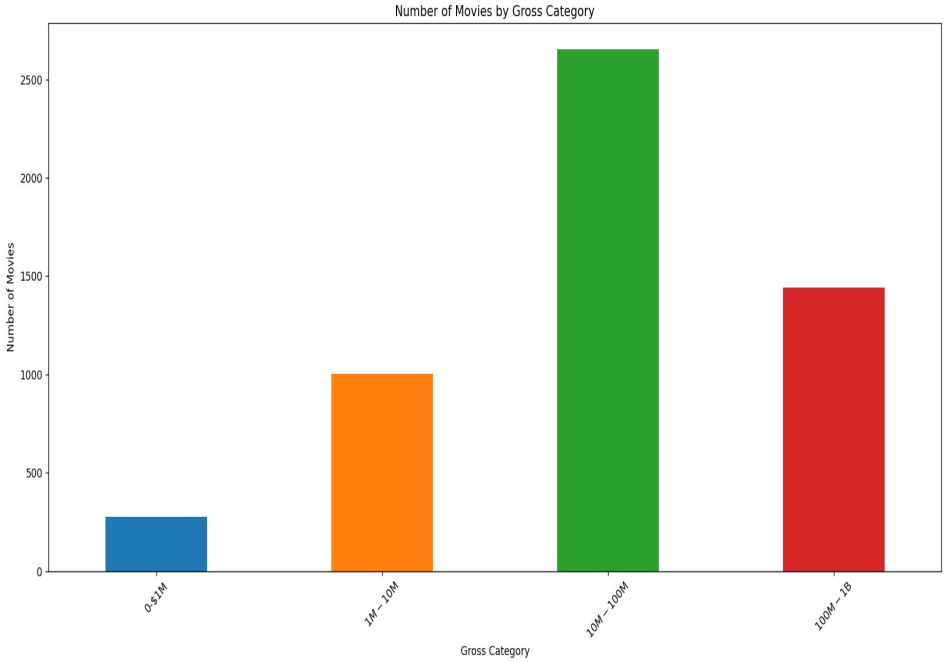
- Imagine you are a filmmaker or head of a movie production house and you have a big question: what makes a movie a blockbuster hit or a flop? You might think it depends on the star power of the actors, the vision of the director, the budget of the production, or the genre of the story. Or you might 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.

- 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 hyper-parameter 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

Dataset

1. We have used:
   a. Movie Industry Dataset
   b. IMDb 5000 Movies Multiple Genres Dataset
   c. IMDb 5000 Movies Dataset
   d. Top 500 Movies Budget
2. Features:
   a. Name
   b. Rating
   c. Genre
   d. Year
   e. Released
   f. Score
   g. Votes
   h. Director
   i. Writer
   j. Star
   k. Country
   l. Budget
   m. Company
   n. Runtime





Data Analysis

name	0
rating	77
genre	0
year	0
released	2
score	3
votes	3
director	0
writer	3
star	1
country	3
budget	2171
gross	189
company	17
runtime	4
dtype: int64	

Fig 1: Null Values

K Best Features

SelectKBest is a feature selection method in Scikit-Learn. It selects features according to the k highest scores of a specified scoring function. It's a way to select the 'k' best features in your dataset, where 'k' is a parameter you choose.

Null Values

In the Movie Industry Dataset, there are 2,247 null values across the 11 parameters totalling 7669. Since budget and gross are our main parameter and output, we dropped those datasets and we were left with 5422 datasets.

Feature	Score
rating_PG-13	194.402250
rating_R	337.959444
genre_Action	239.210623
genre_Animation	276.657909
genre_Comedy	116.786195
director_Anthony Russo	238.515051
director_James Cameron	127.387274
writer_Christopher Markus	199.157551
writer_James Cameron	108.337247
star_Chris Pratt	105.223686
star_Daisy Ridley	120.605232
star_Daniel Radcliffe	102.613119
star_Robert Downey Jr.	151.485500
company_Lucasfilm	110.297606
company_Marvel Studios	496.764152
company_Pixar Animation Studios	107.263914
company_Walt Disney Pictures	172.259012
year	440.975532
score	282.397728
votes	3292.085413
budget	6569.008340
runtime	446.121279

Fig 2: K Best Features

Models

1. We have used:
   a. **Linear Regression:** Linear Regression is a statistical approach for modelling the relationship between a dependent variable and one or more independent variables.
   b. **Decision Tree:** A Decision Tree is a decision support tool that uses a treelike model of decisions and their possible consequences. It is one way to display an algorithm that only contains conditional control statements.
   c. **Bagging:** Bootstrap Aggregating, often abbreviated as Bagging, is a meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.
   d. **Random Forest:** Random Forests is a learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
   e. **XGBoost:** XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.
   f. **Gradient Boosting:** Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Evaluation Metrics

1. **R^2 Score:**

   The R^2 score, also known as the coefficient of determination, is a statistical measure that shows the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model.

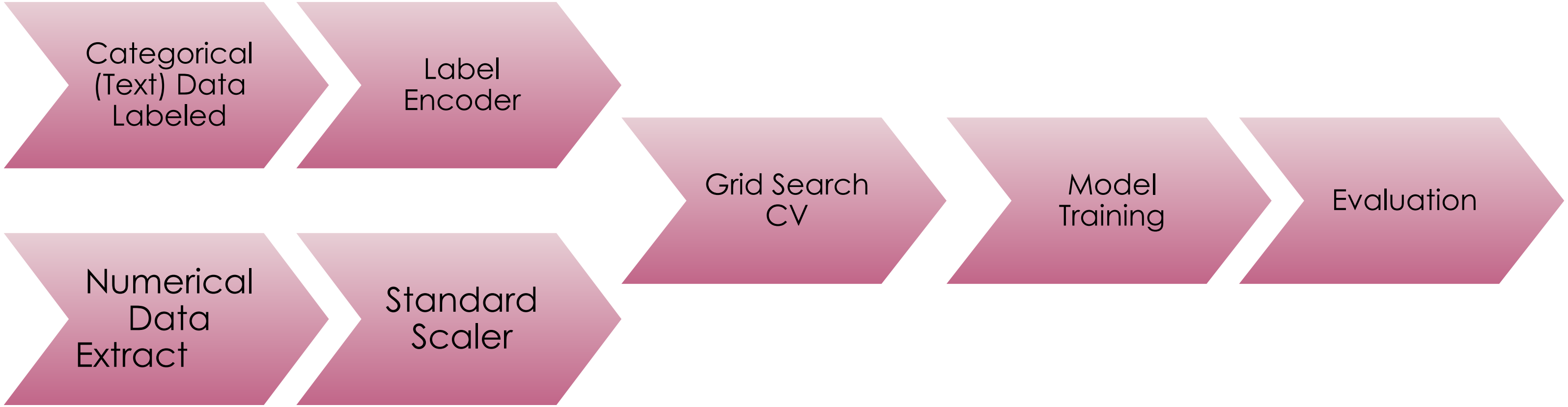
2. **Mean Absolute Percentage Error Error (MAPE):**

   The Mean Absolute Percentage Error (MAPE) is a statistical measure used to assess the accuracy of a forecasting method in predictive studies.It is the mean of all absolute percentage errors between the predicted and actual values.It provides an understanding of the prediction error in terms of the percentage of the actual values. A lower MAPE value indicates a better fit of the data.Also MAPE can be interpreted as the inverse of model accuracy, but more specifically as the average percentage difference between predictions and their intended targets in the dataset. For example, if your MAPE is 10% then your predictions are on average 10% away from the actual values they were aiming for.

Model Evaluation

Model	Training R^2	Training MAPE	Testing R^2	Testing MAPE
Linear Regression	0.6553	35.23%	0.6706	18.49%
Decision Tree	0.8664	13.00%	0.6947	4.60%
Bagging	0.8583	13.32%	0.7719	5.67%
Gradient Boosting	0.9158	10.57%	0.8242	5.69%
XGBoosting	0.9079	9.70%	0.8102	5.53%
Random Forest	0.8728	14.29%	0.7786	5.33%

Network Flow



- Preprocessing , Model Training and Validation

Visualization