Movie Box Office Prediction Using Machine Learning Models

1st Tasin Mohammad
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

2nd Nahian Tasnim
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

3rd Sanjana Mehjabin
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

4th Samiha Tahsin
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

5th Protiva Das
Computer Science and Engineering
Brac University
Dhaka, Bangladesh

***Abstract*—Machine learning algorithms have recently been utilized successfully to extract interesting patterns from large amounts of data and support decision-making in corporate settings. In this study, we attempt to forecast the pre-release box office performance of movies using the ability of such algorithms. The challenge of categorizing a film into one of many groups based on its income is simplified to the difficulty of predicting the box office take for a film. We provided a unique method for building and utilizing some parameters for movies, removing the assumption of movie independence made by machine learning methods. In this paper, a dataset of almost five thousand and five hundred movies is used. Four algorithms are used for predicting the box office scores. These are Linear regression, Decision tree, Ridge Regression, and Support Vector Classifier. This paper is to predict the potential box office success of any movie in terms of US dollar income using some key features.**

***Index Terms*—Machine Learning, Prediction, Linear Regression, Decision Tree, Ridge Regressor, SVC, Data Analysis, Movies, Box Office.**

I. Introduction

Predicting the box office performance of a movie can be challenging, as it depends on a variety of factors such as the movie's genre, budget, marketing efforts, critical reception, and competition from other films. Some of the key factors that can influence a movie's box office performance include:

- **The movie's genre:** Some genres, such as action, adventure, and superhero films, tend to perform well at the box office.
- **The movie's budget:** A high budget can indicate a larger marketing campaign and a higher level of production quality, which can contribute to a movie's box office success.
- **Marketing efforts:** A strong marketing campaign can help generate buzz and attract audiences to the movie.
- **Critical reception:** If a movie receives positive reviews from critics, it can help drive word-of-mouth recommendations and boost box office performance.
- **Competition:** A movie's box office performance may also be influenced by the other films that are released at the same time.

It is important to note that box office predictions are never certain and can be influenced by several other external factors.

There are various machine learning techniques that can be used to predict the box office performance of a movie, such as linear regression, decision trees, and neural networks. In this paper, we used four machine learning models to predict the box office scores of movies, and they are: Linear Regression, Decision Tree, Ridge Regression, and Support Vector Classification (SVC).

II. Methodology

A. Datset

For this paper, we are using a dataset titled 'IMDb 5000+ Movies & Multiple Genres Dataset' from Kaggle consisting of the following columns:

- Movie_Title
- Year
- Director
- Actors
- Rating
- main _genre
- side _genre
- Runtime(Mins)
- 'Censor'
- Total_Gross

B. Workflow

Fig. 1. Work Flow Diagram

- **Data Collection:** We have collected a secondary dataset called 'Movie Dataset' from Kaggle consisting 5558 rows and 10 columns.
- **Preprocessing:** To get the best outcome we modified our dataset in the following ways:

  - **Null Value Removal:** We dropped some rows to remove the null values from our dataset.
  - **Selecting Important Features:** We have dropped the columns of the features which do not affect the result and columns with a high correlation which may induce the result biasedly.
  - **Encoding:** As our models can't work with non-numerical values for some features with non-numerical values we have done label encoding while for other features containing non-numerical values one-hot encoding was done.
- **Splitting:** At first we split our whole dataset as features (X) and target (Y). After which we have split the dataset (Both X and Y) randomly as:
  - Training data (75%)
  - Testing data (25 %)
- **Training Data:** After splitting we have fitted our training dataset in different models (Such as 'Linear regression', 'Decision tree', 'Ridge regression', and 'Support vector classifier'). By fitting our training data in different models we trained our data to produce the best outcome.
- **Testing:** After training, we tested the output of various models using our test data and calculated different accuracy scores.

For the entire process, we used different types of libraries such as 'pandas', 'NumPy', 'sklearn', 'matplotlib' etc. We used a supervised learning method here.

Before preprocessing we had 10 columns (Movie Title, Year, Director, Actors, Rating, Runtime (Mins), Censor, main genre, side genre, Total Gross).

For preprocessing we first dropped the columns 'Movie Title' and 'side genre'. The reason behind it is that 'Movie Title' does not affect the 'Total Gross' and 'side genre' may induce the result biasedly. After that, we dropped the rows which contained any null value. Consequently, only 4393 rows and 8 columns remained. We have taken only the first actor's name from the list of names of actors as the values of the 'Actors' column and changed its name to 'Lead Actor'. Next, we removed $ and 'M' from the values of the Total Gross column to make them numerical. We have done label encoding on the 'Censor' column and labeled 'A' as 0, 'U' as 1, and 'UA' as 2.

We had then done one-hot encoding to columns 'Year', 'Director', 'Lead Actor', and 'main genre'. Consequently, the total number of rows and columns became 4393 and 3412 respectively. Then we saved all features as 'X' and our target as 'Y'. Later we split the whole data into training data (75%) and testing data (25%) randomly. We trained our training data using 4 different models

- **Linear Regression:** Linear regression is a statistical method that can be used to predict the value of a continuous variable based on the values of one or more other variables.
- **Decision Tree:** A decision tree is a machine learning algorithm that can be used to predict a target variable based on the values of one or more predictor variables. Decision trees are a flexible and easy-to-use machine learning technique that can handle both continuous and categorical variables. They are also easy to interpret, as the decision tree can be visualized as a flowchart that shows the decisions made at each step in the process.
- **Ridge Regression:** Ridge regression is a statistical method that can be used to predict the value of a continuous variable based on the values of one or more other variables. It is similar to linear regression, but it includes a regularization term that helps to prevent overfitting.
- **Support Vector Classifier:** Support vector classification (SVC) is a machine learning algorithm that can be used to predict a categorical variable based on the values of one or more predictor variables. SVC is a powerful machine learning technique that can handle both continuous and categorical variables and is effective at finding non-linear relationships between the predictor variables and the target variable.

Next, we tested our test data using those models. We used different performance measures (mean absolute error, mean squared error, root mean squared error, explained variance score, etc.) for every model to get a clear idea of how accurately these models can predict the outcome.

C. Data Visualization

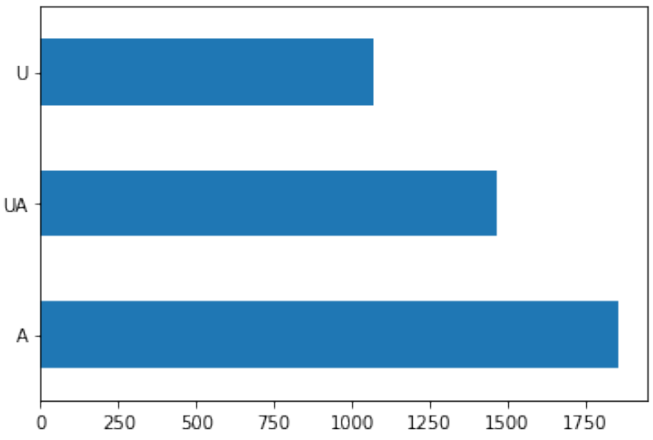

Fig. 2. Bar Chart for Censor column

This bar shows the distribution of Censor ratings throughout the dataset, with most movies having a Censor rating of A (Adult). This is followed by UA, which has the second highest frequency. And finally U, which has the smallest presence among the three categories

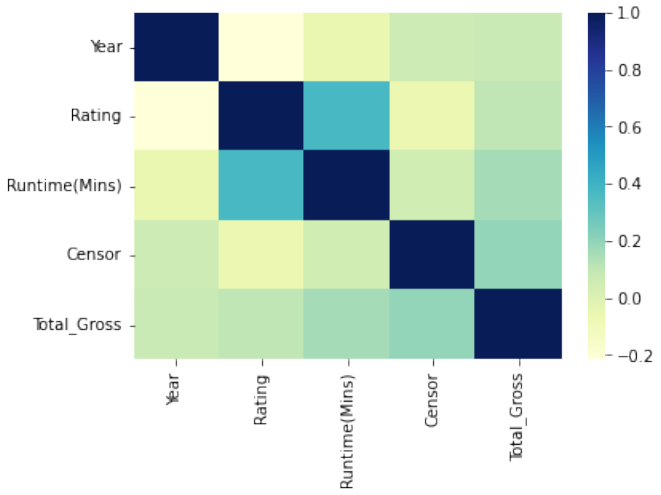

Fig. 3. Heatmap

If we visualize the correlation of features with numerical values using a Heatmap we can see 'Year' is least correlated with 'Total Gross'. Besides, 'Rating' and 'Runtime' are highly correlated with each other. This means if 'Runtime' increases the 'Rating' will increase, if the 'Runtime' decreases the 'Rating' will also decrease, and vice-versa.

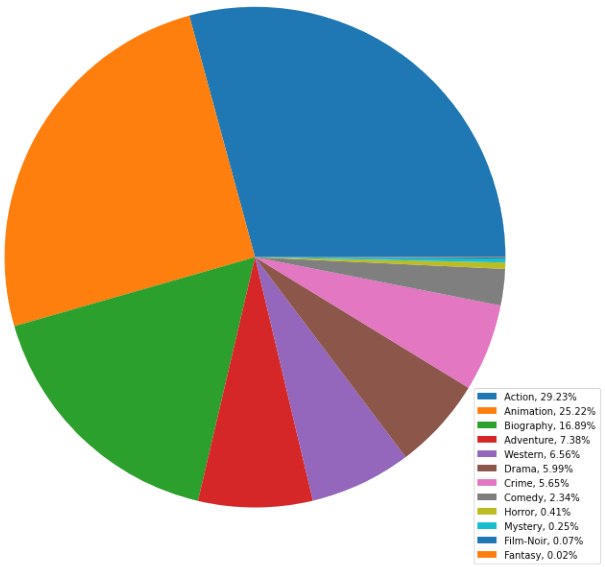

Fig. 4. Pie Chart for main genre column

The Pie Chart demonstrates the distribution of different film genres in the dataset. Action movies consist of 29.23% of all movies, while Animation is close behind with 25.22%. The smallest presence in the genre column is Fantasy, accounting for merely 0.02% of movies.

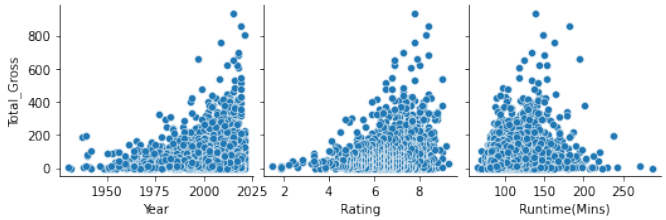

Fig. 5. Scatter Plot

The above graph plots the Total Gross with Year, Rating, and Runtime. This shows a few interesting trends. Firstly, it can be seen that the Total Gross for movies has steadily increased over the years, with movies that were released in the last two decades having the largest box office numbers. Secondly, there appears to be a positive correlation between movies and ratings as per the graph. This is because movies with higher ratings earned the most at the box office. Finally, movies with a runtime of 100-150 minutes seem to perform better at the box office, with most of the higher box office values belonging in that range.

III. Result Analysis

To analyze the accuracy of our models we used five different performance metrics.

- **Coefficient of Determination (r^2):** This demonstrates how well the model has predicted the unknown samples. The score ranges from 0.0 to 1.0 but can also be negative.
- **Mean Absolute Error:** This demonstrates the dataset's average absolute error across all data points.
- **Mean Squared Error:** This demonstrates the dataset's average squared error across all data points.
- **Root Mean Square Error:** This is the square root of the Mean Squared Error and demonstrates the average error across all data points in the same units as the target variable.
- **Explained Variance Score:** This demonstrates how well our model can account for the variations in the dataset. The score ranges from 0.0 to 1.0, with 1.0 representing a flawless model.

The results obtained from each of the four models varied immensely in accuracy and precision.

A. Linear Regression

Linear Regression was the first model we chose to implement due to its simplicity and relatively short training time.

TABLE I
Results from Linear Regression

Metric	Value
Coefficient of Determination	0.0596
Mean Absolute Error	57.75
Mean Squared Error	6588.42
Root Mean Square Error	81.17
Explained Variance Score	0.0612

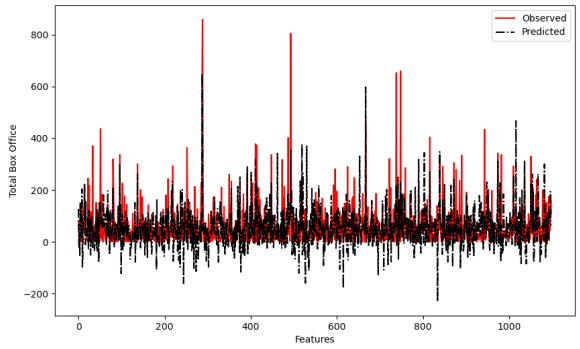

Fig. 6. Linear Regression: Observed vs Predicted

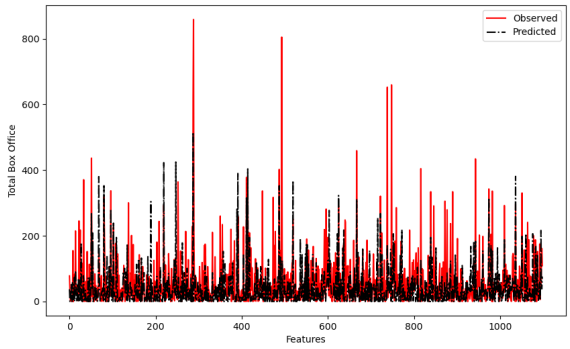

Fig. 7. Decision Tree: Observed vs Predicted

The Linear Regression model did not perform very well when it came to predicting the box office scores for the test samples. This is demonstrated by the low coefficient of determination of just 0.0596 which indicates that the model was able to accurately predict the box office scores of only 5.96% of movies in the test sample. The mean absolute error is also quite large and illustrates an average deviation of 57.7 million dollars between the observed and predicted scores. The low Explained Variance Score of 0.061 proves that the model was not able to account for the variations between the train and test samples very well.

The graph above shows the overlap between the observed box office scores and the box office scores predicted by the Linear Regression model in relation to the features. This further emphasizes the poor performance of the Linear Regression as seen by the low overlap in values between the observed and predicted scores.

B. Decision Tree Regressor

Our next model was a Decision Tree Regressor which is very often an effective method for predictive modeling.

TABLE II
RESULTS FROM DECISION TREE

Metric	Value
Coefficient of Determination	0.0997
Mean Absolute Error	44.44
Mean Squared Error	6307.27
Root Mean Square Error	79.42
Explained Variance Score	0.145

The Decision Tree model did not do particularly well either as we observe that only 9.97% of the films in the test sample had box office scores that could be reliably predicted by the model, which is demonstrated by the poor coefficient of determination of 0.0997. The average discrepancy between the observed and predicted scores, as shown by the mean absolute error, is 44.4 million dollars, which is quite large. Moreover, the model's inability to adequately account for the variations between the train and test data is demonstrated by the low Explained Variance Score of 0.145.

The Decision Tree model was slightly more accurate than Linear Regression in terms of predicting the box office scores.

C. Ridge Regression

TABLE III
RESULTS FROM RIDGE REGRESSION

Metric	Value
Coefficient of Determination	0.399
Mean Absolute Error	43.178
Mean Squared Error	4205.99
Root Mean Square Error	64.85
Explained Variance Score	0.402

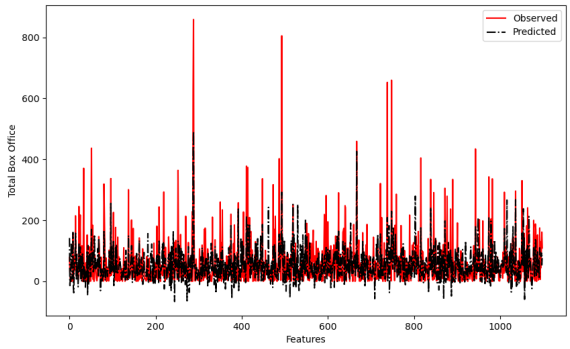

Fig. 8. Ridge Regression: Observed vs Predicted

The Ridge Regression model proved to be the most accurate in predicting the box office scores among all the four models we had implemented. With a Coefficient of Determination of 0.399, the model was able to accurately predict the box office score of 39.9% of films in the test sample. The Mean Absolute Error for the model was the lowest among the models with a deviation of 43.178 million dollars from the observed box office scores. Furthermore, it had the highest Explained Variance Score, at 0.402, which indicates that it was able to

account for the variations between the train and test data better than any other model.

D. Support Vector Classifier (SVC)

TABLE IV
RESULTS FROM SVC

Metric	Value
Coefficient of Determination	0.00637
Mean Absolute Error	53.913
Mean Squared Error	9244.97
Root Mean Square Error	96.15
Explained Variance Score	−0.132

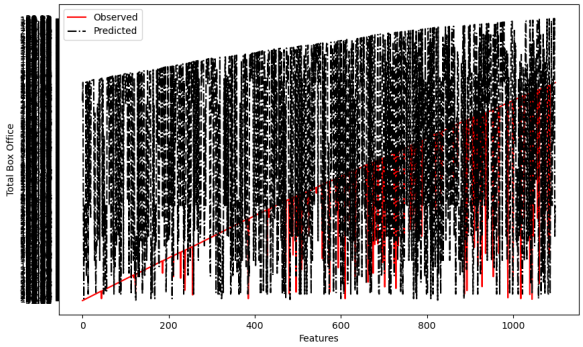

Fig. 9. SVC: Observed vs Predicted

The SVC model was by far the least accurate model we used. The Coefficient of Determination was very close to 0, at 0.00637. This means that the model was only able to accurately predict the box office scores of 0.637% of movies in the test sample, which translates to just 35 out of the 5500 movies in the dataset. The Mean Absolute Error however was only the second highest among all the other models, with Linear Regression being the highest at 57.7 million dollars. This means that although most of the predictions were inaccurate, the predicted scores were closer to the actual scores when compared to the Linear Regression model. The Explained Variance Score was the lowest among the four models, which shows that the model was almost fully unable to take into account the variations between the train and test data.

The graph above also shows the large discrepancy between the observed and predicted box of scores, with very little overlap between the two

E. Combined Results

TABLE V
RESULTS FROM ALL MODELS

Metric	Linear Regression	Decision Tree
Coefficient of Determination	0.0596	0.0997
Mean Absolute Error	57.75	44.44
Mean Squared Error	6588.42	6307.27
Root Mean Square Error	81.17	79.42
Explained Variance Score	0.0612	0.145
Metric	**Ridge Regression**	**SVC**
Coefficient of Determination	0.399	0.00637
Mean Absolute Error	43.178	53.913
Mean Squared Error	4205.99	9244.97
Root Mean Square Error	64.85	96.15
Explained Variance Score	0.402	−0.132

IV. CONCLUSION

The film industry generates billions of dollars in revenue each year. The question of what makes a film successful has been asked many times over the years by large corporations such as Twenty-First Century Fox and Universal Studios, which award million dollar prizes to those who can improve their recommendation and prediction algorithms. We can see from the results of this project that accurately predicting the gross box office for any film is a challenging task. The results are still insufficient for commercial use at this time since a model for industrial use must maintain a certain level of accuracy which is far more precise than the results found here. However, with more advanced machine learning models and a larger dataset with more defined features, the level of error can be significantly reduced.