# Final Report 562

Li Hui Cham Nicholas Boyer Christine Hu Raheq Hassan

April 2024

GitHub link for this project.

## 1 Introduction

The global film industry represents a significant economic sector, generating billions of dollars in revenue annually. Understanding and predicting box office performance is crucial for stakeholders, including producers, distributors, marketers, and financial analysts. These predictions can help in making informed decisions regarding budget allocations, marketing strategies, and release timings, which are critical to a film's financial success.

In this project, we explore the intriguing challenge of predicting box office revenues using detailed movie data, including opening earnings, total gross earnings, release dates, and distributor information. Our dataset, derived from "Top_Highest_Openings.csv," includes 1000 movies, fully prepared and ready for analysis with no missing values. Our exploratory analysis offered initial insights into the factors influencing box office success and prepared the groundwork for deeper statistical examination. We approached the prediction task using a variety of machine learning models, starting with basic linear regression to establish a baseline and progressively incorporating more techniques like polynomial and ridge regression. We also ventured into advanced ensemble methods such as random forests, gradient boosting machines, and XGBoost, alongside exploring neural networks to capture complex patterns. By training and validating these models, we aim to not only predict film revenues effectively but also understand which factors most significantly influence box office outcomes. This project sheds light on the analytical potential within the film industry and showcases how machine learning can be applied to real-world economic forecasting.

## 2 Related Work

In "Top Grossing Movies EDA, Clustering, and Modeling," Angela Cao looks into which distributors often top the charts and explores the performance of different movie genres over time. Angela employs a mix of data visualization, clustering, and regression analysis to uncover patterns and predict outcomes. The project is rich with intuitive graphs and leverages popular data science tools like Pandas, Plotly, and Scikit-Learn to bring the data to life and make predictions more accessible [3].

In "Movie Analysis," Divyansh Aggarwal takes us on an insightful journey through the data-driven world of cinema. He skillfully manipulates movie data to uncover trends in ratings, popularity, and financial success. His approach includes cleaning and organizing the data, followed by exploratory analysis where he visually breaks down the elements that make movies successful or not. It is a practical example of how data science can be applied to better understand the dynamics of the movie industry [1].

## 3 Method

### 3.1 Dataset

We use the dataset "Top Grossing Movies Dataset" from Kaggle [2] for this project. The dataset contains information on the top-grossing movies at the box office, with their release dates, opening weekend earnings, total gross earnings, percentage of the total gross, the number of theaters they were shown in, average earnings per theater, and the distributor. The dataset has 1000 rows (observations) and 8 columns (features).

### 3.2 Data Cleaning & Feature Engineering

We want to predict the box office revenue based on various features in the dataset. First, we clean the dataset to deal with N/A values (if any) and perform feature engineering. For example, we combine the distributors with less than 100 observations into one distributor called Others. Then, we visualise the data for exploratory data analysis to better understand the relationships between the features and their distributions. Furthermore, we convert the Date variable to the individual features Day, Month and Year.

### 3.3 Machine Learning Models

For predictive modeling, various machine learning methods were employed to model the data. These include linear regression, ridge regression, polynomial regression, neural networks, and gradient boosting descent. We split the training and testing data 80% to 20%. We used the following inputs:

- Opening Revenue
- Number of Theaters
- Average Opening per Theater
- Date of Opening: Day, Month, Year

Our goal is to predict the total revenue based on this data. We report the results and analysis of these models in the following section.

## 4 Results

### 4.1 Exploratory Data Analysis

Upon analysis, we observe that Walt Disney Studios Motion Pictures has the highest opening weekend earnings and highest total gross earnings, both followed by Warner Bros. This is due to Warner Bros being the highest individual movie distributor, distributing 171 movies in total and Walt Disney being the third highest at 147 movies in total. We use a histogram to visualise the distribution of numerical data. The total gross earnings (mean = USD 155.5 million), opening earnings (mean = USD 47.4 million) and average earnings per theater during opening weekend (mean = 3428.65 theaters) present a right-skewed distribution, peaking slightly to the left of the center. However, the distribution shape of percentage total gross earnings and total number of theaters are bell-shaped, indicating a normal distribution for both variables.

Next, we look into the time series data. June is the month with the highest number of movies released, followed by July and November. To find out the pattern of gross earnings across 30 years, we use line charts for visualisation. As the number of movies released every year increases, the total gross earnings of the movies increases with it.

We use scatter plots to identify the relationships between variables and determine whether there is a correlation between two variables. Both scatter plots of Opening vs Total Gross and Average vs Total Gross show positive linear relationships in upward trends. This indicates positive correlation between the variables. This is further justified by a correlation heatmap where we observe a strong positive correlation between Total Gross and Opening ($r = 0.88$), as well as Total Gross and Average ($r = 0.86$).

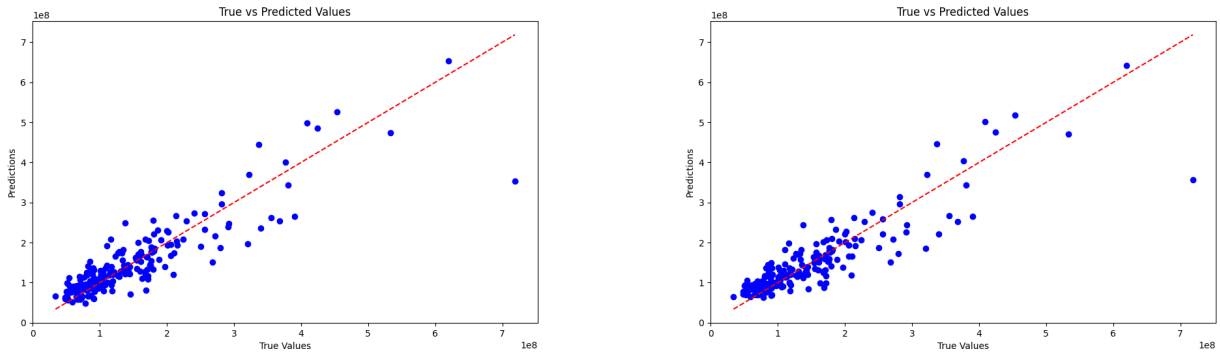
### 4.2 Machine Learning Models

#### 4.2.1 Linear Regression

The first model we used as a baseline model is linear regression. The true vs predicted values are shown in figure 1a. The linear regression had a mean squared error (MSE) of $2.21316 \times 10^{15}$. The most significant outlier is Top Gun: Maverick, with a true total revenue of \$718732821 and predicted revenue of \$352997044. The linear coefficients and intercept of this model are shown in table 1.

Table 1: Linear regression coefficients and intercept.

| Input | Coefficient |
|---|---|
| Total Opening | 1.8764813694969302 |
| Theaters | 18840.659980333803 |
| Average Opening | 3826.7271874296566 |
| Day | 389269.56417618436 |
| Month | 3465958.6599470438 |
| Year | -1863223.7295686738 |
| Intercept | 3666635681.512526 |




(a) Linear model of the total revenue prediction.

(b) Ridge regression model of the total revenue prediction.

Figure 1: Linear based models.

#### 4.2.2 Ridge Regression

The next model we used was ridge regression for regularization. The true vs predicted values are shown in figure 1b. The ridge regression had a mean squared error (MSE) of $2.19631 \times 10^{15}$. The regularization parameter of $\alpha = 10000$ provided improved results over the linear model and reduced over fitting.

#### 4.2.3 Advanced models

The advanced models we investigated include polynomial fits, random forests, gradient boosting regression, XGBoost, and neural network. The MSEs for these models are listed in table 2. The best performing model, neural network, is shown in figure 2. The details of the neural networks are:

- 2 hidden layers with "elu" activation function
- 200 nodes per layer
- Loss is MSE
- Optimizer is Adam
- Epochs=60 (early stopping)
- Data is normalized before input

This set up of the neural network was chosen based on known techniques to improve performance. The "elu" activation function allows for derivatives when $x$ is less than zero, in contrast to "relu". A large number of nodes and 2 hidden layers were chosen to improve performance. Using the standard MSE loss and Adam optimizer, the training was short to prevent over fitting. Another key aspect was pre-training normalization that removed size differences between inputs. These design features allowed the neural network to achieve the highest performance.

The next best model was the 3rd degree polynomial, which is the cubic of all the input data. Further increases in polynomial order appeared to over fit to the data and increase the MSE.

Table 2: MSEs for advanced models.

| Model | MSE |
|---|---|
| 2nd degree Polynomial | $2.28680 \times 10^{15}$ |
| 3rd degree Polynomial | $2.13119 \times 10^{15}$ |
| 4th degree Polynomial | $2.20779 \times 10^{15}$ |
| Neural Network | $2.11993 \times 10^{15}$ |
| Gradient Boosted Regression | $2.40597 \times 10^{15}$ |
| Random Forests | $2.17689 \times 10^{15}$ |
| XGBoost | $2.73705 \times 10^{15}$ |

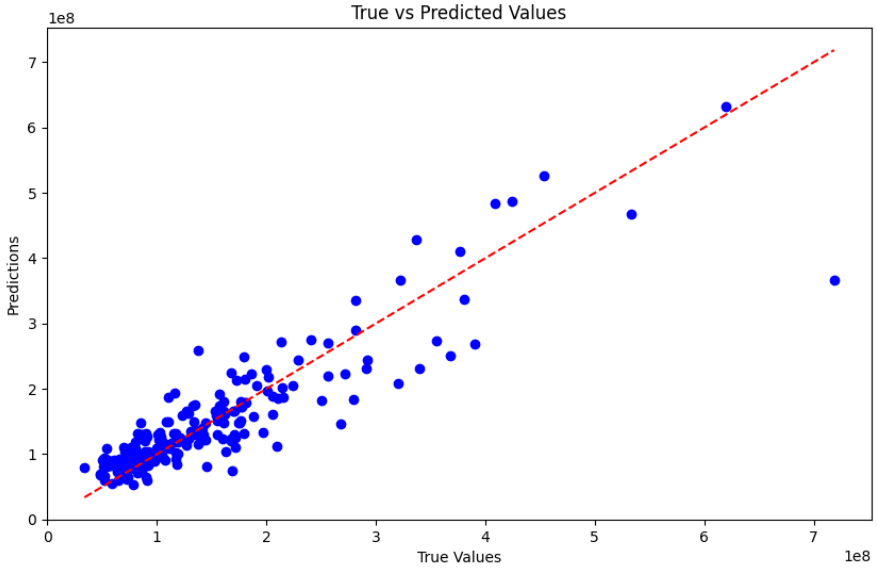



Figure 2: Neural Network model of total revenue prediction.

# 5 Conclusion

In this report, we analyze various initial properties of movies in relation to their total gross revenue at the box office. We identify a strong positive correlation with opening revenue, the average opening per theater, and the year of release; as well as a weak positive correlation with the number of theaters. Furthermore, each month has significant variance in the number of movies released, possibly suggesting differences in both demand and competition. We then use these inputs to compare machine learning models for predicting gross revenue. After establishing the baseline of basic regression models, we test multiple advanced machine learning models and determine that the best performing model is a neural network.

Our results are not indicative of box office earnings overall. The "Top Grossing Movies Dataset" does not account for movies that fail at the box office; the lowest gross revenue in the dataset is $33,889,684. In addition, the MSE is quite high, although it may be influenced by the notable outlier of Top Gun: Maverick. Nevertheless, given these caveats, our report shows that total gross film revenue can be predicted with simple quantitative data.

# References


[1] Apr. 2024. URL: https://www.kaggle.com/code/divyanshagg27/movie-analysis.

[2] Apr. 2024. URL: https://www.kaggle.com/datasets/akankshaaa013/top-grossing-movies-dataset.

[3] Angela Cao. *Top grossing movies EDA, clustering, & prediction.* Apr. 2024. URL: https://www.kaggle.com/code/angc1998/top-grossing-movies-eda-clustering-prediction/notebook.