Predicting US Movies Gross Using Linear Regression

Leena AlQasem

Leenabdulh@gmail.com

Randa Mohammed

randa1414@gmail.com

Abstract:

Prediction of the correct box office gross is essential for the movie industry and decreasing market risk. The success and failure of the movie depend on movie-related variables. Therefore, the goal of this project is to create a machine learning model that will predict the US movies gross taking into consideration many features. The features include IMDb rating, certification of movie, number of votes, release year, movie duration, and metascore. Exploring data scraped from the IMDb websites to see which features affect the prediction the most.

Design:

The progress of this project will be as follows. First gathering data using web scraping, then exploratory Data Analysis (EDA) will be performed on the data we gathered. Finally, preparing data to be used in different regression models. In this project, we analyze the relationship between the target variable “Gross” and the features “Movie name, release year, duration, rating, votes, metascore, certificates, and genres”.

Data:

The dataset used in this project was extracted from IMDb website. Web scraping with beautiful soup was used in order to extract data about movies from the IMDb website. There were 3000 movies with 9 features initially. The dataset becomes 2902 records with 20 features after we have done some cleaning and feature engineering on the original dataset.

Algorithms:

- Feature Engineering:

    1. Convert categorical features to dummy variables (Certificates column).
    2. Subtracting interactive terms in order to make the feature more relevant (Year column).

- Models:

We have explored many regression machine learning models in order to choose the best for our case. The models are linear regression, K-fold linear regression, Polynomial regression, Ridge regression, and lasso regression. The data was splitted into 60 percent training, 20 percent validating, and 20 percent testing. The model we selected was polynomial regression with degree 2 because it shows the best score between all models.

Best Model: **Polynomial regression with degree 2:**

Polynomial Training Score: 0.61368493

Polynomial Validating score: 0.59439209

Polynomial Testing score: 0.67247904

Tools:

- **Technologies:** Jupyter Notebook, Python.
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Request, BeautifulSoup, and Sklearn.

Communication:

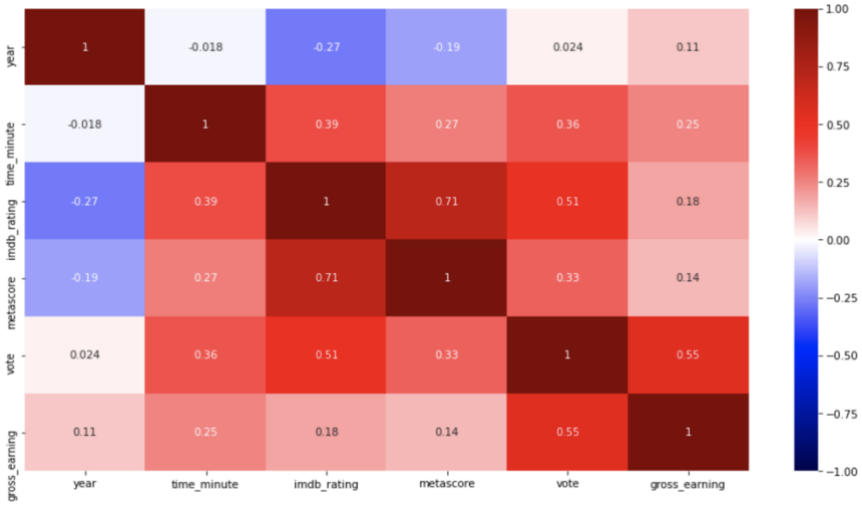

Figure 1 Heatmap correlations

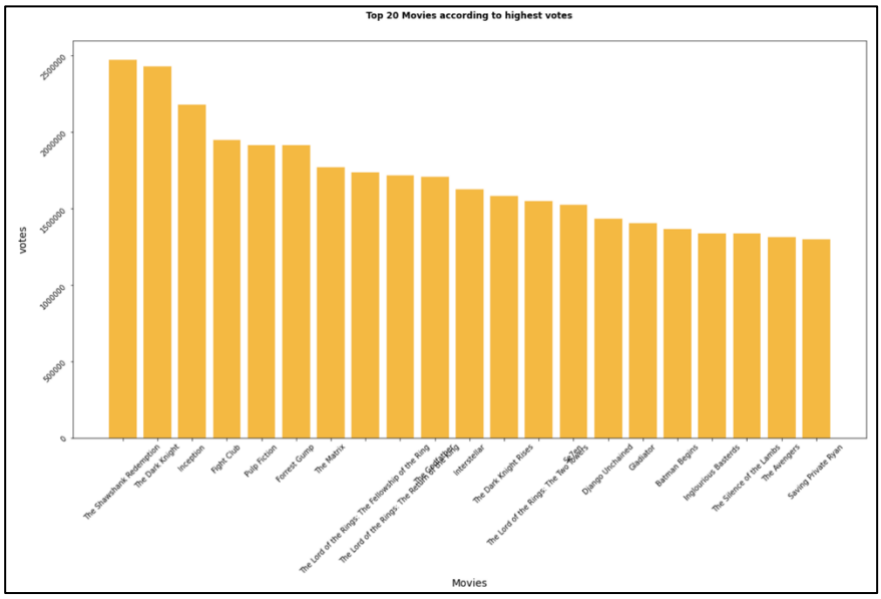

Figure 2 Top 20 Movies Based on Votes

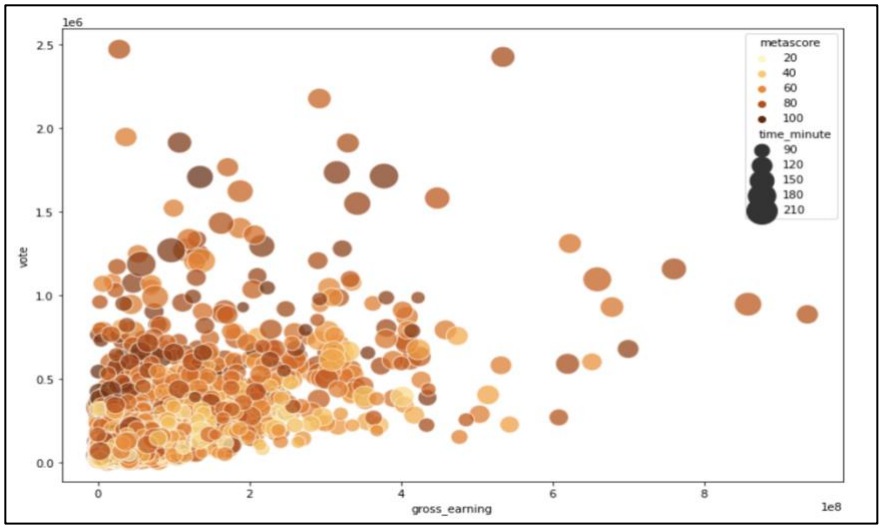

Figure 3 Gross Based on Metascore, Rating, and Duration

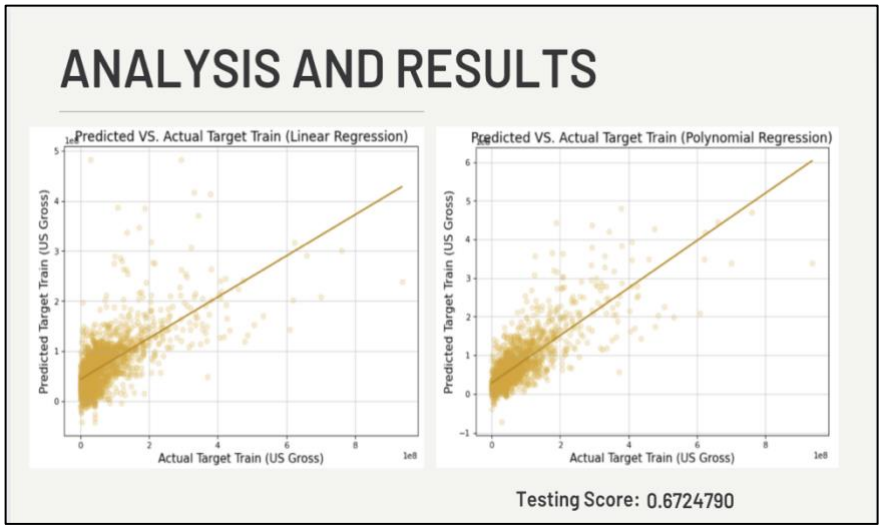

Figure 4 Predict VS. Actual Target for Linear and polynomial Regression