Predicting Movies' Gross Earnings using Linear Regression

Project Report by: Himani Kaushik

Abstract

Every industry is experiencing financial crunch because of the pandemic, including the movie business. Hence, to decrease the financial risk, prediction of the gross earnings of a movie is essential to the movie industry. Therefore, the goal of the project is to create a regression model that will predict US movies' gross after considering several features. The features include IMDb rating, certification, duration of the movie, number of votes and metascore. Anyone interested in financing or investing in the movie, including individual producers, film studios, private investors, etc. will benefit from this model.

Design

The project is designed to webscrape the information of the movies from IMDb website and then utilize the data to build and select the most accurate predictive model for gross earnings of a movie based on its features. Several regression models like linear regression, ridge regression, lasso regression, etc. will be utilized to evaluate the training and validation scores of the dataset to choose the model with the best metric values and run the test data.

Data

The data is scraped from IMDb.com website using Python libraries and packages including Request and BeautifulSoup. Initially, the data had 3000 movies with 9 features. After the data cleaning, an interaction feature was added, and the irrelevant columns were eliminated during feature engineering. This resulted in 2951 movies with 20 features.

Algorithms

- Feature Engineering:
    - Categorical feature (certificates) was converted to dummy variables.
    - Interactive terms were subtracted to make the feature (year) more relevant.

- Models:
    - Several regression models were explored to select the best predictive model.
    - Data was split into 60% training, 20% validating and 20% testing.
    - Following are the results of various models:

Regression Models	Training Score	Validation Score
Normal Linear Regression	0.389285939	0.461151975
K-fold Linear Regression	0.389285939	0.379048417
Polynomial Regression Degree 2	0.00471088	-0.07468319
Polynomial Regression Interaction	0.50838276	0.3800856
Ridge Regression	0.46113795	0.38928325
Ridge Regression Cross-Validation	0.389283247	0.379102316
Lasso Regression Cross-Validation	0.389285939	0.379048764
Ridge Regression on polynomial interaction only	**0.61451453**	**0.54744346**

- The model that was selected was Ridge regression on polynomial interaction only features.

```
Ridge Regression on poly interaction only validation Score:
0.54744346
Ridge Regression on poly interaction only Training Score: 0
.61451453
Ridge Regression on poly interaction only Testing Score: 0.
59082610
```

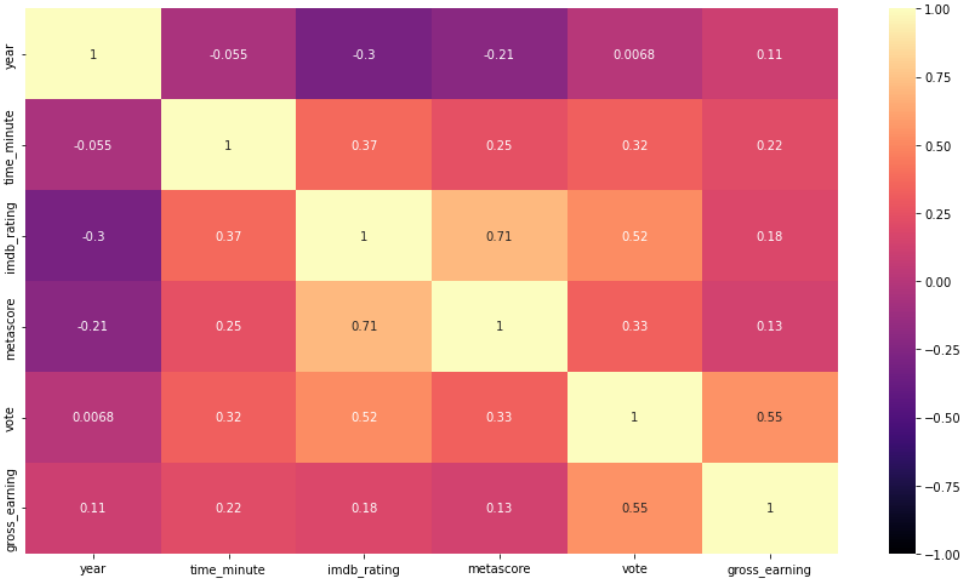
Tools

- Python libraries including Request, BeautifulSoup, Pandas and Numpy for web scraping, parsing and data manipulation.
- Scikit-learn for modeling and evaluation metrics.
- Matplotlib and Seaborn for data visualizations.

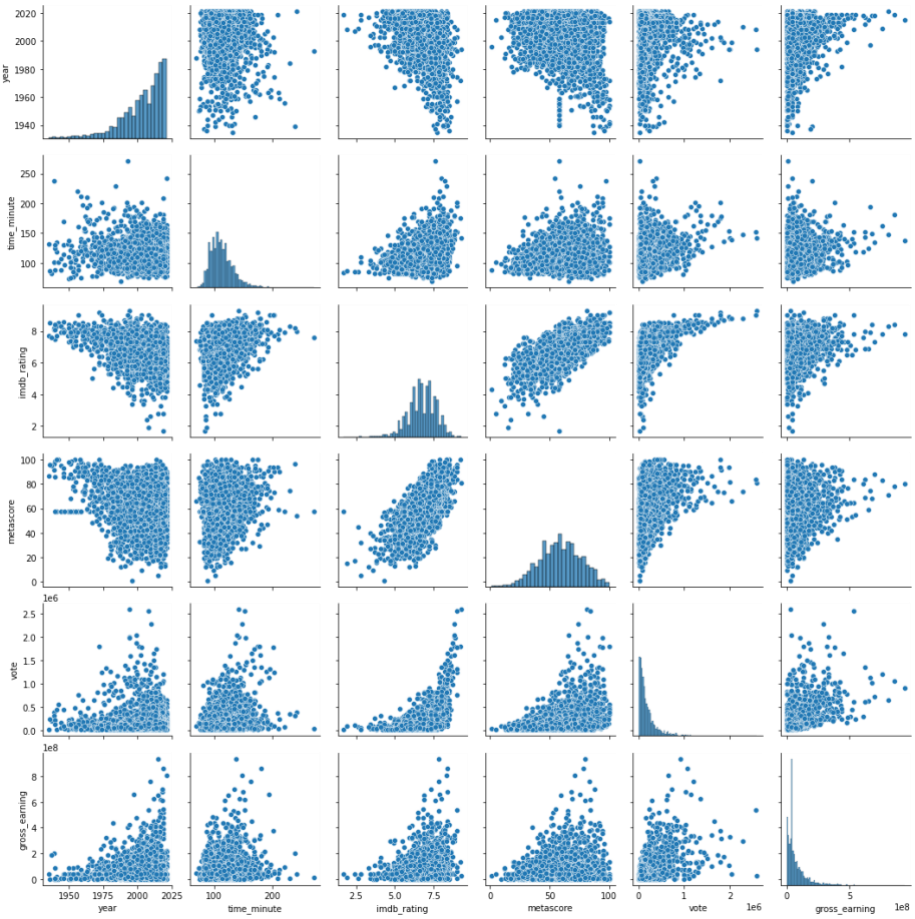
Communication

As stated above, Ridge Regression on polynomial interaction only features is the best model for prediction of gross earnings. Following charts were utilized to analyze and communicate the results:

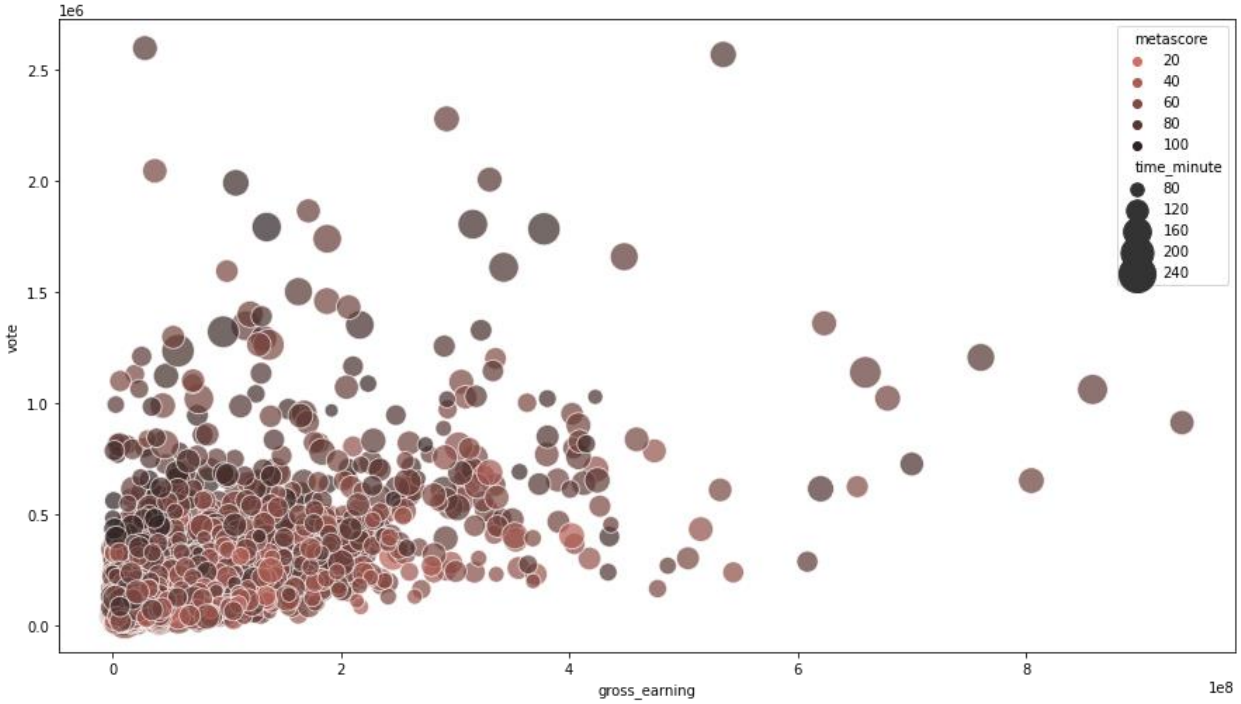
Heatmap Correlations



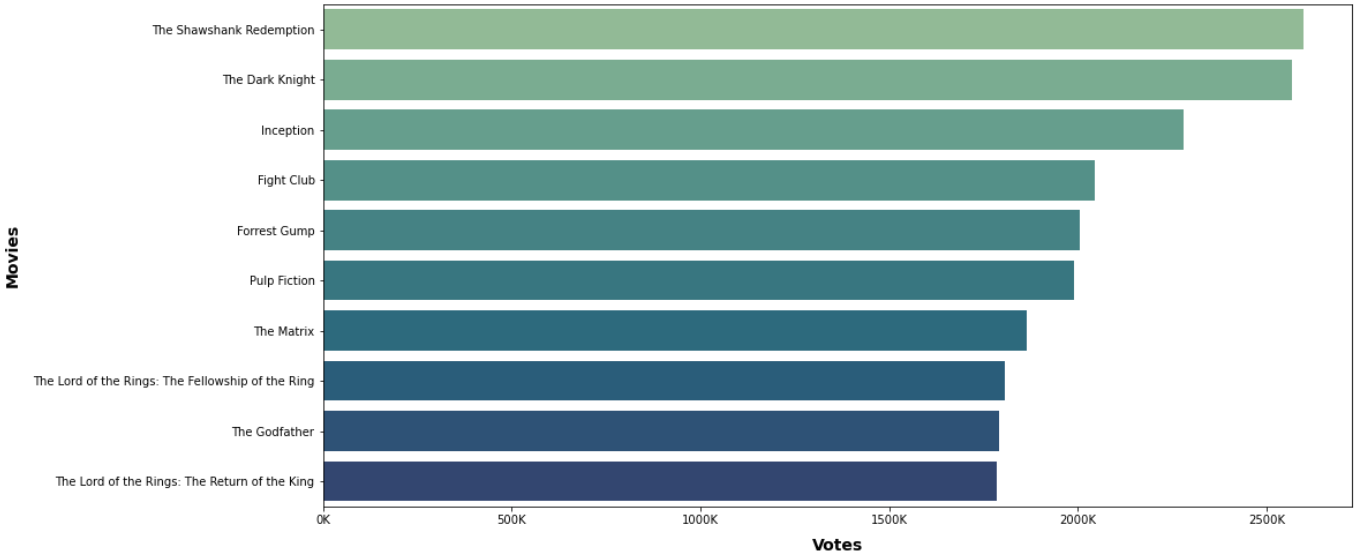
Pairplot of features and target variable

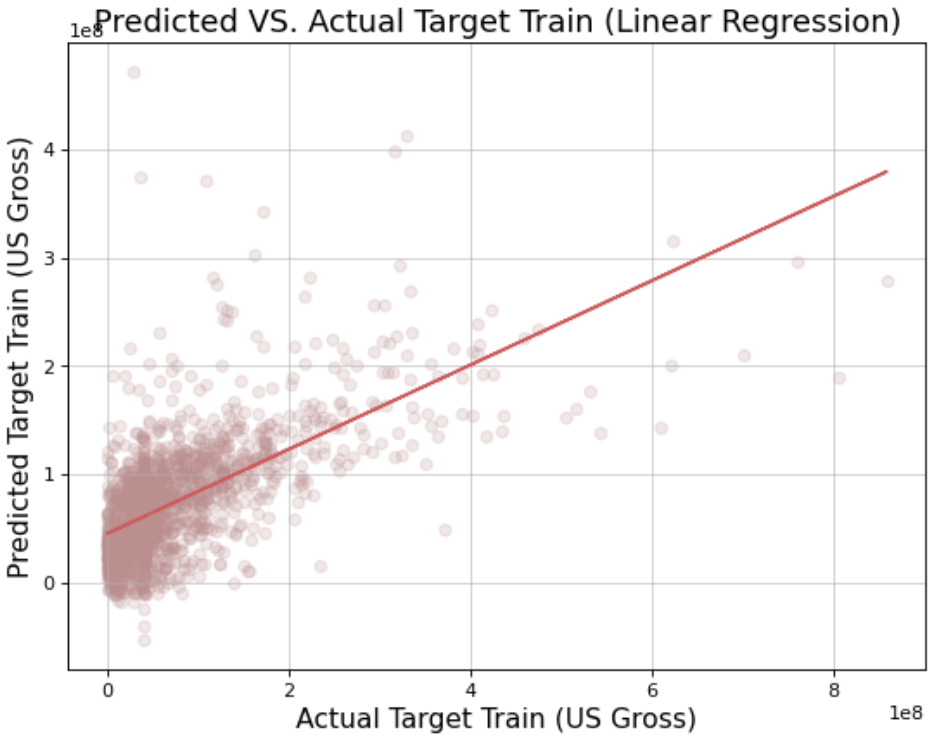


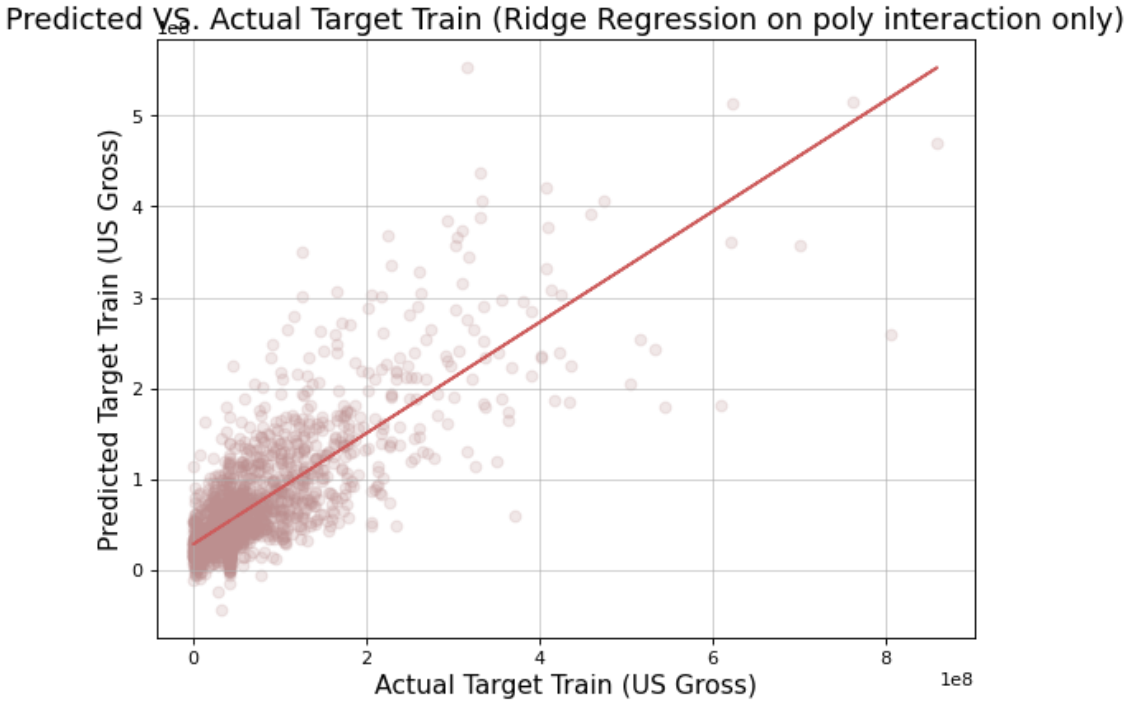
Gross based on metascore, rating and duration



Top 10 Movies by Votes on IMDB




Predicted VS. Actual Target Train (Linear Regression)
1e8
Predicted Target Train (US Gross)
Actual Target Train (US Gross)
1e8
0
1
2
3
4
0
2
4
6
8


Predicted VS. Actual Target Train (Ridge Regression on poly interaction only)
1e8
Predicted Target Train (US Gross)
Actual Target Train (US Gross)
1e8
0
1
2
3
4
5
0
2
4
6
8