Project-2-Linear-Regression with movies based on novels.

Description of project goals

The project required web scrapping and analysis of resulting data using regression, the form of regression was based on the data collected and the feature engineering necessary to analyze the data in regression. This project used simple linear regression with K-fold cross validation to predict the Target worldwide gross using the Features listed below.

Features and Target Variables:

Target:

Worldwide gross.

Features:

Title, Rating, Runtime, IMDB scores, Gernes, The theaters, Total gross, Budget, worldwide gross, and Domestic gross.

Data Used

- IMDB website
- Box Office Mojo.
- The Numbers.

Tools Used

- Numpy

- Pandas
- Seaborn
- Beautiful Soup
- Sklearn

Impacts in the scope of the project:

To create a prediction model for Worldwide gross.

To determine the salient features for predicting this target.

Workflow

Created databases using IMDB, the Numbers webpage and Box Office Mojo, which are noted in their respective jupyter notebooks.
Make a heatmap for all features showing the correlation relation.
Clean data by removing duplicate and non-values. Scale the features and use dummy variables.
Split data train, test, and validation using CV.
Features engineering where we create new variable (Primetime) indicating if a movie was released during a typical successful time of year summer or holiday season. We remove total gross feature to avoid data lake. Using Lasso regression to eliminate uninfluential features.

Work result:

X= Rating, Runtime, IMDB_score, Budget, Theaters, Primetime.

Y= Worldwide Gross.

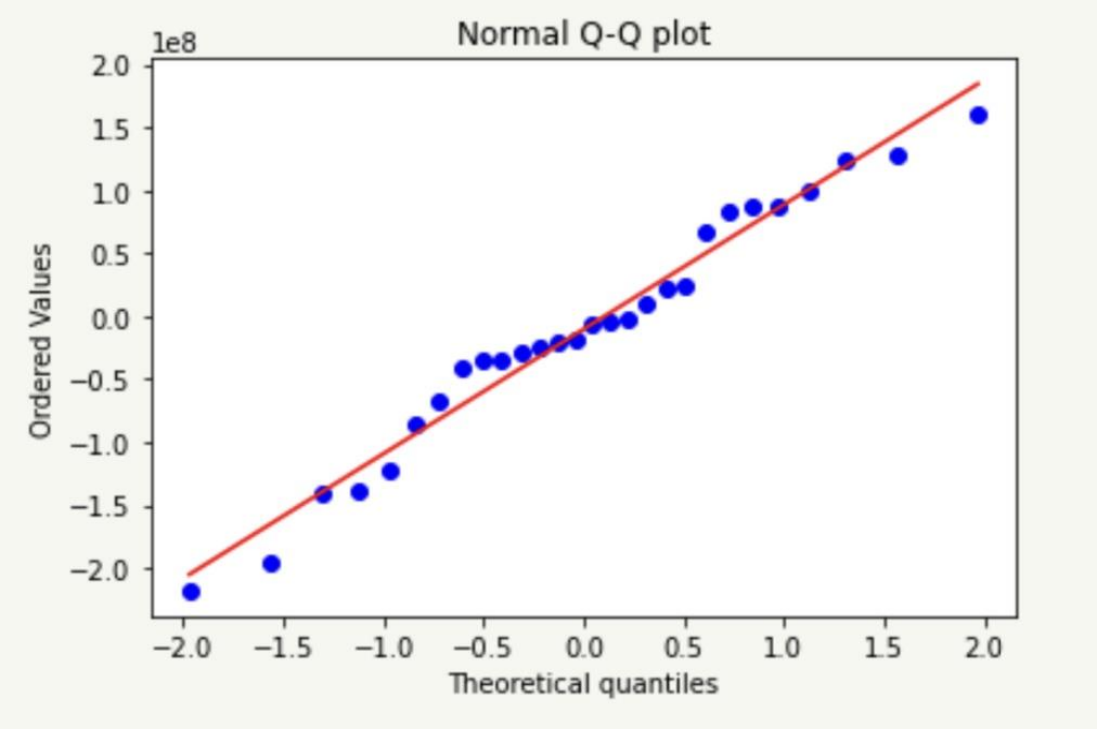
the metrics	Score	R^2
Linear Regression	0.54	0.31
Lasso	0.54	0.31

X= Budget, Primetime, Theaters, Domestic Gross.

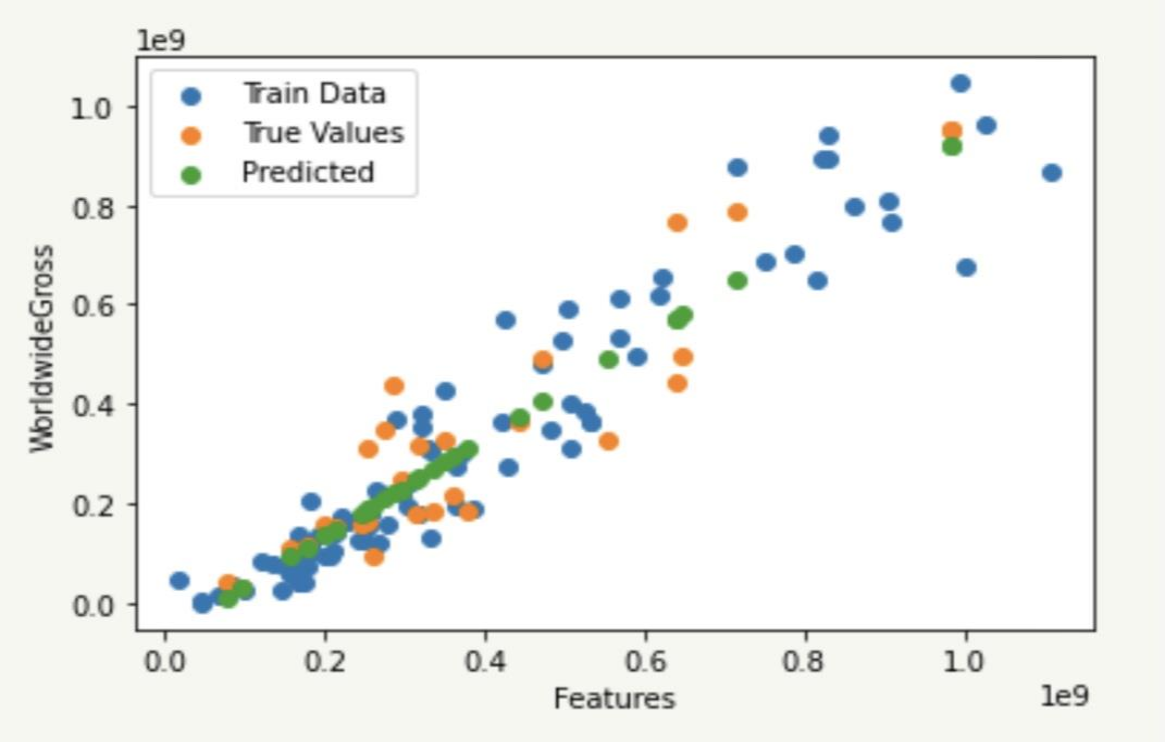
Y= Worldwide Gross.

the metrics	Score	R^2
Linear Regression	0.92	0.85
Lasso	0.92	0.85

Error based on the features:



Linear regression model:



Future steps:

Add more features.

Define different movies category.

TEAMWORK:

Rana Alqahtani.

Khulud Alshamrani.

Ohoud Albabtain.