Preprocessing:

1-Applying some NLP techniques to title, overview, and tagline columns as removing punctuation, making all words lowercase, tokenization, removing stops words, stemming, lemmatization, and removing numeric

2-Then extracting information in dictionary columns by converting them to lists in columns of keywords, genres, spoken languages, production companies, and production countries using split method

3-We've used MultiLabelBinarizer to transform these columns into binary-encoded representations.

We have used a global MultiLabelBinarizer to use in two functions, 'dictionaryPreprocessing' one for the training data that contains fit and transform, and 'dictionaryPreprocessing_test' for the test data that contains only transform from splitting using the get dummies method.

4-Applying label encoding on other columns as status, original language using label encoder functions

5-Converting the column of the Homepage into numeric values using astype function

6-Scaling the columns to be in the same range by subtraction the minimum of the data divided by the difference between maximum and minimum for columns budget, revenue, viewer count, release date, runtime, vote count, original language

Feature selection:

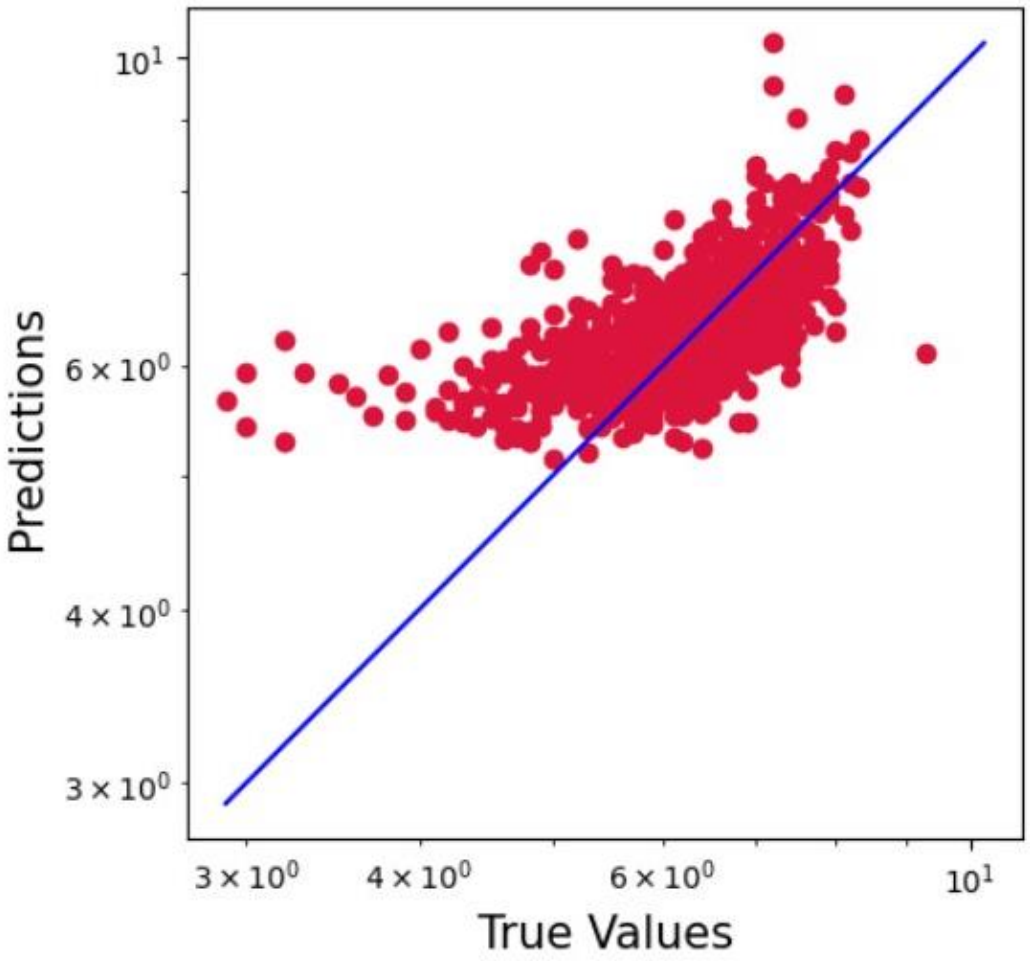
1- In the feature selection we've used 'VarianceThreshold' feature selection technique to remove the columns with low variance from the preprocessed data. By comparing the model performance metrics such as accuracy or other relevant metrics, the effect of this feature selection is way better.

Regression Techniques used:

1-Linear Regression
2-Random Forest Regression
3-Ridge Regression

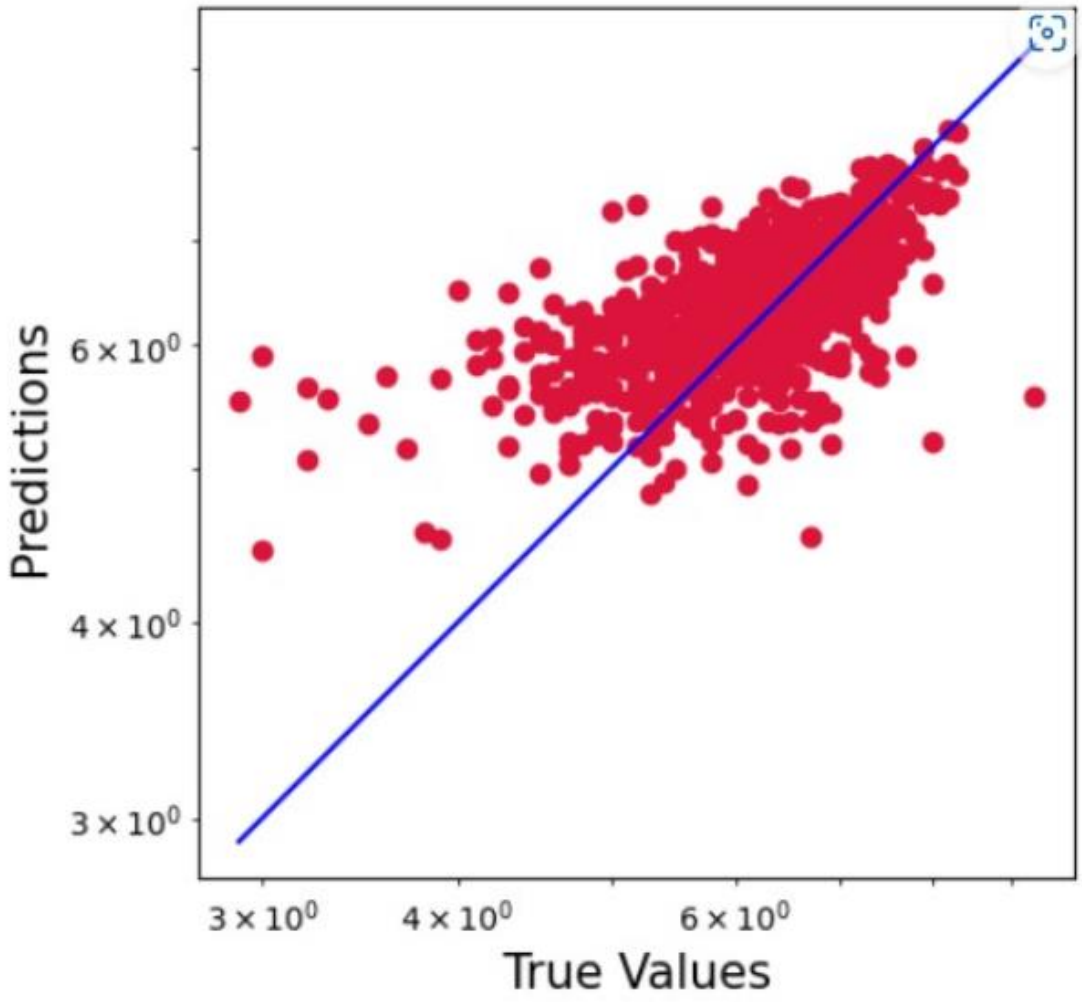
Linear Regression:

Mean Square Error linear regression 0.4969167696463531
Score of linear regression: 0.3519218019363397
Score2 of linear regression: 0.3519218019363397



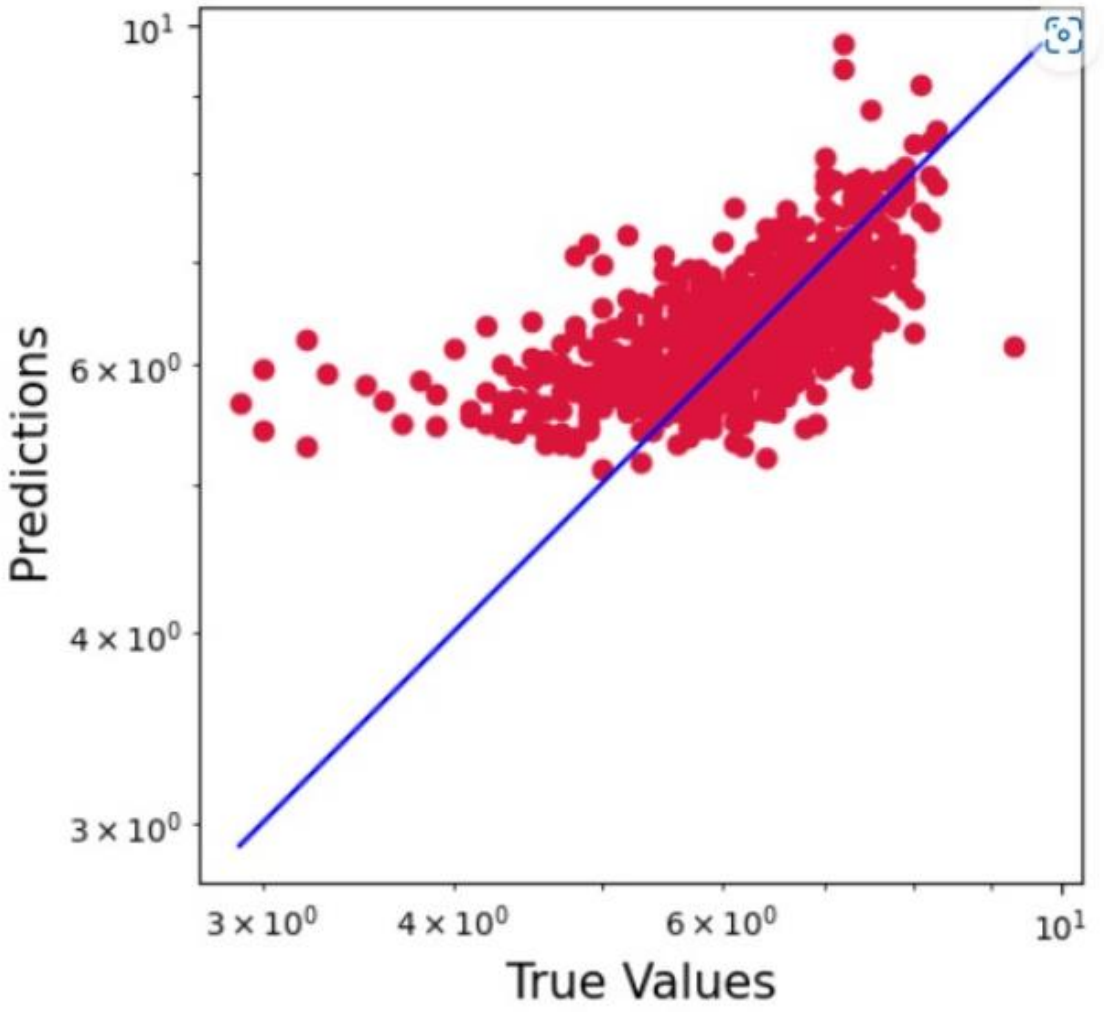
Random Forest:

Mean Square Error random forest: 0.48142488925438603
Score of random forest: 0.3721262919884445
Score2 of random forest: 0.3721262919884445



Ridge Regression:

Mean Square Error RidgeCV 0.4850252144256773
Score of RidgeCV: 0.36743075263058733
Score2 of RidgeCV: 0.36743075263058733



Features used:

The features used after feature selection are:

1- Budget
2- Homepage
3- Id
4- Original language
5- Viewer count
6- Release date
7- Revenue
8- Runtime
9- Vote count

In addition to that, the features that resulted from the conversion of the list of dictionaries into separate columns weren't all used but some of them were selected based on the selection of columns that have a sum of more than 100 which means that it has significant values in only rows less than 100 rows

Size of test and train data:

The train data is 80% of the whole data

The test data is 20% of the whole data

The problems faced us in this milestone and its solution:

1- The columns title, overview, tagline, and original title were of long sentence string so we needed to separate them into single words in separate columns which were processed using NLP techniques such as removing stop words, removing punctuation, tokenization, stemming, and lemmatization
2- The columns of keywords, genres, production countries, production languages, and spoken languages were a list of dictionaries so we needed to convert them to a string by splitting them by name
3- There were a large number of columns which has no effect on the prediction of data so we started eliminating them