Preprocessing:

1-Applying some NLP techniques to title, overview, and tagline columns as removing punctuation, making all words lowercase, tokenization, removing stops words, stemming, limitation, 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-Applying one-hot encoding on the resulting columns 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. We started selecting the appropriate columns by choosing columns that have a significant effect on the data by calculating the sum of values in these columns and if they didn’t exceed the value of 50 we drop them
2. Then ANOVA feature selection technique is applied to data resulting from the NLP process and dictionary conversion to string
3. Then correlation technique is applied to the rest of the columns
4. The top features from both techniques are then combined to be used for training of the models

Regression Techniques used:

1-Linear Regression

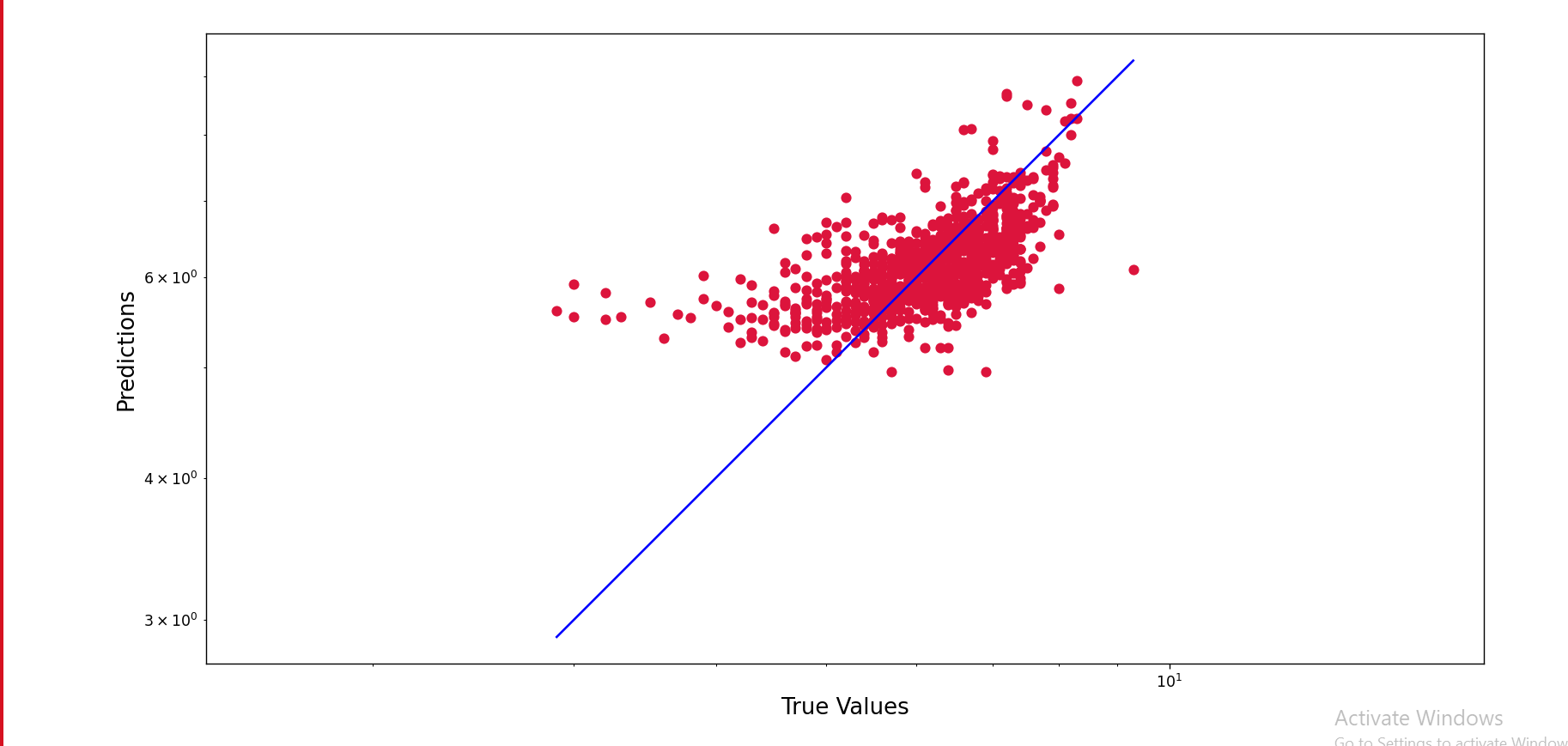
2-Random Forest Regression

3-Ridge Regression

Linear Regression:

It has a mean square error with value of: 0.4493781515826159

And accuracy with value: 0.4139216051531025



Random Forest:

It has a mean square error with value of : 0.4061276578947368

And accuracy with value: 0.47032884219312565



Ridge Regression:

It has a mean square error with value of : 0.4514282242138632

And accuracy with value 0.4112479031210623



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