Movie Popularity Prediction Project

Team ID : CS_60

Members:

روب يبه محمد مصطفى محمد

20201700281

روان محمد أحمد البدوي سيد

20201700279

رنا عفت أحمد عفيفي

20201700269

يوسف اشرف محمد محمد

20201701003

معتز أحمد محمد محمد

20201700849

مصطفى أحمد عبدالرحيم أحمد

20201700823

1- Pre-processing Techniques:

Data Cleaning:

We first applied this on the (home page) column, where if a link exists the value is one, otherwise if there is no hyperlink to the home page the value is zero, we also did the same with the (original language) column where if the language is English then the value is one, otherwise the value is zero, and finally, same with the (status) column, if the movie is released, the value is one, otherwise the value is zero.

We dropped the day values from all the release dates as they are irrelevant data and only used the year and month values as independent variables in our dataset.

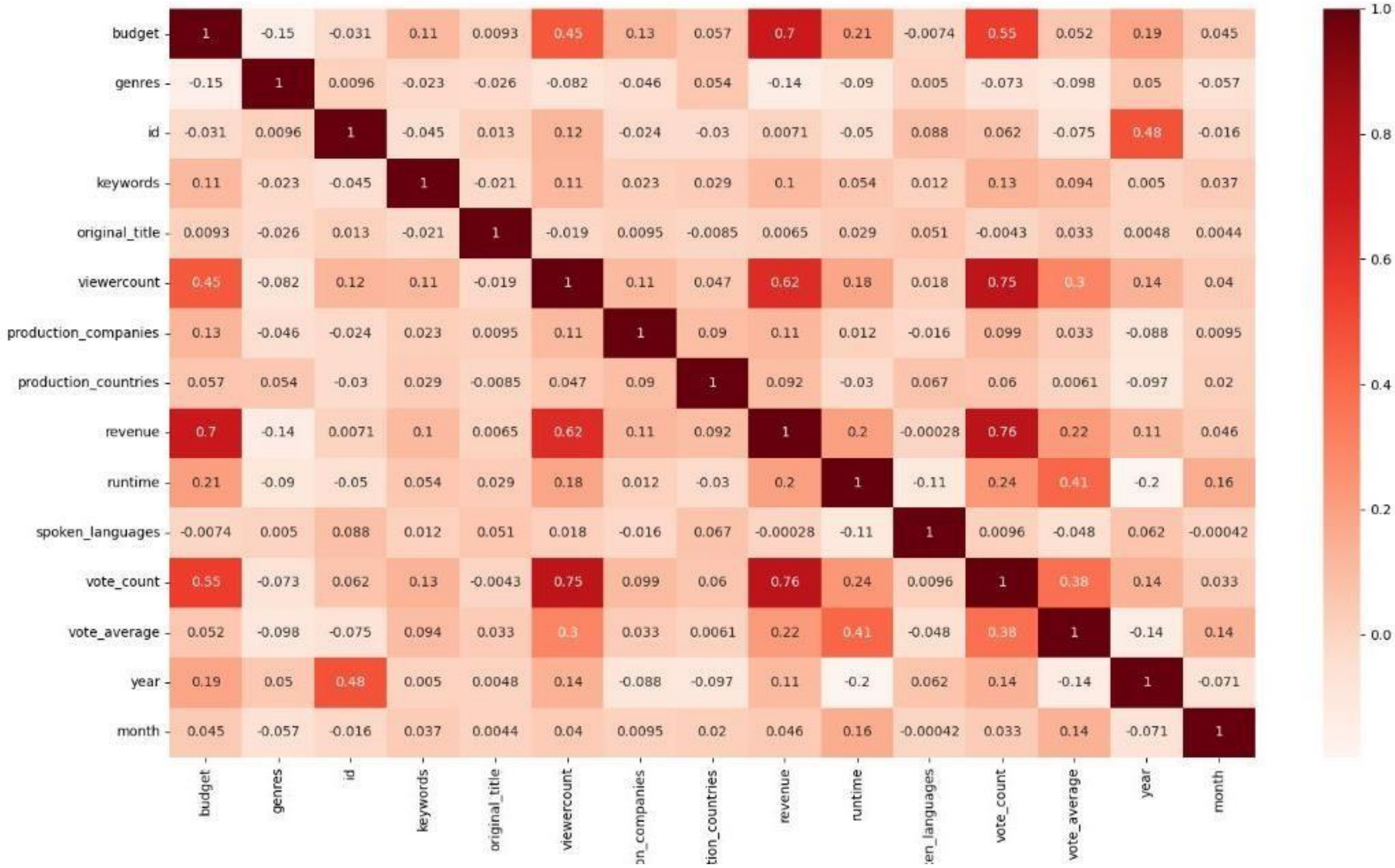
We applied feature encoding on these following columns : (genres , production companies , production countries , keywords , overview , original title , title , tagline , spoken languages) to transform categorial values into numerical ones.

Normalization:

We applied normalization using (MinMaxScaler) on all the columns to round up all the dataset numbers in the same range for better accuracy in the used models.

2- Dataset Analysis:

After applying correlation on the dataset we concluded that the (vote count) variable and (runtime) are the most two variables affecting the prediction.



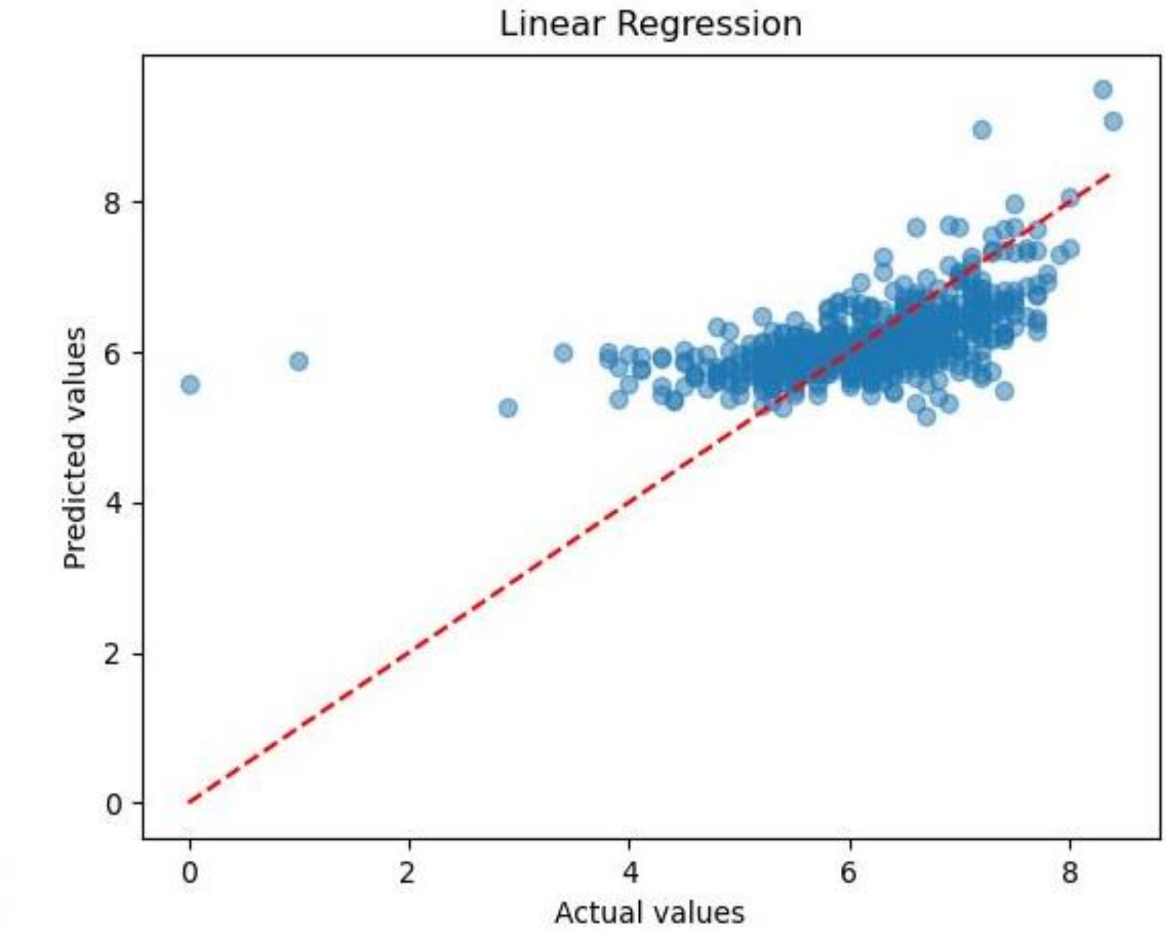
	budget	genres	id	keywords	original_title	viewercount	on_companies	tion_countries	revenue	runtime	en_languages	vote_count	vote_average	year	month
budget	1	-0.15	-0.031	0.11	0.0093	0.45	0.13	0.057	0.7	0.21	-0.0074	0.55	0.052	0.19	0.045
genres	-0.15	1	0.0096	-0.023	-0.026	-0.082	-0.046	0.054	-0.14	-0.09	0.005	-0.073	-0.098	0.05	-0.057
id	-0.031	0.0096	1	-0.045	0.013	0.12	-0.024	-0.03	0.0071	-0.05	0.088	0.062	-0.075	0.48	-0.016
keywords	0.11	-0.023	-0.045	1	-0.021	0.11	0.023	0.029	0.1	0.054	0.012	0.13	0.094	0.005	0.037
original_title	0.0093	-0.026	0.013	-0.021	1	-0.019	0.0095	-0.0085	0.0065	0.029	0.051	-0.0043	0.033	0.0048	0.0044
viewercount	0.45	-0.082	0.12	0.11	-0.019	1	0.11	0.047	0.62	0.18	0.018	0.75	0.3	0.14	0.04
production_companies	0.13	-0.046	-0.024	0.023	0.0095	0.11	1	0.09	0.11	0.012	-0.016	0.099	0.033	-0.088	0.0095
production_countries	0.057	0.054	-0.03	0.029	-0.0085	0.047	0.09	1	0.092	-0.03	0.067	0.06	0.0061	-0.097	0.02
revenue	0.7	-0.14	0.0071	0.1	0.0065	0.62	0.11	0.092	1	0.2	-0.00028	0.76	0.22	0.11	0.046
runtime	0.21	-0.09	-0.05	0.054	0.029	0.18	0.012	-0.03	0.2	1	-0.11	0.24	0.41	-0.2	0.16
spoken_languages	-0.0074	0.005	0.088	0.012	0.051	0.018	-0.016	0.067	-0.00028	-0.11	1	0.0096	-0.048	0.062	-0.00042
vote_count	0.55	-0.073	0.062	0.13	-0.0043	0.75	0.099	0.06	0.76	0.24	0.0096	1	0.38	0.14	0.033
vote_average	0.052	-0.098	-0.075	0.094	0.033	0.3	0.033	0.0061	0.22	0.41	-0.048	0.38	1	-0.14	0.14
year	0.19	0.05	0.48	0.005	0.0048	0.14	-0.088	-0.097	0.11	-0.2	0.062	0.14	-0.14	1	-0.071
month	0.045	-0.057	-0.016	0.037	0.0044	0.04	0.0095	0.02	0.046	0.16	-0.00042	0.033	0.14	-0.071	1

3- Regression Techniques:

We used four different regression techniques.

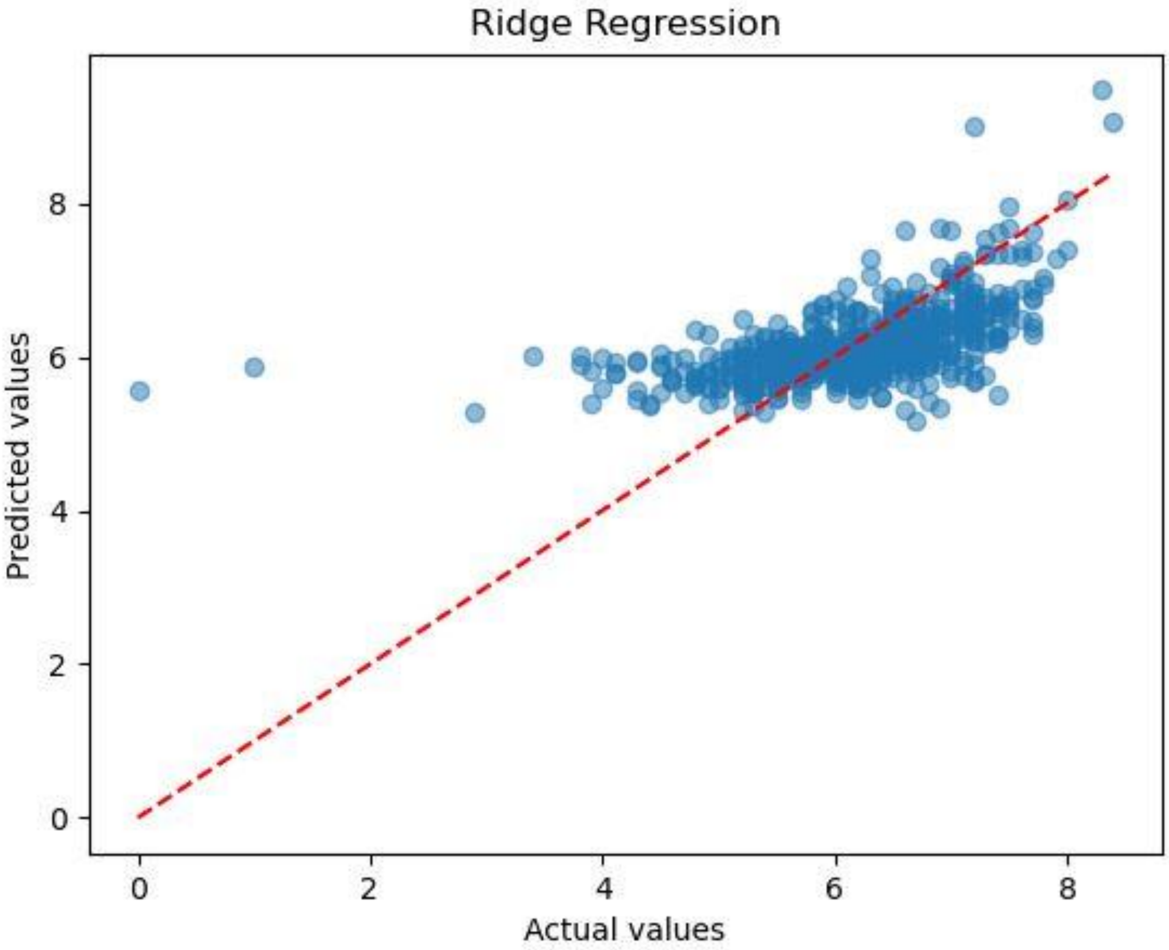
1- Linear Regression:

MSE: 0.5234652932134146



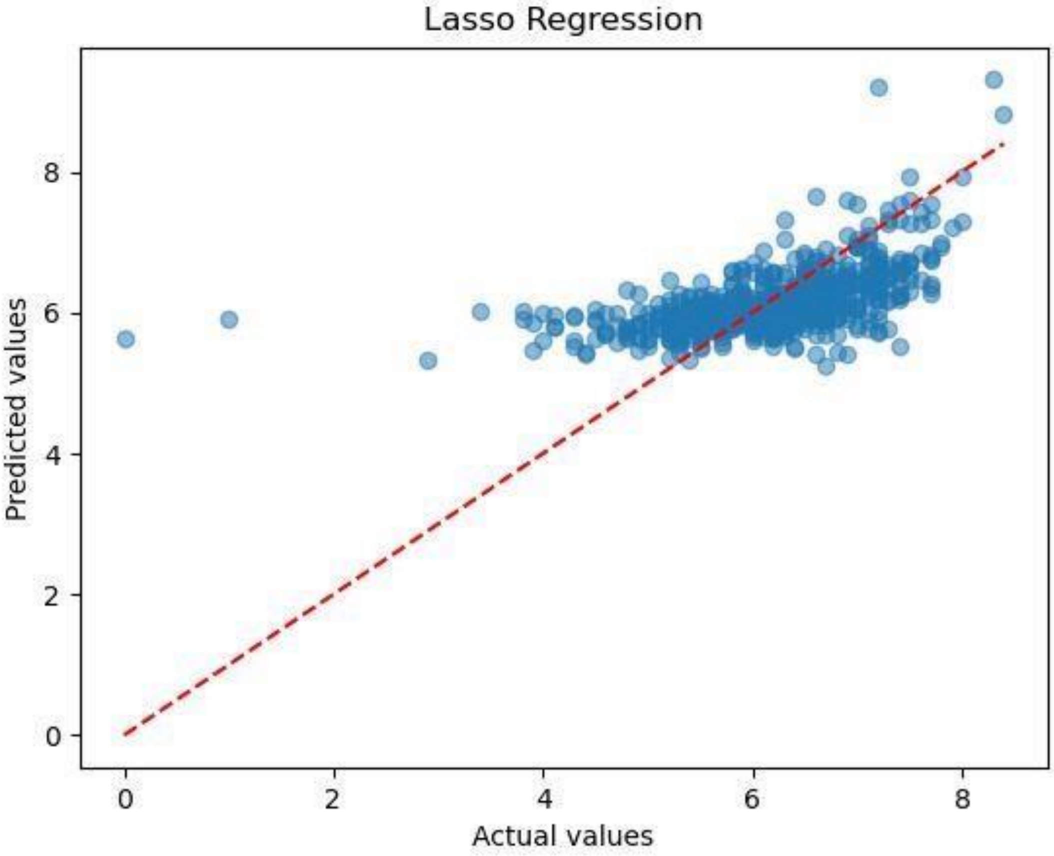
2- Ridge Regression:

MSE: 0.5238328161608443



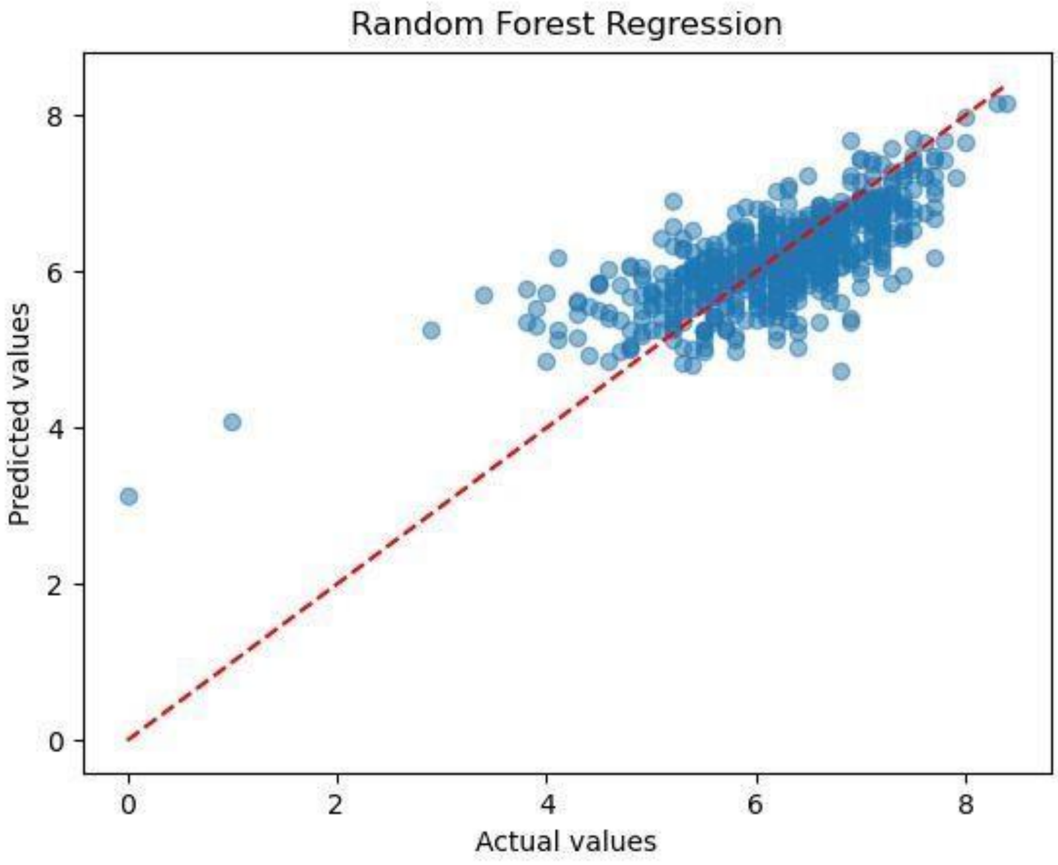
3- Lasso Regression:

MSE: 0.5314973760011776



4- Random Forrest Regression:

MSE: 0.3755459523026317



4- Feature Selection:

The variables dropped are:

(Overview , tagline, original language, title and status) as the prediction values and accuracy were highly affected by them and better without them.

Y-predict value is vote average.

5- Train and Test:

Train set size is 80% of the dataset and the test set size is 20% of the dataset.

6- Project Conclusion:

Our first thoughts about the project were that the (budget) variable and the (genre) variable will be the most effective variables on the prediction output and the accuracy values, meanwhile this was disproved as the most effective variables turned out to be (vote count) and (runtime).