Documentation

I started the project with uploading the dataset (movie_metadat.csv) into python Jupyter Notebook. During the data pre-processing phase, all the N/A values were replaced by the respective means of the columns in the dataset. All the numerical values were also converted into float type for better readability by the Python.

Exploratory Data Analysis

The correlation between all the variables was checked and is reported in the Correlation Matrix as well as in the Scatter Matrix form. No strong multicollinearity was found among the explanatory variables. This is also evident from the Scatter Plot figure. Further, we plot the histograms of all the variables in order to find the pattern of their respective density functions.

Modeling

To start with, we run OLS regression model using imdb_score as the dependent variable. The results are shown in the ipynb file. The significant (p value = 0.00) explanatory variables in the analysis includes num_critic_for_reviews, duration, director_facebook_likes, actor_3_facebook_likes and movie_facebook_likes. However, the value of R square is 14.2%. So, we will try to improve this model using other advanced regression methodologies.

We split the dataset into training and testing dataset with 80:20 ratio. Then we run various model such Linear Regression, Lasso Regression, Decision Tree Regressor and Random Forest Regressor.

Final Results

We find out that Random Forest Regressor has the best R square (29.45%) value among all the models and least MSE value (0.93266). Therefore, we chose Random Forest Regressor for our analysis of movie datasets.