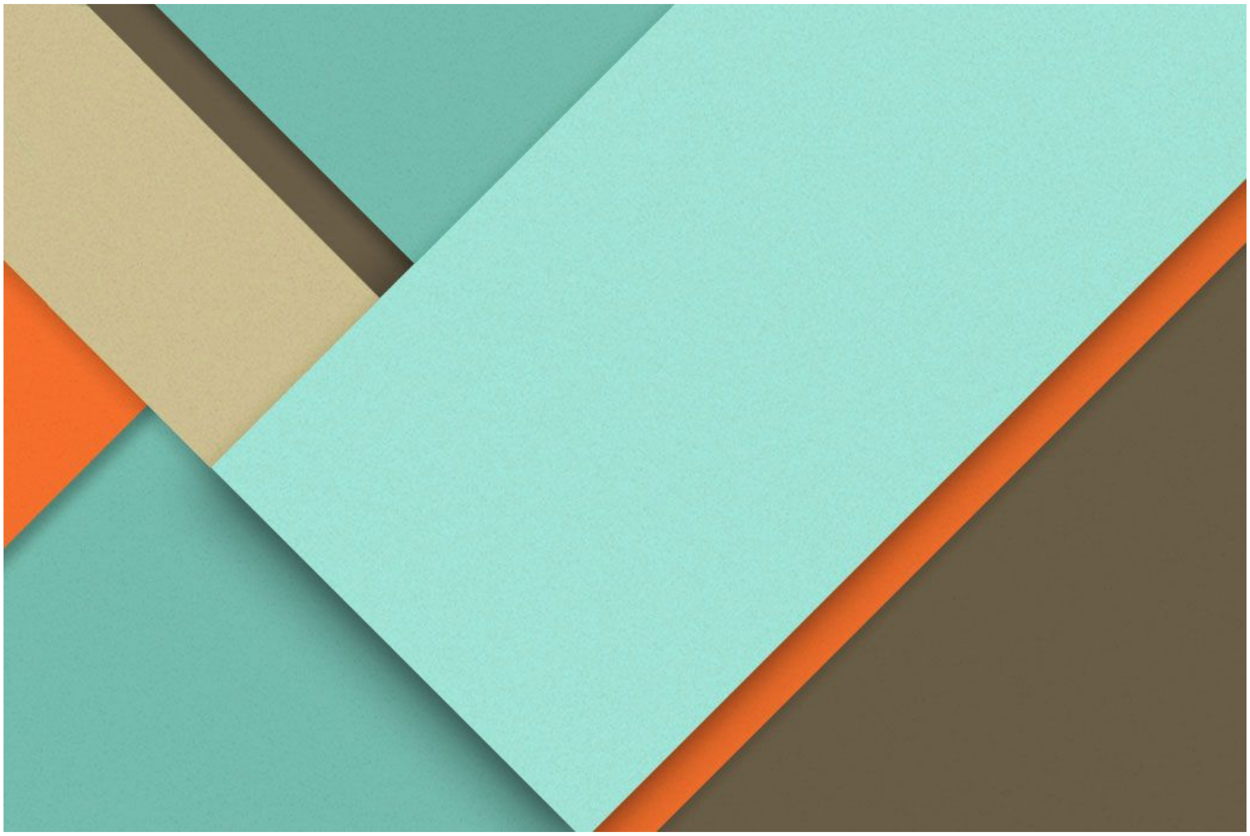


IMDB Rating Predictor

By team ‘Mankind’

Kartik Gupta IMT2016128

Divyanshu Khandelwal IMT2016065

Akash Sharma IMT2016124

Problem Statement

Given the popularity of director and actors, number of critic reviews etc. for different movies. Predict the IMDB rating of a movie.

Data set

Selection of the data was done on the basis of the various features given in the dataset as the most of the features were essential for the prediction of IMDB rating in this dataset. We used google dataset search engine to find the data set. We were able to find it on "data.world".

Link to the data set :
https://data.world/data-society/imdb-5000-movie-dataset

Dataset is comprised of 5042 data points with following features:

Movie title
plot keywords
director_name
no_of_critic_reviews
duration of the movie
facebook likes of director
actor1 name
actor1 facebook likes
actor2 name
actor2 facebook likes
actor3 name
actor3 facebook likes
total facebook likes of the whole cast
movie IMDB link
gross collection
genre
no. of voted IMDB users
no. of actors projected on the poster
aspect ratio
year of releasing
language
country
content rating

Approach

Before watching a movie, viewers consider various factors. Oftenly, genre and popularity of the movie and the lead actor, directors and the cast associated to it becomes the factors to attract mass crowd towards the theatres.

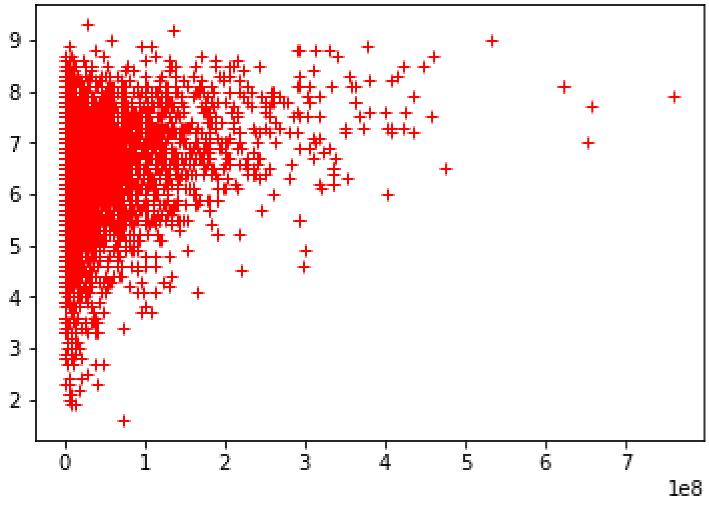
The language in which the movie is projected is also a key factor as more the no. of viewers, more of them are likely to watch a film.

Budget data comes in handy as it also results in gaining popularity among the viewers because of their astounding budgets.

Considering these thoughts in mind we finalized the features like facebook likes of the cast, budget, language and tried to a build a model on that.

Analysis of data and preprocessing

From the domain knowledge and a bit of feature engineering we ended up using with 13 features for our model training. For example after visualisation we concluded that gross of the movie is not related to the IMDB rating of the movie.



Gross vs IMDB rating

Also we dropped columns like name of director and actors which were not at useful for our predictions, as we were using facebook likes as a representative of how good a director or actor is.

Since the original data set was raw. It was having many missing values values and outliers. So preprocessing was an important task to be performed. We used following things to fill up the null values corresponding to the feature :

1.Filled with mean of all data points as the data was not much varying:

No_of_critic_reviews
duration of the movie

2. Since no. of likes was very important feature, we dropped the data points with null values:

facebook likes of director
actor1 facebook likes
actor2 facebook likes
actor3 facebook likes
total facebook likes of the whole cast

3.Filled with median of all data points

no. of voted IMDB users

4.Filled with mode of all data points

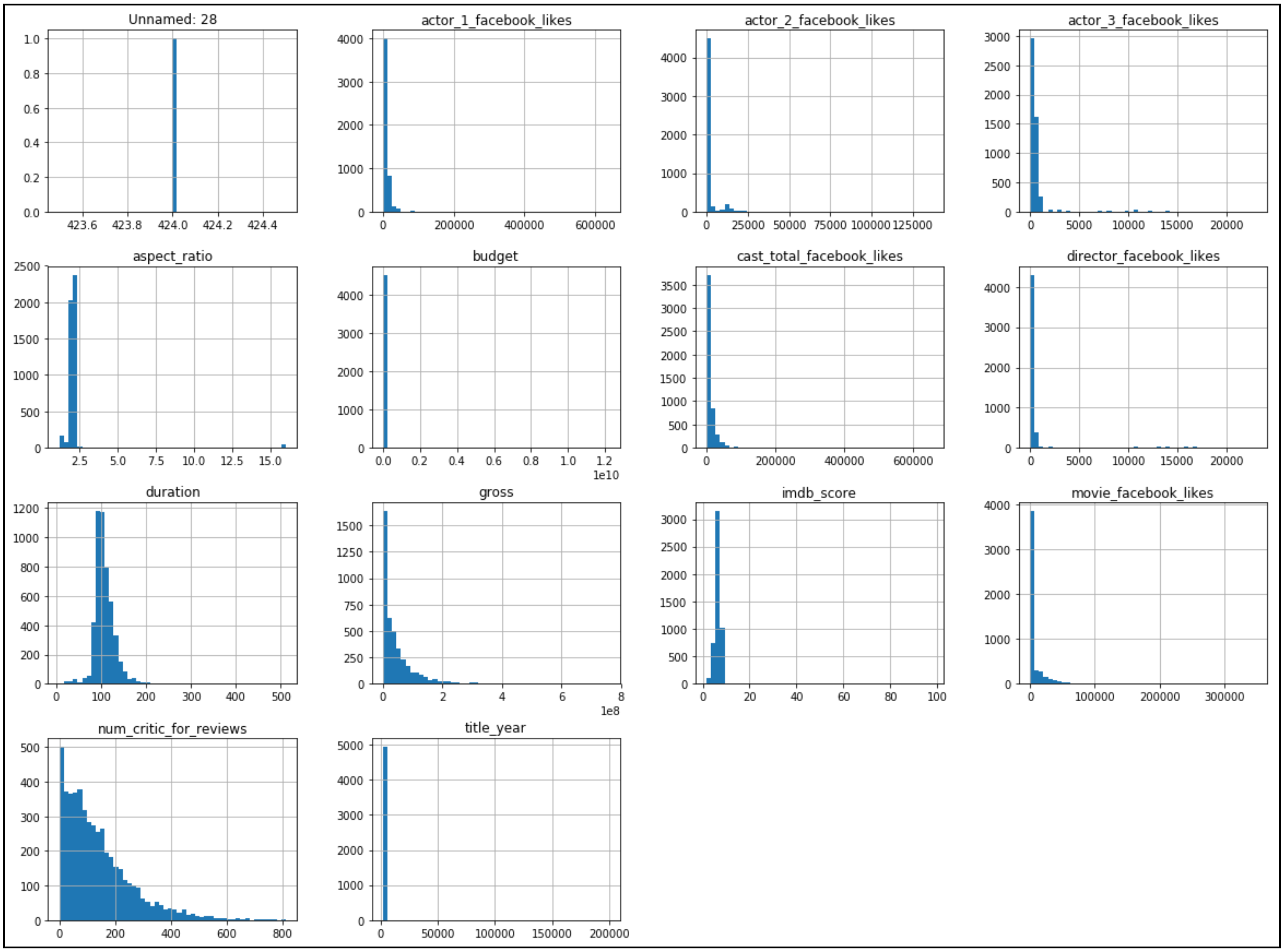
no. of actors projected on the poster
Color

5.Manually filled the data as no. of such data points were very less:

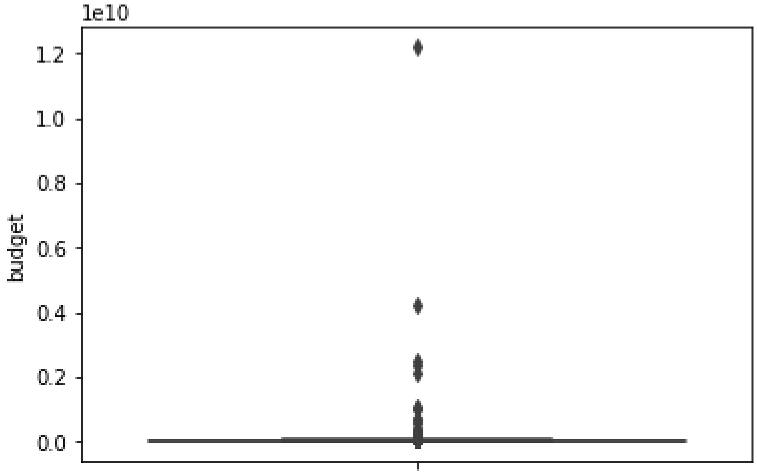
year of releasing
Language
Country

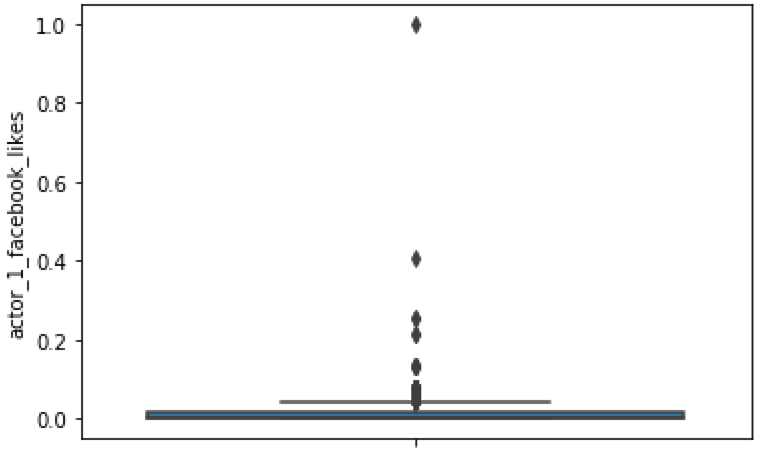
To **analyse** the data we plotted each feature with the label IMDB rating to get a rough estimate of how the IMDB rating is changing with respect to the a feature.

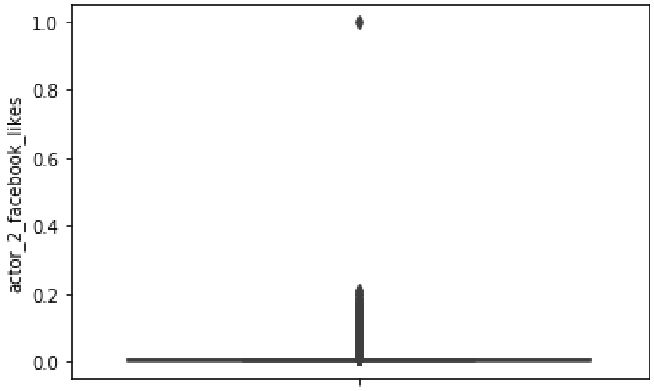
On plotting the histograms of the original data we found that in most of the features the data is concentrated in small range. So we to spread it up we used normalised the values of most of the features by (max -min) value of the data points corresponding to that feature.



Also when plotted box plots for different features, we found that that there are many outliers in that feature which need to be removed.







For the train data to fit in any of the model we import it was necessary the data is always in numerical form. But since we were having some categorical features like color, country and language. So we used label encoding to get them to numerical values.

Model Building

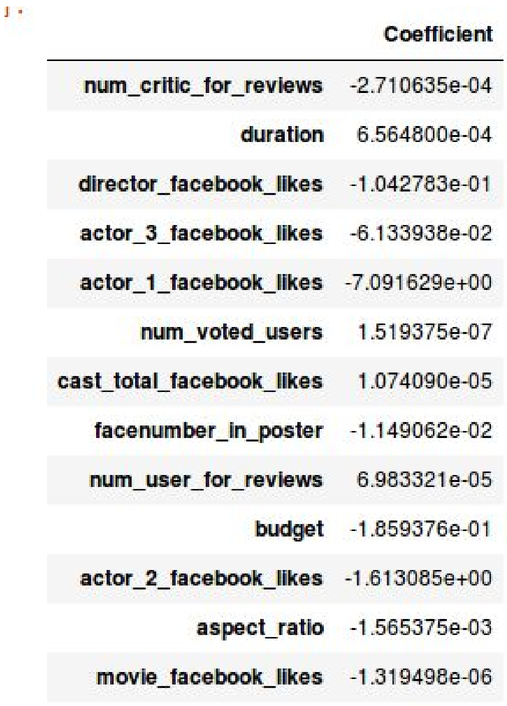
We first tried predicting exact IMDB rating of the movies through n dimensional regression models .

Linear regression model

We first used linear regression model where we were getting very less accuracy score as this model is very sensitive to the outliers of the features.

Random forest

Then we looked into dependency of IMDB rating on each feature. We out the following conclusion:



	Coefficient
num_critic_for_reviews	-2.710635e-04
duration	6.564800e-04
director_facebook_likes	-1.042783e-01
actor_3_facebook_likes	-6.133938e-02
actor_1_facebook_likes	-7.091629e+00
num_voted_users	1.519375e-07
cast_total_facebook_likes	1.074090e-05
facenumber_in_poster	-1.149062e-02
num_user_for_reviews	6.983321e-05
budget	-1.859376e-01
actor_2_facebook_likes	-1.613085e+00
aspect_ratio	-1.565375e-03
movie_facebook_likes	-1.319498e-06

We then converted our regression problem to classification by putting the integral rating as classes and rounding of the corresponding values.

Training and Validation

Results are summarized in the table

- First, the accuracy was calculated by letting the model only predict the exact rating by regression.
  - Which lead to following scores

Model	Score(%)
Linear Regression	26.3
Random forest	36

- Then we decided to turn the problem into classification problem by taking 10 classes 1-10 i.e the rating.
  - Again,the accuracy was calculated by letting the model only predict the exact rating.
    - Which lead to following scores

Model	Score(%)
Logistic Regression	33.8
SVC	34.7
Naive Bayes(Gaussian)	33.11

  - Secondly, the accuracy was calculated by letting the model only predict the range of rating i.e with the error of +-1, so for e.g if the rating predicted was 8 then the accuracy was tested if the actual label was between 8-1 to 8+1.

- Which lead to these results

Model	Score(%)
Logistic Regression	82.71
SVC	77.87
Naive Bayes(Gaussian)	78.45

Conclusion

Finally, the Conclusion that I made from the results

- Predicting the rating does not only depends upon the popularity factors as these features may affect the openings of the movies, but the key factor is the story of these as people rate the movies on that only.

- There seem very less relation between the statistics and the ratings since some actors may be good but the movie's facebook likes wouldn't be that high because of less publicity of the movie. IMDB ratings can be still high even if the a small mob watches a movie but rate it high.
- Logistic Regression gave the best results as compared to other classification algorithms.