IMDb Movie Rating Prediction

Balakrishnan Nagaraj

May 24, 2022

1 Introduction

We will be using IMDB 5000 Movie Data set to build models that predict the IMDB score based on a bunch of features associated with a movie. IMDB score is an important measure of how good a movie is and is maintained by IMDb (Internet Movie Database) owned by Amazon. As of March 2022, the database has records of over 600,000 movies and over 6.5 million TV episodes.

2 IMDb 5000 Movie Data

The data set has 5043 movie records with 28 attributes. Some of the attributes are movie title, director name, actor name, genre, content rating, country, language, facebook likes for the director, actors and movie, IMDb score etc. A complete list of the attributes can be found in appendix A.

3 Data Visualizations

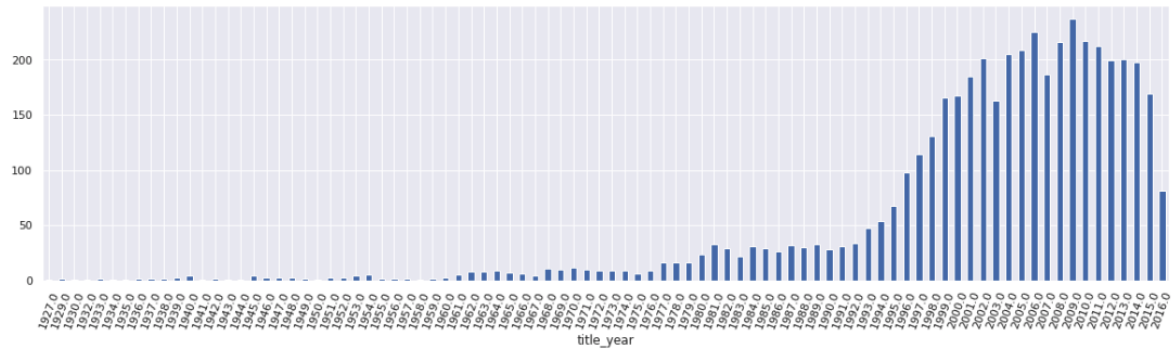

Figure 1: Distribution of movies by year.

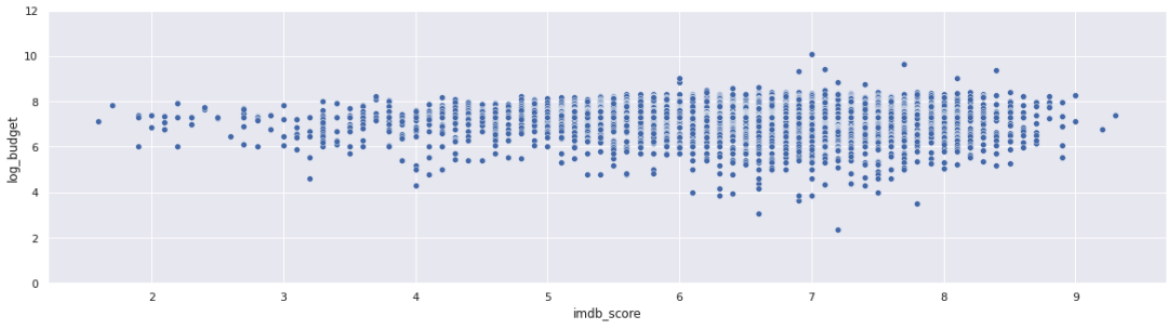

Figure 2: Budget (in log scale) vs IMDb score.

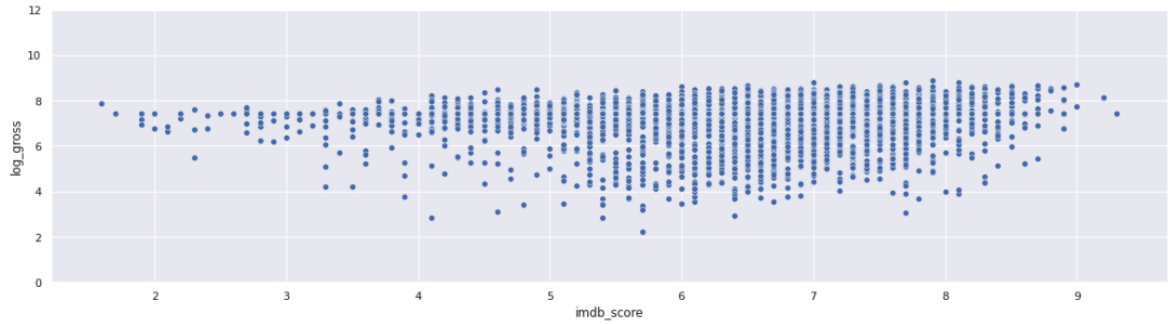

Figure 3: Gross (in log scale) vs IMDb score.

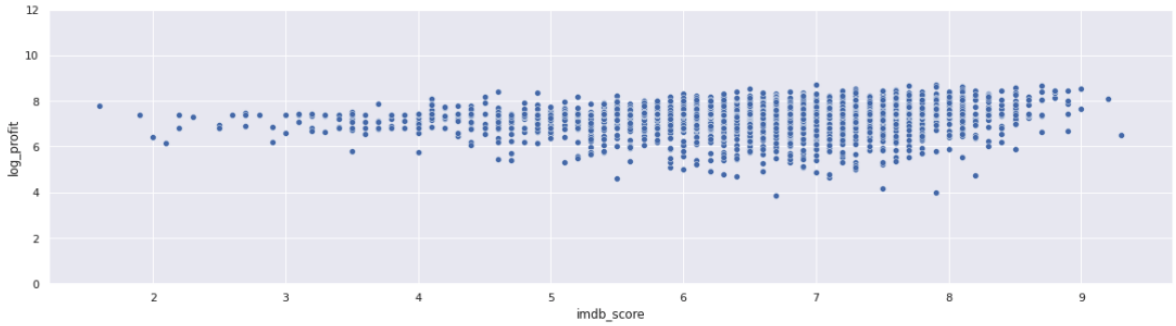

Figure 4: Profit (in log scale) vs IMDb score.

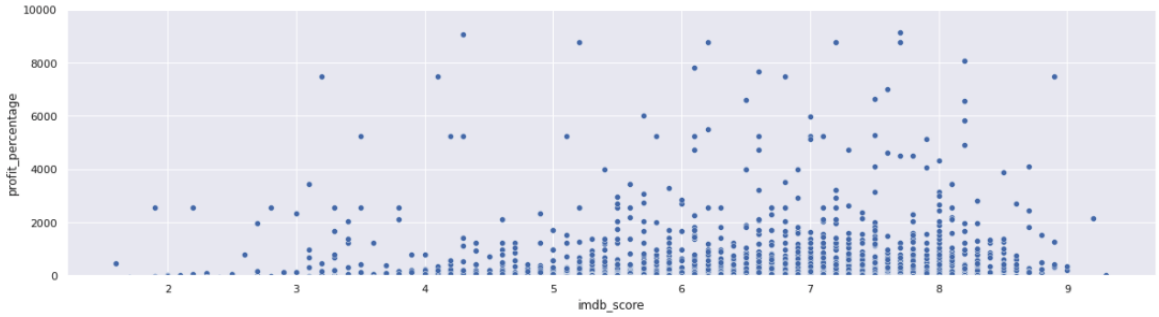

Figure 5: Profit percentage vs IMDb score.

4 EDA and Data Preprocessing

EDA and data preprocessing work is available in IMDB_mrp_EDA_and_Preprocessing.ipynb.

Steps involved in getting the data ready for model training and selection are:

1. Initial data exploration
2. Getting rid of unnecessary features
3. Dealing with missing values (either by deletion or filling in)
4. One-hot encoding categorical variables
5. Train-test split
6. Scaling and normalization of quantitative features
7. Removing highly correlated features

The following variables were used in the final train, val and test data sets:

1. **Categorical variables**: Top 4 countries, top 4 content ratings, 23 genres
2. **Quantitative variables**: budget, num_critic_for_reviews, duration, director_facebook_likes, actor_3_facebook_likes, actor_1_facebook_likes, gross, num_voted_users, facenumber_in_poster, actor_2_facebook_likes, movie_facebook_likes

The categorical variables were selected based on their relevance and on the feasibility of one-hot encoding them without adding too many features. The quantitative variables were selected by using their joint plots with IMDb score and assessing the usefulness of the variable in predicting the IMDb score.

Special attention was paid to the following issues:

1. **Dummy variable trap**: Remove one column of one-hot encoded categorical variable to avoid correlation between features.
2. **Multicollinearity**: Find the correlation matrix of quantitative variables and remove any correlated feature that has a correlation coefficient of greater than 0.7 with any other feature.
3. **Data leakage**: First perform the train-test split and then impute the missing values (in train, val and test set) based only on the train data. This ensures that there is no data leakage from test set to train set.
4. **Feature normalization**: Decision tree based algorithms are scale invariant. But neural networks perform better with normally distributed features. Most of our quantitative variables turned out to have a highly skewed distributions. Hence sklearn PowerTransformer was used to normalize the features.
5. **Curse of Dimensionality**: A conscious choice was made to not use the variables director_name, actor_1_name, actor_2_name and actor_3_name in the final train (val and test) set. The only way to use these features would have been as one-hot encoded categorical variables. This results in over 10000 features. To have over 10000 features with only 5000 data points would result in high variance models. One way to reduce the number of features is to limit only to top-10 (for example) frequent categories for each variable. But this turned out to miss a lot of data records. Hence it was decided to altogether drop these variables.

5 Model Selection

The work concerning model selection is available in the IMDB_mrp_Model_Selection.ipynb notebook.

The following models were trained on the data:

1. Decision Tree Regressor
2. Random Forest Regressor
3. XGBoost Regressor
4. Neural Network

To control the problem of overfitting in the first three models, an optimal value for the hyperparameter max_depth (maximum depth of trees) was found using a simple for loop based search. For the neural network model, we just choose the model that provides the least validation error.

The performance of the models on the test data is tabulated in Table 1.

Based on mean absolute error, we see that the random forest regressor regressor works better than other models, although the XGBoost regressor is a close second.

Model	Mean Absolute Error	Accuracy
Decision Tree Regressor	1.0074	87.31%
Random Forest Regressor	0.9991	89.54%
XGBoost Regressor	1.0044	89.77%
Neural Network	1.0592	79.92%

Table 1: Performance of the models.

6 Conclusion

After a careful data preprocessing, model building and hyperparameter tuning, it was found that random forest regressor works best for the IMDb score prediction task. Further hyperparameter tuning would allow us to get more juice out of the models.

A IMDb Movie Data

IMDb Movie Data has the following attributes:

1. Color: Movie is black or coloured
2. Director_name: Name of the movie director
3. num_critic_for_reviews: No of critics for the movie
4. duration: movie duration in minutes
5. director_facebook_likes: Number of likes for the Director on his Facebook Page
6. actor_3_facebook_likes: No of likes for the actor 3 on his/her facebook Page
7. actor2_name: name of the actor 2
8. actor_1_facebook_likes: No of likes for the actor 1 on his/her facebook Page
9. gross: Gross earnings of the movie in Dollars
10. genres: Film categorization like 'Animation', 'Comedy', 'Romance', 'Horror', 'Action', etc.
11. actor_1_name: Name of the actor 1
12. movie_title: Title of the movie
13. num_voted_users: No of people who voted for the movie
14. cast_total_facebook_likes: Total facebook like for the movie
15. actor_3_name: Name of the actor 3
16. facenumber_in_poster: No of actors who featured in the movie poster
17. plot_keywords: Keywords describing the movie plots
18. movie_imdb_link: Link of the movie link
19. num_user_for_reviews: Number of users who gave a review
20. language: Language of the movie
21. country: Country where movie is produced
22. content_rating: Content rating of the movie
23. budget: Budget of the movie in Dollars
24. title_year: The year in which the movie is released
25. actor_2_facebook_likes: facebook likes for the actor 2
26. imdb_score: IMDB score of the movie
27. aspect_ratio : Aspect ratio the movie was made in
28. movie_facebook_likes: Total no of facebook likes for the movie

B Further Data Visualizations

We provide a few more data visualizations in this appendix.

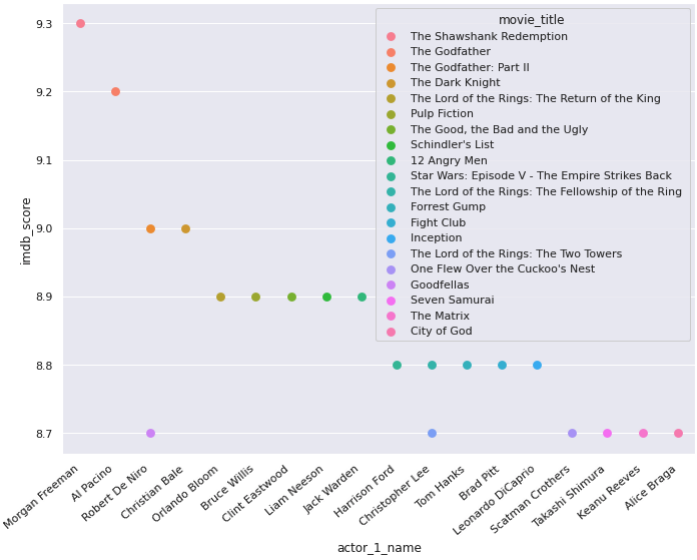

Figure 6: Top 20 lead actors by their movie IMDb scores.

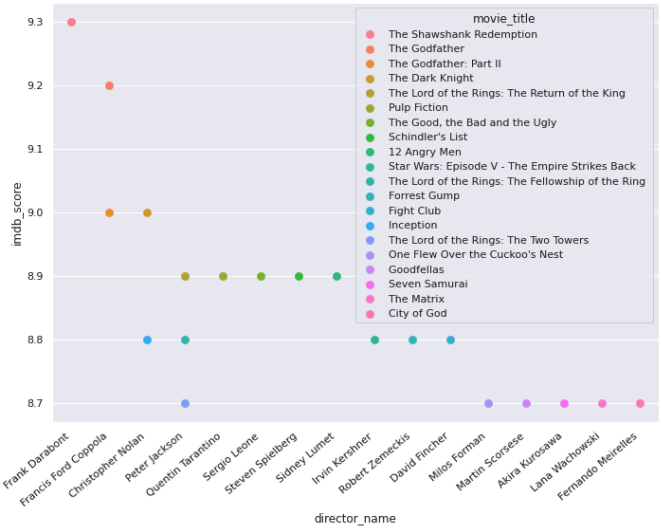

Figure 7: Top 20 directors by their movie IMDb scores.

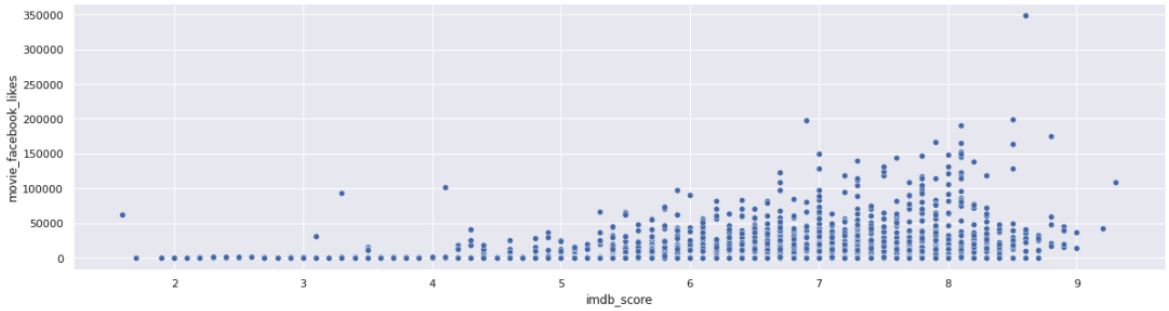

Figure 8: Movie facebook likes vs IMDb scores.