Prediction of Netflix Movie Rating

Damayanti Naik, PhD

Data science Career Track, Capstone Project, December 2020 Cohort

Springboard

Thanks to springboard mentor:
Yadunath Gupta

About Netflix

Netflix is a subscription-based online streaming service that offers streaming of films and television series.

It started in 1997 in USA and now makes its service available in most of the countries in the world.

Netflix recommendation system

Netflix is well known for its efficient recommendation engines providing users choice of movies/shows. The engines work behind the scene and based on:

- Content-based filtering algorithm
- Collaborative filtering algorithm
- Hybrid of both

Problem statement

In Netflix's recommendation system, user's rating plays an important role. Here, I'll build a predictive model to predict movie rating (user review).

To carry out this

- Different predictive models will be developed to predict movies' rating.
- The best one will be selected based on the R2 score (co-efficient of determination)

  i.e how close the actual ratings are to the predicted values.

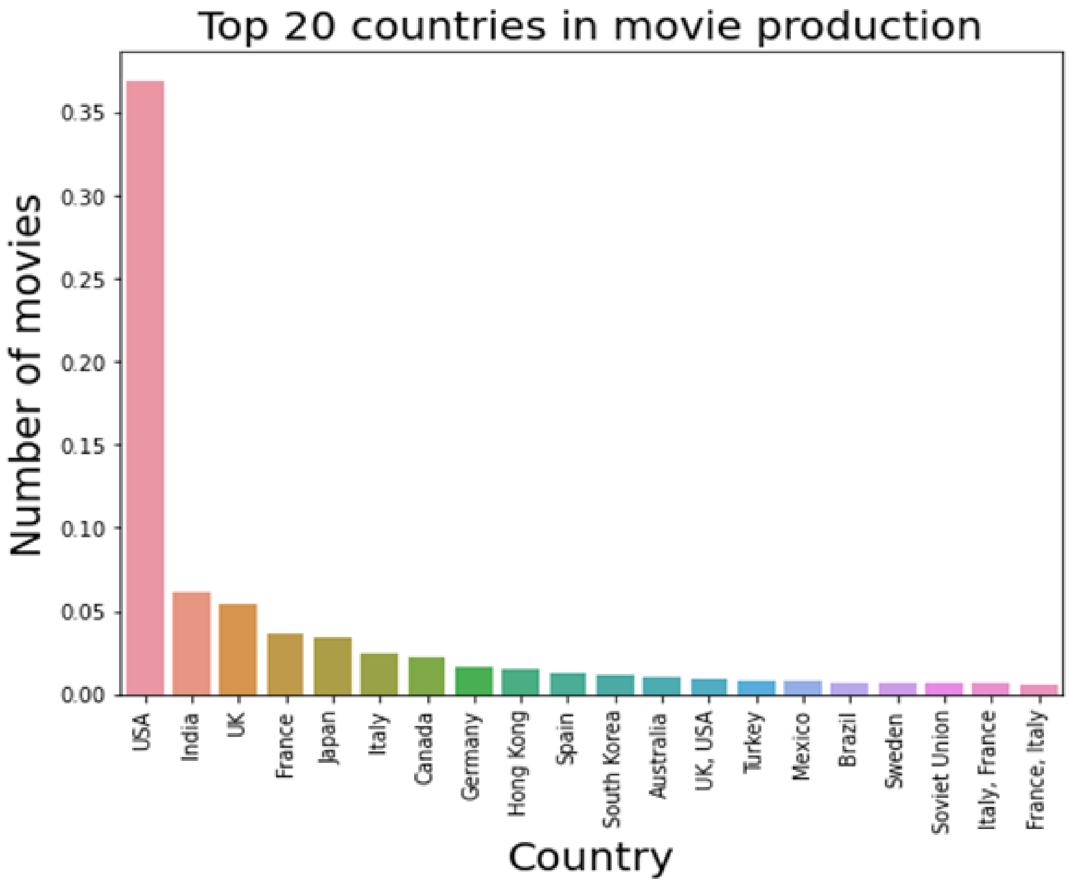
Data

Data has been collected from

- Kaggle
- IMDB

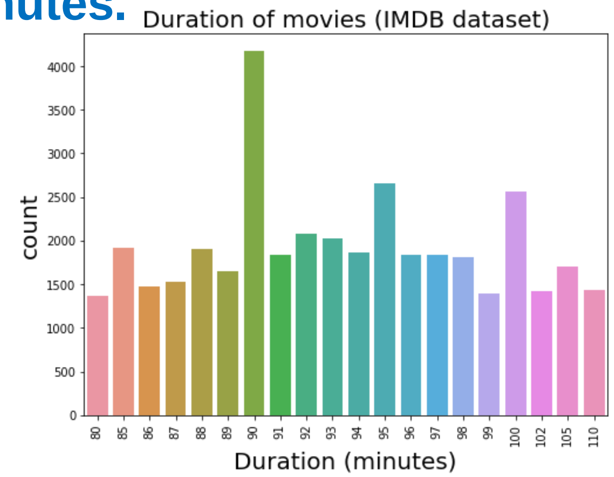
Key findings from Exploratory Data Analysis

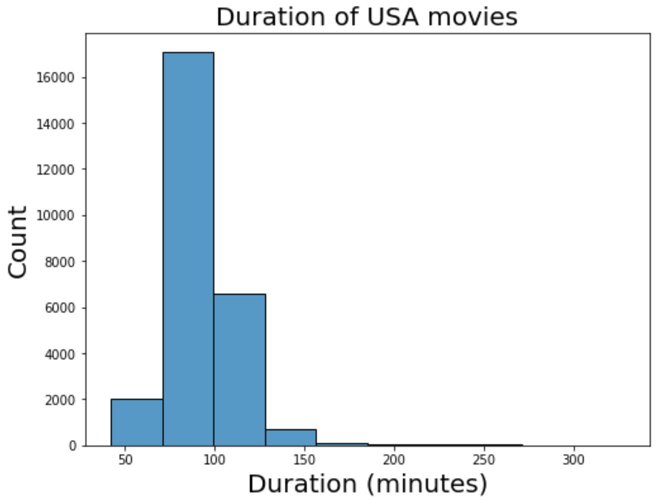
USA ranked Number One in movie production, while India ranked Second followed by United Kingdom in Third place.

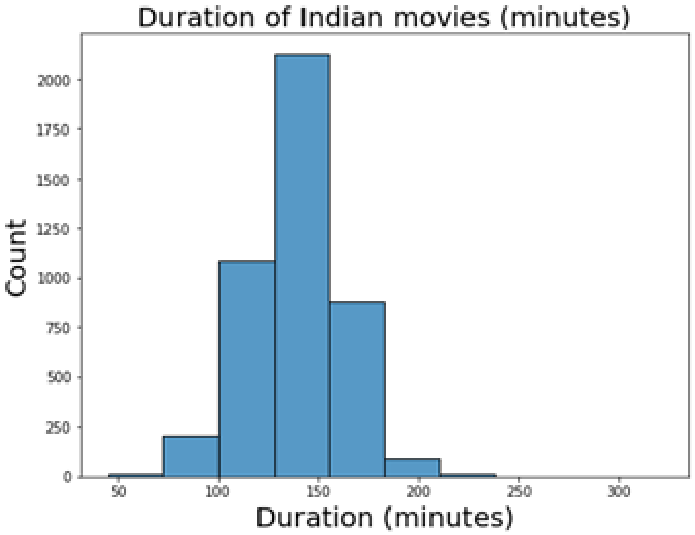


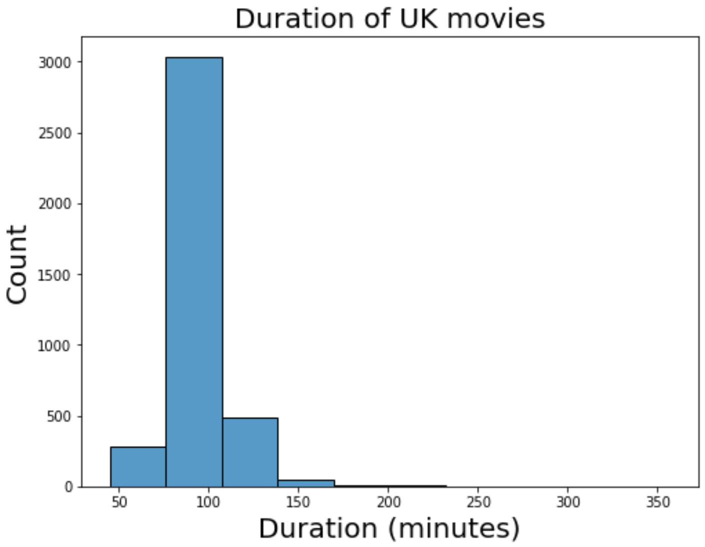
Key findings from Exploratory Data Analysis (continued):

- **Most of the movies are of duration (approximately) 90 mins. However, it varies in different countries. In USA and UK, movies are mostly of 90-100 minutes, whereas in India, movies are mostly of 120-150 minutes.**


Duration of movies (IMDB dataset)


Duration of USA movies


Duration of Indian movies (minutes)


Duration of UK movies

Key findings from Exploratory Data Analysis (continued):

- **These are three top movie production companies:**

  1. **Metro-Goldwyn-Mayer**
  2. **Warner Bros.**
  3. **Columbia Pictures**

  Top movie production company of USA: Metro-Goldwyn-May
  Top movie production company of India: NH Studioz
  Top movie production company of UK: The Rank Organisation

Feature Engineering

Two new datasets were formed out of Kaggle and IMDB dataset before feature engineering:

- **Movies found only in IMDB dataset (not in Kaggle) were used for predictive model building.**
- **Movies common to both were used for testing the model.**

Feature engineering was carried out as follows:

- **Dummie variables were created out of categorical variables: genre, language, actors, directors, production company.**
- **Genre, language have <300 unique values, so dummie values for all these features were created.**
- **Actors, directors, production company have >300 unique values. Out of them, for top 200 unique values, dummie values were obtained.**
- **Standard scaling was carried out on duration time, votes, reviews from users columns.**
- **Outliers are filled with 95th percentile of the values of respective column.**

Model development

Different Regression models were developed to predict the movies rating:

- Simple Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- Gradient Boosting Regressor

Models' performance

- The Simple Linear regression model performance was poor, however it improved significantly when regularization was applied. Among Lasso and Ridge, the later performed best.

- To have better performance, ensemble model Random Forest Regressor was developed. It improved the model performance as compared to Linear regression.

- To achieve much higher performance, ensemble model Gradient Boosting Regressor was also developed. This model had the best performance among all the models.

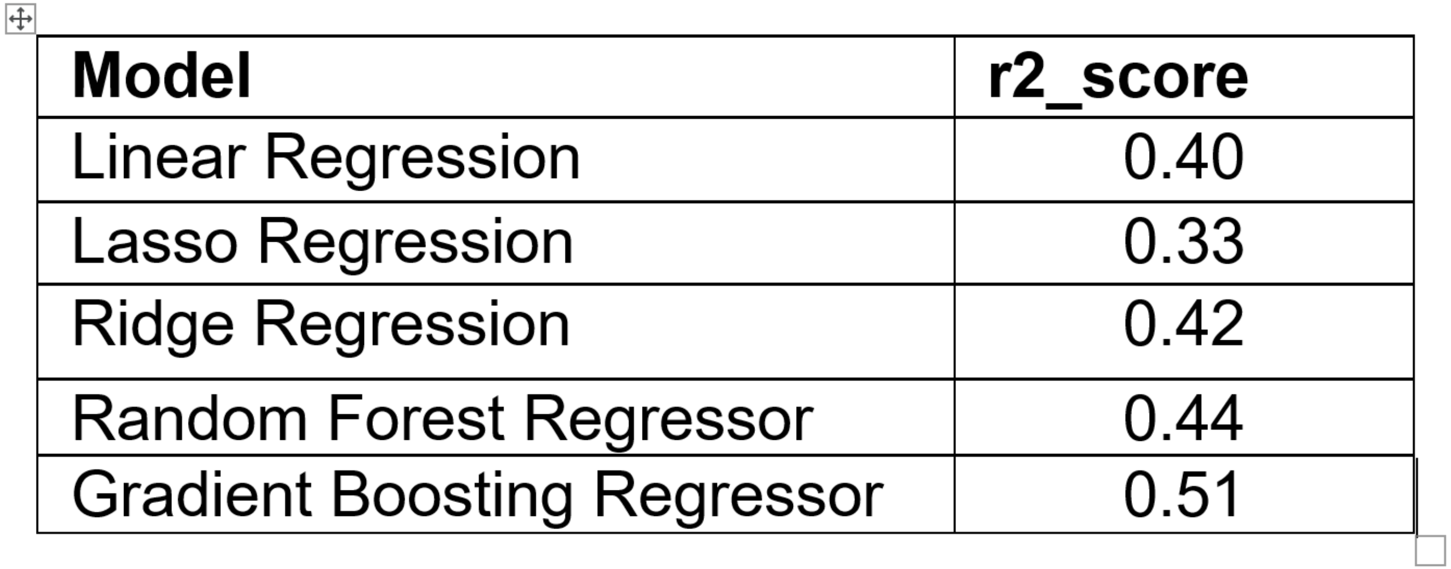
Principal Component Analysis application

To improve the models' performance, PCA (Principal Component Analysis) was applied and models were trained again.

Though it improved Linear Regression's performance and run time, it didn't help both ensemble models.

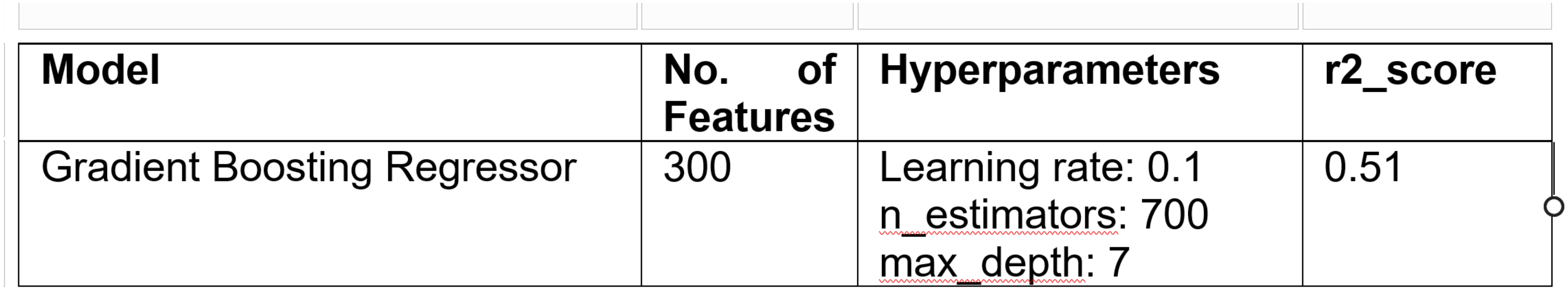
Best model

Gradient Boosting Regressor was the best among all the predictive models.



Model	r2_score
Linear Regression	0.40
Lasso Regression	0.33
Ridge Regression	0.42
Random Forest Regressor	0.44
Gradient Boosting Regressor	0.51

Gradient Boosting Regressor details



Model	No. of Features	Hyperparameters	r2_score
Gradient Boosting Regressor	300	Learning rate: 0.1 n_estimators: 700 max_depth: 7	0.51

Model for future Use

Among all the predictive models, as Gradient Boosting Regressor was the best with R2 score 0.51, it was saved for deployment.

Future Recommendation

The model’s performance can be improved further with Inclusion of more movie features: music quality, picture quality, chorography quality, actors’ ranking, users’ age, etc. Hence data on these should be included in future model building.