IMDB Movie Rating Prediction

Project Report

By Son Nguyen

1. Problem Definition

Question: How can we rate a movie before it is officially released in theatre? Many use their instincts to judge the quality of a film while others rely on critics. However, on one hand it takes considerable amount of time to obtain a reasonable amount of critics' review after a movie is released. On the other hand, human instinct sometimes is unreliable.

Since every year there are thousands of movies produced, is there a better way for us to gauge a movie's merit?

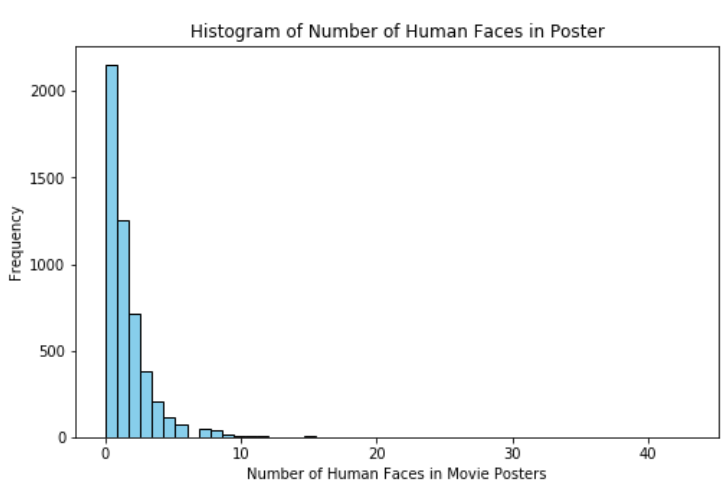
2. The Dataset

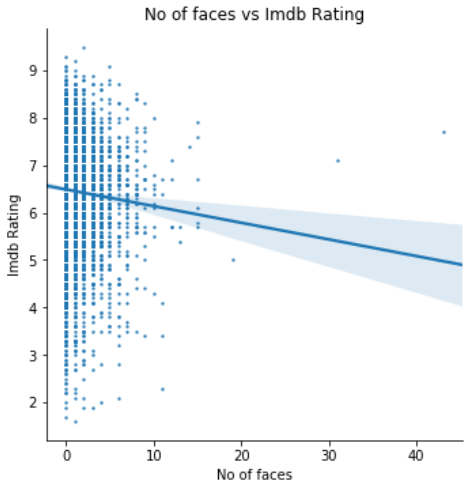
The dataset includes information of over 5000 movies which span over 100 years and 66 countries. Many important information is considered and scraped from the IMDB website such as movie title, director name, cast list, genres, year, duration, etc. Totally, there are 28 different features in the dataset including the IMDB rating.

We will utilize 27 features to make predictions of the movie rating score. Let's dive in some exploratory data analysis.

3. Data Exploration

When a movie is released in cinema, the first thing you notice is its poster. Often people make it as attractive as possible because its purpose is to entice audience to pay for the movie's ticket. One of the features in this dataset is 'facenumber_in_poster', or the number of faces in the poster. Let's see how it correlates with the IMDB score.



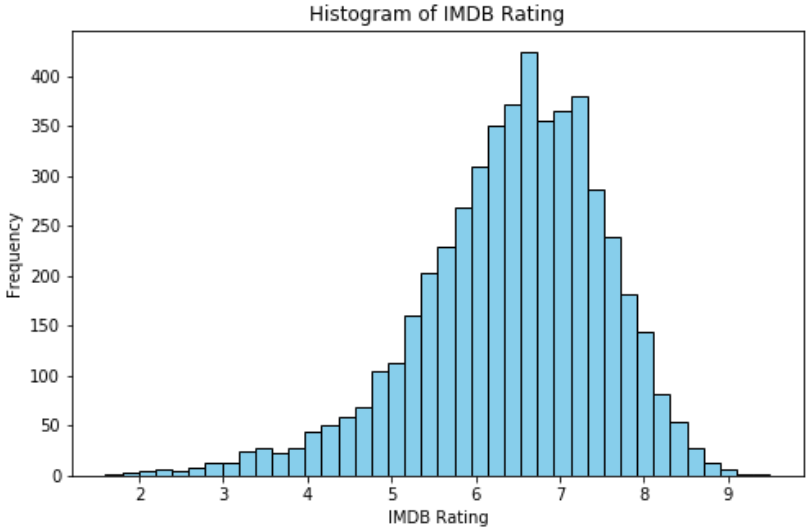


Number of faces in poster vs IMDB Rating

We can see that most of the movie poster does not have any face on it. And the more faces on the poster, the less rating the movie gets. That is described by the trendline in the graph *No of faces vs Imdb Rating.*

Histogram of the IMDB Rating

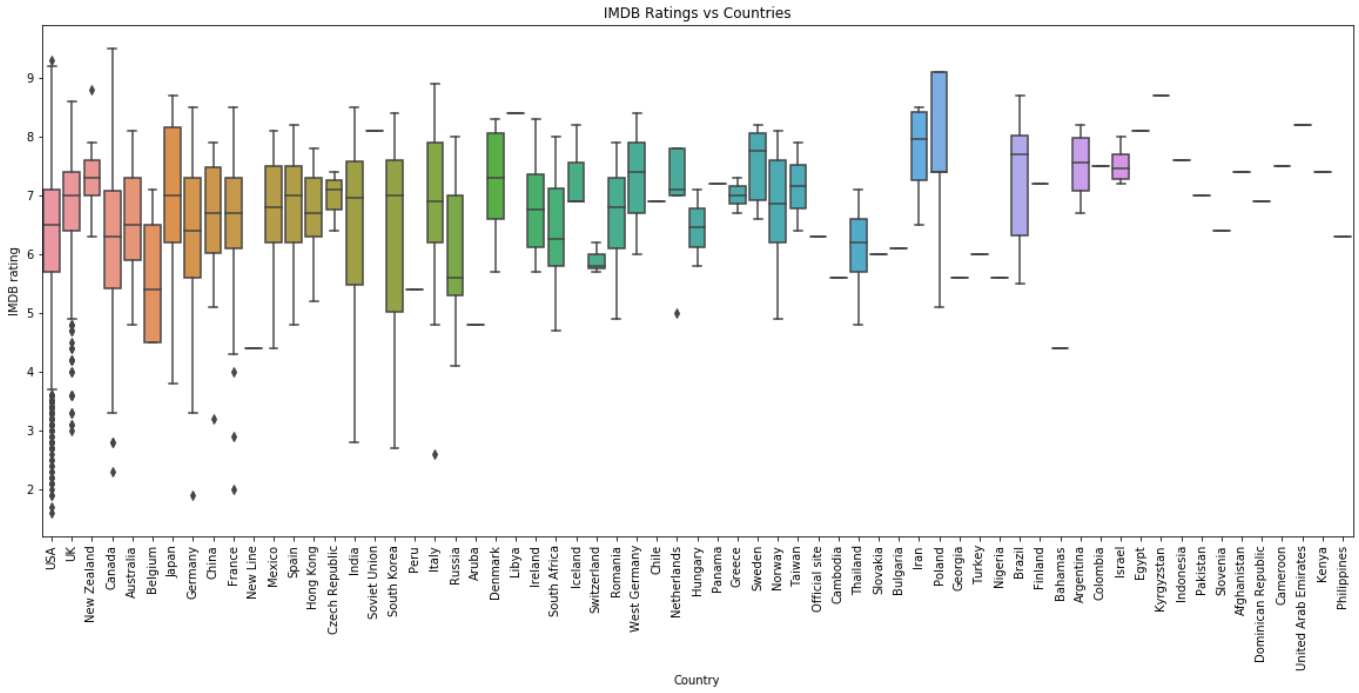
Now, let's see the histogram of the IMDB Rating.



Films with rating higher than 8.0 are in the IMDB top 250, they are great movies. Films with rating from 7.0 to 8.0 are good ones, people can get some perspectives from them, while movies with rating from 1 to 5 are generally considered as bad.

IMDB rating vs Countries

How about the relationship between IMDB rating vs Countries? Is there any interesting discovery? Let's graph it.



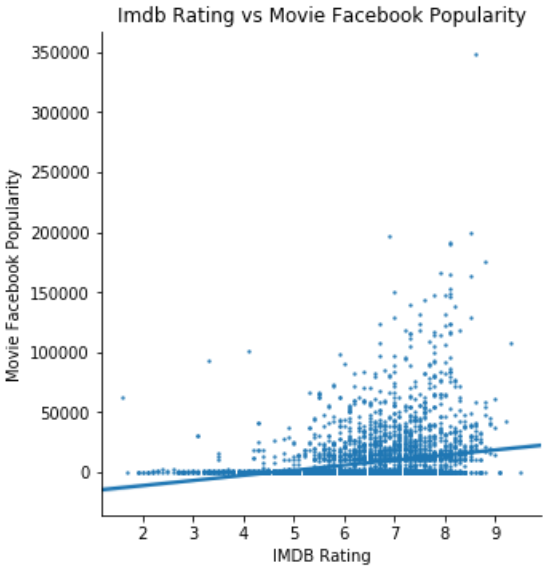
Interestingly, the median IMDB scores of both US and UK are not the highest among all countries. Although most of the movies produced in the past 100 years.

IMDB Rating VS Movie Facebook Popularity

Another interesting question is whether a movie Facebook likes of its social popularity would have any effect on its rating.

So, I draw the scatter plot and the regression line between the movie Facebook popularity and its IMDB Rating.

As we can see in the graph, movies with high Facebook likes tend to have good IMDB ratings as described in the trendline. However, some great movies with very high scores have very low facebook popularity. Those are movies at the right bottom of this scatter plot.

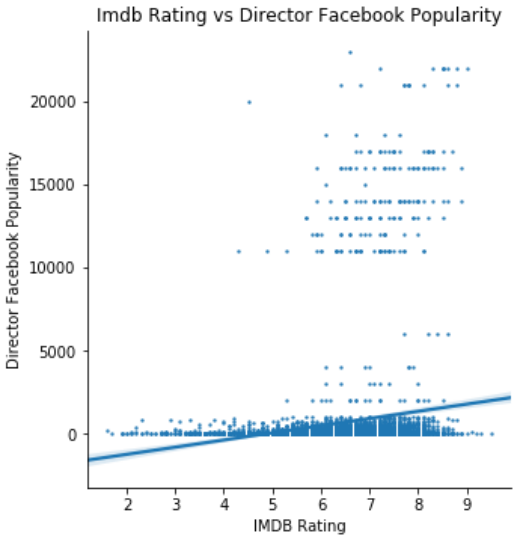


It is understandable to believe that the greatness of a movie is highly affected by its director. How about the relationship between a director social media popularity and his directed movie rating? Let's plot the graph to see.

IMDB score VS Director Facebook Popularity

Directors with more Facebook popularity generally have higher IMDB scores as we can see the positive trendline.

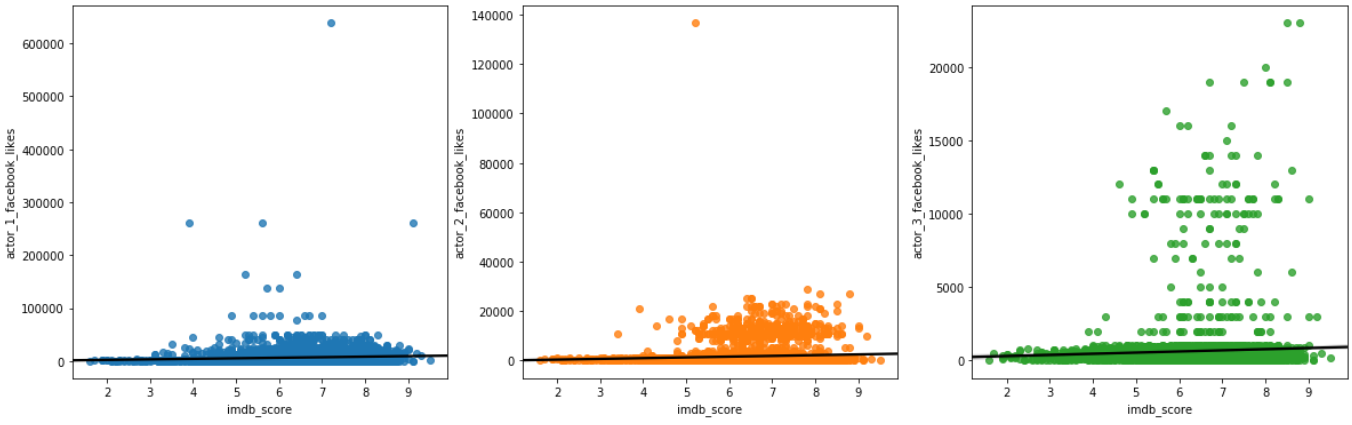
From the plot below, it can be seen that the directors who directed movies of rating higher than 6.0 tend to have more Facebook popularity than the ones who directed movies of rating lower than 6.0.



IMDB score VS top 3 actors/actresses Facebook popularity

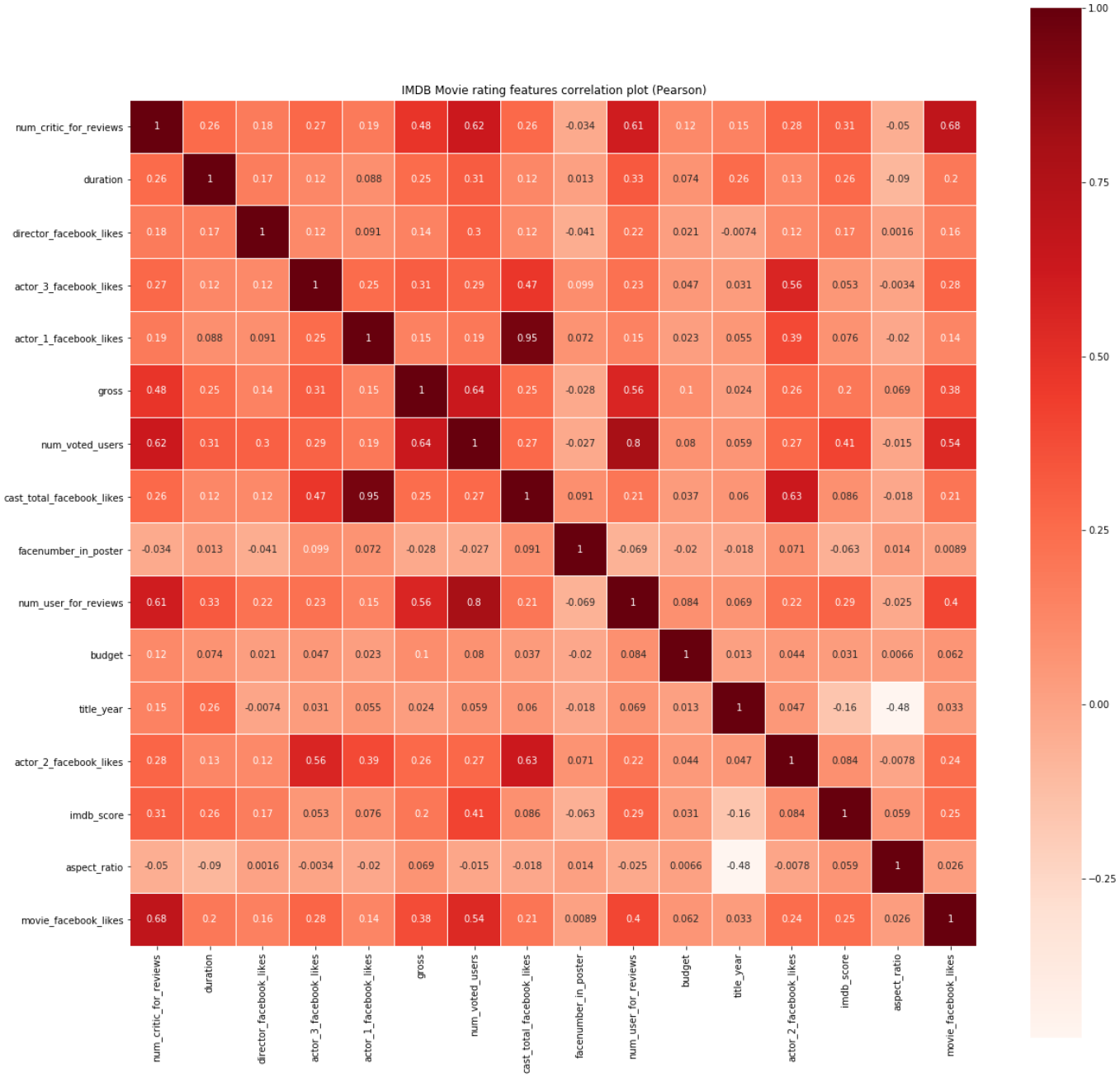
Great actors make a movie great. They are the indispensable parts of movies. How does their Facebook popularity look like?

As we can see the trend line almost flat, leading actors or actresses with high Facebook popularity does not mean that the movie gets great IMDB score.



Feature Correlation Analysis

With 15 continuous variables, I plotted the correlation matrix below.



The matrix gives us some insights that:

- The "cast_total_facebook_likes" has a strong positive correlation with the "actor_1_facebook_likes" and has smaller positive correlation with both "actor_2_facebook_likes" and "actor_3_facebook_likes". This is understandable since the leading actor has the overwhelming Facebook likes as compared to the second and third actors.
- The "movie_facebook_likes" has relatively large correlation with "num_critic_for_reviews", it means that the popularity of a movie in social network can be greatly affected by the critics.

- The "movie_facebook_likes" has relatively high correlation with the "num_voted_users"
- The movie "gross" has strong positive correlation with the "num_voted_users"
- The "num_critic_for_reviews" has high correlation with both "num_voted_users" and "num_user_for_reviews". It means that movies with more attention from users get more critics.
- The "imdb_score" has small and positive correlation with "duration". Long movies tend to have high rating.
- The "imdb_score" has almost no correlation with "budget". Money does not necessarily make a movie great.

4. Data Preprocessing

The columns "aspect_ratio" and "movie_imdb_link" will be dropped because they have little or no effect on the ratings. There are some variables that are not applicable for predicting the Imdb scores, such as "num_voted_users", "gross" because these numbers will be unavailable before a movie is released. So, I drop these columns.

When fitting a multiple linear regression model, removing some variables to reduce multicollinearity is necessary. Since correlation matrix and some regression plots above suggest multicollinearity exists in those numeric variables, I remove the following variables: "cast_total_facebook_likes", "num_critic_for_reviews", and "movie_facebook_likes".

For the purpose of choosing numeric variables to predict the IMDB ratings, I don't take the variable "title_year" since intuitively it does not affect the scores. For text and categorical variables, I specifically choose "genres", "content_rating" for prediction.

There are some numerical variables that I remove rows with missing values because they only account for less than 5% of the sample. Those variables are 'director_facebook_likes', 'actor_3_facebook_likes', 'actor_1_facebook_likes', 'actor_2_facebook_likes'. Missing values from other variables will be replaced by their medians. For text and categorical variables, I replace missing values with None.

After these steps, the dataset is clean. I standardize the features by removing the mean and scaling to unit variance, and then encoding target 'content_rating' labels with value between 0 and number of classes minus 1. Those encoded labels will be binarized in a one-vs-all fashion.

Finally, I split the dataset into training set which contains 80% of the data and the test set containing the rest 20%.

5. Predictive Modeling

Let us build some models. I use Multiple Linear Regression as a baseline model, followed by Random Forest, the Lasso Regression, Ridge Regression, and finally the Support Vector Machine for Regression model.

The metrics used to assess the goodness of fit and the accuracy of our models are R-squared and Mean Square Error (MSE). R-squared is always between 0 and 100%. In general, the higher the R-squared, the better the model fits the data. While MSE represents the difference between the original and predicted values extracted by squared the average difference over the dataset. So MSE is a measure of the quality of an estimator.

Multiple Linear Regression

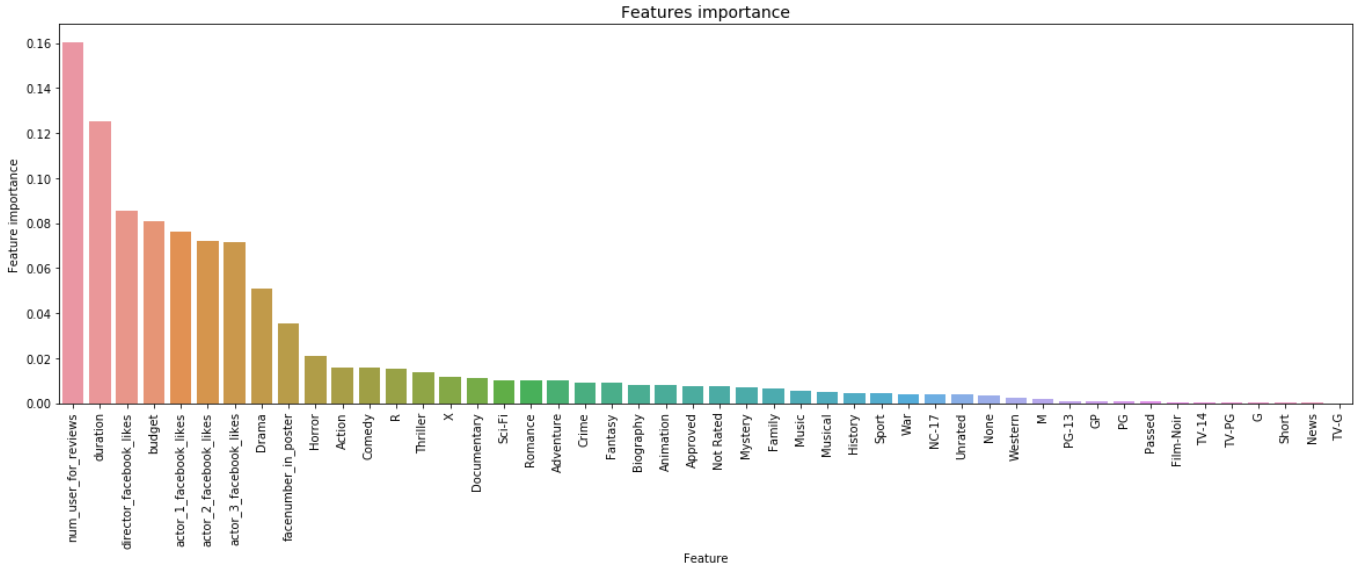
Using Linear Regression from scikit-learn package, this model surprisingly achieves pretty good coefficient of determination R-squared of the prediction: 0.3243, meaning that 32.43% of the variability can be interpreted by the model. The MSE is 0.8338. Some coefficients of this linear model are as follows:

	coefficient
duration	0.21369
director_facebook_likes	0.07567
actor_3_facebook_likes	-0.00931
actor_1_facebook_likes	0.05114
facenumber_in_poster	-0.04524
num_user_for_reviews	0.31823
budget	0.00971
actor_2_facebook_likes	0.00268
Action	-0.22918
Adventure	-0.00141

Random Forest Regression

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning. RandomizedSearchCV is used for the purpose of tuning hyper-parameter, applying 3-fold cross validation on the training set to optimize to get the best hyperparameters.

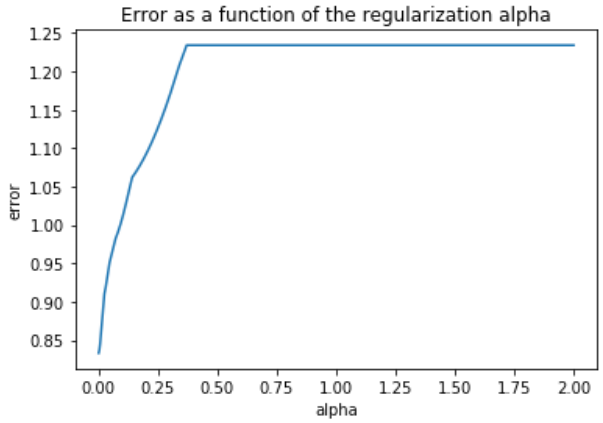
With the best parameters, the random search model has MSE of 0.6709 and multiple R-squared score of 0.4563, meaning that 45.63% of the variability can be explained by this model, while the random forest baseline model only achieves 41.71%. So, the random search model gets improved in accuracy.



As seen in the figure above, visualizing the feature importance of this model reveals that number of users for reviews, duration and director Facebook popularity are the most important features that affect a movie rating, respectively. The next will be budget, leading actor Facebook popularity and supporting actor social media popularity.

The Lasso

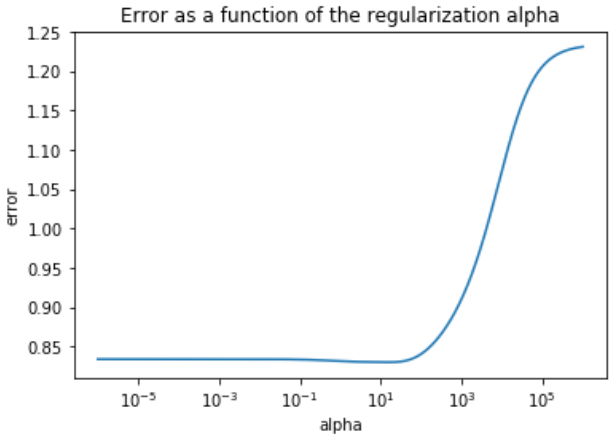
LASSO regression is a variation of linear regression specifically adapted for data that shows heavy multicollinearity. LASSO regression uses L1 regularization, meaning it weights errors at their absolute value. GridSearchCV is used for the purpose of tuning hyper-parameter, applying 5-fold cross validation on the training set to optimize to get the best hyperparameters.



The result is that the Lasso grid search model gives us better R-squared score at 0.3230, meaning that 32.3% of the variability can be explained in this model, while the lasso baseline model without tuning only gets 18.19%. We can visualize the graph of error as a function of the regularization alpha and see that the best alpha which gives us best MSE is near zero (the precise value of the best alpha is 0.0021).

Ridge Regression

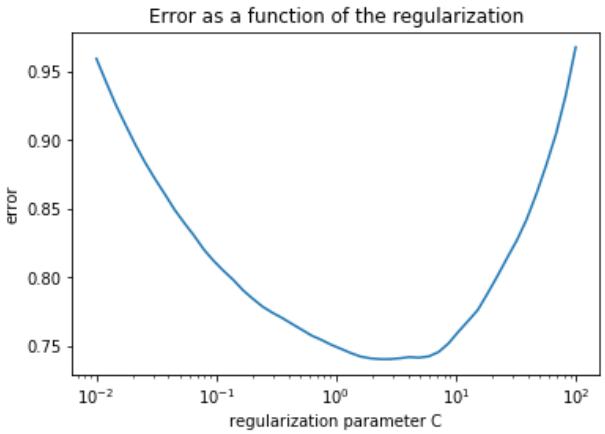
Ridge regression is very similar to LASSO regression in that it applies shrinking. But the largest difference between the two is that Ridge utilizes L2 regularization. Similar to the applying of the Lasso, we use GridSearchCV with 5-fold cross validation to do hyper-parameter tuning.



The Ridge regression grid search model gives us a slightly better R-squared score at 0.3272 and Mean Square Error at 0.8302. We can see the graph of MSE as a function of the regularization with the best alpha for this model is found at 6.507.

Support Vector Machine for Regression

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems. After searching for the best hyperparameters with RandomizedSearchCV, we got the best SVR model with R-squared of 0.4, meaning that 40% of variability can be interpreted by the model. It's better than the baseline model which gets R-squared at 0.3715 and it also gets lower mean square error MSE.



By visualizing the graph of MSE as a function of the regularization C, we can see that the best C value is at around 6 (the precise value of optimal regularization C is 2.811769).

6. Conclusion

From five machine learning algorithms we have conducted, we get the result table as follows:

	Linear Regression	Random Forest	The Lasso	Ridge Regression	Support Vector Machine
R-squared	0.32430	0.45631	0.32299	0.32723	0.40014
Mean Square Error	0.83381	0.67092	0.83543	0.83020	0.74023

- Random Forest Regressor gets the best results with the highest R-squared, meaning that it achieves the highest variability that can be interpreted by the model. Random Forest also gets the lowest Mean Square Error MSE.
- The second-best model here is the Support Vector Machine for regression. Even with this one, we can improve more on its performance if we have more computing power.
- It is suggested that we can achieve better results with collecting more and more data from the IMDB Movie database and trying more non-linear models such as Gradient Boosting Algorithm, XGBoost, or Neural Networks.
- From the results above, Random Forest Regression will be our algorithm of choice.