IMDB Movie Score Prediction

Xiangge Meng, Chenfeng Nie, Xinyao Liu, Yan Wei, Yinwen Shao

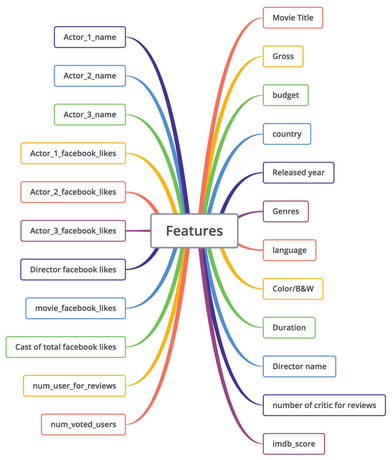
JOHNS HOPKINS UNIVERSITY

Overview

Movies which received high score from viewers always received a high box office receipts. Hence, predict whether a movie is a good movie or not can be a useful tool for investors to decide whether they need to invest lots of money on that film.

Given the voluminous information pertaining to each movie, we conjecture that there exists a relationship between a movie's average user rating and some of its various attributes (which include Director, Actors, Box Office Gross, etc.).

Insight of Dataset



Key Facts:

I. Data collected from IMDB and Facebook (see reference)
II. Training Dataset
  A. 80% of total: 4033
III. Testing Dataset
  A. 20% of total: 1010
IV. Target is IMDB score
V. 23 features in raw dataset
VI. Contain categorical features

Approach

Data **Preprocessing:**

I. Dimension Reduction. Some variables such as "num_voted_users" and "movie_facebook_likes" are not applicable for prediction because these values are unavailable before a movie is released.
II. Since dataset has feature in String format, we used the one-hot and the ASCII encoding to convert strings into numeric values.
III. Normalize all feature on the scale of 0 to 1

Data Visualization:

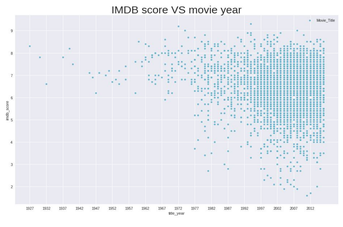
I. Investigating the correlation of each feature

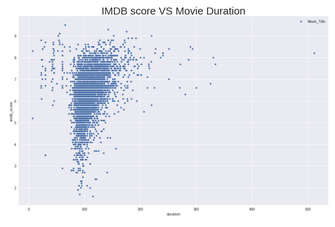
Model Construction:

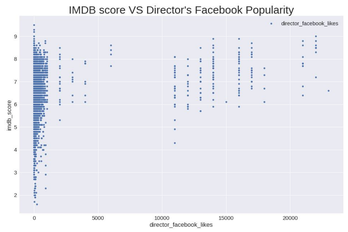
I. We trained 7 different models: Naive Bayes, Random Forest, Decision Tree, AdaBoost, Multi-Layer Perceptron, SVM, KNN.
II. We used cross-validation to determine hyper-parameters of different models, and trained each models 10 times to get average performance.

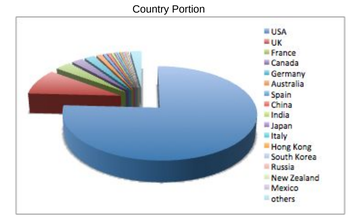
Data Visualization

First, let us has a quick insight on our raw dataset, and following are what we get:

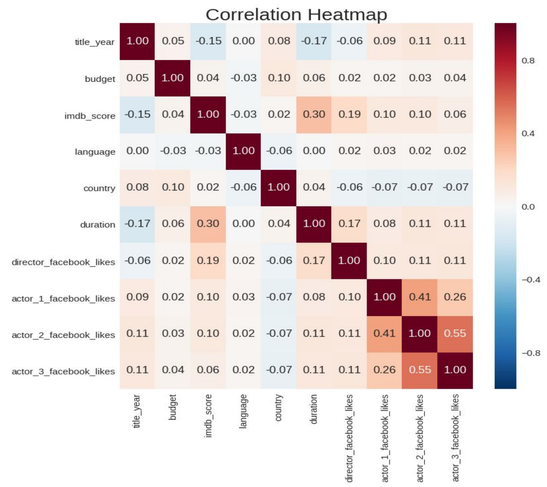








From the above plots, we could have an intuition of what are the features that influenced the IMDB scores most. For example, movies released earlier than 1970s and movies directed by famous directors(have more facebook likes) tend to have a relatively higher scores. However, it only gave us an intuition of what features influenced the IMDB scores more. To be more accurate and objective, we calculated the correlations between the features, and here is the results.

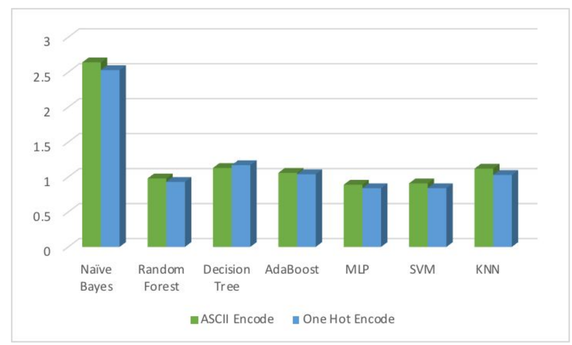


According to the correlation heatmap, we could conclude that IMDB score is most related to the duration, followed by director facebook likes and released year.

Results

For evaluating the result, we use the margin error of the prediction and the real IMDB scores of the movie. And we calculate it by average the absolute value of the difference between prediction and real value.

Margin of Error	Naive Bayes	Random Forest	Decision Tree	Ada Boost	MLP	SVM	KNN
ASCII Encoded	2.64	0.98	1.13	1.06	0.89	0.91	1.12
One Hot Encoded	2.53	0.93	1.17	1.04	0.84	0.84	1.03



Discussion

For data processing, the title of movie and name of director and actors have little correlation to IMDB score, so we removed these features to reduce noise. And since the name of actors and directors have strong correlation with their facebook likes number, we delete them to avoid the overfitting on this features.

For feature encoding, One Hot Encoding has better performance than ASCII Encoding. Since by using multiple dimensions to represent one feature, it can keep similarity of features in same categories.

Different models have similar performance. Naive Bayes model has worst result since it has too strong assumption to feature. SVM and MLP are slightly better since it can fit more complicated model.

References

1. https://nycdatascience.com/blog/student-works/web-scraping/movie-rating-prediction/
2. https://github.com/aksh4y/IMDb-Rating-Prediction
3. www.imdb.com
4. www.facebook.com