IMDB score prediction

Grigor Keropyan

Department of Mathematics and Mechanics at Yerevan State University

Abstract

- Motivation: Producers can use this tool to have successful films and Customers can predict score of the film
- Experiments: Linear Regression, Ridge Regression, Decision Tree, Random Forest, XGBoost, SVM, LightGBM
- Results: One of the best models is Random Forest which achieved 0.44 R^2 score and 0.82 MSE on test set. Best model is LightGBM with $R^2 = 0.45$ and 0.80 MSE on test set

Data

- Data is taken from Kaggle (Few csv files for textual data and jpg files for posters).
- Left the movies which original language is English and have released after 1980.
- In the dataset all NA's are dropped.
- Total 18850 movies. For some models it is splited 70-15-15 as train-validation-test sets and for other 80-20 as train-test. For latter case cross validation is used.

Feature Selection

- Input feature categories: text (synopses), images (posters) and others (runtime, genre, director, actors)
- For posters, using OpenCV extracted number of human faces in posters, means and standard deviations of RGB and HSB
- For crew and cast, extracted top three actors and director

Data	Type	Dimension	Example
Actors	Categorical	1671	Tom Hanks
poster	Numerical	13	Number of faces = 0
Director	Categorical	491	Howard Deutch
Runtime	Numerical	1	121 (minutes)
Genre	Categorical	23	Documentation
Synopses	Categorical	3712	Once

Models

- Linear Regression: $\theta^* = argmin_\theta \frac{1}{n}\Sigma_{i=1}^{n}(\theta^T x^{(i)} - y^{(i)})^2$
- Linear Regression: $\theta^* = argmin_\theta \frac{1}{n}\Sigma_{i=1}^{n}(\theta^T x^{(i)} - y^{(i)})^2 + \lambda||\theta||^2$
- XGBoost: $\mathcal{L}(\phi) = \Sigma_i l(\hat{y}_i, y_i) + \Sigma_k \Omega(f_k)$, where $\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2$
- Decision Trees and Random Forest: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- LightGBM: our best model:
  LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient and have a few advantages.

A former Roman General sets out to exact vengeance against the corrupt emperor who murdered his family and sent him into slavery.

- Director: Ridley Scott
- Actors: Russell Crowe, Joaquin Phoenix, Connie Nielsen
- Genre: Action, Adventure, Drama
- Runtime: 155 min

LightGBM

Results

Method	Train MSE	Test MSE	Train R^2	Test R^2
Linear Regression	0.4366	1.3918	0.7043	0.0668
Ridge Regression	0.5333	0.9824	0.6389	0.3413
Decision Tree	0.8440	0.9153	0.4285	0.3863
Random Forest	0.8277	0.8290	0.2543	0.4441
XGBoost	0.5160	0.8375	0.6506	0.4384
SVR	0.5735	0.9294	0.6117	0.3768
LightGBM	0.6488	0.8067	0.5592	0.4513

- The first six models were trained on 13194 training samples, while hyperparameters were selected using 2828 validation samples and evaluated on 2828 test samples. However, in the case of LightGBM dataset is splited into train and test set with 15080, 3770 samples in each.
- LightGBM is the best model among all of the models.

Comparison with previous work

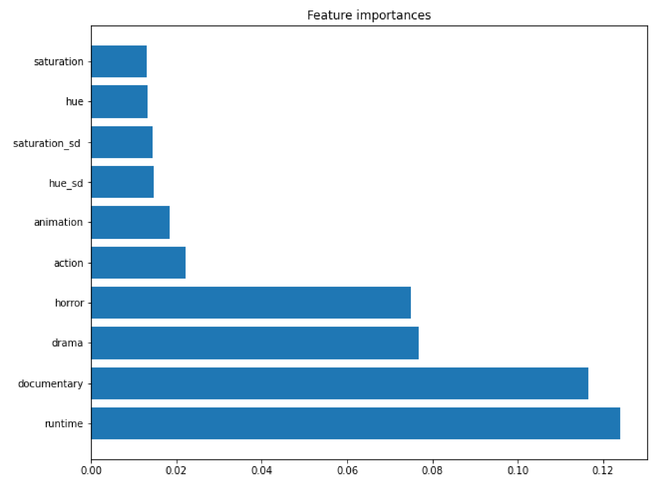
In the Table 1 the model is compared with the Stanford students paper [1]. In this models out Random Forest achieved better accuracy in the sense of MSE and R^2 compared with [1]. Moreover, the LIghtGBM model achieved 0.8 MSE and 0.45 R^2 which is better than our Random Forest model.

Table 1:Comparison with [1]

Method	our MSE	[1] MSE	our R^2	[1] R^2
Linear Regression	1.3918	0.9302	0.0668	0.3745
Ridge Regression	0.9824	0.8775	0.3413	0.4099
Decision Tree	0.9153	0.8959	0.3863	0.3975
Random Forest	0.8290	0.8546	0.4441	0.4253
SVR	0.9294	0.8765	0.3768	0.4109

Feature Importance

Feature importance in the Random Forest model is shown in the following figure. From the bar plot we can conclude that runtime and specific genre is important for the IMDB score.



References

[1] Yichen Yang et al., *"Predicting Movie Ratings with Multimodal Data"*, http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26260680.pdf.