IMBD Movie Rating – Linear Regression Project

Abstract

The goal of this project was to create a model that can evaluate and predict a rating for a new movie. Such a model is meant to be used by a marketing department at a film studio to assess the likely public perception of a film that hasn't been released, using available information about the film, such as its length, as well as estimates that could be extrapolated based on previous success of similar movies, director, etc.

Data

The two main data sources used were imbd.com, from which I scraped the majority of the data used in the analysis, as well as a database on Kaggle: https://www.kaggle.com/danielgrijalvas/movies. After data cleaning, I arrived at 2112 rows and 28 features. Final model contains 13 features, 5 of which are numeric.

Feature engineering:

1. Converting categorical features to binary dummy variables
2. Created an interaction variable between average rating for director, writer and star, resulting in a unique rating for each combo, based on averages for movies prior to the particular data point.
3. Created a log transformation of the runtime variable to address its distribution
4. Created MPAA rating/genre variables to address the collinearity issue.

Model Evaluation and Selection

I used an 80/20 split between train and test, and used KFold cross validation to evaluation on training portion.

Model progress:

Base model:

Reg RMSE:0.4939452427608857
Reg MAE:0.381660842544367

Ridge RMSE:0.4940531645321911
Ridge MAE:0.3824829488619353

Lasso RMSE:0.4957045395695033
Lasso MAE:0.3819727203085793

R2 training:
Reg mean r^2: 0.641 +- 0.017
Ridge mean r^2: 0.641 +- 0.017
Lasso mean r^2: 0.639 +- 0.017

R2 test:
Teg test: 0.6100502839107935
Ridge reg test 0.5965136188333309
Lasso reg test 0.5998301569323061

MODEL 2: (logged runtime, added combos for MRAA cert + genres)

Reg RMSE:0.47662003656105467
Reg MAE:0.3863691673750603

Ridge RMSE:0.4768093341562867
Ridge MAE:0.3868479377969975

Lasso RMSE:0.4790793106025752
Lasso MAE:0.38880874580992714

R2 train:
Reg mean r^2: 0.633 +- 0.032
Ridge mean r^2: 0.633 +- 0.032
Lasso mean r^2: 0.631 +- 0.030

R2 test:
Simple reg test: 0.6160213761258275
Ridge reg test 0.5966860846747223
Lasso reg test 0.5934150833807972

MODEL 3 (interaction term + reducing dimensionality)

Reg RMSE:0.3719026287009993
Reg MAE:0.2868996664064177

Ridge RMSE:0.37256452290741143
Ridge MAE:0.28760142600016175

Lasso RMSE:0.3787135942683239
Lasso MAE:0.2903289590305702

R2 train:
Reg mean r^2: 0.783 +- 0.012
Ridge mean r^2: 0.783 +- 0.012
Lasso mean r^2: 0.781 +- 0.011

R2 test:
Reg test: 0.7864868990598985
Ridge reg test 0.7845498534763847
Lasso reg test 0.7863469887278335

Tools:

BeautifulSoup and Requests for webscaping
Numpy and Pandas for data manipulation
Scikit-learn and statsmodels for modeling
Matplotlib and Seaborn for plotting

Communication

Slides in a pdf that will be presented during the presentation.