Rotten Tomatoes Ratings Prediction Using Linear Regression

Objective

Predict **Rotten Tomatoes** critics ratings (AKA **Tomatometer**) using Linear Regression

Data Sources

- Rotten Tomatoes for past ratings
- Complete List of Oscar Nominees and Winners for awards and nominations

Features

It is impractical to use categorical data, such as plot synopsis, actor or director names, with linear regression because they would result into thousands of indicator columns. Consequently they have been replaced by aggregating proxy features as follows:

Proxy for	**Feature**	**Rationale/Intuition**
Content	Genre	Rough approximation for plot summary
	MPAA ratings	Restricted movies may have more freedom of artistic expression
	Runtime	Critically acclaimed movies may be longer
Potential for commercial success	Size of the reported cast	RT reports only the cast of certain note (celebrities), so the size of the cast may have a positive correlation with the box office success
	Number of movies to-date[1] for the cast, director and writers	Successful artists are more likely to be cast again
Potential for critical acclaim	Number of awards, wins and nominations to-date[2] for the cast, director and writers	Critically acclaimed artists are more likely to be in critically acclaimed movies
	Release month	Critically acclaimed movies tend to be released before awards seasons.

[1] Each of the counts was stored in a time-series and the value used for a movie was the most recent one prior to the release date.

[2] Ditto

Input Data Set

`data/rt_movies.csv.gz`, 15857 rows

Target

Rotten Tomatoes critics rating, normalized into 0..1 interval.

Exploratory Data Analysis And Feature Selection

Initial data visualization revealed no significant[3] cross-correlation between any of the features and the target so it was not possible to pick one significant feature for baselining. Instead the following tests were performed:

1. OLS with all features
2. OLS with the features with a p-value < threshold (.05 or 0.02)
3. OLS with the features with absolute of the target cross-correlation in the 50 percentile, i.e. only the highly correlated features

The R^2 scores, both raw and adjusted, were compared for the above and since they were very similar (0.87 down to 0.85, but that's expected due to fewer features), the choice was to go with #3 since the least number of features signifies a simpler model which is generally better.

Model Selection

The data set was partitioned 80/20 into training and test subsets.

The training set was used for 5-fold cross-validation with the following models: LR, Ridge and Lasso, without and with 2-degree polynomials; regularization was applied with α in the [10^{-3}, 10^{6}] range with 100 steps.

Each trained model was then scored using the test subset and the one with highest score was selected: Lasso(α=0.001526).

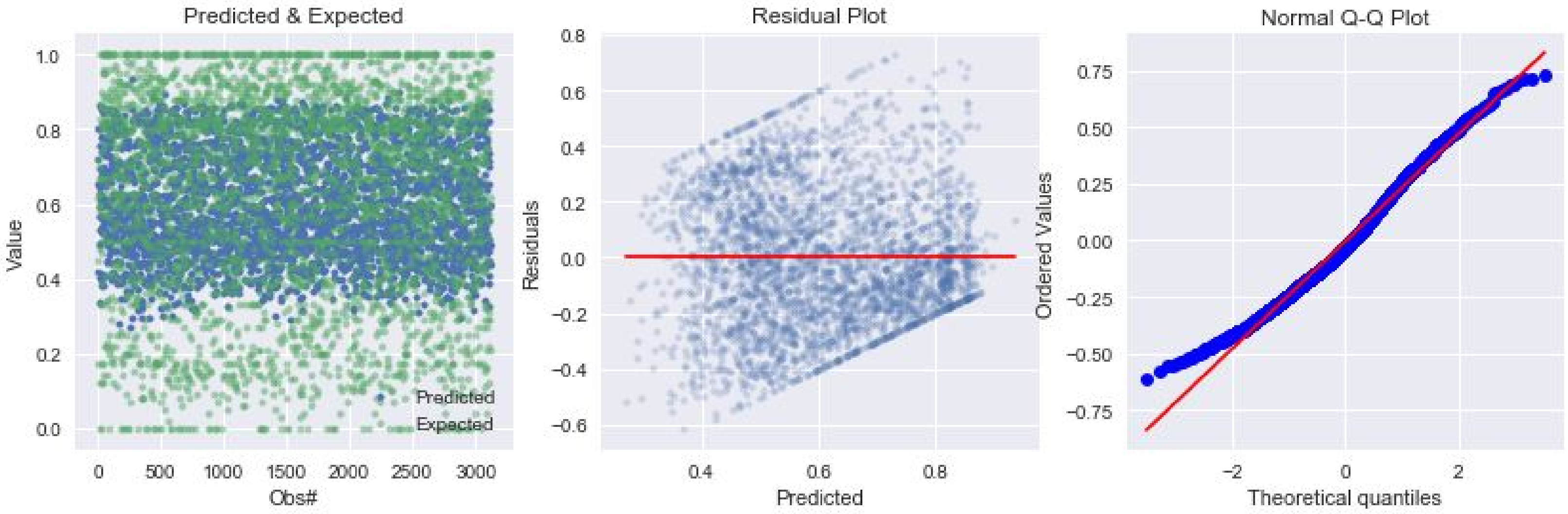
The selected model was then re-trained with the entire training set and the final score was determined for it.

Final Results

Model	Lasso(α=0.001526)
Test Set R^2	0.253
Test Set Mean	0.611
Test Set RMSE	0.239 (39.16% of mean)

[3] Abs value >= .5

Diagnostic Plots



Conclusions

The low R^2 score and the diagnostic plots indicate that there is not enough relevance in the features for accurate predictions with Linear Regression.

A different ML algorithm is required such one that supports NLP in order to use more specific features like plot synopsis and artist names.