National Institute of Technology
Karnataka, Surathkal

PROJECT REPORT

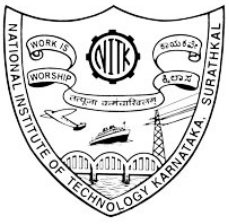
Data Mining and Data Warehousing

Movie Popularity Analysis

By:
Isha Tarte
Senior Year
Computer Science
14CO217

Aparna R Joshi
Senior Year
Computer Science
14CO204

November 15, 2017



1 Summary

The Internet Movie Database (IMDB) is one of the largest online resources for general movie information combined with a forum in which users can rate movies. There is no universal way to claim the goodness of movies. Many people rely on critics to gauge the quality of a film, while others use their instincts. But it takes the time to obtain a reasonable amount of critics review after a movie is released. And human instinct sometimes is unreliable. We investigate the extent to which a movie's average rating can be predicted after learning the relationship between the rating and a movie's various attributes from a training set. We use Ridge and Lasso regression models to predict the rating of a movie before it is released and obtain a mean squared error (MSE) of 0.9332368 and 0.9368678 respectively for the two models while we obtain a residual standard error of 1.022 using multi linear regression.

2 Approach used

2.1 Data mining technique used

2.1.1 Pre-Processing

1. With the use of the Python library scrapy, we scraped the titles of around 5000 movies.

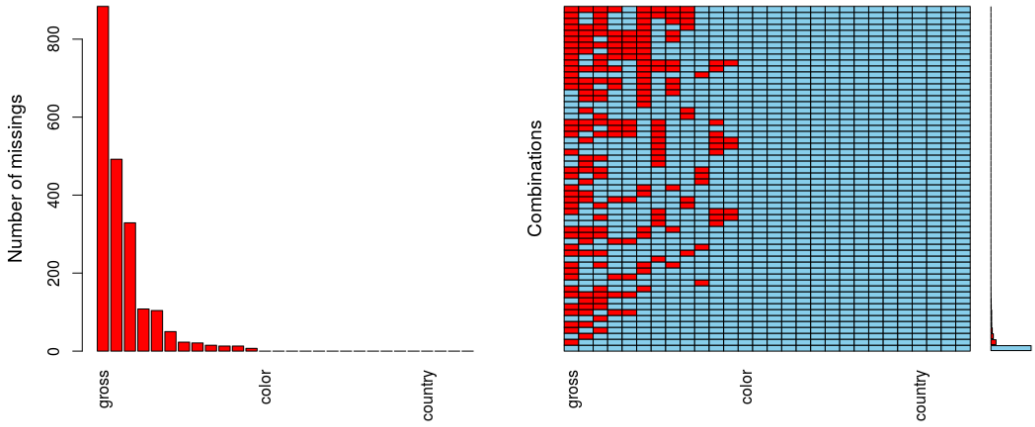
2. Then, we stored these titles into a JSON file.

3. We search for the movies using these titles in IMDB to get the movie links.

4. With the requests module provided by Python, we retrieve the data from the movie pages using the links.

5. We parse through the integrated data, clean it, and store it in a CSV file.

We investigate the missingness of the various attributes, sorting them in descending order. We choose the "good attributes" for training the models based on factors like: timely availability and relevance of attributes, correlation, and missing values.



2.1.2 Models

We use ridge and lasso regression model to predict the movie rating.

Linear Regression We select few good attributes for linear regression. We eliminate cast_total_facebook_likes since it is highly correlated with actor_1_facebook_likes, and attributes as movie_facebook_likes and num_voted_users since these will not be available when a movie is released. Also we ignore the categorical data. The residual error obtained on using Linear Model is 1.022.

Ridge Regression: Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

Lasso Regression Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

The correlation matrix computed revealed that multicollinearity existed between the 15 continuous variables. So, ridge and lasso models were used.

2.1.3 Training and testing the model

We divided the dataset into 70-30 for training and testing. Nine attributes were chosen for the purpose of training. We also performed 10 fold cross validation in order to choose the best lambda over a slew of values. The mean square error associated with this best value of lambda was calculated.

2.2 Data set used

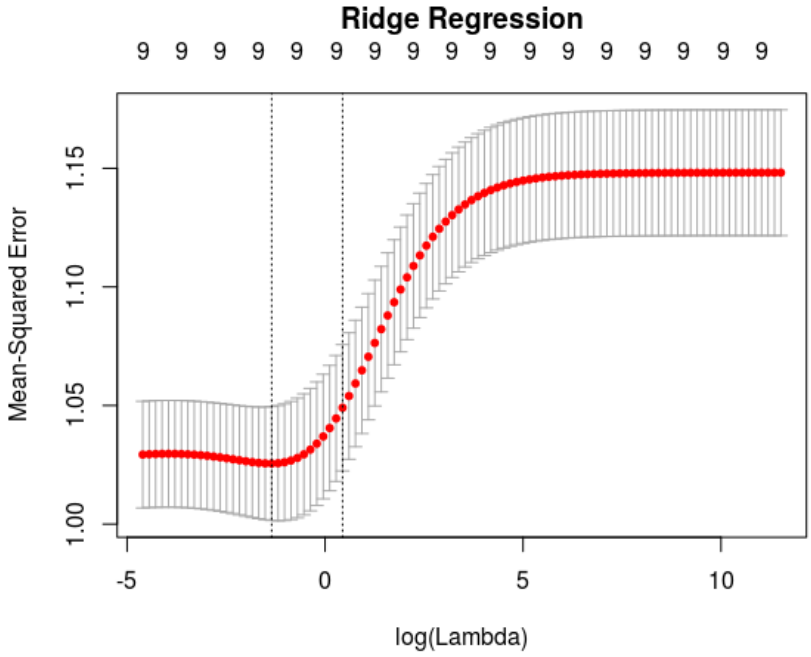
We use an aggregated dataset consisting of 5000 movies comprised of 25 attributes. The dataset was obtained by scraping the movie titles followed by acquiring characteristic information about them from IMDB. The attributes scraped include movie title, director name, cast attributes, facebook likes for actors and the director, genres etc. The scraped attributes also includes the IMDB rating for each of the movies, and we use this data to train our model and later, predict ratings for test movies. The features of the dataset are as follows:

1. director_facebook_likes
2. actor_3_facebook_likes
3. actor_1_facebook_likes
4. gross
5. cast_total_facebook_likes
6. facenumber_in_poster
7. content_rating
8. budget
9. actor_2_facebook_likes
10. imdb_score

3 Result and Discussions

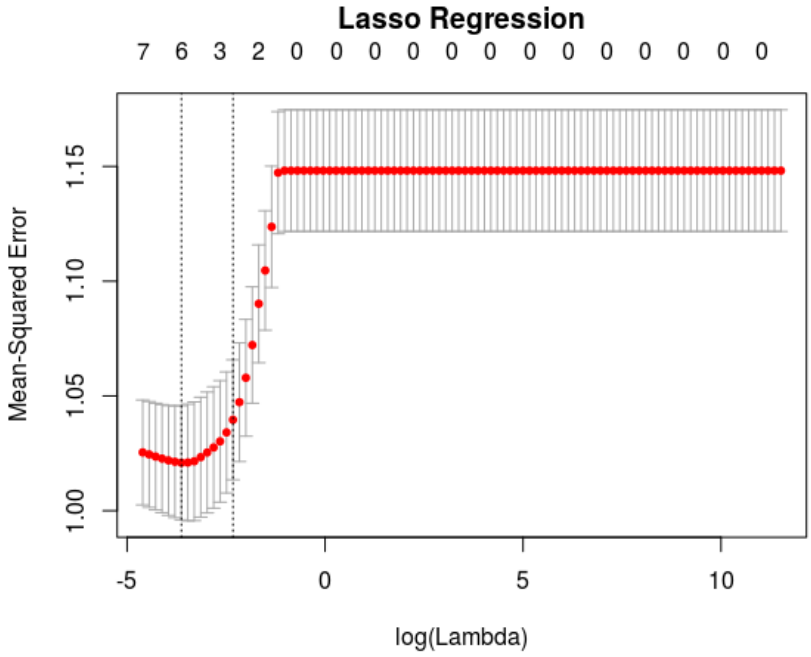
We obtain an MSE value of 0.9332368 with Ridge regression. The coefficients of the various attributes obtained are as follow,

Attribute	Coefficient
intercept	$6.289819e+00$
director_facebook_likes	$3.984165e-05$
cast_total_facebook_likes	$2.797322e-07$
actor_1_facebook_likes	$2.332429e-06$
actor_2_facebook_likes	$5.400163e-06$
actor_3_facebook_likes	$-2.143775e-05$
movie_facebook_likes	$8.854214e-06$
facenumber_in_poster	$-2.526755e-02$
gross	$1.543812e-09$
budget	$2.175598e-11$



We obtain an MSE value of 0.9368678 with Lasso regression. The coefficients of the various attributes obtained are as follow,

Attribute	Coefficient
intercept	$6.285633e+00$
director_facebook_likes	$4.057852e-05$
cast_total_facebook_likes	$-$
actor_1_facebook_likes	$1.424556e-06$
actor_2_facebook_likes	$-$
actor_3_facebook_likes	$-4.769225e-06$
movie_facebook_likes	$9.899551e-06$
facenumber_in_poster	$-1.890144e-022$
gross	$1.434691e-09$
budget	$-$



4 Conclusion

Obtaining an integrated dataset from internet sources, we were able to identify the most influential attributes for movie popularity analysis. We divide our scraped data of about 5000 movies with relevant features into train-

ing and testing sets in a ratio of 70:30. We initially performed multivariate linear regression on the data and obtained an error of 1.022. Following that, we performed Ridge and Lasso regression on the data which resulted in considerably lesser error. This could be attributed to the fact that the attributes "cast_total_facebook_likes" and "actor_1_facebook_likes" were found to be highly correlated (the attribute cast_total_facebook_likes will likely rise with an increase in the value of the attribute actor_1_facebook_likes). This is termed as "Multicollinearity". Models as Ridge and Lasso can be used when features selected tend to be highly correlated with one another, as linear regression or Ordinary Least Squares might result in high variance.

5 References

- Armstrong, Nick, and Kevin Yoon. Movie rating prediction. Technical Report, Carnegie Mellon University, 2008.
- Augustine, Achal, and Manas Pathak. User rating prediction for movies. Technical report, University of Texas at Austin, 2008.