Abstract

A Publishing company has received various manuscripts and wants to select which manuscripts to publish. To make a decision it needs a model that predicts the rating of the book.. I worked with numerical and categorical data available on goodreads website leveraging feature engineering and linear regression model to predict book rating with an MAE of 0.1130(Ratings scale 0-5). The book rating was highly correlated to the square of author rating. The final model developed using ridge regression was used.

Data

36 Pages,1800 records with 12 data points were scraped with help of BeautifulSoup to gather book rating and author features. Error handling was built in the program to compensate for missing data. EDA was performed on scraped data to cater for missing or null values. There was 1 Categorical Feature and 7 Numerical features.

Algorithm

Feature Engineering:

1. Creation of additional variables were :
    - Percentage author presence in top 1800 records,
    - Square of author rating
    - Interaction between no of rating and reviews
2. Converting categorical variable to binary dummy variable

Models

All 3 models namely Linear Regression,Lasso Model and Ridge were explored during the course of feature enginerring. Initially with large no of features and correlated data Lasso Model was used which helped to minimize the number of features. When fewer features which were not correlated were present, simple regression was opted to make a conclusion. For conclusion when few features and high correlation existed Ridge model was used with 3 features and 1 interaction variable. Major variance in target(Book Rating) was explained by 1 feature i.e. Author Average Rating. The other significant features did not explain major variance in the target. **Model Building Code still needs to be cleaned and organized, due to time crunch I am not able to complete it and plan to do after submitting all coursework for the module.

Model Evaluation and Metrics

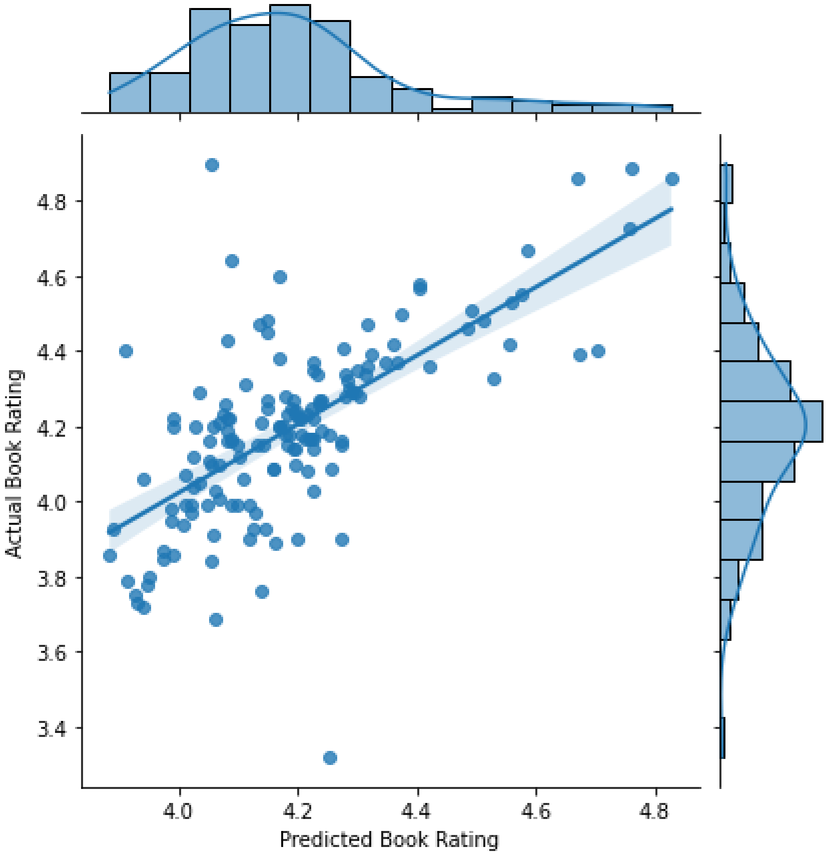
Entire dataset of 1500 records was split into 64/16/20 Training/Validation/Test Data Set and R2 and MAE was used to finalize a better performing model. Lasso and Stats Model P values were used to decide which features to drop. K Fold cross validation was performed while evaluating performance in Lasso and Ridge Regression Models.

Tools

Numpy and Panda for manipulating the data
Scikit-learn, Lasso for modeling
Matplotlib and seaborn

Communication

Following is a chart describing predicted Book Rating vs Actual Book Rating. R2 = 0.4914, MAE = .1130


Actual Book Rating
Predicted Book Rating