Predicting Book Ratings with User-generated Tags

James Lee
University of Virginia

Abstract

Predicting the rating that a reader will give a book can be used to help people pick their next book to read. In this project, we utilize the user-generated tags placed on books to try and predict the rating a user would give a book.

Background

Goodreads is a massive online community where people can review, rate, and tag books. The book ratings go from 1 to 5, with 5 being the highest rating. Tags are arbitrary strings that users say is relevant to a certain book. Because users can give arbitrary tags to any book, one must take care to account for typos and nonsensical tags such as '-a-0–' or 'fatnasy'. However, in the aggregate, we believe that the user-generated tags could be used to quantify the characteristics of a book.
Specifically, we hypothesize that a book can be described as a mixture of several base genres. Then, because people often have specific tastes in book genres, this mixture can be used to predict the rating a user would give the book.

Objectives

First, we want to transform the messy user-generated tags into useful information about each book. We want to see if the tags can be used to describe the overall genre of a book. Once we have a "genre distribution" for every book, we can create models to predict the rating a user would give a specific book.

Goodbooks 10k Dataset

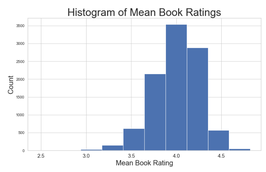
The goodbooks data set contains user ratings for 10,000 popular books from a site similar to goodreads. Basic identifying information such as author, title, and publishing date is included. Furthermore, on the website, users can tag a book with an arbitrary string to indicate topics or ideas that the book is related to. This data contains the list of all possible tags, as well as the top 100 most popular tags for each book.

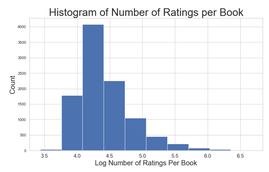
Data Highlights

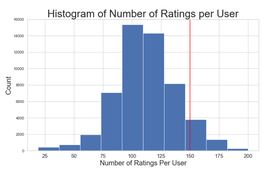
- 10,000 books
- 53,000 users
- 6,000,000 ratings
- 34,252 unique tags
- 100 tags per book
- 19-200 ratings per user
- messy tag strings

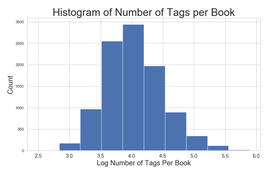
Data Processing

The tags had to be extensively cleaned in order to be useful. After filtering out unnecessary tags and combining variations such as 'sci-fi' and 'science-fiction', the top 10 most popular tags were used as the genres. Also, we manually selected the top genres in the non-fiction category and added them to the genre list. Tags included: biography, business, childrens, classics, fantasy, historical-fiction, humor, memoir, mystery, non-fiction, paranormal, romance, science, science-fiction, young-adult.



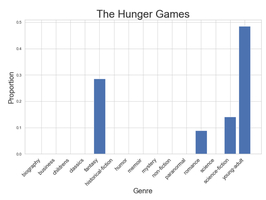


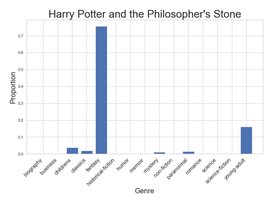


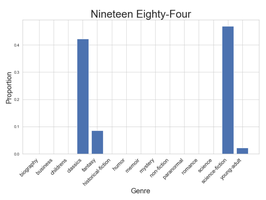


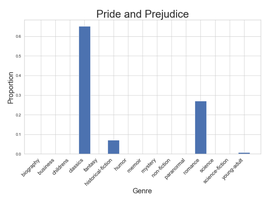
Genre Distribution

With the base genres, we can now calculate the genre distribution for a given book. As shown in the selected results below, the distributions seem reasonable for anyone familiar with the books.

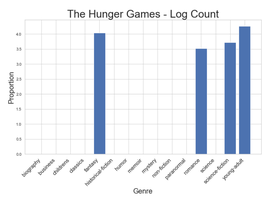


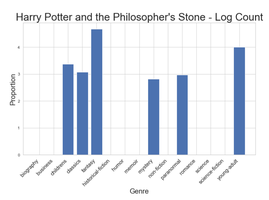


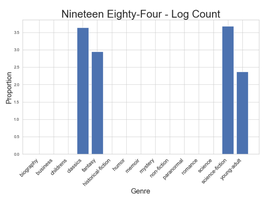


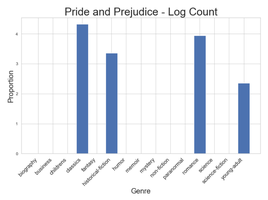


Using the genre proportions as the predictors caused some issues with collinearity and resulted in poor predictive performance. Instead, we use the log counts of the genres as the predictor variables as seen in the plots below.









Model

The biggest roadblock was how to create a single model to describe the relationship between the user and book genres and the predicted rating. It is reasonable to think that users will have different preferences for the genres so treating users as a categorical variable in a linear regression without interactions would not make sense. Hence, we decided to see how well a single model per individual could perform as well as a small polynomial regression model that contained the user-genre interactions.

1. Linear regression model per person
2. Polynomial regression model for n users

1. Linear Regression

$$r_{ui} = \beta_{i,0} + \beta_{i,1}x_{\text{bio}} + \beta_{i,2}x_{\text{bus}} + \cdots + \beta_{i,15}x_{\text{ya}}$$

2. Polynomial Regression

$$r_{ui} = \beta_0 + \beta_1 x_{\text{user1}} + \beta_2 x_{\text{user2}} + \cdots + \beta_{n+1}x_{\text{bio}} + \beta_{n+2}x_{\text{bus}} + \cdots + \beta_{n+15}x_{\text{ya}} + \cdots + \sum_i \sum_j \beta_{ij} x_{\text{user}_i} x_{\text{genre}_j}$$

Single User Model

For the linear regression model of the user with the most ratings, we see that the important predictors are:

OLS:

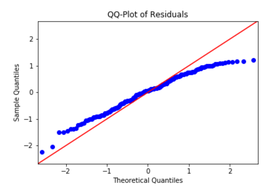
- fantasy: 0.12
- historical-fiction: -0.12
- paranormal: -0.18

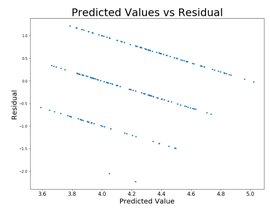
Lasso:

- childrens: -0.01
- fantasy: 0.07
- historical-fiction: -0.06
- humor: 0.03
- paranormal: -0.11
- science-fiction: 0.03

However, as $R^2 = 0.168$ and RMSE$= 0.69$, the fit is not great. For reference, the base model using the user's mean as a predictor has an RMSE of 0.75. With 5-fold cross validation to estimate the test error:

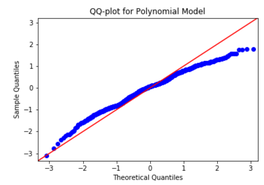
Model	RMSE	R^2
Linear	0.78	0.14
Lasso	0.77	0.11
Constant	0.76	0.07

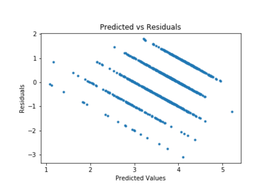




Polynomial Model

We run a degree 2 polynomial model for the top 5 users with the most ratings by adding the user-genre interactions to the linear model from above. The number of predictors jump from 15 to 96 as a result with about 1000 observations. The resulting linear regression model was unstable, but using Lasso regression we see that a total of 28 predictors had non-zero coefficients with an test RMSE of 0.97.





Conclusion

Based on our experiments, although the models were interesting and could possibly be used to determine how genres can effect a users predicted rating, they seemed to have limited predictive power. An interesting point was that none of the models performed particular better than the baseline model.

LaTeX TikZposter