In-depth Analysis

At first, I was thinking about using linear regression for my Goodreads dataset. Using book’s total number of reviews and total number of ratings and their genres to predict how popular a book is. I would use bootstrap inferential statistics to resample the sample data we have from Goodreads and run 10,000 re-sampling to find the correlation between the total number of review and total number of ratings for each genre to predict its average.

Later, I realized that it does not really make sense. The average of a book or a genre is unpredictable. Many factors could factor it such as how long the book has been published, how good its advertisement is, and how many people were willing to spend their time to rate the book on Goodreads and the state of mood when the person gives the rating. Also, even if I were able to use an algorithm to predict the average rating of a book, it still does not make sense because knowing the average rating of a book does not create much values for publishers or readers. If I were to proceed with my previous idea, I would be using multi linear regression models for my machine learning technique.

Later, I stumble upon the idea of building a recommendation system. There are many ways of building a book recommender. One is based on the most popular books overall, which would be base on the most popular votes by users. However, this kind of recommender does not provide customization for each users. Everyone would get the same recommendation which would not be fun anymore.

A more complex ones is using K-nearest neighbor. The recommender would give out recommendations base on user’s input. And base on the input user provide, the system would find the most similar book base on the book’s average rating, genre, total rating and more to return the recommendation list. This model is better than previous one but the model would be bias toward books that are in the same genre, publication date because books with largest total rating tend to be publishing earlier than other books. So this model is not what I want.

Then my mentor suggested the collaborative filtering technique. The collaborative filtering technique can be user base, content base or item base and more. I will be using user base. The recommender would first find the similar users based on the rating the person gives to the books he or she has rated and the ratings those similar users have rated their books, then the recommender would use the ratings that other users have provided to predict whether this person would like or dislike those books that he or she hasn’t read yet. This recommender might suggest a book or a genre of book the person has not been exposed to. This recommendation system would create more values for both users and publishers than previous two versions.

Because this is a recommender, the machine learning in-depth analysis does not provide much information overall. It provides useful information when an user would like to seek out books that they would like to read next. It is not like linear regression model or multi - regression model that tells us how each addition on one element of a house, ex bed room or bathroom, the total housing price would go up by five thousand dollars. Or a clustering machine learning model that tells us the distinctions belong one group from another. The recommender only provides information when we provides the input.