**Predicting the New York Times Best Sellers List**

*Lydia Skrabonja*

The NYT Book API

Initial research into the New York Times Books API was heartening. It seemed as if it was well constructed and would be fairly easy to work with. There are multiple lists to choose from so I chose what I believed would be most inclusive, “Combined Print and E-Book Fiction”. Unfortunately, my initial assessment would prove to be more wishful thinking. Through some strange hiccups in my data I found a few odd things with their API set up.

The first strange feature of the API was how it dealt with books that ranked below 16. All of those books indicated that this was their first week on the list, even when that wasn’t true. In order to counter the effects this may have on my predictions, I excluded all the entries after the fifteenth ranked book.

The second anomaly in the API I discovered had to do with the rank the book held the previous week. In the JSON object the API returned, it listed the rank the book held last week. If the book had been ranked 16th or below, it was listed as if it had not appeared on the list the week before. Strangely, the time the book spent on the bottom of the list factored into its “Weeks on the List” count, but not the rank it held. This made the historical data more valuable than that field, and therefore I no longer dealt with it.

The third strange thing about the API was the historical data itself. If a book entry indicated it was on the list for 32 weeks, sometimes it would only return 4 results pertaining to that list. As a proxy for speed, I chose to use the historical data, as I assumed after a certain number of weeks, the data would not matter as much. A more precise model could possibly be gleaned from forgoing the use of the historical data and building a data set from repeated iterations of getting the list from various weeks. Since I wanted my model to be workable over multiple weeks, I thought it was an appropriate thing to cut out.

I was initially interested in the column that was intended to indicate bulk ordering. According to the site, some books are bulk ordered and that was indicated by a dagger on the pages or website. I believed that would be a telling measure of whether or not a book was going to last on the list. As it turned out, none of the books ever indicated they were bulk ordered. I threw that column out.

The last struggle with the API came in the form of encoding. The API call could not handle various symbols often found in books, so I stepped through and found their HTML equivalent to match them up.

Initial Research

Looking around the data revealed some interesting initial tidbits. For example, about a third of all books on my selected books are making their first appearance. I find predicting books that are about to appear on the list to be out of the scope of the project, so I eliminated those results.

Another interesting part of the data was the number of weeks displayed on the list. On a given list, the median number of weeks books have been on it ranks in as 3. Additionally, the mean number of weeks is 9.7. This, coupled with the previous mention of a third of the books being first timers, suggests that most books appear to have been on the list for a month or less, and there are a small handful of books with staying power. These books have been on for multiple months and swing the mean towards them.

First Models

I worked two models a bit at the same time. I wanted one continuous model that could later be rounded to rank numbers and one categorical model. I decided to start with simple linear regression for the continuous model. For the linear regression, the initial tests are promising. Mostly cleaned data is returning an adjusted R-Squared of 0.70. I am currently in the process of running an MSE on it, after running against the problem of string variables in the more malleable forms of linear regression.

For the categorical model, I have decided to start with a Random Forest, given its reputation for being higher than most other models in general scenarios. I have run into issues with the string variables. I am looking into preprocessing solutions as well as alternative categorical options and hope to have more to offer soon.

Next Steps

The immediate next steps for me is cleaning up and improving the two models I am currently working on. I have not yet folded in the fact that books fall off the list. All my current training data assumes that the books come out as being something, so I have to find a way to build that in as well. Besides that, I would like to try at least 2 more models before being confident in presenting my findings. After the models are completed I plan on making slides and visuals so as to give an interesting presentation. I was thinking a heat map of the list if it makes sense, where my model tends to put things.