Predicting New York Times Bestsellers

A Meta-study for Understanding Modern Readership and Publishing Industry

By Monica Anuforo, Cherie Xu, Alice Yang

Stanford University

Introduction

The New York Times Bestsellers List is an influential listing for publications and significantly dictates the repertoire of work that many people are exposed to. We want to identify salient features of a NYT-Bestseller-viable book in particular genre, given the zeitgeist and current events. We hope to use our results to better understand consumer preferences for literature and non-fiction. As the market for books becomes progressively more commerc -ialized, we want to see if the criteria for a popular book has shi -fted to reflect a change in communal taste. It would be interest-ing to observe if there is any inherent arbitrariness in the comm-ercial viability of books.

Figure 1: How a NYT Bestseller is Made

Problem and Purpose

The goal of our project is to predict the sale statistics of a book = using key characteristics that we identify based on recurring trends and, in turn, determine future New York Times Bestseller. Beyond the predictive component of the task, there will be ano -ther valence to this project. The predictive component will be the validation of the actual functionality of the program i.e. gen -erating salient features of a commercially viable book. What we are ultimately interested in and want to also examine are the salient features and their implications of modern readership and societal trend of reading.

Input and Output Behaviors

Data Acquisition

Relevant data were scraped from multiple sources to allow the extraction of comprehensive features. We obtained data from Open Library which is a public API affiliated with Archive.org and the library of congress. From the OL dataset, we extracted information for 10,000 of randomly chosen book to serve as the non-examples of NYT bestsellers. We also scraped the available data from New York Times' book API which gave us information of all NYT bestsellers starting from 2008 (total of 4620). Between the two sources, we were able to obtain basic description for the books such as ISBN, author, short descriptions, genre. We also obtained reviews and ratings from GoodRead for each selected novel.

Feature List

- Title
- Page number
- Author
- publication year
- publication month
- Descriptions NLP
- Average rating
- Rating counts
- Text_reviews_count
- Full Reviews
- Shelf-Categorization
- Labels and their counts

Activation Function $y_i = tanh(v_i)$

$$Loss(\hat{y}, y, W) = \frac{1}{2}||\hat{y} - y||_2^2 + \frac{\alpha}{2}||W||_2^2$$

Adam
$$(m_t)_i = \beta_1(m_{t-1})_i + (1-\beta_1)(\nabla L(W_t))_i,$$
$$(v_t)_i = \beta_2(v_{t-1})_i + (1-\beta_2)(\nabla L(W_t))_i^2$$
$$(W_{t+1})_i = (W_t)_i - \alpha\frac{\sqrt{1-(\beta_2)_i^t}}{1-(\beta_1)_i^t}\frac{(m_t)_i}{\sqrt{(v_t)_i}+\varepsilon}.$$

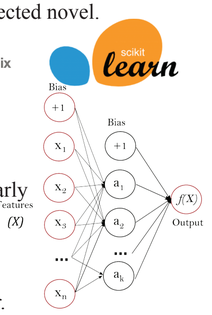
Default Values $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$

Hidden Layers
Neurons per layer

$$N_{hidden} = \frac{N_{sample}}{\alpha DOF}$$
$$N_{neuron} = \frac{2}{3}DOF$$

Methodology and Models

The general task of the project is classification of whether a book has the potential to become a New York Times bestseller. The model selected for our project is a multivariate classifcation model. We choose to implement this multilayer perceptron regression, as it has the capability of handling non-linearly separable inputs and sparse input. We suspect that some of the features, especially ones generated from the review and description of the novel, i.e. natural language features, will not be linearly separable given the context in which certain words appear.



Multi-layer perceptron classification is a supervised learning technique that utilizes backpropagation and activation functions which are non-linear. Activation functions for the hidden layers of the multi-layer perceptron are based on the hyperbolic tangent. We will be utilizing Sum of square error with a regularization factor for the weights. Loss will be minimized and the weights updated based on Adam algorithm.

Results and Error Analysis

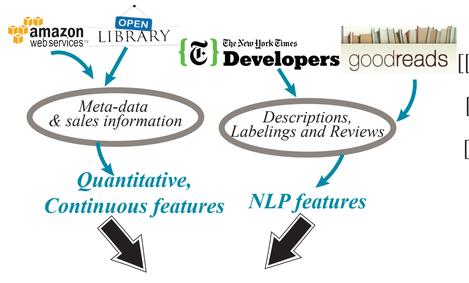


Figure 2: Our Pipeline: an Multi-source Effort

Book: The Rooster Bar
{'publisher = Doubleday' = 1,
'description = Three students at a sleazy for–profit law school hope to expose the student–loan banker who runs it.' = 1,
price = 15, 'author = John Grisham' = 1, prev_bestseller = 1, page_num = 368, published_date = 2017–10–19, 'List_name = Combined Print and E–Book Fiction' = 1, etc.}

Figure 3: Example Entry of Data

```
[[-0.00122341 -0.17727802  0.15325789 ...,  0.03149813 -0.00407913
  -0.12654206]
 [-0.2232641  -0.78030652  0.60285576 ..., -0.10697831 -0.18738892
  -0.41709224]
 [-0.05389842 -0.13956173  0.11173908 ..., -0.02092262 -0.07529071
  -0.14939164]....
```

Figure 4: Sample Weights for Classification

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Figure 5: Potential Confusion Matrix

We have collected the following: for the training data set, we have gotten a score of 0.6936 which correspond to the sum of square error 121967845.874. For the t esting data set we have the score which is the R-square value as 0.463 which corre-spond to sum of square error 3647539.831.

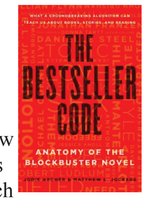
Discussion

We hope to retrain the neural nets with more accurate and complete data and complete sets of features including audience response derived from Goodread. We have yet to implement the feature where the algorithm a ccounts for the key current event during the year a book is published.

We will be expanding the numbers of books that we examine. Our next goal after the inclusions of more features is to optimize neural nets and to prune the hidden layers and their nodes. Hopefully with more data, the signal from the features can be more salient. We will also experiment with different solver such as Adam and see if there is any significant improvement in performance.

Cultural Implication of Our Model

This method of analysing bestsellers tells us what we may already know about our society. Archer and Jockers showed predictability in what sells within the text but did not look so much meta-textuality.The authors acknowledge, but do not give much credit to, the influence of reviews, splashy covers, big-name blurbs, and marketing budgets;We wondered how much did post-production really influence the craft and how there is always something looming over people's taste that is no longer individual. It is not so much about the text but presiding texture that determines taste and pre-ference well before we begin reading.



Future Directions and Limitations

Throughout this process, we were shocked by the opacity behind the bestseller systems. Despite the availability of APIs, the APIs themselves rarely present in a concise manner any useful or salient information. We were surprised that there is a lack of systematic organization. NYT itself does not publish or consolidate most of their historical data or reviews. Nielsen bookscan is another untransparent agent that is utterly unwilling to share information about sales.We want to imp-rove upon our NLP data using more sophisticated techniques than Ngrams

References

Archer, Jodie, and Ramón Saldívar. Reading the Bestseller: An Analysis of 20,000 Novels.2014. Print.

W. Y. Chong, B. Selvaretnam and L. K. Soon, "Natural Language Processing for Sentiment Analysis: An Exploratory Analysis on Tweets," 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, Kota Kinabalu, 2014, pp. 212-217

Piramuthu, Selwyn, Michael J. Shaw, and James A. Gentry. Classification Using Multi-Layered Perceptrons. Pittsburgh, Pa: Carnegie Mellon University, the Robotics Institute, 1990. Print.