JOURNAL OF
EMERGING INVESTIGATORS

Article

A Novel Model to Predict a Book's Success in the New York Times Best Sellers List

Matthew Lee[1], Siddhant Arora[2], Johann Lee[3], and Rohan Vaidya[4]

[1]San Marino High School, San Marino, CA , [2]Clements High School, Sugar Land, TX, [3]The Lawrenceville School, Lawrenceville, NJ, [4]Dougherty Valley High School, San Ramon, CA

SUMMARY

The popularity of media, such as books and music, has historically been considered difficult to forecast. This popularity is important in determining the success that can be achieved once the media is published. Therefore, we aim to evaluate the extent to which this fact holds true, as we propose that these public opinion trends are quite deterministic in nature. We investigated the important non-textual attributes determining a book's popularity, including but not limited to a book's previous ranking in the New York Times Best Sellers list and its popularity in searches. We then constructed two models: a generalized classifier of a successful Best Seller and a predictor of a book's weekly rank on the New York Times Best Sellers list. The reasonable accuracy of our classification and regression models suggest that book popularity is indeed deterministic. These findings point towards definitive characteristics that can help creators produce successful works.

INTRODUCTION

Even with the advent of technology, reading remains a popular pastime for Americans as the average American reads 12 books every year. Last year, 695 million books were sold in the United States alone (1). Even with so many books to read, only a select few become extremely popular. The exclusivity of book popularity made us wonder whether it is possible to predict which books will be popular. In the past, studies have generated algorithms to predict the popularity of other media such as research papers and songs with a high degree of precision.

FutureRank, an algorithm that uses citation information, authors, and publication time to predict the future ranking of a scientific paper, was able to rank the first 25% of retrieved articles with 100% precision (2). Another model plotted the path a song takes through the Billboard Top 100 by predicting the $n+1^{th}$ rank from the nth previous rankings with very low error (3).

Given the success in predicting popularity in other forms of media, it is likely that by considering the right attributes, book popularity can be predicted as well. Several of these attributes have already been highlighted in past research. Six literary features – action, measure of emotion, personalities of major characters, themes, romaticization, and simplicity – can determine what makes a best-selling novel to an accuracy of 82% (4). Characteristics like the season published or genre, which intuitively appear crucial to book success, do indeed play a large role in a book's New York Times Best Seller list rank; however, other more subtle properties like initial rank also heavily influence the book's future ranks (5). Such features have been used to manually classify Best Sellers against non-Best Sellers with an accuracy of 86% (6).

Model	Accuracy
Random Forest	98.4560570%
SVM (Linear Kernel)	92.3396674%
Naive Bayes	90.1425178%
SVM (Sigmoid Kernel)	71.6745843%

Table 1. Accuracy of the different models tested.

Despite identification of contributing features, there has yet to be a model that predicts a book's success. Based on past studies, we believed that the various characteristics of a book have a strong correlation to its success and popularity. We hypothesized that if a book hits the New York Times Best Sellers, its relative weekly performance will strongly correlate with future performance. Therefore, our research was divided into two subsections: the classification model and the regression model. Our classification model determined whether a specified book would make the New York Times Best Sellers or not. If the book did indeed make it, our regression model predicted the path of the book's weekly ranking. We believe this two-fold method best accounted for how most books never make it onto the New York Times Best Sellers. Our classifier and predictor are the first to focus specifically on the success of literature based on nonliterary features. Our models are also a proof of concept that prediction of the New York Times Best Sellers is possible.

RESULTS

We built a classifier to classify Best Sellers from non-Best Sellers and a forecasting model to predict the path of the book on the New York Times Best Sellers in order to determine what non-literary features impacted popularity most. For the classification model, we used non-literary characteristics of a novel as features for several classification algorithms. For the forecasting model, we used three time series as features for a time-series based regression model. Our example book was *Winter of the World* (ISBN:0525952926), a New York Times Best Seller.

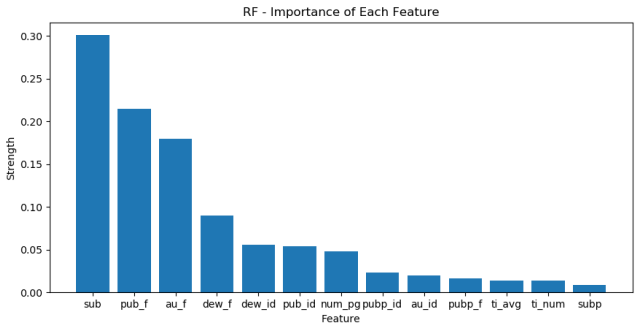

Figure 1. Random Forest Model-Relative Importance Scale and Visualization. The strength of each factor in the random forest model (a measure of how often the factor was considered in the model's nodes). To train this model, 6,000 books were used. The features are from left to right: subject, publisher frequency, author frequency, Dewey Decimal number frequency, Dewey Decimal number, publisher, number of pages, publishing location, author, publish place frequency, title average word length, title number of words, and subject place.

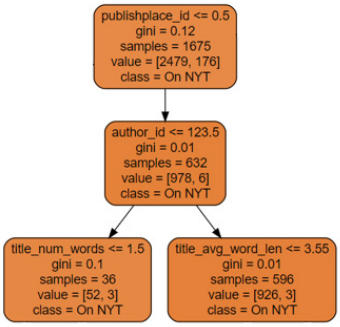

Figure 2. Sub-section of the RF Model Decision Tree. Decision tree framework can be represented using a graph-like model of yes/no statements. The RF model used in this paper contained several hundreds of nodes.

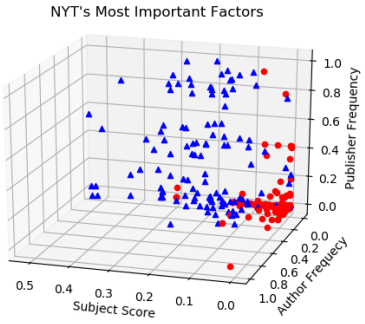

Figure 3. 3-D Plot of the Most Important Features. Reducing the k-dimension data to 3, this 3-dimensional plot constructs a visualization of the high separability between books that made the list and books that did not.

Classification

Winter of the World was correctly classified as a Best Seller. Several models were used to classify this book. Random forest classification with 100 estimators seemed to produce the best fit to the training data with an accuracy of around 98.46%, indicating that of the 1684 testing entries, only 29 were misclassified (**Table 1**). Each classification model utilized all of the 13 features: subject score, publisher frequency, author frequency, Dewey Decimal frequency, Dewey Decimal ID, publisher ID, number of pages, publish place ID, author ID, publishing place frequency, title average word length, title number of words, and subject place score. Frequencies measured the occurrence of their respective categories among all the books surveyed. Other IDs and scores represented features of a book in a numerical format. The relative importance of each of these features (how much they impacted a book's classification) was determined (**Figure 1**). A portion of our Random Forest Decision Tree nodes and the if-statements that determine how a book will be classified was also generated (**Figure 2**).

Surprisingly, we found that the textual features we considered, title average word length and title number of words, played little effect on whether a book was classified as a Best Seller or not. Non-literary features dominated in terms of relative importance (**Figure 1**). Therefore, we took the three most important features (subject score, publisher frequency, and author frequency) and constructed a 3D plot to better visualize the distinction between Best Sellers and non-Best Sellers (**Figure 3**). Finally, a confusion (error) matrix was presented to view the false positive/negative rate of the random forest model (**Figure 4**). Type I and II error rates for the RF model were quite low, with approximately 50 false positive and 200 false negatives out of the entire 6,000 book dataset.

Regression

We defined the best model as the model with the lowest cumulative mean absolute error (MAE) at week 10. The four models we tested were determined by the different combinations of (m, k) we considered. Each model considered past rank, weekly percent genre (the ratio of the number of books in that book's genre to the total number of books), and Google Search Ranks index.

The MAEs of the various possible combinations per week were plotted (**Figure 5**). An (m, k) of (2, 2) consistently has the lowest MAE at week 10 (4.449), making it the best (m, k) by our metric. By running our model with various combinations of features, we generated three different predicted paths (**Figure 6**). Overall, considering all three features provided the best fit prediction.

We ran a Student's t-test on the β coefficients (2, 2) model to determine whether our features had a statistically significant relationship with the ranks and generated a p-value of $3.55 * 10^{-8}$. Given that our value was much lower than the commonly accepted alpha of 0.001, we rejected the null hypothesis, allowing us to conclude that there was a relationship between previous rank, genre, and Google Trends index with current rank.

DISCUSSION

For all models, we accounted for overfitting by performing cross validation. The sample size of our data was small enough that overfitting was possible, so cross validation

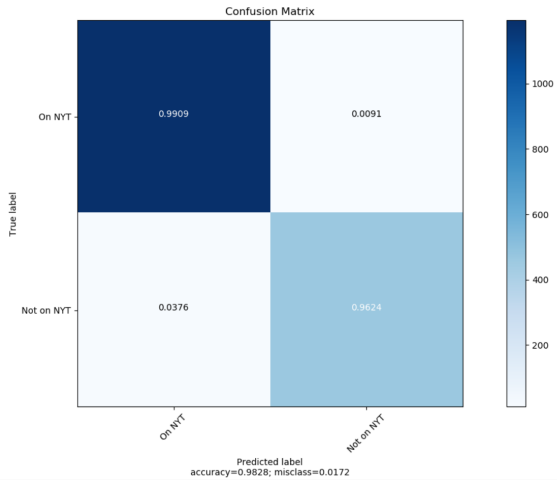

Figure 4. Confusion (Error) Matrix. Confusion matrix of the random forest model. Rows represent the true label and columns represented the predicted label. Color intensity represents number of books in that category.

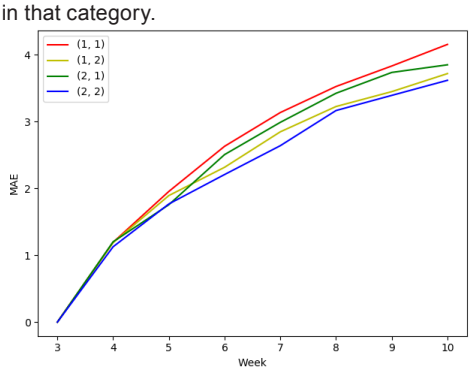

Figure 5. MAE per Week. Mean absolute error of our training model while considering multiple combinations for the degrees of both the regression and time series models. The x-axis represents the week on the Best Sellers and the y-axis represents of the mean absolute error of the ranks predicted.

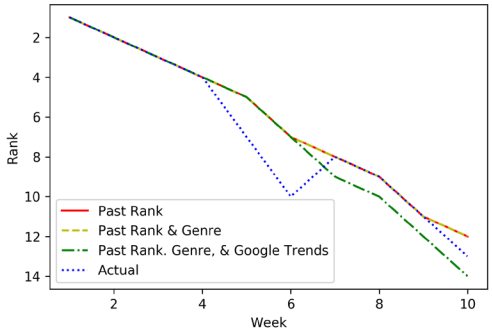

Figure 6. Actual vs Predicted Path for Winter of the World. An example of the predicted path against actual path for Winter of the World, taken with past rank only, past rank and genre, and past rank, genre, and Google Trends. The x-axis represents the week on the Best Sellers and the y-axis represents the rank predicted on the Best Sellers list.

ensured that our models were still able to generalize. We implemented five-fold cross validation, which split our dataset into five randomly assigned groups. Each unique group was used as a validation data set while the model trained on the other four. The model was then evaluated based on prediction performance on the validation data set. From this process, we were able to determine how extensively we needed to train to optimize error and generalization capabilities. By using classification models like random forest and support vector machines (SVMs), we were able to predict the chance of a book reaching the New York Times Best Seller List with high probability given a set of certain features. Unlike prior approaches using textual analysis and NLP algorithms, we introduced a novel approach of classification where we only considered external discrete features. Our results suggest that if a book includes certain characteristics, it is able to reach the New York Times Best Sellers without regard to its actual content. By using multilevel quadratic regression, we were able to create a model that predicted the path of a book through the New York Times Best Sellers with a MAE of 3.449 when predicting 10 weeks ahead. Our analysis and statistical significance provide strong evidence suggesting that book popularity is deterministic in nature and can be represented as a function of past rank, weekly percent genre, and search popularity. We are not aware of any previous studies utilizing our methodology to predict book ranking and our experiment suggests such prediction is indeed possible.

The limitations our research faced can be split into two main categories: application programming interface (API) restrictions and data restrictions. Many of the APIs (interfaces used to collect data) we considered using were unusable or were heavily rate limited. On the data restriction side, several considered sources were either hidden behind paywalls or defunct. Apart from overcoming the limitations we faced, future research could implement other models of time series regression such as conditional random fields and hidden Markov models. In addition, our research could be applied to other forms of media in order to recognize whether similar properties apply and whether future rank is similarly deterministic. We can also make our models more accurate by incorporating more features.

The two models suggest that book success has little to do with the actual content of the book; if the book fits into certain categories, it is likely to become popular. Although cultural phenomena like book popularity are generally regarded as random, our models indicate that the underlying mechanisms for success are more mathematical than they appear.

MATERIALS AND METHODS

Technologies

Data manipulation was done with the Python library pandas, which allowed for data parsing and imputations for any missing data (8). For the book rankings regression, we made use of both scikit-learn's built-in ridge regression and numpy's polyfit; likewise, scikit-learn's classification tools gave us great ease in analyzing and constructing model (9).

Data

The New York Times Best Sellers dataset, which returned the top 20 Best Sellers each week from 2008-2018, made up the core of our data (3). Additional data came from the

Goodreads-10k book dataset, which gave us random assorted books (10). From our datasets, we were able to obtain two critical details for our time series: the path it took through the New York Times Best Sellers in terms of rank and the ratio of genres each week. The third detail of our time series, search relevance, was obtained using the Google Trends API. We found the first and last week each ISBN appeared on the New York Times Best Sellrs, and by using this interval and the book title, we were able to compile the search relevance of each book from one month before the first week to one month after the last week. The final data was split as follows: 70% to train, 15% to validate, and 15% to test.

For our classification model, we created one list of ISBNs that made the New York Times Best Sellers and another list of ISBNs that did not. We then collected details about the books, ranging from book weight to cover. One issue we encountered was the large number of categories some books belonged to. Our solution was to generate a hash map that mapped the categories present in the New York Times Best Sellers to their respective appearance frequency. We then represented the categorical data as the average of the frequencies of all the categories. Alongside this, we also added original IDs to the author as a measure of mapping authors to their respective titles. Through this data discretization method, we were able to represent categorical data as numerical data. The final data was split in the same ratio as the regression data.

Classification

Although the initial dataset was composed of nearly 14,000 verifiable books, valid information was only accessible for 6,000 books. These books could be described by 13 features: subject, publisher frequency, author frequency, Dewey Decimal number frequency, Dewey Decimal number, publisher, number of pages, publishing location, author, publish place frequency, title average word length, title number of words, and subject place. Contrary to a previous approach. we instead applied less importance on the 'textual values' of a novel and used categorical external factors purely for classifying the books onto the New York Times List (7).

We were able to numericize our frequencies of data by applying a tweaked version of the Natural Language Processing bag-of-words process, in which the encapsulating dictionary was built only from information of books that have been on the New York Times's Best Seller List in the past decade. Thus, a frequency score was given as follows:

$$F_i = \frac{f_i}{max(f)}$$

where f is the frequency of a certain term in the New York Times list. Frequencies scores, ranging from 0 to 1, were calculated for the author, publisher, Dewey Decimal, and publishing place columns. Next, in order to account for large-scale authors, publishers, and genres, an ID for each of category served as an efficient way of recognizing prevalent authors in the New York Times list. This helped predict with high probability that books by these big-name columns would likely achieve success, providing an alternative to the frequency score.

One area of interest was the subject field, which provided a list of anywhere from zero to a hundred unique words and phrases to identify the books by. Once again, we applied the alternate bag-of-words scheme to obtain a total dictionary of unique subject counts strictly from the New York Times books. After eliminating extraneous data fields like “Fiction” and “Protected DAISY” (as they were close to two orders of magnitude more common than other subjects and were not specific descriptions of the book), we calculated a subject score from a book's subject list S as:

$$S_i = \frac{\sum_0^n f_i}{n}$$

where the subject list's length is given as n and the frequency of a subject in the New York Times's dictionary is given as f_i.

This approach is also applied to calculate a subject place score (which is a data field that relays additional information about the setting of a book).

For the classification, we tested SVMs, naive Bayes, and random forests with a 70/15/15 train/validate/test split and compared their accuracies. Support vector machines with kernel trick raise the data by a dimension, then separate the classes with a hyperplane. Naive Bayes models utilize the Bayes Theorem with assumption of independence between the features. Random forest models are a collection of decision trees with feature bagging.

Regression

Initially, we turned the weekly top charts data into a hash map that mapped from each book to a list of its rankings over the weeks in which it traversed the top charts. In addition to that, we took the genres of the books on the New York Times Best Sellers and compiled a 2-D array that described the percentage of books of each specific genre in the Best Seller list per week. For search trends, due to the Google Search Trends index being relative to other searches at the same time period, many books had relative search strengths of 0. To counteract this, we took the data of the books which had more than 25% 0s and generated cumulative weekly searches instead. This helped mitigate the effect of sparse datasets, as seven occurrences of 0 would translate to one occurrence while concurrent zeros with at least one positive value would translate to a positive integer. As a result, sparse data was condensed while retaining relative search interest.

In order to run regression on all three of our time series, we needed to convert them into non-temporal numeric values. Hence, we ran regression on each time series and used the coefficients of each time series regression as features in the overall regression model. Due to the Central Limit Theorem, for n number of samples where $n > 30$, the distribution is a close approximation of the normal distribution (11). Since we have 400 samples for our error terms, which is more than 30,

we pass normality.

Initially, we modeled each of our features - past rank, weekly percent genre, and Google Search Trends index - by the function:

$$f(x) = \sum_{i=0}^{k} a_i$$

where *k* is the degree of regression, and *x* is the past data. We set random initial a_i and applied gradient descent until we reached a local minimum cost. We then took each ai from each model, and fed it as parameters into the level 2 model:

$$f(x) = \sum_{i=1}^{m} \sum_{j=1}^{k} b_i a_j^i + c$$

where *m* is the degree of regression for the level 2 model, *k* is the degree of regression for the time series model, and *c* is a constant. After, we ran gradient descent on the level 2 model with 5-fold cross validation to avoid overfitting. To determine which *m* and *k* were the best, we tried (*m*, *k*) combinations: (1, 1), (1, 2), (2, 1), and (2, 2). We did not test values of *m* and *k* beyond 2 because of a strong possibility of overfitting our data. Error was the difference between predicted rank and actual rank, giving us the mean absolute error:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

A better MAE indicates a more accurate predictor and a better (*m*, *k*) as well. We suspected that certain features were not impactful and ran models with the following feature combinations: past rank, past rank and weekly percent genre, past rank, weekly percent genre, and Google Search Trends index. The change in MAE between features was an intuitive indicator for which features mattered, while the collection of bij was a more technical measure of importance.

To predict the entire path of the book, we took the prediction of the rank in the $n+1^{th}$ week from the previous n weeks of data, starting with $n = 3$ (as 3 points is the minimum number of points needed for a quadratic line), assuming that it is the actual next value, and appended it to the data. We then re-ran the models again, with the updated data set, and got a new prediction for the $n+2^{th}$ week. We repeated this until we have predicted *w* weeks ahead, and the predictions of ranks from 3 to *w* weeks trace a path that the book rank will take. To determine whether our model was statistically significant, we ran a Student's t-test to determine whether our explanatory variables had a statistically significant impact on the response.

ACKNOWLEDGEMENTS

We gratefully acknowledge Lina Kim, Director of Pre-College Programs at University of California Santa Barbara for her administration of the Science and Engineering Research Academy program, which supported this research. We also express gratitude towards Shadi Mohagheghi, Rachel Redberg, and Angela Zhang for their advice during this research and their review of this manuscript. We would also like to thank the Journal of Emerging Investigators for giving us the feedback needed to improve our manuscript and the opportunity to publish.

Received: October 13, 2019
Accepted: May 06, 2020
Published: May 23, 2020

REFERENCES

1. Watson, Amy. "Unit sales of printed books in the United States from 2004 to 2018 (in millions)." *Statista*, 14 Jan. 2019. https://www.statista.com/statistics/422595/print-book-sales-usa/.
2. Sayyadi, Hassan, and Lise Getoor. "FutureRank: Ranking Scientific Articles by Predicting their Future PageRank." *Proceedings of the 9th SIAM International Conference on Data Mining*, 2009, pp 533-544.
3. Cibils, Cristian, *et al*. "Predicting a Song's Path through the Billboard Hot 100." Stanford CS229, 2015, pp 1-3.
4. Harvey, John. "The Content Characteristics of Best-Selling Novels." *Public Opinion Quarterly*, vol. 17, no. 1, 1953, pp. 91-114.
5. Yucesoy, Burcu, *et al*. "Success in Books: A Big Data Approach to Bestsellers." *EPJ Data Science* 7, 2018, pp 2-3.
6. Archer, Jodie. Reading the Bestseller: An Analysis of 20,000 Novels. 2014. *Stanford University*, PhD dissertation.
7. Archer, Jodie, and Matthew L. Jockers. The Bestseller Code: Anatomy of the Blockbuster Novel. *St. Martins Griffin*, 2017.
8. McKinney, Wes, *et al*. "Data Structures for Statistical Computing in Python." I*n Proceedings of the 9th Python in Science Conference*, 2010, Vol. 445, pp. 51–56.
9. Pedregosa, Fabian, *et al*. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 2011.
10. Zając, Zygmunt. Goodbooks-10k. *Github*, 27 Nov. 2017. https://github.com/zygmuntz/goodbooks-10k.
11. "The Central Limit Theorem." *Florida State College at Jacksonville*, 2019, https://guides.fscj.edu/Statistics/centrallimit/. Accessed: 14 July 2019.