UNIVERSITY OF CALIFORNIA

Los Angeles

Predicting New York Times Bestselling Fiction Books Using Text & Image-Based Machine Learning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Anum Iqbal Damani

2025

ABSTRACT OF THE THESIS

Predicting New York Times Bestselling Fiction Books
Using Text & Image-Based Machine Learning

by

Anum Iqbal Damani
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2025
Professor Yingnian Wu, Chair

This analysis focuses on predicting New York Times (NYT) bestselling fiction books, specifically on the Combined Print & E-Book Fiction list, using book descriptions and book covers. Approximately ten years of data from the NYT Books API and Open Library API is collected, and an exploratory data analysis is performed. All machine learning models are evaluated using performance metrics in the classification report, including accuracy, weighted average precision, and weighted average recall. In addition, the model performances are assessed using the precision-recall curve, ROC curve, and confusion matrix. Although the dataset is fairly balanced, NYT bestsellers are rare in reality, so the classification threshold is tuned for each model and precision is prioritized. Seven text-based models are implemented by utilizing cleaned book descriptions as input, which are transformed using either TF-IDF or BERT. Based on the overall performances, the text-based models with the best performances are Logistic Regression with BERT and XGBoost with TF-IDF, with strong accuracies at 92% and 91%, respectively. Next, five image-based models are implemented by extracting image features from book covers. The overall performances of the image-based models indicated that the XGBoost model with CLIP embeddings and the ResNet50 model are the strongest, with accuracies 87% and 83%, respectively. The best text-based and image-based models are then considered for multimodal modeling. Four multimodal models are implemented and compared. The Logistic Regression model with BERT & CLIP features was the top-performing multimodal model, achieving a strong accuracy of 97%.

The thesis of Anum Iqbal Damani is approved.

David Anthony Zes

Vivian Lew

Frederic R. Paik Schoenberg

Yingnian Wu, Committee Chair

University of California, Los Angeles

2025

To my supporters

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

For decades, The New York Times (NYT) bestseller lists have served as indicators for book popularity, reflecting and influencing readers' preferences. After browsing through the shelves of bookstores or libraries, one will quickly notice that several books receive praise by NYT. Although NYT bestselling books are prominent throughout bookstores and libraries, the books that receive this special recognition are actually quite rare. Every year, around three million books are published, but the number of books that are featured on the NYT bestselling lists is under five hundred books [23].

My motivation for this analysis is to gain a better understanding of the books that receive this special recognition. In order to do this, it is crucial to investigate the relationship between book features and NYT bestseller status. The descriptions and covers of books can be influential in a reader's decision to select a book to read. This analysis aims to address the following research questions. To what extent can text-based features, specifically extracted from text descriptions, predict NYT bestsellers? Additionally, how accurately can image-based features, specifically extracted from book covers, predict NYT bestsellers? Finally, how effective are multimodal features, which are combined textual and visual features, in predicting NYT bestsellers?

While the NYT has many bestseller lists, this analysis focuses solely on the Combined Print & E-Book Fiction list, which considers print and e-book sales of fiction books. My focus on the Combined Print & E-Book Fiction list stems from the widespread popularity of both print and e-book formats among readers. Moreover, the Combined Print & E-Book Fiction list provides a holistic perspective on fiction book success. For the purpose of this analysis, I have defined the term "bestseller" to refer to books that have been featured on the NYT Combined Print & E-Book Fiction list between January 4, 2015 to January 19, 2025. Furthermore, the

books that have not been featured on the NYT Combined Print & E-Book Fiction list during this date range are labeled as "non-bestseller." A book that is labeled as "non-bestseller" in the dataset could have received recognition on a different NYT bestselling list, such as "Hardcover Fiction."

This analysis revolves around a binary classification problem with fairly balanced classes. In this dataset, there are 4,666 observations in total, where 2,216 observations are bestsellers and 2,450 observations are non-bestsellers. To account for the rarity of NYT bestsellers in real life, my approach involved tuning the classification threshold for all models to above 0.5, if possible. Through this approach, I prioritized precision, while aiming to maintain a reasonable recall around 0.8. Although it is not ideal to miss bestsellers, it can be argued that the cost of false positives is greater. It can be risky to falsely predict a book as a bestseller due to the amount of resources that are allocated in the book publishing process. Through the implementation of text-based, image-based, and multimodal models in this manner, I hope that my exploration of this topic provides valuable and applicable insights about fiction book success across print and e-book formats.

Chapter 2

Data

2.1 Data Collection Methods for Tabular Dataset

2.1.1 Bestsellers

The New York Times bestselling fiction book dataset was collected using the New York Times Books API [22] [16]. This process involved obtaining a NYT API key from the NYT Developer portal and generating an API request for the specified bestseller list of interest for this analysis: "Combined Print & E-Book Fiction." Next, the bestseller lists were downloaded using the desired date range. The date range selected for this data is January 4, 2015 to January 19, 2025. Specifically, the variables obtained are: title, author, list publication date, book image URL, and book description.

The NYT Books API also provides other information about bestsellers, such as rank, number of weeks on the list, ISBN numbers, and publisher name; however, these variables were excluded from the analysis. Note that the NYT Books API includes *list_publication_date*, which is the publication date of the list in which the book was featured. In order to obtain the book publication year, it was necessary to fetch this information from Open Library API [11]. Open Library API allows users to connect to a large database containing information pertaining to more than 39 million works [12]. A function was written to obtain the book publication date from Open Library API for every row in the dataset, specifically using the title and author information. The variable name for the resulting column for the book publication date is called *year*. Finally, the *list_publication_date* variable was dropped from the dataset.

In terms of data cleaning, the duplicate observations were dropped from the dataset. There were 76 observations with missing book descriptions, so these observations were dropped. There are 352 observations with missing values in the *year* column, but these observations were

kept in the dataset. A new column for the class label 1 was added to this dataset. The resulting NYT fiction bestseller dataset includes 2,216 observations. To summarize, the columns of the bestseller dataset are:

- ***title***: The title of the book
- ***author***: The author(s) of the book
- ***year***: The publication year of the book
- ***book_image***: The URL for the book cover image
- ***description***: The description of the book
- ***label***: The label column contains the value 1

2.1.2 Non-Bestsellers

Recall that the NYT bestseller dataset spans ten years of data, from 2015 to 2025. In the NYT bestseller dataset, there are several books published prior to 2015 that were featured on the NYT Combined Print & E-Book Fiction bestseller list in 2015. Specifically, *The Alchemist* by Paulo Coelho was originally published in 1988 but it was featured on the NYT bestselling fiction list in 2015. Another book featured on the NYT bestselling fiction list in 2015 is *Sharp Objects* by Gillian Flynn, which was originally published in 2006. Lastly, a book that was published in 2011 and featured on the NYT bestselling fiction list in 2015 is *The Martian* by Andy Weir. It is important to note that the film adaptation of *The Martian* was also released in 2015. The book *The Martian* by Andy Weir is an example of a "backlist revival," a term that is prevalent in the book publishing industry which emphasizes surges in popularity, or prolonged popularity, of books years after their publication date. It is possible for the popularity of a book to be driven by an upcoming film adaptation.

The non-bestselling fiction book dataset was collected using Open Library API, specifically the Subjects API. The data spans 2010 to 2024, and this date range was carefully selected. To mitigate the issue of data sparsity in Open Library API, the non-bestseller dataset includes more than ten years of data. Given the backlist revivals present in the bestseller dataset, and to allow for a fairer comparison, the lower bound for the date range of the non-bestseller dataset is chosen to be 2010. The upper bound for the date range for the non-bestseller list is 2024, rather