# No One Ever Said “Don’t judge a book by its features”

By Kristina Barounis

# 1.
# The Outline

# Data Sources

**New York Times API**

- For books that were on any NYT best selling list between 2017 and the present

**Web scraping Goodreads**

- For books that were not on an NYT list
- For all features

# Features & Observations

**Scraped**

- Title
- Author
- Rating
- Genre
- Publisher
- Publish date
- Number of pages

**Engineered**

- Series (Y/N)
- Top author (Y/N)
- Top 5 publishing company (Y/N)
- Month of publishing

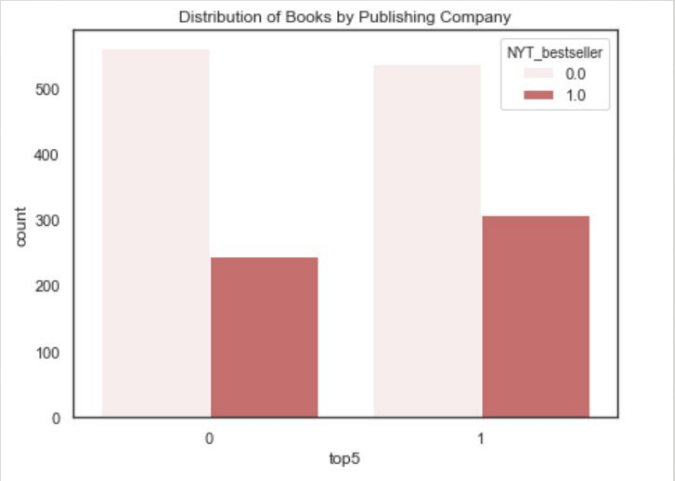
**Result**

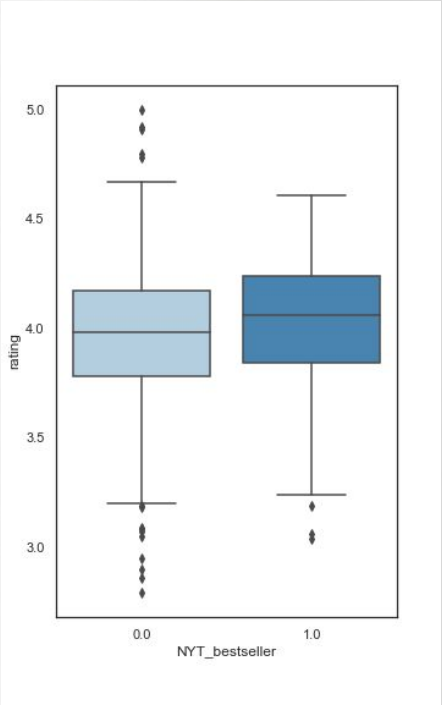
- Total of 37 columns in the dataset
  - Mostly categorical
  - 1 continuous: rating
- 1646 observations
  - 551 bestsellers
  - 1095 non-bestsellers

# How important are the top 5 publishing companies?

***Of the 551 bestselling books in the data set, 307 were published by the top 5...***

***But they also only publish 6% more bestsellers than everyone else***

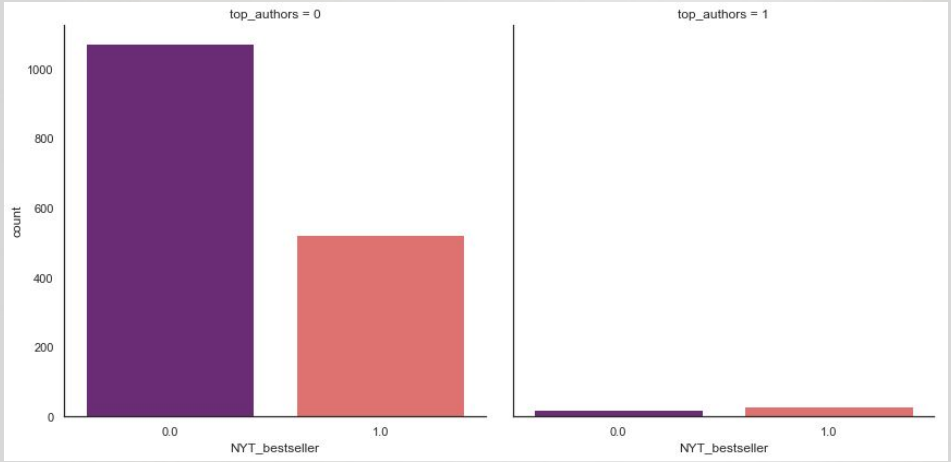


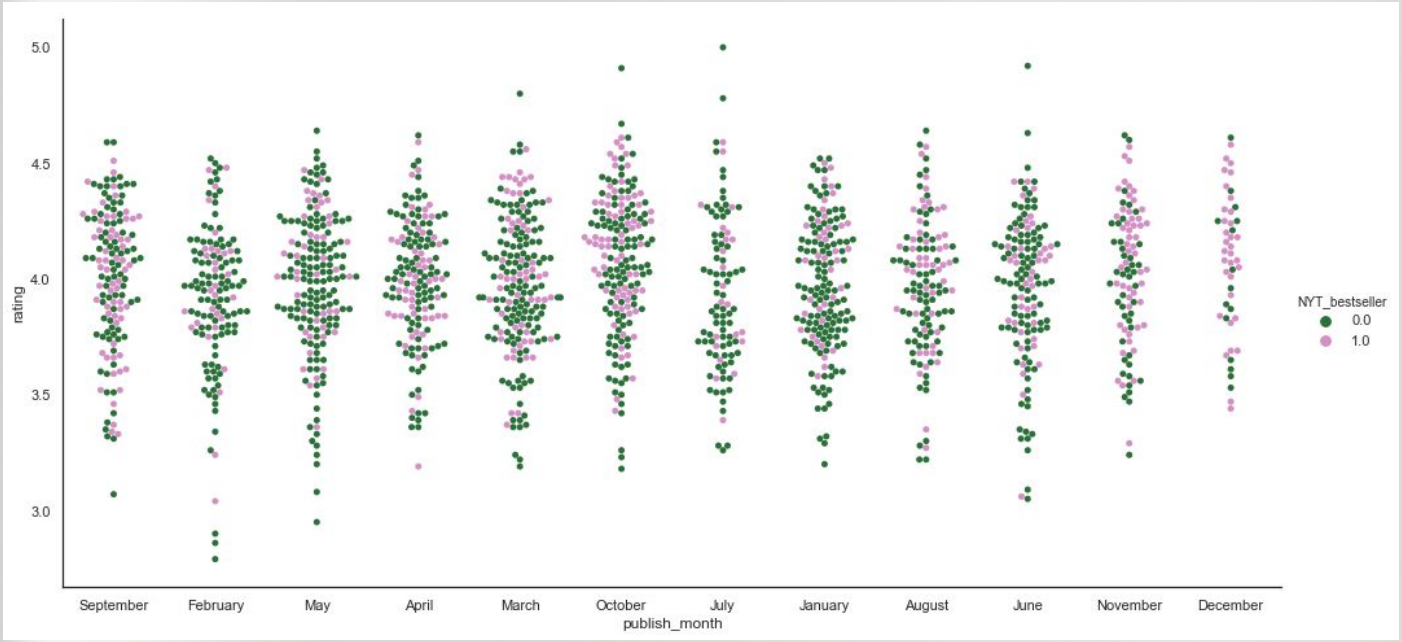



***Average rating for bestsellers is slightly higher - but this could be due to the "NYT stamp of approval"?***

***Forbes highest earning authors 2017 and 2018:***

- J. K. Rowling
- James Patterson
- Jeff Kinney
- Dan Brown
- Stephen King
- Nora Roberts
- John Grisham
- Paula Hawkins
- E.L. James
- Danielle Steel
- Rick Riordan
- Michael Wolff

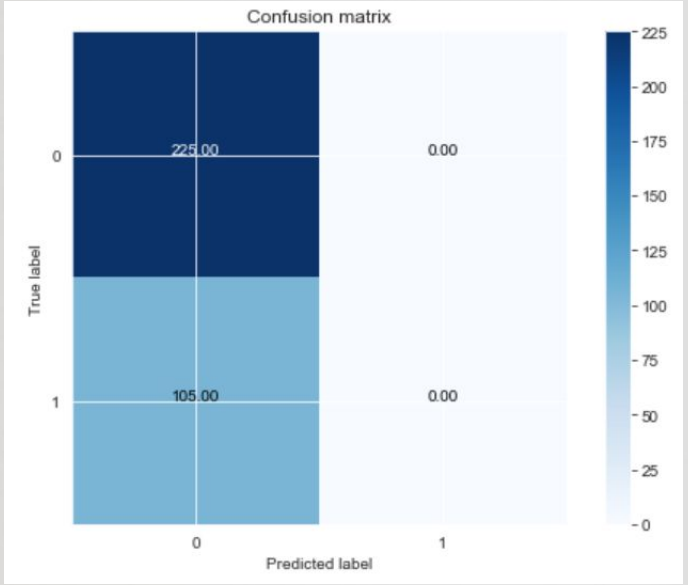



5.0
4.5
4.0
3.5
3.0
rating
September
February
May
April
March
October
July
January
August
June
November
December
publish_month
NYT_bestseller
0.0
1.0

# 2.
# The First Draft

# Baseline Model: Dummy Classifier

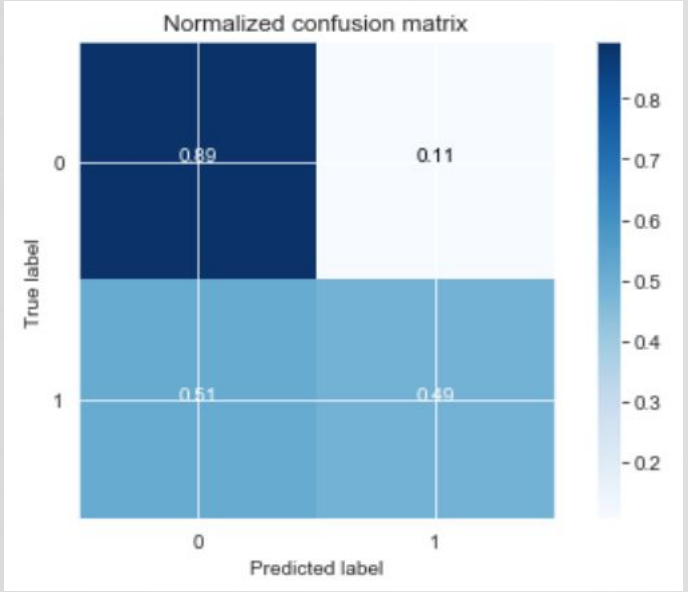
- 68% accuracy by picking the most frequent class, i.e. not a best seller



# 3.
# The Final Draft

# Final Model: Logistic Regression

- 75% accuracy
- Default hyperparameters turned out to be optimal
  - C: 1 (C=1/ƛ)
  - Penalty: L2 (Ridge)



# The CliffNotes

| Positive - higher likelihood | Negative - lower likelihood |
|---|---|
| Top authors | Horror |
| Politics | Science fiction |
| Publishing in December | Fantasy |
| Biographies | Short stories |
| Publishing in November | Fiction |
| Publishing with Penguin Random House | Young adult |

# Ready for publishing?!

- “The Institute” by Stephen King
  - Not a series
  - 4.36 rating
  - Top author
  - Published September
  - Not published by a top 5 company
  - Horror genre
- Correct!

- “Double Down: Game Change 2012” by Mark Halperin & John Heilemann
  - Part of a series
  - 3.87 rating
  - Not top authors
  - Published November
  - Published by Penguin
  - Politics genre
- Incorrect