Finding the Time to Read

Jake Kamen

Kamen.23@osu.edu

Department of Physics, The Ohio State University

Motivation

Authors want their book to be successful. The problem is that understanding how and when to sell is a very different skill from the ones needed to write a novel. This report aims to understand the when.

Goals:

1) Understand how the average ratings of a genre changes over time. Apply this method to month and year long scales
2) Analyzing the number of ratings a genre gets each month. Use an ARIMA model to forecast the future popularity of a genre

Dataset

Goodreads :

- Collection of over 2.4 million books
- Genres
    - Romance (392,851 books)
    - Fantasy and Paranormal (325,216)
    - History, Historical Fiction, and Biography (398,156)
    - Mystery, Thriller, and Crime (316,452)
    - Children's Books (392,851)
    - Non-Fiction (338,284)
    - Poetry (50,961)
    - Young Adult (230,047)
- Map their genre collection onto the full Goodreads data catalog
- Columns of data
    - Book ID number
    - Titles
    - Authors
    - Genre classification
    - Publication day, month, and year
    - Average Rating
    - Number of ratings
    - 20+ columns of other data
- Publication day, month, and year reworked into a time series

Ex.

3209319	6066819
To Have and Have Not	Best Friends Forever
Author id: 1455	Author id: 9212
History:12, Mystery:9	Romance:23
8-1-2006	7-14-2009
3.57	3.49
45	51184

Methods

ARIMA

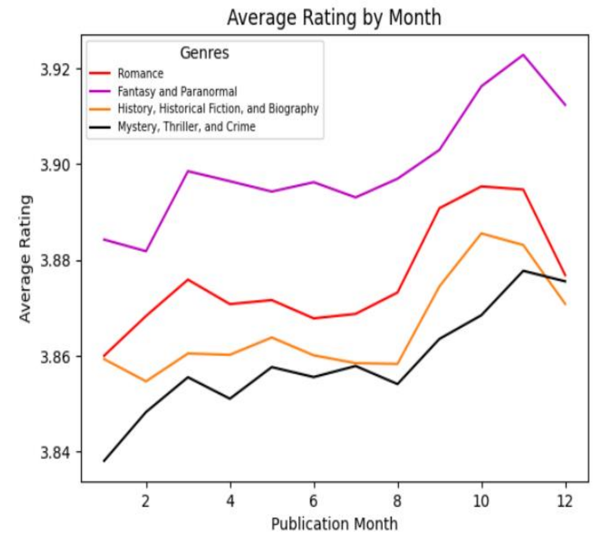
- Model used to forecast future engagement
- Auto Regression (AR) - the output variable depends linearly on its own previous values
- Integrated (I) - he difference between consecutive data points in time, forcing data to become stationary
- Moving Average (MA): follows trends and past data using residual errors, to create forecasts

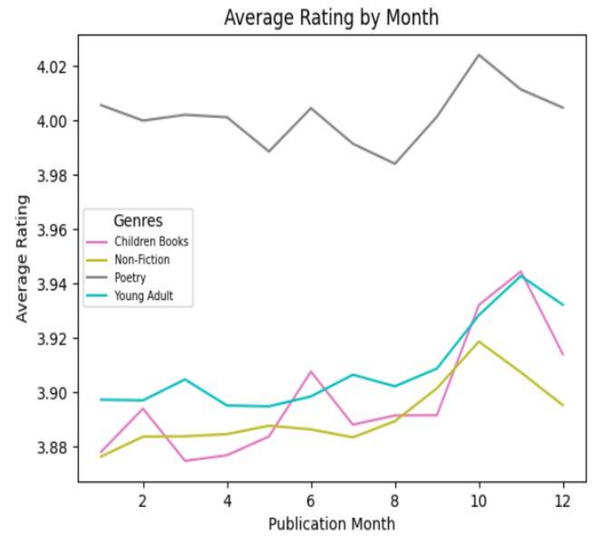
3 Parameters:

- p: number of lagged observations used during auto-regression
- d: degree of differencing
- q: order of the moving average, using lagged forecast errors

Results

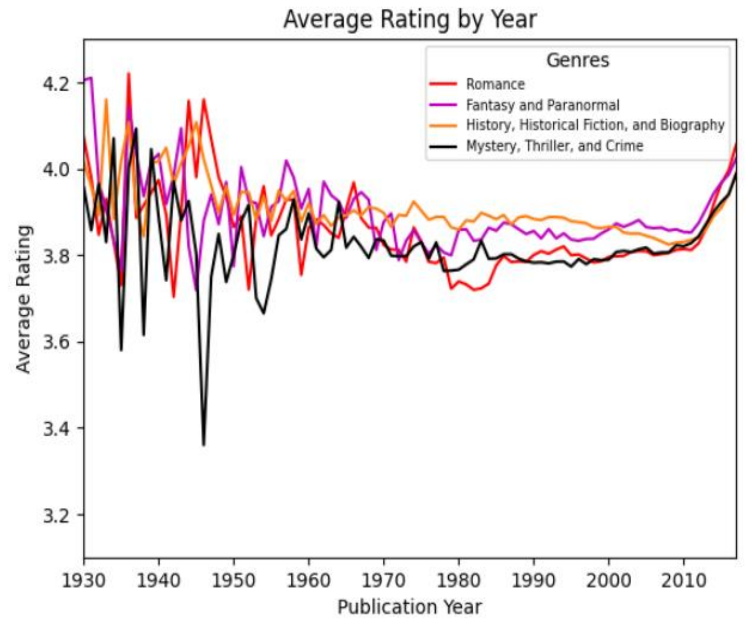
Average Ratings by Month

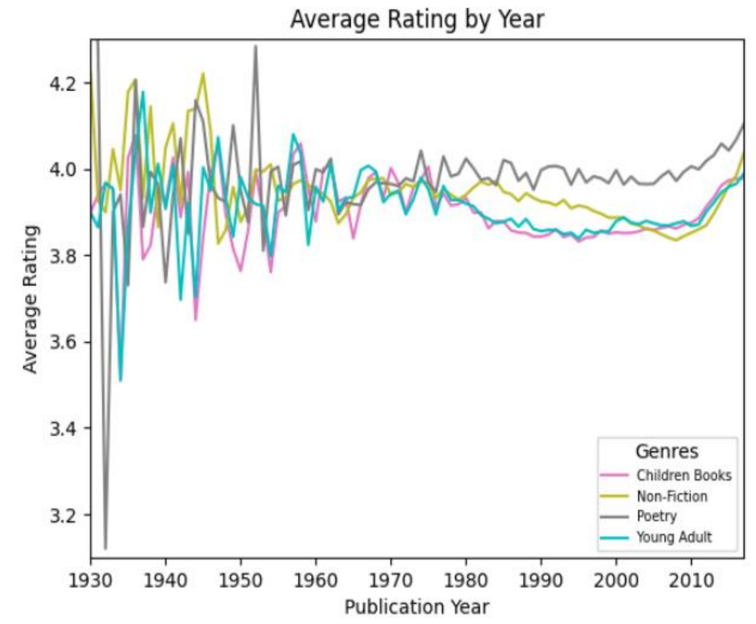




Average ratings continuously increase throughout the year, peaking in October. Higher average rating for poetry all year long

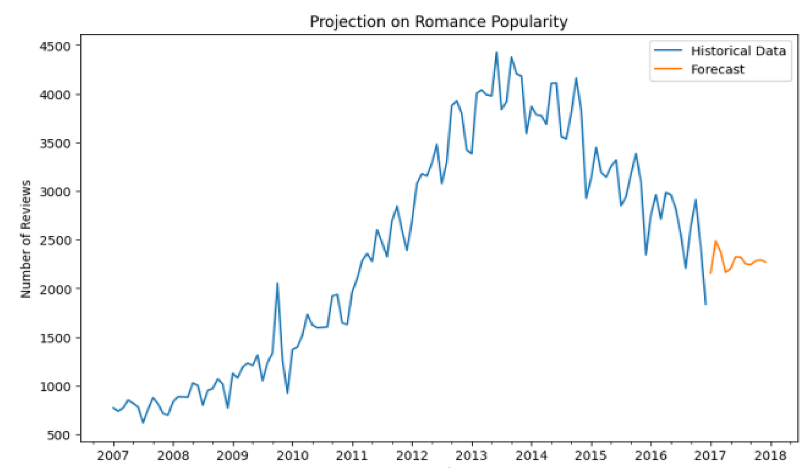
Average Ratings by Year

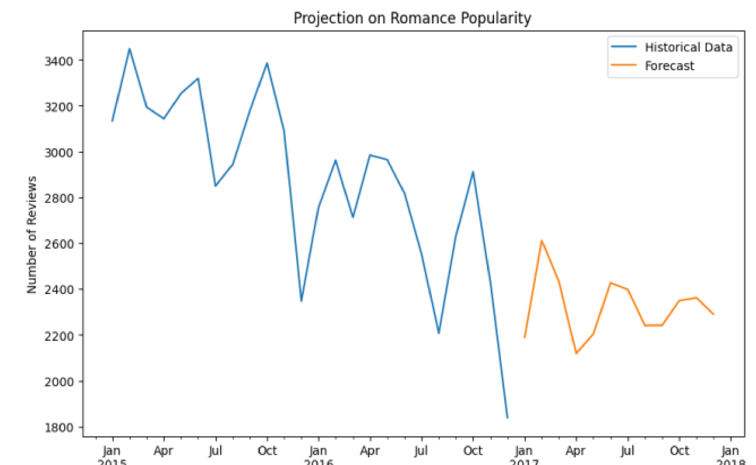




- Lots of noise before 1960's
    - Data inadequacy due to insufficient information
    - Fantasy peak in 1938 from The Hobbit by J.R.R. Tolkien
    - Mystery peak in 1934 from Murder on the Orient Express
- More books published and cataloged after 1960's
    - Ratings begin to stagnate due to volume of information
- 2010 begins an upward trend in average ratings
    - Increased usage in online rating platforms
    - Self-publishing leads to wider range of genres and topics

Engagement by Month



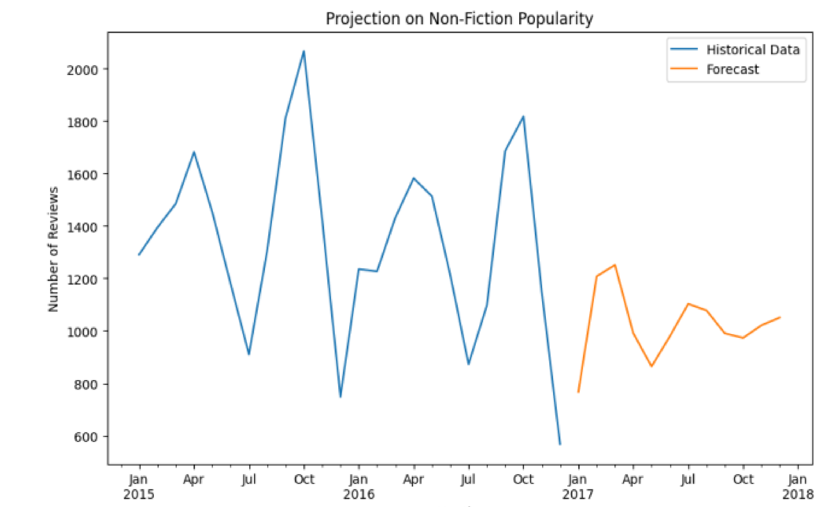


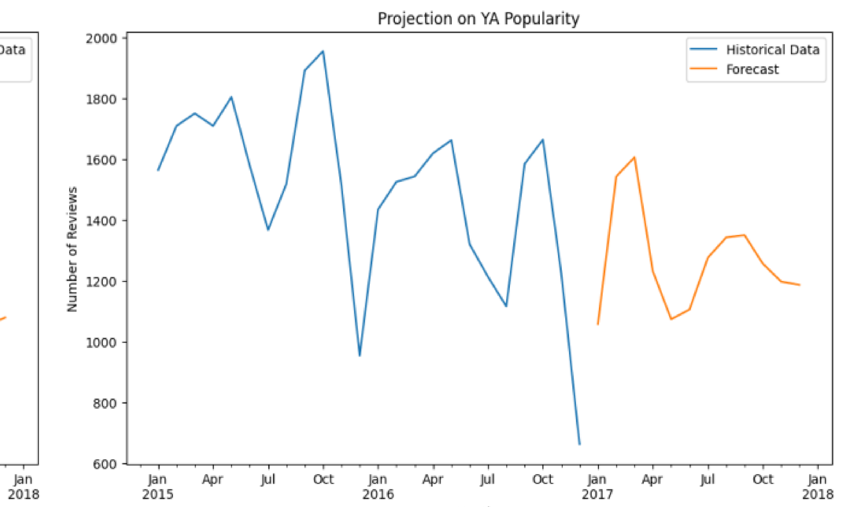
- Periodic reader engagement peaks for genres
    - Romance in February (Valentine's Day)
    - Poetry in April (National Poetry Month)
- Set up ARIMA to forecast genre popularity
    - Set parameters (p,d,q) to (3,1,0)
    - Forecast from two years of historical data
    - Separate by train and test subsets
    - Metric to use is Mean Absolute Deviation (MAE)
        - Measures errors between paired observations
    - Percent Error: Difference between predicted (forecast) value and actual value
- Apply ARIMA across all genres and minimize percent error
    - Rework p-value until percent error is minimized

Discussion

	Romance	Fantasy	History	Mystery	Children	Non-Fiction	Poetry	YA
Avg Review Count	2877	2028	1746	2054	756	1339	200	1495
p-value	2	4	2	3	10	2	9	7
MAE	339	262	289	317	403	189	342	442
Percent Error	11.8	12.9	16.5	15.4	53.3	14.1	171	29.5

- True p-values used above, opposed to the initially set value of 3, were decided through trial to minimize MAE
- Most genres produce good models, keeping the percent error around 15% or less
- Poetry model is insufficient to explain the data, even with an increased p-value for a more complex regression model
    - Small monthly average review count makes prediction hard
- Children's books also does poorly, although not as bad as poetry
    - Again, significantly less data to forecast on and abnormal peaks
- Interesting comparison between Non-Fiction and Young Adult books
    - While both sets look similar, the necessary p to minimized percent error for each is drastically different
    - Reveals the complexity of ARIMA and nuance for choosing parameters





Conclusions and Future Work

Two main ideas:

- Understanding the average ratings of genres over time
    - By Month:
        - Increases throughout the year
        - Genre specific peaks in different months
    - By Year:
        - Pre-1960 sporadic movement
        - Smoothing as data catalog grows
        - Upturn since 2010
- Forecasting genre engagement into the future
    - ARIMA model to predict genre popularity
        - Understanding the auto-regression by p-values
        - Running models to minimize the percent error

Future Work:

- Dive deeper into average rating shits over the past century, identifying 'classics' and analyzing the changes in genre approval from this
- Run additional forecasting models parallel with ARIMA to compare model effectiveness
- Introduce other metrics which reveal a model's goodness of fit

References:

[1] Wang, Xindi, et al. "Success in books: predicting book sales before publication." EPJ Data Science 8.1 (2019): 1-20.

[2] Maity, Suman Kalyan, et al. "Understanding book popularity on goodreads." Proceedings of the 2018 ACM International Conference on Supporting Group Work. 2018.

[3] Sachdeva, Hansika, Ujjwal Puri, and S. Poornima. "Predicting the popularity of books before publication using machine learning." AIP Conference Proceedings. Vol. 3075. No. 1. AIP Publishing, 2024.