THE OHIO STATE UNIVERSITY

Finding the Time to Read

Project Category: GoodReads
Physics 5680, Autumn 2024

Author(s): J. Kamen

December 12, 2024

Abstract

Time plays an incredibly important role in the popularity of books. Whether that be the year a book was published or simply the season it was published in, time affects how an audience absorbs material. Some genres may be more in fashion than others due to world events or other popular media. When people buy books and for what reason affects their genre choices, ratings, and sales too. Thus, this report analyzes two major situations, 1) the average ratings of genres each year over the past century or so, and 2) the number of overall ratings per genre per month. The first tells us about how a genre's popularity has changed over history, whether that be due to popular books from the genre or just increased genre approval by society. It gives us insight about what genres are best for publication. The second situation reveals when most ratings occur throughout the year for each genre, helping us to understand when it would be best to publish and when readers are engaged with that genre the most. An ARIMA prediction model is used for this, forecasting peak reader engagement in the coming months by understanding past review count patterns.

1 Introduction

We will explore the GoodReads dataset, a catalog of more than 2.4 million books. Goodreads itself is the world's largest website for readers and book recommendations. Their recommendation engine helps prospective readers find books that may be of interest to them by using a collection of data, such as book ratings and genre classification. However, the company does nothing to help authors whose books are contained on the site. This brings us to our main problem.

Every author wants their book to be successful. How to actually make that happen though, well that's a lot harder. The problem is that understanding how to sell, or even when to sell, is a skill very different from the ones needed to write a novel. So opposed to learning about what books to read from the audience's perspective, let's take a dive into the author's world, where every review and sale can make a difference.

We begin by looking at the two main focuses of this essay. Our first describes how average ratings for genres have changed over long periods of time, scaling average ratings into years, and comparing them with over a century's worth of data. By understanding which genres are popular and when, we can get a better understanding of society. We can pull out specific years of peaked ratings, created by books that are now renowned as classics. It is also worth it to contextualize the genres within history to understand why some, like mystery, took longer to be accepted by society.

The second focus describes a shorter period look at genres, visualizing how audience engagement fluctuates from month to month. For this task, we will take the number of ratings a genre is getting per month, not just the average rating, and train a model to predict future engagement. Our algorithm itself receives [book id, genre, publication day, publication month, publication year, number of ratings, average rating, title, authors] as inputs. Processed through an ARIMA network, which runs auto-regression on the data in order to predict the future number of ratings by month, we can visualize the audience engagement we're looking for.

2 Related Work

Many publishers and companies have sought out to analyze the best time to publish specific books and what sorts of genres are worth pursuing. To be able to understand the success of a book before it is even published is essentially the key goal of the publisher to begin with, as profit is their main goal. Therefore, a great many papers have been written to solve this problem.

The most information dense and expansive report on this issue was by Wang and Xindi [1] in 2019. They began by clustering genres, first by separating fiction and non-fiction and then, similar to the approach in this report, separating books into much smaller genres. They then graphed by topics and keywords to create a visually connected network between genres and classification ideas. To begin fitting their data they used a number of different models including Linear Regression, K-Nearest Neighbor (KNN), Neural Networks (NN), and Learning to Place (L2P). By using a large number of different model types, they could compare ROC curves and analyze each AUC score to determine the best model for the situation. Since this report isn't classification based, following the same AUC/ROC pathway as Wang and Xindi is not reasonable. Instead, we will use a Mean Absolute Error (MAE) approach, another metric for understanding model proficiency. Additionally, although Wang and Xindi's paper [1] used many different models to determine the best approach, the amount of thought put into the many networks for their report is astounding. Even while it produces fairly good results, the sheer time to produce the same output in this report seems unreasonable.

Studies analyzing book catalogs to predict popularity have been produced for other purposes as well. Suman Kalyan Maity [2] uses the Readers Choice Awards, a nomination and voting sector of the Goodreads site. Their approach is to see if they can predict, based on the characteristics of a book, how many votes they would receive, and therefore the books popularity. Maity's paper helps to discuss books popularity through internal attributes, outputting votes and projected engagement. This paper, however, uses engagement as an input, outputting larger-scale projections on entire genres. Another paper, by Hansika Sachdeva [3], uses various text features of a book to predict popularity. Unlike the first two papers mentioned, however, Sachedva's paper focuses strongly on the publisher's perspective, discussing how the meta-data of a novel may predict the success of it for the publishers. All 3 above, regardless of their focus and methods, use unbelievably large groupings of data though, helping to defend their approaches independently.

3 Dataset

The Goodreads dataset, as described above, is a collection of over 2.4 million books. The dataset was collected in late 2017 from GoodReads.com and provide us with an unbelievable amount of books and chunks of information for each. Wan Mengting, a principal research scientist at Microsoft's office of applied research, and her team initially created these datasets for their own research on "Behavior Chains"[4] and "Large-Scale Review"[5]

In order to begin analyzing this enormous assortment of data, we begin by splitting our 2.4 million books by genre, pulling out the 8 largest categories used by the database: 1) Romance, 2) Fantasy and Paranormal, 3) History, Historical Fiction, and Biography, 4) Mystery, Thriller, and Crime, 5) Children's Books, 6) Non-Fiction, 7) Poetry, and 8) Young Adult. By doing this, its easier to understand how more discrete genre collections change over time, mainly how genre approval has changed over many years. To create these genre sets, it was necessary to map a separate Goodreads genre collection onto the full Goodreads data catalog, as the large 2.4 million book file had no column information about genre classification. Once concocted, the

data could be split into the 8 genres described above.

	book id	title	authors	genres	pub day	pub month	pub year	avg rating	ratings count
1	6066819	Best Friends Forever	author id: 9212	romance: 23, mystery, thriller, crime: 10	14	7	2009	3.49	51184
2	6066814	Crowner Royal (Crowner John Mystery, 13)	author id: 37778	history, historical fiction, biography: 38, mystery, thriller, crime: 38	6	4	2009	3.93	186
3	7203847	The Tommyknockers	author id: 3389, author id: 105602, role: Narrator	fiction: 1205, mystery, thriller, crime: 330, fantasy, paranormal: 247	13	5	2010	3.48	45
4	1902202	Dead in the Morning (Patrick Grant, 1)	author id: 190988	mystery, thriller, crime: 23	1	12	1975	3.3	52
5	3209319	To Have and Have Not	author id: 1455, author id: 1004238, role: Narrator	history, historical fiction, biography: 12, mystery, thriller, crime: 9	1	8	2006	3.57	45

Table 1: The first 5 lines of the Mystery, Thriller, and Crime Book Dataset

The genre collections built include a huge combination of numerical, categorical, textual, and temporal data points. Columns describe book id numbers, titles, authors, genre classification, publication day, month, and year, average rating, number of ratings as well as another 20+ pieces of information per book. For our sake, however, the data frames have been limited to only the 9 necessary columns mentioned above, as they contain all the information needed to train any model and solve our problem at hand. Additionally, to process these subsets for usage, clearing all book entries that had no information about publication (pub) day, month, or year, was needed, as these would not be useful to for the study. To better understand the data frame's set up, a table of the first five columns of the Mystery, Thriller, and Crime dataset has been applied above. As you can see, the data is much easier to visualize in this fashion and now, using pandas data frame manipulation, can be analyzed.

4 Methods

ARIMA is the main machine learning model and predictor used to analyze the Goodreads genre datasets. ARIMA itself is a regression analysis which takes multiple changing variables and follows patterns to determine future expectation values. The acronym itself helps to further describe its process. Auto Regression (AR) describes a modeling where the output variable depends linearly on its own previous values. Integrated (I) has to do with calculating the difference between consecutive data points in a time series, allowing for a non-stationary time series of data to become a stationary one. Moving Average (MA) relates observations with residual errors, following trends and past data to create the forecasts in the future. Thus, we get ARIMA.

ARIMA has 3 parameters that deeply affect our output as well. Noted as p, d, and q (p,d,q), the input value for each of these throughout the rest of the report affects the effectiveness of ARIMA on each genre's forecast. P represents the number of lagged observations used during auto-regression for the model. Essentially, a larger p value builds a more complex model, obtaining finer details in the data. The downside, however, is that increasing p can also increase the possibility of over-fitting the data. The game is to rework the value of p in our training of the model and graphing in order to maximize the model's efficiency. Within the results section below, each genre is described as using a specific p-value. These values were gained, as described here, by reworking its value until the mean absolute error (MAE) of the model is minimized.

The second input parameter, denoted as d, describes the degree of differencing in the model. Differencing is the process of subtracting the previous value by the current value in a time sequence in order to make the data stationary (a constant mean and variance over time). So d itself represents the amount of times that the data needs to be differenced in order to make it stationary. A value of 0 means that the data is already stationary whereas a high value means that there is a strong trend over time which must be accounted for in the forecasting of the data. For our purposes, a value of 1 for all d, regardless of genre, is the best choice. To remove trends or seasonality within our data is fairly simple because there are clear and repetitive movements in month to month and year to year time periods.

The final input into ARIMA, called q, expresses the order of the moving average. Its value decides the number of lagged forecast errors that are used within the model. So a larger q value includes more past forecast errors for the predictions. By doing this, the model smooths each step in its prediction as it regresses along, with the goal of filtering out noise in a data set. In this study, the value will always be left at 0, as there is no clear noise which disrupts our forecasting and no need for averaging over multiple time observations.

5 Results/Discussion

5.1 Average Rating Over Time

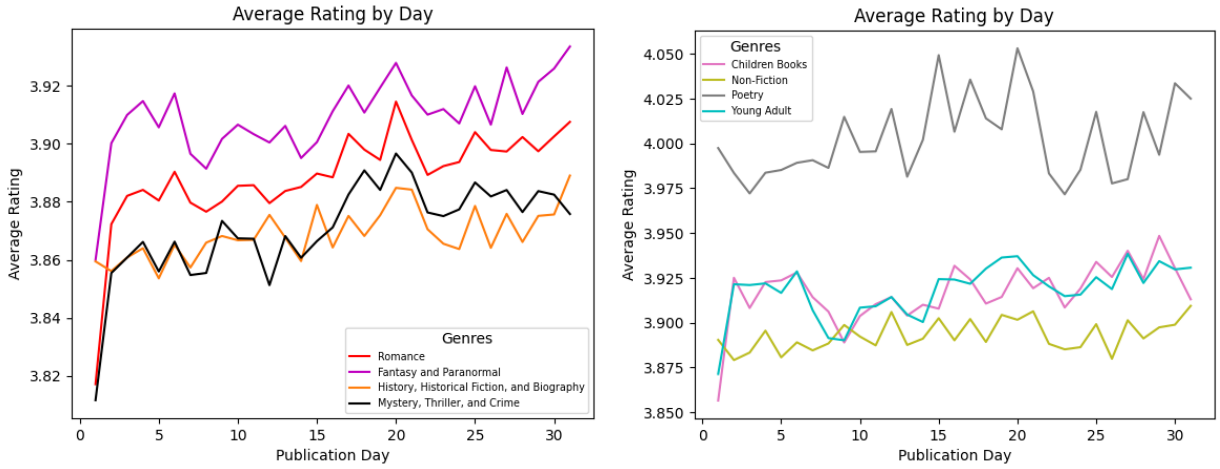

Figure 1: The average rating against publication day. Observe how the average rating seems to increase throughout the month in every genre. Additionally, each has peak ratings around day twenty.

The first goal of this report is to understand how the average rating of a genre changes over time. From the properties of our dataset, we can understand this through three time intervals, ratings by day, month, and year. Figure 1 displays all 8 genres, 4 genres on each graph for better visualization, for average rating against publication day. By looking at the plots, notice how, regardless of genre, the average ratings increases as the days of the month increase. Some increase more rapidly then others, such as the quick and continuous increase in the average rating for the Fantasy and Paranormal genre, but all indisputably have higher ratings on day thirty then on day one. Another interesting feature that appears is a spike, or at least a maximum, on day twenty for pretty much every single genre. The first of the depictions makes it quite clear the peak on this day of the month. Now, the actual reason behind this is fairly unclear. In searching for an answer, I came up empty handed, finding no investigation on this interesting trait. Perhaps more people start reading a book at the beginning of a month or perhaps there is some psychological explanation.

The second comparison is between the average rating and publication month. These graphs, indicated in Figure 2, provide more compelling information for the publisher. Similar to the average rating by day plots, the average rating by month also seems to increase from the beginning to the end of the year. A peak for each genre is achieved either in October or November. There are additional maximums in other places throughout the year too. Children's books have an enormous increase in the month of June compared to its surrounding points. This is most likely from the influx of personal reading for children over the summer, due to summer reading programs and more free time among the age group. Another notable feature is the significantly higher average ratings for poetry then every other genre. Poetry, while having the lowest number of engaged readers and books in the catalog, seems to outperform always. This seems to be from the fact that, while unpopular compared to other genres, poetry has an extremely dedicated reading base, meaning those who read the genre often are well-versed in much of the material. Those who care about the genre tend to enjoy reading it more, and thus rate it more highly across the board.

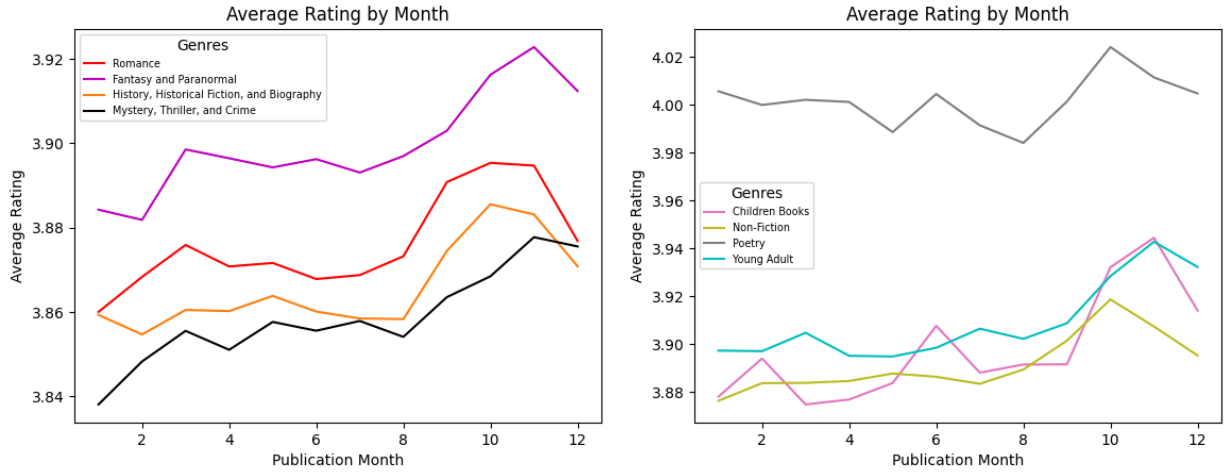

Figure 2: The average rating against publication month. Every month has peaks within either October or November. Poetry seems to have a significantly higher rating then every other genre in every month

The final comparison, and main one when looking at genre changes over time, is a genre's average rating by year. For both graphs in Figure 3, there is an incredible amount of noise which begins to level out between the 1960's and 1970's. This is purely a numerical inadequacy from the dataset because of the scarcity of books per genre before then. Since there are many more books after these decades, both that have been written and that have been engaged with by the public, the chaos in these graphs begin to fade. There are still interesting pieces of information that can be taken from this time period though. For example, although hard to spot with the genres overlaid, there is a significant peak in the fantasy and paranormal chart in the year 1938. By further investigation, it's discovered that both 'The Hobbit' by J.R.R. Tolkien and Disney's 'Snow White and the Seven Dwarfs' hit the market that year. For Snow White, it gained exceeding popularity because Disney had come out with Snow White the movie, its first princess movie, just a year prior. As for The Hobbit, not only are all The Lord of the Rings books hailed as masterpieces, but J.R.R. Tolkien is attributed with creating the entire genre of high fantasy himself. There are also peaks in the mystery genre around the start of the graph too. Agatha Christie, one of the most renowned mystery authors in history, published two books in this time, Murder on the Orient Express in 1934 and Death on the Nile in 1937. These books alone explain the incredible spikes in the genre at that time.

As more books began to be published and cataloged after the 60's, we begin to loose this easily noticeable 'classic book' tracker. The average ratings for every genre begin to stagnate, not moving more then a percent

or two each year. That is until 2010. From here, every single genre begins an upward trend, continuing to increase in average rating up until the end of the data in 2017. This feature, unlike the spike on day twenty for the average rating by day plots, is actually very explainable. For one, online viewing platforms, such as GoodReads, have allowed readers to more easily rate and review books. By having this tool, the audience can now choose what they read more carefully, taking into account thousands of other reader's opinions too. They may also be more likely to choose a previously highly rated book to read, only leading to a continuing increase in this average rating trend. A boom in self-publishing also plays into the shift. Since more people can publish, a wider range of genres and voices are available. People can now choose to read books with extremely specific messages and find content that more easily aligns with their interests.

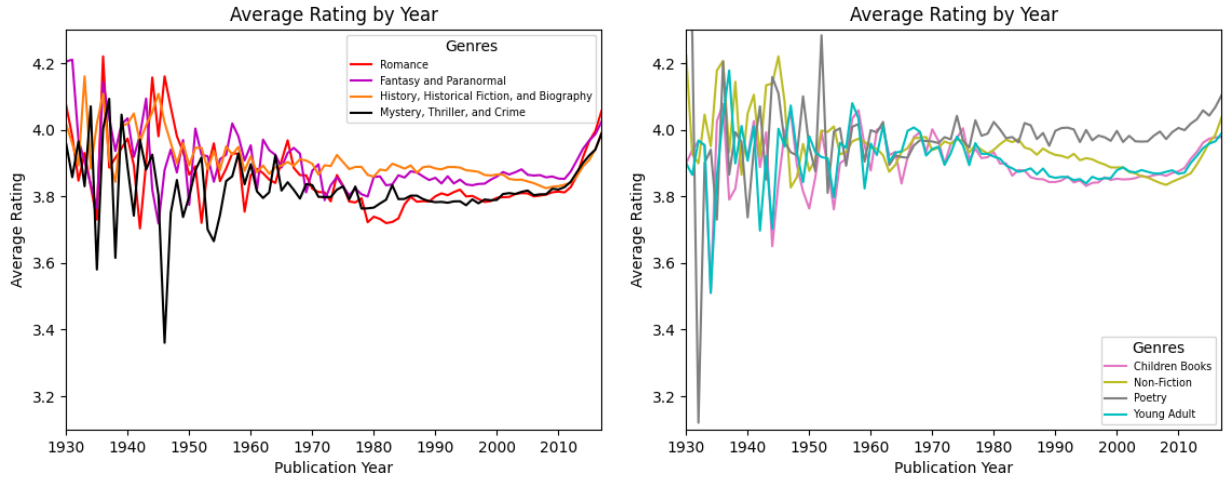

Figure 3: The average rating against publication year. Initial variability in each genre is do to lack of data. Graphs smooth as engagement and the number of books increases in the catalog. Notice how each genre begins an upward trend in its average rating around 2010.

5.2 Engagement Over Time

The second goal of this report is to analyze the number of ratings a genre gets per month. While the average ratings of a genre is very important, the actual number of readers engaged with it matters even more. It's a more direct approach to understanding how good a book does since the count of ratings directly translates into the number of people reading the book. And unlike looking at the average ratings by genre, where most of the trends are shared between them, the number of ratings given for each genre by month have extremely different trends. For example, when looking at Romance books, the most ratings come in during the month of February. This comes to no surprise as Valentine's Day and thoughts of love are common in the month. Poetry exhibits a strong peak in April. This should also be thought of as reasonable as April is designated as National Poetry Month here in the US. Schools often teach the subject in this month, library outreach focuses on it, and bookstores host events to promote the genre.

There are some trends that share similarities to our average ratings over time section though too. October again has peaks in engagement throughout every genre. When investigating this situation, we find many understandable explanations. Firstly, October is National Book Month, a period where reading is increasingly promoted. The weather is also a driving factor, as the turn from outdoor activities to indoor ones leads people to their bookshelves. Also, most national book awards are given out during the month, such as the National Book Awards, so publishers and authors will release soon before the awards to keep the title fresh in people's minds.

Now that we have analyzed this data for each genre, we can begin a process of predicting their engagement in the future. This will be done using an ARIMA model, as described extensively in Section 4 - Methods. To prepare the data for the model, we cut the data into two new subsets, one which keeps only the last ten years of data and one which keeps only the last two years of data. Any more then ten years and the regression relies too much on data that is no longer applicable. Any less time then two years and the regression doesn't have enough data to visualize yearly trends. Once done, publication day, month, and year columns can be reworked into one continuous time series, a necessary step to allow ARIMA to regress on the data from month to month.

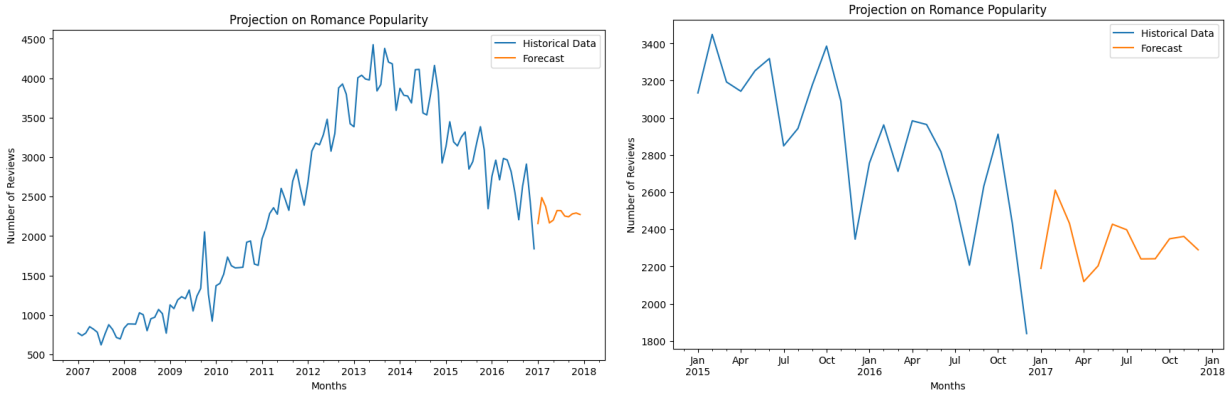

Figure 4: These graphs show the projection on romance novels for the number of reviews it should expect per month. The first displays a ten year look at engagement for the genre and the second displays a two year perspective to more easily resolve month to month patterns.

We begin by training the ARIMA model on the two year Romance genre data set, shown in Figure 4. After splitting our data into train and test, we run the model on the train set, keeping the ARIMA parameters (p,d,q) as (3,1,0). We then apply the model fit to a forecast using the length of the train set. To understand how good our model fits the data, we will use metric called the mean absolute error (MAE). What this metric does is it collects all the differences between the predicted value for a month and the actual value for the month, taking into account the number of data points, and outputs a numeric value. The output describes the expected deviation any predicted point may have from its actual value. So if there are 1000 predicted reviews in a specific month and the MAE is 100, we would expect the real number of reviews for that month to be off from the predicted by 100 reviews. Now, with a metric to analyze how good a model ran, we can take the MAE and divide it by the average rating count, returning a percent error for the model. This error tells us how far any given month's actual rating count may be from the predicted rating count for that month.

With the entire model now built and a process by which we can see the goodness of the model, we can begin changing the parameters, specifically the p-value, in order to minimize out percent error. This process was simply one of trial and error, trying values between 1-10 until the minimum percent error was revealed. Actually finding it was easy, as all p-values continuously sloped towards the minimum percent error when changing them. For Romance, a minimum was achieved when using a p-value of 2. This reveals to us that the Romance genre took very little complexity to explain the data as best as possible. Any greater p and you begin over-fitting the model on the data, any less and you begin missing some of the trends that Romance novels exhibit.

	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

Table 2: Describes the mean absolute error (MAE) when an ARIMA model was trained and tested on each specific genre subset. It also shows the percent error that the MAE represents, comparing the MAE to the average review count of a genre over the two-year time period in question.

Everything works now: Plotting the data, applying ARIMA to regress on the data, gaining information about the goodness of fit for the model, and refining the parameters to achieve the best percent error. All there is to do is reapply this process to each of our 8 genres. Once done, you can begin to compare the results. Table 2 displays all this information, including the average review count of the genre per month over the past two years, the optimized p-value found, the MAE as a result, and the calculated percent error from our model. Notice how most of the genres produce fairly good models, keeping the error around 15 percent or less. And while this doesn't seem all that good, for most data points, a 15 percent increase or decrease still keeps it in the position it previously was, as in above, between, or below the points around it. There are, however, a few straying genres. The most obvious is poetry. Due to its low engagement, a lack of data calls for a very hard prediction model. The model itself follows the poetry monthly trends pretty closely, but a small average rating count makes it hard to be sure. Children's books do fairly poorly as well, although not nearly as bad as poetry. Again, it seems a lack of information and engagement led to percent errors that were sub-optimal. These genres do share a couple features in common that explain some of the regression issues. When looking at the average ratings by month in figure 2, both Children's books and poetry have abnormal peaks within February and particularly within June. Their number of ratings per month also changes more drastically from month to month, making it harder for regression to map trends and smooth noise.

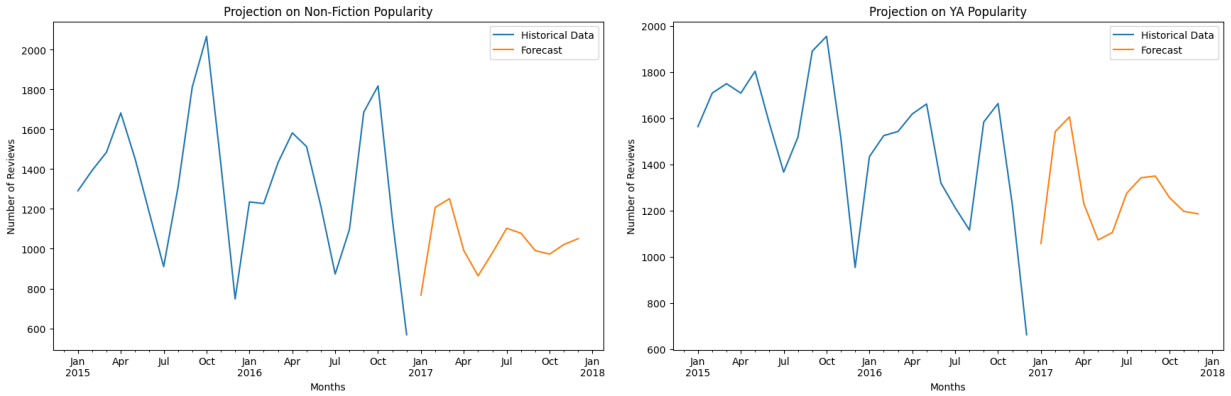

Figure 5: These graphs compare the projection of young adult books to the projection of Non-Fiction books. While both look similar, Non-fiction books minimizes percent error with a p-value of 2 whereas young adult books minimize percent error at a p-value of 7. This contrast reveals how similar looking data can need a different set of parameters. It also helps to show the complexity of the ARIMA model, and how even slight changes in the inputs can affect the necessary parameters for a good output.

6 Conclusions/Future Work

Within this report, we addressed two main ideas, understanding the average ratings of genres over different time periods and forecasting genre engagement into the future. The first teaches us about when genres have been popular in society. The day to day ratings of books helps to reveal a monthly cycle of the human experience, how reading in different parts of the month affects the audiences absorption of the material. Looking on a month scale we see a similar trend, that later in the year, like later in the month, is a better time to read. Then, when extending our view to a yearly basis, we can see how a genre has developed over the last century. We can visualize how specific books may have affected the early development of a genre and how the creation of recognized 'classics' spike the rating of a year. It's also important to recognize that, while engagement is different by genre, the average ratings patterns of different genres are similar. So even though poetry gets higher average ratings then the rest of the genres, it too follows a fairly constrained average rating pattern, peaking on the twentieth day of the month, peaking in October, and curving upwards post-2010. These patterns teach us about when its best to publish, when people are most likely to rate your book highly, and what the general public thinks about a genre.

The second focus of this report, on forecasting genre engagement, teaches us much about both publishing and society too. By understanding the periodic tendencies of different genres, publishers can decide which genres are best to publish and when. Romance books peak in February, Poetry in April, and Children's Books in June. These peaks are not random or unexplainable either. The influence by society, through designated genre months or school curriculum, and the influence by the time of the year, such as the holidays or seasonal changes, significantly affect a genres popularity. So we analyze these features and adapt techniques, such as the ARIMA model used in this report, to predict how well a genre will do in the future. And by running ARIMA across different genres, with different p-values, we gain insight into how the model works and how we can best map out future engagement. It tells us how the size of a data frame affects the auto-regression on the data. It teaches how to predict most genre engagement to under a 15 percent error. Now, of course whole this process benefits the publisher, but it also helps authors to understand what's popular in society and how releasing their book at different times can affect how it's received. In full, and as an iteration from the start, this report is for the authors, not the readers. It's written in hopes that authors can better understand their audience, rework how they think of the business aspects of their career, and improve both the rating and engagement of their future projects.

For future work on the subject, much can be done to improve. For starters, with more time, one could dive deeper into the shifts of average genre ratings over the past century. Peaks early in Figure 3 can be extracted and their influence on the genres development understood. As these plots begin to smooth out, smaller peaks and troughs can be drawn out for similar analysis. Genre approval can also be looked at, seeing how other media, such as movies, affected the rating of its book counterparts. The biggest rework to this report could be to continue to work on engagement over time. Other models could be built and ran parallel to ARIMA, comparing their effectiveness to represent the data. More metrics could be used to understand the goodness of each model, in the hopes to refine the percent error of the data. Additional genres could be used in the analysis and other data sets beyond just GoodReads could be taken to see if the results found from their site are consistent across other online rating and reading platforms. Much of this improvement just requires more time but I'm sure as online databases grow and the book market continues to expand, the need for understanding future engagement will only grow.

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.

[4] Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys'18. [bibtex]

[5] Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL'19. [bibtex]