Business Analytics Report

How can the publishing industry leverage analytics on new releases to forecast and increase sales?

September 2021

Coraline Duval
CID: 01930178

Word count: 4965

Abstract

While books and writings were one of the main entertainment sources up to the start of the 21st century, they are today not able to compete with social media and streaming platforms in today's attention economy. One of the reasons is that publishing industries are not leveraging consumers data at nearly the same level as their other media competitors.

This report explores two stages of the release of a book to understand and predict a book sales performance through its popularity. First, during the pre-release period, the publisher has control over some of the book features and can, using data, optimize them to maximize the popularity of the book. Then, during the post-release period, a publisher can use user-generated data around the book such as ratings or reviews to predict the overall performance of the book.

Through linear regressions, it was found for the pre-release features that neutral colors for the covers, a shorter title conveying a positive sentiment, the publication in March or October, and a form of eBook format led to higher sales. For post-release, using reviews in the 3 months following publication, an average positive sentiment conveyed by the text reviews is linked to higher sales volumes.

These results show how data can be used by publishers to make their products, books, more appealing to customers, and how it can help them grow their sales. One main limitation of this report was finding sources of data, especially data regarding book sales. While publishers might feel like protecting their data advantages them against other publishers, they should consider grouping forces to help the industry as a whole reduce the gap they have with other entertainment media on the uses of data and artificial intelligence in their businesses.

Table of Contents

Introduction

Books and writings have been a continuous part of the media scenery since the 11th century (Briggs & Burke, 2009), first through the clergy, but democratized over time, thanks to the invention of the printing press. In the last centuries and decades, media has diversified itself from radio and television to more recently social media. By the start of the 20th century reading newspapers and books were a large source of daily entertainment, for both the working and upper classes (Smith, 2021). With the rise of social media in the 21st century and the use of data-centric algorithms to promote them and maximize user engagement on the platforms, reading seems to have lost some of its attractiveness as an entertainment possibility. This is further validated as the percentage of Americans having read at least one book decreased by 10% between 1999 and 2014 (Weissmann, 2014).

The publishing industry has made efforts to adapt to new consumption patterns, with e-readers becoming mainstream from 2007 with the release of the first Amazon Kindle, but also through audiobooks and the use of social platforms to promote books, as well as the creation of new formats, such as for the book The Little Girl Who Lost Her Name, a book that could be digitalized to add the name of the child reading the book (Belton & Wall, 2015). However, they have not yet embraced the use of artificial intelligence and machine learning to the same level as social media and streaming platforms have which gives them a disadvantage in today's attention economy. A report for the Publishers Association (frontier economics, 2020) writes that: "AI investments in the sector has just begun" and "the sector must overcome a number of investment barriers, including … general awareness of the benefits of AI". As social media and streaming platform personalize their content to appeal to their users, such as Netflix which personalize the artwork for a movie or a show depending on the user's preference (Chandrashekar, et al., 2017), publishers could use a similar approach to understand how to market their books best to optimize sales and become a competitive media against with social media and streaming platforms.

This report will explore different ways publisher can leverage analytics and different machine learning techniques when releasing new books to ensure it is as popular as possible and sells appropriately. It will more specifically focus on what the publisher can have an impact on before publishing, such as the title, description, date of publication, format, and book cover. The goal is to understand, once a book has been written by a certain author, with already a fixed genre and a fixed number of pages, what the publisher can do so that this book becomes as popular as possible, as other media publishers such as streaming platforms or social media do to showcase their content to maximize user engagement. The report will then focus on the post-release time of a new book and use ratings and natural language processing on user reviews written close to the publication to understand their role in book sales. The two processes are differentiated as they represent two time-separated stages of a newly released book. The first stage should thus be independent of the second one to not create bias in the analysis.

Literature review

As one of the major media, and as more advanced techniques of machine learning developed, books and book sales have been at the center of an increasing number of studies.

Wang, et al. (2019) develops that fiction and non-fiction have different patterns in book sales, and uses feature importance analysis to find that author, publishing houses and genres play an important role in understanding how books sell. This is reinforced by Schimidt-Stölting, et al. (2011) in a study focusing on Hardcover and Paperback Editions in Germany that found, thanks to a regression model, that the performance of hardcover and paperback editions is driven by the popularity of the author and the publisher, specific genres, and book cover designs. Specifically, this research found that book covers can have conflicting results whether they are in paperback or hardcover, reinforcing the idea that the format of the book plays an important role in how a book should be marketed and published.

Yucesoy, et al. (2018) select the specific case of New York Times bestsellers books and finds that fiction sells more than non-fiction, which goes against the industry trend that non-fiction sells better than fiction books (The Publishers Association, 2019). The research also finds that popular fiction authors tend to have multiple popular books, thus making the author an important feature when predicting book sales. This research also shows that there is a link between early sales performance and long-term sales performance, leading to the separation of the two in the later stages of this report.

In a study of the factors influencing sales of new book release (d'Astous, et al., 2006), the idea developed by other research that the reputation of the author and the publisher, as well as an attractive book cover, are central factors to get buyers to buy new releases when in a bookshop is reinforced. However, this research delves deeper as it finds that the author's reputation is an important factor in increasing sales only for technical non-fiction books, which shows that nuance is needed to explore this topic, and understand book sales per genres, format, authors and publishers, as sales results can differ between them all.

Finally, Chevalier & Mayzlin (2006) note that while a good review on a book-selling website seems to increase sales, cross-referencing between websites can also be done and its impact would be harder to study. It also shows conflicting results on which star rating (1 or 5 stars) leads to an increase in sales, as an added 5-stars rating can increase sales, but the overall number of 1-star ratings seems to have a bigger impact. It also notes that reviews have a higher impact than ratings on the number of books sold, even though they require more implication from customers.

I. Data acquisition and pre-processing

A. Book metadata

To retrieve data about newly released books a few datasets were available.
Kaggle offered a few options but generally few features were included in the datasets and none of them included review text data which would be an issue for the post-release analysis.

Goodreads, a book social media website where users can interact with each other and their readings, previously offered an API but stopped providing new keys in December 2020. However, users' public shelves were scraped in 2017 for academic use (Wan & McAuley, 2018) (Wan, et al., 2019) and can be found online (Wan, 2019). This data is very complete and includes:

- 2,360,655 books and their associated:
  - 829,529 authors,
  - 23,606,655 genres,
  - 400,390 series,
  - 1,521,962 works (i.e., count of books without counting different editions)
- 15,739,967 reviews,
- 228,648,343 user-book interactions (if the book is read if a rating was given or a review was left by a user about a certain book).

Detail about the features can be found in the Appendix.

Apart from book sales, this dataset included all the features needed for the analysis and offered the luxury of having a large number of books, which could be filtered as needed for this analysis without ending up with a sample too small.
The books include all types of genres, from fiction to non-fiction, through comics and coloring books, and are published before December 2017.

B. Sales data

Finding book sales data proved to be more difficult.
Book sales can be defined through different metrics. For physical books, sales can be measured through (Michel, 2016):

- The number of copies of a book printed,
- The number of copies of a book that were sent to stores,
- The number of copies sold to readers (hard to measure as some books sent to store will be returned),
- The number of copies as recorded as sold by Nielsen BookScan.

The first three are not shared by the largest publisher today as they are most likely considered data that competitors could use to their advantage. The last, Nielsen BookScan uses book ISBNs to record when a book is sold through different retailers such as Amazon, Barnes & Nobles, Apple, and many more retailers throughout the world (Nielsen Bookscan, 2017). It is generally considered that the BookScan number underestimates the number of sales (Michel, 2016). Moreover, BookScan is marketed for professional use and costs seem to be a minimum of £2,000 to access book sales data and were thus not a possibility for this report.

Book sales can also include eBooks or audiobooks sales, which are also not widely available to the public, for similar reasons as for physical books.
No data that could be scraped or dataset already available was found giving sales numbers for a wide array of books (more than 1,000 books), thus alternatives needed to be explored.

Kaggle offered a dataset containing Amazon sale rank data for print and Kindle books (Lurig, 2018), which is the ranking of the book in each of the categories. This ranking is based on the number of copies sold by Amazon in the last few hours or days. Amazon Standard Identification Number (ASIN), which are amazon product numbers was how each book was identified. The Goodreads data also contained ASIN, even though only 20% of books had an ASIN recorded, and when the Goodreads data and amazon ranking were joined through the ASIN, only 930 books were left, which was considered not enough for analysis (as no filter had yet been applied to the type of book). The two datasets were also joined through a book title match (as 100% of books in the Goodreads dataset had a title) but still only 1,478 books resulted from the join.

Finally, Google Trends was selected as an alternative to book sales data. It would offer the possibility of searching most book titles in the Goodreads dataset, the possibility to select the category 'Books & Literature' to narrow down the search, and the possibility to search for monthly data since 2006, offering a large timeframe for the analysis.
However, for books with similar titles (as for series, or simply two authors giving the same name to a book), results would not be optimal as no separation of sales between the two could be done. It would also not be a direct translation of sales numbers, as Google Trends only express what Google's users have searched online, which could have led to a sale or not. It also does not take into account people who directly go to the e-commerce website of their choice to make a book purchase. Nevertheless, Google Trends translate an interest in a book, which would then lead to sales. While not proven, it is generally believed that a higher Google Trends number is translated to a higher number of sales, and the opposite would also hold. In the rest of this report, Google Trends data will be considered sales data.

Finally, while publishers consider book sales data a competitive advantage that they don't want to share with the rest of the industry, they might want to reconsider that stance. Social media and major internet companies make some of their data available to the public (and thus their competitors), which can be argued, has allowed for collective research and knowledge to be furthered and benefited those companies in the long run. Moreover, a paper by Botelho (2018) details the case of buy-in in the investment industry that sharing knowledge among competitors added more value overall. A parallel could be drawn between an uncertain buy-in in investment and the signing of an author for the publication of a book. Sharing knowledge, including sales data, between publishing houses could enable each publisher to make more informed decisions, thanks to a wider array of data available, which would then lead to an increase in value for all the publishing houses and give them a chance at competing in the current entertainment sphere.

C. Data pre-processing

To process the data, spark, and most specifically pyspark was used for the first loading of the data and pre-processing. Indeed, spark can process a large amount of data, while the pandas module of python did not manage to load all the data points within a classic RAM computer capacity.

Books from Goodreads were chosen as published between January 2010 - when Google was already widely popular and it became relatively widespread to buy books online, which would create good quality Google Trends data - to December 2017, where the Goodreads data ended. Only books published in English were selected, as this would ensure that most of the titles, descriptions, and reviews would be analyzable in English. It was also decided to focus only on first editions, as to try not to have too much Google Trends data from the publication of different editions, and also to be able to consider 'newly released' books as books unknown to the public previously. It was decided to select books that had more than 5 reviews on Goodreads, as to not have too many self-published books, not backed by a publisher, that could skew the results of the analysis. Finally, it was decided to select only a few genres of books, that would fit the entertainment type of books that could replace time spent on social media or streaming platforms. The selected genres are fiction, fantasy and paranormal, mystery, thriller and crime, romance, and young adult. The excluded genres are poetry, non-fiction, comics and graphic, children, history, historical fiction, and biography. The genres were not defined by Goodreads, but by user's popular shelves and thus, most books had multiple genres. The main genre for a book was selected as the genre being added to the most shelves and the filtering was done on that main genre.

On the resulting books (about 36,000) I reformatted the titles to exclude series numbers or shorten them as somebody might do when searching for the book on Google as Google Trends tended to work best on shorter titles. Finally, I regrouped the monthly resulting Google Trends data into the sum of Google Trends over the first 6 months after publication and then from 6 months to 3 years, to differentiate between short-term and long-term sales data.

The resulting 19,759 books (excluding those without any Google Trends results) were then selected through pandas for the analysis.

II. Pre-release analysis

A. Features and model

To understand the impact a publisher's choice can have on the popularity of a book and thus its sales, a few aspects that were available in the Goodreads data were studied. The format of the books regrouped under 7 formats:
- Kindle edition,
- eBook (including other eBook brands except Amazon's Kindle),
- paperback,
- hardcover,
- audiobooks (covering all types of audios),
- online free books,
- mixed format (such as library binding or both eBook and paperback).

As detailed in the previous part, these are the format of the first editions, released first to the public. It is the less viable feature as, especially as eBooks become more common over the years, physical and eBook formats can be released at the same time, which Goodreads does not account for. Overall, the dataset has 59.8% of eBooks (Kindle and eBooks in the format above) against 32.5% of paperback and hardcover combined, which might not be representative of

The date a book is published is also generally a publisher's decision. This report will not try to identify seasonal books (Christmas-themed for example) but rather looks more globally if some publication dates (through month only for simplicity) seem to lead to an increase in sales.

The book title and description are two features a publisher will have at least some control over. While an author might prefer a certain title, optimizing the title, the first this with the cover that a potential customer will see about the book, is essential. Similarly, descriptions can take different forms and tones, and thus is an important feature to the release of a book. More specifically, two features are the length of words in the title and the description which give a first insight into each of these. While it could seem that shorter titles grab the customer's attention more easily, it will be interesting to confront this hypothesis to the dataset. Then, punctuation and common words that are generally neutral in sentiment analysis were filtered out before finally taking the lemmas of all the words left. On these, the sentiment intensity analyzer from the nltk library was used to determine if the titles and descriptions conveyed a positive or negative emotion to the potential customer.

Finally, two features were extracted from the covers. First, the position of the title and author name on the cover. While it seems from empirical testing that this text takes generally most of the width of the page, it is generally placed at different levels on the cover: at the top, the center, or the bottom of the cover. Generally, the text is placed in a combination of these positions (top and bottom or top and middle for example), thus they are not mutually exclusive. The position of the text was extracted using the urllib.request and PIL libraries to get the images from the URLs given in the dataset and load them and the position was extracted using the pytesseract library, which is a Python wrapper for Google's Tesseract optical character recognition tool. Finally, the main colors in the cover were extracted from the cover using the ColorThief and

colorname libraries, which get the three principal colors from the covers and convert them to general colors name so they can be grouped and compared without the issue of having specific RGB codes or obscure color name. While the extraction of the colors seemed to work well, the extraction of the text was not quite as optimal, mostly due to the very poor resolution of the cover images.

Using all the features developed, a linear regression was performed on the dummies variables version of the features using the sklearn library. A random train sample of 80% was selected from the dataset. While for categorical analysis, the Google Trends data of the first 6 months and then up to 3 years after publication were used, for the linear regression, the prediction parameter will be the total of the two to make the result more comprehensive.

B. Results

The multiple linear regression on all the detailed features from the previous part led to an R^2 score of 0.35, which shows that these features alone are not enough to get a good prediction regarding the popularity of a book. The details of the OLS Regression results can be found in the Appendix. If a p-value of 0.05 is selected, then 7 features would not be statistically significant. These include the number of words in the description (p-value=0.056) as well as the sentiment of the description (p-value=0.062), the three different levels of the text on the book cover (lower: p-value=0.108, middle: p-value=0.445, higher: p-value=0.892) which can be explained by the fact that the quality of the extraction of the position of the text was not particularly accurate from the low resolution of the image (it seemed that some text was not detected or that sometimes text was detected in the wrong area). The color 'magenta' in the cover colors has a p-value of 0.063 and the Hardcover format has a p-value of 0.062, which from running a correlation matrix does not seem to come from any correlation between the features.

A second linear regression without the text position features led to all the other features lowering their p-values (only 'magenta' and the sentiment of the description were still not statistically significant with p-values of 0.057) however the R^2 score stayed the same. The root mean squared error was also calculated and was for both iterations of the model the same: 682 for the train set (80% of data) and 677 on the test set, which seems rather high values that demonstrate that the model is not optimal.
Nevertheless, some learning can be done from the features coefficients:

Both white and grey on the book cover seem to lead to higher sales (respectively of coefficients 128 and 112) than other colors, with yellow having the lowest coefficient of all the colors (coefficient = 34). It thus seems like neutral colors sell best.

While the coefficients are high for the months published, they are all in a similar range (between 155 and 220) which does not lead to a clear recommendation as to which months to publish to get higher sales, though March and October (respectively of coefficients 220 and 207) seems to be a better month to publish a new book than December and September (respectively of coefficients 158 and 157). To understand the role of the month of publication better, the graph below was drawn, which shows the average sales (divided over short and long term) by month by each type of genre of books. Thus, while March is a good month for romance books, it is not the best

month to publish a mystery, thriller, or crime book and October seems a good month for sales up to 6 months after the release over all genres. June and April also seem like particularly good months for young adult books.

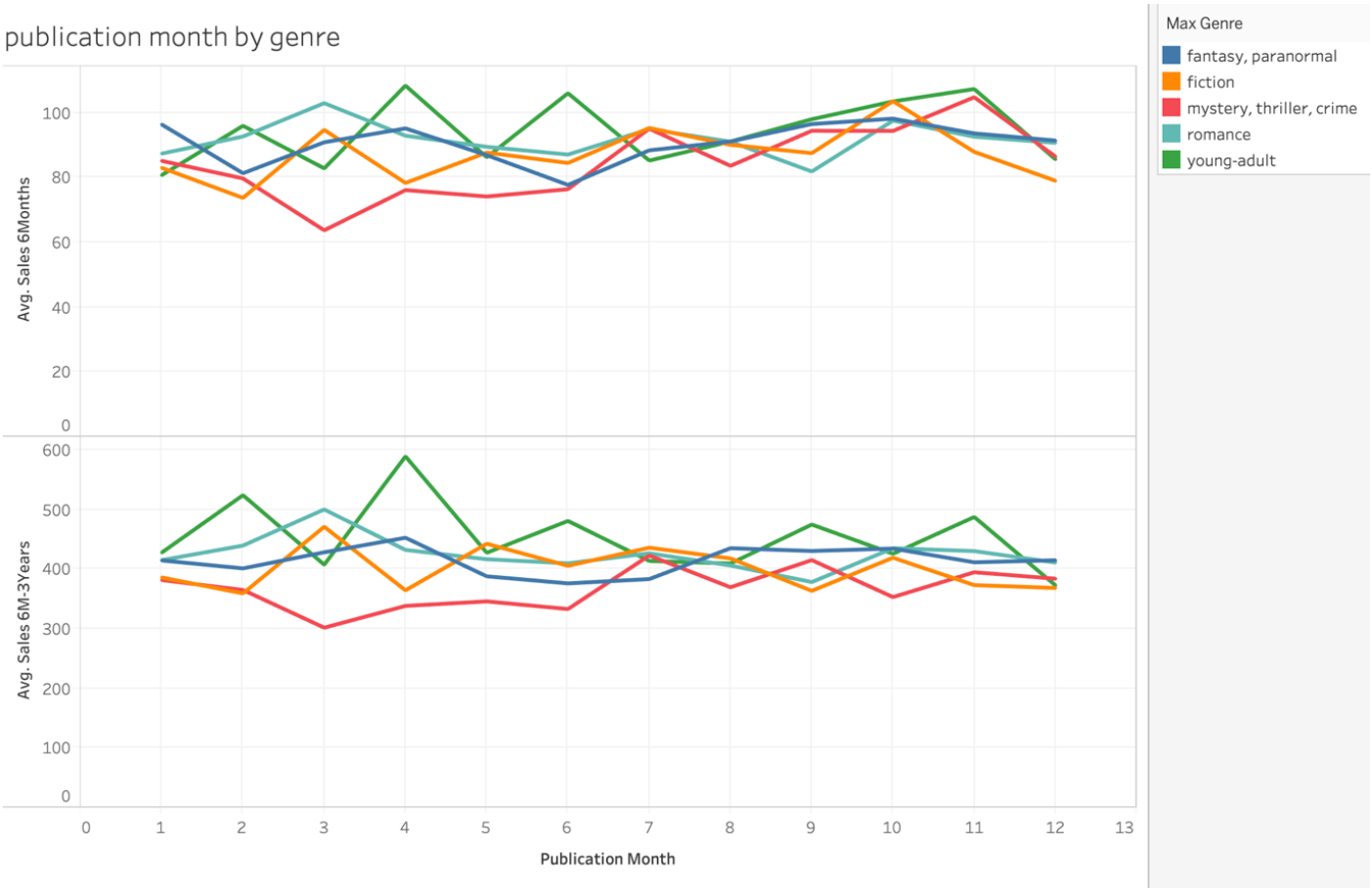

Figure 1: average book sales per month per genre

It can also be noted that the longer a title is, the less a book will sell, though is a smaller proportion than the impact of previously discussed features (coefficient = -5.4) and the length of the description does not seem to have an important impact on sales (coefficient = 0.13). Similarly, having a positive title had a much more positive influence on sales than having a positive title (respectively of coefficients 122 and 25).

Finally, the format leading to the most sales is the 'Kindle Edition' (coefficient = 123), while paperbacks lead to more sales than (coefficient = 54) hardcovers (coefficient = 8.6), which could be explained by pricing reasons, as paperbacks are generally cheaper than hardcovers. Figure 2 below also shows that for short-term sales, hardcovers sell more for books more than 500 pages, while paperbacks sell better for short books under 100 pages. Kindle Editions and eBooks are quite consistent over all genres. Audiobooks also perform well on longer books above 500 pages, which might be easier to listen to than to read. Figure 3 also shows that audiobooks sell

much better when the book is part of a series, while hardcovers perform better on books that are not part of a series.

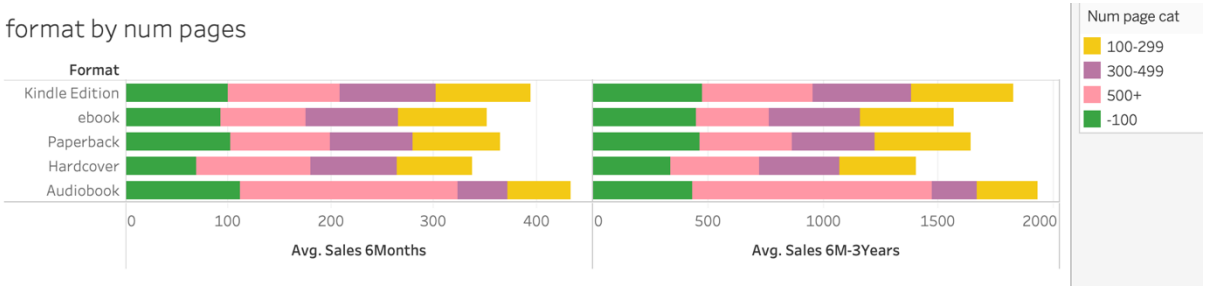

Figure 2: Average book sales per format per number of pages
Note: the total averages are less than for figure 3 because of NaN values in the number of pages.

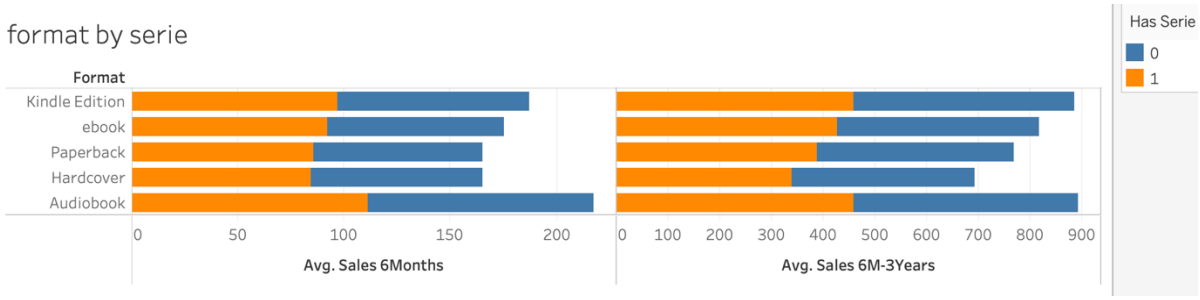

Figure 3: Average book sales per format whether the book is part of a series

III. Post-release analysis

A. Features and model

For the post-release analysis, the reviews and interactions datasets were the main datasets used. The former contains all the written reviews of the users, as well as if the books were read (and when), the rating given by the user, and how many votes and comments each review has. The latter contains the rating given to the book, if the book was read (and when), as well as a part of the written review if there is one. Thus, all the reviews are included in the interactions. Moreover, a rating of 0 indicates that the user did not give the book a rating.

All the reviews and interactions for the books in the datasets were selected up to 3 months after the publication of the book. 3 months allowed for a reasonable amount of time for customers to buy, read and write a review about the book, without allowing for too much word of mouth yet. The goal is to understand if those "early adopters" of the book are representative of what the public will think of the book (and thus how popular it will become). This filter reduced the reviews to 138,279 reviews with on average 12 reviews per book and the interactions to 1,344,629 with an average of 89 interactions per book.

To the methodology used in the previous section, the text reviews pre-processing also added filtering on whether they contained any emoji (Rose, 2021). As emoji convey emotions to the users, it is important not to ignore them and take their meaning into account. Then the same Sentiment Intensity Analyzer process was used to determine the sentiment of the review.
Then, reviews were regrouped by book over the number of reviews of positive and negative sentiments, the average sentiment score for the book, the average length of the reviews for the book, the number of interactions where the book was read, and the number of reviews from a score of 1 to 5.

Finally, these 10 features will be used in a linear regression model to predict the total number of sales over the 3 years after the release.

B. Results

The multiple linear regression led to an R^2 score of 0.21, which shows that while reviews and user-driven action in the 3 months following the book release can help predict some of the sales but would not be enough to have reliable predictions of sales. The detailed results of the first regression can be found in the Appendix. At the 5% significance level, the number of ratings of 1 and 2 stars and the number of positive reviews were not statistically significant features, and the regression was run again without them. Computing the correlation matrix between the number of positive reviews and the average sentiment of the reviews showed a low correlation between the two but a high correlation between the number of positive reviews and the number of ratings of 5 stars. The R^2 score did not change, and the root mean square errors for the train (80% of the data) and the test data were 748 and 771, which is, as in the model for the pre-release predictions, too high to make accurate sales predictions. Regarding the importance of each feature, it seems that the more users read the book,

the lower sales go (coefficient = -1.95), which seems counter-intuitive, but could be explained by the fact that sometimes ratings are given while the book is not noted as read (either by user mistake or by users rating books that they have not read). A rating of 3 stars seems to have a bigger impact on sales than ratings of 4 and 5 stars (coefficients respectively 4.3, 1.6, and 2.1), which also seems surprising. A potential explanation could be that some of those ratings are accompanied by a review that led to insights to potential customers. Finally, while a high average sentiment score seems to have a very large importance in increasing sales (coefficient = 418), having many negative reviews also seems to have higher sales, though in a lesser proportion (coefficient = 7). Overall, the biggest factor in increasing sales is having a high average sentiment (thus positive) among all the reviews of the book. By running a regression exclusively on this feature, it showed that it explained 19% of sales (as the R^2 score is 19), a decrease of 2% from the full model used at the start of this section. Thus, post-release sales can be explained mainly using the average sentiment of all reviews for a book in the first three months after publication, while ratings don't seem to have a high impact of sales.

Conclusion and recommendations

Publishers can have an impact on the popularity, and thus the sales, of a book about to be released by choosing neutral colors and avoiding yellow as the main colors on the book cover. While the placement of the title and the author was inconclusive, a shorter title conveying a positive sentiment can increase sales. Publishing in March and October also led to better sales than in September and December. Finally, offering the book in a Kindle or eBook format, and as a paperback rather than hardcover would also benefit sales. Considering reviews in the 3 months following publication, the main driver of sales is a generally positive sentiment conveyed by the text reviews. Finally, publishers should also use the nuances in the results of the data if additional filters are first used, such as a book's genre, number of pages, or even through information about the author or the publisher itself.

Better sales data and book metadata would lead to clearer conclusions, and the predictions models could be refined further, either through using different models or through more specific book features, that capture how a book is marketed more clearly. Google Trends data could also be used in further research as a feature expressing the popularity of a book.

Nevertheless, this report still provides publishers with guidance on how to use data to market their books to attract as many customers as possible in the quest of becoming a media able to compete with social media or streaming platforms.

While publishers might currently start using their data for similar studies, it would benefit the industry as a whole if data would be shared so that further research could be done to understand customers' preferences and behaviors regarding book consumption for entertainment, which would, in turn, most likely benefit all publishing houses. While Goodreads, an Amazon owed website is currently not pushing for data insights, other similar actors are developing, such as the StoryGraph where users can have access to statistics regarding their book consumption and could pave the way a better data usage in the publishing industry.

References

Belton, P. & Wall, M. (2015) Did technology kill the book or give it new life?. *BBC News Technology of Business*, 14 August.

Botelho, T. L. (2018) Here's an Opportunity: Knowledge Sharing Among Competitors as a Response to Buy-in Uncertainty. In: *Organization Science 29 (6).* s.l.:Informs, pp. 1033-1055.

Briggs, A. & Burke, P. (2009) Introduction. In: *A Social History of the Media.* Third Edition ed. s.l.:polity, p. 9.

Chandrashekar, A., Amat, F., Basili, J. & Jebara, T. (2017) Artwork Personalization at Netflix. *Netflix Technology Blog*, 7 December.

Chevalier, J. A. & Mayzlin, D. (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research,* 43(3).

d'Astous, A., Colbert, F. & Mbarek, I. (2006) Factors influencing readers' interest in new book releases: An experimental study. *Poetics,* 34(2), pp. 134-147.

frontier economics (2020) *People Plus Machines The role of Artificial Intelligence in Publishing,* s.l.: Publishers Association.

Lurig, M. (2018) *Amazon sales rank data for print and kindle books.* [Online]
Available at: https://www.kaggle.com/ucffool/amazon-sales-rank-data-for-print-and-kindle-books/version/3?select=amazon_com.csv
[Accessed 13 August 2021].

Michel, L. (2016) *Everything You Wanted to Know about Book Sales (But Were Afraid to Ask).* [Online]
Available at: https://electricliterature.com/everything-you-wanted-to-know-about-book-sales-but-were-afraid-to-ask/
[Accessed 10 August 2021].

Nielsen Bookscan (2017) *BookScan Panel UK.* [Online]
Available at: https://www.nielsenisbnstore.com/documents/BookScanPanelUK.pdf
[Accessed 20 August 2021].

Rose, C. D. G. (2021) *A Tutorial on Performing Sentiment Analysis in Python 3 Using the Natural Language Toolkit (NLTK).* [Online]
Available at: https://medium.com/geekculture/a-tutorial-on-performing-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk-40e5b35ab440
[Accessed 25 August 2021].

Schmidt-Stölting, C., Blömeke, E. & Clement, M. (2011) Success Drivers of Fiction Books: An Empirical Analysis of Hardcover and Paperback Editions in Germany. *Journal of Media Economics,* 24(1), pp. 24-47.

Smith, B. G. (2021) Life in the Modern City. In: *Europe in the Contemporary World, 1900 to the Present.* s.l.:Bloomsbury Publishing.

The Publishers Association (2019) *Yearbook 2019,* London: The Publishers Association.

Wang, X. et al. (2019) Success in books: predicting book sales before publication. *EPJ Data Science,* 31(8).

Wan, M. (2019) *UCSD Book Graph.* [Online]
Available at: https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home
[Accessed 3 August 2021].

Wan, M. & McAuley, J. (2018) Item Recommendation on Monotonic Behavior Chains. In: S. Pera, M. D. Ekstrand, X. Amatriain & J. O'Donovan, eds. *Proceedings*

of the 12th {ACM} Conference on Recommender Systems, RecSys 2018,, October 2-7, 2018. Vancouver, BC, Canada: ACM, pp. 86-94.
Wan, M., Misra, R., McAuley, J. & Nakashole, N. (2019) Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, p. 2605–2610.
Weissmann, J. (2014) The Decline of the American Book Lover. *The Atlantic*, 2014 January.
Yucesoy, B., Wang, X., Huang, J. & Barabási, A.-L. (2018) Success in books: a big data approach to bestsellers. *EPJ Data Science,* 7(7).

Appendix 1: Goodreads data structure

The Goodreads data is in json format. An entry example for each file extracted from the USCD Book Graph (Wan, 2019) is given below:

- books:

{'isbn': '1591935857',
 'text_reviews_count': '4',
 'series': [],
 'country_code': 'US',
 'language_code': '',
 'popular_shelves': [{'count': '2', 'name': 'picture-books'},
  {'count': '2', 'name': 'ducks'},
  {'count': '1', 'name': 'online-reading-in-the-stacks'},
  {'count': '1', 'name': 'nature'},
  {'count': '1', 'name': 'children-books'},
  {'count': '1', 'name': 'animal-books'},
  {'count': '1', 'name': '17909-books'},
  {'count': '1', 'name': 'to-read'}],
 'asin': '',
 'is_ebook': 'false',
 'average_rating': '4.29',
 'kindle_asin': '',
 'similar_books': [],
 'description': '',
 'format': 'Hardcover',
 'link': '<https://www.goodreads.com/book/show/27036533-jump-little-wood-ducks>',
 'authors': [{'author_id': '13195', 'role': ''},
  {'author_id': '30853', 'role': 'Photographs'}],
 'publisher': 'Adventurekeen',
 'num_pages': '36',
 'publication_day': '24',
 'isbn13': '9781591935858',
 'publication_month': '2',
 'edition_information': '',
 'publication_year': '2016',
 'url': '<https://www.goodreads.com/book/show/27036533-jump-little-wood-ducks>',
 'image_url': '<https://images.gr-assets.com/books/1473603845m/27036533.jpg>',
 'book_id': '27036533',
 'ratings_count': '7',
 'work_id': '47077776',
 'title': 'Jump, Little Wood Ducks',
 'title_without_series': 'Jump, Little Wood Ducks'}

  - author:

{'average_rating': '3.51',
 'author_id': '2943855',
 'text_reviews_count': '634',
 'name': 'Kat Menschik',
 'ratings_count': '4599'}

  - works:

{'books_count': '2',

'reviews_count': '33',
'original_publication_month': '',
'default_description_language_code': '',
'text_reviews_count': '4',
'best_book_id': '378460',
'original_publication_year': '',
'original_title': 'The Wanting of Levine',
'rating_dist': '5:7|4:4|3:2|2:0|1:0|total:13',
'default_chaptering_book_id': '',
'original_publication_day': '',
'original_language_id': '',
'ratings_count': '13',
'media_type': '',
'ratings_sum': '57',
'work_id': '368291'}

- series:

{'numbered': 'true',
'note': '',
'description': 'War Stories was a comic book series written by Garth Ennis.',
'title': 'War Stories',
'series_works_count': '5',
'series_id': '834955',
'primary_work_count': '4'}

- user-book interaction:

{'user_id': '8842281e1d1347389f2ab93d60773d4d',
'book_id': '6565837',
'review_id': 'c6c803a462ea21452ffc35b46093ada8',
'is_read': False,
'rating': 0,
'review_text_incomplete': '',
'date_added': 'Thu Aug 17 15:15:28 -0700 2017',
'date_updated': 'Thu Aug 17 15:15:35 -0700 2017',
'read_at': '',
'started_at': ''}

- reviews:

{'user_id': '8842281e1d1347389f2ab93d60773d4d',
'book_id': '18245960',
'review_id': 'dfdbb7b0eb5a7e4c26d59a937e2e5feb',
'rating': 5,
'review_text': 'This is a special book. It started slow for about the first third, then in the, [...] through the decay of neutrons. Also, a high-energy cosmic ray entering the atmosphere may destroy thousands of such miniature universes....'',
'date_added': 'Sun Jul 30 07:44:10 -0700 2017',
'date_updated': 'Wed Aug 30 00:00:26 -0700 2017',
'read_at': 'Sat Aug 26 12:05:52 -0700 2017',
'started_at': 'Tue Aug 15 13:23:18 -0700 2017',
'n_votes': 28,
'n_comments': 1}

Appendix 2: OLS linear regression results for the pre-release model

OLS Regression Results

Dep. Variable:	total	R-squared (uncentered):	0.352
Model:	OLS	Adj. R-squared (uncentered):	0.350
Method:	Least Squares	F-statistic:	276.2
Date:	Fri, 03 Sep 2021	Prob (F-statistic):	0.00
Time:	01:10:07	Log-Likelihood:	-1.2558e+05
No. Observations:	15807	AIC:	2.512e+05
Df Residuals:	15776	BIC:	2.515e+05
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
count_title	-5.5649	1.300	-4.279	0.000	-8.114	-3.016
count_description	0.1188	0.056	2.105	0.035	0.008	0.229
lower	20.4175	11.347	1.799	0.072	-1.824	42.659
middle	10.3102	11.239	0.917	0.359	-11.719	32.340
higher	0.3731	11.179	0.033	0.973	-21.539	22.285
white	126.2389	21.697	5.818	0.000	83.710	168.767
black	73.6821	18.408	4.003	0.000	37.601	109.763
green	58.8084	22.973	2.560	0.010	13.778	103.839
yellow	31.7482	12.765	2.487	0.013	6.727	56.769
red	57.4191	12.213	4.702	0.000	33.481	81.357
magenta	46.4826	21.738	2.138	0.033	3.874	89.092
gray	110.3931	12.323	8.958	0.000	86.238	134.548
blue	48.2575	15.820	3.050	0.002	17.249	79.266
cyan	58.6793	13.588	4.319	0.000	32.046	85.313
publication_month_2	177.2669	25.878	6.850	0.000	126.543	227.991
publication_month_3	215.2572	25.500	8.441	0.000	165.274	265.240
publication_month_4	187.1719	25.460	7.352	0.000	137.268	237.076
publication_month_5	179.7106	25.096	7.161	0.000	130.519	228.903
publication_month_6	161.4461	24.853	6.496	0.000	112.731	210.161
publication_month_7	179.1020	25.991	6.891	0.000	128.158	230.046
publication_month_8	178.1784	25.189	7.074	0.000	128.806	227.551
publication_month_9	153.1405	24.823	6.169	0.000	104.485	201.796
publication_month_10	203.6518	24.503	8.311	0.000	155.622	251.681
publication_month_11	192.2787	25.600	7.511	0.000	142.100	242.458
publication_month_12	154.1880	27.264	5.655	0.000	100.748	207.628
format_Hardcover	4.2030	23.995	0.175	0.861	-42.830	51.236

format_Kindle Edition	118.7732	19.639	6.048	0.000	80.279	157.267
format_Paperback	49.7660	20.393	2.440	0.015	9.794	89.738
format_ebook	84.1959	19.659	4.283	0.000	45.662	122.730
comp_score_description_pos	24.4843	11.133	2.199	0.028	2.661	46.307
comp_score_title_pos	120.2632	13.643	8.815	0.000	93.521	147.005

Omnibus:	3905.398	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7704.737
Skew:	1.524	Prob(JB):	0.00
Kurtosis:	4.553	Cond. No.	2.36e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.36e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Appendix 3: OLS linear regression results for the post-release model

OLS Regression Results

Dep. Variable:	total	R-squared (uncentered):	0.211
Model:	OLS	Adj. R-squared (uncentered):	0.210
Method:	Least Squares	F-statistic:	322.5
Date:	Fri, 03 Sep 2021	Prob (F-statistic):	0.00
Time:	00:41:37	Log-Likelihood:	-97257.
No. Observations:	12103	AIC:	1.945e+05
Df Residuals:	12093	BIC:	1.946e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
is_read	-3.2456	1.093	-2.970	0.003	-5.388	-1.104
rating_1	2.4465	3.744	0.654	0.513	-4.892	9.785
rating_2	3.7942	3.520	1.078	0.281	-3.105	10.694
rating_3	4.4137	1.855	2.379	0.017	0.777	8.051
rating_4	3.3792	1.344	2.515	0.012	0.746	6.013
rating_5	3.3301	1.190	2.797	0.005	0.997	5.664
count_text	0.5568	0.043	12.830	0.000	0.472	0.642
compound_text	417.1713	15.457	26.990	0.000	386.873	447.469
pos_review	0.3188	0.860	0.371	0.711	-1.367	2.005
neg_review	7.2788	2.610	2.788	0.005	2.162	12.396

Omnibus:	2483.403	Durbin-Watson:	1.878
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4400.554
Skew:	1.331	Prob(JB):	0.00
Kurtosis:	4.282	Cond. No.	530.

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
747.5715743802352
772.3178188999102