Analyzing the Factors of Novel Success

Taylor Callahan

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

July 11, 2023

Business Problem

In the highly competitive publishing industry, aspiring authors face significant challenges in landing a literary agent and achieving traditional publication for their novels. With literary agencies receiving hundreds of queries each month and offering representation to only a handful of authors annually, the odds of success are dishearteningly low. To shed light on the factors that contribute to the success of a novel, this analysis aims to examine various elements such as author characteristics, book genres, and other book attributes to determine their impact on book ratings. By identifying the key factors that influence a book's rating, this analysis will provide valuable insights for both industry professionals seeking new talent and authors striving to secure traditional publishing opportunities.

Background/History

The quest to understand the factors that contribute to the success of a novel has long intrigued authors, literary agents, and publishers. Throughout history, countless authors have sought to unravel the elusive code that determines whether a book will capture the hearts of readers and achieve commercial success. However, the publishing industry has always been a challenging and subjective landscape, with the fate of a manuscript often resting in the hands of literary agents and publishers who must make difficult decisions based on limited information (Yucesoy et al., 2018).

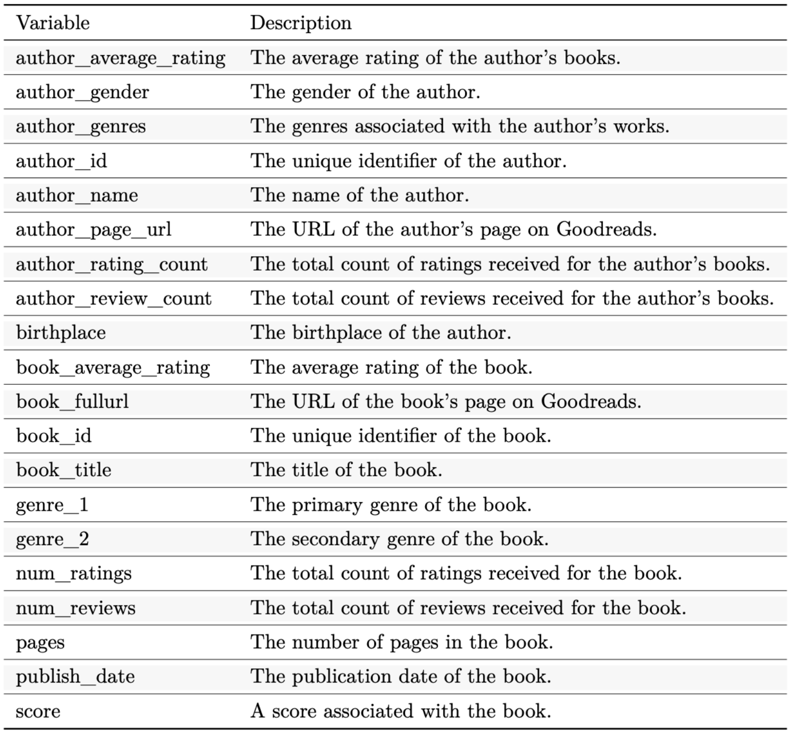
In the past, traditional publishing gatekeepers relied heavily on their intuition and personal preferences when selecting books for publication. This subjective approach often led to biases and overlooked promising authors or stories that didn't fit within established norms. With the advent of digital platforms and self-publishing options, the publishing landscape has undergone significant changes. However, the allure of traditional publishing and the desire for the validation and support it provides remains strong among many authors. Consequently, there is still a pressing need to uncover the underlying factors that can influence the success of a novel in the traditional publishing world. By delving into the historical context and examining data-driven insights, this analysis aims to shed light on this enduring problem and provide valuable guidance to both aspiring authors and industry professionals.

Literary agencies have produced general numbers, suggesting they receive around 200 to 400 queries per month, with that number rising by up to 50% post-COVID. Of those queries, agencies typically only request a full manuscript from around 10% and offer representation to between 5 to 20 per year. This leaves aspiring authors with less than a 1% chance of traditionally publishing their book, a number that is commonly known in the industry (Literary Representation: By the Numbers, 2020).

Data Explanation

The dataset used for this analysis is sourced from Goodreads, a popular online platform for book lovers. It consists of information on 20,000 books found on the platform's most popular list. The dataset, available on Kaggle, contains both author and book attributes. Author attributes include the average rating, gender, genres, ID, name, website URL, rating count, review count, and birthplace. Book attributes encompass the average rating, URL, ID, title, genres, number of ratings, number of reviews, pages, and publish date. Further information on these features can be found in Table 1 below, where “book_average_rating” is the outcome variable.

Table 1: Data Dictionary



Variable	Description
author_average_rating	The average rating of the author's books.
author_gender	The gender of the author.
author_genres	The genres associated with the author's works.
author_id	The unique identifier of the author.
author_name	The name of the author.
author_page_url	The URL of the author's page on Goodreads.
author_rating_count	The total count of ratings received for the author's books.
author_review_count	The total count of reviews received for the author's books.
birthplace	The birthplace of the author.
book_average_rating	The average rating of the book.
book_fullurl	The URL of the book's page on Goodreads.
book_id	The unique identifier of the book.
book_title	The title of the book.
genre_1	The primary genre of the book.
genre_2	The secondary genre of the book.
num_ratings	The total count of ratings received for the book.
num_reviews	The total count of reviews received for the book.
pages	The number of pages in the book.
publish_date	The publication date of the book.
score	A score associated with the book.

While the dataset provides a comprehensive overview of authors and their books, it has certain limitations. Firstly, the dataset focuses only on books from Goodreads' most popular list, potentially introducing a bias towards popular or highly rated books. Additionally, the dataset may not capture the entire spectrum of literary genres, as it relies on categorization provided by the platform. Another limitation is the presence of missing values, errors, or inconsistencies that need to be addressed through data cleaning techniques. Furthermore, the dataset does not include explicit features for the length of the book title, the first word of the title, or the number of words in the title. However, these features can be derived during the data preparation stage.

To prepare the data for analysis, several steps will be undertaken. Firstly, missing values will be addressed using imputation techniques to ensure data completeness. Additionally, new features will be created, including the length of the book title, the first word of the title, and the number of words in the title. These additional features can provide valuable insights into the impact of the book title on the rating. Finally, data cleaning and transformations will be performed to ensure the reliability and quality of the dataset for subsequent analysis.

The dataset provides a rich source of information about authors and their books. However, it comes with limitations related to potential biases, missing values, and the need for data preparation. By addressing these limitations and performing necessary data cleaning and transformation steps, including the creation of additional features, the dataset can be prepared for analysis to uncover the factors that contribute to the success of a novel.

Methods and Analysis

Prior to conducting the modeling, exploratory data analysis (EDA) was performed to gain insights into the dataset. The EDA aimed to understand the distribution of book ratings, the frequency of common first words in book titles, the primary genres with high ratings, the relationship between average book rating and title characteristics, and any potential differences in ratings based on author gender.

A histogram was created to visualize the distribution of book ratings. The histogram showed a unimodal distribution, with the majority of ratings centering around 4. This indicates that the ratings tend to be positive, with a concentration towards the higher end of the rating scale. This can be seen in Figure 1.

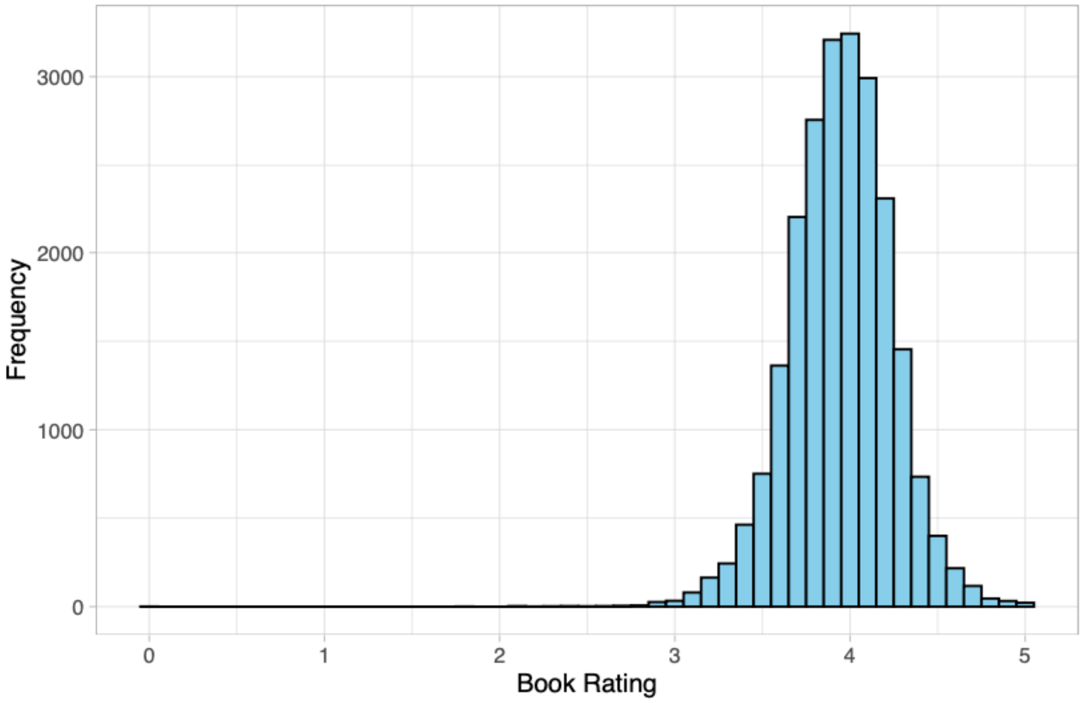
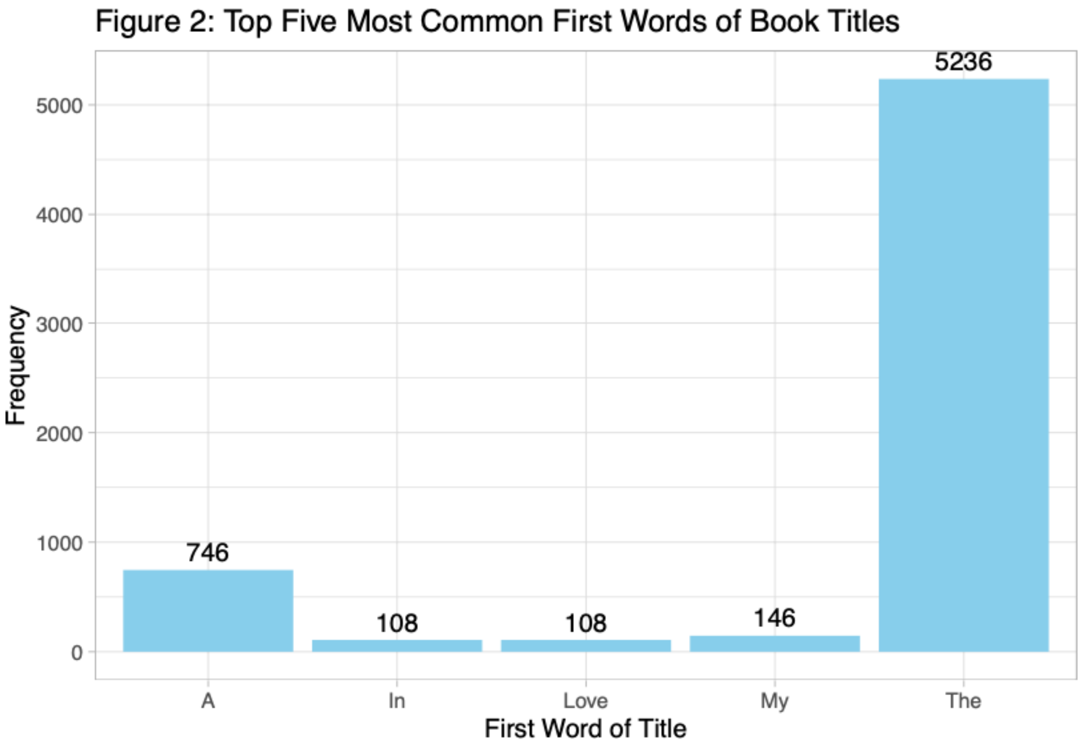


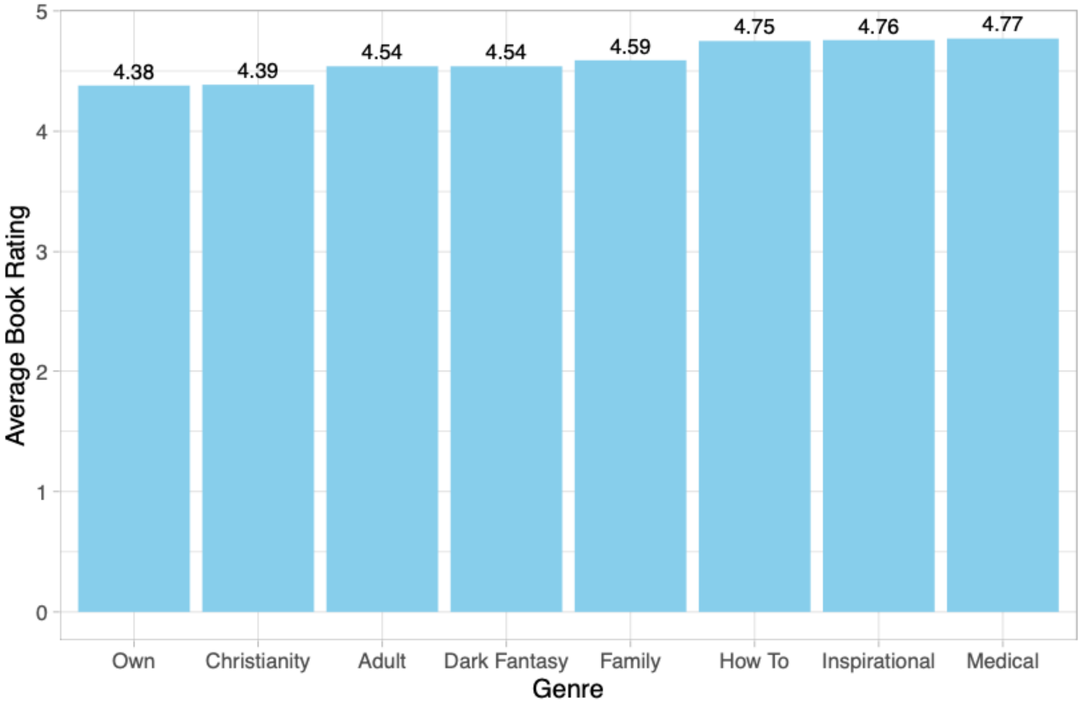
Figure 1: Histogram of Book Ratings

To explore the common first words in book titles, a bar chart was generated displaying the top five most frequent first words. The analysis revealed that "The" was the most common first word, followed by "A", "My", "In", and "Love". This information provides insights into the typical patterns and themes present in book titles. This can be seen in Figure 2.



Another bar chart was created to examine the primary genres with the highest ratings. The top eight highly rated primary genres were identified as "Medical", "Inspirational", "How To", "Family", "Dark Fantasy", "Adult", "Christianity", and "Own", which refers to "OwnVoices". "OwnVoices" refers to literature where the protagonist shares the same marginalized identity as the author, providing an authentic representation of the character's experiences. This can be seen in Figure 3.

Figure 3: Average Book Ratings for Top Eight Highest−Rated Genres



Scatterplots were generated to investigate the relationship between average book rating and the number of words in the title, book title length, and number of pages. None of these scatterplots showed a clear linear relationship, suggesting that these factors alone may not be strong predictors of the average book rating. These can be seen in Figures 4, 5, and 6.

Figure 4: Average Book Rating vs. Words in Title

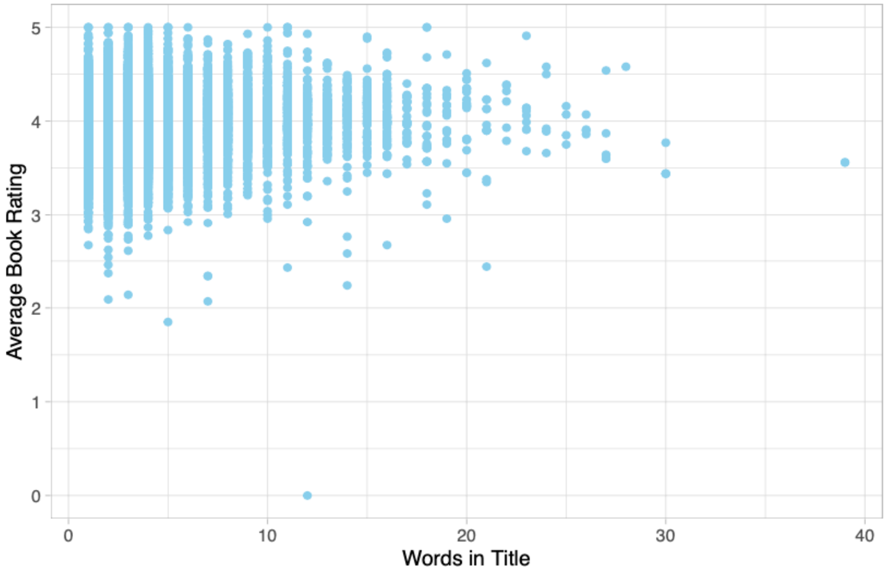


Figure 5: Average Book Rating vs. Book Title Length

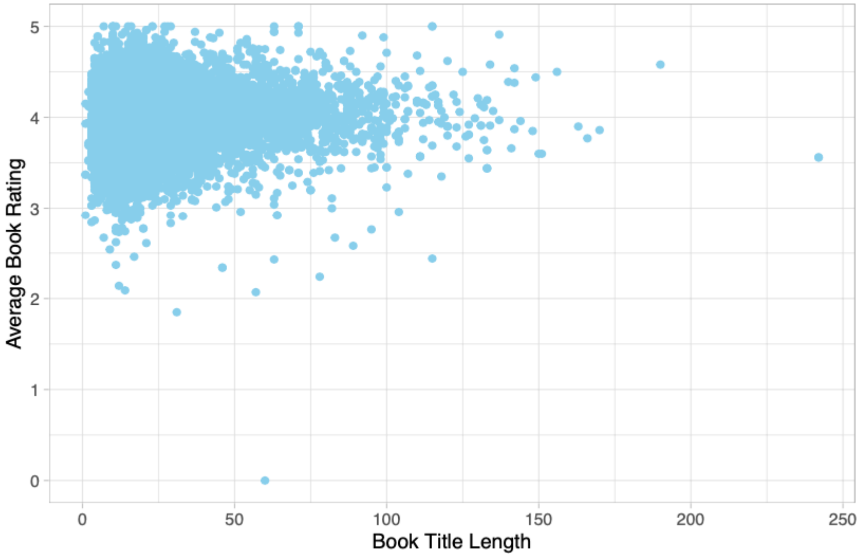
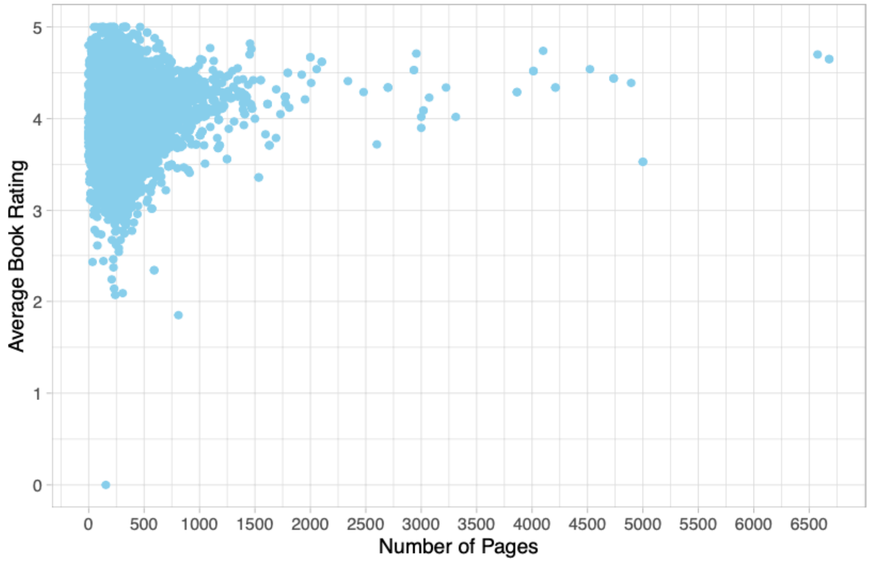
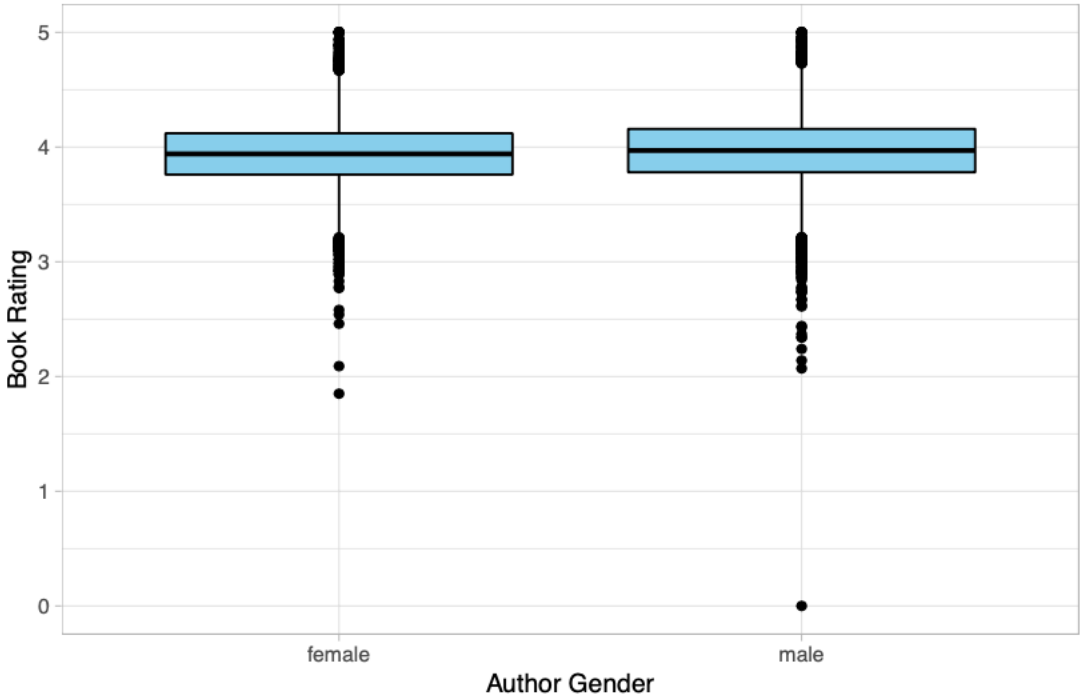


Figure 6: Average Book Rating vs. Number of Pages



A box plot was constructed to compare average ratings by author gender. The analysis revealed no significant differences in ratings based on gender, indicating that author gender does not appear to have a substantial impact on the average book rating. This can be seen in Figure 7.

Figure 7: Book Ratings by Author Gender



Overall, the visual EDA provided valuable insights into the data. The distribution of book ratings, common first words in book titles, highly rated genres, and the absence of gender-based differences in ratings contribute to a better understanding of the dataset. These findings can inform subsequent modeling and analysis steps to predict the factors that contribute to the success of a novel.

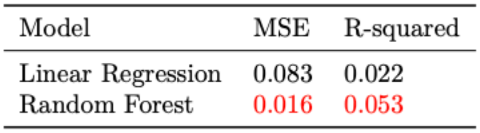
Fitting a linear model and random forest model allows us to explore the relationship between the independent variables and the outcome variable in the dataset. A linear model assumes a linear relationship between the predictors and the response variable. It provides coefficients that indicate the strength and direction of the relationship between each feature and the average book rating. This model is useful in understanding the individual impact of each feature and their combined effect on the outcome. However, linear models may not capture complex nonlinear relationships present in the data (Nathans et al., 2012).

On the other hand, a random forest model is an ensemble learning algorithm that combines multiple decision trees. It is capable of capturing nonlinear relationships, interactions, and complex patterns in the data. Random forests are advantageous for their ability to handle high-dimensional data with many features and their robustness against overfitting. In this analysis, the random forest model is particularly useful because it provides feature importance scores, which allow us to identify the relative impact of each feature on the book rating (Haddouchi et al., 2019).

To evaluate the performance of the models, mean squared error (MSE) and R-squared were chosen as evaluation metrics. MSE measures the average squared difference between the predicted and actual book ratings, providing a measure of the model's predictive accuracy. R-squared, also known as the coefficient of determination, represents the proportion of variance in the book ratings explained by the model. Higher R-squared values indicate better fit and capture of the variability in the data.

In comparing the linear model and random forest model, it was found that the random forest model had a smaller MSE and a higher R-squared value. This suggests that the random forest model outperformed the linear model in terms of predictive accuracy and capturing the variability in the book ratings. These results can be seen in Table 2. However, it is important to note that both models had relatively low R-squared values, indicating that the models explain only a small portion of the variance in the book ratings. This suggests that there may be other factors not captured by the available features that influence the book ratings.

Table 2: Model Evaluation Metrics



Model	MSE	R-squared
Linear Regression	0.083	0.022
Random Forest	0.016	0.053

By examining the feature importance scores obtained from the random forest model, it was determined that the page count had the greatest impact on the book rating. This suggests that the length of a book may play a significant role in its rating. These results can be seen in Table 3. However, it is important to interpret these results with caution, as the models performed poorly overall and the R-squared values were low. This indicates that the available features in the dataset might not be sufficient to fully explain the complexity and nuances that contribute to the success of a novel. Further exploration and consideration of additional factors may be necessary to gain a more comprehensive understanding of the rating dynamics.

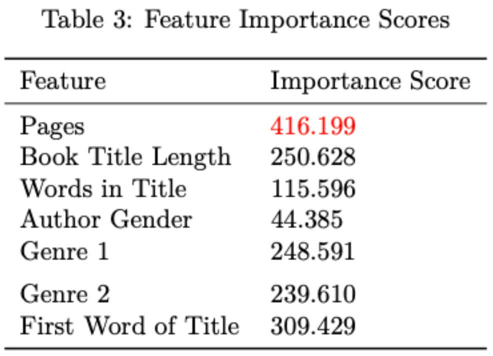


Table 3: Feature Importance Scores

Feature	Importance Score
Pages	416.199
Book Title Length	250.628
Words in Title	115.596
Author Gender	44.385
Genre 1	248.591
Genre 2	239.610
First Word of Title	309.429

Conclusion

The analysis of the dataset provided valuable insights into the factors that contribute to the success of a novel. The exploratory data analysis (EDA) revealed interesting patterns and trends in the data, such as the distribution of book ratings, common first words in book titles, highly rated genres, and the absence of gender-based differences in ratings. However, the models fitted for prediction, including the linear model and random forest model, performed poorly overall, indicating that the available features may not fully capture the complexities that influence book ratings. Nonetheless, the random forest model showed better predictive accuracy and provided feature importance scores, identifying the page count as the feature with the greatest impact on the book rating.

The insights gained from this analysis can be practically applied in several ways. For literary agents and publishers, understanding the factors that contribute to the success of a novel can help in the selection process when evaluating new manuscripts. By recognizing the influence of the page count, agents can consider the length of a book as a potential factor affecting its marketability and appeal to readers. Additionally, the knowledge of highly rated genres can guide agents in identifying book categories that tend to resonate well with readers.

For aspiring authors, these findings can offer guidance in the pursuit of traditional publishing opportunities. Knowing that book ratings tend to concentrate around the higher end of the scale can motivate authors to aim for excellence in their writing. Understanding the common first words in book titles can help authors craft titles that are catchy and align with readers' expectations.

Assumptions

The analysis assumes that the dataset from Goodreads provides a representative sample of books and ratings. It also assumes that the dataset is free from significant biases that could impact the results. Additionally, the assumption is made that the selected features, although limited, provide meaningful insights into the factors influencing book ratings.

Limitations and Challenges

The dataset focuses on books from Goodreads' most popular list, potentially introducing a bias towards highly rated or popular books. The absence of certain features, such as explicit title length and the number of words in the title, required the creation of derived features. However, these derived features might not fully capture the nuances of the original variables. The low R-squared values in both models suggest that there may be other unaccounted factors influencing book ratings, indicating the need for additional variables or data sources.

Future Uses/Additional Applications

Future studies could expand the dataset to include a broader range of books and ratings from different sources, providing a more comprehensive understanding of the factors impacting book success. Additionally, the analysis could be extended to include more advanced modeling techniques or alternative approaches to capture the complexities of book ratings.

Recommendations and Implementation

Based on the findings, it is recommended that literary agents and publishers consider the page count as a factor when evaluating manuscripts. Authors can focus on optimizing the length of their books to align with market expectations. Additionally, authors can leverage the knowledge of highly rated genres to guide their writing choices and target audience preferences.

Ethical Assessment

The analysis of book ratings raises ethical considerations regarding bias and representation. The dataset should be examined for any potential biases that could perpetuate inequalities or favor certain groups. Additionally, it is crucial to ensure that the analysis respects the privacy and anonymity of authors and readers. Care should be taken to handle personal data appropriately and adhere to relevant data protection regulations.

Questions

1. How reliable and representative is the Goodreads dataset used in the analysis?
2. Can the findings from this analysis be generalized to the broader publishing industry?
3. Are there specific factors that were found to have a significant impact on book ratings other than the page count?
4. What are the potential limitations of using average book ratings as the outcome variable?
5. How did the analysis address or mitigate potential biases in the dataset?
6. Can the analysis provide insights into the relationship between specific genres and book ratings?
7. Did the analysis consider the influence of other variables, such as author characteristics, on book ratings?
8. How do the low R-squared values in the models affect the validity of the findings?
9. Were there any unexpected or counterintuitive findings in the analysis?
10. How might the findings and insights from this analysis be practically applied by authors, literary agents, and publishers in their decision-making processes?

References

Haddouchi, M., & Berrado, A. (2019, October). A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)* (pp. 1-6). IEEE.

Literary Representation: By the Numbers. The Darling Axe. (2020, May 17). https://darlingaxe.com/blogs/news/by-the-numbers

Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: a guidebook of variable importance. *Practical assessment, research & evaluation*, *17*(9), n9.

Yucesoy, B., Wang, X., Huang, J. *et al.* Success in books: a big data approach to bestsellers. *EPJ Data Sci.* **7**, 7 (2018). https://doi.org/10.1140/epjds/s13688-018-0135-y

Code Appendix

```r
library(ggplot2)
library(kableExtra)
library(dplyr)
library(stringr)
library(gridExtra)
library(ranger)
library(knitr)

# load data
df <- read.csv("good_reads_final.csv")

# Create the data dictionary
data_dict <- data.frame(
  Variable = c("author_average_rating", "author_gender", "author_genres",
               "author_id", "author_name",
               "author_page_url", "author_rating_count", "author_review_count",
               "birthplace",
               "book_average_rating", "book_fullurl", "book_id", "book_title",
               "genre_1", "genre_2",
               "num_ratings", "num_reviews", "pages", "publish_date", "score"),
  Description = c("The average rating of the author's books.",
                  "The gender of the author.",
                  "The genres associated with the author's works.",
                  "The unique identifier of the author.",
                  "The name of the author.",
                  "The URL of the author's page on Goodreads.",
                  "The total count of ratings received for the author's books.",
                  "The total count of reviews received for the author's books.",
                  "The birthplace of the author.",
                  "The average rating of the book.",
                  "The URL of the book's page on Goodreads.",
                  "The unique identifier of the book.",
                  "The title of the book.",
                  "The primary genre of the book.",
                  "The secondary genre of the book.",
                  "The total count of ratings received for the book.",
                  "The total count of reviews received for the book.",
                  "The number of pages in the book.",
                  "The publication date of the book.",
                  "A score associated with the book."),
    stringsAsFactors = FALSE
)

# Format the Data Dictionary table using kableExtra
data_dict_table <- kable(data_dict, format = "latex", booktabs = TRUE,
                         caption = "Data Dictionary") %>%
  kable_styling(latex_options = c("striped", "hold_position"),
                full_width = FALSE) %>%
  row_spec(1, bold = FALSE) %>%
  collapse_rows(columns = 2, valign = "top") %>%
  column_spec(1, border_right = FALSE)
```

```r
# Print the Data Dictionary table
print(data_dict_table)

# Remove leading/trailing whitespace and newline characters from the
# book_title column
df$book_title <- str_trim(df$book_title, side = "both")

# Add book_title_length feature to df
df <- df %>% mutate(book_title_length = nchar(book_title))

# Add first_word_of_title feature to df
df <- df %>% mutate(first_word_of_title = word(book_title, 1))

# Add words_in_title feature to df
df <- df %>% mutate(words_in_title = str_count(book_title, "\\w+"))

# Create a histogram of book ratings
ggplot(df, aes(x = book_average_rating)) +
  geom_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +
  labs(x = "Book Rating", y = "Frequency",
       title = "Figure 1: Histogram of Book Ratings") +
  theme_light()

# Count the frequency of first words
first_word_freq <- table(df$first_word_of_title)

# Sort the frequencies in descending order
sorted_freq <- sort(first_word_freq, decreasing = TRUE)

# Extract the top five most common first words
top_words <- names(sorted_freq)[1:5]

# Subset the dataframe for the top five words
top_df <- subset(df, first_word_of_title %in% top_words)

# Create a bar chart of the top five most common first words
ggplot(top_df, aes(x = first_word_of_title)) +
  geom_bar(fill = "skyblue") +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  labs(x = "First Word of Title", y = "Frequency",
       title = "Figure 2: Top Five Most Common First Words of Book Titles") +
  theme_light()

# Calculate the average book rating by genre
avg_ratings <- aggregate(df$book_average_rating, by = list(df$genre_1),
                         FUN = mean)
colnames(avg_ratings) <- c("Genre", "Average_Rating")

# Sort the average ratings in descending order
avg_ratings <- avg_ratings[order(avg_ratings$Average_Rating,
                                 decreasing = TRUE), ]

# Select the top eight highest-rated genres
```

```r
top_genres <- avg_ratings$Genre[1:8]
top_avg_ratings <- avg_ratings[avg_ratings$Genre %in% top_genres, ]

# Create a bar chart of average book ratings for the top eight
# highest-rated genres
ggplot(top_avg_ratings, aes(x = reorder(Genre, Average_Rating),
                            y = Average_Rating)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_text(aes(label = round(Average_Rating, 2)),
            vjust = -0.5, color = "black", size = 3) +
  labs(x = "Genre", y = "Average Book Rating",
       title = "Figure 3: Average Book Ratings for Top Eight Highest-Rated Genres") +
  theme_light()

# Scatterplot: average_book_rating vs. words_in_title
ggplot(df, aes(x = words_in_title, y = book_average_rating)) +
  geom_point(color = "skyblue", fill = "skyblue") +
  labs(x = "Words in Title", y = "Average Book Rating") +
  ggtitle("Figure 4: Average Book Rating vs. Words in Title") +
  theme_light()
# Scatterplot: average_book_rating vs. book_title_length
ggplot(df, aes(x = book_title_length, y = book_average_rating)) +
  geom_point(color = "skyblue", fill = "skyblue") +
  labs(x = "Book Title Length", y = "Average Book Rating") +
  ggtitle("Figure 5: Average Book Rating vs. Book Title Length") +
  theme_light()

# Convert 'pages' to numeric and remove non-numeric values
df$pages <- as.numeric(as.character(df$pages))
df <- df[!is.na(df$pages), ]

# Scatterplot: average_book_rating vs. pages
ggplot(df, aes(x = pages, y = book_average_rating)) +
  geom_point(color = "skyblue", fill = "skyblue") +
  labs(x = "Number of Pages", y = "Average Book Rating") +
  ggtitle("Figure 6: Average Book Rating vs. Number of Pages") +
  scale_x_continuous(breaks = seq(0, max(df$pages), by = 500)) +
  theme_light()

# box plot of ratings by gener
ggplot(df, aes(x = author_gender, y = book_average_rating)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(x = "Author Gender", y = "Book Rating",
       title = "Figure 7: Book Ratings by Author Gender") +
  theme_light()
# Linear regression model
linear_model <- lm(book_average_rating ~ pages + book_title_length +
                     words_in_title, data = df)

# Fit the ranger random forest model with importance option
random_forest_model <- ranger(book_average_rating ~ pages + book_title_length +
                                words_in_title + author_gender + genre_1 +
                                genre_2 + first_word_of_title, data = df,
                              importance = "impurity")

# Prediction using linear regression model
linear_pred <- predict(linear_model, newdata = df)

# Prediction using random forest model
random_forest_pred <- predict(random_forest_model, data = df)$predictions

# Calculate MSE and R-squared for linear regression model
linear_mse <- mean((df$book_average_rating - linear_pred)^2)
linear_r_squared <- summary(linear_model)$r.squared

# Calculate MSE and R-squared for random forest model
random_forest_mse <- mean((df$book_average_rating - random_forest_pred)^2)
random_forest_r_squared <- random_forest_model$prediction.error
# Create a data frame with model names and evaluation metrics
evaluation_table <- data.frame(
  Model = c("Linear Regression", "Random Forest"),
  MSE = c(linear_mse, random_forest_mse),
  R_Squared = c(linear_r_squared, random_forest_r_squared)
)

# Highlight the lowest RMSE and highest R-squared values in red
evaluation_table$MSE <- ifelse(evaluation_table$MSE
                                == min(evaluation_table$MSE),
                                sprintf("\\textcolor{red}{%.3f}",
                                        evaluation_table$MSE),
                                sprintf("%.3f", evaluation_table$MSE))

evaluation_table$R_Squared <- ifelse(evaluation_table$R_Squared
                                     == max(evaluation_table$R_Squared),
                                     sprintf("\\textcolor{red}{%.3f}",
                                             evaluation_table$R_Squared),
                                  sprintf("%.3f", evaluation_table$R_Squared))

# Set the column names
colnames(evaluation_table) <- c("Model", "MSE", "R-squared")

# Print the table using kable with LaTeX formatting
kable(evaluation_table, format = "latex", booktabs = TRUE, escape = FALSE,
      caption = "Model Evaluation Metrics") %>%
  kable_styling()
# Get feature importance using importance()
importance_scores <- importance(random_forest_model)

# Create a data frame with feature names and importance scores
importance_table <- data.frame(Feature = names(importance_scores),
                               Importance = importance_scores,
                               row.names = NULL)

# Capitalize new words and replace underscores with spaces in feature names
importance_table$Feature <- gsub("_", " ", importance_table$Feature)
importance_table$Feature <- tools::toTitleCase(importance_table$Feature)
```

```
# highlight greatest score
importance_table$Importance <- ifelse(importance_table$Importance
                                == max(importance_table$Importance),
                                sprintf("\\textcolor{red}{%.3f}",
                                        importance_table$Importance),
                                sprintf("%.3f", importance_table$Importance))

# Set the column names
colnames(importance_table) <- c("Feature", "Importance Score")

# Print the table using kable with LaTeX formatting
kable(importance_table, format = "latex", booktabs = TRUE, escape = FALSE,
      caption = "Feature Importance Scores") %>%
  kable_styling()
```