Analyzing the Factors of Novel Success

A Study by Taylor Callahan

Table of contents

01

Introduction

Problem and Objective

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 less than a 1% chance of success.

Objective

The objective of the analysis is to identify key factors that impact book ratings in the competitive publishing industry, offering valuable insights for aspiring authors and industry professionals seeking new talent.

History and Prior Research

Current Trends

There are a rising number of queries to literary agencies, increasing demand for diverse voices and genres, and a continued preference for positive book ratings among readers.

Prior Research

Prior research in the publishing industry has focused on understanding subjective selection processes, the impact of author characteristics on success, and the influence of genre preferences on book ratings and sales.

Future Projections

Future projections suggest a continued shift towards digital publishing platforms, increased use of data analytics for book marketing, and a growing emphasis on diverse and inclusive storytelling to cater to changing reader preferences.

Data

Data Source

Source

The data comes from Goodreads, a popular online platform for book lovers, and is available on Kaggle.

Contents

The dataset contains author and book attributes such as book and author rating, author gender, book genre, title, etc.

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 reviews 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.

Variable	Description
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.

Data Preparation

Missing Values

Imputation techniques were used where possible. Other missing data was removed.

New Features

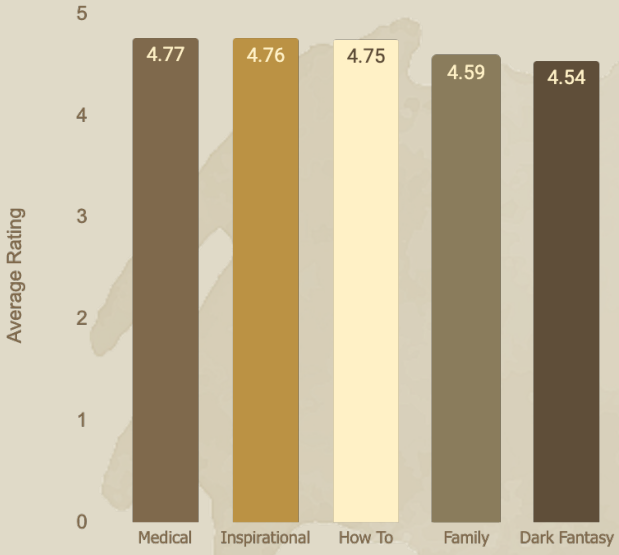
Features were created for length of book title, first word of title, and number of words in title.

Data Cleaning

Data cleaning ensures the reliability and quality of the dataset.

Methods

Exploratory Data Analysis



The top five genres based on average rating are medical, inspirational, how to, family, and dark fantasy.

Book Title

Most books begin with the word ”The”, followed by “A”, “My”, “In”, and “Love”.

Correlation

No clear linear correlation between rating and number of words in title, book title length, or number of pages.

Gender

No clear difference in average rating based on an author’s gender.

Outliers

There were the most outliers present in page number, with a few books exceeding 6,500 pages.

Modeling

Linear

Assumes a linear relationship between the predictors and the response variable.

Result 2.2%

Percent of the variation explained by the model.

0.083
MSE

0.022
R-Squared

Random Forest

Capable of capturing nonlinear relationships, interactions, and complex patterns.

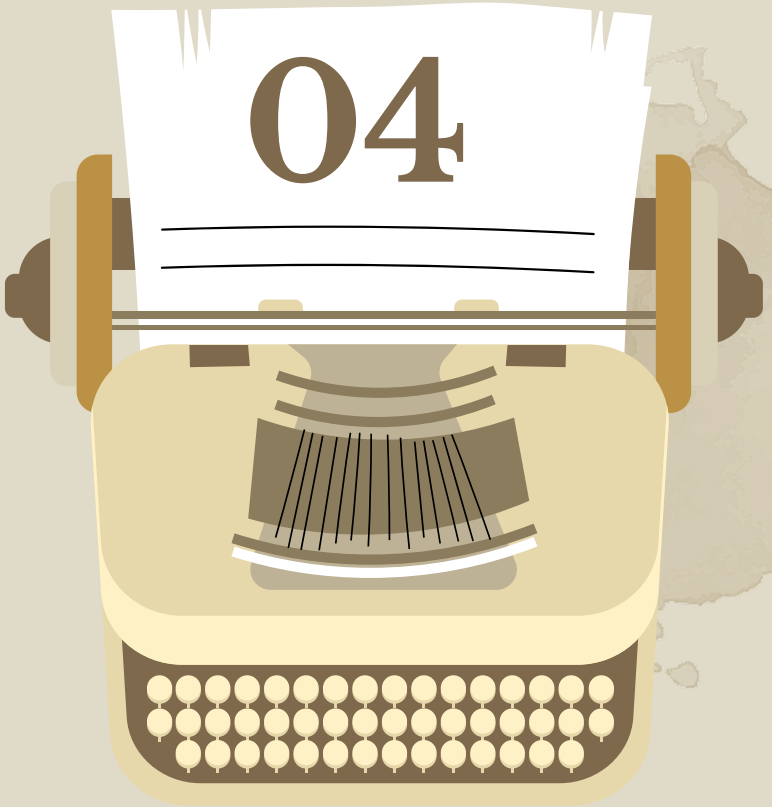
Result 5.3%

Percent of the variation explained by the model.

0.016
MSE

0.053
R-Squared

Limitations



Limitations and Challenges

Insufficient Features

The available features in the dataset might not fully capture the complexities and nuances that influence book ratings, leading to relatively low R-squared values and limited explanatory power of the models.

Dataset Bias

The dataset focuses on books from Goodreads' most popular list, potentially introducing a bias towards highly rated or popular books, which may not represent the entire publishing landscape.

Conclusion

Model Selection and Practical Use

Model Selection

The final model selection favored the random forest model over the linear model due to its ability to capture nonlinear relationships and better predictive performance, despite the overall low R-squared values.

Practical Use

Literary agents can consider page count when evaluating manuscripts, and authors can leverage insights on highly rated genres and catchy book titles to increase their chances of success in traditional publishing.

Thank You