SC1015 _ ECDS _ Group 10

Mini Project

Predict Best Seller

by Chu Gia Han, Nguyen Phuong Thuy, Nguyen Duy

Contents

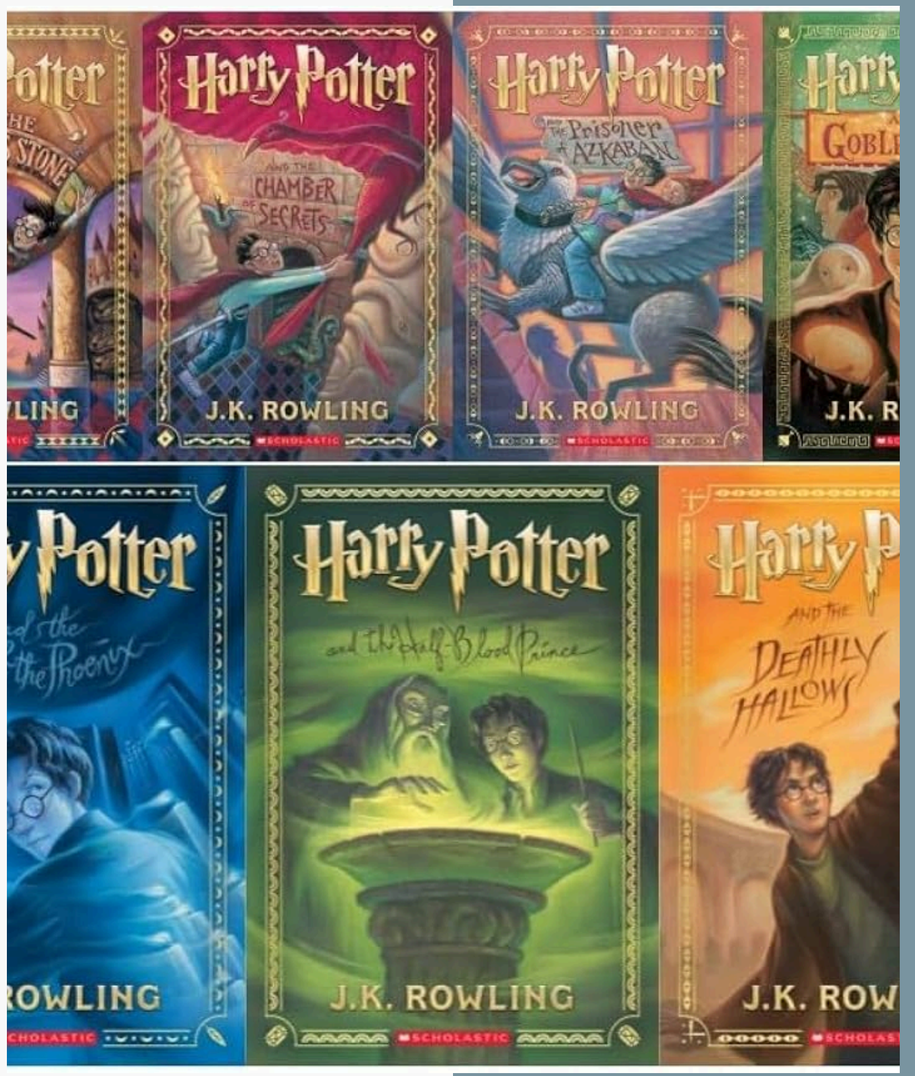
Motivation

- This is a small project for our course SC1015 (Introduction to Data Science and Artificial Intelligence), we decided to choose books as our main focus base on the data set we found.
- This project explores a dataset of books with the goal of uncovering patterns in publishing trends, bestseller characteristics, and author profiles using data visualization, text mining, and machine learning techniques.

Best Seller Book

Harry Potter

- Author: J.K. Rowling (British author)
- Genre: Fantasy, Adventure, Coming-of-age, Young Adult
- Language: Originally published in English
- Publisher: Bloomsbury (UK), Scholastic (US)
- Publication Period: 1997 – 2007 (7 books total)
- Total Books Sold: Over 600 million copies worldwide
- Best-Selling Status: One of the best-selling book series of all time

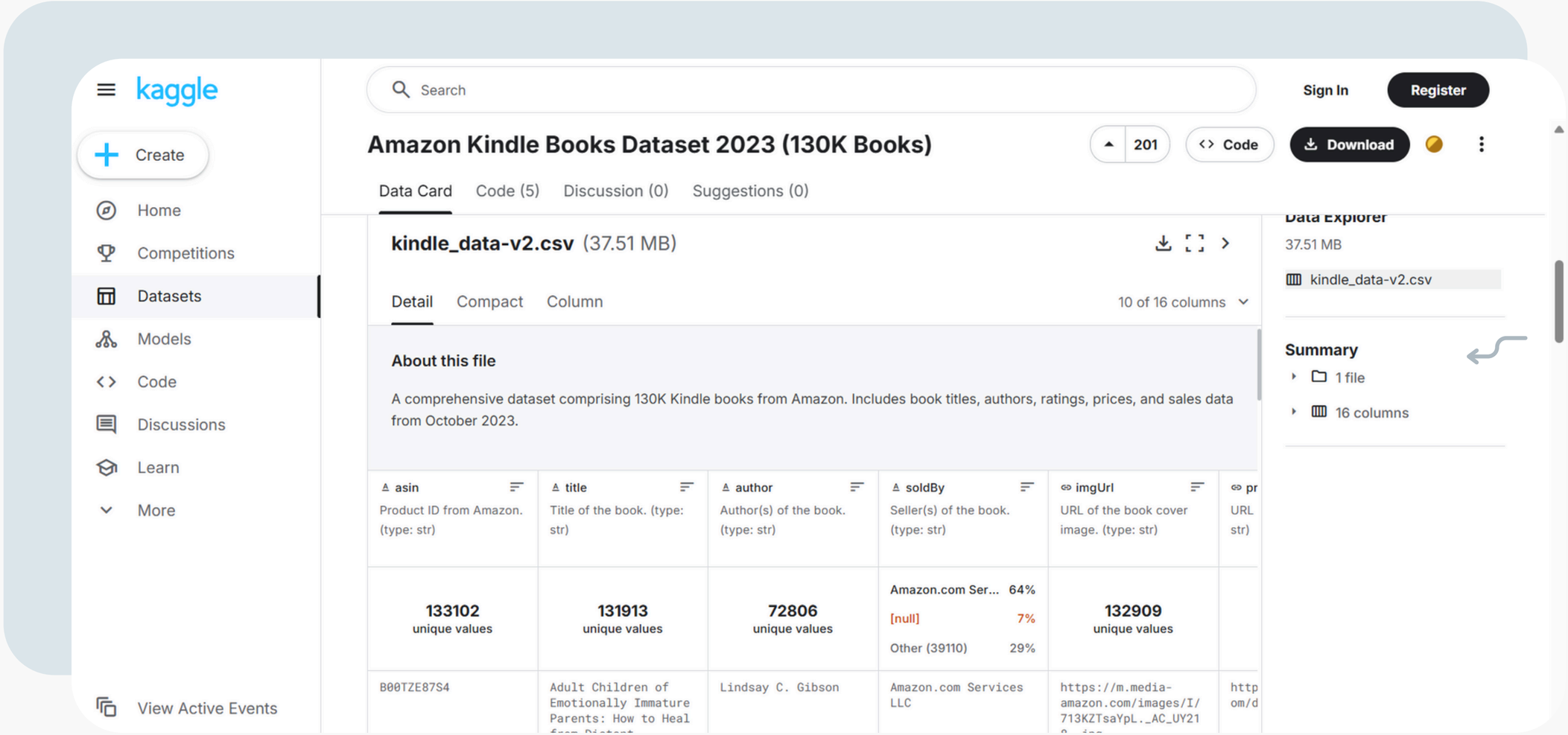


Problem Definition

Can we tell if the book will be the best seller based on it writer, review and title?

Which model will fit the best to predict in this problem?

DataSet


kaggle
Search
Sign In
Register
Create
Home
Competitions
Datasets
Models
Code
Discussions
Learn
More
View Active Events
Amazon Kindle Books Dataset 2023 (130K Books)
201
Code
Download
Data Card
Code (5)
Discussion (0)
Suggestions (0)
kindle_data-v2.csv (37.51 MB)
Detail
Compact
Column
10 of 16 columns
About this file
A comprehensive dataset comprising 130K Kindle books from Amazon. Includes book titles, authors, ratings, prices, and sales data from October 2023.
asin
Product ID from Amazon. (type: str)
title
Title of the book. (type: str)
author
Author(s) of the book. (type: str)
soldBy
Seller(s) of the book. (type: str)
imgUrl
URL of the book cover image. (type: str)
133102
unique values
131913
unique values
72806
unique values
Amazon.com Ser... 64%
[null] 7%
Other (39110) 29%
132909
unique values
B00TZE87S4
Adult Children of Emotionally Immature Parents: How to Heal
Lindsay C. Gibson
Amazon.com Services LLC
https://m.media-amazon.com/images/I/713KZTsaYpL._AC_UY21
Data Explorer
37.51 MB
kindle_data-v2.csv
Summary
1 file
16 columns

Data Preparation

1. **Remove the blank box in "author" column**

```
#Remove row blank author
book_data_cleaned = book_data.dropna(subset=['author'])
book_data_cleaned.info()
book_data_cleaned.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 133102 entries, 0 to 133101
Data columns (total 16 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   asin                133102 non-null  object
 1   title               133102 non-null  object
 2   author              132677 non-null  object
 3   soldBy              123869 non-null  object
 4   imgUrl              133102 non-null  object
 5   productURL          133102 non-null  object
 6   stars               133102 non-null  float64
 7   reviews             133102 non-null  int64
 8   price               133102 non-null  float64
 9   isKindleUnlimited   133102 non-null  bool
 10  category_id         133102 non-null  int64
 11  isBestSeller        133102 non-null  bool
 12  isEditorsPick       133102 non-null  bool
 13  isGoodReadsChoice   133102 non-null  bool
 14  publishedDate       84086 non-null   object
 15  category_name       133102 non-null  object
dtypes: bool(4), float64(2), int64(2), object(8)
memory usage: 12.7+ MB
```

Data Preparation

2. Remove unnecessary columns

```
#Remove unnecessary columns
columns_to_remove = ["soldBy", "imgUrl", "productURL", "reviews","publishedDate"]
book_data_cleaned = book_data_cleaned.drop(columns=columns_to_remove)
book_data_cleaned.info()
book_data_cleaned.head()
```

KEEP	REMOVE
• asin • title • author • stars • price • isKindleUnlimited • category_id • isBestSeller • isEditorsPick • isGoodReadsChoice • Category_name	• soldBy • imgUrl • productURL • reviews • publishedDate

Data Preparation

3. Check for duplicate books

```
#Check for duplicate books using asin
duplicate_count = book_data_cleaned.duplicated(subset=['asin']).sum()

if duplicate_count > 0:
    book_data_cleaned = book_data_cleaned.drop_duplicates(subset=['asin'], keep='first')
    print(f"Removed {duplicate_count} duplicate ASINs. File updated successfully.")

else:
    print("No duplicate ASINs found. The file remains the same.")
```

4. Remove wrong asin values

```
#drop asin that is int (wrong format)
book_data_cleaned = book_data_cleaned[~book_data_cleaned['asin'].astype(str).str.isdigit()]
book_data_cleaned.info()
```

5. Save book_data_cleaned into csv file

```
book_data_cleaned.to_csv('book_data_cleaned.csv', index=False)
```

Exploratory Data Analysis

Rating

Bestseller Status

Catergory

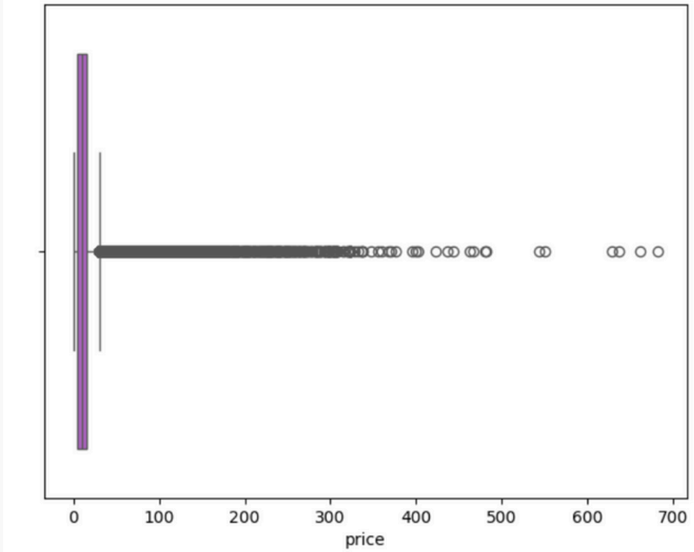
Price

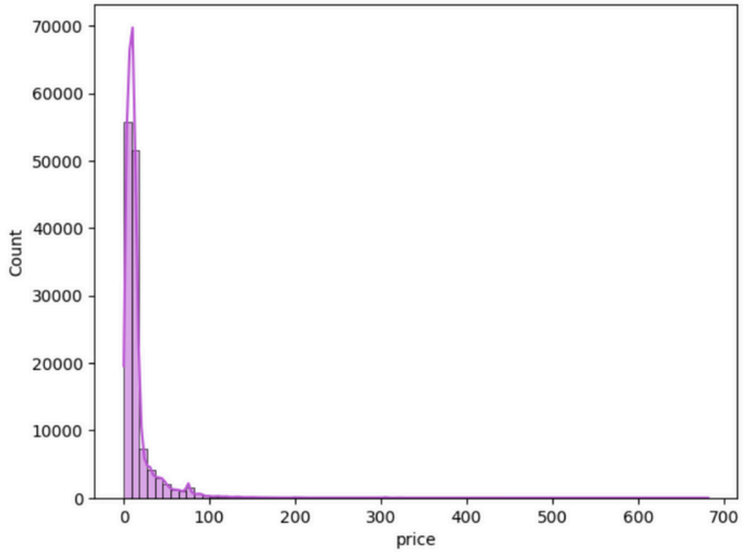
Author

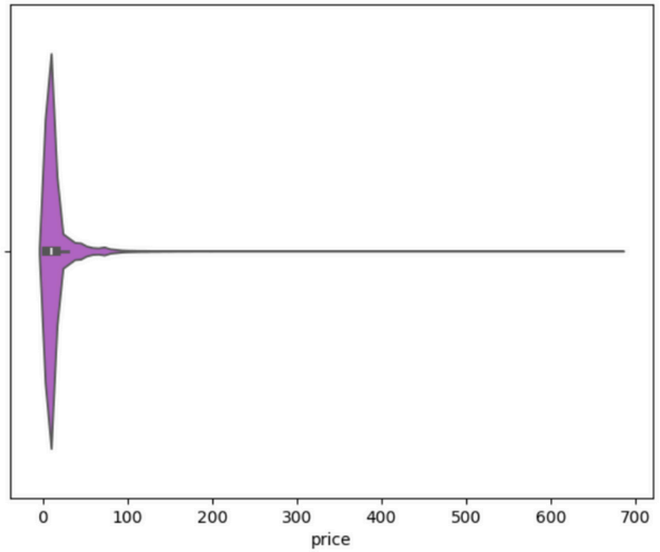
Univarate

Book Price

- Most books are priced under $20, with a peak around $5–$10.
- The price distribution is right-skewed, with some high outliers.
- Bestsellers tend to fall in the $5–$15 range.
- Extremely cheap or expensive books are less likely to be bestsellers.

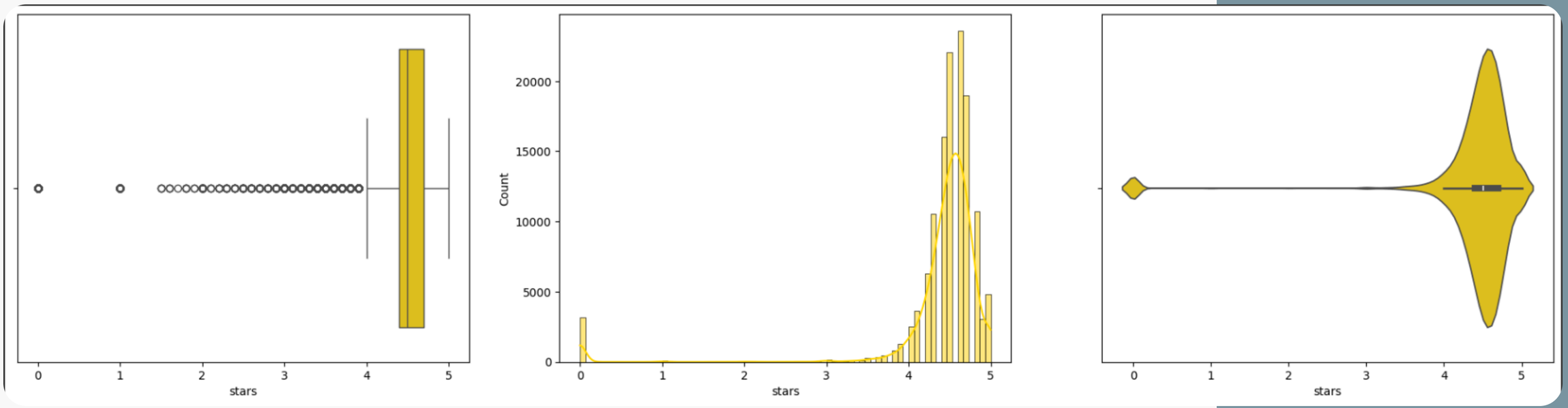




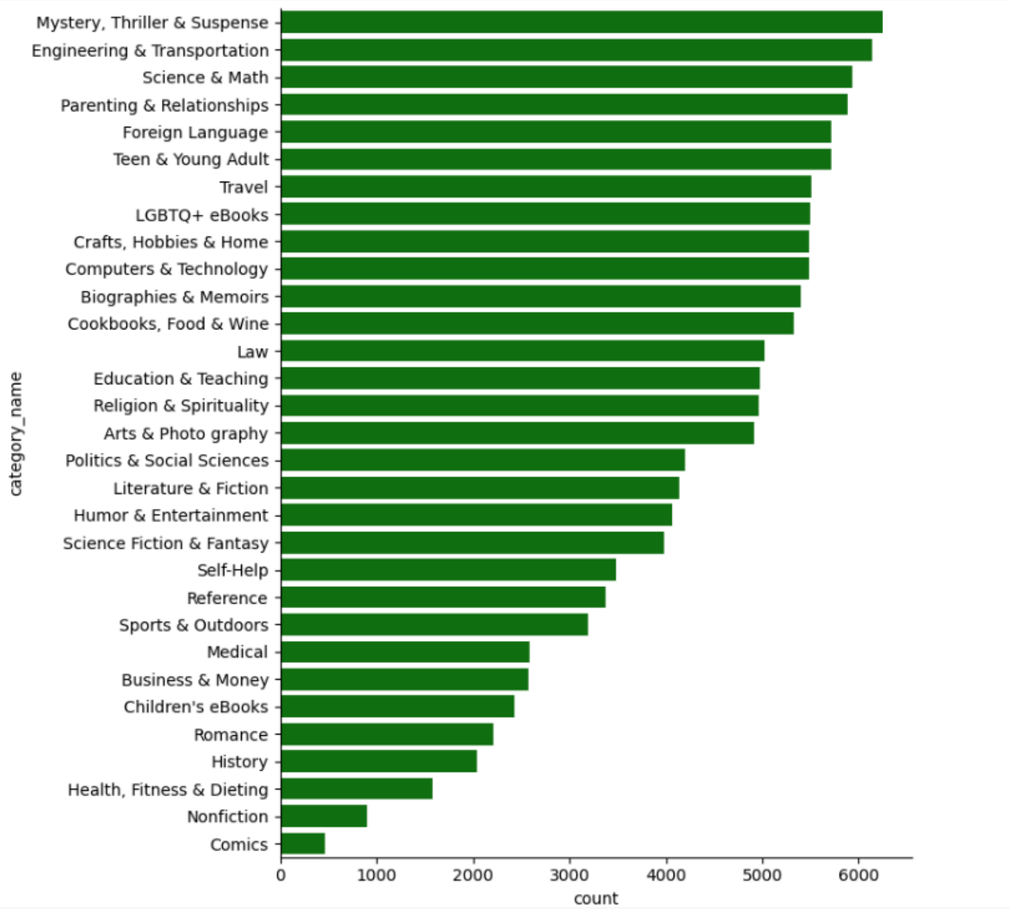


Rating Stars

- Most books have ratings between 4.2 and 4.8, showing a high concentration of positive reviews.
- Very few books are rated below 4.0.
- Bestsellers tend to have slightly higher ratings, often above 4.5.
- Rating alone isn't enough to predict success, but high ratings are a strong positive signal.

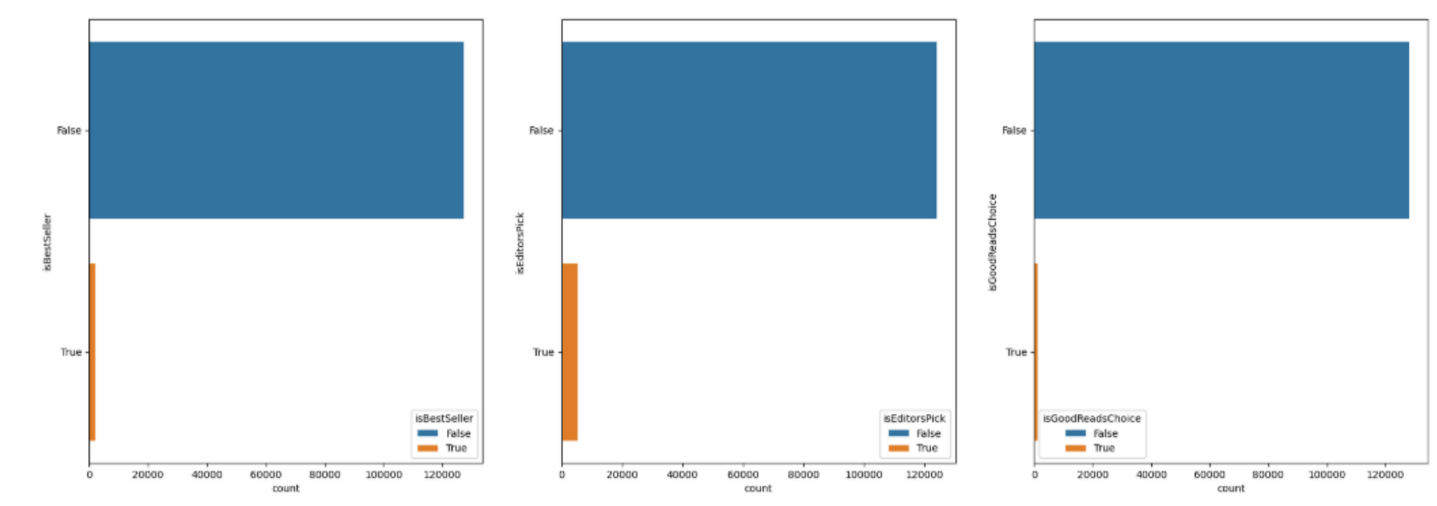


Genres



- Top Categories: Mystery, Thriller & Suspense (6245) and Engineering & Transportation (6138) dominate, reflecting high demand.
- Middle-Tier: LGBTQ+ eBooks (5493) and Crafts, Hobbies & Home (5488) highlight a variety of interests.
- Lower-Tier: Nonfiction (896) and Comics (471) show niche or limited content.
- Observation: Strong focus on entertainment and technical categories; potential growth in underrepresented areas.

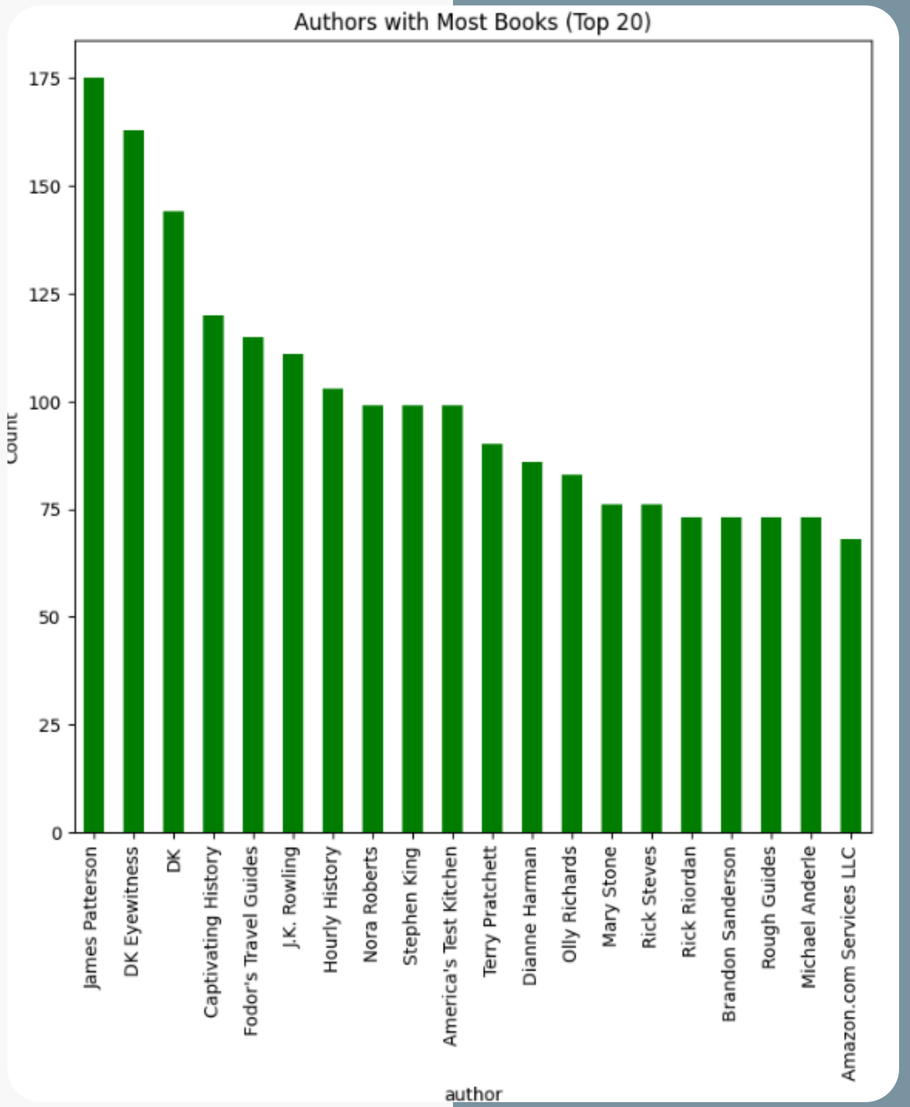
Category



- BestSellers: Very few books (2,141) are bestsellers, showing exclusivity.
- Editors' Picks: A larger but still small group (5,408), indicating notable recognition.
- Goodreads Choices: The rarest category (1,328), emphasizing its prestige.
- Observation: Most books don't belong to these "prestige" labels, highlighting their rarity.

Authors

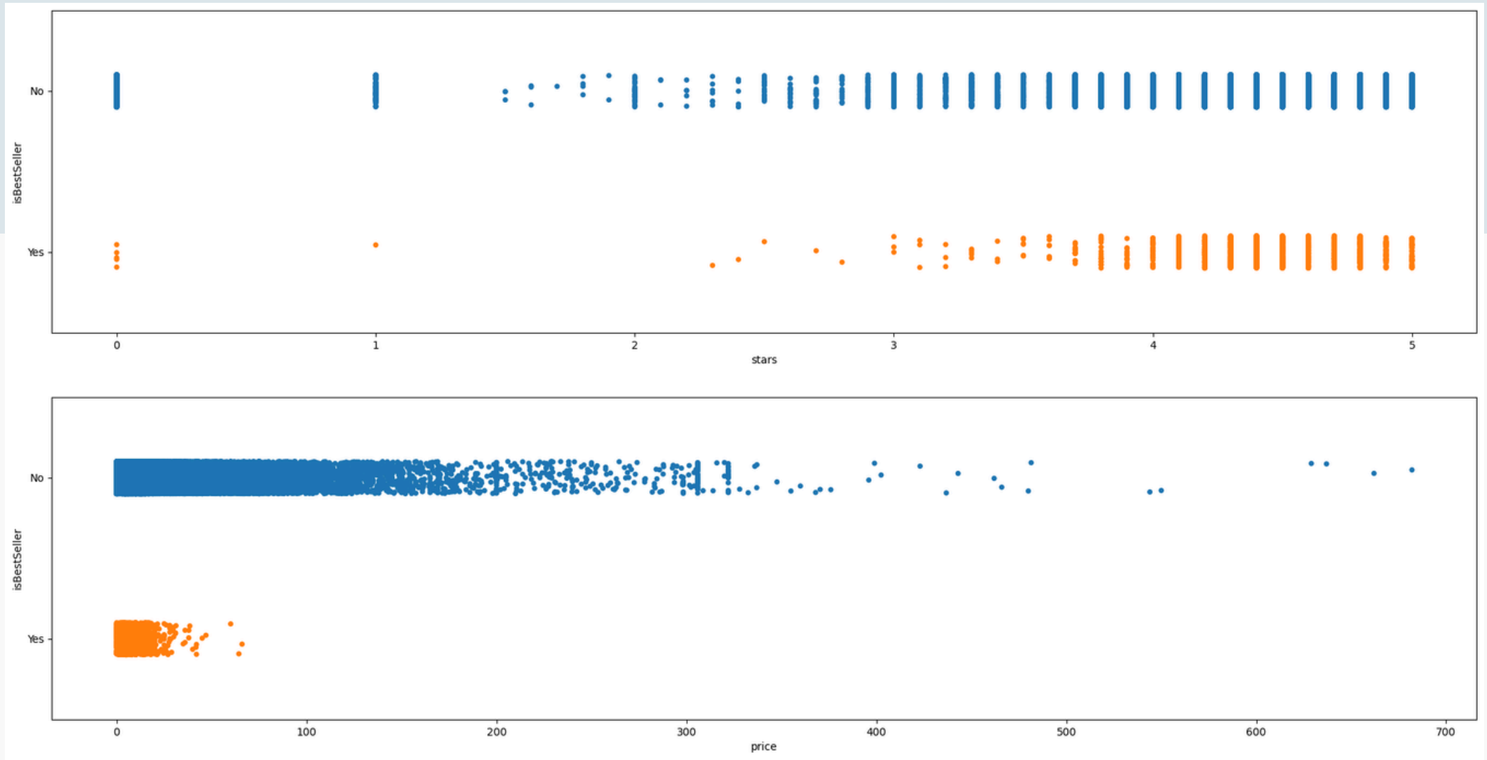
- Number of unique authors is 72096, which is a significant amount, implying that there is a wide variety of writers.
- The top authors vary widely in genre, showcasing diversity in literary production and reader interest. James Patterson, DK, and J.K. Rowling are standout names.



Bivariate

Numeric Variables vs isBestSeller

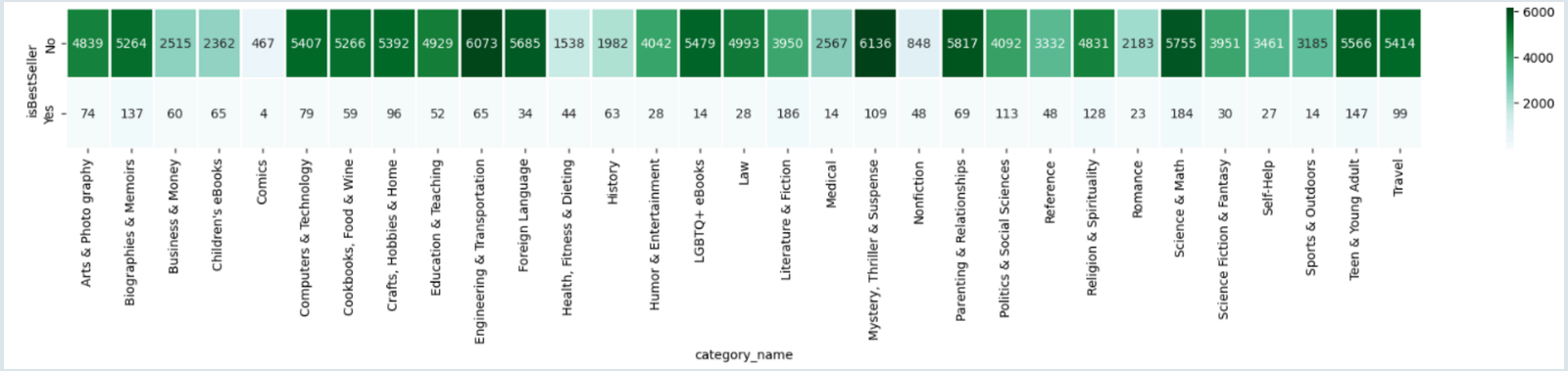
- **Stars Distribution:** Bestseller books generally cluster around higher star ratings, while non-bestsellers show more variability.
- **Price Distribution:** Non-bestsellers span a wider range of prices, while bestsellers are concentrated in specific price points.

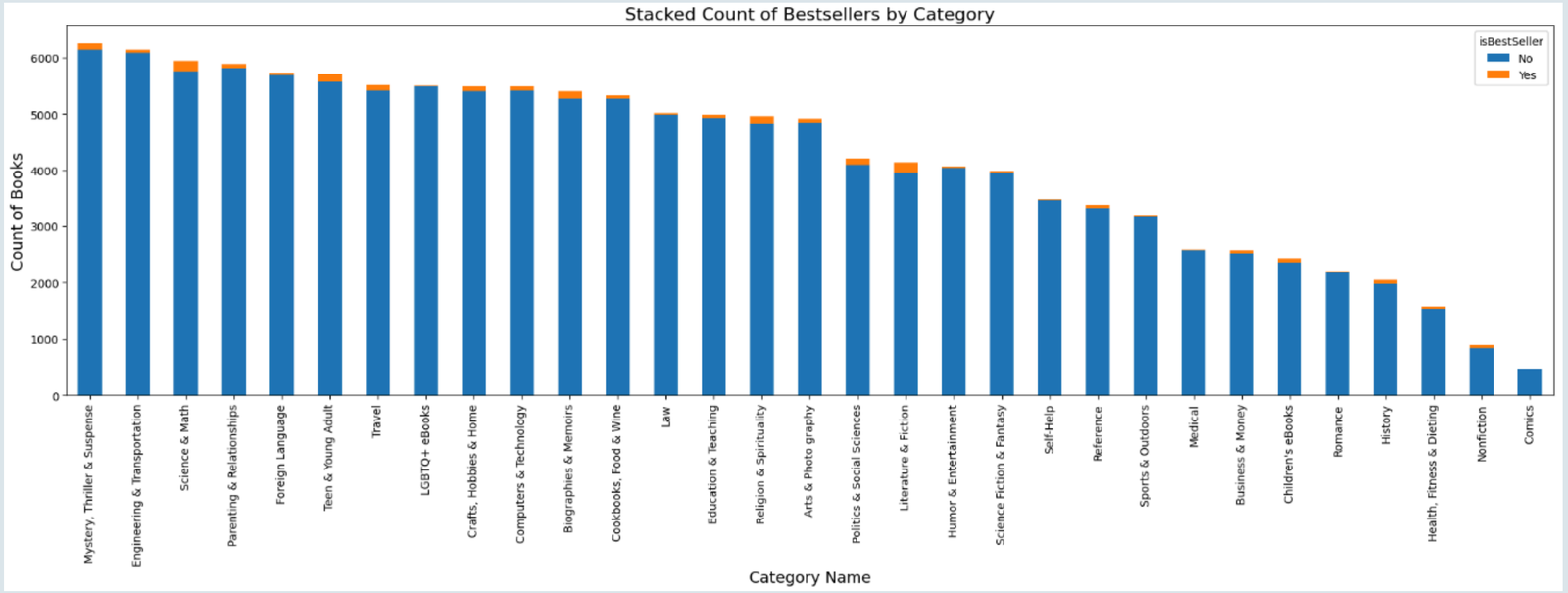


Category vs. isBestSeller

Genres like **Science & Math** and **Literature & Fiction** a higher proportion of bestsellers

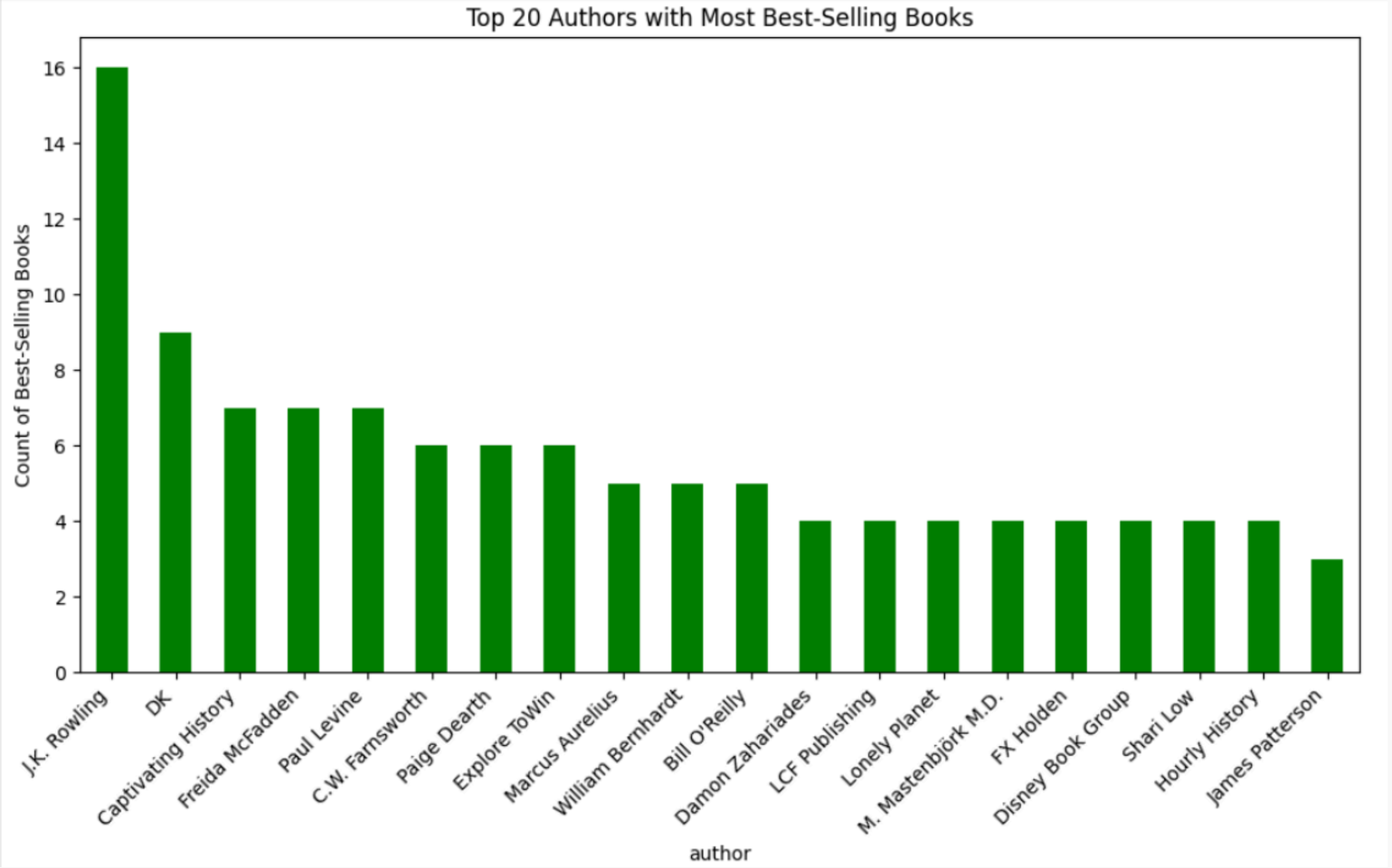
→ **Genre plays a role in a book's popularity**, likely due to reader demand and market size.





Authors with most best selling books

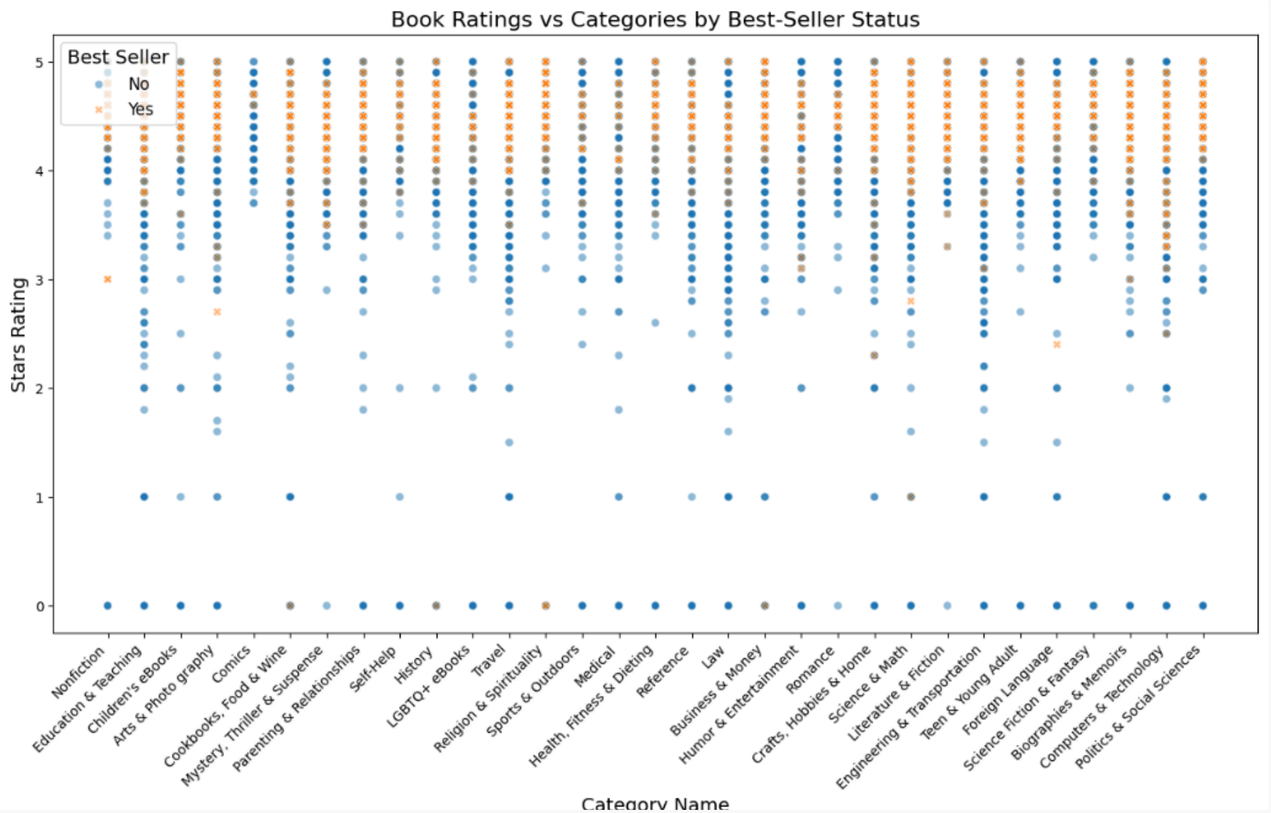
The chart highlights the strong presence of J.K. Rowling at the top, with a steady decline in bestseller counts among other authors. This distribution suggests certain authors consistently captivate readers and maintain their popularity over time.



Multivariate

1. Star Ratings Across Categories:
   - Most categories display star ratings: **4 - 5**
   - **Mystery, Thriller & Suspense** and **Science Fiction & Fantasy:** high ratings for both best-sellers and non-best-sellers.
2. Best-Seller Status Comparison:
   - Best-sellers: dominate high ratings in **Parenting & Relationships** and **Engineering & Transportation.**
   - Non-best-sellers: appear **more dispersed** across the rating range

   → greater variability in quality.



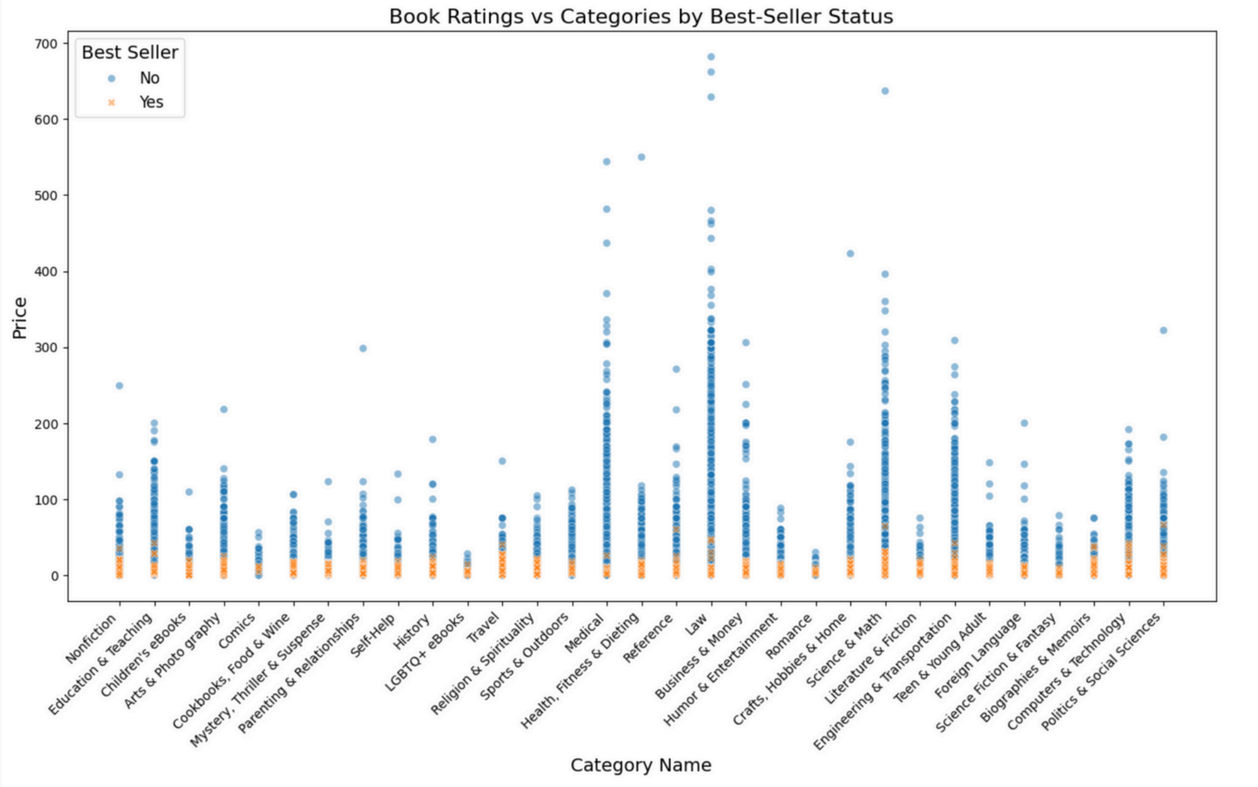
Multivariate

1.Best-Sellers:

- Concentrated in low ranges of prices

2. Non-Best-Sellers:

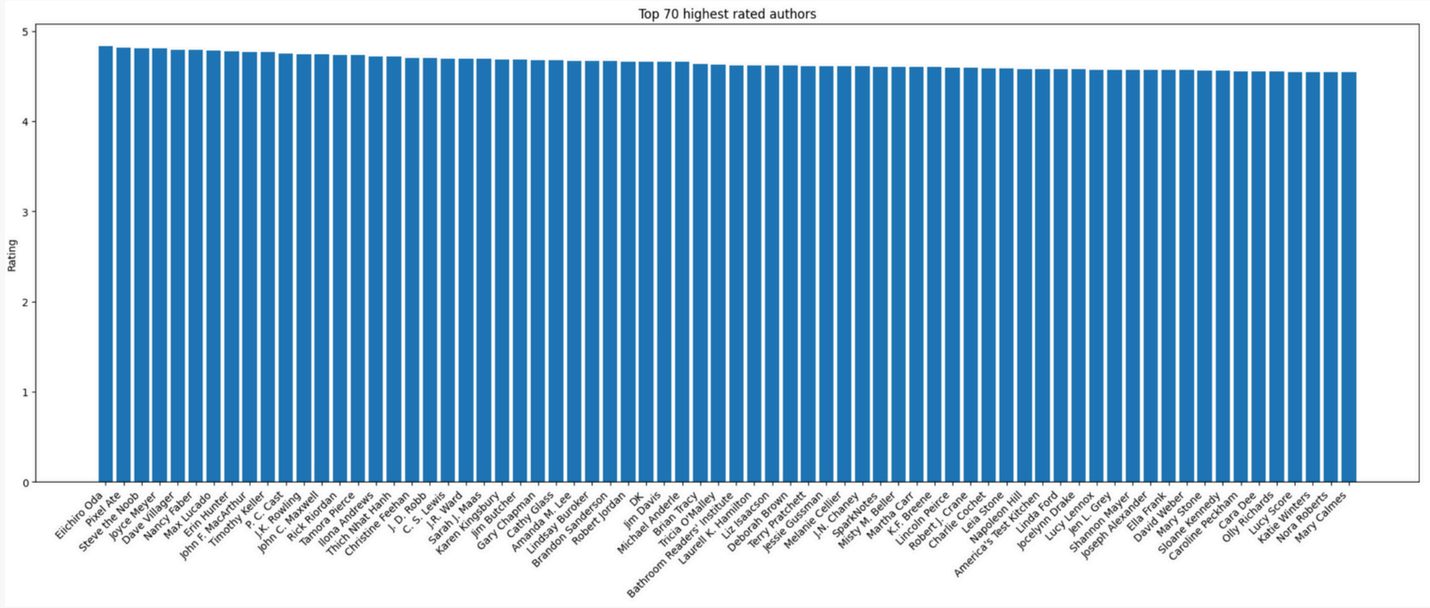
- Wider spread in prices, with some reaching over **700**.
- Variability suggests non-bestsellers often explore diverse pricing strategies.



Feature Engineering

Create “top_author” column

1. Category:
- Written at least 30 books
- Top 70 authors with highest stars rating

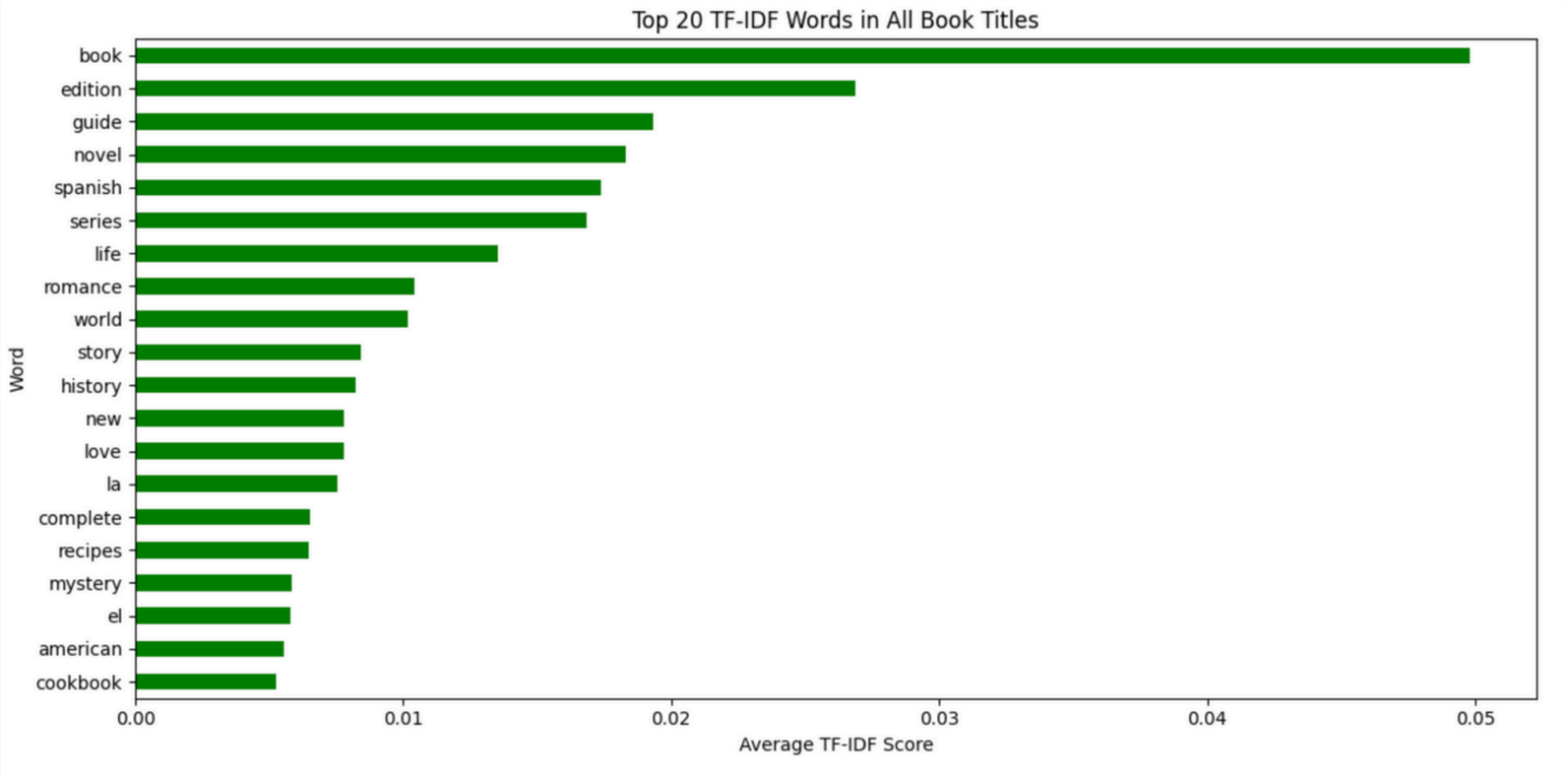


top_author
False
False
False
False
False

Feature Engineering

Create Term Frequency–Inverse Document Frequency (TF-IDF) vector for book titles

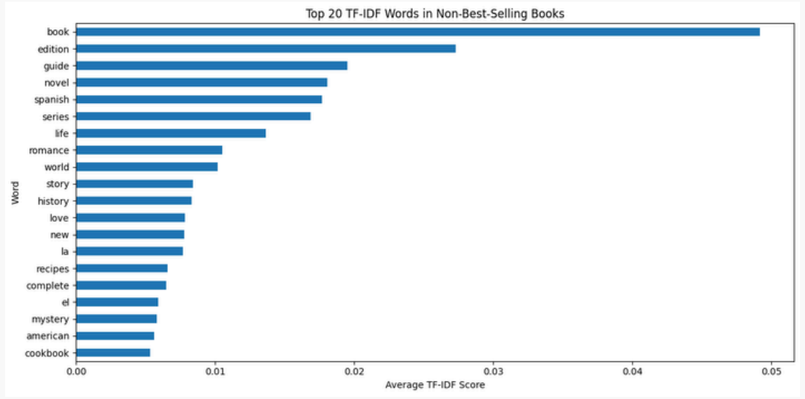
- Calculate TF-IDF for all titles and create a new DataFrame
- Mean TF-IDF scores for each word

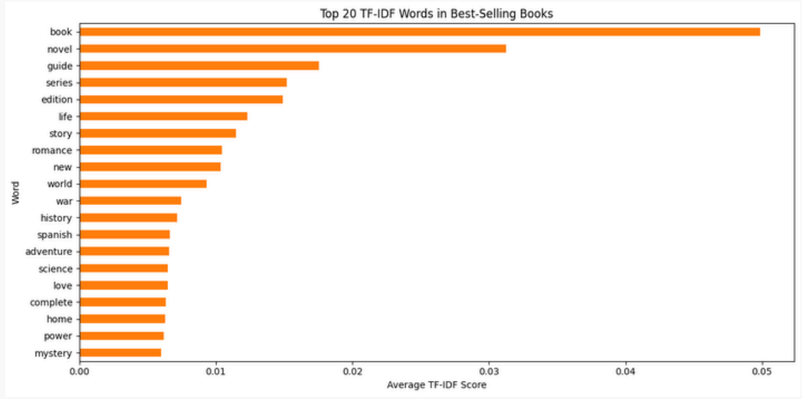


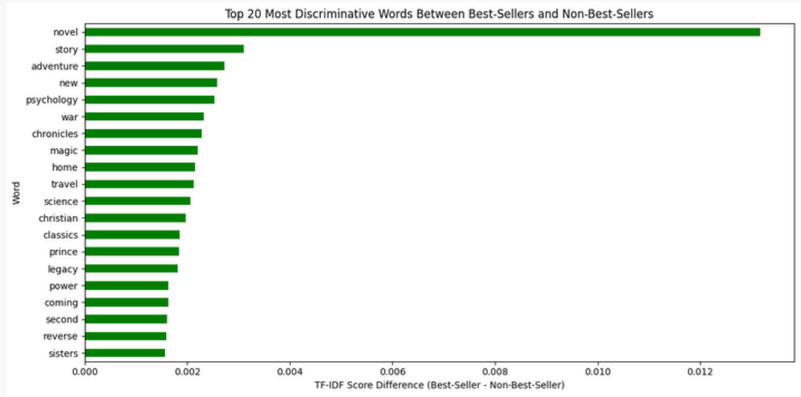
Feature Engineering

Create Term Frequency–Inverse Document Frequency (TF-IDF) vector for book titles

- Mean TF-IDF score for each words in bestseller books and non bestseller books
- Differences between TF-IDF score for bestseller books and non best seller books' title







Feature Engineering

One Hot Encoding

- Use pd.get_dummies
- Drop 1 column to avoid multicollinearity

```
#One Hot Encoding for genres
book_data_encoded = pd.get_dummies(book_data_cleaned, columns= ['category_name'])
pd.set_option('display.max_columns', None)  # Show all columns
pd.set_option('display.width', None)
shape = book_data_encoded.shape
print('DataFrame dimension: ', shape)
book_data_encoded.info()
book_data_encoded.head()
```

category_name_Arts & Photo graphy	category_name_Biographies & Memoirs	category_name_Business & Money	category_name_Children's eBooks	category_name_Comics	category_name_Computers & Technology	category_name_Cookbooks, Food & Wine	ca
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	
False	False	False	False	False	False	False	

Model for Predicting Best Seller books

Classification Tree

Model Training

- Feature Combination:
  - Text features (TF-IDF) and numeric features are combined into a single matrix for training and testing.
- Decision Tree:
  - A DecisionTreeClassifier is trained using predictors to give the response.

Visualization:

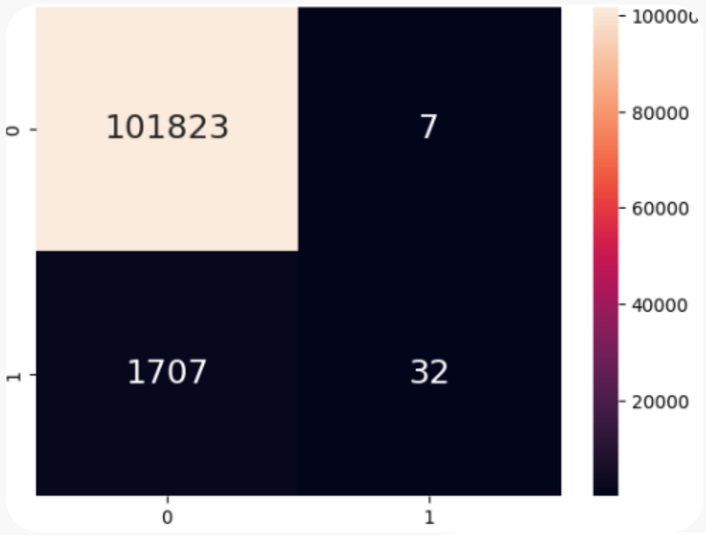
- Decision Tree Plot:
  - Tree visualization is created with clear labels for feature names and class outcomes ("No," "Yes").
  - Enables interpretation of classification rules and feature importance.

Data Preparation:

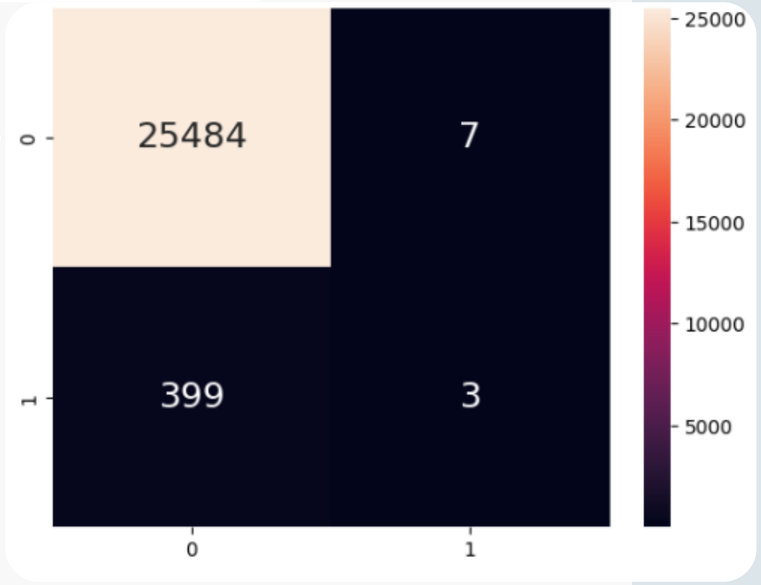
- Response and Predictors:
  - Predictors: Title + Author +Genre + Star rating + Price
  - Response: isBestSeller.
- Train-Test Split:
  - Dataset split: 80% training, 20% testing, ensuring robust evaluation.
- TF-IDF Vectorization:
  - Text data (title) is transformed into numeric features using the TF-IDF method.
  - Sparse matrices are created for numeric features for efficiency.

Model 1:

Train Data Analysis



Test Data Analysis



Classification Tree with max depth equal to 4

- Model 1 predicts significantly accurate results for non-bestseller books. However, it misclassified most of the best-selling books. This is because the dataset contains only a small proportion of bestseller books.
- To improve our model, we will balance the dataset by choosing some representatives for non-bestseller books while resampling bestseller books.
- **Accuracy of 98.35% in Train Data and 98.43% in Test Data Analysis.**

Data Balancing

Elbow Method

Find the point when Within Cluster Sum of Squares stops decreasing significantly, which gives the ideal number of clusters for data.

K-Means Clustering Implementation

Use the k-means clustering model to group the books into clusters.

Adding Cluster Labels and Previewing Data

Create a new column "cluster" in book_data_clustering.

Resampling Non-Bestsellers Across Clusters

From each cluster, we would select 4% of the books.

Resampling bestseller books

Resample bestseller books data so that they have the same amount of data as non-bestseller books.

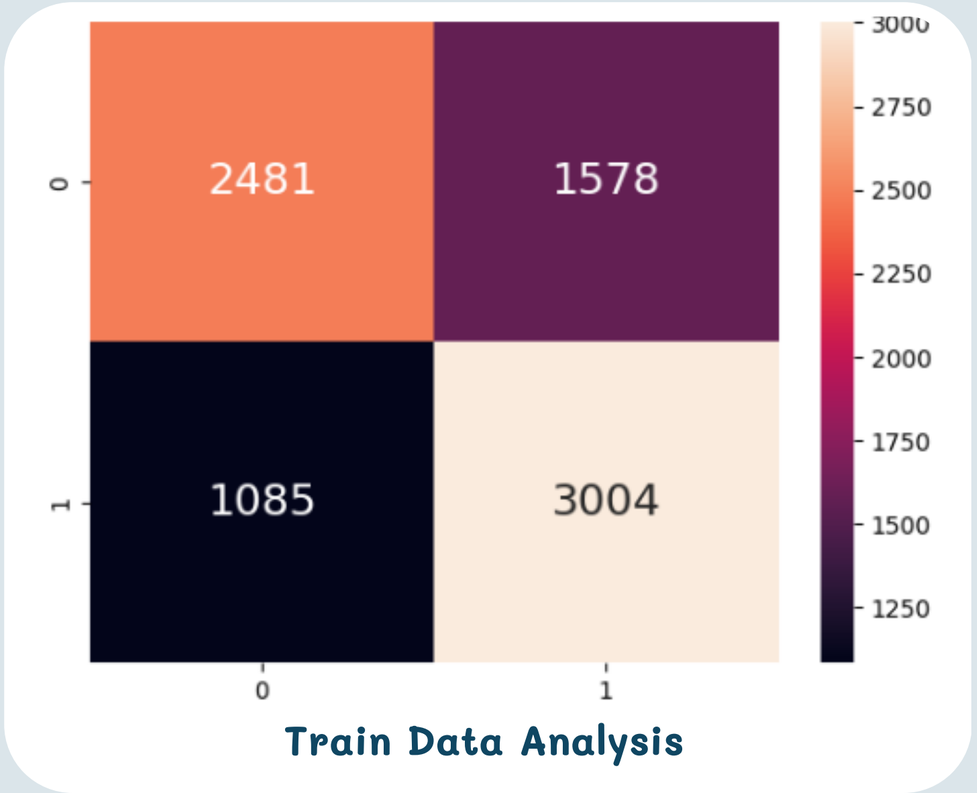
Balancing Bestseller and Non-Bestseller Classes

Combine the sampled_bestseller DataFrame and sampled_non_bestseller DataFrame to create the balanced DataFrame

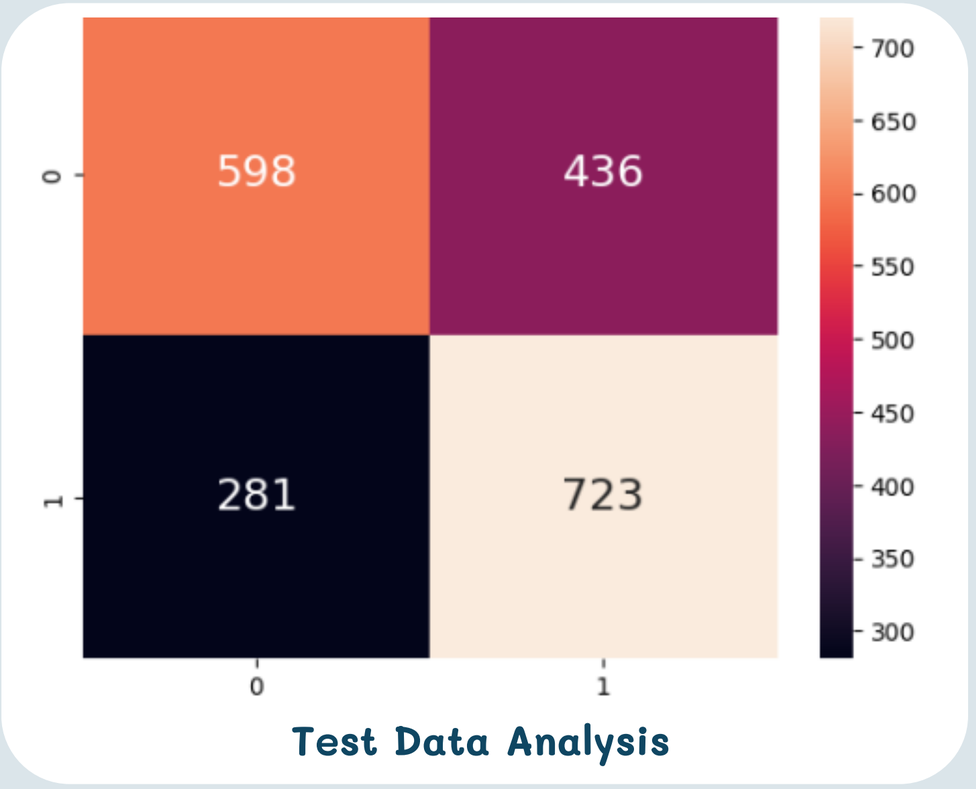
Model 2

Use the book_data_balanced DataFrame and Classification Tree model (max depth = 4)

- The book_data_balanced DataFrame significantly increases True Positive Rate and decreases False Negative Rate.
- Although the Accuracy, True Negative Rate decrease and False Positive Rate increases, it is because the original data was dominated by non bestseller books, not because the model was better.
- **Accuracy of 67.32% in Train Data, and 64.82% in Test Data Analysis.**



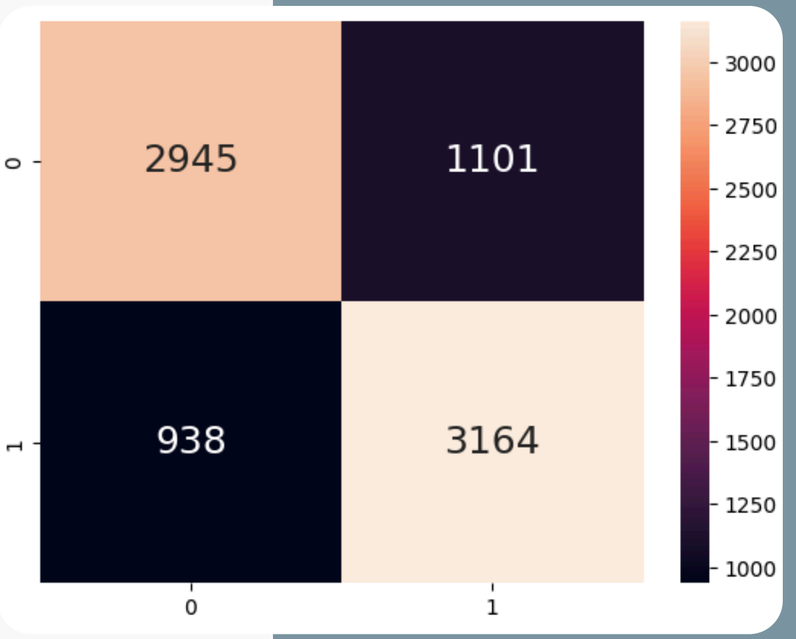
Train Data Analysis



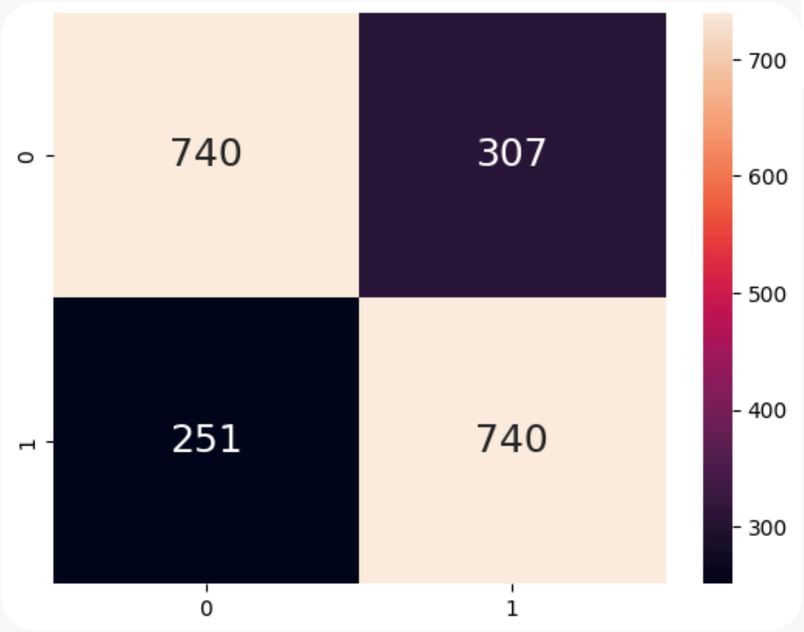
Test Data Analysis

Use the book_data_balanced and Classification Tree Model, but find the optimal depth of Tree.

- Hyperparameter Tuning for Decision Tree
- The new classification tree model would have max depth equal to 10.
- Improve the accuracy for both train data and test data.
- **Accuracy of 74.98% in Train Data and 72.62% in Test Data Analysis.**



Train Data Analysis



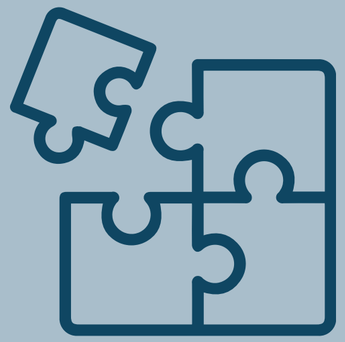
Test Data Analysis

Model 3

Random Forest Tree

Decision Trees

- Random Forest builds an ensemble (a collection) of decision trees, each trained on different subsets of data.



Bootstrap Aggregation

- Each tree is trained on a random sample with replacement (bootstrap sample) from the original dataset.

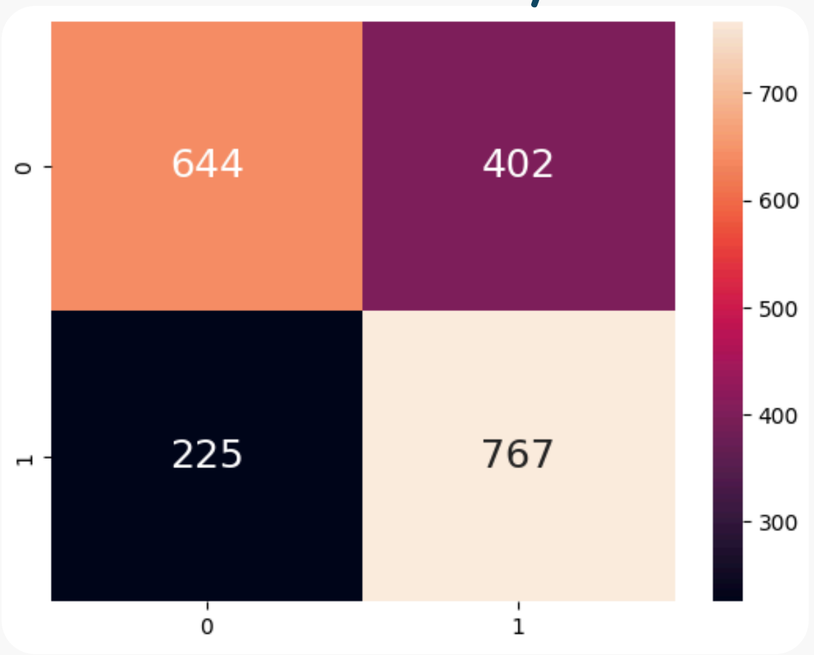
Feature Randomness

- At each split in the tree, only a random subset of features is considered, adding diversity among trees and reducing correlation.

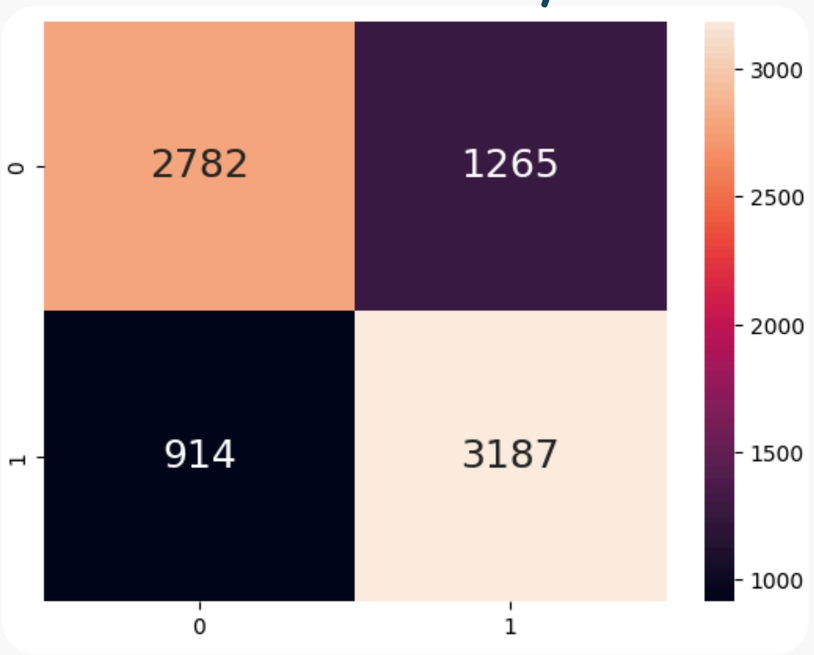
Model 4

Random Forest with max depth equal to 4 and 100 trees

Test Data Analysis



Train Data Analysis



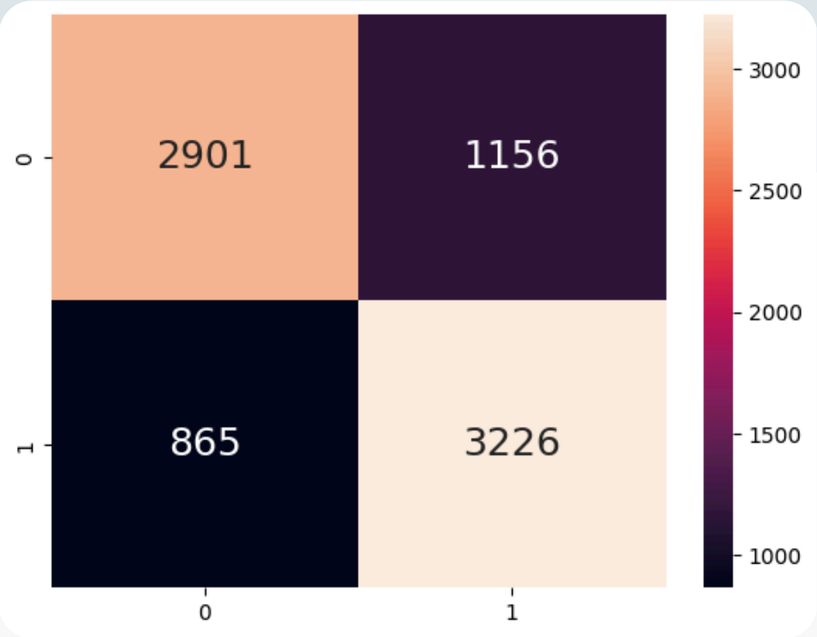
- Changing from Classification Tree to Random Forest does not significantly improve the model's performance. However, we have just created Random Forest with max depth of 4 and 1000 trees. We can change the parameters to see the differences.
- **Accuracy of 73.26% in Train Data and 69.23% in Test Data Test.**

Model 5

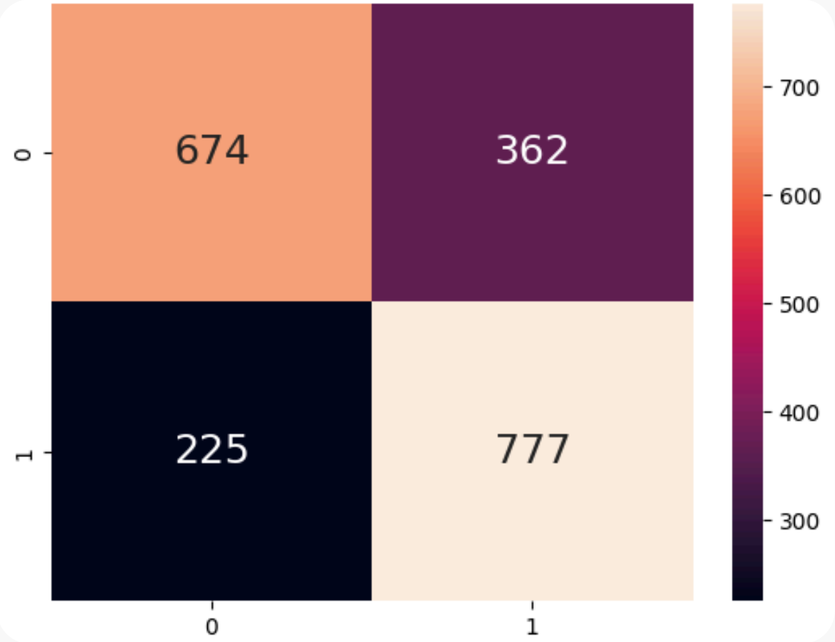
Random Forest Model (1000 trees, max-depth: 4)

- When we increase the number of trees in Random Forest to 1000, the performance of the model 5 increase in both predicting bestseller and non bestseller books comparing to model 4.
- **Accuracy of 75.20% in Train Data and 71.20% in Test Data Analysis.**

Train Data Analysis

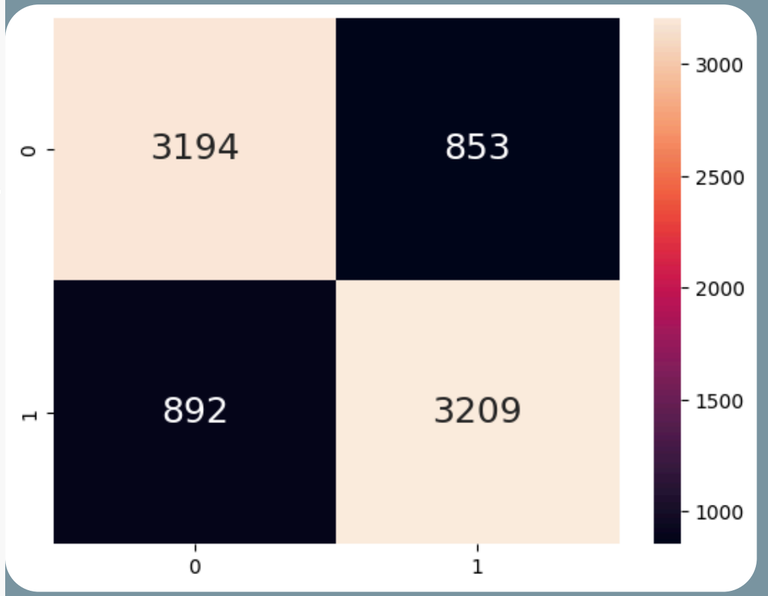


Test Data Analysis

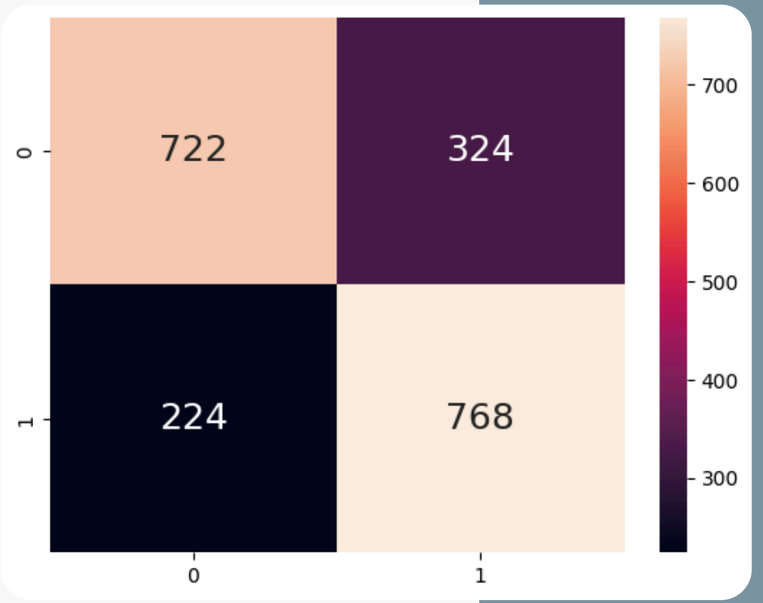


Model 6

Train Data Analysis



Test Data Analysis



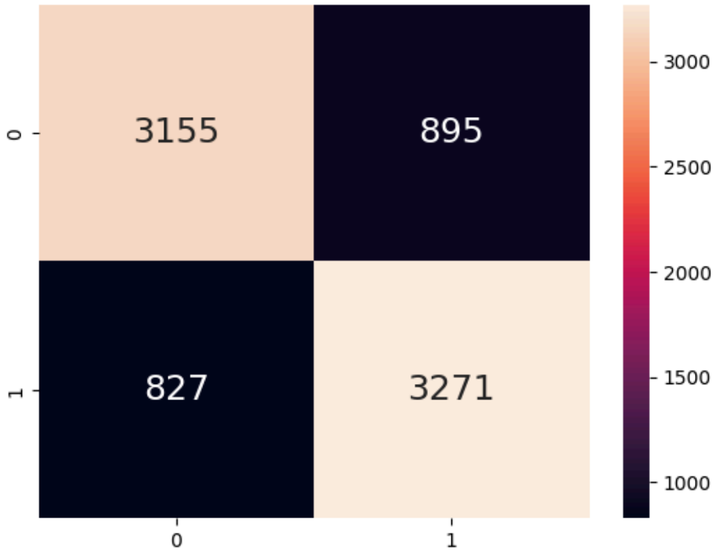
Random Forest Model (100 trees, max-depth: 10)

- When we increase the max depth of trees in Random Forest to 10, the performance of the model 6 increase in both predicting bestseller and non bestseller books comparing to model 4. Moreover, we can see that the improvement when increasing max depth is larger than when increasing number of trees.
- **Accuracy of 78.58% in Train Data and 73.11% in Test Data Analysis.**

Model 7: GridSearchCV

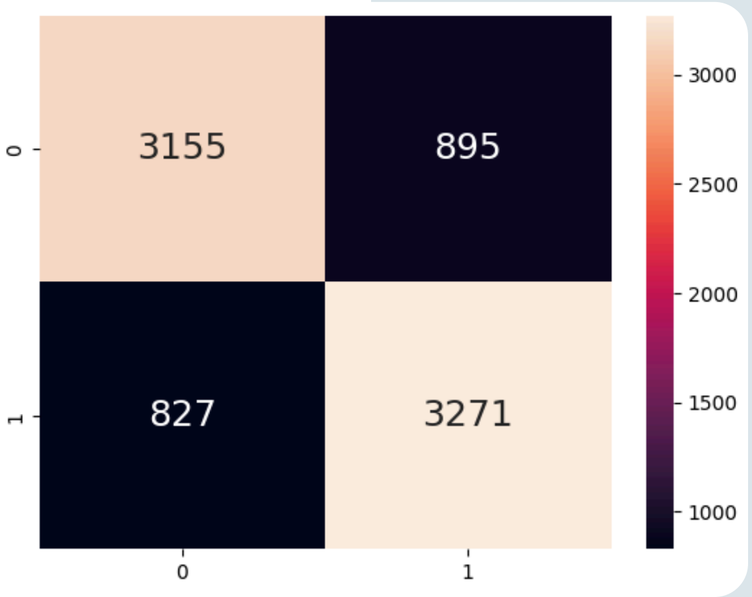
- **Objective:**
  - Optimize the parameters of the Random Forest Classifier (n_estimators and max_depth) to achieve the best accuracy for predicting bestseller status.
- **Hyperparameter Grid:**
  - n_estimators: Represents the number of trees in the Random Forest, ranging from 100 to 1000 in increments of 100.
  - max_depth: Indicates the maximum depth of trees, evaluated from 2 to 10.

- **Cross-Validation:**
  - 5-fold CV: Ensures robust evaluation by splitting the training data into 5 subsets, training on 4 subsets, and validating on the remaining 1, repeated 5 times.
  - Scoring Metric: Accuracy is used to assess model performance.
- **Training the GridSearchCV:**
  - hpGrid.fit(X_train_combined, y_train): Trains Random Forest models across all combinations of n_estimators and max_depth specified in the grid.



Model 7: GridSearchCV

- Using GridSearchCV, we found that the optimal Random Forest Model is the model with max depth equal to **10** and **800** trees. Comparing to previous models, this model have higher accuracy in both train set and test set, as well as more balance true positive rate and true negative rate.



- **Accuracy of 78.87% in Data Train and 74.24% in Test Data Analysis.**

Conclusion

- Price is a strong indicator, as reflected across models.
- Text-based features provide valuable context
- Resampling imbalanced data ensured better representation of minority classes
- Random Forest Classifier, when fine-tuned using hyperparameter tuning (e.g., GridSearchCV), performed consistently well.
- Decision Trees provide interpretable results, clearly outlining factors, but require constraints to prevent overfitting.
- Combining textual and numeric features: comprehensive approach, enabling acceptable accuracy and recall rates.

What We Learned

Data visualization is key for spotting trends and outliers you can't see in raw numbers.

Handling the imbalanced datasets

Being a bestseller isn't just about content — it's also about language, branding, and potentially timing.

Text data holds a lot of predictive power — even simple TF-IDF scoring can reveal big differences in language between successful and average books.

Clustering adds structure to complex datasets and can help categorize patterns in consumer markets.

References

- **Dataset:** https://www.kaggle.com/datasets/asaniczka/amazon-kindle-books-dataset-2023-130k-books

Thank you