Predict Your Upcoming Audiobook Performance

Audiobooks have become increasingly popular in recent years, as people seek convenient and immersive ways to consume literature. With the rise of platforms like Audible and Spotify, audiobook consumption has become more accessible than ever before. As a result, there has been a growing interest in understanding the factors that influence listeners' ratings of audiobooks. My project involves analyzing and building a machine learning model to predict the ratings of a new audiobook. By leveraging data on factors such as the price, the length of the audiobook, and the title, and more, I aim to develop a model that can accurately predict how well a new audiobook will be received by listeners. This research has the potential to provide valuable insights into the factors that influence audiobook ratings and could help publishers and authors better understand their audience and improve their products.

1. Data

This Kaggle dataset is gathered from Audible and represents audiobooks from 1998 until 2025.

https://www.kaggle.com/datasets/snehangsude/audible-dataset

2. Method

An audiobook rating is typically represented by a ranking system of 1 to 5 stars. To be considered successful, an audiobook would ideally receive 5 stars. Regardless of the number of ratings received, the overall rating is determined by averaging the stars. In this project, we will evaluate the success of an audiobook based on whether it has received a 5-star rating.

3. Data Cleaning

Data Cleaning Notebook: https://github.com/expl0ding/Capstone2/blob/main/Capstone2_DataClean.ipynb

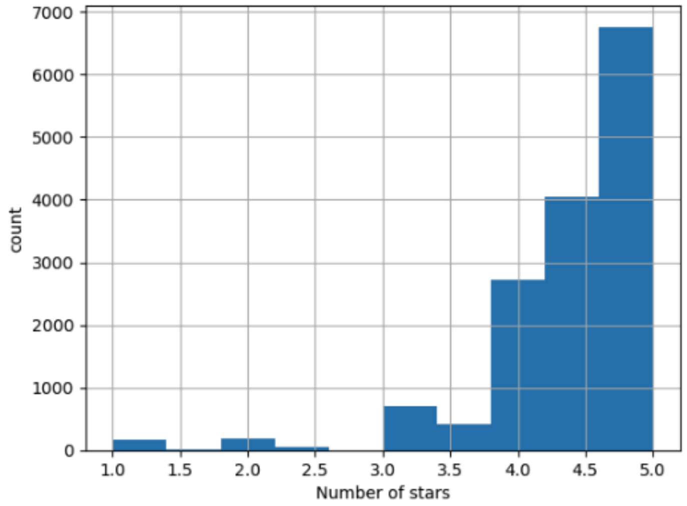
Problems and Solutions:

1. Problem 1: The number of stars and number of ratings were in the same column along with non-numerical characters. There were also columns that had ‘Not Rated Yet’. Solution: I built a for loop to split up the “stars” and “ratings” information, as well as remove the strings, and keep only the numerical values. The columns with “Not Rated Yet” were filled with the value 0.
2. Problem 2: The time column was filled with values in this format: “2 hrs and 20 mins”. Solution: I wrote a for loop that converted the time column into a numerical integer converted hours to minutes.
3. Problem 3: The date was in this format: “23-02-11” Solution: I extracted and kept only the year.
4. Problem 4: The narrator and author rows had to be cleaned to not include “Writtenby:” and “Narratedby:” Solution: I wrote a loop to eliminate “Writtenby:” and “Narratedby:”

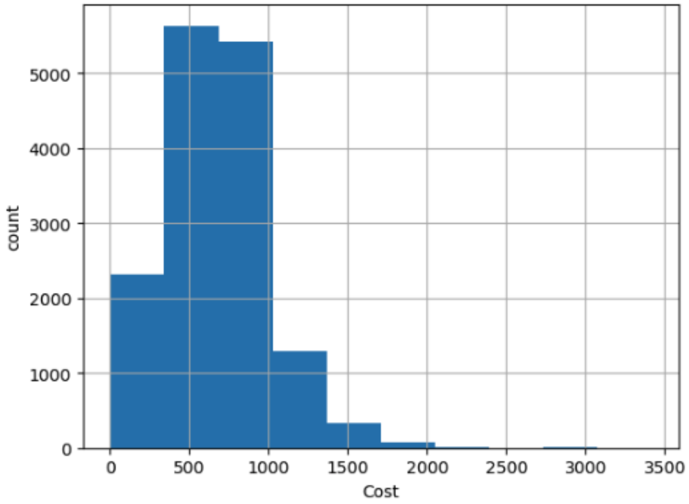
4. EDA

EDA Notebook: https://github.com/expl0ding/Capstone2/blob/main/Capstone2_EDAv2.ipynb

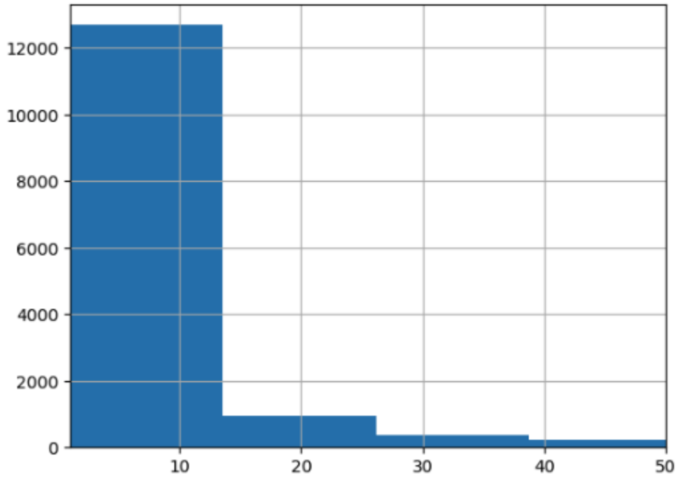
Most audiobooks seem to have received above 4.0 stars.



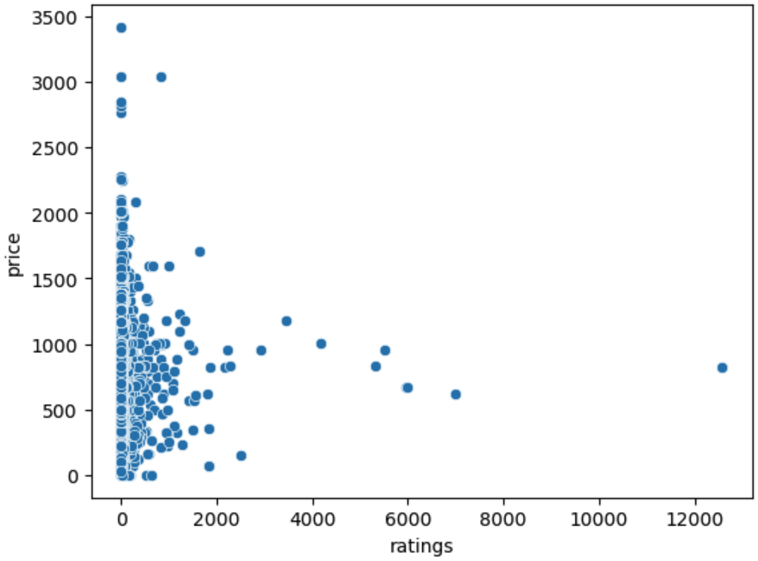
And most books are under 12USD /100 IR



Many of the audiobooks had 10 or fewer ratings



There were a couple outliers that had well over 2000 ratings as shown here.



Those books with an oddly large number of ratings are all non-fiction, with some of them being biographies and some of them being self-help/educational books.

- Elon Musk by AshleeVance
- Becoming by MichelleObama
- Atomic Habits by JamesClear
- Sapiens by YuvalNoahHarari
- 21 Lessons for the 21st Century by YuvalNoahHarari
- The Psychology of Money by MorganHousel
- Rich Dad Poor Dad by RobertT.Kiyosaki
- The Psychology of Money by MorganHousel
- Atomic Habits by JamesClear
- How to Win Friends and Influence People by DaleCarnegie
- Can't Hurt Me by DavidGoggins
- Ikigai by HéctorGarcía, FrancescMiralles
- The Subtle Art of Not Giving a F*ck by MarkManson
- Life's Amazing Secrets by GaurGopalDas

4. Machine Learning

Modeling Notebook: https://github.com/expl0ding/Capstone2/blob/main/Capstone2_Modeling-Final_v2.ipynb
Model Metrics File: https://github.com/expl0ding/Capstone2/blob/main/Capstone2_ModelMetrics.txt

After experimenting with several models, such as Decision Tree, Logistic Regression, Random Forest, Cat Boost Classifier, and Gradient Boosting, the Gradient Boosting model had the highest accuracy of 57%, making it the most suitable model for our classifier problem.

Based on this model, the most important features for accurate classification are whether the book title includes an adjective and the length of the audiobook. It is worth noting that older books tend to receive lower ratings, but this should not be a cause for concern as future books cannot be published in the past.

5. Final Predictions + Improvements

The final model has been designed to offer a comprehensive solution to audiobook success prediction. Users will input all the features of their upcoming audiobook, ranging from the title and author to the narrator, genre, length, and more, to determine whether it is likely to be successful or not. This advanced feature is particularly useful for publishers, authors, and producers who want to make informed decisions about which audiobooks to invest in and promote. This approach saves time, resources, and money by allowing users to avoid investing in audiobooks that are unlikely to perform well. Overall, the final model is an indispensable tool for anyone looking to succeed in the audiobook industry by making informed decisions based on data-driven insights.

Adding the genre feature to the dataset would be a significant improvement to the accuracy of the model. Some of the most-rated outliers were in the 'non-fiction' category, indicating that genre plays a crucial role in the success of an audiobook. By including the genre feature, the model could accurately capture the nuances and characteristics unique to each genre, enabling it to make more precise predictions.

Despite this limitation, the model is still effective in making predictions about the success of audiobooks based on other features included in the dataset. The insights provided by the model can still be valuable for stakeholders in the audiobook industry, however, as the dataset evolves and new features are added, the accuracy of the model will continue to improve, making it an even more powerful tool for predicting audiobook success.