Alice Fan

Capstone 2

Project Report

The audiobook market is growing rapidly and understanding the characteristics and preferences of an audiobook consumer is crucial in the success of releasing an audiobook. My client is an author who wants to release an audiobook. They want to uncover key findings within the industry before their release.

The goal of this analysis is to uncover any patterns and trends in a dataset from Audible to gain a deeper understanding of the marketplace. We will investigate if there are any correlations between audiobook length, price, language, ratings and other variables. We will also identify factors that influence the success of audiobooks. Through this analysis, we will be able to provide actionable and valuable insight for the client for their future release.

Here is the link to the Audible dataset: <https://www.kaggle.com/datasets/snehangsude/audible-dataset>

Outline of Steps

1. Import the dataset into Jupyter Notebook for further analysis.
   1. Investigate and explore the data
   2. Notice any missing values and inaccuracies – decide on if these need to be addressed immediately or after some deeper analysis
2. Understand overall structure of the data (identify outliers) by exploring the distribution of values for each column using data visualization techniques with histograms & bar charts.
   1. Create visualizations for each column and decide which to explore further
3. Investigate the relationship between different variables in the dataset.
   1. Create scatter plots to examine the relationship between various features:
      1. Audiobook Length and Price
      2. Length of Time Since Release vs. Ratings/Stars – do we need to adjust the analysis based on time periods?
      3. Time Since Release vs Price
      4. Title vs Number of Ratings/Stars
      5. Price vs Number of Ratings
      6. Release Date vs Number of Ratings/Stars, etc.
      7. Authors vs Ratings/Stars
      8. Narrators vs Ratings/Stars
   2. Create a correlation matrix to see the relationships between all the variables in the dataset.
   3. Explore the Titles
      1. Explore the variations in:
         1. Length of title
         2. If title includes verbs, nouns, numbers, dates
4. Use statistical analysis to identify patterns or trends in the data.
   1. Regression analysis to identify factors that influence the number of ratings a book receives
   2. Cluster analysis to group similar books based on their characteristics
5. Build models to explore and determine features and how it would perform
   1. Split training and testing data
   2. Random Forest Model/Cross Validation
   3. Fit data into model
   4. Create visualizations for interactive dashboards for ease of sharing findings with the client
6. Extract insights from title, author and narrator columns using text mining techniques
   1. sentiment analysis to determine the overall tone of the books
   2. topic modeling to identify common themes among the books
7. Summarize your findings and provide recommendations based on your analysis.
   1. Identify key features
   2. Identify ‘golden standard’ of audiobooks as it relates to each feature