Project Title	**Audible Insights: Intelligent Book Recommendations**
Skills take away From This Project	Python scripting, data cleaning, Exploratory Data Analysis (EDA), Machine Learning,NLP, Streamlit, AWS deployment
Domain	Recommendation Systems

Problem Statement

Design a book recommendation system that retrieves book details from given datasets, processes and cleans the data before applying NLP techniques and clustering methods and builds multiple recommendation models. The final application will allow users to search for book recommendations using a user-friendly interface deployed with Streamlit and hosted on AWS.

Business Use Cases

Personalized Reading Experience:

- Help readers discover books tailored to their preferences based on their reading history, genres, or authors.

Enhanced Library Systems:

- Libraries and bookstores can use recommendations to improve book borrowing/sales based on popular or similar reads.

Improved Author/Publisher Targeting:

- Provide authors and publishers with data-driven insights about popular genres, reader preferences, and high-demand books.

Reader Engagement:

- Increase reader engagement by suggesting trending or highly rated books across different categories.

Approach

1. **Data Preparation:**
    - Use two provided datasets containing book information, ratings, and user interactions.
    - Merge the datasets based on common attributes like book names and authors.

2. **Data Cleaning:**
    - Handle missing or inconsistent data by imputing or removing incomplete records.
    - Standardize formats for fields like genres and ratings.
    - Remove duplicate records.

3. **Exploratory Data Analysis (EDA):**
    - Analyze book genres, ratings distribution, and other trends.
    - Visualize key insights like the most popular genres, top-rated books, and trends in publication years.

4. **NLP and Clustering:**
    - Apply NLP techniques to extract features from book titles, descriptions, or reviews.
    - Use clustering algorithms (e.g., K-means or DBSCAN) to group books based on similarities in features.

5. **Recommendation System Development:**
    - **Build recommendation models using:**
        - Content-Based Filtering (based on book features like genres, descriptions, etc.)
        - Clustering-Based Recommendations
        - Hybrid Approaches

- Compare these models using evaluation metrics such as precision, recall etc.

6. Application Development:

- **Build a user interface using Streamlit to:**
  - Input user preferences (e.g., favorite genres or books).
  - Display personalized book recommendations.
  - Visualization and EDA.

7. Deployment:

- Deploy the application to AWS, ensuring accessibility and scalability.

Data Flow and Architecture

1. **Data Preparation:**
   - Merge and clean the provided datasets.
   - Store the processed data locally or in an AWS S3 bucket.
2. **Processing Pipeline:**
   - Clean and preprocess the data using Python libraries like pandas.
   - Perform feature engineering for recommendation models.
3. **Model Training:**
   - Develop models using libraries like scikit-learn or Surprise.
   - Save trained models for deployment.
4. **Deployment:**
   - Use Streamlit to create a user-friendly front end.
   - Host the application on AWS EC2 or Elastic Beanstalk.

Datasets:

Dataset 1: Audible_Catlog.csv

Dataset 2: Audible_Catlog_Advanced_Features.csv

Dataset Explanation:

Dataset 1: This dataset contains detailed information about various books, including:

- **Book Name:** Title of the book.
- **Author:** Name of the author.
- **Rating:** Average rating of the book.
- **Number of Reviews:** Count of user reviews.
- **Price:** Cost of the book.
- **Description:** Brief description or synopsis of the book.
- **Listening Time:** For audiobooks, total duration.
- **Ranks and Genre:** Rankings across various categories and genres.

Dataset 2: This dataset provides complementary information about books, including:

- **Book Name:** Title of the book.
- **Author:** Name of the author.
- **Rating:** Average rating of the book.
- **Number of Reviews:** Count of user reviews.
- **Price:** Cost of the book.

Exploratory Data Analysis (EDA)

Analyze trends and insights in the dataset, including:

- Distribution of ratings across genres.
- Most common genres and authors.
- Relationship between book ratings and review counts.
- Trends in publication years.

Key Visualizations:

- Bar charts for popular genres.
- Heatmaps to visualize correlations (e.g., ratings and reviews).
- Line charts for publication trends over time.

Questions to Be Answered:

Easy Level:

1. What are the most popular genres in the dataset?
2. Which authors have the highest-rated books?
3. What is the average rating distribution across books?
4. Are there trends in publication years for popular books?

5. How do ratings vary between books with different review counts?

Medium Level (Combination of Different Tables/Models):

1. Which books are frequently clustered together based on descriptions?
2. How does genre similarity affect book recommendations?
3. What is the effect of author popularity on book ratings?
4. Which combination of features provides the most accurate recommendations?

Scenario Based:

1. A new user likes science fiction books. Which top 5 books should be recommended?
2. For a user who has previously rated thrillers highly, recommend similar books.
3. Identify books that are highly rated but have low popularity to recommend hidden gems.

You can answer the above questions and also do your own analysis.

Results

By the end of this project, learners will achieve:

1. A merged and cleaned dataset ready for analysis and modeling.
2. Extracted text features using NLP and clustered books based on similarities.
3. Multiple recommendation models developed and evaluated.
4. A functional recommendation system accessible via a Streamlit interface.
5. A deployed application hosted on AWS.

Project Evaluation Metrics:

Data Cleaning Process:

- Evaluate how missing data, duplicates, and inconsistencies are handled.

Model Performance:

- Assess recommendation accuracy using metrics like RMSE, precision, and recall.

Application Functionality:

- Ensure the Streamlit interface is user-friendly and responsive.

Deployment Quality:

- Verify the application's accessibility and scalability on AWS.

Technical Tags:

Python, Machine Learning, Recommendation Systems, NLP, Clustering,Streamlit, AWS Deployment,Recommendation Systems

Deliverables:

Data Preparation:

- Evaluate how well the datasets are merged and cleaned.

NLP and Clustering:

- Assess the quality of feature extraction and clustering methods.

Model Performance:

- Evaluate recommendation accuracy using metrics like precision, recall, and RMSE.

Application Functionality:

- Ensure the Streamlit interface is user-friendly and responsive.

Deployment Quality:

- Verify the application's accessibility and scalability on AWS.

Timeline:

NOTE (for batch DS-WD-E-B15): Timeline for this project has been given 3 weeks (i.e. deadline 10th Feb), kindly give preference to your final project,

complete and submit your final project and then you can start working on this project.

References:

Project Live Evaluation Metrics	Project Live Evaluation
EDA Guide	Exploratory Data Analysis (EDA) Guide
Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
AWS recordings	AWS
Streamlit recordings (English)	Special session for STREAMLIT(11/...
Streamlit recordings (Tamil)	https://us06web.zoom.us/rec/share/JTr7DywhE1-SarjyIHBSCn4qnl7_uvJH6IGk06qAlkE0Ny1o_rqcq5FRFKuo93dm.iyM2o6l0h9aTUkNE
Streamlit documentation	Install Streamlit
Project Orientation (English)	Project Orientation Session : Audible...
Project Orientation (Tamil)	

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.
Note: Book the slot at least before 12:00 Pm on the same day

Timing: Monday-Saturday (4:00PM to 5:00PM)

Booking link :https://forms.gle/XC553oSbMJ2Gcfug9

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.
Note: This form will Open only on Saturday (after 2 PM) and Sunday on Every Week

Timing:
For DS and AIML
Monday-Saturday (05:30PM to 07:00PM)

Booking link : https://forms.gle/1m2Gsro41fLtZurRA

Evaluation Metrics : Project Live Evaluation

Project Created By	Verified By	Approved By
Vinodhini	Nehlath Harmain	Nehlath Harmain