A Session Bases Cross-Domain Recommendation System

Objective: The objective of this project is to build a recommendation system using content-based filtering combined with session-based tracking to enhance user experience. By analyzing item features and user interactions during a session, the system delivers personalized and relevant suggestions. The goal is to improve engagement and satisfaction by understanding user preferences in real time without relying on historical data from other users.

Introduction: This project involves the development of a web-based recommendation system that utilizes user interaction data—such as clicks and the time spent on items during a session—to provide personalized recommendations. The system adopts a **session-based tracking** approach, focusing on real-time user behavior rather than long-term user profiles. A **content-based filtering** technique is implemented to analyze item features and suggest similar content based on user preferences. Designed as a **cross-domain recommendation system**, it delivers recommendations across diverse categories including **Movies**, **Music**, and **Books**, enhancing the overall user experience by offering relevant suggestions across multiple domains.

Tech Stack:

- **Machine Learning:** NumPy, Pandas, Matplotlib, Seaborn, PyTorch
- **Frontend:** ReactJS, CSS
- **Backend:** Flask
- **Deployment:** Azure/Streamlit

Data:

- **Movie Data-base: IMDB Dataset**
- **Music Data-base: Spotify Dataset**
- **Books Data-base: GoodReads Dataset**

Data Collection & Preprocessing:

Data Collection:

The datasets for this project were sourced from HuggingFace, an open-source platform that provides high-quality, publicly available datasets. Three separate datasets were used to represent the different domains of the recommendation system—Movies, Music, and Books. These datasets were selected for their richness in content features, user interaction data, and domain diversity, making them suitable for building a cross-domain recommendation model.

Dataset Description:

- **Movies Dataset:**
    - **Dimensions:** 1078 rows and *12* columns
    - **Key Variables:** 'names', 'date_x', 'score', 'genre', 'overview', 'crew', 'orig_title', 'status', 'orig_lang', 'budget_x', 'revenue', 'country'
    - **Data Types:** Strings, Floats, Date
    - **Missing Values:**
        - **'Genre':** 85 missing values were filled with `"Unknown"`.
        - **'Crew':** 56 missing entries were replaced with `"Not Available"`.

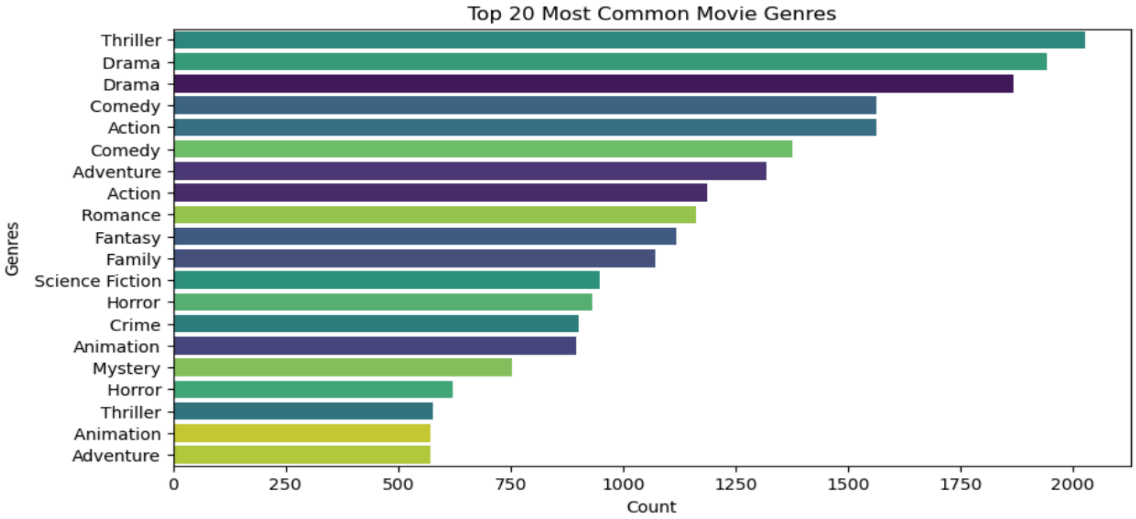
- **Music Dataset:**
    - **Dimensions:** 114000 rows and *21* columns
    - **Key Variables:** 'artists', 'track_name', 'popularity', 'track_genre'
    - **Data Types:** Strings, Integers
    - **Missing Values:**
        - Dropped a single row with missing data to retain clean structure.

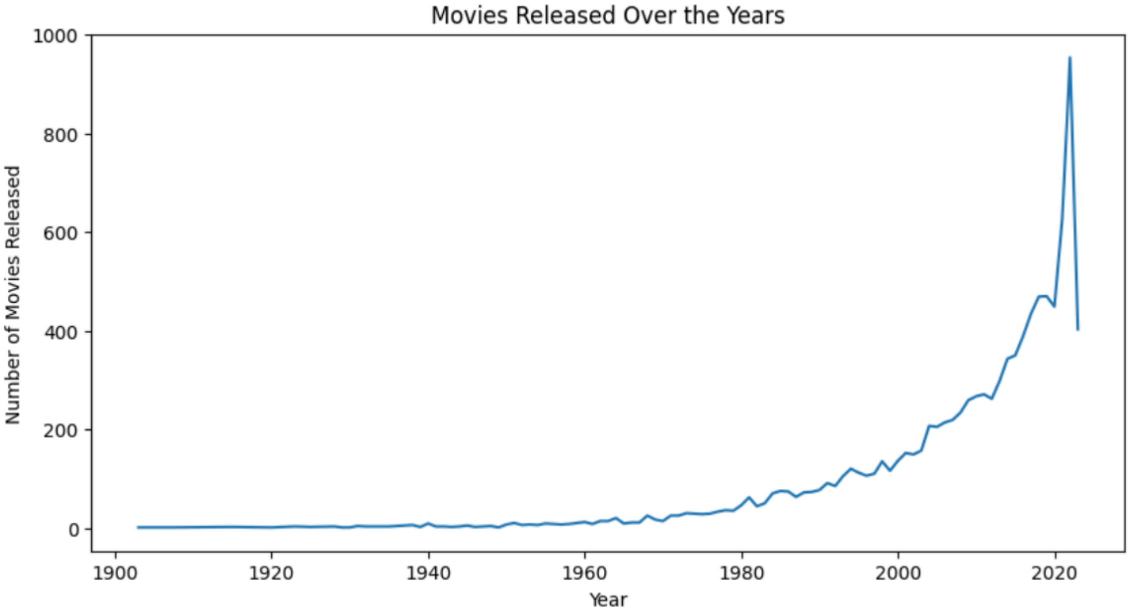
- **Books Dataset:**
    - **Dimensions:** 10000 rows and *8* columns
    - **Key Variables:** 'Book Title', 'Author', 'Genres', 'Description', 'Avg Ratings'
    - **Data Types:** Strings (titles, authors, genres), Floats (ratings)
    - **Missing Values:**
        - **'Description':** Missing values were filled with the placeholder text `"No Description Available"`.

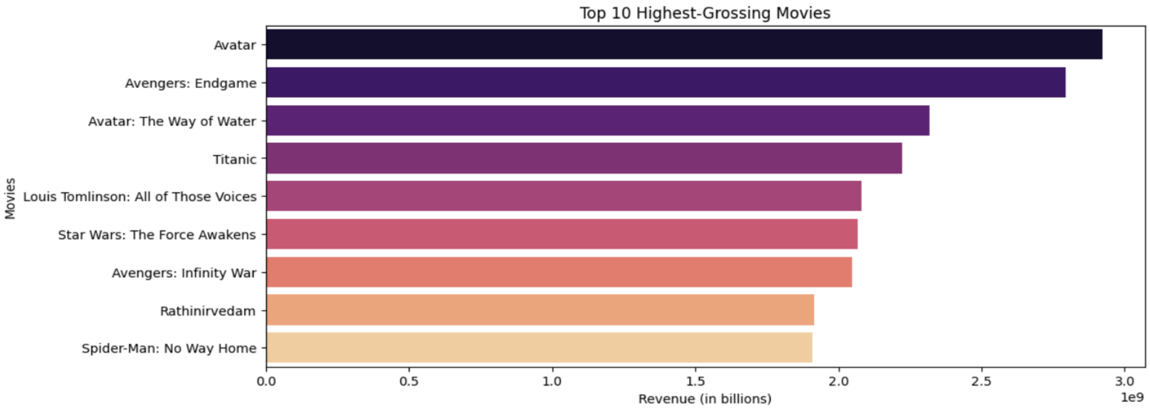
Exploratory Data Analysis & Visualizations:

Movies:

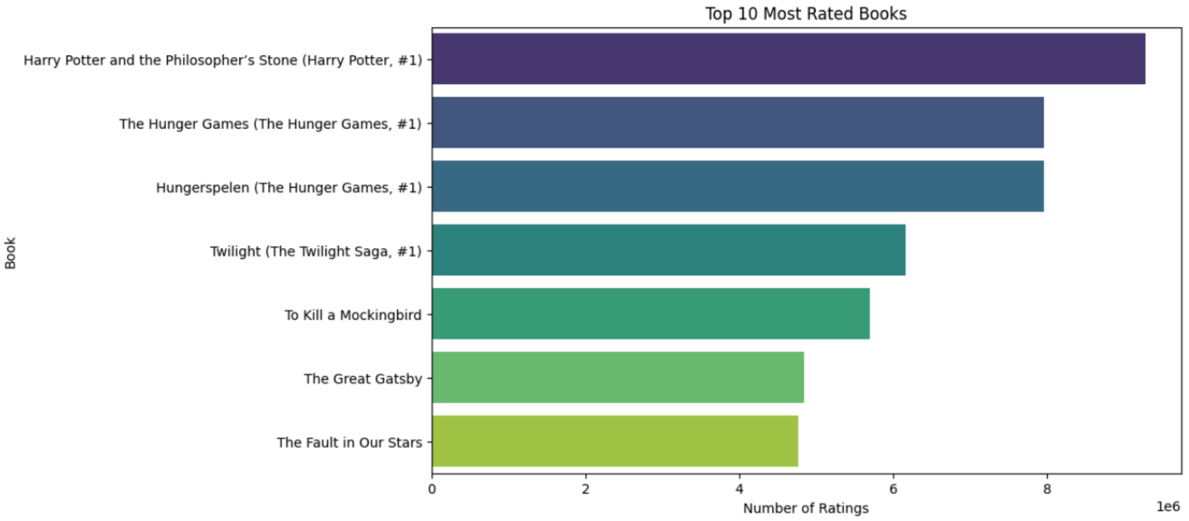
Top 20 Most common Movie Genres

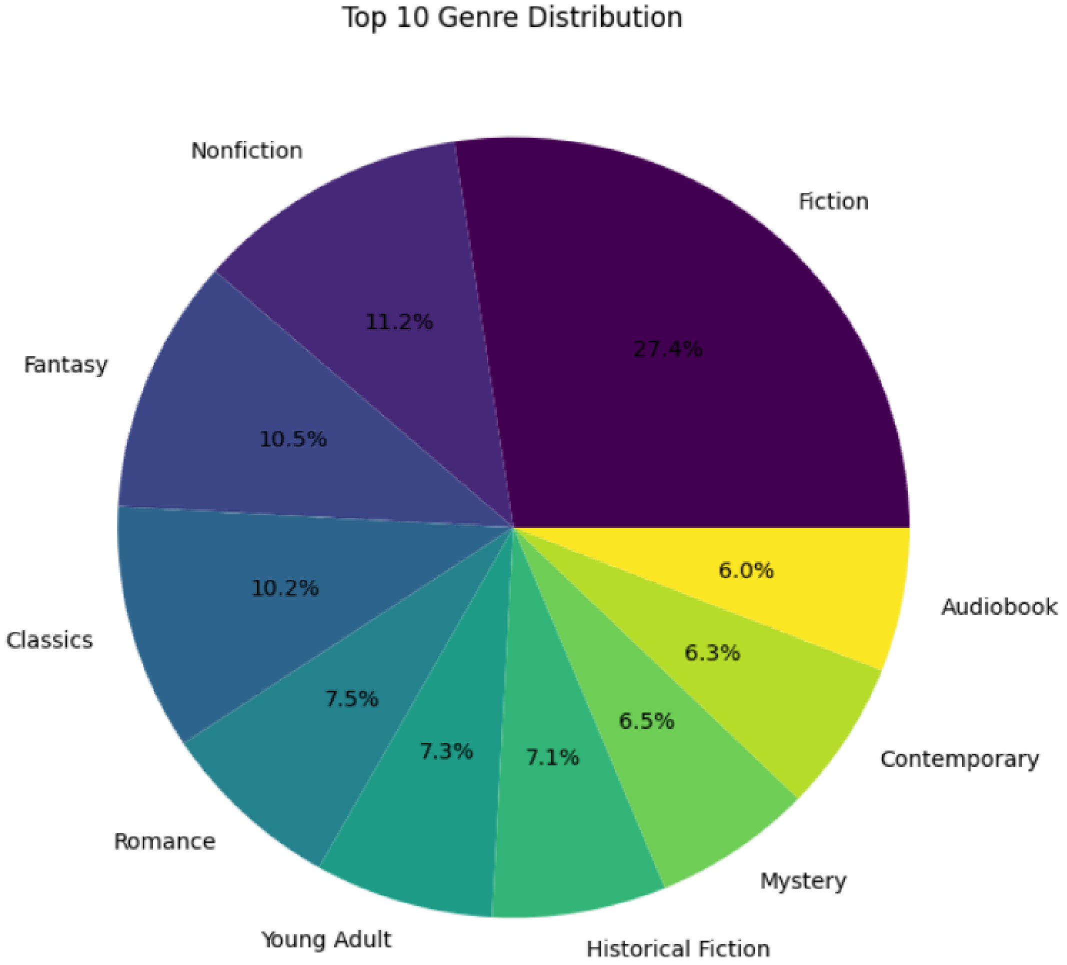


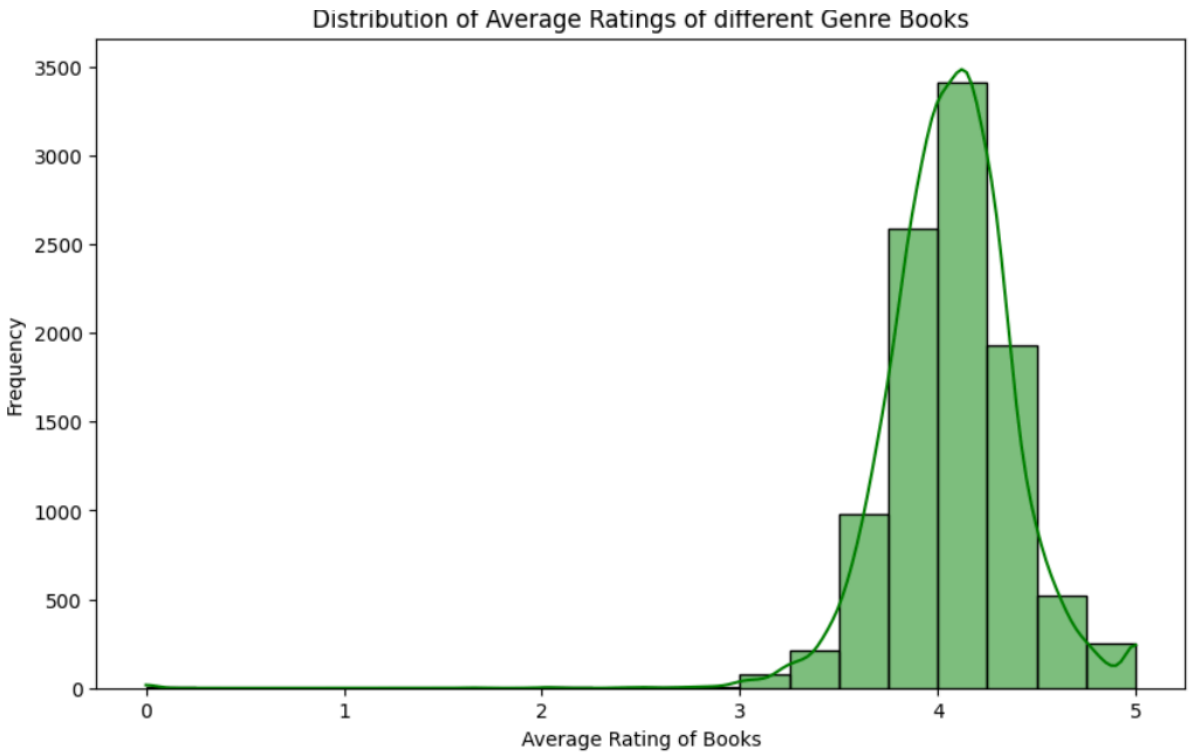




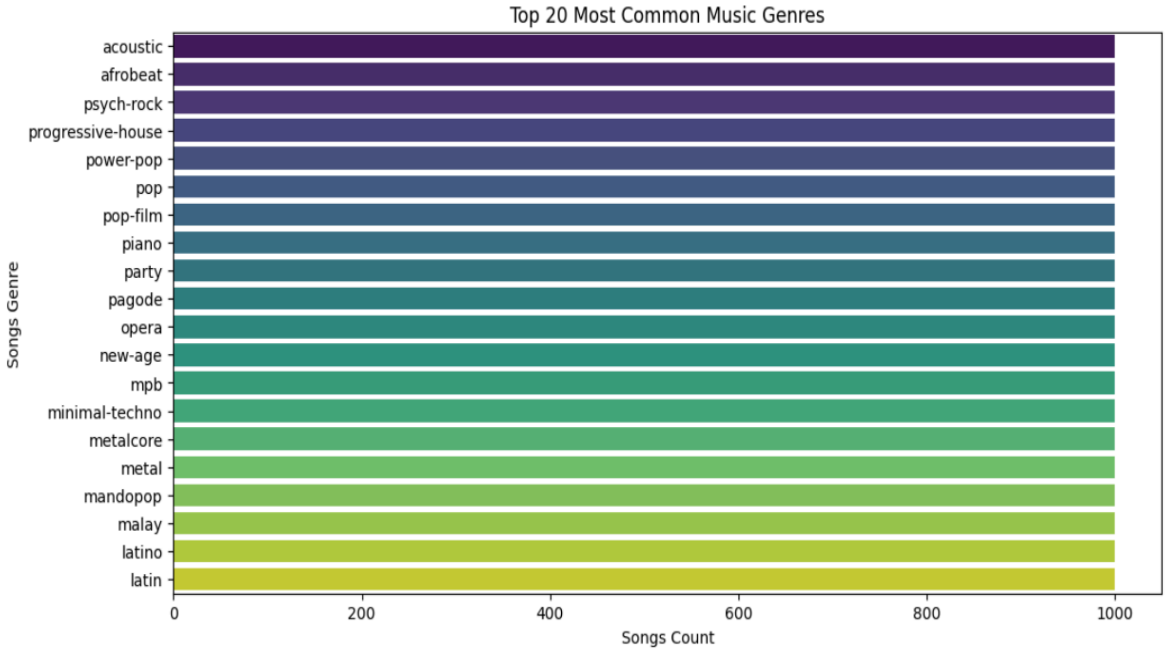
Books:

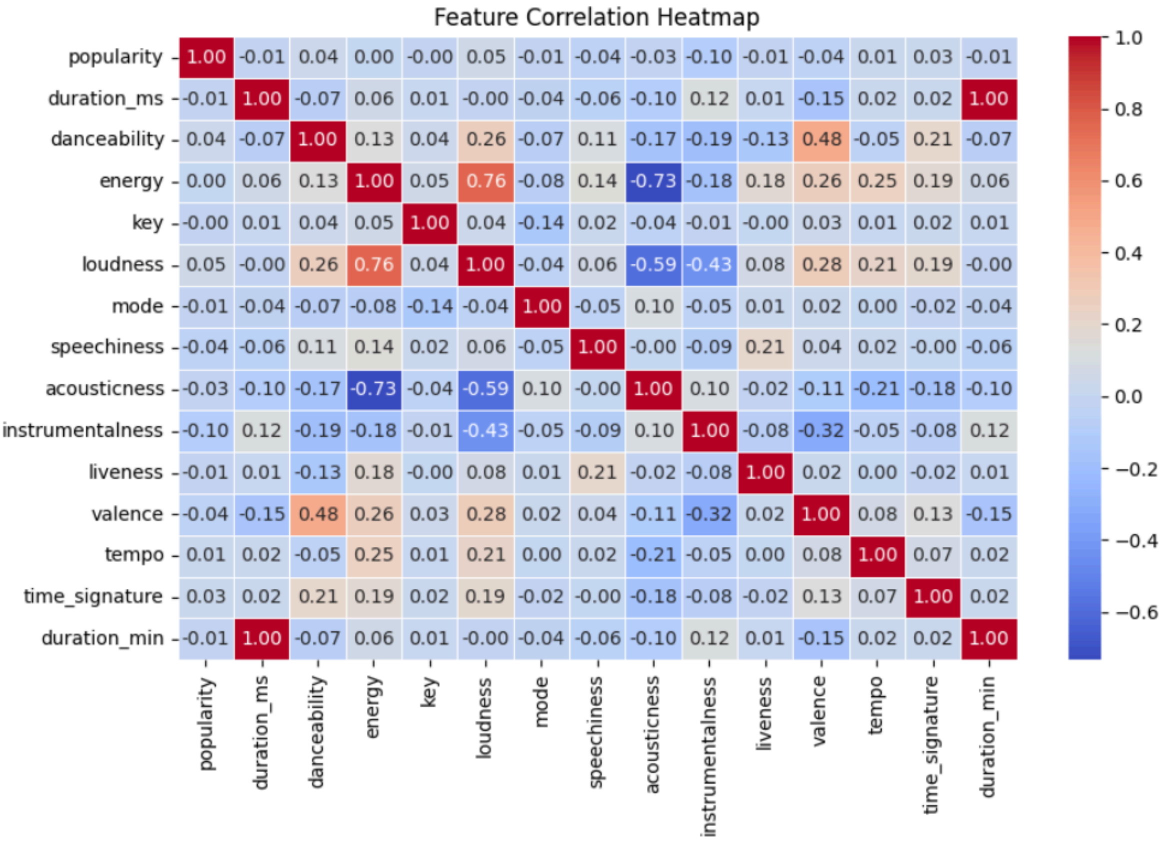

Top 10 Most Rated Books


Top 10 Genre Distribution


Distribution of Average Ratings of different Genre Books
Frequency
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
Average Rating of Books

Music:

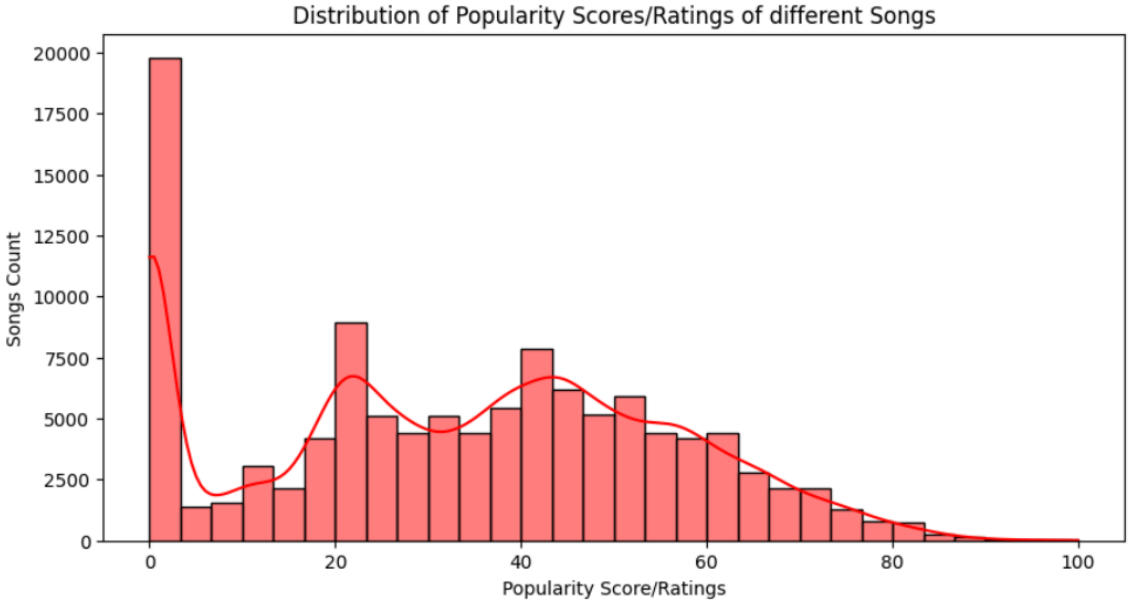

Top 20 Most Common Music Genres
Songs Genre
acoustic
afrobeat
psych-rock
progressive-house
power-pop
pop
pop-film
piano
party
pagode
opera
new-age
mpb
minimal-techno
metalcore
metal
mandopop
malay
latino
latin
0
200
400
600
800
1000
Songs Count



Feature Correlation Heatmap

	popularity	duration_ms	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	duration_min
popularity	1.00	-0.01	0.04	0.00	-0.00	0.05	-0.01	-0.04	-0.03	-0.10	-0.01	-0.04	0.01	0.03	-0.01
duration_ms	-0.01	1.00	-0.07	0.06	0.01	-0.00	-0.04	-0.06	-0.10	0.12	0.01	-0.15	0.02	0.02	1.00
danceability	0.04	-0.07	1.00	0.13	0.04	0.26	-0.07	0.11	-0.17	-0.19	-0.13	0.48	-0.05	0.21	-0.07
energy	0.00	0.06	0.13	1.00	0.05	0.76	-0.08	0.14	-0.73	-0.18	0.18	0.26	0.25	0.19	0.06
key	-0.00	0.01	0.04	0.05	1.00	0.04	-0.14	0.02	-0.04	-0.01	-0.00	0.03	0.01	0.02	0.01
loudness	0.05	-0.00	0.26	0.76	0.04	1.00	-0.04	0.06	-0.59	-0.43	0.08	0.28	0.21	0.19	-0.00
mode	-0.01	-0.04	-0.07	-0.08	-0.14	-0.04	1.00	-0.05	0.10	-0.05	0.01	0.02	0.00	-0.02	-0.04
speechiness	-0.04	-0.06	0.11	0.14	0.02	0.06	-0.05	1.00	-0.00	-0.09	0.21	0.04	0.02	-0.00	-0.06
acousticness	-0.03	-0.10	-0.17	-0.73	-0.04	-0.59	0.10	-0.00	1.00	0.10	-0.02	-0.11	-0.21	-0.18	-0.10
instrumentalness	-0.10	0.12	-0.19	-0.18	-0.01	-0.43	-0.05	-0.09	0.10	1.00	-0.08	-0.32	-0.05	-0.08	0.12
liveness	-0.01	0.01	-0.13	0.18	-0.00	0.08	0.01	0.21	-0.02	-0.08	1.00	0.02	0.00	-0.02	0.01
valence	-0.04	-0.15	0.48	0.26	0.03	0.28	0.02	0.04	-0.11	-0.32	0.02	1.00	0.08	0.13	-0.15
tempo	0.01	0.02	-0.05	0.25	0.01	0.21	0.00	0.02	-0.21	-0.05	0.00	0.08	1.00	0.07	0.02
time_signature	0.03	0.02	0.21	0.19	0.02	0.19	-0.02	-0.00	-0.18	-0.08	-0.02	0.13	0.07	1.00	0.02
duration_min	-0.01	1.00	-0.07	0.06	0.01	-0.00	-0.04	-0.06	-0.10	0.12	0.01	-0.15	0.02	0.02	1.00

```
plt.show()
```



Distribution of Popularity Scores/Ratings of different Songs

MileStone 2

2. Feature Engineering:

2.1 Dataset Overview

Three datasets were sourced using the HuggingFace `datasets` library:

Dataset Source	Domain	Original Columns Extracted
IMDB Movie Dataset	Movies	`names`, `orig_title`, `genre`, `overview`, `score`, `crew`, `date_x`
Spotify Tracks	Music	`track_id`, `track_name`, `track_genre`, `popularity`, `artists`
Goodreads Books	Books	`Book`, `Genres`, `Description`, `Avg_Rating`, `Author`

2.2 Feature Unification

To bring all datasets into a shared format, columns were standardized:

Unified Column	Description	Source Columns
`Item_ID`	Unique ID prefixed with content type	`names`, `track_id`, `Book`
`Title`	Title of the item	`orig_title`, `track_name`, `Book`
`Genre`	Genre(s) of the item	`genre`, `track_genre`, `Genres`
`Description`	Brief summary of content	`overview`, `Description`

`Popularity`	Rating/score/popularity	`score`, `popularity`, `Avg_Rating`
`Creator`	Actors, Artists, or Authors	`crew`, `artists`, `Author`
`Timestamp`	Time information for session simulation	`date_x`, generated
`Item_Type`	Type of content	"Movie", "Music", "Book"

1.3 Feature Construction

- Item_ID
  - Constructed by prefixing item type to original identifiers (e.g., `Movie_`, `Music_`, `Book_`).
  - Ensures global uniqueness across datasets.
- Timestamp Handling
  - Missing timestamps (especially for books/music) were filled using static values to enable session ordering.
- Session Simulation
  - Simulated user interactions using:
    - Random session IDs (`Session_ID`) assigned across all entries.
    - Random `Action_Type`: `Clicked`, `Searched`, `Scrolled`.
- Output: `session_events.csv`
  - Final dataset with 10 columns and thousands of rows.
  - Unified, clean, and ready for sequential modeling.

2.2 Feature Importance via Model Structure

Instead of manual selection:

- GRU model learns feature representations via embedding layers.
- Item_ID is the key feature encoded numerically and learned over time.

3. Model Selection

3.1 Selected Model: GRU4Rec (Gated Recurrent Unit for Recommendation)

Rationale for Selection

The GRU4Rec model, based on Gated Recurrent Units (GRUs), was selected due to its effectiveness in session-based and sequential recommendation tasks. The key motivations are:

- **Sequential Pattern Learning**: GRUs capture dependencies in time-ordered data, making them well-suited for modeling user interaction sequences.
- **Session-based Suitability**: Unlike user-centric models, GRUs can function effectively in session-only environments.
- **Computational Efficiency**: GRUs are faster and more efficient than LSTMs, while maintaining comparable performance.
- **State-of-the-Art Performance**: Widely recognized in literature as one of the foundational deep learning models for next-item recommendation tasks.

3.2 Model Architecture

The architecture of GRU4Rec is summarized below:

Layer	Description
Embedding Layer	Converts item indices into dense 128-dimensional vectors.
GRU Layer	A GRU with 256 hidden units to capture sequential dependencies.
Fully Connected	A linear layer that maps GRU outputs to vocabulary size for next-item prediction.

Model Parameters:

- Embedding dimension: 128
- Hidden dimension: 256
- Sequence length: 50 (padded)
- Vocabulary size: Derived from number of unique Item_IDs

4. Training Strategy

4.1 Data Loader Configuration

- Custom PyTorch `Dataset` class for session pairs (input-target).
- Batched using `DataLoader` with custom padding-based `collate_fn`.

4.2 Loss Function and Optimizer

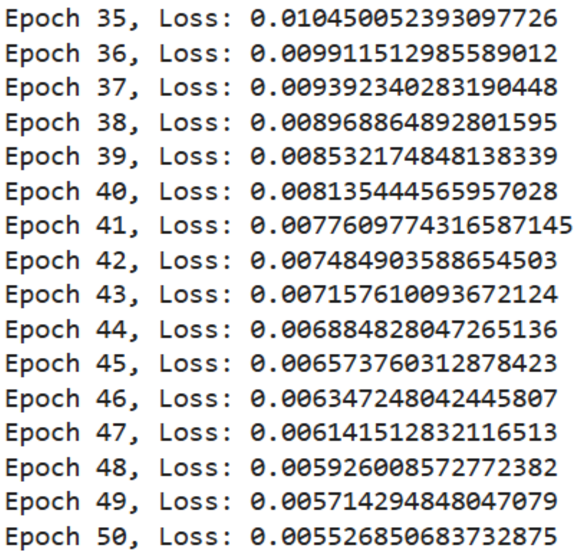
- **Loss Function**: Categorical CrossEntropyLoss
- **Optimizer**: Adam (Learning Rate = 0.001)
- **Batch Size**: 64
- **Epochs**: Up to 50, with early stopping

4.3 Early Stopping

- Implemented to avoid overfitting.
- Patience = 5 epochs (stop training if no improvement in validation loss for 5 consecutive epochs).
- Best model saved as `best_session_rec_model.pth`.

4.4. Training Metrics and Logs

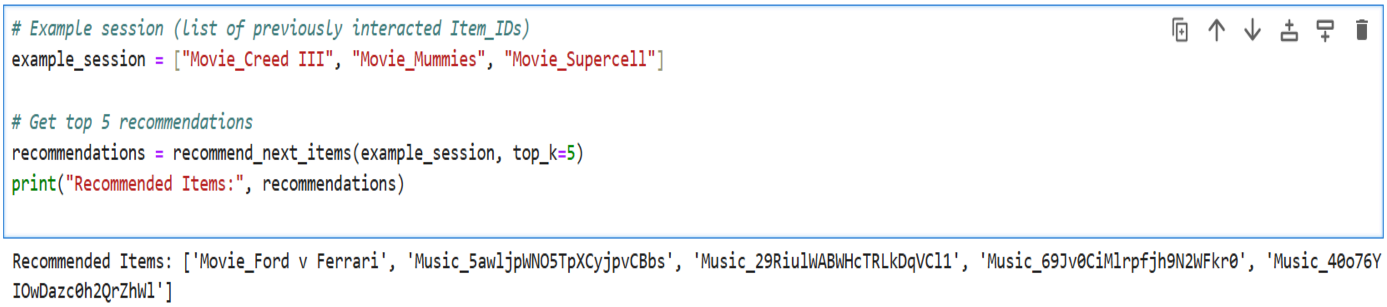
- **Best Loss Achieved**: ~0.0050 (Epoch 47)
- **Total Parameters**: Trained with GPU acceleration; model supported with `DataParallel` for multi-GPU environments.



```
Epoch 35, Loss: 0.010450052393097726
Epoch 36, Loss: 0.009911512985589012
Epoch 37, Loss: 0.009392340283190448
Epoch 38, Loss: 0.008968864892801595
Epoch 39, Loss: 0.008532174848138339
Epoch 40, Loss: 0.008135444565957028
Epoch 41, Loss: 0.0077609774316587145
Epoch 42, Loss: 0.007484903588654503
Epoch 43, Loss: 0.007157610093672124
Epoch 44, Loss: 0.006884828047265136
Epoch 45, Loss: 0.006573760312878423
Epoch 46, Loss: 0.006347248042445807
Epoch 47, Loss: 0.006141512832116513
Epoch 48, Loss: 0.005926008572772382
Epoch 49, Loss: 0.005714294848047079
Epoch 50, Loss: 0.005526850683732875
```

4.5 Testing

The model successfully generalized across different content domains (movies and music), recommending relevant next items based on prior session history.



```
# Example session (list of previously interacted Item_IDs)
example_session = ["Movie_Creed III", "Movie_Mummies", "Movie_Supercell"]

# Get top 5 recommendations
recommendations = recommend_next_items(example_session, top_k=5)
print("Recommended Items:", recommendations)
```

```
Recommended Items: ['Movie_Ford v Ferrari', 'Music_5awljpWNO5TpXCyjpvCBbs', 'Music_29RiulWABWHcTRLkDqVCl1', 'Music_69Jv0CiMlrpfjh9N2WFkr0', 'Music_40o76Y
IOwDazc0h2QrZhWl']
```