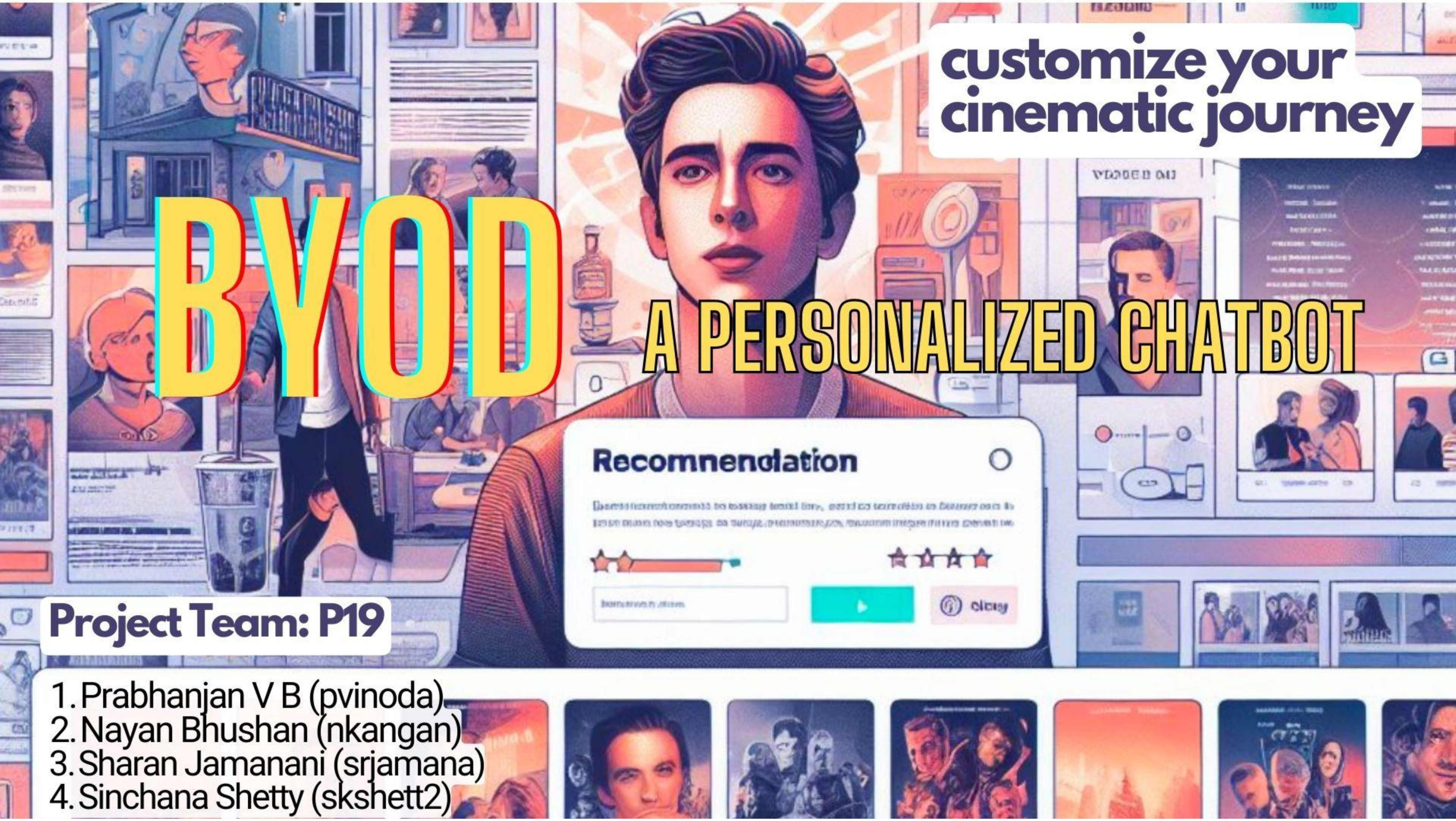


Project Link : https://github.ncsu.edu/pvinoda/engr-ALDA-Fall2023-P19

Problem Statement

- In the dynamic world of streaming services, the user experience is paramount.
- Personalized movie recommendations play a pivotal role in shaping user satisfaction and engagement.
- **The Netflix Prize dataset**, with its extensive collection of user ratings and diverse movie choices, adds a layer of complexity to the recommendation task.

The dataset is comprised of fields such as:

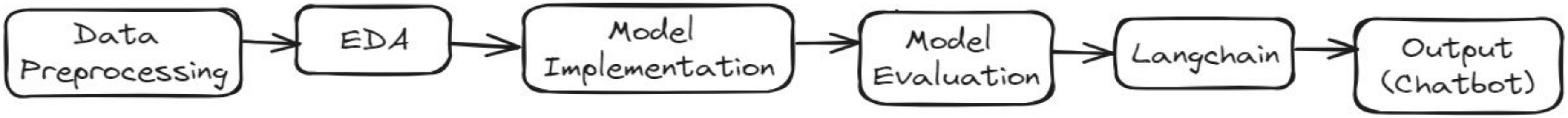
- **Movie Id**
- **Customer Id**
- **Ratings**
- **Date**

Background :

Breese.,(1998) introduced the crucial concept that users with similar rating histories exhibit similar preferences in future interactions. Collaborative Filtering (CF) techniques leverage this principle, utilizing user similarities to provide personalized movie recommendations.

Vito Xituo Chen.,(2016) applies Single Value Decomposition to break-down user-item interactions, predicting missing ratings and tailoring recommendations from these hidden factors.

Workflow :



Data Preprocessing

Data Collection: Acquired a range of netflix dataset from kaggle categorized information as customer_id, ratings, dates.

Data Splitting: The dataset has been divided into 5 parts. Due to resource constraints the results are evaluated for first part.

Data Cleaning: Handled missing values, discrepancies, and duplicates while also standardizing movie titles for consistency across all datasets.

Exploratory Data Analysis - 1

Dataset Overview: Utilized summary statistics to comprehend the dataset's structure and dimensions.

Data Aggregation and Filtering: Aggregated data for movies and customers, filtered out less significant entries, and controlled dataset significance through thresholds.

Pivot Table and Visualization: Organized data via pivot tables for analysis and visually represented rating distributions.

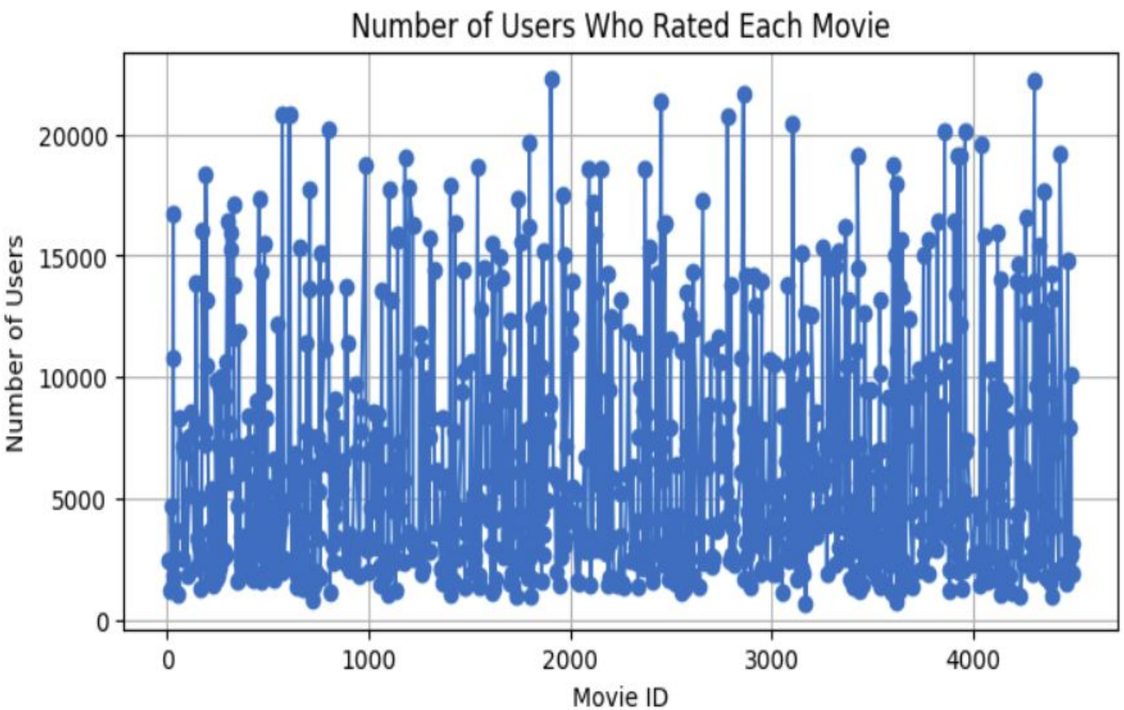
Exploratory Data Analysis - 2

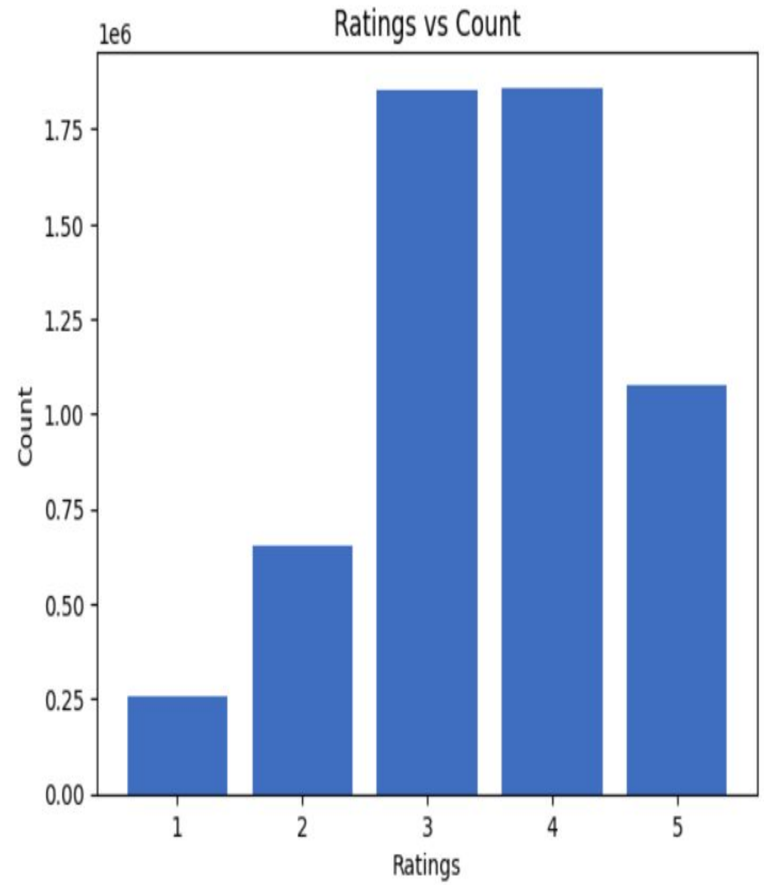
Visualization of Users per Movie: Explored user engagement patterns by visualizing the number of users who rated each movie.

Statistical Analysis: Assessed distribution normality using Q-Q plots and gained insights into rating spread through box plots.

Rating Statistics: Calculated mean ratings and explored patterns in ratings for each movie to derive insights.

EDA - Data Visualizations





Models Implementation

Inputs : Movie id, Customer id, ratings, date

Output : Estimate Score, Movie title

Algorithms Implemented : KNN, SVD

Cross-Validation : 10 fold

Evaluation Matrix : RMSE, MAE

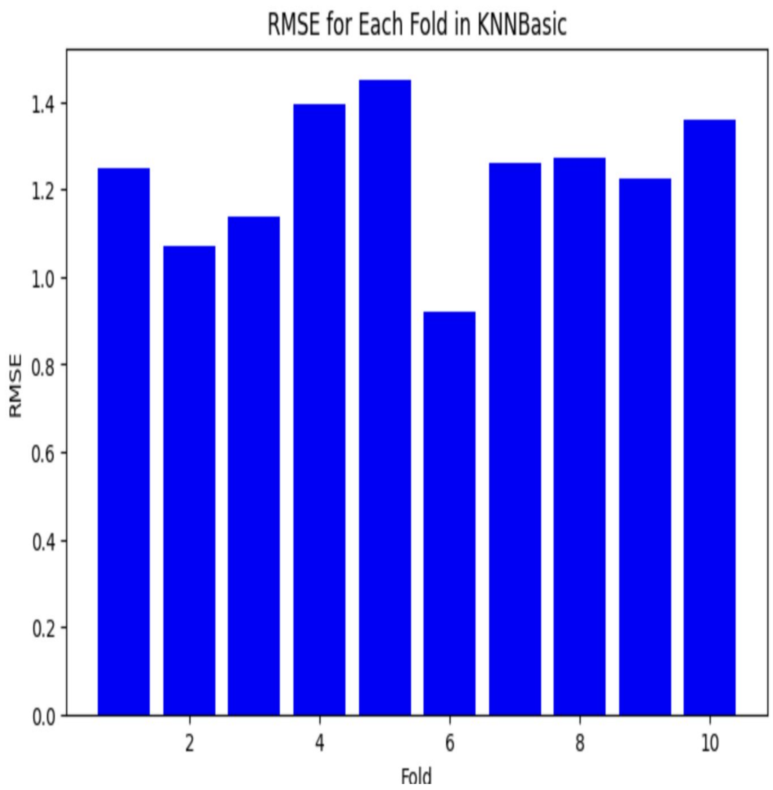
Model 1: K Nearest Neighbors (KNN)

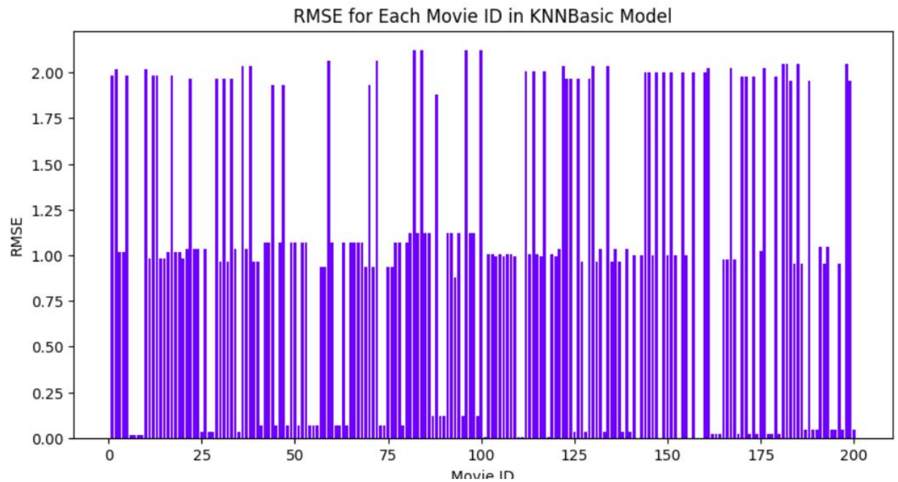
- A collaborative filtering technique based on similarity metrics.

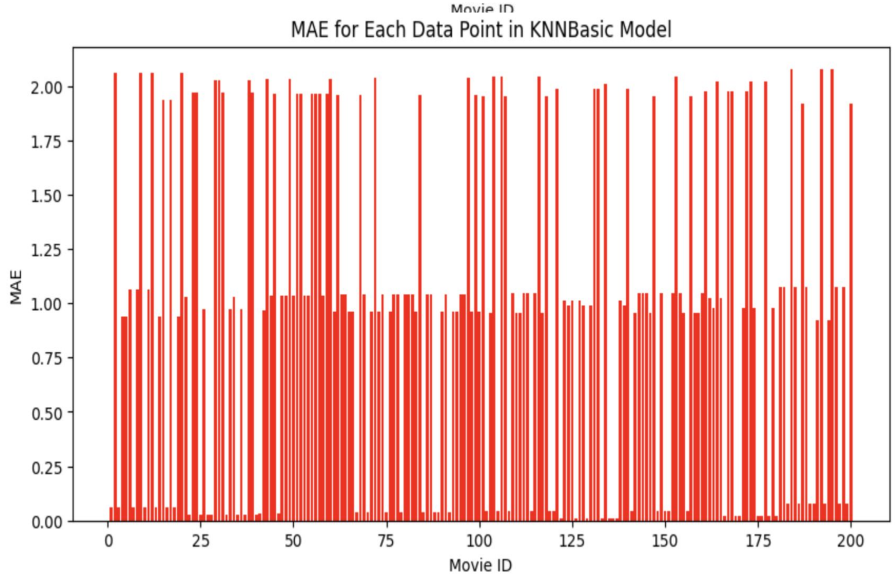
- Finds k similar users by assessing movie ratings to establish a prediction mechanism that recommends movies based on the liked movies of their closest neighbors.

- KNN is chosen for movie recommendations due to its non-parametric, content-based approach, using similarity measures on movie features for intuitive suggestions

Graphical Representation of Evaluation Matrix







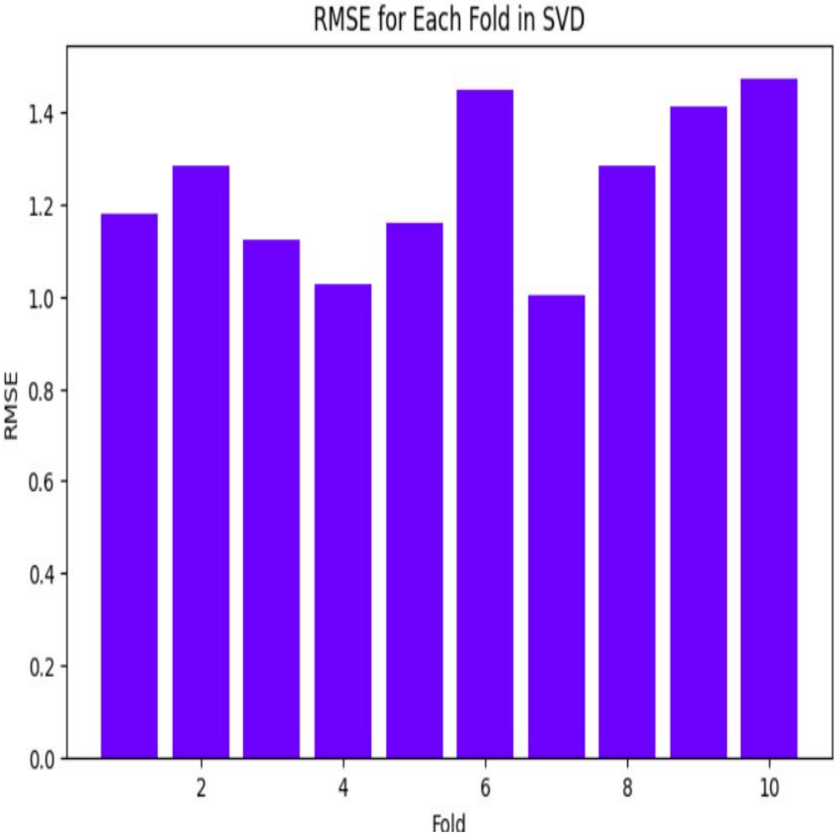
Model 2 : Singular Value Decomposition (SVD)

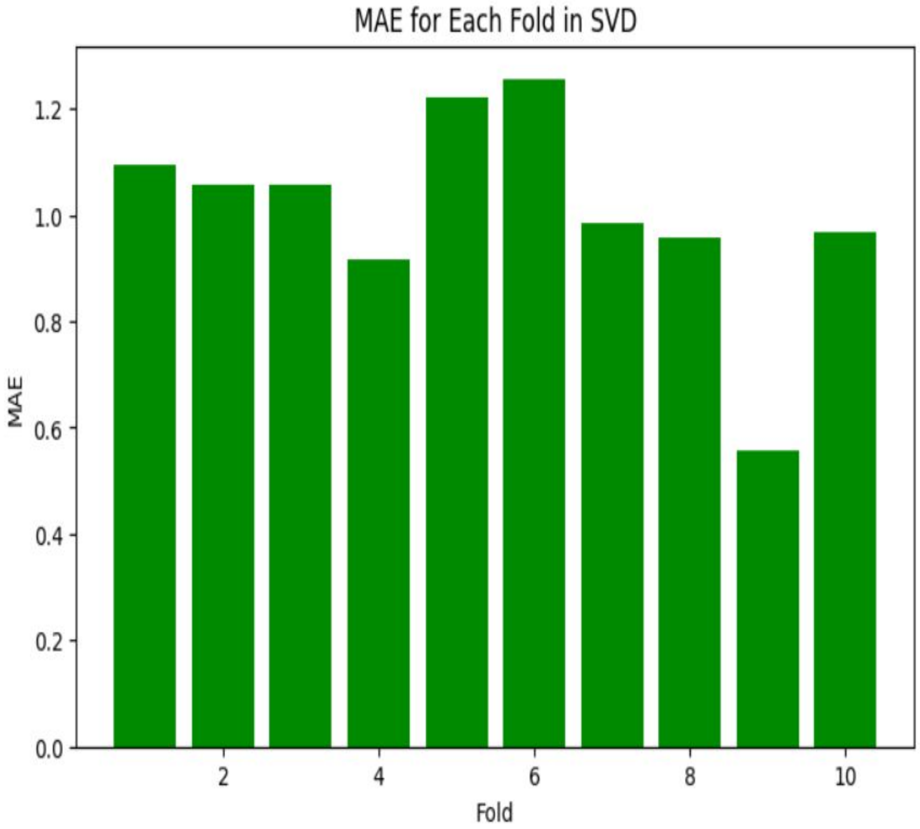
- SVD is used in collaborative filtering-based recommendation systems to handle the sparsity of user-item rating matrices and provide meaningful recommendations.

- It helps in grasping relationships between users and items in a more abstract and interpretable way

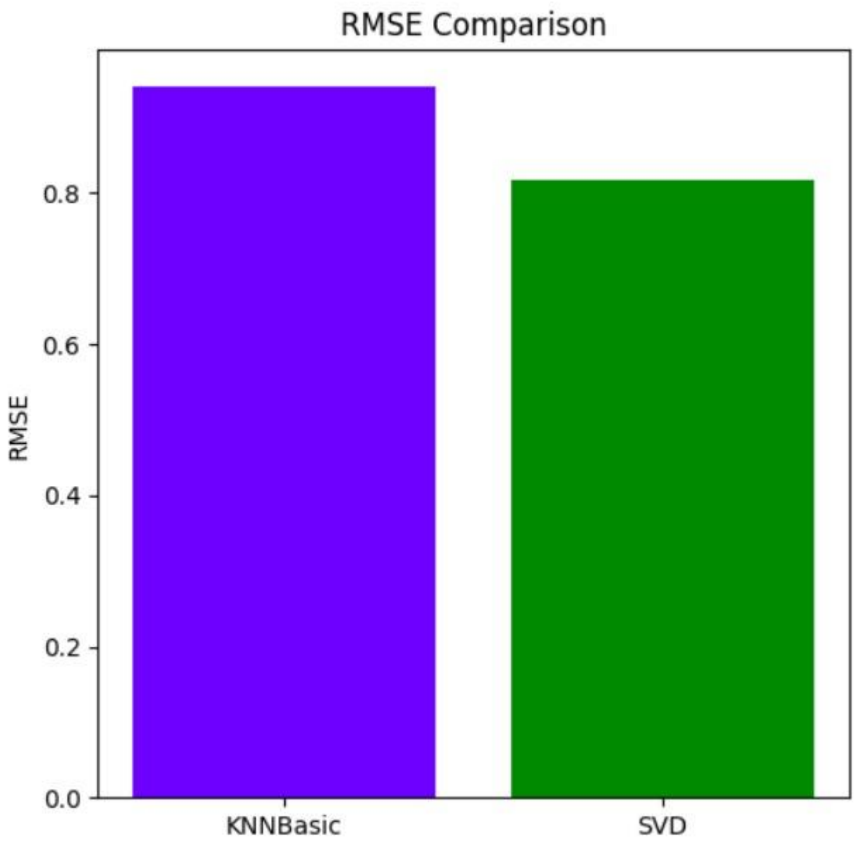
- This model is effective in capturing underlying patterns in user-item interactions.

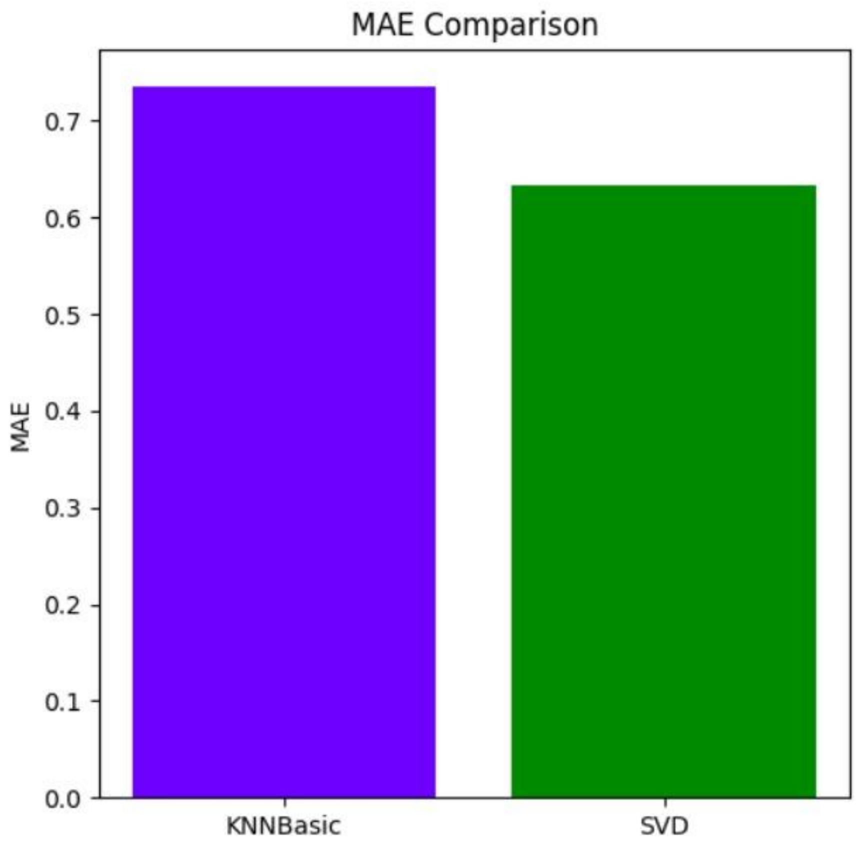
Graphical Representation of Evaluation Matrix





Best Model?

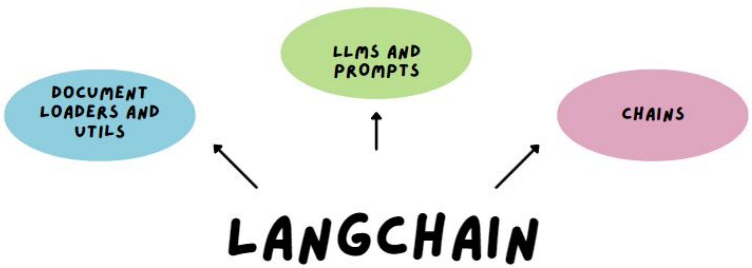




Best model : SVD

Langchain: What is it and Why is it relevant ?

LangChain is a framework for developing applications powered by large language models (LLMs).



- It enables the creation of context-aware applications that can leverage the power of LLMs to solve complex problems.
- We can leverage LangChain's built-in **fine-tuning** tools to efficiently train your LLM model on your custom data.

Using Transformers in generating Model Response

- Transformers are not mere disguised robots! In NLP, they're potent models, excelling in understanding word relationships via attention mechanisms.

- We are streamlining interactions with SentencePiece and AutoTokenizer. Text-To-Text tasks are being used for seamless response generation.

- In conclusion, we've scratched the surface of transformers in natural language generation. This technology is continually evolving, and exploring its potential can lead to various NLP applications.

Chatbot depicting an NLP Response based on the model output

Out[114]:

Chatbot: Welcome To BYOD. Please enter the customer id for processing movie recommendations

User:

Enter the customer id here…

Chat!

```
\LLM Output:
You are likely to enjoy this movie. Lost: Season 1 The Simpsons:
Season 6 Lord of the Rings: The Fellowship of the Ring The
Godfather The Sixth Sense The Simpsons: Season 3 Family Guy:
Freakin' Sweet Collection CSI: Season 1 The West Wing: Season 1
Family Guy: Freakin' Sweet Collection CSI: Season 1 CSI: Season 2
Family Guy: Freakin' Sweet Collection CSI: Season 1 CSI: Season 2
 Family Guy: Freakin' Sweet Collection CSI: Season 1
```

Conclusion, Takeaways and future scope

Conclusion and Takeaways :

- From this project, we have learnt to practically implement the data mining concepts learnt in class.
- Thorough preprocessing, including handling missing values and data type conversions, laid a strong foundation for subsequent analysis and modeling.
- The project's holistic approach, from preprocessing and EDA to model comparison, underscores the importance of a well-rounded methodology.

Future scope:

- Implement real-time adaptability to cater to evolving user behaviors and preferences.
- Expand natural language processing techniques using fine-tuned language models from Huggingface.

If you have any questions feel free to ask or mail:

nkangan@ncsu.edu

srjamana@ncsu.edu

pvinoda@ncsu.edu

skshett2@ncsu.edu