Enhancing Interpretability in Diverse Recommendation Systems through Explainable AI Techniques

Meera Narvekar[1], Krish Bharucha[2] , Varun Vishwanath[3], Neel Gabani[4], Shaun Fernandes[5]

[1,2,3,4,5]SVKM's Dwarkadas J Sanghvi College of Engineering, Mumbai, India
Email: meera.narvekar@djsce.ac.in[1], krishbharucha03@gmail.com[2] , varunvis2903@gmail.com[3], neelgabani24ng@gmail.com[4], shaunferns94@gmail.com[5]

Received: 18.01.2024 Revised: 11.02.2024 Accepted: 24.02.2024

ABSTRACT
This paper explores the application of XAI methodologies, particularly focusing on the utilization of the Shapley Additive explanations (SHAP) framework, and implement it into three distinct recommendation systems with explainability: matrix factorization, content-based filtering, and collaborative filtering. Using a novel blend of SHAP values and a multimodal Large Language Model (LLM), namely GPT-4, we highlight a unique methodology utilized for understanding the decision-making processes underlying recommendation algorithms. The exploration of SHAP values reveals granular insights into the factors which influence individual recommendations, embiggening users understanding of the suggestions provided by these algorithms. Leveraging a multimodal LLM further augments interpretability by providing a detailed yet succint explanation of SHAP-derived insights. By laying bare the inner working of the chosen recommendation models, our research seeks to foster transparency and increased user control in the domain of recommendation systems.

Keywords: Recommendation Systems, Matrix factorization, Content-based filtering, Collaborative filtering, xAI, SHAP, GPT-4

1. INTRODUCTION

In this era, recommendation systems play a crucial part in helping users navigate through the enormous collection of online content and services. From personalized movie suggestions on streaming platforms to customer specific product recommendations on e-commerce websites, these recommendation systems are now an essential part of our digital experience, consciously and subconsciously shaping our choices and preferences [1].

As recommendation systems continue to expand across various domains, the importance of trust and transparency intheir operation cannot be overstated. Every day, millions of users rely increasingly on algorithmic suggestions to discovernew content, make purchasing decisions, and explore complex information spaces. However, the fact that recommendation algorithms are essentially black boxes raises concerns abouttheir fairness, accountability, and potential biases. In an age where AI-driven decisions influence various aspects of ourlives, establishing trust in recommendation systems becomes paramount.

This paper thus introduces explainable AI (XAI), a recently expanding field aimed at decoding the "black box" of machine learning models and providing insights into its decision-making processes [2]. XAI when applied to recommendation techniques, offers a means to shed light on why specific itemsor content are recommended to users, which helps solidify user understanding. By elucidating the underlying factors which determine recommendations, XAI not only empowers users to make knowledgeable choices but also allows developers to address potential issues in the algorithm such as biases or improper feature selection.

This research paper explores diverse recommendation systems spanning collaborative filtering, content-based filtering, and hybrid approaches. We not only look at the implementation and evaluation of these systems but also on integration of explainable AI techniques to increase their interpretability. By listing the decision-making rationale behind recommendation algorithms, our target is to bridge the gap between users and algorithms, and facilitate more knowledgeable interactions in recommendation-driven environments. By following empirical analysis and case studies, this paper explores the efficiency of XAI methods in enhancing the interpretability of various ecommendation paradigms and discuss implications for trust and transparency in AI-driven systems.

2. LITERATURE SURVEY

The ability of recommendation systems to be interpreted or understood is highly significant in various domains, including health, financial, and criminal justice. The development of Explainable AI (XAI) techniques is opening up a pathway toward enhancing the interpretability of recommendation systems across several applications. This survey provides an overview of recent research, synthesizing results to better understand the state of XAI techniques in recommendation systems and identifies possible future research directions.

Linardatos et al. (2020) [3] proposed a comprehensive literature review and taxonomy of approaches for interpretability in machine learning. Their research acts as a reference for theoreticians and practitioners to gain exposure to various XAI methods and their programming implementations. It provides a foundational understanding of interpretability methods in machine learning and sets the stage for further in-depth exploration of diverse XAI techniques in recommendation systems.

Sewada, Ranu, Jangid, Ashwani, Kumar, Piyush, and Mishra, Neha (2023) [4] explored the relevance of XAI across various domains, including healthcare, finance, and criminal justice. The authors considered various XAI methodologies and techniques, such as LIME and SHAP, evaluating their interpretability against computational efficiency and accuracy. Their study highlights the potential of XAI techniques to enhance interpretability in diverse recommendation systems and emphasizes the need for further exploration of these methods across different domains. Furthermore, Sewada et al. presented a survey and framework that categorizes XAI design goals and their corresponding evaluation methods, providing insights into how different XAI user groups' design goals can be mapped to evaluation methods. This framework serves as a valuable resource for understanding the design and evaluation of XAI techniques, laying the groundwork for future research in this area.

In the context of sentiment analysis, Xu, Feiyu, Uszkoreit, H., Du, Yangzhou, Fan, Wei, Zhao, Dongyan, and Zhu, Jun (2019) [5] proposed a commonsense-based neurosymbolic framework that overcomes limitations and provides fully interpretable and explainable results. This approach offers new perspectives on enhancing interpretability in recommendation systems by leveraging neurosymbolic techniques, opening avenues for future research in developing neurosymbolic XAI techniques for diverse recommendation systems.

The work by Buccinca, Zana, Lin, Phoebe, Gajos, Krzysztof Z., and Glassman, Elena L. (2020) [6] proposed an efficient and effective Tuning framework for Aligning LLMs with Recommendations, known as the TALLRec framework, which significantly enhances the recommendation capabilities of LLMs in the movie and book domains. This study demonstrates the potential for enhancing recommendation systems through specialized XAI techniques, paving the way for further research in this domain.

While existing research provides valuable insights into XAI techniques for enhancing interpretability in diverse recommendation systems, there are still knowledge gaps that warrant further investigation. For instance, further research should explore the applicability of XAI techniques in specific domains such as healthcare and finance, and develop specialized XAI methods tailored to these applications. Additionally, the ethical implications of XAI in diverse recommendation systems remain an important area for future research, particularly concerning

transparency and accountability. Moreover, the development of standardized evaluation metrics for XAI techniques across different domains and applications is essential to ensure their effectiveness and reliability.

In conclusion, this literature review on the application of explainable artificial intelligence in enhancing interpretability across different types of recommendation systems provides a comprehensive understanding of the current research landscape. The synthesis of findings from previous studies indicates that XAI techniques hold great promise for improving interpretability in recommendation systems across various domains. However, future work must address existing gaps and explore the ethical implications of XAI in diverse recommendation systems to develop more effective and reliable XAI techniques.

3. METHODOLOGY

A. Dataset

Some of the key steps in the methodology that have beenfollowed in this research start with choosing matrix factorization, content-based filtering, and collaborative filtering asthree different recommendation systems and extend to theirwidespread use. These techniques are representative of different recommendation approaches and provide a broad basis onwhich to apply XAI techniques.

Then, the SHapley Additive exPlanations framework is integrated into each of these systems. The respective SHAP valuesare computed as a way of quantifying the contribution everyfeature has made to recommendations made by algorithms.This will be one very important step towards showing

whatgranular factors influence every recommendation with thepurpose of providing model decision-making transparency.
Improving Interpretability of SHAP-Derived Insights: Forimproving the interpretability of SHAP-derived insights, amultimodal Large Language Model is used-GPT-4. The LLMwill generate long, short explanations of the SHAP values thattranslate numeric insights into user-friendly narratives. Closingthe gap between complex model outputs and the end-user'scomprehension, the recommendations are more accessible andeasy to understand.
Besides disclosing the internal mechanisms of recommendation algorithms, this combination of SHAP values and GPT-4is a novel approach toward enhancing the transparency anduser control in recommendation systems. By laying bare thedrivers of recommendations, this work aims at empoweringusers with deeper insights into the suggestions made henceeliciting greater trust and responsibility in AI-driven systems.
The study makes use of the ML-100k dataset, it is a verypopular dataset [7], for recommendation systems research thatis based from the MovieLens database. User-item interactions,including timestamps and ratings, are gathered from a widerange of users in this dataset.
Matrix Factorization: The ML-100k dataset is put in aSurprise trainset object designed for the Matrix Factorizationtechnique, and a Singular Value Decomposition algorithm istrained by the dataset. We will have to use matrix factorizationin extracting latent factors which will be equivalent to user anditem embeddings. These factors are then used for the user-itemratings.
Content-Based Filtering: For Content-Based Filtering, theML-100k dataset is pre-processed and augmented with itemfeatures extracted from the movie information. These features include genre information, such as Action, Adventure,Comedy, Drama, etc. The dataset is converted into a pandasdataframe and merged with movie information, enablingthe extraction of item features. Features of the items arethen processed using TF-IDF representation. This basically
quantifies how important a feature is within the dataset and,in turn, how relevant a feature is to a particular movie.
Collaborative Filtering: The collaborative filtering resultswere conducted using the SVD model trained on the ML-100kdataset. ML-100k is a dataset that contains interaction databetween users and items, including ratings and timestamps.The dataset was pre-processed, and a full training set was constructed for the SVD model to train latent factors representingusers and items. Using this trained model, it is then possibleto predict ratings for all items by one randomly chosen user.At heart, the predictions are the basis for recommendationsmade by collaborative filtering; that is, the system shouldrecommend items that have higher predicted ratings for theuser.
Dataset Preprocessing: Data preprocessing includes datacleaning, filtering, and feature engineering steps performed onthe dataset before passing it to the model for training andevaluation. Data cleaning then refers to the handling of missingvalues by making data consistent and wiping out redundantinformation. Filtering might involve selecting some subset ofusers or items based on some criteria to reduce computationalcomplexity or focus on some specific user segments. Featureengineering is the process of taking raw data and constructing
features that can be fed into recommendation algorithms. Thisdataset further gets divided to become both a training datasetand a test dataset, thus allowing the possibility of evaluationand model performance testing.
Subset Selection: In order to make post-training computation easier and more scalable, the dataset is selected for experimentation in analyzingexplainability. With a representativesample of user-item interactions, ratings, and item features,this subset shall be used for a comprehensive evaluationof recommendation algorithms, while avoiding computationalresource constraints.
The preparation and preprocessing steps of the datasetsshould aim at laying an appropriate foundation with which totrain and evaluate the recommendation algorithms in a robustand scalable manner while maintaining the interpretability ofthe experimental framework.

B. Recommendation Algorithms Studied

1) **Matrix Factorization**: Matrix factorization, in particular Singular Value Decomposition, is a core method for recommendation systems that model user-item interactions to make personalized suggestions [8] [9]. SVD is one of the ways of building recommendation systems via matrix factorization in an attempt to learn latent factors signifying users and items from observed user-item interactions. SVD factorizes a user-item interaction matrix into three matrices: a user matrix, an item matrix, and a diagonal matrix of singular values. These matrices describe the latent factors underlying user preferences and item characteristics. Therefore, it, in approximating an original matrix by lower-dimensional representation, consequently leads to the reduction of the

dimensionality of the user-item space while preserving important patterns or relationships. Singular Value Decomposition is understood through thegiven formula:
$R = U\Sigma V^T$
Where:
R is the user-item interaction matrix,
U is the user matrix containing latent factors representing users,
Σ is the diagonal matrix of singular values representing the importance of latent factors,
V^T is the item matrix containing latent factorsrepresenting items.

2) **Content Based Filtering**: It is a recommendation approach focusing more on theintrinsic characteristics of items (e.g., movies, products)and users' preferences to make recommendations thatare personalized. Unlike collaborative filtering, whichrelies on previous interactions among users and items,content-based filtering utilizes item features or attributesin order to suggest recommendations. [9]In content-based filtering, each item can be described bya set of features or attributes, such as genre, keywords,or metadata. In addition, user profiles are created basedon their preferences—often derived from their interactions with items or explicitly provided preferences.Recommendations then become feature similarity-basedof items to the user profile. This is essentially a recommendation of items that are similar in content towhat the user had liked before. Several methods existfor this computation of closeness or similarity such ascosine similarity and TF-IDF (Term Frequency InverseDocument Frequency).
In practical terms, SVD learns latent factors such as userpreferences for specific features (for example, moviegenres) and item attributes (e.g., movie ratings). Byleveraging these learned factors, recommender systemsbased on SVD can predict user ratings for items thatthey have not interacted with yet, enabling personalizedrecommendations.
The formula for content-based filtering can be represented as: score(u, i) = similarity(user profile(u), item features(i))
Where:

- score(u, i) represents the predicted score or relevance of item i for user u,
- user profile(u) represents the profile of user u, of ten derived from their past interactions or explicitly provided preferences,
- item features(i) represents the features or attributes of item i,
- similarity(·, ·) represents the similarity measure used to compute the similarity between the user profile and item features.

TF-IDF Vectorizer: In our implementation, we utilisedTerm Frequency-Inverse Document Frequency as ameans of computing similarity between data points.TF-IDF is a common technique used to transform textdata into numerical vectors, often used in natural language processing tasks like document classification andinformation retrieval. When applied to content-basedfiltering for recommendation systems, TF-IDF can beapplied to represent item features (e.g., movie genres)as numerical vectors [10].
The TF-IDF representation assigns weights to each term(feature) in a document (item) based on its frequencywithin the document and its importance across all documents. TF-IDF is given by the following formula:
$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$
where:

- TF(t, d) represents the term frequency of term t indocument d,
- IDF(t, D) represents the inverse document frequency of term t across all documents in corpus D.

Cosine Similarity:Cosine similarity is a metric which is used to computethe similarity between two vectors by determining thecosine of the angle between them. Applied to content-based filtering, we used cosine similarity in order toquantify the similarity between a user profile (whichrepresents user preferences) and item features (whichrepresents item attributes).The formula for cosine similarity between two vectorsA and B is given by:

$$\text{Cosine\_Similarity} = \frac{A.B}{\|A\|\|B\|}$$

Where:

- A · B represents the dot product of vectors A andB,
- ‖A‖ and ‖B‖ represent the Euclidean norms ofvectors A and B, respectively.

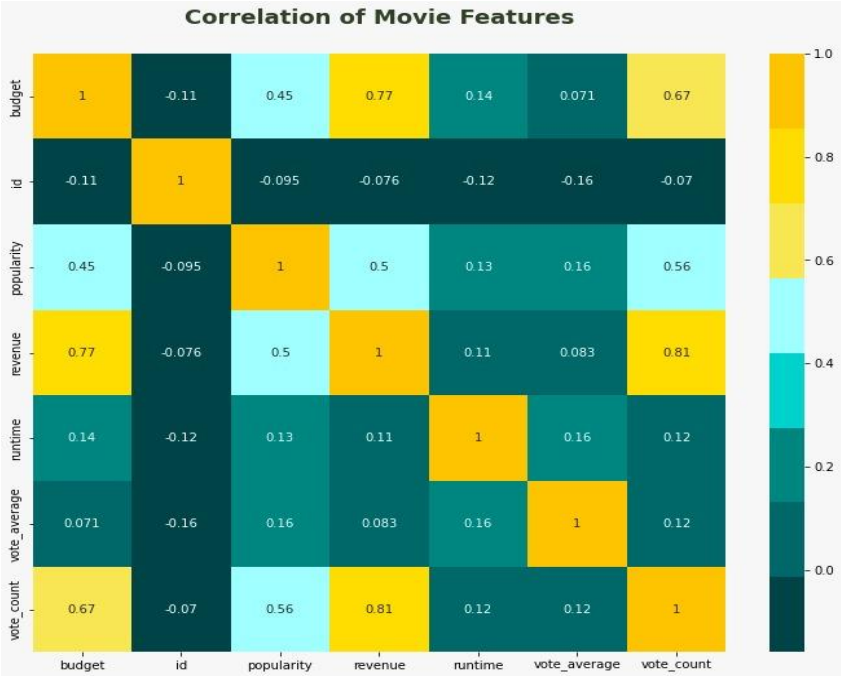

Fig. 1: Heat Map for Content Based Filtering

3) **Collaborative Filtering**: The predominant technique in recommendation systems is collaborative filtering, where the system predicts personalized recommendations to users based on their past interactions and tastes. Such models synthesize knowledge from the audience to predict new items likely to be of interest or utility to the user. We begin by introducing collaborative filtering and describing a formula for calculating predicted ratings, the methodology used to make recommendations [11].
In collaborative filtering, the predicted rating $\widehat{r_{ui}}$ for useru and item i is computed as a weighted sum of theratings given by similar users to item i. Mathematically,it can be expressed as:

$$\widehat{r_{ui}} = \bar{r}_u + \frac{\sum_{v \in N(u)} \mathrm{sim}(u, v).(r_{vi} - \bar{r}_v)}{\sum_{v \in N(u)} \mathrm{sim}(u, v)}$$

where

- $\widehat{r_{ui}}$ is the predicted rating of user u and item i
- $\bar{r}_u$ is the average rating given by user u
- r_{vi} is the rating given by user v to item i
- $\bar{r}_v$ is the average rating given by user v
- $N(u)$represents the set of users similar to user u
- $\mathrm{sim}(u, v)$is the similarity between users u and v, typically computed using metrics like cosine similarity or Pearson correlation.

To compute predicted ratings using collaborative filtering, the following steps areperformed:

i. **Data Preparation**: Load the MovieLens-100k dataset in the Surprise library. This is a tabular form of data containing user-item interactions.
ii. **Model Training**: Train the Singular Value Decomposition (SVD) model on the full dataset. SVD is type of the matrix factorization models majorly used in collaborative filtering.
iii. **Random User Selection**: Generate a random user ID to simulate a user for which recommendations will be generated.
iv. **Predicting Ratings**: Predict ratings on all items from the dataset of a randomly selected user. This is done by using the trained SVD model to predict the ratings.

C. Explainable AI (XAI)

The techniques within the domain of eXplainable AI (XAI)make an attempt to give transparency in machine learningmodels in a way that makes the decision process understand-
able. This is important since black-box models are hard tointerpret, which in turn increases a lack of trust and potentialbiases. XAI helps it to be possible that users understand whycertain decisions are made,

thereby increasing confidence andgiving stakeholders the opportunity to pinpoint and rectifypotential models [12].

XAI techniques play a big role in adding transparency andinterpretability to the recommendation systems, putting anend to the black box normally associated with these machinelearning models [4]. Incorporation of techniques such as theSHapley Additive exPlanations (SHAP) framework offers theuser insights into the recommendations made by the algorithms.

SHAP values are among the leading tools within the XAItoolbox, which allow for the well-systematic evaluation ofeach input feature's contribution to the output of a model.For a recommendation engine, such features could be userpreferences or item attributes, among other factors importantto the recommendation process. Computing SHAP values foreach recommendation offers the possibility of gaining deepinsight into the rationale behind each single recommendation[13].

The implementation of SHAP involves several tangiblesteps:

1) Training Recommendation Models: Our next step will be to train the recommendation models using the selected algorithms on the prepared dataset. For this, algorithms come into range like matrix factorization techniques: Singular Value Decomposition (SVD), content-based filtering, and collaborative-filtering methods.
2) Generating SHAP Values: Once we have trained the models, SHAP values are computed for each recommendation made by the model. We calculate these using the SHAP values of each feature in the input data and then display them in such a way that all the black boxes in the recommendation can be opened up.
3) Visualizing SHAP Values: We have applied relevant visualization techniques that make the presentation of SHAP values interpretable and understandable. Summary plots, or individual feature attribution plots, can be used to display the effect of any one feature on the model's output.
4) Analyzing SHAP Results: Finally, SHAP results were analyzed to unveil patterns, biases, or anomalies in the recommendation process. Through the evaluation of SHAP values for different recommendations, we get an insight into how the model was making decisions and what room there would be for further improvement or finetuning.

By integrating SHAP-based XAI techniques in our recommendation systems, users become more empowered tocomprehend and trust the recommendations being given tothem. This transparency gives a better experience not only tousers but also helps stakeholders identify issues and improvethe general performance of the recommendation algorithms.

D. Utilizing multimodal LLMs to improve explainability

Our setting enhances the interpretability of recommendationsystems by building on new advances in Multimodal LargeLanguage Models (LLMs) specifically using GPT-4. LargeLanguage Models are state-of-the-art models in a series of veryinfluential methods for very hard benchmarks. One differenceof the Multimodal LLMs is that they support different modalities, like text, images, and graphs. Built on deep learningand natural language processing, these models utilize advancedtechniques for the comprehension and generation of human-like responses across data types.

To improve the interpretability of our recommendationsystems, we leverage recent developments in the field andincorporate GPT-4 into our framework. The incorporationprocess is detailed as follows:

1) Graph Generation: A SHAP-based explanation graph, showing the contribution of each feature towards the recommendation model output, is generated. Generate graphs as generated output.
2) Input Preparation: Generate graphs as generated output.
3) Prompt Design: Prepare a prompt tailored to guide the GPT-4 model in analyzing and interpreting the explanation graphs properly.
4) Model Inference: Pass the explanation graphs and prompt through the GPT-4 model for inference.
5) Explainability Analysis: Analyze the output provided by GPT-4, with clear and concise explanations about the interpretability of the SHAP-based graphs.

4. RESULTS

A. Matrix Factorization

For understanding, the x-axis is utilized to plot distinctfeatures of the model that have been applied to predict userratings for items. Given these, the features can be furthercategorized as follows:

User and Item IDs: These are individually unique identifiersfor users and items.

User-Item Interaction Features: These describe how usersinteract with items.

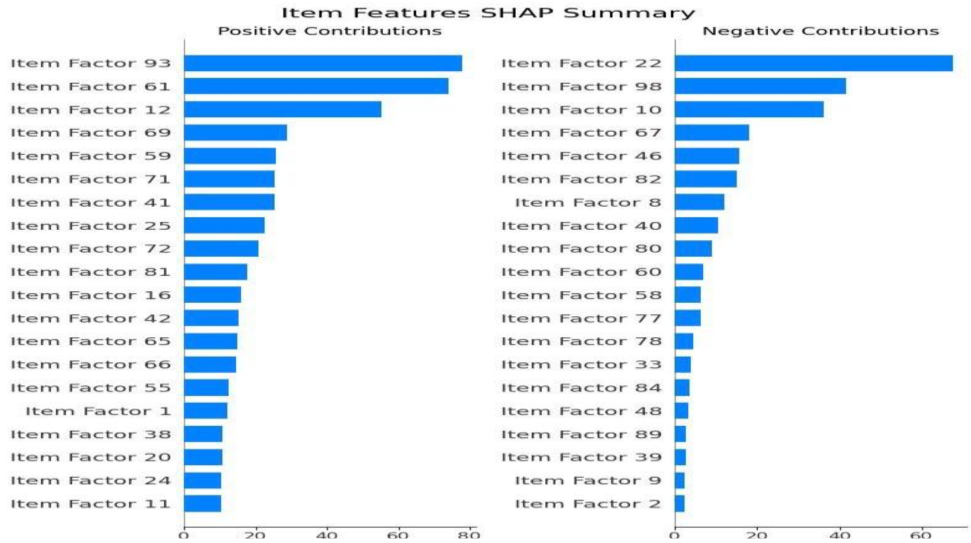

Fig. 2. Matrix Factorization- Highlighting SHAP VALUES and Average impact on model

The value of the SHAP values corresponding to a feature givesthe average magnitude of feature importance in a model'srating prediction for any user-item pair. SHAP values representthis feature's importance towards the prediction in either apositive or negative direction. Here, the absolute average valueis taken, so you can interpret it as how important a feature isin general for the model's output.

Higher values on the y-axis for a specific feature indicatethat, on average, changes to that feature have a more significantimpact on predicted ratings. For instance, a high value for "Item ID 10" suggests that the model heavily relies on a user'sinteraction with item 10 (or its inherent qualities captured byassociated latent factors) when making predictions. Lower values on the y-axis suggest that the feature has a relatively minorimpact on the overall magnitude of the model's predictions.This may indicate that the model finds other features moreinformative for making accurate predictions.

B. Content Based Filtering

Item Attributes

X-axis: Textual descriptions, genre categories, demographicinformation about actors/directors (for movies), and so on,depending on the domain of our recommender system.

Y-axis: The y-axis represents the average absolute SHAPvalue. The SHAP value for a specific feature and data point(user-item pair) indicates the impact (positive or negative) thatfeature has on the model's predicted rating for that data point.Averaging the absolute values provides a sense of the overallmagnitude of a feature's influence on the model's output.

Interpretation: If "action" has a high SHAP value, it suggeststhe model prioritizes the "action" genre when recommendingmovies to users who have watched action movies in the past.

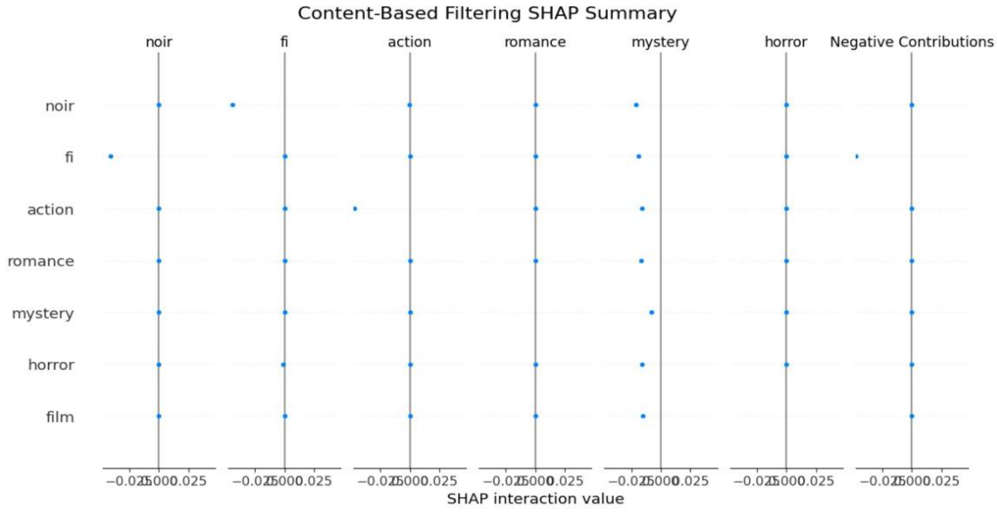

Fig. 3. Content Based Filtering- Highlighting SHAP VALUES and Average impact on model

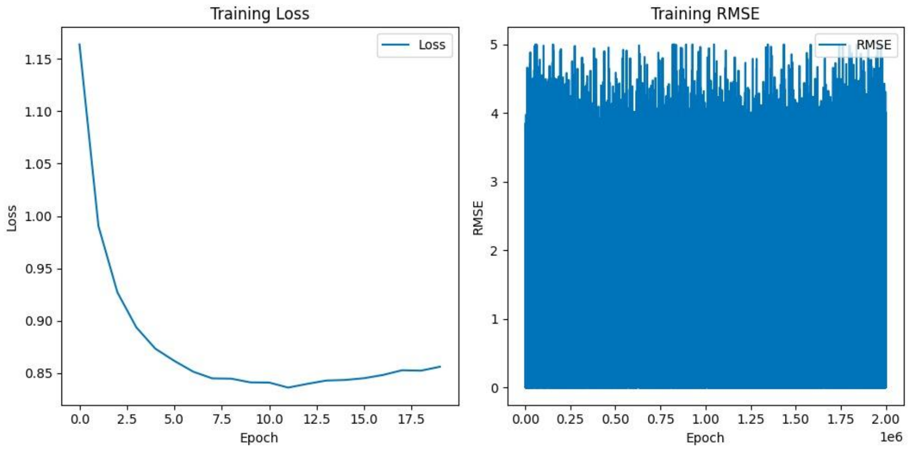

Fig. 4. Loss Function vs Epoch Graph for Collaborative Filtering

C. Collaborative Filtering

The features observed can be described as the following:

User ID: Encodes a specific user.

Item ID: Encodes a specific item.

User-Item Interaction Features: These capture interactionsbetween users and items, potentially including:

- Explicit ratings provided by users for items.
- Implicit interaction data like clicks, views, or purchases.

Interpretation of the Axes:

X-axis: The x-axis represents the SHAP values, rangingfrom negative (blue) to positive (red), with zero in the center (white). Negative values indicate the feature contributesnegatively to the predicted rating (reduces the rating), whilepositive values indicate a positive contribution (increases therating).

Y-axis: The y-axis represents the feature names or indicesdepending on the specific plot. Example: Imagine a dot for "Item ID 10" positioned relatively high on the positive side(red). This suggests that for many user-item interactions, userinteractions with item 10 (high ratings, frequent views, etc.)contribute positively to the predicted ratings for other items.This might indicate that users who liked item 10 also tend tolike similar items.

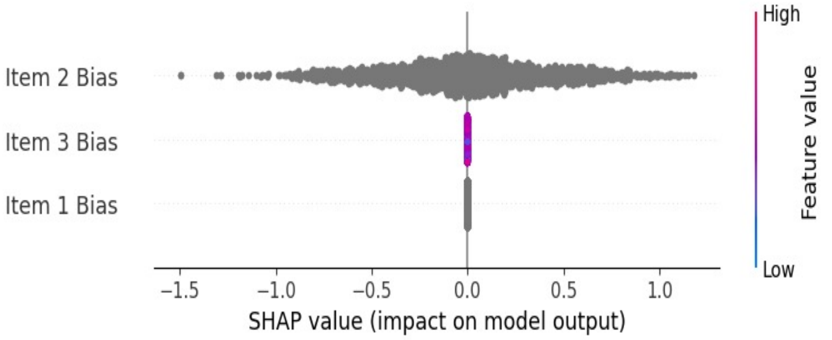

Fig. 5. Collaborative Filtering- Highlighting SHAP VALUES and Average impact on model

D. Explainability using GPT-4

As mentioned before, after obtaining individual graphs foreach recommendation system process, we passed the graphs tothe multimodal GPT-4 model along with a curated and testedprompt. The prompt and the interpretable response given byGPT-4 are given below.

1) **Prompt utilized**:

   You are an AI model which helps insimplifying and explaining researchon Explainable AI in RecommenderSystems.

   You will be given information inthe form of graphs about matrixfactorization, Content-based, andCollaborative filtering methods.Provide insight about SHAP values and
   why they are used.

Based on the calculated SHAP valuesfor all three methods, give aneasy-to-understand observation ofhow the particular value AFFECTS theparticular recommendation based onthe graphs provided for EACH method.

2) Response Generated

<START>

SHAP values are utilized because they offer a clear andquantifiable way to determine how each feature withina model influences the model's output. This is crucialfor transparency and trust, particularly in recommendersystems where decisions directly affect user experience.

1. Content-Based Filtering SHAP Summary

Graph Overview: This graph illustrates SHAP interaction values among various genres in a content-based filtering system. Each dot represents how muchthe interaction between two genres contributes to themodel's predictions.

Observations:

- Negligible Interaction Values: Most genre interactionsare clustered around zero, indicating minimal impact onthe recommendations. This suggests that while genrefeatures are used, their direct interactions do not significantly influence the recommendation output.
- Subtle Genre Effects: For a user, this means the systemdoesn't heavily rely on combined genre preferencesto make recommendations, possibly focusing more onindividual genre traits or other item attributes.

2. Item Bias in SHAP Values

Graph Overview: This visualization targets the biasassociated with specific items, with SHAP values indicating the direction and magnitude of their impact onthe model's output.

Observations:

- Significant Negative Bias for Item 1: The prominentnegative SHAP value for Item 1 indicates a strong biasagainst recommending this item, regardless of its contentor user preferences. It suggests an underlying factor inthe model or data that leads to consistently lower scoresfor this item.
- Minimal Bias for Other Items: Other items showa neutral bias, implying that their recommendationsare more influenced by user preferences or itemcharacteristics rather than inherent biases.

3. Item Features SHAP Summary from MatrixFactorization

Graph Overview: This chart shows the positive andnegative contributions of item factors (latent featuresderived from matrix factorization) to the model's predictions.

Observations:

- Positive Factors: High SHAP values for certain factors(e.g., Item Factor 93, 61) suggest these aspects of theitems strongly resonate with user preferences, drivingrecommendations.
- Negative Factors: Conversely, factors like Item Factor22 and 98 detract significantly from an item's likelihoodof being recommended, indicating traits that are generally unfavorable or irrelevant to user interests.

Conclusion:

These insights demonstrate the power of SHAP valuesin making machine learning models more interpretable.In recommender systems, understanding these contributions:

- Enhances Model Transparency: Allows developers toexplain why certain recommendations are made, increasing user trust.
- Guides Model Improvement: Pinpoints which features(or biases) have the most impact, suggesting whereadjustments or rebalancing might be needed to improverecommendation quality and fairness.

</END>

5. CONCLUSION

This work explores the important area of improved interpretability and transparency in recommendation systemsvia techniques extracted from Explainable AI (XAI). Withthe framework of SHapley Additive exPlanations (SHAP),we unlocked valuable insights into decision-making processesfrom a diverse variety of recommendation algorithms. Througha careful methodology of model training, the generation ofSHAP values, and visualization, we have presented the concrete steps to understand the explanation results.

The additional integration of a multimodal Large LanguageModel, as exemplified by the GPT-4 model, has further increased the power of explanation our approach offers manifold.

By letting the model learn how to critically analyze SHAP-generated graphs and communicate their meaning in clear,brief language, we bridged the gap between outputs of a blackbox model and human understanding.

Our exploration has underscored the significance of explainability in recommendation systems, not only in fostering usertrust and satisfaction but also in facilitating model refinementand domain insights. The findings presented herein lay asolid foundation for future research endeavors aimed at furtheradvancing the interpretability and usability of recommendationsystems.

REFERENCES

[1] C. Lucchese, C. I. Muntean, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini, "Recommender systems," in Encyclopedia ofMachine Learning and Data Mining, 2021. [Online]. Available:https://api.semanticscholar.org/CorpusID:1381259

[2] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang,S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpretingblack-box models: A review on explainable artificial intelligence,"Cognitive Computation, vol. 16, no. 1, pp. 45–74, 2024. [Online].Available: https://doi.org/10.1007/s12559-023-10179-8

[3] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainableai: A review of machine learning interpretability methods," Entropy,vol. 23, no. 1, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[4] R. Sewada, A. Jangid, P. Kumar, and N. Mishra, "Explainable artificialintelligence (xai)," international journal of food and nutritional sciences,2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260436532

[5] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainableai: A brief survey on history, research areas, approaches and challenges,"in Natural Language Processing and Chinese Computing, J. Tang, M.-Y.Kan, D. Zhao, S. Li, and H. Zan, Eds. Cham: Springer InternationalPublishing, 2019, pp. 563–574.

[6] Z. Buc¸inca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasksand subjective measures can be misleading in evaluating explainableai systems," in Proceedings of the 25th International Conferenceon Intelligent User Interfaces, ser. IUI '20. New York, NY, USA:Association for Computing Machinery, 2020, p. 454–464. [Online].Available: https://doi.org/10.1145/3377325.3377498

[7] F. M. Harper and J. A. Konstan, "The movielens datasets: History andcontext," ACM Trans. Interact. Intell. Syst., vol. 5, no. 4, dec 2015.[Online]. Available: https://doi.org/10.1145/2827872

[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques forrecommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[9] M. Curmei, W. Krichene, L. Zhang, and M. Sundararajan, "Privatematrix factorization with public item features," Proceedings of the 17thACM Conference on Recommender Systems, 2023. [Online]. Available:https://api.semanticscholar.org/CorpusID:261823311

[10] M. Manwal, D. Rawat, D. Rawat, K. C. Purohit, andT. Choudhury, "Movie recommendation system using tf-idf vectorizer and bag of words," 2023 12th InternationalConference on System Modeling & Advancement in ResearchTrends (SMART), pp. 163–168, 2023. [Online]. Available:https://api.semanticscholar.org/CorpusID:267773787

[11] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua,"Neural collaborative filtering," Proceedings of the 26thInternationalConference on World Wide Web, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13907106

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: A surveyon explainable artificial intelligence (xai)," IEEE Access, vol. 6, pp.52 138–52 160, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52965836

[13] A. Zern, K. Broelemann, and G. Kasneci, "Interventional shapvalues and interaction values for piecewise linear regression trees," inAAAI Conference on Artificial Intelligence, 2023. [Online]. Available:https://api.semanticscholar.org/CorpusID:259746710