Journal of Data Science and Modeling, Vol. 2, No. 2, 119-147, June 2024

Research Manuscript

Intelligent Travel Recommendations Using Neural Collaborative Filtering for Touristic Landmarks of Iran

Mohammad Hossein Zolfagharnasab,
Latifeh PourMohammadBagher*, Mohammad Bahrani
Faculty of Computer Science, Department of Mathematics, Statistics, and Computer,
Allameh Tabataba'i University, Tehran, Iran.

Recieved: 10/11/2024 Accepted: 16/04/2025

Abstract: This study introduces a tailored recommendation system aimed at enriching Iran's tourism sector. Using a hybrid model that combines neural collaborative filtering (NCF) with matrix factorization (MF), our approach leverages both demographic and contextual data of the combined tourist-landmark (4177 samples) to provide personalized tourism recommendations. Empirical evaluations on the implemented methods show that the hybrid model outperforms factorization techniques, achieving a test F1 score of 0.84, accuracy of 0.90, and a test error reduction from 0.83 to 0.37. Feature vector integration further improved test recall by 17%, underscoring the model's robustness in capturing user-item relationships. Further analysis using t-SNE as well as visual analyses of embedding structures confirm the system's ability to generalize patterns in latent space; thereby, mitigating the cold-start problem for new tourists or unvisited landmarks. This study also contributes a structured dataset of Iranian landmarks, user ratings, and supplementary contextual data for fostering future research in culturally specific intelligent recommender systems. For implementation details, refer to the GitHub repository at https://github.com/MsainZn/Collaborative_Filtering_Tourism_Landmarks.

Keywords: Neural collaborative filtering; Matrix factorization; Recommender Systems; Deep Information Retrieval Systems; Intelligent Tourist Management.

*Corresponding Author: l_pmb@atu.ac.ir

1. Introduction

Iran's tourism industry has become a hidden gem, with its rich historical sites and diverse natural landscapes making it an ideal destination for cultural and nature tourism. Cities like Isfahan, Shiraz, and Kurdistan showcase Iran's historical and cultural heritage, while regions such as Gilan, Mazandaran, and Azerbaijan offer stunning natural beauty as well as serve as special tourist destinations.

Despite the variety of attractions, tourists struggle to discover their interests, relying on limited recommendations or inflexible technologies like expert systems, resulting in suboptimal experiences and missed opportunities for personalized exploration (Zolfagharnasab *et al.* (2025)). Over the past years, progress in this area has been limited due to four main issues: a lack of open datasets, intense market competition, outdated systems in existing studies, and privacy concerns discouraging data sharing.

Given the outlined limitations, this study undertakes four fundamental objectives to address the described challenges:

- **Introducing a native dataset.** Collecting a sufficient, standard dataset of Iranian tourist spots along with user ranking scores to facilitate future research and system development in the tourism sector.
- **Comparing different methods.** Evaluating and comparing various recommendation techniques, from matrix factorization up to neural-based models, to determine the most effective approach for tourism recommendations.
- **Addressing the cold-start problem.** Utilizing implicitly learned vector embeddings to overcome the cold-start problem for new users and landmarks, ensuring effective recommendations even with limited initial data.
- **Feasibility of pre-trained models.** Investigating the potential of using pre-trained models to enhance the accuracy and efficiency of the recommendation system.

With these objectives in mind, the remainder of this study is structured in seven more sections. Section 2 presents the *literature review*, covering relevant studies in the field and providing benchmarks for comparison. Section 3 introduces the *collected dataset*, followed by the necessary *preprocessing steps* in Section 4. The *proposed method* is outlined in Section 5, explaining the methodology, implementation, and evaluation metrics. Section 6 presents the *results and discussion*, detailing the experimental results and their analysis. Finally, Section 7 provides the *conclusion*, summarizing key findings and suggesting potential future research directions. For more interested readers, the statistical analysis performed on the dataset is also presented in the *Appendix*.

2. Related Work

Collaborative filtering (CF) has a profound impact on consumer decisions, making it a central focus of research, particularly in personalized retrieval systems (Ricci *et al.* (2015)). This area of study has consistently attracted a diverse range of researchers due to its critical role in generating recommendations that align with individual user preferences. As a foundational technology in industries like e-commerce, entertainment, and social media, CF is essential for delivering tailored experiences and driving user engagement (Schafer *et al.* (2007)).

The initial breakthroughs in CF models came with the development of Matrix Factorization (MF) techniques (Koren (2009)). These techniques addressed the challenge of sparsity in the user-item interaction matrix by transforming it into lower-dimensional hidden embeddings. These embeddings captured latent factors that represented the underlying relationships between users and items, thus allowing the system to predict interactions even in cases where explicit data was sparse. Matrix factorization helped uncover patterns that were not immediately visible, making it a crucial step in personalized recommendations (Mnih and Salakhutdinov (2008)).

Despite their success, the reliance of MF-based approaches on linear interactions between latent factors limited their ability to capture complex, non-linear relationships (Rendle (2010)). This limitation spurred further research into more sophisticated methods, including Neural Collaborative Filtering (NCF), which leveraged the flexibility of neural networks to model the non-linear interactions between users and items (He *et al.* (2017)). By using neural networks, NCF allowed for more accurate predictions, particularly in sparse data scenarios or when modeling highly complex user preferences (Cheng *et al.* (2016)).

Due to the capabilities and potential of the NCF approach in modeling complex user-item relationships, variations of this method, such as DeepFM (Guo *et al.* (2017)), Neural Attention Collaborative Filtering (Chen *et al.* (2017)), and Graph Neural Collaborative Filtering (Wang *et al.* (2019)), quickly found their way into industry. These models are particularly effective in handling large-scale, sparse datasets while incorporating additional features such as user demographics, behavior patterns, and item characteristics, thus improving the overall relevance and engagement of the recommendations provided to users. In addition to the standard functionality, tech companies like Amazon, Netflix, and Facebook complemented their baseline models with feature extraction models for processing multi-modal inputs (audio, video, pictures, etc.) to enhance their recommendation engines, enabling them to provide more personalized and accurate content recommendations (Zhou *et al.* (2018)). Studies in this area are also found interesting and fruitful due to the development of Large Language Models (LLMs) and Vision

Transformers for guiding information retrieval (Zolfagharnasab *et al.* (2024)). Nevertheless, it should be noted that the CF component is yet to be the primary part that associates the user-item relation, and the LLMs and feature extractors are generally used to provide more semantic information.

Despite the described evolution (from matrix factorization to neural-based approaches), studies such as Ferrari Dacrema *et al.* (2019) have also provided invaluable arguments, highlighting persistent issues, including weak baselines, inadequate hyper-parameter tuning, and challenges in comparing methods across diverse datasets as factors that might change the scale in favor of traditional MF methods as well. As a result, the research community is advised to assess the effectiveness of the CF methods based on the application's context and the problem description.

In tourism applications—the focus of this study—the impact of Recommender Systems (RS) has been profound, as they play a crucial role in guiding travelers toward personalized experiences (Fesenmaier *et al.* (2006)). Many studies have examined the role of CF in improving the effectiveness of tourism recommendations. For instance, in the study by Zanker *et al.* (2008), the authors explored how collaborative filtering can be used to tailor recommendations for tourism packages, showing that CF-based approaches significantly improve user satisfaction by matching individual preferences with travel itineraries. In another example, Ricci and Del Missier (2011) conducted a study that examined the use of CF in tourism, where the system was employed to suggest personalized destinations based on the preferences of similar users. Their work demonstrated that CF not only enhances the quality of recommendations but also increases user engagement by offering a more interactive and user-friendly experience. Similarly, Kabassi (2010) explored the use of CF in cultural tourism, where recommendations were personalized based on user interests in various cultural landmarks. The study demonstrated that CF-based recommendation systems help increase user satisfaction by suggesting culturally relevant sites and improving the overall tourism experience.

These examples underscore the importance of RSs in the tourism industry, where personalized travel experiences have become essential for enhancing customer satisfaction and driving engagement. Unfortunately, in terms of native works, there is a notable scarcity of research and development within local contexts, which can be traced back to the unavailability of open datasets. This gap highlights the need for more targeted efforts in developing native datasets, models, and approaches that address the unique preferences and behaviors of tourists, as well as the diverse range of tourist attractions within the country.

3. Dataset

The dataset used in this study comes from domestic trips organized by a private travel agency, which stored the information in Excel files. Due to industry competition and privacy concerns, data from the past four years (since the COVID-19 pandemic) has been withheld, and the dataset covers the period from 2012 to 2020. The dataset includes essential details such as the name of the visited landmark, users' city of birth, date of birth, and the rating assigned to the attraction (1–5). To increase flexibility and ensure consistency with standard database formats, the dataset is divided into three distinct tables: user information, tourist landmark data, and a rating table. The general dataset properties are presented in Table 1, and the CSV files corresponding to the tables are also published in Repository. The following section offers a more detailed analysis of the statistical information within the dataset.

Table 1: General Dataset Properties

Table	**Column Information**	**Count**	**Size**
Landmark	ID, Name, Category, City, Province, Payment	309	17 KB
Tourist (user)	ID, City, Province, Age	200	5 KB
Rating	ID, Attraction ID, User Rating	4177	42 KB

4. Data Preprocessing

This section outlines the preprocessing steps carried out to prepare the dataset for subsequent model implementations.

Age Binning. The continuous age data was discretized into specific age groups. Tourists are classified into three categories: young (20–39), middle-aged (40–59), and elderly (60 and above). This classification is based on the approximate retirement age and societal norms within Iranian culture.

Price Categorization. Landmark ticket prices were divided into three categories: free, inexpensive (under 500K), and expensive (over 500K) IR-Rials. This categorization is based on the principle that the landmark ticket fee should not exceed 10% of the average hotel price, which is around 5M IR-Rial per night.

Encoding Categorical Features. To enable efficient processing during model training, each categorical feature was converted into numeric labels. Table 2 shows the encoding scheme applied to each feature. Although tourist and landmark IDs were encoded earlier, they are included in the table for completeness. A dictionary data structure was used to manage these transformations, ensuring the reversibility

of the changes.

Table 2: Feature Encoding Used in the Dataset

Feature Name	Original Data	Encoded Format
Age	Young, Middle-aged, Elderly	[0, 1, 2]
Provinces	31 provinces labeled with words	[0, 1, ..., 30]
Cities	117 cities with tourist attractions	[0, 1, ..., 116]
Prices	Free, Inexpensive, Expensive	[0, 1, 2]
Categories	Historical, Natural, Cultural, Recreational, Architectural	[0, 1, 2, 3, 4]
Tourist IDs	200 tourists in the dataset	[0, 1, ..., 199]
Landmark IDs	309 tourist attractions in the dataset	[0, 1, ..., 308]

Data Splitting. The standard random data splitting method is unsuitable for RS, as such systems are not explicitly trained to predict for completely new users or items. Instead, a masking technique is employed, in which part of a user's data is hidden during training and revealed during testing. This ensures that all users and items are present in both phases, allowing for accurate system evaluation. In this study, 80% of each user's ratings were used for training, while the remaining 20% were reserved for testing, ensuring robust performance evaluation.

Following the described preprocessing steps, the dataset is prepared for pipelining into the models and generating recommendations.

5. Proposed Method

This section details the proposed methods, covering key aspects such as mathematical modeling, key parameters, and model implementation. The techniques used range from traditional mathematical methods like matrix factorization to modern architectures like the tower model. The aim is to establish a foundation for comparing different collaborative filtering and hybrid models in later chapters.

Standard matrix factorization (MF) is a technique that decomposes the user-item rating matrix R into two lower-dimensional matrices representing latent factors of users and items. The goal is to find two matrices, $P \in \mathbb{R}^{m \times k}$ and $Q \in \mathbb{R}^{n \times k}$, where m and n are the numbers of users and items, respectively, and k is the dimension of the latent factors (embedding vectors). The predicted rating $\hat{r}_{ui}$ for user u and item i is calculated as the dot product of the user and item latent vectors, as shown in Eq. (5.1):

$$\hat{r}_{ui} = P_u^\top Q_i = \sum_{f=1}^{k} P_{uf} Q_{if}, \tag{5.1}$$

where P_u and Q_i are the user-specific and item-specific latent vectors, respectively. The optimization objective is to minimize the difference between the predicted ratings $\hat{r}_{ui}$ and the actual ratings r_{ui} by minimizing the squared error loss, as in Eq. (5.2):

$$\mathcal{L} = \sum_{(u,i)\in\mathcal{K}} (r_{ui} - \hat{r}_{ui})^2 + \lambda \left(\sum_u \|P_u\|^2 + \sum_i \|Q_i\|^2 \right), \tag{5.2}$$

where $\mathcal{K}$ is the set of known user-item interactions, and λ is a regularization term that prevents overfitting by controlling the magnitudes of the learned latent vectors. The key parameter in this technique is the embedding dimension k, which determines the size of the latent factors P_u and Q_i. However, the standard MF model has several limitations:

- **No Contextual Information**: It only estimates ratings without incorporating additional contextual information such as item descriptions or user preferences.
- **No Bias or Initialization Control**: It does not account for user or item biases, proper weight initialization, or constraints to ensure valid rating predictions within a specified range.
- **Cold-Start Problem**: It lacks mechanisms to handle scenarios where new users or items have insufficient historical data for accurate predictions.

To address these limitations, several modifications were implemented in the Modified Matrix Factorization (MMF) model. These include the introduction of user and item bias terms, a global bias, weight initialization, and the use of a sigmoid function. The prediction formula of MMF is given by Eq. (5.3):

$$\hat{r}_{ui} = 5.5 \cdot \sigma(\mu + b_u + b_i + P_u^\top Q_i), \tag{5.3}$$

where μ is the global bias, b_u and b_i are the bias terms for user u and item i, and σ is the sigmoid function. This formulation incorporates biases, improved weight initialization, and prediction normalization to enhance accuracy and stability. Despite these improvements, MMF still lacks the ability to incorporate additional user and item information or address the cold-start problem.

To overcome these challenges, Factorization Machines (FM) were implemented. FM extends MMF by modeling pairwise interactions between features, enabling more complex and flexible modeling. Unlike MMF, FM can incorporate contextual information about users and tourist attractions, thereby addressing the cold-start problem through content-based filtering techniques. The prediction function of FM is defined in Eq. (5.4):

$$\hat{y}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j, \tag{5.4}$$

where w_0 is the global bias, w_i is the weight for the i-th feature, and $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \mathbf{v}_i^\top \mathbf{v}_j$ represents the interaction between the latent vectors of features i and j. The terms x_i and x_j correspond to the values of the input features. By embedding features into latent vectors, FM can effectively generalize to previously unseen users and items. Different embedding vectors were assigned to each input feature, as outlined in Table 3.

Table 3: Summary of Embedding Parameters (Model Inputs)

Feature Name	**Embedding Dimension**	**Trainable Parameters**
Age	4	4×3
Provinces	8 for user, 8 for item	16×200
Cities	8 for user, 8 for item	16×117
Cost	4	4×3
Landmark Categories	4	4×5
Tourist IDs	16	16×200
Landmark IDs	16	16×309
Total	**76**	**13,260**

Despite these advantages, FMs are limited to modeling only pairwise interactions and are incapable of capturing non-linear relationships among features. To address this limitation, Generalized Factorization Machines (GFM) were implemented. GFM extends the capabilities of FM by introducing more flexible interaction functions capable of modeling non-linear and higher-order interactions among features. The GFM prediction function is defined in Eq. (5.5):

$$\hat{y}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \Phi(\mathbf{v}_i, \mathbf{v}_j) x_i x_j + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} \Psi(\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k) x_i x_j x_k + \dots, \tag{5.5}$$

where Φ and Ψ represent generalized functions that model interactions between latent vectors. In this study, triplet interactions were considered, and a non-linear Rectified Linear Unit (ReLU) function was applied, as formulated in Eq. (5.6):

$$\Phi(\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k, \dots) = \sigma(\mathbf{W}[\mathbf{v}_i; \mathbf{v}_j; \mathbf{v}_k; \dots] + \mathbf{b}), \tag{5.6}$$

where σ denotes the ReLU activation, $\mathbf{W}$ is the weight matrix, and $\mathbf{b}$ is the bias vector. While GFM is capable of modeling both linear and non-linear feature

interactions, it remains limited by the number of interactions that can be feasibly computed.

To overcome these constraints, neural networks (NN) were utilized due to their ability to capture complex, non-linear relationships between users, items, and additional features. Unlike MF or GFM, which rely on predefined interaction structures, NNs can learn these patterns directly from the data through multiple hidden layers and non-linear activation functions (Zolfagharnasab and Damari (2024)). In this study, a neural network architecture with five layers was implemented. The input consists of the embedding vectors for users and items ($\mathbf{E}_u$ and $\mathbf{E}_i$), and the hyperbolic tangent (tanh) function is employed as the activation. The properties of the layers are presented in Table 4.

Table 4: Architecture of the Neural Network-Based Model

Layer Number	**Input Size**	**Output Size**	**Bias Size**	**Num. Parameters**
Layer 1	76	128	128	9,856
Layer 2	128	32	32	4,128
Layer 3	32	16	16	528
Layer 4	16	8	8	136
Final Layer	8	1	1	9
Embedding	-	-	-	1,632
Total	-	-	-	**16,289**

While neural networks encapsulate the strengths of earlier methods, studies such as Ferrari Dacrema *et al.* (2019) have highlighted their instability during training, which can lead to inconsistent performance.

To mitigate these issues, a hybrid model—commonly referred to as the *tower architecture*—was implemented for user-item modeling. As illustrated in Figure 1, this architecture integrates the strengths of both Matrix Factorization (MF) and Neural Networks (NN) by leveraging their respective embedding vectors. This design aims to effectively capture both linear and non-linear relationships between users and items, thereby enhancing predictive reliability and overall accuracy. The configuration details of the tower model are presented in Table 5.

By combining GMF and NN, the hybrid model is capable of capturing a broader spectrum of interactions and patterns in the data. This integration addresses the limitations inherent in individual models and provides a more robust and accurate recommendation framework.

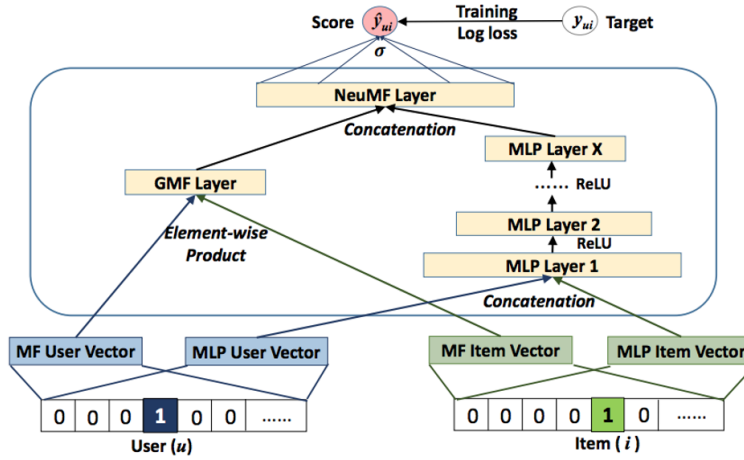


Figure 1: A schematic of the tower model (adapted from He *et al.* (2017)).

Table 5: Hybrid Model (Tower Architecture) Properties

Layer Component	Input Size	Output Size	Number of Parameters
Matrix Factorization (GMF)	76	1	13,260
Neural Network	76	1	29,549
Final Merging Layer	2	1	2
Total	–	–	**42,811**

6. Results and Discussion

This section presents an analysis and evaluation of the performance of the implemented models. First, the accuracy and reliability of the models' predictions are assessed to validate the core functionality of the recommender system (RS). A comparative analysis is then conducted using feature vectors. Finally, two critical issues are examined: the feasibility of employing pre-trained models and strategies for addressing the cold-start problem.

6.1 Model Comparison

Figure 2 illustrates the performance trends of the implemented models across 200 epochs. The training error curves exhibit a consistent decline for all models as training progresses, indicating effective learning. Models such as standard MF and NN converge rapidly in the initial epochs, demonstrating efficiency in minimizing error. In contrast, GFM shows slower convergence, likely due to its increased computational complexity. Notably, the hybrid (Tower) model achieves the lowest training error, underscoring its superior ability to extract meaningful patterns

from the data.

Test error trends follow a similar pattern, albeit with variability in convergence rates. The standard MF model shows a relatively gradual decrease in error, highlighting its limited capacity to generalize to unseen data. This is expected given its simplicity and constrained feature interaction capabilities. In contrast, GMF, NN, and the Tower architecture attain substantially lower test errors, indicating stronger generalization performance. Eventually, all models stabilize, suggesting effective mitigation of overfitting, supported by prior hyperparameter tuning.

The accuracy metrics across training and test phases reveal additional insights. While most models demonstrate progressive accuracy gains with increasing epochs, MF and FM display some fluctuations, suggesting potential difficulty in consistently balancing precision and recall. On the other hand, NN and Tower models maintain higher and more stable accuracy levels, indicative of their robust predictive performance.

Recall values mirror the trends seen in accuracy, with GMF and Tower models outperforming others in both training and testing sets. These results imply better identification of true positives, a critical factor in recommendation quality.

Finally, the F1 score, which harmonizes precision and recall, improves steadily for all models throughout training and testing. The Tower architecture achieves high F1 scores early, reflecting its effectiveness in balancing false positives and false negatives. GMF and FM exhibit more gradual gains, likely due to slower convergence in this balance. The test F1 scores show close competition between the Tower and NN models, with the Tower model maintaining a slight edge.

In conclusion, the Tower architecture emerges as the top-performing model, excelling across all key evaluation metrics including error reduction, accuracy, recall, and F1 score. Its balanced design allows it to outperform other approaches in capturing both linear and nonlinear user-item interactions. NN and GMF also deliver strong results but fall slightly short of the hybrid model's comprehensive performance.

To facilitate a clearer comparison of the implemented models, Table 6 presents a summary of evaluation metrics and the optimal performance achieved during both the training and testing phases. Among the models, the Tower and NN architectures consistently deliver superior results across key metrics. However, the hybrid nature of the Tower model results in slightly longer training times compared to simpler models such as MF, with the exception of GFM.

As anticipated from Eq. 5.5, the GFM model incurs increased training time due to the computational complexity of modeling triadic feature interactions. Despite this added complexity, its relatively constrained nonlinear structure limits its predictive power when compared to NN-based models. This observation reinforces

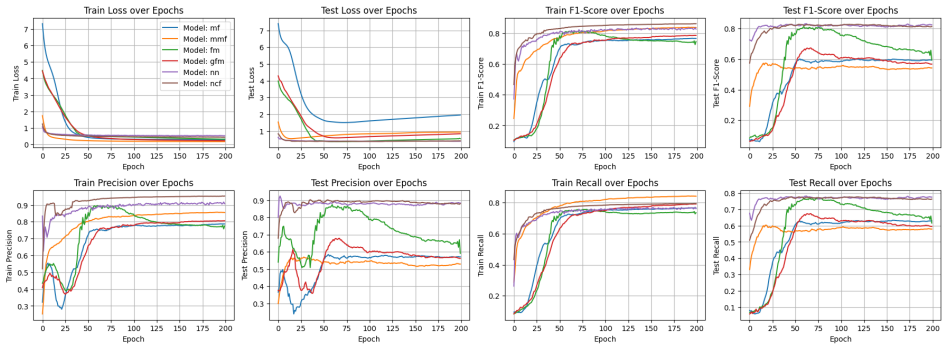


Figure 2: Performance comparison of different models.

an important insight: greater computational demands do not necessarily translate into improved model accuracy.

Table 6: Summary of model performance (train, test) across various metrics.

Model Name	MF	MMF	FM	GFM	NN	Hybrid
Error	(0.300, 1.825)	(0.191, 0.834)	(0.520, 0.3716)	(0.430, 0.596)	(0.531, 0.389)	(0.465, 0.384)
Accuracy	(0.778, 0.576)	(0.837, 0.545)	(0.896, 0.861)	(0.741, 0.671)	(0.911, 0.880)	(0.947, 0.896)
Recall	(0.759, 0.631)	(0.821, 0.589)	(0.750, 0.761)	(0.700, 0.667)	(0.765, 0.777)	(0.787, 0.775)
F1 Score	(0.759, 0.599)	(0.817, 0.555)	(0.813, 0.804)	(0.716, 0.666)	(0.831, 0.828)	(0.855, 0.837)
Training Time (s)	18.68	16.82	17.879	34.840	18.182	26.700

To further enhance interpretability, Figure 3 visualizes the simulation results as discrete histograms for each model across all evaluation metrics. A key observation is the similarity of evaluation metrics for the Tower and FM models across both training and testing sets, underscoring their strong generalizability. Additionally, the pronounced differences in error values and squared error metrics between the standard MF and generalized MF models highlight the performance improvements gained through feature augmentation. These disparities further emphasize the contributions of enhancements such as boundary functions (e.g., sigmoid), refined initialization strategies, and the inclusion of user and item bias terms.

6.2 Model Validation

While the evaluation metrics discussed earlier offer a broad overview of the models' performance, they do not provide a detailed analysis of the reliability of the predicted outputs. This section delves into this aspect of model validation.

Table 7 presents 40 randomly selected locations, along with their average user ratings and the model's predicted ratings. The results show that the estimated ratings closely align with the actual ratings in the dataset, suggesting that the

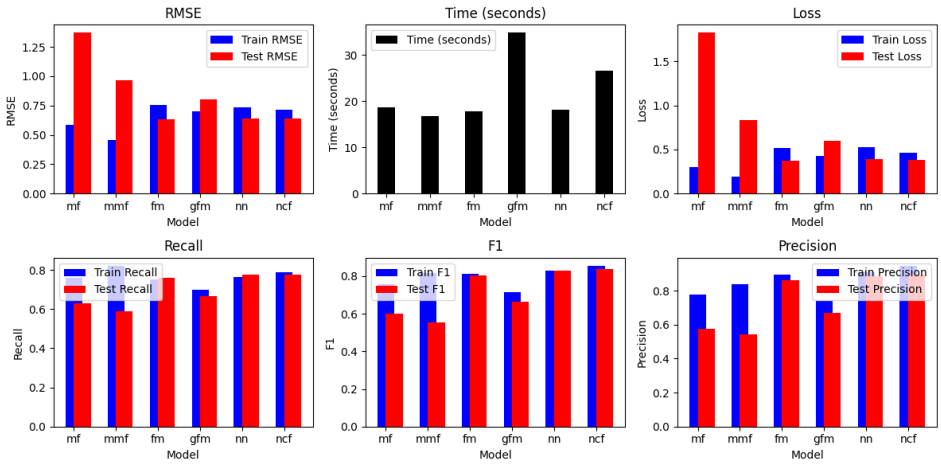


Figure 3: Performance comparison of different models using histogram.

model's predictions are reliable. For example, high-traffic landmarks such as Imam Reza Shrine and Naqsh-e Jahan Square exhibit predicted ratings that match the real ratings, likely due to their higher number of user reviews.

On the other hand, locations like Birjand Palace exhibit significant discrepancies between the actual and predicted ratings. This discrepancy can be attributed to the small number of ratings (only three), which limits the model's ability to generalize effectively. For instance, Birjand Palace received two ratings of 1 and one of 3, leading to distributional challenges that affect the prediction accuracy.

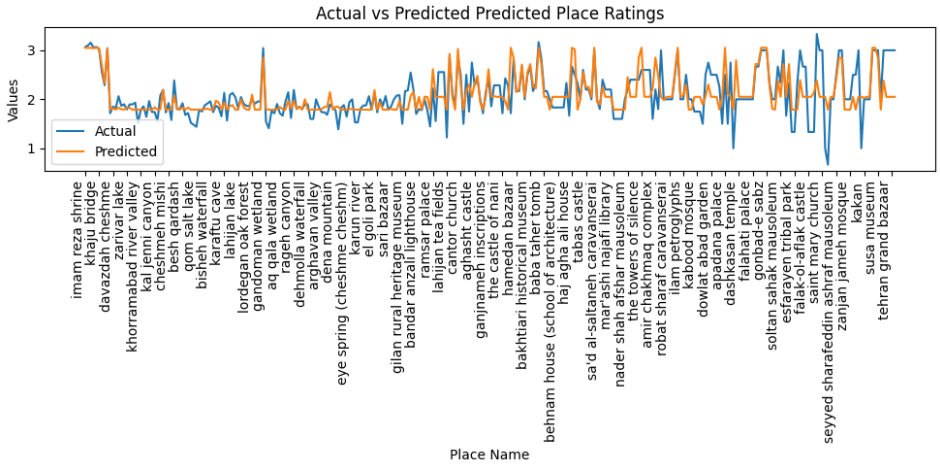


Figure 4: Comparing model average prediction with average of the actual ratings for the visited landmarks.

Table 7: Average Ratings vs Model Predictions for 40 Landmarks Selected from Dataset

Location	Pred	Label	Location	Pred	Label
Tange Tizab	2.855	2.8966	Imam Reza Shrine	4.0588	4.0671
Namakdan Cave	2.8181	2.8966	Naqsh-e Jahan Square	4.0652	4.0982
Zarivar Lake	2.8953	2.7931	Hafez Tomb	4.0555	4.156
Badab Soort Springs	2.8183	2.931	Pasargadae	4.056	4.0571
Nayband National Park	2.7801	2.5172	Ali Qapu Palace	4.0651	4.0673
Khorramabad River Valley	2.7838	2.7857	Khaju Bridge	4.0652	4.0297
Ashuradeh Island	2.8343	2.8571	Chehel Sotoun Palace	3.5324	3.56
Alangdareh Forest Park	2.7801	2.6429	Vakil Bazaar	3.0859	3.2887
Anzali Lagoon	2.7815	2.9643	Si-o-se-pol Bridge	4.0652	4.043
Pashtkouh Protected Area	2.7801	2.7407	Churt Lake	2.7801	2.7143
Birjand Palace	3.525	1.667	Falak-ol-Aflak Castle	3.014	2.333
Kal Jenni Valley	2.7817	2.7407	Dozdeh Cheshmeh	2.848	2.8529
Ghasemabad Forest	2.9606	3.0741	Rudkhan Spring	2.8874	2.8182
Hamun Lake	2.8354	2.5556	Kaboudval Waterfall	2.9168	3.0625
Ali Sadr Cave	3.1264	3.1923	Lar Dam	2.818	2.8667
Khorramabad Sarab	2.8185	2.7308	Golestan National Park	2.8169	2.9
Ferdowsi Tomb	3.014	4	Chaharfasl Bathhouse	3.014	4
Kish Island	3.014	2.6	Milad Tower	4.0641	4
Naqsh-e Rustam	3.434	3.6	Mugan Lagoon	3.277	3.5
Tabas Palace	3.853	3.2	Green Dome	2.780	2.333

To further visualize the model's accuracy in predicting user scores, Figure 4 compares the model's performance in estimating ratings for various landmarks against the actual values. Consistent with the results in Table 7, the system performs well for most popular landmarks, closely matching the real values. However, it exhibits higher errors for lesser-visited locations. For example, attractions such as the Jameh Mosque of Qorveh, which has conflicting ratings (4, 1, and 1), demonstrate greater uncertainty, resulting in a larger deviation from the real mean.

It is worth noting that enhancing model performance could be achieved by excluding infrequent data points and outliers, potentially improving prediction accuracy. However, this approach would involve removing lesser-visited locations, which was not implemented in this study.

To gain a deeper understanding of our model's performance, Figure 5 presents a histogram of residuals alongside density plots for actual and predicted ratings. The blue and red lines represent the densities of actual and predicted values, respectively, while the black histogram illustrates the residuals, highlighting the deviations between actual and predicted ratings. The density distributions of

actual and predicted ratings are closely aligned, suggesting that our model effectively captures the statistical properties of the dataset. The residuals are centered around zero, indicating generally good performance. However, the deviations in the histogram also reveal some prediction errors, primarily due to the decision not to remove outliers. Additionally, the system tends to assign an average score of 3 to items with limited information (i.e., few records). This behavior could be improved by either increasing the data volume or excluding records with low data counts, although this was not done due to the limited size of the dataset.

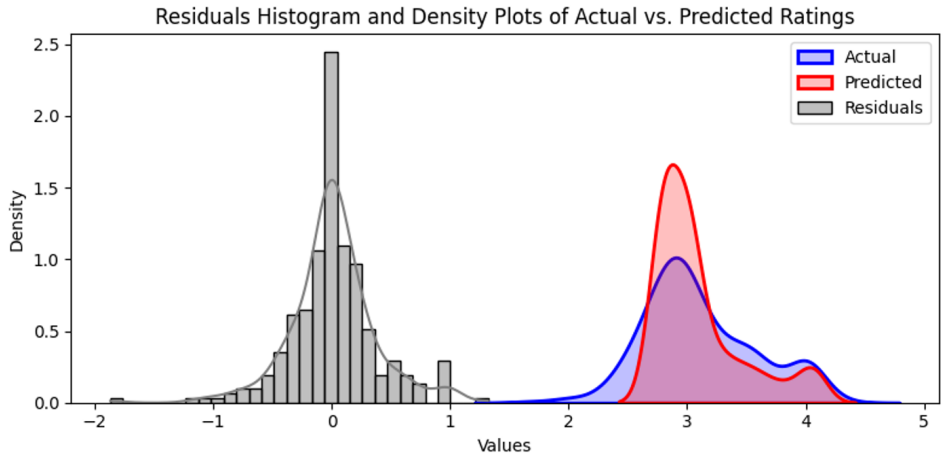


Figure 5: Histogram of residuals and density of actual vs. predicted ratings.

For the final assessment, Figure 6 presents a Bland-Altman plot, a technique for comparing the differences between actual and predicted values against their mean. In this plot, the vertical axis shows the differences, while the horizontal axis displays the mean. The dashed red line at zero represents the mean difference, with gray dashed lines indicating the 95% agreement limits. Data points are shown as blue dots, where a concentration near the horizontal axis indicates small discrepancies, and higher densities above the red line suggest larger errors. Most data points cluster around the mean difference line, indicating that system predictions are generally accurate. However, points farther from this line signify larger prediction errors, with positive differences representing over-predictions and negative differences representing under-predictions.

With confidence in the reliability of the predictions, the next section explores the feasibility of pre-training in the hybrid model.

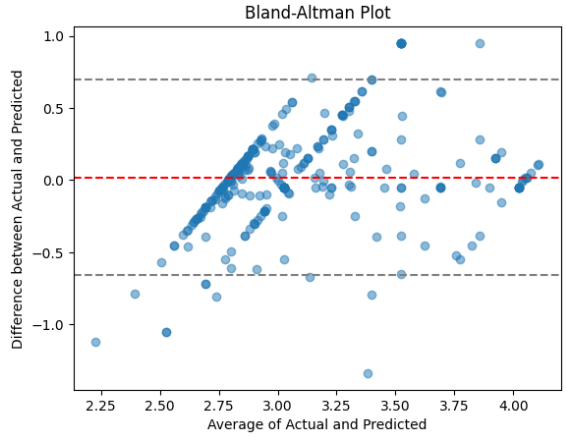


Figure 6: Bland-Altman plot comparing mean real and predicted ratings.

6.3 Impact of Pre-training

Based on existing literature, leveraging pre-trained components enables the implementation of independent training strategies, accelerates convergence, improves generalization, and enhances user-item pattern recognition (Zolfagharnasab *et al.* (2024)). With this in mind, the objective of this section is to assess the impact of pre-trained components on the performance of the hybrid model. Given the dataset and model sizes, it is important to note that even slight performance improvements or declines can have a more noticeable impact in larger-scale projects.

To evaluate the impact of pre-training, we established four scenarios: (1) the hybrid model is trained from scratch, with all parameters initialized anew; (2) the NN is trained first, then frozen, followed by training the tower model, where only the independent parameters of the hybrid and FM models are updated; (3) the process is reversed for the FM: the FM is trained and frozen first, then only the independent parameters of the hybrid model and NN are updated; (4) the pre-trained parameters of both the FM and NN are frozen, allowing only the limited parameters of the hybrid model to be tuned during training.

As shown in Figure 7, pre-training notably influences the convergence trajectory of our models, ultimately enhancing their performance. Initially, the loss curves indicate that all models exhibit decreasing loss over time. However, it is noteworthy that the fully pre-trained models (Scenario 4) and the FM (Scenario 3) converge more quickly, starting with lower error rates. This suggests that the FM model contributes more to convergence delays in this architecture compared to the NN model, which leads to a greater loss in system performance. This occurs because the FM model inherently lacks the capability to model multidimensional,

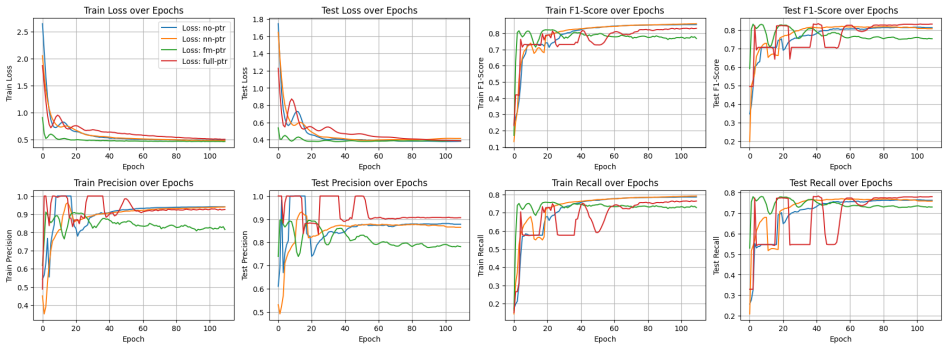


Figure 7: Impact of pre-training hybrid components on overall performance.

nonlinear relationships. Once the FM parameters are frozen (post-initial training), its flexibility to integrate with the NN model is removed, resulting in a more substantial performance loss and increased sensitivity to overfitting in Scenario 3.

The precision and recall graphs further illustrate the superiority of the pre-trained models, particularly the fully pre-trained model (Scenario 4), over the others. This advantage is crucial for large-scale and industrial applications, as it indicates the model's robustness against overfitting observed in the other scenarios. However, fluctuations in this model's performance persist up to epoch 60, as it seeks to balance outputs from the base models. Despite this, the F1 score charts show that the fully pre-trained (Scenario 4) model consistently outperforms the other models, demonstrating that separately pre-training the base models not only mitigates negative impacts on system performance but also slightly enhances performance metrics and increases resistance to overfitting.

6.4 Cold-Start Treatment

The cold-start problem is a key challenge in CF approaches, particularly when encountering new users, items, or contexts without historical interaction data. In such cases, traditional CF methods struggle, as they depend on past user-item interactions for accurate recommendations. To alleviate this issue, models usually incorporate additional contextual data, such as user demographics or item features, allowing the system to infer preferences based on similar user or item characteristics, which enhances recommendation accuracy even in the absence of prior interactions.

Table 8 demonstrates the impact of incorporating feature vectors on enhancing the matrix factorization model's performance. The results show that, although training metrics slightly decrease, the model's generalization on the test set improves significantly. For example, adding feature vectors reduces the test error

Table 8: Impact of Feature Vectors on Model Performance (Training, Testing)

Model	Error	Precision	Recall	F1 Score
Without Feature Vectors	(0.191, 0.834)	(0.837, 0.545)	(0.821, 0.589)	(0.817, 0.555)
With Feature Vectors	(0.520, 0.371)	(0.896, 0.861)	(0.750, 0.761)	(0.813, 0.804)
Difference	(0.329, 0.462)	(0.06, 0.316)	(0.071, 0.172)	(0.004, 0.249)

from 0.834 to 0.371 and boosts test accuracy from 0.545 to 0.861, indicating the model's increased capacity to capture complex interactions and improved generalization. In short, the use of feature vectors raises the F1 score in the test set from 0.555 to 0.804, achieving a 25% improvement in balancing precision and recall. These results confirm that feature vectors enhance the model's understanding of user-item relationships and significantly improve generalization and performance on unseen data.

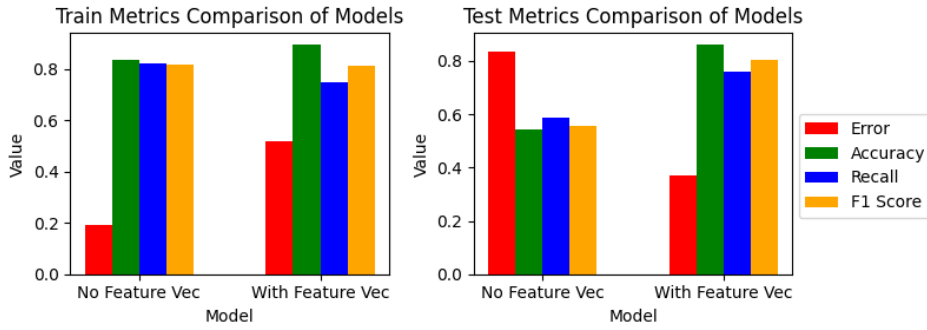
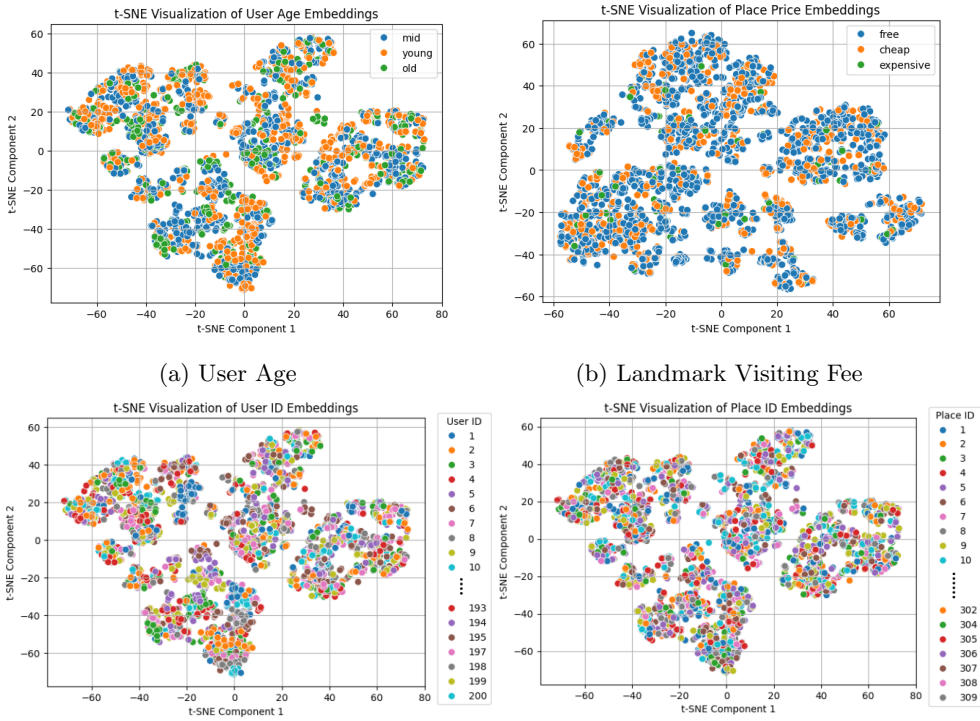


Figure 8: Impact of utilizing additional features (excluding user & landmark IDs).

To better visualize this impact, Figure 8 demonstrates how incorporating additional feature vectors enhances the model's performance. As shown, the addition of supplementary feature vectors notably reduces error while boosting accuracy, recall, and F1 score, indicating improved learning and greater generalizability. Furthermore, minor variations in the training set suggest that the model with feature vectors is capturing deeper patterns within the data.

Having confirmed the effectiveness of supplementary demographic information on the model's performance metrics, the next phase involves analyzing the embedded vectors, which provide valuable insights into user-item relationships within the latent space. To achieve this, the embedded vectors are mapped into a two-dimensional space using the t-distributed stochastic neighbor embedding (t-SNE) technique. Unlike common methods, such as Principal Component Analysis (PCA), t-SNE preserves local structures and relative distances between neighboring points, making it particularly well-suited for capturing non-linear relationships.



(a) User Age

(b) Landmark Visiting Fee

(c) User-ID

(d) Landmark-ID

Figure 9: Models Perception concerning the features.

Regardless of the reduction technique, Figure 9 illustrates how the input data is clustered in the embedding space based on various input features, such as demographic information. Upon observing the reduced feature space, it is evident that the mapping for all inputs is consistent. However, a closer inspection reveals that similar features tend to be clustered together. For instance, data related to users' age shows that young and middle-aged individuals are grouped closely. Similarly, in the cost-related chart, items are clustered into distinct categories of free, low-cost, and high-cost.

Table 9: Assessing RS Recommendations with Sporadic Inputs (Cold-Start)

Input Type	**Input Details**	**1st Rec.**	**2nd Rec.**
Location	Kish Island	Neor Lake	Ashuradeh Island
User Age	Sites Suitable for Seniors	Tabriz Historic Bazaar	Shushtar Hydraulic System
Cost	Low-Cost, High-Rated Sites	Sepahdar School	Besh Gardesh

Notably, the clustering is not limited to auxiliary features and demographic

vectors, as shown in Figure 9, where both user and item (tourist-location) data also exhibit clustering. However, due to the large number of tourists and locations, as well as provincial tourism, distinct labels could not be displayed; instead, color coding was employed. Despite this, the results indicate a clear grouping among users/items, showcasing the model's ability to identify similar instances within these groups. In summary, the embedded vectors are well-structured within the model's latent space, facilitating the differentiation and comparison of new tourists or locations, which effectively addresses the cold-start problem.

By leveraging feature embedding vectors, this study addresses the cold-start problem through four primary steps, given entries of i, where input information can range from age groups, landmark categories, or other relevant features:

1. Map input information into the embedding space utilized by the RC.
2. Compare the similarity between the mapped data and new user/item entries.
3. Rank the information based on its similarity to the input data.
4. Present the ranked information to the user (tourist).

By following these steps, the system can generate effective recommendations with minimal information on new inputs, thereby addressing the cold-start problem. For instance, given a new location similar to Kish Island or an older adult age group, the system can recommend prominent sites and cost-effective, highly-rated locations, as illustrated in Table 9. This concludes the investigations undertaken in this study.

7. Conclusions

This study developed a recommendation system using neural collaborative filtering to enhance tourism experiences in Iran by providing personalized suggestions based on demographic and contextual data. Leveraging travel data from a travel agency, it structured information on user demographics, landmarks, and ratings. The research compared matrix factorization, neural networks, and a hybrid model, using pre-training to improve accuracy and embedding vectors to address the cold-start problem. Key findings include the introduction of a unique dataset for Iranian tourism, a high-performing hybrid model with an F1 score of 0.84 and accuracy of 0.90, and a 17% improvement in recall due to feature integration. Embedding analysis confirmed the effectiveness of cold-start recommendations, while pre-trained sub-models enhanced training efficiency and reduced overfitting. This research establishes a foundation for future culturally specific tourism recommendation systems by offering models, datasets, and baseline evaluations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence this paper.

Acknowledgment

The authors gratefully acknowledge the support of the travel agency for providing the dataset used in this study.

References

Chen, J., Zhang, H., He, X., Nie, L., Liu, W., and Chua, T.-S. (2017), Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–344.

Cheng, H.T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., and Shah, H. (2016), Wide & Deep Learning for recommender systems, *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 7–10.

Ferrari Dacrema, M., Cremonesi, P., and Jannach, D. (2019), Are we really making much progress? A worrying analysis of recent neural recommendation approaches, *Proceedings of the 13th ACM Conference on Recommender Systems*, ACM, RecSys '19, September. http://dx.doi.org/10.1145/3298689.3347058.

Fesenmaier, D.R., Werthner, H., Wober, K., and Dimitrios, B. (2006), Destination recommendation systems: Behavioral foundations and applications, *Annals of Tourism Research,* **33(3)**, 103-122.

Guo, H., Tang, R., Ye, Y., Li, Z., and He, X. (2017), DeepFM: A factorization-machine based neural network for CTR prediction, *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., and Chua, T.-S. (2017), Neural collaborative filtering, *Proceedings of the 26th International Conference on World Wide Web*, 173–182.

Kabassi, K. (2010), Personalizing recommendations for tourists, *Telematics and Informatics*, **27(1)**, 51–66.

Koren, Y. (2009), The BellKor solution to the Netflix grand prize, *Netflix Prize Documentation*, 1–10.

Mnih, A., and Salakhutdinov, R. (2008), Probabilistic matrix factorization, *Proceedings of the 20th Conference on Neural Information Processing Systems*, 1257–1264.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021), Learning transferable visual models from natural language supervision, *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Rendle, S. (2010), Factorization machines, *Proceedings of the 2010 IEEE International Conference on Data Mining*, 995–1000.

Ricci, F., and Del Missier, F. (2011), Supporting Travel Decision Making through Personalized Recommendation, *Information and Communication Technologies in Tourism 2011*, Springer, Vienna, 31-41.

Ricci, F., Rokach, L., and Shapira, B. (2015), Recommender systems: Introduction and challenges, *Recommender Systems Handbook*, Springer, 1–34.

Schafer, J.B., Frankowski, D., Herlocker, J., and Sen, S. (2007), Collaborative filtering recommender systems, *The Adaptive Web*, Springer, 291–324.

Wang X., He, X., Wang, M., Feng, F., and Chua, T.-S. (2019), Neural graph collaborative filtering, *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174.

Zanker, M., Jessenitschnig, M., and Fuchs, M. (2008), Automated Semantic Annotations for Tourism Information Systems Using Collaborative Filtering Approaches, *Information Technology & Tourism,* **10(4)**, 309-325.

Zhou, G., Mou, N., Fan, Y., Pi, Q., Bian, W., Zhou, Z., and Gai, K. (2018), Deep interest network for click-through rate prediction, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1059–1068.

Zolfagharnasab, M.H., Bahrani, M., Hamed Saghayan, M., and Masoumi, F.S. (2024), Exploring a Novel Multi-Channel Structure to Improve Facial Expression Recognition on Occluded Samples Using Deep Convolutional Neural Network, *Journal of Artificial Intelligence, Applications and Innovations*, National Association of Artificial Intelligence of Iran, **1(2)**, 26–41.

Zolfagharnasab, M.H. and Damari, S. (2024), A Comparative Analysis of Machine Learning Models in News Categorization, *U. Porto Journal of Engineering*, **10(3)**, 23–38.

Zolfagharnasab, M.H., Damari, S., Soltani, M., Ng, A., Karbalaeipour, H., Haghdadi, A., Saghayan, M.H., and Matinfar, F. (2025), A novel rule-based expert system for early diagnosis of bipolar and Major Depressive Disorder, *Smart Health*, **35**, 100525.

Zolfagharnasab, M.H., Freitas, N., Gonçalves, T., Bonci, E., Mavioso, C., Cardoso, M.J., Oliveira, H.P., and Cardoso, J.S. (2024), Predicting Aesthetic Outcomes in Breast Cancer Surgery: A Multimodal Retrieval Approach, *Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care*, Springer, 137–147.

Appendix: Statistical Analysis

We begin our dataset analysis by assessing the tourist landmark CSV, which contains information on 301 tourist attractions in Iran. This table includes details about the attractions, their categories, entrance ticket prices, and the cities and provinces where they are located, offering valuable statistical insights. Figure 10 shows the distribution of attractions by category, with natural and cultural sites being the most numerous, totaling around 100. Historical and architectural landmarks follow, while recreational parks are less common, likely because travel agencies prioritize natural and historical sites over parks.

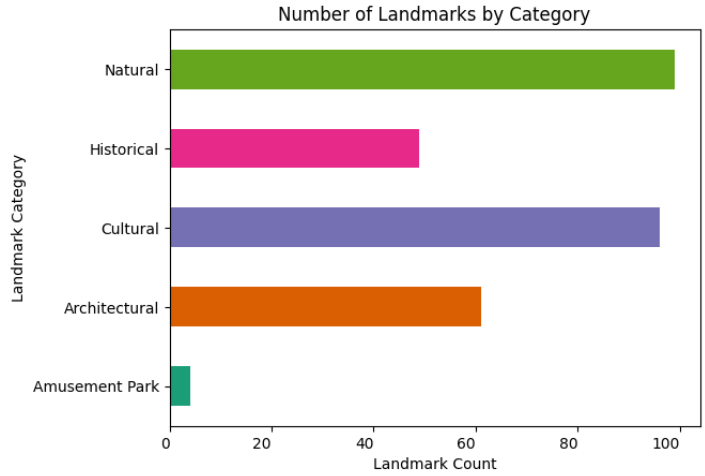


Figure 10: Number of tourist spots per category in the dataset.

Next, we analyze the distribution of tourist attractions by province, as shown in Figure 11. Historically significant provinces like Fars, Isfahan, and those known for natural beauty, like Gilan, have the most attractions. In contrast, provinces like Alborz and Markazi have fewer tourist spots.

Following the landmark CSV, we examined tourist information, including age and residence, from the user information CSV. Figure 13 shows the distribution of tourists by province. Tehran residents take the most trips, followed by those from industrial provinces like East Azerbaijan, Isfahan, and Fars. This trend is unsurprising, given factors such as income levels, population size, and lifestyle. Less affluent areas and provinces with international borders report fewer trips, possibly due to lower domestic travel interest.

Since the rating CSV contains only the tourist ID, attraction ID, and the rating given by the tourist, the data alone offers limited interpretability. However, as shown in Figure 12, a rating of 3 is overwhelmingly the most frequent, suggest-

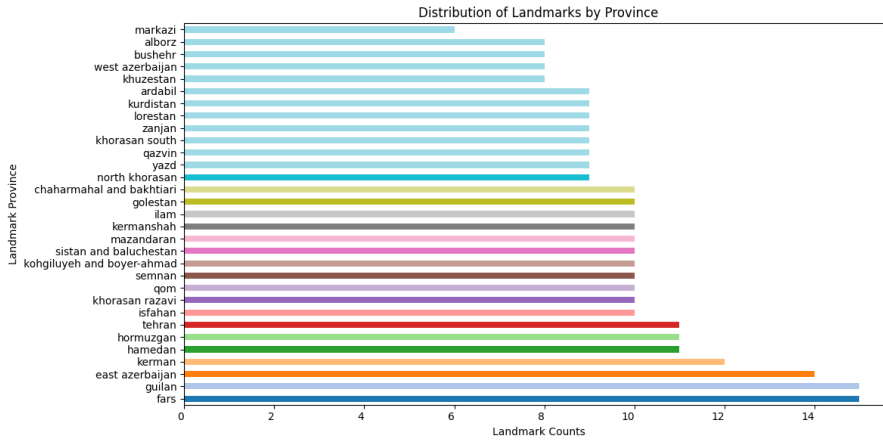


Figure 11: Distribution of tourist spots by province.

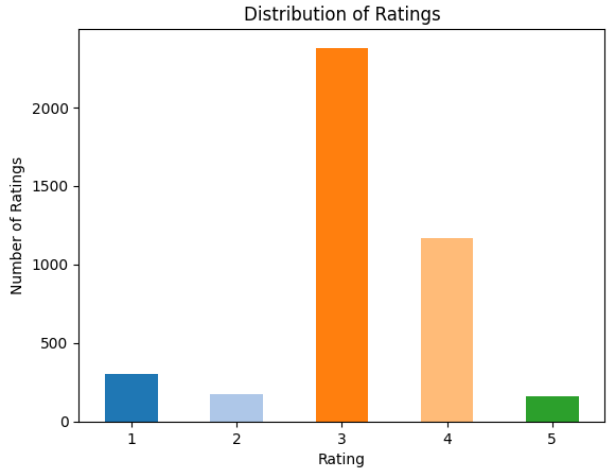


Figure 12: Distribution of tourist ratings.

ing that many tourists were moderately satisfied with their visits. A rating of 3 typically corresponds to "neutral" or "average," which is likely why a significant number of users opted for this middle score. The second most common rating is 4, reflecting "fairly good" satisfaction, indicating a generally positive experience among tourists. Fewer tourists expressed extreme satisfaction (rating 5) or slight dissatisfaction (rating 2), while the number of tourists who were entirely dissatisfied (rating 1) is relatively small. It is also notable that dissatisfied tourists tend to give the lowest rating (1) rather than choosing a 2. Similarly, only a small percentage of tourists awarded the highest score (5), likely due to negative secondary

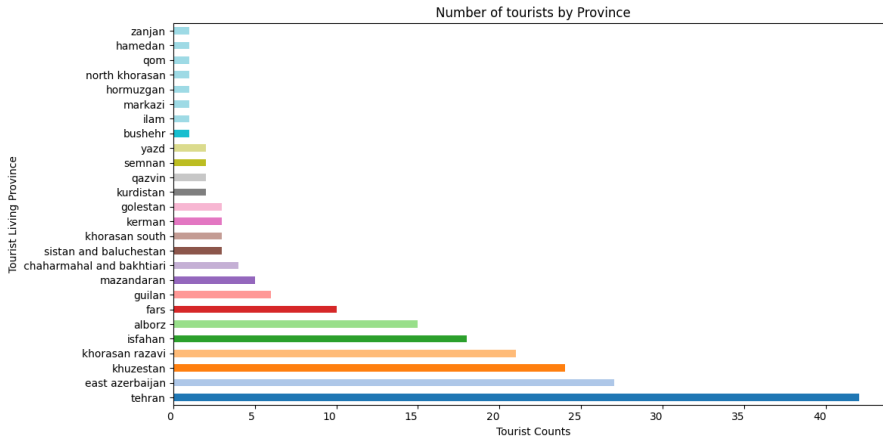


Figure 13: Tourist distribution by place of residence.

factors during their travels.

It is also worth noting that the histogram in Figure 12 closely resembles a normal (Gaussian) distribution, commonly seen in data derived from natural interactions or complex systems. This suggests that the dataset follows a statistically valid distribution, making it suitable for analysis with normal models.

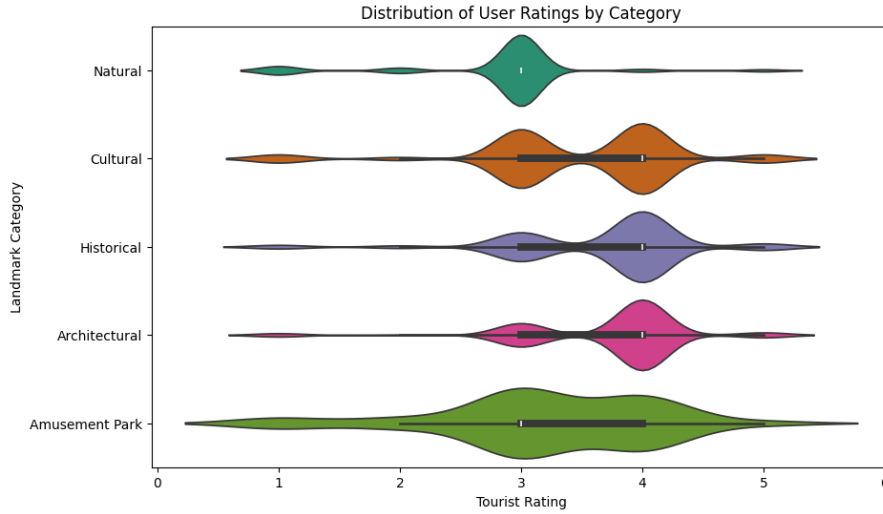


Figure 14: Distribution of tourist ratings by category.

After analyzing each table individually, a combined statistical analysis of all three datasets was conducted. The analysis begins by examining the distribution of user ratings across different categories of tourist attractions, as represented by

violin plots in Figure 14. As depicted, categories such as cultural, historical, and architectural attractions tend to have higher average ratings compared to natural attractions. While natural attractions have the lowest average rating, their distribution forms a single peak centered around a rating of 3, indicating consistent user feedback for this category. This consistency may suggest that services at natural attractions are generally weaker, especially in open environments. In contrast, architectural and historical attractions show narrower distributions, indicating more uniform ratings. Furthermore, the broader distribution for recreational centers, which have only four tourist spots, may appear larger due to the small sample size and should not be mistaken for a typical statistical distribution.

Next, an important question regarding tourists' preferences in relation to their age was assessed, as shown in Figure 15. As expected, categories such as historical, cultural, and architectural attractions tend to attract older tourists, particularly those over the age of 50. In contrast, tourists visiting natural attractions are generally younger, with the majority falling within the 30 to 45 age range, suggesting that younger individuals and families prefer these destinations. While natural attractions display a broader age range, recreational parks show the lowest average age. However, conclusions about this category are limited due to the small sample size. Overall, the presence of tourists from a wide range of age groups helps minimize potential bias in training the recommendation system.

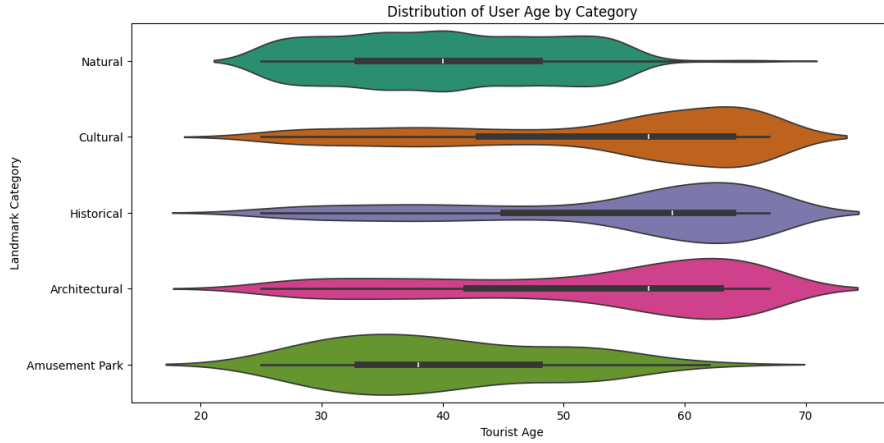


Figure 15: Tourist age distribution by category of tourist attractions.

To determine the most and least popular tourist destinations, a thorough statistical analysis of tourist ratings across attractions was conducted. The findings indicated that prominent sites, such as the Tomb of Ferdowsi, Urmia Bazaar, and Nasir al-Mulk Mosque, were the most visited and highly rated, receiving ratings approximately 0.83 points above the overall average of 3.21. In contrast, lesser-

known sites like the Green Dome and Falak-ol-Aflak Castle, with only a few ratings, were among the least popular attractions, scoring 0.84 points below the average. However, due to the limited number of ratings for these lesser-known locations, it is difficult to draw reliable conclusions about their true popularity.

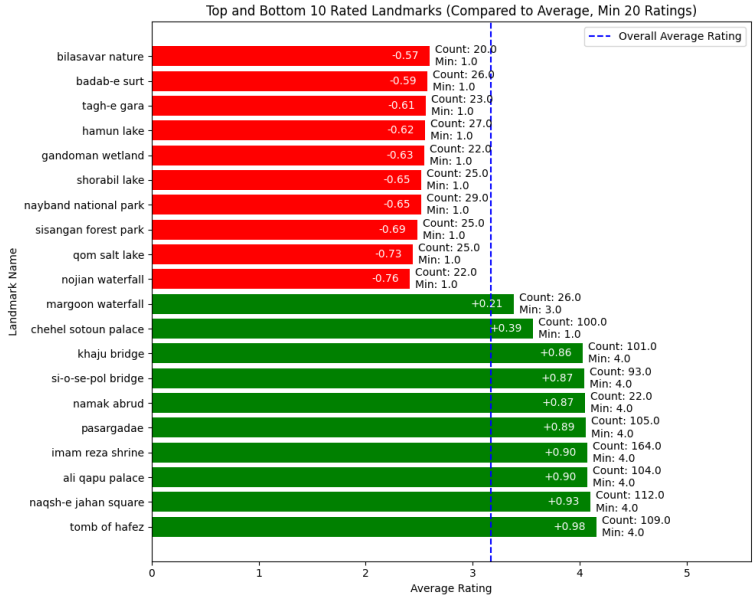


Figure 16: Distribution of tourist ratings for attractions with at least 20 ratings.

In light of this, Figure 16 emphasizes the importance of focusing only on tourist attractions that have received at least 20 ratings when compiling lists of popular and unpopular locations. This approach ensures that the preferences of a broader audience are considered, minimizing the impact of individual subjective opinions. As shown in the figure, highly frequented sites such as the Imam Reza Shrine, Ali Qapu Palace, the Tomb of Hafez, Pasargadae, and the scenic area of Namak Abrood stand out as some of the most popular destinations. In contrast, less visited locations like Takht-e-Garah, Lake Namak Qom, and the natural area of Bilesavar, each with approximately 20 ratings, are identified as the least popular, scoring 0.75 points below the overall average. This underscores the importance of considering both the volume of ratings and the average scores when evaluating tourist destinations.

The final analysis focuses on identifying the most and least popular provinces, as illustrated in Figure 17. This figure presents the average rating for each province along with the number of ratings received. The provinces of Isfahan (529 ratings), Fars (383 ratings), and Razavi Khorasan (218 ratings) emerge as the most popular

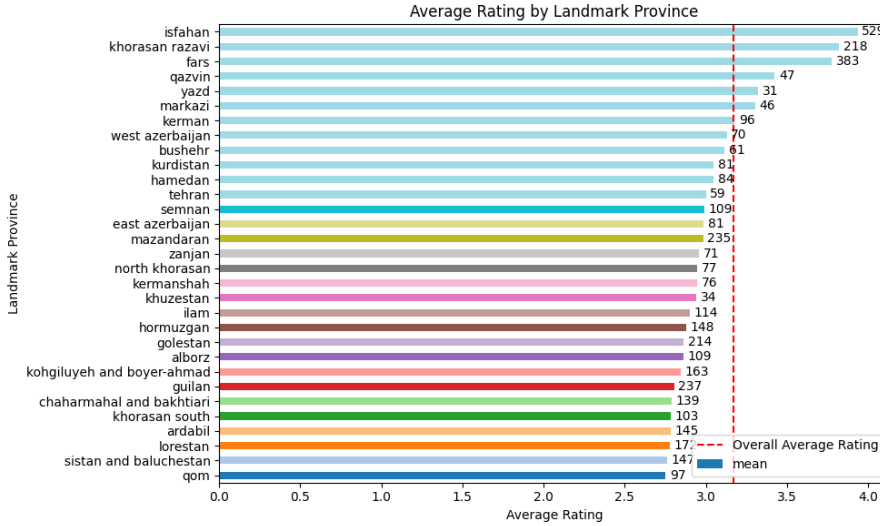


Figure 17: Distribution of tourist ratings by province.

tourist destinations. This is likely due to the religious significance of the Imam Reza Shrine in Razavi Khorasan and the abundance of historical and architectural landmarks in Isfahan and Fars. Notably, despite attracting a significant number of tourists, Tehran does not rank among the top-rated provinces, which aligns with expectations. Provinces known for natural attractions, such as Gilan and Golestan, also do not score highly, potentially due to lower service quality or tourists' preference for agency-organized tours. On average, most provinces do not significantly deviate from the overall mean rating, indicating a relative uniformity in tourist experiences across Iran's provinces. This uniformity presents a challenge for the recommendation system, which must enhance correlations between input features to effectively differentiate tourist attractions and provinces.