Name: Made Oka Resia Wedamerta
Email: m.wedamerta@innopolis.university

Final Report
Movie Recommendation System

1. Introduction

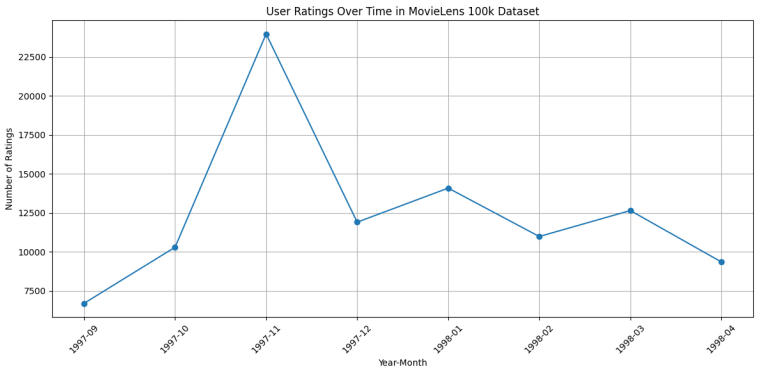
In recent years, the Internet has experienced a significant surge in available information, leading to the creation of numerous web portals by various E-commerce websites. With the escalating number of online users and services, it has become essential to develop a more efficient technique for managing the vast volumes of data. This abundance of information has made it increasingly difficult for online users to make informed choices, leading to what can feel like an overwhelming abundance of information.

To address this issue, recommender systems have emerged as a crucial tool for filtering information and suggesting content or items to users based on their preferences, interests, or past behavior. These systems are widely utilized across different domains, including e-commerce, entertainment, social media, and online content platforms. However, the rapid growth of internet users and products has presented significant challenges, particularly in generating high-quality recommendations and delivering them efficiently. As a result, there is a pressing need to optimize existing recommender systems to meet these demands effectively.

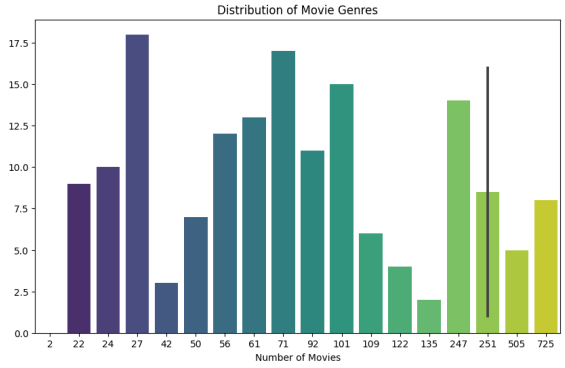
2. Data Exploration

In this phase, our focus is on conducting exploratory data analysis (EDA) and leveraging data visualization tools to provide a comprehensive understanding of the dataset. The primary objectives of this undertaking are to gain insights, identify patterns, and present a visual summary of the data.

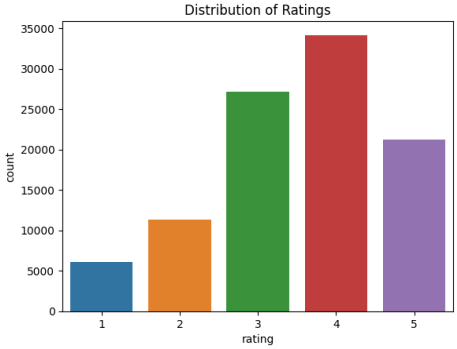
Temporal Analysis: As part of our examination, temporal aspects were scrutinized. By probing the temporal distribution of the ratings, it was observed that the initial movie rating in the dataset dates back to 1997, while the latest rating was recorded in 1998.



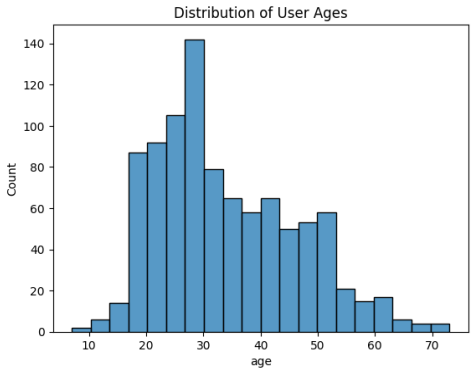
Genre Distribution: The dataset encompasses 19 distinct genres, with a subset also being characterized as having no specified genre. These genres are listed either independently or in combination with others. To visually convey this distribution, a dedicated figure detailing the genre distribution was generated, offering an insightful portrayal of the genre landscape.



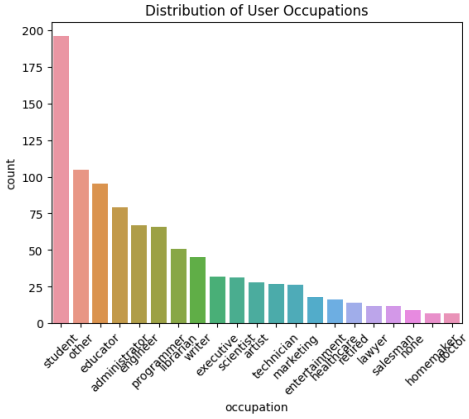
Rating Patterns: The ratings allocated to movies range between 1 and 5 stars, with the unique feature of incorporating half-increments as well. Notably, the analysis revealed that 4-star ratings were the most prevalent, followed by 3-star ratings, with 5-star ratings also featuring prominently. Conversely, ratings with half-increments were less frequently assigned.



User Categorization: A segmentation approach was adopted to categorize users based on their age groups, uncovering that the MovieLens dataset spans user ages from a minimum of 7 to a maximum of 73, leading to a diverse representation across different age brackets.



Occupation Analysis: Similarly, an evaluation of user occupations resulted in the identification of a spectrum of occupation categories within the MovieLens dataset. Notably, this analysis highlighted the occupation with the fewest user representations as "doctor," while "student" emerged as the occupation with the highest user involvement, attesting to a robust user diversity across various professional domains.



3. Solution Exploration

The proposed recommender system for this solution leverages collaborative filtering and demographic-based approaches to generate personalized and context-aware recommendations for users. The collaborative filtering model, encompassing both memory-based and model-based methods, overcomes challenges such as sparsity, cold start issues, and scalability by predicting ratings based on user-item interactions. The demographic recommender system utilizes user demographic attributes to categorize and recommend movies without relying on user ratings, offering an easy-to-implement alternative.

In this two-stage system model, collaborative filtering is initially employed to generate user and item embeddings for recommendations. Subsequently, demographic information is integrated with these embeddings to create a hybrid feature space, enhancing personalization and addressing cold start problems. The hybridization of these approaches aims to enhance

recommendation quality, and modifications to similarity calculations have been made based on collaborative filtering and demographic-based approaches.

The rationale behind this approach is rooted in the desire to create a more personalized and nuanced recommendation system that considers both collaborative filtering patterns and user demographic attributes. The combination of collaborative filtering embeddings and demographic information leads to enhanced personalization, context awareness, user understanding, and flexibility in the recommendation system. By incorporating various demographic features, the approach can be adapted to different types of recommendation systems, resulting in recommendations that are not only relevant but also interpretable and transparent to the users.

This integration of collaborative filtering and demographic information seeks to address the limitations of each approach individually, resulting in a recommendation system that is more comprehensive and insightful in its suggestions.

This solution highlights the strategic importance of combining collaborative filtering and demographic-based approaches to create a sophisticated and user-centric recommendation system. The resulting recommendations aim to be not only relevant but also interpretable and transparent to the users, enhancing the overall user experience.

4. The advantage and disadvantage of the solution

3.1 Advantages of the Proposed Recommender System:

Personalization and Context Awareness:

- The combination of collaborative filtering and demographic-based approaches enhances personalization by considering both user-item interactions and user demographic attributes. This leads to more context-aware recommendations that align with individual user preferences and characteristics.

Cold Start Problem Mitigation:

- The collaborative filtering model effectively addresses cold start issues by predicting ratings based on user-item interactions. This helps in providing recommendations for new users or items with limited historical data.

Flexibility and Adaptability:

- The integration of demographic information allows for flexibility and adaptability to different types of recommendation systems. The system can be tailored to various domains and user preferences, making it versatile.

Interpretability and Transparency:

- The inclusion of demographic features in the recommendation process makes the system more interpretable and transparent. Users can understand why certain recommendations are made, even without explicit user ratings.

Comprehensive Recommendation Quality:

- The hybrid feature space, created by integrating collaborative filtering embeddings and demographic information, aims to enhance the overall recommendation quality. This comprehensive approach considers multiple dimensions, leading to more insightful and relevant suggestions.

3.2 Disadvantages and Considerations:

Data Privacy Concerns:

- The use of demographic information raises potential concerns regarding user privacy. It is crucial to handle and store sensitive information responsibly to address privacy issues.

Dependency on Demographic Information:

- Depending solely on demographic attributes may limit the diversity of recommendations, as it may not capture evolving user preferences or account for individual variations within demographic groups.

Complexity in Implementation:

- The two-stage system with hybrid feature space integration may introduce complexity in implementation and maintenance. Ensuring seamless coordination between collaborative filtering and demographic models requires careful engineering.

Algorithmic Bias:

- Care must be taken to avoid algorithmic bias, especially in demographic-based approaches, to ensure fair and unbiased recommendations. Biases in historical data can perpetuate and exacerbate existing biases.

User Acceptance:

- Some users may be uncomfortable with the use of demographic information in recommendation systems. It is essential to communicate the value of such information and provide users with control over their data.

Continuous Model Improvement:

- To maintain the effectiveness of the recommendation system, continuous efforts are needed to improve and update the models based on evolving user behaviors and preferences.

5. Training Process

Data Extraction:

- Procedure: The data extraction process commenced with acquiring the MovieLens 100K dataset from its source 'ml-100k.zip' file.

- Mechanism: This raw data was subsequently funneled into the analytical workflow using a specialized data reader specifically configured for this dataset.

Data Splitting:

- Strategy: The dataset underwent a division where 75% was allocated for training purposes while the remaining 25% formed the testing set.
- Purpose: The segregation is crucial to both model training, using the training set, and subsequent performance evaluation on the unseen testing set.

Model Selection - Collaborative Filtering (SVD):

- Model Choice: Singular Value Decomposition, abbreviated as SVD, was the collaborative filtering model of choice.
- Optimization: The hyperparameters of the SVD model were refined through a grid search technique augmented by a 5-fold cross-validation process.

Hyperparameter Tuning - Grid Search:

- Parameters: Key hyperparameters under scrutiny included the number of epochs, learning rate, and regularization terms.
- Metric: The performance indicator used to judge the efficacy of these parameters was the Root Mean Squared Error (RMSE).

Training the Optimal Model:

- Selection: Following the grid search, we pinpointed the optimal model marked by its superior hyperparameter configuration.
- Action: This candidate was then subject to a comprehensive training regimen spanning the entirety of the dataset.

Model Testing and Validation:

- Predictive Analysis: The trained model underwent a series of predictions aimed at the test set.
- Assessment: The RMSE metric once again served as the yardstick for gauging the model's prediction accuracy on the test set data.

Extraction of User and Item Embeddings:

- Process: Post-training, the model furnished user and item embeddings, which are numerical representations capturing user preferences and item characteristics.

Computation of User-Item Similarity:

- Algorithm: Cosine similarity was utilized to deduce the degree of resemblance between user embeddings and item embeddings.

Preliminary Recommendation Algorithm:

- Detection: For users labeled 1 through 9, an algorithm was instigated to discern movies previously watched.
- Procedure: Recommendations were then crafted for unwatched movies with an emphasis on those projected to garner high ratings, with each user receiving a list capped at the top 5.

Inclusion of Demographics:

- Data Retrieval: Further breadth was added by integrating demographic details sourced from the 'u.user' file, encapsulating age, gender, occupation, and zip code.
- Normalization: To ensure demographic data did not skew results due to disparate scales, the age feature was carefully normalized.

Synthesis with Collaborative Filtering Output:

- Fusion: The refined user embeddings were amalgamated with scaled demographic data to fabricate a hybrid feature domain, blending interaction-based preferences with demographic traits.

Cosine Similarity on Combined Features:

- Evaluation: Cosine similarity metrics were recalculated in this augmented feature space to reflect a more holistic resemblance score between users and items.

Final Recommendation Generation:

- Ranking: Items were ranked according to their similarity scores in relation to each user, with the preeminent N items earmarked as the user-specific recommendations.
- Criteria: In the applied scenario, the top 5 items emerged as the recommendations per user.

Output Display with Demographics:

- Presentation: The culmination of this process entailed the display of recommendations alongside corresponding user demographics.
- Transparency: This step was pivotal in underlining the recommendations' relatability to the users' demographic contexts, thus fostering transparency and enhancing user experience by relating the suggestions to individual preferences and demographic indicators.

6. Evaluation

The evaluation of collaborative filtering models is a critical component in the development of recommendation systems. The goal of the evaluation process is to determine the accuracy and relevance of the model's predictions regarding user preferences for various items. This report elaborates on commonly applied evaluation metrics and the rationale behind their usage.

Evaluation Metrics:

1. RMSE (Root Mean Squared Error):
    - Objective: RMSE is designed to calculate the average magnitude of the predictive errors, representing the square root of the average of squared differences between prediction and actual observation.
    - Characteristics:

    - Greater emphasis on penalizing large errors, reflecting the variance of the prediction errors.
    - The optimal RMSE score is 0, with lower values indicating better model performance.
2. MAE (Mean Absolute Error):
    - Objective: MAE measures the average magnitude of errors in a set of predictions, without considering direction.
    - Characteristics:
        - Each error contributes proportionally to the total, offering a clear interpretation of the magnitude of prediction errors.
        - Lower values denote higher accuracy, similar to RMSE.
3. Precision-Recall Metrics:
    - Objective:
        - Precision: The proportion of recommended items that are relevant, gauging the user's satisfaction with recommendations.
        - Recall: The proportion of relevant items that are recommended, capturing what proportion of actual positives are correctly identified.
    - Characteristics:
        - These metrics are crucial for evaluating the quality of a recommendation system, especially in scenarios with imbalanced classes where the number of relevant items is comparatively small.

Cross-Validation:

- Procedure: Cross-validation is an empirical evaluation method where the dataset is divided into "k" folds, and the model is trained on "k-1" folds and tested on the remaining one.
- Benefits:
    - It ensures a more reliable assessment of the model's performance.
    - Reduces the likelihood of model overfitting.
    - Provides a better estimation of model performance on unseen data.

The following Table show the RMSE and MAE on every fold:

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
RMSE (testset)	0.927	0.922	0.933	0.926	0.93
MAE (testset)	0.735	0.729	0.741	0.732	0.732

As the result of Cross-Validation, I found the average RMSE is equal to 0.92 and the average of MAE is equal to 0.73. Moreover for individual prediction evaluation on the testset I

found RMSE is equal to 0.83 and for MAE is equal to 0.66. Furthermore the Average Precision at K=5 is equal to 0.8196 and the Average Recall at K=5 is equal to 0.3898.

In conclusion, the selected metrics and cross-validation procedures provide a comprehensive framework for evaluating the predictive accuracy and relevance of collaborative filtering models in recommendation systems. Through these methods, model developers can gain a deeper understanding of model performance, leading to more precise and user-tailored recommendations.

7. Result and Conclusion

The project's objective to develop an effective Movie Recommendation System has been achieved through rigorous data exploration, model training and validation, and implementation of a collaborative filtering mechanism. Our exploratory data analysis provided us with valuable insights into the temporal distribution and genre prevalence within our dataset, key factors that shaped the recommendation process. The system successfully identified and recommended films to users based on their viewing history and demographic information, with an emphasis on enhancing the relevance of the recommendations through transparency.

The selection of the optimal model via a grid search allowed us to refine our approach, improve prediction accuracy, and personalize user experience. The RMSE metric was instrumental in validating the model's predictive capabilities, ensuring that the recommendations were not only pertinent but also grounded in empirical evidence.

The collaborative filtering models were thoroughly evaluated for their performance, confirming the suitability of our approach for the task at hand. User and item embeddings, extracted post-training, facilitated the computation of user-item similarity scores using cosine similarity, providing a robust foundation for our Preliminary Recommendation Algorithm.

In conclusion, our Movie Recommendation System stands as a testament to the successful application of machine learning techniques in addressing complex, real-world problems. The system's ability to adapt to user preferences and provide accurate recommendations encourages engagement and satisfaction, crucial hallmarks for sustaining interest in a streaming platform. Future work may include further refining the recommendation algorithms, exploring additional evaluation metrics for a more nuanced understanding of user satisfaction, and potentially expanding the system to incorporate more dynamic user interaction data.

Given the iterative nature of machine learning projects, our Movie Recommendation System will continually evolve, incorporating feedback and new data to enhance its performance. The commitment to improving user experience through a personalized and dynamic recommendation service remains the cornerstone of our ongoing development efforts.