Comparative Analysis of Four Recommendation System Methods: LightGCN, BiVAE, NeuMF, and GMF

V. Nithin Reddy
Roll No: SE22UCSE278
Department of Computer Science and Engineering
Team Id: 5555
Mahindra University, Hyderabad, India
Email: se22ucse278@mahindrauniversity.edu.in

***Abstract*—Recommender systems are pivotal in curating personalized user experiences, yet their performance varies significantly across architectures and scenarios. This paper presents a comprehensive comparative analysis of four recommendation system approaches: LightGCN (Light Graph Convolutional Network), BiVAE (Bilateral Variational Autoencoder), NeuMF (Neural Matrix Factorization), and GMF (Generalized Matrix Factorization). I implement and evaluate these models using the MovieLens 100K dataset, focusing on recommendation accuracy, computational efficiency, recommendation diversity, popularity bias, and cold start performance.**

Through rigorous experimentation with different hyperparameter configurations, I reveal significant differences in model behavior and performance. BiVAE demonstrates exceptional cold start handling and recommendation diversity, while LightGCN achieves superior accuracy for established users and items but struggles with cold start scenarios. NeuMF offers balanced but computationally expensive performance, and GMF provides a simple, efficient baseline with limited expressiveness.

I. Introduction

Recommendation systems have become essential components of online platforms, helping users navigate vast content collections by providing personalized suggestions. These systems are challenged by the cold start problem, which refers to scenarios where recommendations must be made for new users with limited interaction history or for new items with few user engagements.

This paper evaluates four distinct approaches to recommendation systems:

1) **LightGCN**: A graph-based model that simplifies traditional Graph Convolutional Networks (GCNs) for recommendation tasks by removing feature transformation and nonlinear activation, focusing solely on neighborhood aggregation in the user-item interaction graph.
2) **BiVAE**: A generative model based on variational autoencoders that employs a probabilistic approach to model user-item interactions, capturing distributions rather than point estimates.
3) **NeuMF**: A neural network-based approach that combines GMF with Multi-Layer Perceptron (MLP) in a hybrid architecture, allowing for both linear and nonlinear modeling of user-item interactions.
4) **GMF**: A discriminative approach with minimal complexity that generalizes traditional matrix factorization within a neural network framework.

I implement each model using the MovieLens 100K dataset and evaluate their performance across standard recommendation metrics and the cold start problem. My analysis goes beyond performance numbers to discuss practical considerations including training efficiency, model complexity, and recommendation diversity.

II. Background and Problem Definition

A. Recommendation System Fundamentals

The core task of recommendation systems is to predict users' preferences for items they have not yet interacted with. Formally, given a set of users $U = \{u_1, u_2, ..., u_m\}$ and a set of items $I = \{i_1, i_2, ..., i_n\}$, along with observed interactions $R = \{r_{ui} | u \in U, i \in I\}$, the goal is to predict unobserved interactions $\hat{r}_{ui}$ for user-item pairs without interactions.

B. The Cold Start Problem

The cold start problem represents one of the most challenging aspects of recommendation systems. It occurs in two primary scenarios:

1) **User Cold Start**: When new users join the system with very few or no interactions, making it difficult to understand their preferences.
2) **Item Cold Start**: When new items are added to the system with very few or no interactions, making it difficult to recommend them to appropriate users.

Addressing the cold start problem requires models that can effectively leverage limited interaction data and potentially incorporate auxiliary information.

C. Evaluation Metrics

To evaluate the performance of recommendation models, I use four standard metrics:

1) **Mean Average Precision (MAP)**: Measures the ranking quality of recommended items, focusing on the overall precision of the ranked list.
2) **Normalized Discounted Cumulative Gain (NDCG)**: Evaluates the ranking quality considering the position of relevant items, with higher positions contributing more to the score.
3) **Precision**: Measures the relevance of recommended items, calculated as the ratio of relevant recommended items to the total number of recommended items.
4) **Recall**: Measures the coverage of relevant items, calculated as the ratio of relevant recommended items to the total number of relevant items.

III. Model Architectures

A. Light Graph Convolutional Network (LightGCN)

LightGCN is a graph-based approach that models the user-item interaction matrix as a bipartite graph. It simplifies traditional GCNs by removing feature transformation and non-linear activation, focusing exclusively on neighborhood aggregation.

Key features of LightGCN include:

- Initial user and item embeddings
- Layer-wise propagation that updates embeddings based on neighborhood information
- Layer combination mechanism that integrates representations from different propagation steps

LightGCN typically optimizes Bayesian Personalized Ranking (BPR) loss or binary cross-entropy (BCE) loss with L2 regularization. The graph structure provides additional implicit regularization, contributing to robust performance even with minimal explicit regularization.

B. Bilateral Variational Autoencoder (BiVAE)

BiVAE employs a generative approach to recommendation by modeling the data distribution of user-item interactions. It consists of two variational autoencoders: one for users and one for items.

The key components of BiVAE include:

- Encoder networks that map user and item vectors to latent distributions
- A latent space where users and items are represented
- Decoder networks that reconstruct user-item interactions
- A variational inference framework that optimizes the Evidence Lower Bound (ELBO)

The probabilistic nature of BiVAE allows it to model uncertainty in user preferences and capture complex interaction patterns. The ELBO objective comprises a reconstruction loss term and a Kullback-Leibler divergence regularization term, providing an elegant theoretical framework for balancing reconstruction accuracy against overfitting.

C. Neural Matrix Factorization (NeuMF)

NeuMF extends GMF by incorporating a multi-layer perceptron component alongside the matrix factorization pathway. This hybrid architecture enables the capture of both linear and non-linear interaction patterns.

Key features of NeuMF include:

- User and item embedding layers
- GMF path that models linear interactions
- MLP path that models non-linear interactions
- Fusion layer that combines the outputs of both paths

NeuMF employs point-wise (BCE) or pair-wise ranking losses with L2 regularization on model parameters. It additionally incorporates dropout in its MLP component as a form of regularization.

D. Generalized Matrix Factorization (GMF)

GMF represents a discriminative approach with minimal complexity. It generalizes traditional matrix factorization within a neural network framework, modeling user-item interactions through element-wise products of latent factors.

Key features of GMF include:

- User and item embedding layers
- Element-wise product of embeddings
- Output layer that predicts user-item interactions

GMF offers computational efficiency due to its simple architecture but may lack the expressive capacity to model complex user-item interactions.

IV. Experimental Setup

A. Model Configurations

For my comparative analysis, I configured each model with specific hyperparameters:

LightGCN Configuration:

- Network Architecture: 3 layers
- Learning Rate: 0.005
- Weight Decay: 0.0001
- Regularization Parameter: 0.0001
- Top-K Items: 10
- Evaluation Frequency: Every 5 epochs

BiVAE Configuration:

- Latent Dimensions: 50
- Encoder Structure: [100]
- Activation Function: tanh
- Likelihood Model: Poisson
- Learning Rate: 0.005

NeuMF Configuration:

- Number of Factors: 4
- Layer Sizes: [16, 8, 4]
- Learning Rate: 0.005

GMF Configuration:

- Number of Factors: 4
- Learning Rate: 0.005

B. Cold Start Analysis Methodology

I employed two distinct approaches to analyze the cold start problem systematically:

Approach 1: Interaction Bucket Analysis: To systematically analyze the cold start problem, I categorized both users and items into distinct buckets based on their interaction counts:

- User buckets: 1-28, 29-49, 50-92, 93-170, and 171+ interactions
- Item buckets: 1-4, 5-15, 16-42, 43-99, and 100+ interactions

This bucketing approach allowed me to systematically analyze how model performance varies with different levels of interaction sparsity, providing insights into both user and item cold start scenarios. The cold start analysis was done on models with configurations of 50 epochs and 1024 batch size.

Approach 2: Binary Cold Start Classification: In this approach, I defined cold-start users specifically as those having 25 or fewer interactions in the training data. This binary classification (cold-start vs. warm-start) provides a clear threshold for evaluating how well different recommender models perform when encountering users with very limited history.

For this analysis, I used:

- Training Data: Contains user-item interaction data used to train the models
- Test Data: Contains user-item interactions used for evaluation
- Top-K Recommendations: Contains the top-K recommendations made by each model for each user

Performance was evaluated using Recall@10 and NDCG@10 metrics, calculated separately for both cold-start and warm-start users to clearly highlight the performance gap.

V. Experimental Results

A. Overall Performance Metrics

I conducted a comprehensive evaluation of all four recommendation models across multiple configurations to understand both their performance characteristics and efficiency trade-offs.

TABLE I: LightGCN Performance Across Configurations

Epochs	Batch Size	MAP	NDCG	Prec.	Train(s)	Pred(s)
50	1024	0.140	0.458	0.400	127.12	0.343
50	256	0.130	0.434	0.380	361.85	0.096
100	1024	0.130	0.437	0.386	233.89	0.331
100	256	0.115	0.393	0.347	718.98	0.360

TABLE II: BiVAE Performance Across Configurations

Epochs	Batch Size	MAP	NDCG	Prec.	Train(s)	Pred (s)
50	256	0.099	0.232	0.198	20.83	2.02
50	1024	0.152	0.365	0.325	11.72	1.81
100	256	0.174	0.410	0.358	25.79	1.80
100	1024	0.168	0.401	0.349	24.15	1.79
500	128	0.187	0.439	0.382	155.48	1.78

TABLE III: NeuMF Performance Across Configurations

Epochs	Batch Size	MAP	NDCG	Prec.	Train(s)	Pred(s)
50	256	0.101	0.197	0.178	309.0	3.64
50	1024	0.100	0.196	0.177	216.0	2.84
100	256	0.095	0.189	0.170	625.4	3.02
100	1024	0.096	0.191	0.172	438.2	2.97

TABLE IV: GMF Performance Across Configurations

Epochs	Batch Size	MAP	NDCG	Prec.	Train(s)	Pred(s)
50	256	0.115	0.213	0.191	254.1	44.04
50	1024	0.117	0.215	0.192	189.3	49.27
100	256	0.115	0.211	0.187	508.0	48.63
100	1024	0.119	0.218	0.196	380.7	0.51

B. Performance Analysis and Insights

A thorough examination of these results reveals several insights about the performance characteristics of the four recommendation models:

1) Accuracy Metrics:

- **LightGCN** consistently achieves the highest NDCG (up to 0.458) and Precision (up to 0.399) among all models, demonstrating its superior ability to rank relevant items highly. The performance advantage is most pronounced in the 50-epoch, 1024 batch size configuration, suggesting that LightGCN efficiently captures user-item interaction patterns without requiring extensive training.
- **BiVAE** shows remarkable improvement with increased training epochs, reaching peak performance with 500 epochs (NDCG: 0.439, Precision: 0.382, MAP: 0.187). This pattern indicates BiVAE's capacity to continually refine its probabilistic representation with more training, unlike other models that may plateau or overfit.
- **GMF** exhibits moderate performance across configurations with its best configuration (50 epochs, 1024 batch size) achieving NDCG of 0.215 and Precision of 0.192, substantially lower than LightGCN and BiVAE but higher than NeuMF.
- **NeuMF** consistently underperforms compared to other models, with its highest NDCG at 0.191 and Precision at 0.177. Interestingly, increasing training epochs from 50 to 100 actually decreased performance slightly, indicating possible overfitting due to its limited expressiveness.

2) Computational Efficiency:

- **BiVAE** demonstrates exceptional computational efficiency, with training times as low as 11.72 seconds (50 epochs, 1024 batch size) and consistently fast prediction times around 1.8 seconds. Even its most intensive configuration (500 epochs) completes training in 155.48 seconds, comparable to LightGCN's simpler configurations.
- **LightGCN** shows moderate training efficiency that varies significantly with batch size. The 256 batch size configurations require 2.8-3.1× longer training time compared to the 1024 batch size configurations with the same number of epochs. However, prediction time remains consistently fast (¡ 0.36 seconds).

- **NeuMF** offers relatively efficient prediction times (around 2.8-3.6 seconds) but with training times 18-24× longer than BiVAE's most efficient configuration.
- **GMF** exhibits the poorest efficiency in prediction time which reaches nearly 50 seconds in some configurations. This represents a 25-30× slower prediction time compared to BiVAE and LightGCN, making it potentially problematic for real-time recommendation scenarios, despite having training times comparable to NeuMF.

C. Recommendation Diversity

The diversity metrics reveal substantial differences in the exploration capabilities of the four models:

TABLE V: Recommendation Diversity Metrics

Model Config	Avg. Popularity	Unique Items
LightGCN 50e, 1024b	137.14	545
LightGCN 50e, 256b	121.16	696
LightGCN 100e, 1024b	121.83	697
LightGCN 100e, 256b	116.24	753
BiVAE All configs	45.68	1642
NeuMF 50e, 1024b	110.94	786
NeuMF 50e, 256b	103.70	873
NeuMF 100e, 1024b	105.11	855
NeuMF 100e, 256b	103.57	877
GMF 50e, 1024b	129.01	603
GMF 50e, 256b	129.40	606
GMF 100e, 1024b	128.07	612
GMF 100e, 256b	127.49	616

- **BiVAE** demonstrates exceptional diversity with an average popularity score of 45.68 (60-67% lower than other models) and recommends 1,642 unique items (1.9-3× more than other models). This superior diversity likely stems from its probabilistic framework, which naturally explores a broader range of the latent space during sampling.
- **LightGCN** shows the strongest configuration dependence, with unique items increasing from 545 to 753 (38% improvement) and average popularity decreasing from 137.14 to 116.24 (15% reduction) when moving from config1 to config4. This suggests that increasing epochs and reducing batch size helps LightGCN escape local optima in the graph structure.
- **NeuMF** exhibits moderate diversity that improves with smaller batch sizes. The 256 batch size configuration recommends 11% more unique items than the 1024 batch size variant, indicating that smaller gradient updates help capture more diverse user-item relationships.
- **GMF** shows limited diversity improvement across configurations, with minimal changes in average popularity (127.5-129.4) and unique items (603-616), suggesting an inherent limitation in its representation capacity.

D. Popularity Bias Analysis

The popularity bias metrics reveal clear differences in how models distribute attention across the popularity spectrum:

TABLE VI: Popularity Bias Metrics

Model Config	Popular Ratio
LightGCN 50e, 1024b	0.571
LightGCN 50e, 256b	0.476
LightGCN 100e, 1024b	0.472
LightGCN 100e, 256b	0.445
BiVAE All configs	0.200
NeuMF 50e, 1024b	0.425
NeuMF 50e, 256b	0.385
NeuMF 100e, 1024b	0.389
NeuMF 100e, 256b	0.383
GMF 50e, 1024b	0.534
GMF 50e, 256b	0.538
GMF 100e, 1024b	0.525
GMF 100e, 256b	0.529

- **BiVAE** exhibits remarkably low popularity bias with a popular ratio of only 0.20, meaning that just 20% of its recommendations come from the top 20% most popular items. This balanced distribution suggests that BiVAE's probabilistic generative approach effectively models the entire item space rather than concentrating on frequently observed interactions.
- **LightGCN** demonstrates the highest popularity bias, with popular ratios ranging from 0.44 to 0.57. This indicates that LightGCN's graph structure tends to amplify existing interaction patterns, as popular nodes naturally accumulate more connections and influence in the propagation process. The bias decreases with smaller batch sizes and more epochs, suggesting that these configurations help propagate influence to less-connected nodes.
- **NeuMF** shows moderate popularity bias (0.38-0.42) that improves with smaller batch sizes, likely because smaller updates help capture more subtle interaction patterns beyond the dominant popular items.
- **GMF** displays consistently high popularity bias (0.52-0.54) with minimal variation across configurations, indicating an inherent tendency to focus on popular items regardless of training parameters.

E. Cold Start Performance

1) Approach 1: Interaction Bucket Analysis Results: For user cold start scenarios:

TABLE VII: LightGCN User Cold Start Performance

User Bucket	Count	MAP	NDCG	Precision
1-28	1097	0.192	0.328	0.198
29-49	1792	0.159	0.306	0.255
50-92	3080	0.140	0.388	0.348
93-170	6068	0.118	0.540	0.491
171+	12971	0.091	0.724	0.698

- **LightGCN** shows relatively strong performance for users with very few interactions (MAP of 0.19 for users with 1-28 interactions) but exhibits a counterintuitive pattern where MAP decreases as user interactions increase. NDCG and Precision increase consistently with more user interactions.

TABLE VIII: BiVAE User Cold Start Performance

User Bucket	Count	MAP	NDCG	Precision
1-28	1097	0.000	1.000	0.587
29-49	1792	0.961	1.000	0.908
50-92	3080	0.625	1.000	1.000
93-170	6068	0.323	1.000	1.000
171+	12971	0.181	1.000	1.000

TABLE IX: NeuMF User Cold Start Performance

User Bucket	Count	MAP	NDCG	Precision
1-28	1097	0.071	0.141	0.101
29-49	1788	0.056	0.131	0.120
50-92	3072	0.074	0.196	0.150
93-170	6056	0.120	0.230	0.221
171+	12878	0.176	0.308	0.292

TABLE X: LightGCN Item Cold Start Performance

Item Bucket	Count	MAP	NDCG	Precision
1-4	168	0.000	0.000	0.000
5-15	762	0.000	0.001	0.001
16-42	2258	0.003	0.008	0.005
43-99	5678	0.010	0.026	0.020
100+	16142	0.185	0.464	0.379

TABLE XI: BiVAE Item Cold Start Performance

Item Bucket	Count	MAP	NDCG	Precision
1-4	168	0.985	1.000	0.164
5-15	762	0.993	1.000	0.206
16-42	2258	0.979	1.000	0.342
43-99	5678	0.916	1.000	0.503
100+	16142	0.712	1.000	0.827

TABLE XII: NeuMF Item Cold Start Performance

Item Bucket	Count	MAP	NDCG	Precision
1-4	175	0.037	0.054	0.016
5-15	1215	0.044	0.074	0.031
16-42	3474	0.069	0.125	0.063
43-99	6987	0.075	0.131	0.085
100+	13040	0.129	0.234	0.188

- **BiVAE** exhibits perfect NDCG across all user buckets, suggesting excellent ranking quality. It shows extremely high precision for all user buckets, with MAP values following an unusual pattern.
- **NeuMF** shows modest but consistent performance across interaction buckets, with performance generally improving as user interactions increase.

For item cold start scenarios:

- **LightGCN** completely fails for items with very few interactions (1-4 bucket shows zero for all metrics) and shows minimal performance until the 100+ interaction bucket, exhibiting strong popularity bias.
- **BiVAE** demonstrates exceptional performance for cold start items, with MAP and Recall exceeding 0.98 even for items with only 1-4 interactions, and perfect NDCG across all item buckets.
- **NeuMF** shows modest but measurable performance even for the coldest items, with performance increasing gradually across all metrics as item interactions increase.

2) Approach 2: Binary Cold Start Classification Results: For this approach, cold-start users were defined as those having 5 or fewer interactions in the training data, with warm-start users being all others. The evaluation focused on the NDCG@10 and Recall@10 metrics.

TABLE XIII: Cold vs. Warm Start Performance Comparison

Model	Epochs	Batch Size	Cold NDCG@10	Warm NDCG@10	Cold Recall@10
LightGCN	50	1024	0.316	0.514	0.317
BiVAE	500	128	0.335	0.530	0.335
BiVAE	50	1024	0.217	0.385	0.236
NCF	50	1024	0.156	0.266	0.160

Key observations from the binary cold start analysis:

- **BiVAE** with 500 epochs and 128 batch size outperforms other models for cold-start users, achieving a Cold NDCG@10 of 0.3348 and Cold Recall@10 of 0.3351.
- **LightGCN** performs quite well with Cold NDCG@10 of 0.3160 and Cold Recall@10 of 0.3169, showing better performance than expected given the item cold-start results from the first approach.
- **NCF/NeuMF** shows the poorest performance on cold-start users, with Cold NDCG@10 of 0.1556 and Cold Recall@10 of 0.1597.
- All models show a significant performance gap between cold-start and warm-start users, with BiVAE (500 epochs) and LightGCN having an NDCG gap of approximately 0.195 and 0.198 respectively.

VI. Performance Analysis and Insights

The comprehensive evaluation of LightGCN, BiVAE, NeuMF, and GMF across accuracy metrics, computational efficiency, recommendation diversity, and popularity bias reveals distinct strength-weakness profiles that make each model suitable for different recommendation scenarios.

Each architecture exhibits a clear performance-diversity tradeoff. LightGCN's graph-based approach delivers superior ranking accuracy (NDCG up to 0.458) through its effective propagation of collaborative signals, but simultaneously demonstrates the highest popularity bias (up to 0.57). This suggests that graph-based models naturally amplify existing interaction patterns, creating an accuracy-novelty tension. Conversely, BiVAE's probabilistic framework uniquely balances high accuracy (NDCG 0.439) with exceptional exploration capability (popularity bias of only 0.20), challenging the conventional wisdom that accuracy and diversity are fundamentally opposed.

The models also reveal a fascinating efficiency-expressiveness spectrum. BiVAE combines sophisticated probabilistic modeling with remarkably efficient computation (training times as low as 11.72 seconds), while NeuMF's hybrid architecture paradoxically results in both limited

expressiveness and poor computational efficiency (prediction times approaching 50 seconds). This demonstrates that architectural complexity does not necessarily translate to either better performance or higher computational costs.

Most notably, the cold start evaluation exposes fundamental differences in how these architectures generalize from limited data. While LightGCN excels with established interactions but completely fails for new items (zero performance for items with 1-4 interactions), BiVAE maintains robust performance even in extreme sparsity conditions (MAP exceeding 0.98 for cold start items). This suggests that generative probabilistic approaches may be inherently better suited for uncertainty modeling than discriminative or graph-based methods.

VII. Comparative Analysis

A. Architectural Differences

The four models represent fundamentally different architectural approaches to recommendation:

- **LightGCN** explicitly models higher-order connectivity between users and items through graph structure, focusing solely on neighborhood aggregation.
- **BiVAE** employs a generative probabilistic framework, modeling latent variables as distributions rather than point estimates.
- **NeuMF** combines two components: a GMF path for linear interactions and an MLP path for non-linear interactions.
- **GMF** uses a simple element-wise product mechanism to model user-item interactions.

These architectural differences lead to significant variations in how the models learn from interaction data and handle cold start scenarios.

B. Cold Start Handling

The models exhibit distinct behaviors in cold start scenarios:

- **BiVAE** demonstrates remarkable resilience to the cold start problem for both users and items, maintaining high performance even with minimal interaction data. It can effectively recommend items with as few as 1-4 interactions. Both cold start analysis approaches confirm BiVAE's superior performance, with the binary approach showing it achieves the highest Cold NDCG@10 (0.3348) and Cold Recall@10 (0.3351).
- **LightGCN** performs reasonably well for cold start users but completely fails for cold start items in the first analysis approach, showing zero performance for items with 1-4 interactions and requiring approximately 100+ interactions for items to be effectively recommended. However, the binary approach shows competitive Cold NDCG@10 (0.3160) and Cold Recall@10 (0.3169) values, suggesting its performance may be better than expected with a different definition of "cold-start."
- **NeuMF** shows moderate performance for both cold start users and items, with gradual performance degradation as interaction counts decrease. It requires around 16-42 interactions to begin providing reasonable recommendations. The binary approach confirms its poorer cold start performance with significantly lower Cold NDCG@10 (0.1556) and Cold Recall@10 (0.1597) metrics.
- **GMF** struggles with both user and item cold start problems, showing the poorest overall performance in these scenarios.

C. Training and Inference Efficiency

My experimental results reveal significant differences in computational efficiency:

- **BiVAE** demonstrates the highest efficiency in both training and prediction, with training times as low as 11.72 seconds for the 50-epoch configuration and consistent prediction times around 1.8 seconds.
- **LightGCN** shows moderate training efficiency but fast prediction times, making it suitable for real-time recommendation scenarios once trained.
- **NeuMF** exhibits the poorest efficiency, particularly in prediction times which reached nearly 50 seconds in some configurations, making it less suitable for real-time recommendation scenarios.
- **GMF** offers reasonable efficiency due to its simpler architecture but still lags behind BiVAE in both training and prediction times.

D. Recommendation Diversity

The models show substantial differences in recommendation diversity and popularity bias:

- **BiVAE** excels in providing diverse recommendations, with the lowest average popularity score, highest number of unique items recommended, and lowest popularity bias. This suggests BiVAE would provide users with more varied recommendations and better exposure to the long tail of items.
- **NeuMF** shows moderate diversity metrics, with reasonable numbers of unique items and moderate popularity bias.
- **LightGCN** exhibits higher popularity bias, particularly in configurations with fewer epochs, suggesting a tendency to favor popular items.
- **GMF** demonstrates the poorest diversity metrics, with high average popularity and high popularity bias, indicating a strong tendency to recommend already-popular items.

E. Model Complexity and Paradigm

The four architectures represent distinct recommendation paradigms with varying complexity:

- **LightGCN** represents a discriminative graph-based approach of moderate complexity. By removing feature transformation and non-linear activation functions from traditional GCNs, it focuses exclusively on neighborhood aggregation operations. This simplification enables efficient learning of user and item embeddings directly from the interaction graph structure.

- **BiVAE** adopts a generative probabilistic framework with higher computational complexity due to its encoder-decoder architecture and distribution modeling requirements. Unlike discriminative models, BiVAE captures underlying data distributions rather than merely discriminating between observed and unobserved interactions.
- **GMF** represents a discriminative approach with minimal complexity, generalizing traditional matrix factorization within a neural network framework. This simplicity makes it computationally efficient but potentially limits its expressive capacity.
- **NeuMF** extends GMF with a multi-layer perceptron component, increasing model complexity to a moderate level. This hybrid architecture enables the capture of both linear and non-linear interaction patterns.

F. Training Objective and Regularization

Each model employs distinct training objectives and regularization techniques:

- **LightGCN** typically optimizes Bayesian Personalized Ranking (BPR) loss or binary cross-entropy (BCE) loss with L2 regularization. The graph structure provides additional implicit regularization, contributing to robust performance with minimal explicit regularization.
- **BiVAE** optimizes the Evidence Lower Bound (ELBO), comprising a reconstruction loss term and a Kullback-Leibler divergence regularization term. This Bayesian approach provides an elegant theoretical framework for balancing reconstruction accuracy against overfitting.
- **GMF** and **NeuMF** both employ point-wise or pair-wise ranking losses with L2 regularization. NeuMF additionally incorporates dropout in its MLP component as regularization. These models rely more heavily on explicit regularization techniques compared to LightGCN and BiVAE.

G. Auxiliary Data Handling

The models differ substantially in their capacity to incorporate auxiliary information:

- **BiVAE** exhibits strong native capacity for auxiliary data integration, allowing the incorporation of content features directly into its encoder architecture. This flexibility enables BiVAE to handle multi-modal data and utilize side information.
- **LightGCN**, in its standard form, provides limited native support for auxiliary data, focusing primarily on the user-item interaction graph. While extensions exist to incorporate node features, the basic LightGCN model does not naturally accommodate side information.
- **GMF** similarly offers limited capacity for auxiliary data integration in its standard formulation. The simple element-wise product mechanism does not provide natural pathways for incorporating additional features.
- **NeuMF** offers moderate auxiliary data handling capabilities through its MLP component, which can potentially incorporate features with architectural modifications.

H. Hyperparameter Sensitivity

All four models require careful hyperparameter tuning, but with varying levels of sensitivity:

- **NeuMF** appears particularly sensitive to architectural choices given its multiple components. The number of factors, layer sizes, and learning rates significantly impact its performance, as shown by the variations in my experimental results.
- **BiVAE** demonstrates moderate sensitivity to hyperparameters, particularly the latent dimension size and encoder structure. However, it shows remarkable stability across different batch sizes.
- **LightGCN** is moderately sensitive to the number of graph convolution layers and regularization parameters, but robust to small changes in embedding size.
- **GMF** shows the least hyperparameter sensitivity among the four models, with relatively consistent performance across different configurations.

VIII. Discussion and Insights

A. Key Observations

Based on my comprehensive analysis, I draw the following conclusions about each model's characteristics:

1) **BiVAE** emerges as the superior model for addressing the cold start problem, particularly for item cold start scenarios. Its probabilistic framework enables effective modeling of uncertainty in user preferences and item attributes, allowing it to make reasonable recommendations even with minimal interaction data. Additionally, BiVAE demonstrates the best recommendation diversity and lowest popularity bias, suggesting it provides users with more varied and potentially serendipitous recommendations. Its strong native capacity for auxiliary data integration further enhances its versatility.
2) **LightGCN** shows strong performance for scenarios with sufficient interaction data but struggles significantly with cold start items. Its graph-based approach effectively captures collaborative signals through higher-order connectivity, but this strength becomes a limitation when dealing with items that have few connections in the graph. LightGCN would be most suitable for established systems with rich interaction data, though its limited ability to incorporate auxiliary information represents a potential drawback for certain applications.
3) **NeuMF** offers balanced but moderate performance across different scenarios. Its hybrid architecture provides flexibility but comes with increased computational costs, particularly for prediction. NeuMF's high sensitivity to hyperparameter configuration requires careful tuning to achieve optimal performance. This model might be appropriate when moderate performance across all interaction levels is acceptable and computational resources for prediction are not constrained.
4) **GMF** serves as a simple baseline with reasonable efficiency but limited performance. Its straightforward

architecture makes it computationally efficient but limits its expressive capacity for modeling complex user-item interactions. GMF shows the lowest hyperparameter sensitivity among the models, making it easier to implement and tune.

B. Model Selection Guidelines

Based on my findings, I propose the following guidelines for selecting recommendation models based on specific application requirements:

- **For new platforms with minimal interaction data:** BiVAE is the clear choice due to its exceptional cold start performance and ability to provide diverse recommendations even with limited data.
- **For established platforms with rich interaction histories:** LightGCN would be appropriate due to its strong performance for users and items with sufficient interactions, though care should be taken to address its bias toward popular items.
- **For balanced scenarios with moderate computational resources:** NeuMF offers a reasonable compromise between performance and complexity, though its higher prediction times should be considered for real-time applications.
- **For resource-constrained environments:** GMF provides a simple and efficient baseline, though with notable performance limitations.

C. Practical Implications

Several practical considerations emerge from my analysis:

1) **Minimum interaction thresholds:** My results indicate that LightGCN requires approximately 100+ interactions for items to be effectively recommended, while BiVAE can work with as few as 1-4 interactions. This has significant implications for new item promotion strategies.
2) **Training-inference trade-offs:** BiVAE offers the best balance between training and inference efficiency, making it particularly suitable for dynamic environments where both aspects are important.
3) **Recommendation diversity:** The significant differences in diversity metrics suggest that model selection should consider not just accuracy but also the breadth and variety of recommendations provided to users.
4) **Popularity bias:** All models except BiVAE exhibit some degree of popularity bias, which may reinforce existing consumption patterns rather than exposing users to new content. This should be considered when designing recommendation strategies.

IX. Conclusion

My comparative analysis of LightGCN, BiVAE, NeuMF, and GMF highlights the strengths and limitations of these distinct recommendation approaches. BiVAE demonstrates superior cold start performance and recommendation diversity, LightGCN excels with sufficient interaction data, NeuMF offers balanced but computationally expensive performance, and GMF provides a simple efficient baseline.

The performance hierarchy observed in my experiments (BiVAE $>$ LightGCN $>$ NeuMF $>$ GMF) reflects a progression from simple discriminative models to more sophisticated architectures that better capture the complex, multi-faceted nature of user-item interactions in recommendation systems. The choice between these approaches should be guided by specific application requirements, considering factors such as the importance of cold start handling, computational resources, and the desired balance between accuracy and diversity in recommendations.

References

[1] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in Proc. 43rd Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, 2020, pp. 639–648.

[2] Q.T. Truong, A. Salah, and H.W. Lauw, "Bilateral variational autoencoder for collaborative filtering," in Proc. 14th ACM Int. Conf. Web Search Data Mining, 2021.

[3] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. 26th Int. Conf. World Wide Web, 2017, pp. 173–182.

[4] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," ACM Trans. Interactive Intelligent Syst., vol. 5, no. 4, pp. 1–19, 2015.

[5] B. Marlin, R. Zemel, S. Roweis, and M. Slaney, "A systematic analysis of cold-start recommendations in recommender systems," in Proc. 23rd Conf. Uncertainty Artificial Intelligence (UAI), 2007.