Intepretability of the LightGCN Recommender System on MovieLens Dataset

by Langtian Ma

February 13, 2025

1 Introduction

Recommender systems have become indispensable in digital platforms. However, a growing number of modern recommender systems relies on complex models like deep learning and graph neural networks. Their decision-making processes are often opaque. This lack of transparency undermines user trust, debugging [2], and raises ethical concerns, such as bias and fairness [5]. The ability to explain why recommendatons are generated has thus become a critical research area. This report implements a slightly modified version of LightGCN [4], trains it on the MovieLens 1M dataset [3] and attributes feature importance for specific predictions using the Integrated Gradients method [6]. The approach successfully identifies key features of both users and movies that inluence the predictions.

2 Problem Formulation

Understanding which features most influnce recommendation decisions is crucial. In this report, we trained a modified LightGCN model with additional user features and item features informations on the **MovieLens 1M dataset** [3]. The MovieLens 1M dataset is a widely used benchmark dataset for evaluating recommender systems. It contains **1 millioin ratings** from **6,000 users** on **4000 movies**, collected by the GroupLens research team. Each rating is an integer between 1 (worst) and 5 (best). In addition to the user-movie interactions, the dataset provides user demographic informtion (age, gender, occupation, and zip code) and moview metadata (title and genres). We try to answer key questions such as:

- What user features most affect their recomendations?
- Which movie genres paly the most significant role in recoomendation scores?

The primary beneficiaries of this analysis include:

- End-users, who can better understand why a moview was recommended.
- Developers and researchers, who can debug and improve the model more eaily.

To investigate interpretability, we apply the Integrated Gradients method [6], which attributes importance scores to input features. This allows us to quantify how much different user and movie characteristics contribute to a specific recommendation.

3 Methods Application

3.1 Training LightGCN

LightGCN (Lightweight Graph Convolutional Network) is a simplified yet effective GCN-based model for collaborate filtering on user-item graphs [4]. We use an enhanced LightGCN which incorperate user and item side features. Let U be the number of users, M be the numebr of items, d be the dimension of embeddings, d_u be the dimension of user featrues, and d_m be the dimension of item features. The model is formulated by:

1. Trainbale Embeddings:
    - $\boldsymbol{P_u} \in \mathbb{R}^{U \times d}$ for users.

- $\boldsymbol{P_m} \in \mathbb{R}^{M \times d}$ for items.

2. Feature Projections:

   - User features $X_u \in \mathbb{R}^{U \times d_u}$ are passed through a linear larayer $\boldsymbol{W_u} \in \mathbb{R}^{d_u \times d}$.

   - Item features $X_m \in \mathbb{R}^{M \times d_m}$ are passed through $\boldsymbol{W_u} \in \mathbb{R}^{d_m \times d}$.

The initial embeddings for user u and item m are:

$$\boldsymbol{E_u^{(0)}} = \boldsymbol{P_u^{(0)}} + \boldsymbol{X_u W_u^{(0)}}, \quad \boldsymbol{E_m^{(0)}} = \boldsymbol{P_m^{(0)}} + \boldsymbol{X_m W_m^{(0)}}.$$

These are concatenated into a single matrix $\boldsymbol{E^{(0)}} = \left(\boldsymbol{E_u^{(0)^T}}, \boldsymbol{E_m^{(0)^T}}\right)^T \in \mathbb{R}^{(U+M) \times d}$. Let $\boldsymbol{A} \in \mathbb{R}^{(U+I) \times (U+I)}$ be the normalized user-item interaction matrix where edges encode interactions. The embeddings are propagated through $\boldsymbol{A}$:

$$\boldsymbol{E}^{(k+1)} = \boldsymbol{A E}^{(k)}.$$

After n steps, we average the embeddings $(\boldsymbol{E}^{(0)}, \boldsymbol{E}^{(1)}, \dots)$ to form the final node representation:

$$\boldsymbol{E}^* = \frac{1}{n+1} \sum_{k=1}^{n} \boldsymbol{E}^{(k)}.$$

Then we split $\boldsymbol{E^*}$ to $\boldsymbol{E_u^*} \in \mathbb{R}^{U \times d}$ and $\boldsymbol{E_m^*} \in \mathbb{R}^{M \times d}$. For a user-item pair (u, m), the predicted rating is computed via dot product:

$$\hat{y}_{u,i} = \boldsymbol{E_u^*}(u) \cdot \boldsymbol{E_m^*}(m).$$

A BCEWithLogitsLoss over each observed rating $r_{u,m}$ is used to trian the model:

$$\mathcal{L} = \sum_{(u,m) \in \text{data}} \text{BCE}(\hat{y}_{u,m}, r_{u,m}).$$

3.2 Applying Integrated Gradients

The Integrated Gradients (IG) method is an attribution method used to explain the prediction of neural networks by computing the importance of each input feature. Given a model f and an input x, the importance of each feature x_i is computed by:

$$\text{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha,$$

where x' is a baseline input.

In our approach, we combine user features and item features to form the input x, using the averaged feature values as the baseline input x'. We then apply Integrated Gradients (IG) to attribute the predicted scores for specific movies. The IG method for the recommender system takes approximately 40 seconds to run on a MacBook Air with an M1 chip.

We select a young student (age ≤ 18) and find that the system predicts *Toy Story 2* as his most favored movie. We then run the IG method to determine the contribution of different features to this prediction. As shown in Figure 1, IG provides a clear explanation of why the recommender system suggests a particular movie by identifying the most influential user and movie features in the prediction.

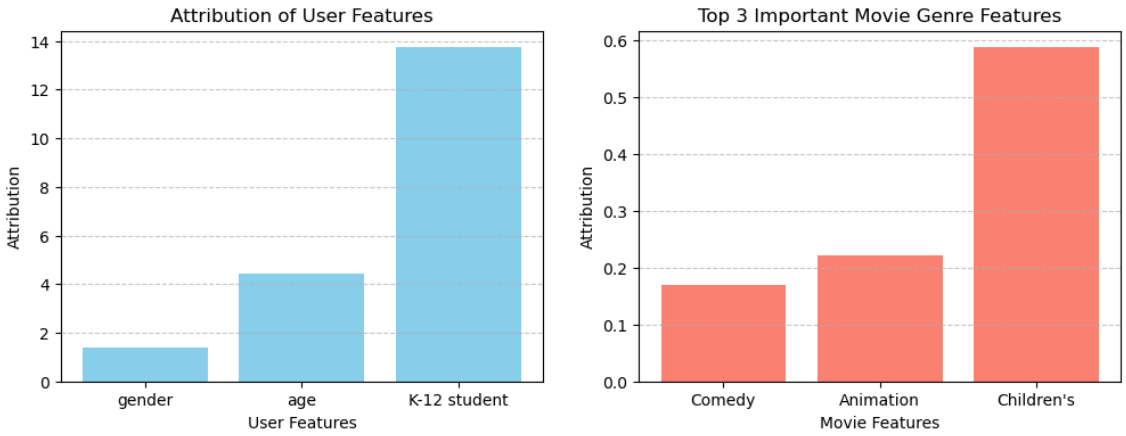

Figure 1. Feature importance of the predicted score that a teenager student would give to the movie "Toy Storey".

The attribution method can also help identify whether a user prefers certain genres. We selected a writer aged 25–30 and analyzed three movies, each including **Fantasy** as one of their genres. As shown in Figure 2, the attribution score for *Fantasy* is significantly higher than that of other genres, suggesting that the user has a strong preference for Fantasy movies.

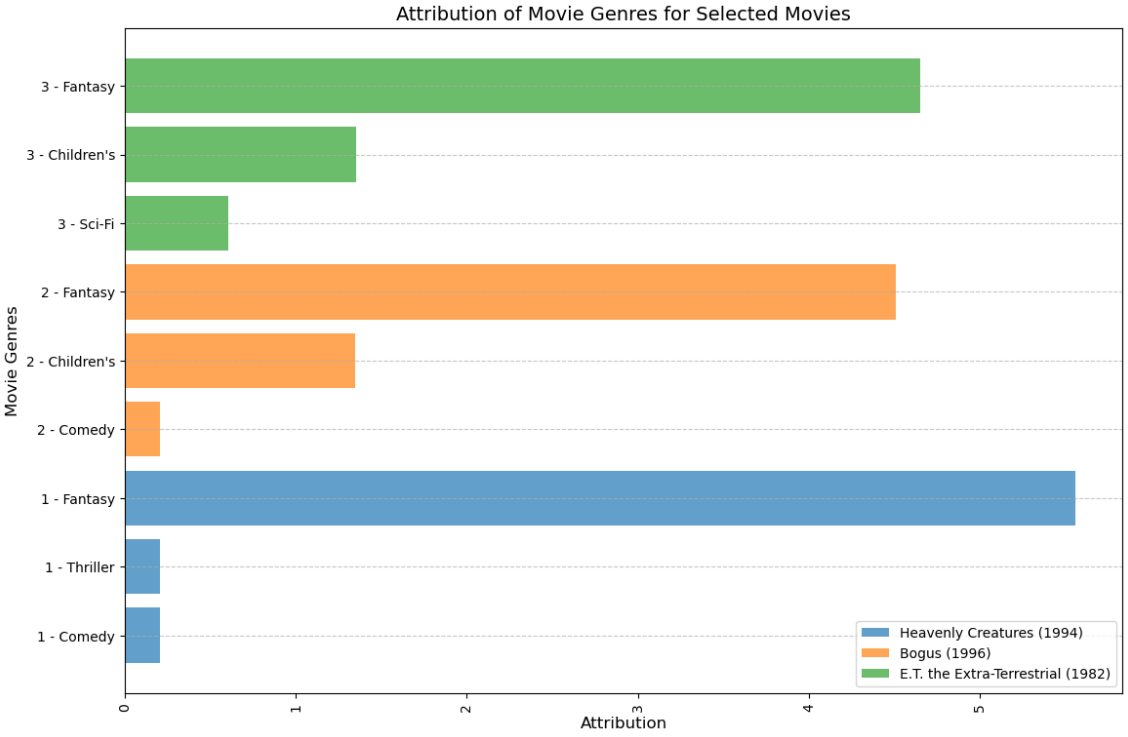

Figure 2. Top three most influential genres for selected movies. Higher values indicate greater impact on the model's recommendations.

4 Discussion

The results of the Integrated Gradients (IG) method provide insightful explanations for the recommender system's predictions. However, certain limitations remain. For example, one might assume that users who favor movies like *Toy Story 2* do so because the film appeals to young children. However, as shown in Figure 1, the user's occupation as a K-12 student receives the highest attribution, while age has a much lower attribution. This discrepancy likely arises because age and student status are highly correlated, leading to confounding in the data. This suggests that the attribution method does not support causal claims. Additionally, the results imply that the model relies on correlations rather than causal relationships, which may indicate poor out-of-distribution generalization [1].

Another observation from Figure 1 is that the scale of attributions for user features is substantially larger than that of movie features, despite feature normalization. While further experiments are needed to reveal the reason, a possible explanation is that one-hot encoding of user occupations results in a sparser feature vector than genre encoding for movies, where each movie can belong to multiple genres. This sparsity could cause the model to assign greater importance to user features during training, leading to the observed disparity in attribution scales.

To conclude, while IG offers valuable insights into the model's decision-making process, its reliance on correlation rather than causation limits its interpretability in real-world scenarios. Further experiments, such as alternative baseline choices or perturbation analysis, could enhance our understanding of the model's behavior and improve its robustness in different contexts.

Bibliography

[1] Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. 4(2):110-115.

[2] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in Information Retrieval. 9(5):355-475.

[3] F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. 5(4):19-1.

[4] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation.

[5] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319-3328. PMLR.