Avoiding Decision Fatigue with AI-Assisted Decision-Making

Sujen Kancherla and Aditya Melkote

December 6, 2023

1 Introduction

We make numerous daily decisions, including what food to eat and what entertainment to choose. A lot of times people start to experience "Decision Fatigue." It refers to a mental state where making choices becomes progressively difficult due to the overload of options. This fatigue can influence decisions, often leading consumers to opt for default choices, especially in scenarios like shopping, where the selection of product features becomes overwhelming. Researchers have noted that decision fatigue plays a significant role in impacting people in poverty, as it drains the mental energy needed for activities that could aid social mobility.

In the context of online shopping and recommendation systems, the goal is usually to align with user preferences and suggest relevant items. Our work, however, emphasizes the importance of preventing decision overload to facilitate quicker and more satisfying choices, thereby reducing decision fatigue. We investigate ways to minimize the number of decisions users must make during online selection processes, like choosing a movie or shopping for kitchen items. Our contributions include a dataset for sequential movie selection decisions. By predicting the rating that a use will give an item, we can recommend a product to a user that they might rate higher.

2 Dataset

2.1 Background

The dataset in question focuses on sequential decision-making in the context of movie review and selection, a task not adequately represented in previous datasets. It involves users sequentially reviewing movies and ultimately selecting one to watch. The dataset records various details such as movie titles, ratings, user IDs, and the users' reasons for their choices. The unique aspect of this dataset is its emphasis on negative interactions and the necessity for users to review at least four movies before making a selection, ensuring thoughtful engagement.

the data was collected with a movie review website used by participants from Amazon Mechanical Turk, who are required to have a high approval rate and a significant number of completed tasks. This approach yielded over 8,000 decisions from more than 800 decision sequences. To enhance the dataset's scale for machine learning algorithm training, additional data was generated using GPT 3.5 and GPT 4, simulating users' sequential movie reviews and selections, both with and without user-specific information derived from the MovieLens dataset. The final dataset comprises a subset of 303 movies selected from the MovieLens Dataset, offering a rich resource for studying sequential decision-making and recommendation systems.

2.2 EDA

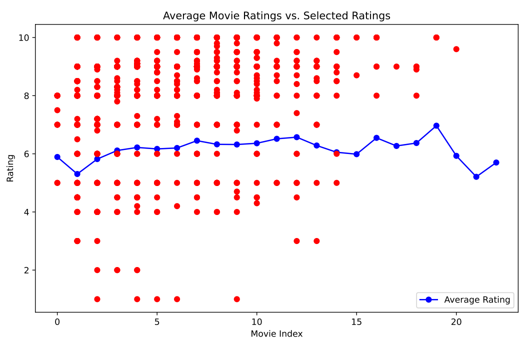

Figure 1: Distribution of Movie Ratings in Decision Sequences. This plot reveals the variability and general trends in user ratings across different decision sequences. The concentration of ratings towards specific parts of the sequence may indicate tendencies in user behavior, potentially influenced by decision fatigue.

The first visualization (Figure 1) presents the distribution of movie ratings in the decision sequences. It allows us to observe how users tend to rate movies throughout their selection process. A key observation is the distribution pattern of ratings, which can be linked to the onset of decision fatigue. The overall rating average has a slight decrease towards the end, showing that people tend to get a decision fatigue and lower their interest overall.

We can also see that there is not much of a pattern in the selected index, the users were picking a movie in the beginning and end of the sequence, The selection almost peaks at around 15 movies, because users are not going to continue for much longer before selecting a movie.

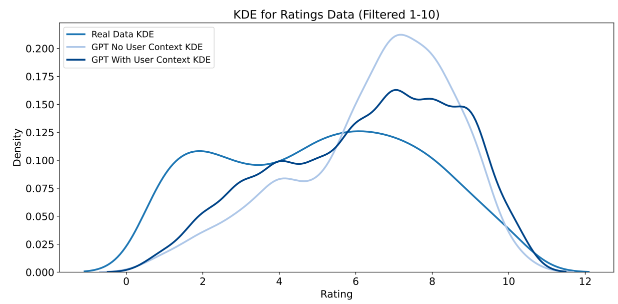

Figure 2: Kernel Density Estimation of Ratings: Comparing Real and GPT-Generated Data. This plot contrasts the density distributions of actual user ratings with those simulated by GPT models, providing insight into the realism and applicability of synthetic data in replicating user behavior.

The second visualization (Figure 2) is a Kernel Density Estimation plot that compares the ratings data from real users with those generated by GPT models. This comparative analysis is crucial for validating the use of synthetic data in our study. Similarities in the density curves would indicate that GPT-generated data closely mimics real user preferences, which is vital for training our AI models. Conversely, differences could highlight areas where synthetic data generation needs refinement to better replicate actual user behavior patterns.

The figure shows that the GPT data with context is pretty similar to the human generated data as the rating distribution is almost the same. We can also see that the difference in the distribution comes from the fact that the GPT generated data has a more skew towards positive ratings. This may be from the fact that OpenAI has put an emphasis on their reinforcement learning layers into safety and positivity in their model. This causes the synthetic data to be more inclined to higher ratings.

3 Related Work

3.1 Decision Fatigue

In previous research, various experiments were conducted to investigate the onset of decision fatigue during sequential decision-making tasks. In one set of experiments, a group of participants examined the pros and cons of different computer features without actually choosing any ([9]). Another group was tasked with configuring a computer by selecting specific features from a provided list, simulating the post-decision phase where decisions are implemented ([11]). A third group was given the freedom to explore and select computer features according to their preferences, differing from the first group who only contemplated options and the second group who merely implemented given choices ([10]). This third task was found to be the most exhausting, with post-hoc measurements indicating that this group experienced the greatest depletion in self-control. Many recommender systems, as discussed in sequential recommendation literature ([15, **?**]), focus on presenting users with items that might appeal to them next. Some online platforms even aim to maximize user engagement time ([14, 3, 13]). However, there is a notable lack of emphasis on facilitating efficient decision-making for users in these systems.

3.2 AI-Assisted Decision-Making

AI technology can be tailored to a range of applications by learning from user data, as supported by various studies [6, 8, **?**]. This AI-assisted decision-making has been implemented in sectors such as healthcare [7] and design [4]. It's also been employed to enhance sequential decision-making processes. One key function of AI in this context is to counteract various biases like position bias in recommendations [12], anchoring bias in college admissions [5], and other cognitive biases [1]. Methods used in AI-assisted decision-making encompass altering the sequence of presented instances [5], providing explanations that encourage active user involvement [1], and tracking the mental models of users [1]. Additionally, cognitive forcing functions, strategies designed to stimulate critical thinking during decision-making [2], have been developed. These functions, particularly in the context of AI explanations [2], have shown a positive impact on the interaction between humans and AI in decision-making scenarios.

4 Predictive Task

In this project, our objective is to predict movie ratings given by users. We will evaluate our model using Mean Squared Error (MSE), which accurately reflects the precision of our predictions by heavily penalizing larger errors. Our baseline models for comparison will be traditional collaborative filtering techniques. Additionally, we will compare these with a sophisticated model from our course to gauge improvements.

We will ensure the validity of our model's predictions by splitting our dataset into training and test sets. Key features for the model will include user demographics, historical ratings, and movie metadata, which will be preprocessed for consistency and to handle missing values. This approach aims to balance theoretical understanding from our course with practical application in predictive modeling.

The primary baseline model used is a simple linear regression model. This model simply predicts the rating the user would give for a title based on the average rating that users across the entire dataset give the movie. Another baseline that was used was a KNN. However, the KNN performed quite bad, leading us to choose the simple linear regression model as a better indicator of a baseline model to compare against the optimal model that we chose, perhaps due to the multicollinearity of some of our features in our data set. The features used in this dataset will be the user IDs, ratings and the imdb titles. Using these features we can accumulate the users to types of movies they will rate good and movies that they will rate lower.

5 Model

The most optimal model that we chose was an SVD with hyperparameter tuning to further optimize the model using cross validation grid search to find the best parameters with the lowest root mean square error. In essence, this model predicts the user's rating on a movie based on the ratings given by all the other users who have also watched that particular movie. As the data set was not enormous, scalability was not an issue that we ran into, and hence this was not a concern when choosing the final model. We did not run into overfitting as an issue either, as we incorporated splitting the data into a train and test data set and did not heavily over tune several of our parameters. Some of the other models that we tried along the way were K-nearest-neighbors , simple linear regressors and deep learning methods, all of which either resulted in sub par RMSE scores or unsuccessful attempts to train the model.

In general, neural networks shine when it comes to not only handling large amounts of data well, but also being able to handle complex non-linear relationships between features, as well as learning features implicitly. However, they come with the downsides of requiring large amounts of data to be able to be accurate, resulting in computationally intensive training that could not be feasible for us. Interpreting the reasoning behind prediction is also difficult with neural networks, due to their self-learning nature that is hidden away. KNN's are very easy to interpret and very simple to make, a perform as a good starting baseline model as well as being flexible as it does not utilize any assumptions about the nature of the data. However, due to their simplicity and poor scalability to larger datasets, as well hyperparameter and multicollinearity sensitivity, they are generally poor final models.

Although some of the other models such as some of the neural networks we tried had the potential to capture any deeper relationships between features and utilize the large amount of features present in our data set, where these models fell short were either overcomplexity resulting in either a very poor tradeoff between the difficult and lengthy time training the models, resulting mostly in complete failures of RMSE's that were not very much better than simpler baseline models, or the opposite, being so simple and basic that the complexity of the features were hardly captured.

6 Results

Our model outperformed the baseline models significantly, with RMSE after tuning the hyperparameters being roughly 1.4378, compared to baseline RMSE's of 2.016 and 2.486 for the linear regression and KNN models, respectively. The mean that the SVD model is significantly better at incorporating the latent factors to predict ratings for user – movie pairs, especially after tuning the n_factors parameter, which represent the latent features that are included. We used the standard number of epochs due to gridsearchCV taking too long when trying to find the optimal number for this parameter, although we optimized the learning rate of the other parameters using the former. The feature representations that work best for this is simply unique user, movie id and user rating for that movie id pair, as the SVD can capture the rest of the features through this. In some of the other models, we used movie_id features such as genre as well as user features that were part of the data set, as well as engineered additional features through performing textual sentiment analysis or finding the most common genre that users watched, which ultimately resulted in overly complex and multicollinearity.

The lengthy and difficult training time that caused some of the other models to fail stemmed from a very complex data set with numerical, categorical and text-based features that made pre-processing for the models extremely difficult and not worth it compared to the trade off in performance that these models gave. This was where the SVD model shined, performing well in terms of metrics as well as handling the complex underlying structure of the data while somewhat implicitly, capturing the latent factors in the data. This was one of the influential factors when choosing SVD over other models other than the having a far better RMSE than some of our others.

References

[1] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, 2022.

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[3] Riccardo Chianella. Addictive digital experiences: the influence of artificial intelligence and more-than-human design, 2021.

[4] Leah Chong, Ayush Raina, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. The evolution and impact of human confidence in artificial intelligence and in themselves on ai-assisted decision-making in design. *Journal of Mechanical Design*, 145(3):031401, 2023.

[5] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. Ai-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[6] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[7] Tamra Lysaght, Hannah Yeefen Lim, Vicki Xafis, and Kee Yuan Ngiam. Ai-assisted decision-making in healthcare: the application of an ethics framework for big data in health and research. *Asian Bioethics Review*, 11:299–314, 2019.

[8] Tariq Mahmood, Ghulam Mujtaba, and Adriano Venturini. Dynamic personalization in conversational recommender systems. *Information Systems and e-Business Management*, 12(2):213–238, 2014.

[9] Grant A Pignatiello, Richard J Martin, and Ronald L Hickman Jr. Decision fatigue: A conceptual analysis. *Journal of health psychology*, 25(1):123–135, 2020.

[10] Evan Polman and Kathleen D Vohs. Decision fatigue, choosing for others, and self-construal. *Social Psychological and Personality Science*, 7(5):471–478, 2016.

[11] Kathleen D Vohs, Roy F Baumeister, Jean M Twenge, Brandon J Schmeichel, Dianne M Tice, and Jennifer Crocker. Decision fatigue exhausts self-regulatory resources—but so does accommodating to unchosen alternatives. *Manuscript submitted for publication*, pages 1–55, 2005.

[12] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 610–618, 2018.

[13] Allison Zakon. Optimized for addiction: Extending product liability concepts to defectively designed social media algorithms and overcoming the communications decency act. *Wis. L. REv.*, page 1107, 2020.

[14] Zhengwei Zhao. Analysis on the "douyin (tiktok) mania" phenomenon based on recommendation algorithms. In *E3S Web of Conferences*, volume 235. EDP Sciences, 2021.

[15] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. Filter-enhanced mlp is all you need for sequential recommendation. In *Proceedings of the ACM web conference 2022*, pages 2388–2399, 2022.