Sallah AI Case Study

Introduction

The purpose of this report is to analyze the data provided, which consists of approximately 4.2 million user ratings for 476K items on a scale of 1 to 5 between 1998 to 2014. The goal is to build a recommendation system to increase user engagement and increase sales by showing the right product to the right user to solve the low AOV (average order value).

Answers

Q1: You have been tasked with building a model and must decide whether to use a CPU or GPU, taking into consideration that GPUs typically cost more.

A1: For this baseline, I have developed an algorithm that runs on a CPU to minimize costs for each customer. I recommend conducting predictions offline and storing them in a hash table for future use to save inference costs.

Q2: Describe whether the model will run continuously or intermittently.

A2: Currently, the model operates intermittently because I believe online learning is not required for this particular scenario. However, my answer may change once I learn more about the business requirements.

Q3: Explain how feedback will be obtained from the model and the retraining process.

A3: We will use A/B testing in production and Shadow evaluation to obtain feedback from the model. For our purposes, training in batches is sufficient, but we can switch to online training with a single function call if needed.

Q4: Discuss the model's precision and recall and suggest ways to improve them.

A4: The model's precision and recall are both above 70%, which is significantly better than random. We can improve the model's performance by adding more data, using more sophisticated algorithms, utilizing the timestamp feature, employing ensemble models, performing hyperparameter search, NVIDIA Merlin, and more. Additionally, we can vary the precision/recall threshold to sacrifice one over the other based on the business requirements. Adjusting the top-k recommendations is another simple way to affect precision and recall dramatically.

Q5: Indicate whether a pre-trained model is being used.

A5: For this baseline, I trained an SVD model from scratch as a starting point. In future iterations, I plan to evaluate different algorithms, both pre-trained and trained from scratch, to determine which is the best fit for our needs.

Method

There are two models provided to solve a “Collaborative filtering” problem, one by using matrix factorization and the other by using a cosine similarity approach. Both models have been trained twice, once by user-based and the other by item-based. All the other details will be explained in the jupyter notebook for better visualization and a step-by-step explanation.

Question: What is the correct type of recommendation model you will build that will give us the most accurate results, explain why you chose it among other types.

Answer: This is a typical collaborative filtering problem. The best fit is user-based collaborative filtering since we have more users than items. However, since the case study required item-based collaborative filtering, I decided to provide both, although they all gave similar results. Also, we don’t have any features related to users or items, so we can’t use content filtering as well.