Week 3 Report

Shuhan Zhang zhang19@sas.upenn.edu Xinyu Zhang joyxyz@sas.upenn.edu
Lexuan Chen amy0305@sas.upenn.edu

Problem Statement

Our objective is to model consumer hotel selection behavior using a **utility-based approach**. If consumers make choices based on perceived utility, we can estimate their preferences through a **multinomial logit model**, where selection probabilities are determined by utility values and a **SoftMax** function. First, we assume consumers share identical preferences and choose from a randomized offer set of **2 to 4 hotels**. This simplified case establishes the foundation for expanding our model to more complex scenarios with heterogeneous preferences and varied conditions.

We aim to minimize the error between the estimated and actual probability of a hotel being chosen. While individual choices vary, consumers exhibit a common probability distribution when presented with the same set of options. Understanding this decision-making process enables platforms like Expedia to improve recommendation algorithms and enhance booking conversions.

Our model's performance will be evaluated using **Root Mean Square Error (RMSE), accuracy,** and **Negative Log-Likelihood (NLL)**. However, challenges arise due to constraints such as limited visibility into the number of hotels a user views and variability in the offer sets displayed. Defining a true probability distribution for every possible set is impractical, too. Instead, we focus on **inferring the utility of each hotel**, providing a scalable and interpretable solution.

Our analysis requires Expedia user booking details and hotel preferences, but data limitations—such as incomplete visibility into the exact offer sets shown—add complexity. By framing the problem in terms of utility estimation rather than direct probability prediction, we aim to develop a robust and adaptable model for capturing consumer decision-making patterns.

Technical Approach

We model consumer hotel selection using a **multinomial logit (MNL) model**, where the probability of choosing hotel from a set is:

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

with *i* representing the latent utility of the hotel. This framework builds on the intuition introduced by Batsell and Polking (1985), who proposed a market share model based on choice probabilities and competitive interactions. Our objective is to minimize prediction error using **Negative Log-Likelihood (NLL)** as the loss function:

$$L = -\sum_{n} \sum_{i \in S_n} y_{n,i} \log P_{n,i}$$

We employ **Adam** as the optimizer due to its efficiency in handling sparse data. The model is implemented in **PyTorch**, following a structured pipeline:

Model Construction → Data Import → Training → Testing & Validation

Performance is assessed using **RMSE, Accuracy,** and **NLL**, comparing predicted probabilities against actual choices. We utilize the **Expedia Hotel Business Database** from Kaggle, but challenges include dataset biases and limited visibility into individual user histories. Our approach prioritizes scalability and interpretability, ensuring robust consumer preference modeling.

Initial Results

The model demonstrates stable convergence, with original and frequency-domain losses plateauing at 0.3581 and 0.1289 and keep consistent from Epoch more than 1000. Both in-sample (original: 0.358, frequency: 0.129) and out-of-sample metrics (original: 0.355, frequency: 0.125) align closely, indicating robust generalization. However, training progress halts prematurely, suggesting inefficiencies in optimizer settings or model capacity.

Our current research operates within a highly simplified experimental framework, resulting in a model that lacks sufficient feature complexity, explanatory depth, and predictive robustness. The current implementation demonstrates marginal performance gains over naive baselines (e.g., simple mean estimation), highlighting its limited utility for real-world applications. To address this, we propose enriching the model by incorporating item-specific attributes (e.g., temporal, spatial, or socioeconomic features) and introducing consumer segmentation strategies. Explicitly modeling heterogeneous preference patterns across user subgroups—such as distinguishing price-sensitive versus quality-driven consumers—could enhance both interpretability and estimation accuracy. These refinements would better align the framework with the nuanced dynamics of real-world decision-making processes.

Next Steps

Our current model assumes homogeneous consumer preferences, though real-world data suggests multiple preference types. A study by Jagabathula et al. (2020) explores choice modeling under heterogeneous consumer behaviors, but its methodology lacks a Python implementation. We aim to bridge this gap by developing a tailored solution. However, several technical challenges remain, including selecting an appropriate solver, improving our proficiency with PyTorch, managing large-scale datasets within GPU constraints, and parallelizing computations for varying choice sets. To refine our approach, we seek further guidance on PyTorch best practices, access to relevant project examples, and alternative modeling strategies for customer segmentation. While no alternative approaches have been identified yet, continued exploration is necessary. Thus we have established a structured framework for optimization and gained insights into formulating the problem efficiently.