Status Report & Risk Assessment

Auteurs :

Vincent	MORIN
Antonin	DOAT
Augustin	MOUTON
Nicolas	LAINE
Alexis	LULIN

Enseignant :

M. MEDEIROS MACHADO

Nous attestons que ce travail est original, qu'il est le fruit d'un travail commun au binôme et qu'il a été rédigé de manière autonome.

PARTIE I. Dataset

A. Synthetic Data Generation Protocol

Due to the insufficient volume and sparsity of the initial dataset, which limited the ability to effectively train the Heterogeneous Graph Information Bottleneck (HGIB) model, we generated a robust synthetic dataset comprising 800 users and 5,200 unique trips spanning the 2015–2025 period. To ensure the data remained coherent and semantically rich, we followed a rigorous generation protocol starting with strict metadata chaining, where user attributes like nationality deterministically influenced cultural nomenclature and continent of origin. We further assigned latent behavioral profiles to each user—such as "Cultural/History" or "Adventure/Nature"—which acted as hidden variables driving the selection of destinations. This edge formation process was governed by a conditional probability model that prioritized profile-matching destinations in 95% of cases to create strong graph community structures, while introducing a 5% random exploration factor to simulate realistic human variability and prevent overfitting. Finally, we enforced "hard" environmental constraints to prevent logical inconsistencies, such as restricting specific accommodation types or transport modes to compatible destinations, thereby ensuring the resulting heterogeneous graph encoded meaningful, learnable relationships rather than random noise.

PARTIE II. EDA

The EDA was performed on the raw dataset containing 5,200 trips. The distributions of numerical variables (age, duration, costs) were analyzed using histograms and boxplots, revealing an average trip duration of around 6 days and a median daily cost of approximately $100. Cross-analyses identified significant seasonal patterns, including a preference for summer travel. Correlations between variables were examined to detect key relationships (duration/cost, age/budget). Finally, segment-based analyses (gender, nationality) identified behavioral differences between user groups, which are essential for the recommendation system.

PARTIE III. HGIB Framework

A. Graph Design & Data Representation

To effectively model complex travel preferences, we constructed a heterogeneous bipartite graph using the PyTorch Geometric framework. The graph topology consists of two distinct node sets—Users and Destinations—connected by directed edges representing individual trips. Unlike traditional recommendation graphs that often rely solely on interaction existence, our design enriches the nodes with intrinsic features to solve the cold-start problem and improve generalization. User nodes are initialized with a composite feature vector containing encoded gender and nationality,

alongside normalized age values to capture lifecycle-related travel habits, while destination nodes incorporate geographic context through encoded country identifiers.

The core innovation of our architecture lies in the edge attribute engineering, which serves as a direct representation of "Travel Behavior." Each interaction edge encapsulates the specific context of a trip, allowing the model to distinguish *how* a user travels, not just *where*. This behavior is modeled through a hybrid data structure: categorical variables (accommodation type, transport mode, and season) are prepared for embedding layers, while quantitative metrics (accommodation cost, daily budget, and duration) are normalized using a Standard Scaler. This separation allows the Graph Neural Network to simultaneously learn semantic preferences and financial sensitivities, providing a holistic view of the user's decision-making process.

PARTIE IV. PreProcessing Pipeline

Since the dataset was already clean, we did not need to handle null values or typos. However, we extracted new features from our columns, such as:

- The season (from the start date).
- The country (from the destination).

For the remaining preprocessing, we encoded all categorical data into numerical values by mapping each unique category to a number.

Finally, we plan to test whether normalizing the numerical data improves the model's performance.

PARTIE V. Style & Streamlit

Streamlit development commenced this week, successfully establishing the application's initial architecture and defining the complete visual style. The design features an orange background with black text, utilizing the *Pynion Script* for the startup name and *Glacial Indifference* for the main content. While this is the team's first extensive use of Streamlit for a multi-page application, development is progressing smoothly, confirming the team is well-positioned to master this tool to create a comprehensive user interface.