Space Optimization Challenge

– Manish Kumar Vuppugandla

Assumptions & Feature Engineering

- Does the "True" value in the "IS_CANCELLED" column indicate that the meeting was canceled?
  - If it does, we should investigate further to predict the likelihood of a booked meeting actually taking place.
  - Assumed otherwise as of now

- 'duration' of the meetings (in minutes)
- 'duration' column into 'counts' (1 count = 30 minutes)

- Question?
  - To only constrict rows using 'Duration' less than or equal to 12 hours.
  - I.e, Duration count <= 24

Choosing Model and Deployment

Several key observations were made during the analysis of the dataset, which influenced the choice of machine learning models and deployment strategies:

- **Data Skewness:** The data exhibited complex skewness patterns. Approximately 90% of room booking data was concentrated in the year 2023, with the months from May to August accounting for 80% of bookings. This skewness needed to be considered during modeling.
- **Meeting Durations:** More than 95% of bookings had durations of 1-2 hours, with a long right tail in the duration distributions. This skewed distribution of meeting durations had implications for model training and prediction.

- **Building Dominance:** A small subset of buildings, approximately 5 out of many, accounted for over 50% of total bookings. When considering 13 such buildings together, they contributed to more than 80% of the 77,000 bookings in the dataset. This building dominance indicated the need for building-specific modeling approaches.

- **Amenities Influence:** Certain amenities offered in meeting spaces, such as 'Dual Monitors,' 'Height Adjustable Desks,' 'Docking Stations,' and 'Keyboard & Mouse,' collectively accounted for over 50% of bookings. Understanding the popularity of specific amenities was crucial for feature selection.

- **Floor Significance:** Floor selection also played a substantial role. Floors 1 to 5 combined accounted for more than 50% of bookings, and roughly 75% of bookings occurred on floors numbered less than 10 (e.g., 1, 2, 3, 4, 5, 6, 7, 8, 9, 10). Contrary to the assumption that people prefer top floors or skyscrapers, the data indicated otherwise.

- Given these observations and the clear boundaries in terms of feature popularity (e.g., floors, amenities, accepted guest count, building ID, time of day, and day of the week), traditional machine learning models such as decision trees and random forests were chosen as suitable candidates. These models could handle the skewed data and capture the specific preferences and patterns evident in the dataset.

- Furthermore, upon closer examination, it was evident that a significant portion of meetings had a duration of approximately 1 hour. Additionally, the meeting capacity, or the number of people accepted to these meetings, typically averaged around 5 individuals. This trend indicated a preference for smaller group sizes when booking meetings.

Model deployment in Production

- Additional preprocessing steps will be required for handling categorical features such as *amenities, time of day, day of the week, month of the year, building ID*, and others.
- Furthermore, as part of further app development, the integration of a user-friendly front-end interface is essential to facilitate these preprocessing tasks.

- Regarding prediction interpretation, it's important to note that *each prediction* corresponds to a count, with each count representing a time duration of *30 minutes*.
- In terms of model deployment, it's advisable to include post-processing steps to convert predictions back into meaningful durations.

- When integrating the model into the app, additional API calls should be added to handle preprocessing before invoking the prediction API. The pretrained model can be packaged along with the app for deployment on the client's device or cloud hosting to facilitate website integration.