MINING HOTEL BOOKINGS TO PREDICT CANCELLATIONS

SEIJI AOYAMA & ELAINE FRENCH

DESCRIPTION

This project aims to explore and identify the key factors that can predict hotel booking cancellation rates. The hypothesis is that factors such as lead time, deposit type, previous cancellation history, and booking channel significantly influence the likelihood of a booking being canceled. By analyzing these variables, the project seeks to develop a model for accurately predicting cancellations.

PRIOR WORK

- One notable prior work using the same dataset is "Predicting hotel bookings cancellation with a machine learning classification model" (2017) by Nuno Antonio, Ana Almeida, and Luis Nunes. This study used sophisticated machine learning classification models to predict hotel booking cancellations and analyzed the factors influencing cancellations. While we aim to learn from their research, we also strive to generate new insights from the dataset.

DATASETS

We are using the hotel booking demand dataset found at

https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data

This data was provided by **two hotels in Portugal: a resort hotel in the Algarve region (H1) and a city hotel in Lisbon (H2)**. The authors of the paper, Nuno Antonio, Ana de Almeida, and Luis Nunes collected and published the data with the cooperation of the hotels.

The dataset has been downloaded on Seiji's local machine.

PROPOSED WORK

Data cleaning:

- Missing Values: Check for missing values, especially in key columns like is_canceled and lead_time. Columns may contain NULL or NaN values. Decide whether to remove or impute these missing values.
- Outlier Detection: Look for outliers, such as unusually high values in lead_time or adr (average daily rate). Consider handling these outliers as they can affect the prediction model.

Data preprocessing

- Encoding Categorical Variables: Encode categorical variables so that they can be used as inputs for the model. This may involve one-hot encoding.
- Data Type Conversion: Convert date-related columns such as arrival_date_year into appropriate formats.

Data integration

- The dataset already contains comprehensive information about hotel bookings, so integrating additional datasets is generally not required.

TOOLS & EVALUATION

- Python library: Pandas, Scikit-learn
- Analytics will be conducted using Jupyter notebook
- We plan to evaluate the model for cancellation rates from multiple perspectives, including cross-validation.