IEOR E4523 Data Analytics – Project Dataset Submission

Team: Pandas Learning

Joseph High (jph2185), Hoa Duong (hhd2115),
Jiawen Li (jl6121), Mary Daudi Kiangi (mdk2170),
Ziyun Lu (zl2961)

Data Description

The dataset consists of a list of 1,000+ hotels and their reviews from various sources such as *TripAdvisor, hotels.com, expedia.com*. The dataset includes hotel location, its name, ratings, review titles, review texts, usernames of the reviewers, hotel features, etc. The dataset contains a total of 70 attributes of various types, including date, categorical, numeric, and string, etc.

Data Source

The CSV files attached were sourced from Kaggle, but were originally sourced from Datafiniti's Business Database, which includes a more expansive version of the data that can be sourced from their API in either CSV or JSON format. We would be able to download this version directly from Datafiniti. Of course, the original dataset was sourced from various sites such as TripAdvisor, hotels.com, expedia.com, etc.

Data Source Links:
https://www.kaggle.com/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews_Jun19.csv
https://www.kaggle.com/datafiniti/hotel-reviews?select=Datafiniti_Hotel_Reviews.csv
https://www.kaggle.com/datafiniti/hotel-reviews?select=7282_1.csv
https://developer.datafiniti.co/docs/business-data-with-python-json

Project Analysis Proposal

Using the reviews data and hotel features, we will perform text mining on the text review data and try to generate and visualize word clouds to identify high frequency keywords in highly positive and negative reviews.
Next, we will develop a variety of models using Natural Language Processing (NLP) like tf-idf and/or Machine Learning (ML) techniques that predict the probability that a user will rate a hotel as positive (or negative) and ultimately classify whether the review is positive or negative. The initial feature set will include user review, hotel features, and geographic location. We can further improve the project by using hotel features (like services provided such as 24-hour front desk, dry cleaning, etc., or location and nearby attractions) to predict whether a user will like or dislike the hotel and provide recommendations. Building upon that, we may be able to predict and provide a score for the hotels, with a higher score representing a greater probability that a hotel will receive good reviews and ratings.

Training a binary or multi-classification model

We will initially split the data into train and test sets. The model development process will include a feature engineering step where features will be transformed if necessary (i.e., one-hot encoding, etc.), a feature selection process whereby features will be chosen based on their contribution to the independent variable (i.e., feature importance), and training a variety of models using different machine learning techniques/algorithms on the data and determining which performs best. Finally, we will evaluate the final selected model's performance on the test set.