HW Assignment 1 (A1): Exploratory Data Analysis (EDA)

CS6140: Machine Learning Fall 2024

Due Date: Tuesday, September ___, 2024

(10 points)

Introduction

In this assignment, you'll explore data from Airbnb, one of the top online platforms for vacation rentals, to gain hands-on experience with **Exploratory Data Analysis (EDA)** and **Feature Engineering**. You will download listings and reviews data for **five cities of your choice** (available from Q2 2024) and analyze it to identify trends, correlations, and patterns that can drive business insights.

You'll also analyze the **Reviews dataset** to engineer features aimed at predicting customer satisfaction and improving service offerings. Be sure to specify the cities you've chosen in your submission.

The assignment follows the STAR methodology: Situation (S) and Tasks (T) are outlined for you. You're responsible for carrying out the Actions (A) and analyzing the Results (R).

Situation (S):

You are a data scientist at an analytics firm specializing in real estate and tourism. Your team is tasked with analyzing Airbnb data to provide insights that will help optimize property listings and improve guest satisfaction. Your task is to focus on Airbnb Listings dataset and the Reviews dataset (and the data dictionary). The data can be downloaded here:

https://insideairbnb.com/get-the-data/

The Listings dataset contains detailed information about each property, including price, location, availability, and host details. The Reviews dataset includes guest comments on various properties, reflecting their experiences. Your objective is to c**onduct exploratory data analysis (EDA)** on the Listings dataset and **perform feature engineering** on the Reviews dataset to derive actionable insights to enhance the Airbnb platform.

Tasks (T):

1. Descriptive Statistics
Calculate summary statistics for numerical features such as `price`, `minimum_nights`, `maximum_nights`, `number_of_reviews`, and `review_scores_rating`. Understand the central tendency, dispersion, and distribution of these variables.

2. Distribution Analysis
Plot histograms or density plots for key numerical features like `price`, `minimum_nights`, and `review_scores_rating`. Analyze the distribution of these features to identify any skewness or outliers.

3. Correlation Analysis
Create a correlation matrix to explore relationships between numerical variables such as `price`, `number_of_reviews`, `availability_365`, and `review_scores_rating`.
Identify any strong correlations that might be useful for predictive modeling or further investigation.

4. Price Analysis
Analyze the distribution of prices across different neighborhoods (`host_neighbourhood`) or room types (if available). Understand which neighborhoods have higher or lower average prices and whether certain neighborhoods are more popular for shortterm or longterm stays.

5. Neighborhood Comparison
Compare the average `review_scores_rating` across different neighborhoods. Determine if certain neighborhoods have consistently higher ratings, which could indicate better or worse guest experiences.

6. Outlier Detection

Identify outliers in the dataset, particularly in price, minimum_nights, and review_scores_rating.

7. Text Length: Create a new feature that measures the length of each review (number of words or characters). Determine if the length of a review correlates with its sentiment or the review scores.

8. Keyword Extraction
Identify and count the occurrence of specific keywords (e.g., "clean," "comfortable," "noisy") in the reviews. Generate new features based on the presence of these keywords, which might influence guest satisfaction.

Actions (A): Perform the tasks listed above. Document your approach, the steps you took, and any assumptions you made. Visualize your findings where applicable.

Results (R): After completing the tasks, summarize the key insights you gained from the EDA and Feature Engineering. Report your findings in the form of a formal report with the following sections: Introduction, Analysis, Limitations, Conclusion.