EMORY UNIVERSITY

Introduction to Business Data Analytics

Group 5
Project proposal

Team members:
Abhishek Saha, Sameen Haroon, Sharon Xu, Zeyu Zhu

Business Problem

- Definition of the business problem: Make prediction of the price of an Airbnb listing based on its features.

  Airbnb is an online marketplace and hospitality services company. Its members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; as a broker, it receives commissions from every booking.
  A major factor for someone choosing to rent a place is its price, which can vary based on multiple factors. The objective of this data set was to predict the price of a listing based on some features / factors of the listing. This data set was part of a competition where the aim was to predict the price of Airbnb listings in major US cities.

- How tackling this business problem would add business value: Price of a listing is by far the topmost reason for a listing to get rented on Airbnb. If the price can be predicted using data science, it significantly increases the chances of getting rentals. It ultimately increases the revenue for the listing owner and the commission for Airbnb, hence increasing both topline and bottom line.

- The typical use scenario: As an owner would provide the features of the listing on Airbnb, the model would predict the price based on those features and historical data for listing rentals. This value would then be provided to the listing owner as suggested price on the Airbnb website.

Modeling Ideas

This project presents a supervised regression problem, since the goal is to predict the price of a listing based on its attributes. Linear regression, decision trees and nearest neighbors can be used to build different predictive models for the listing price, each of which can be evaluated using nested k-fold cross-validation.

- Each Airbnb listing represents a separate data instance, encoded as distinct rows in the data
- The target variable of interest in this case is the log price of a listing
- A total of 28 relevant features are included in this dataset (excluding id column and log data).
  - When looking at factors that influence the price of a listing, it is likely that the location e.g. city and neighborhood variables will prove useful in the model.
  - Similarly, details on the space available e.g. property type (apartment vs. house) and room type (shared room vs. entire home) will capture meaningful variation in price.
  - Some features will require transformation prior to model inclusion as well. For instance, decoding the ‘amenities’ provided by a listing into wireless, television, etc. will be necessary.
  - Automatic feature selection will be included in certain techniques, while additional feature selection may be implemented as needed for other algorithms, such as k-NN.

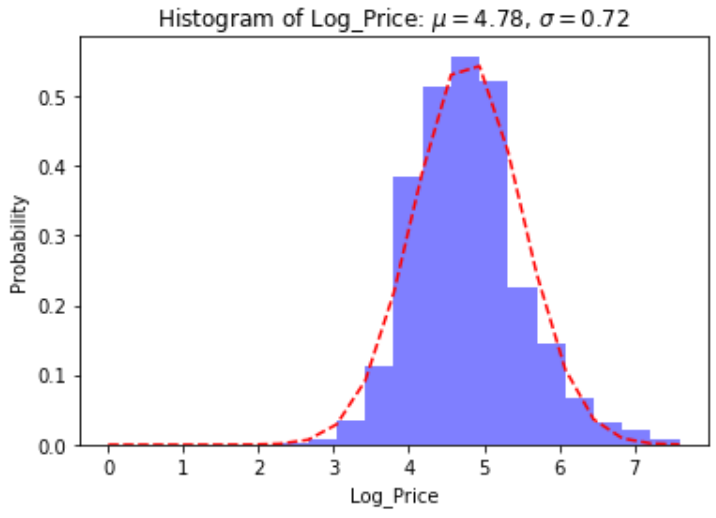
Data Details

- Data source: Airbnb listings in major US cities on Kaggle Dataset
  Link: *https://www.kaggle.com/rudymizrahi/airbnb-listings-in-major-us-cities-deloitte-ml*

- Description of variables: There are 74,111 observations as well as 29 columns in this dataset. The first column is id, the second column is the target variable *log_price*, and the rest 27 variables are features for the listing. Details of variables can be found in the table below.

Variable	**Type**	**Notice**
property_type	Object; categorical	
room_type	Object; categorical	
amenities	Object	Contains NA
accommodates	Integer; numerical	
bathrooms	Float; numeric	
bed_type	Object; categorical	
cancellation_policy	Object; categorical	
cleaning_fee	Boolean; binary	
city	Object; categorical	
description	Object; string	
first_review	Object; date	Contains NA
host_has_profile_pic	Object; binary	Contains NA
host_identity_verified	Object; binary	Contains NA
host_response_rate	Object; string	Contains NA; can be transformed to float
host_since	Object; date	Contains NA
instant_bookable	Object; binary	Can be transformed to boolean
last_review	Object; date	Contains NA
latitude	Float; numerical	
longitude	Float; numerical	
name	Object; string	
neighbourhood	Object; categorical	Contains NA
number_of_reviews	Integer; numerical	
review_scores_rating	Float; numerical	Contains NA
thumbnail_url	Object; string	Contains NA
zipcode	Object; string	Contains NA
bedrooms	Float; numerical	Contains NA
beds	Float; numerical	Contains NA

- Distribution of the target variable: *log_price* obeys normal distribution with mean=4.78, standard deviation=0.72, skewness=0.51 and kurtosis=0.66.



Count	74111
Mean	4.7821
Std	0.7174
Min	0
25%	4.3175
50%	4.7095
75%	5.2204
Max	7.6004
Skewness	0.5147
Kurtosis	0.6606