Predicting House Prices with Machine Learning

A sneak peek into the Airbnb activity in Seattle, WA, USA

GROUP 9:
KRISHI DIVYA DHARSHINI V
BHARATH VISHAL G

Data Set:

- The Seattle Airbnb open dataset 2016-2017 from Kaggle.
- This dataset is part of the Airbnb Inside Initiative.
- The following Airbnb activity is included in this Seattle dataset:
    - Listings, including full descriptions and average review score
    - Reviews, including unique id for each reviewer and detailed comments.
    - Calendar, including listing id and the price and availability for that day

OBJECTIVES:

- ➔ To predict the prices of listings to give insights for future hosts.
- ➔ To study what factors affect the most.
  - ◆ Listing.csv data is used
- ➔ To understand the seasonal pattern of prices.
  - ◆ Calendar.csv data is used
- ➔ To learn the relationship between reviews and prices.
  - ◆ Review.scv data is used

Seasonal Pattern of Prices:

- The period of time in which the prices are maximum is analysed.
- The variation of prices at different times of the week is also studied.

Sentiment Analysis of Reviews:

- The reviews.csv file is used for sentiment analysis.
- The polarity and the **Sentiment type** is generated using the **textblob** module.
- The **average polarity** of a listing is generated to understand the customers' views.
- The relationship between number of reviews are price is analysed.

Sentiment Analysis of Reviews

- The reviews.csv file is used for sentiment analysis.
- The polarity and the **Sentiment type** is generated using the **textblob** module.
- The **average polarity** of a listing is generated to understand the customers' views.
- The relationship between number of reviews are price is analysed.

Exploratory Data Analysis

1. The following features were analyzed categorically with their respective prices using boxplot:
   a. Room type
   b. Property type
   c. Bed type
   d. Host_is_superhost

2. Count Plot is used to analyze:
   a. Host_response_time
   b. Neighbourhood_group_cleansed
   c. review_scores_rating

3. Analyze data using Correlation matrix to extract features.

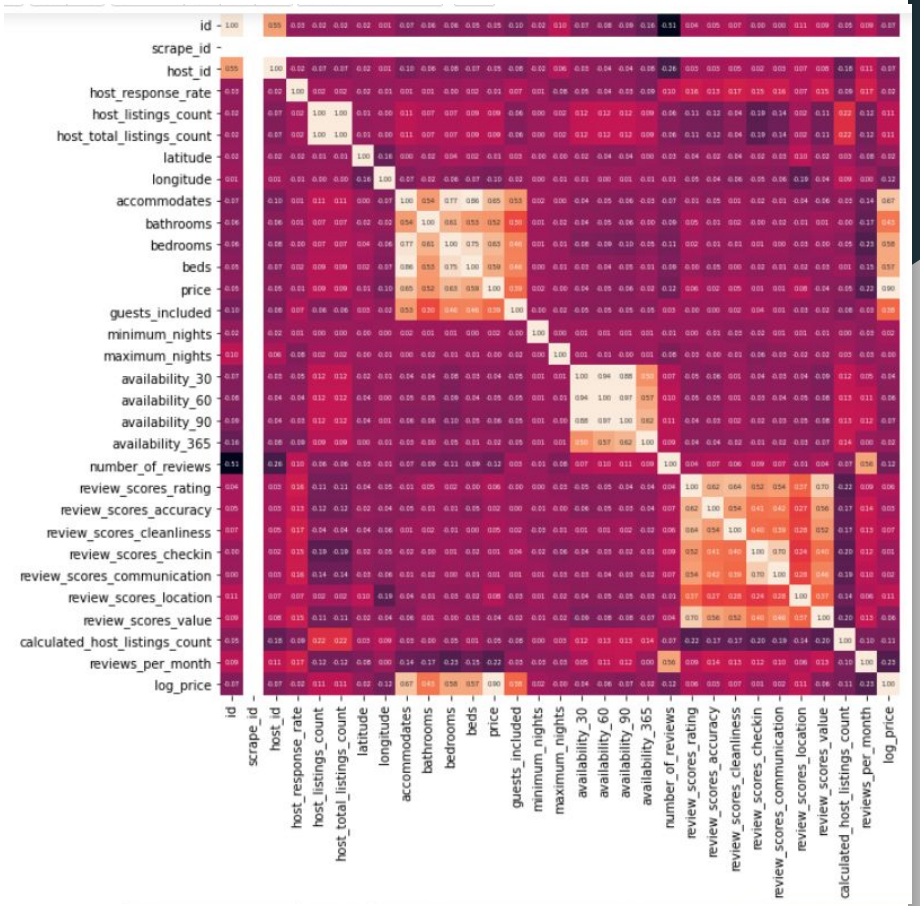
Data Preparation for Training

- Cleaned our target variable - price by casting the data type of price column from **string to float.**
- **Filled Null values** of
  - categorical variables.
  - numerical variables with mean value.
- Apply **Log transformation** of variable price to reduce skewness.

Feature Extraction

- A new dataframe with the features of our interest is created with the help of our EDA.
- The selected features are:

'property_type', 'room_type', 'bathrooms', 'bedrooms', 'bed_type', 'accommodates', 'guests_included','review_scores_rating', 'neighbourhood_group_cleansed', 'log_price'



Scaling the Features:

- StandardScaler() from the sklearn module is used to transform the data such that its distribution will have a mean value 0 and standard deviation of 1.

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

- Features are scaled so that every variable contributes to the model equally.

Encoding Categorical Features:

- ML models and Neural Networks require all input and output variables to be numeric.
- Hence, the categorical variables should be encoded before fitting and training.
- We use the OrdinalEncoding() from the sklearn module for this process.
    - The following features are encoded:'property_type', 'room_type', 'bed_type', 'neighbourhood_group_cleansed'

Reading Images:

- ➔ The column ‘picture_url’ contains the url for corresponding pictures of listing posted by the host.

	picture_url
0	https://a1.muscache.com/ac/pictures/67560560/c...
1	https://a0.muscache.com/ac/pictures/14409893/f...
2	https://a2.muscache.com/ac/pictures/b4324e0f-a...
3	https://a0.muscache.com/ac/pictures/94146944/6...
4	https://a1.muscache.com/ac/pictures/6120468/b0...

- ➔ Using the urllib and skimage module the images from the web pages are read and stored in an array ‘images’
- ➔ The images are resized for uniformity and scaled.

The employed model

- For Numerical data, we employ ANN to train the model
- For images, we employ CNN to extract features
- We concatenate the individual outputs.
- ANN for final prediction of price.