AI PROJECT

Yap Jack (1211103024)
Ashley Sim Ci Hui (1211101022)
Koh Jia Jie (1211102879)
Ng Yi Min (1221303664)

INTRODUCTION

There have been a lot of online marketplaces that changed the way of the people to travel and book for accommodations. Both the host and the customer want to make sure the price of the Airbnb is suitable from their perspective. Therefore, we need to be able to predict the price of the Airbnb accurately to optimize the price strategies for both the host and the customer.

PROBLEM FORMULATION

Goal: Predict the price of a Boston Airbnb based on features such as room type, number of rooms, number of beds and so on.
Dataset used: Airbnb listing data for Boston that includes features for us to train and evaluate the models.
Models used: Linear Regression Model, Decision Tree Regressor Model, Stacking Regressor Model, Gradient Boosting Regressor Model
Workflow: Data preprocessing and pipeline, model training, model evaluation and discussion.

DATA PREPARATION & PROCESSING PIPELINE

Steps:

1. Load & Get Overview of Dataset

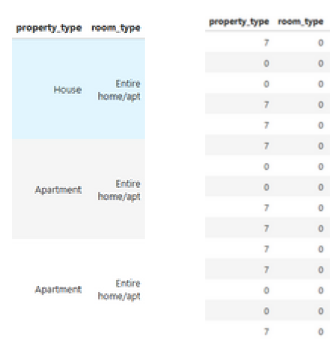
2. Data Cleaning

- Drop unnecessary features such as identifiers & URLs, textual descriptions, constant/nearly constant features, highly missing features, review date features, host details, and calendar & availability features
- Apply missing and duplicate value handling
  - Missing values: 'bathroom', 'bedrooms', and 'beds' are grouped by 'room_type' and filled with the mode of each group and others are filled with the mode of each column.
  - Duplicate values: No duplicate values, so no need for cleaning
- Handle outliers by removing outliers in the 'price' column for each room type
- Filter market for "Boston" and is_location_exact for "t"

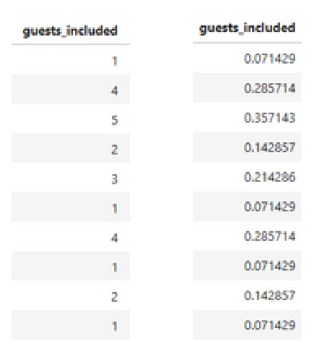
3. Data Transformation

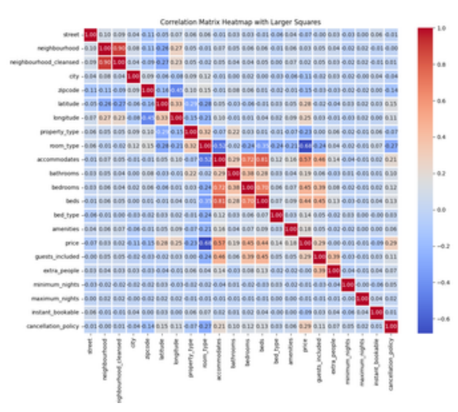
- Data encoding using LabelEncoder
- Data normalization using MinMaxScaler

Before & after data encoding



Before & after data normalization





4. Feature Selection

- Make a correlation matrix to identify relevant features
- Choose features with absolute correlation value > 0.1
- Define features (X) and target (y)

EXPERIMENT

Before we proceed to model training, we do the train-test split. After splitting, the X_train and y_train will be used to fit the models while X_test and y_test will be used to test the performance of the models.

DESCRIPTION OF AI MODELS APPLIED

Linear Regression Model

Predicts the target by finding the best-fitting line through the data and assuming a linear relationship between the features and the target.

Decision Tree Regressor Model

A non-linear model that splits the data into subsets based on the value of the features. It predicts the target by finding the mean value of the target in each subset.

Stacking Regressor Model

An ensemble model that combines multiple regression models such as Linear Regression Model and Decision Tree Regressor Model to improve its predictions.

Gradient Boosting Regressor Model

An ensemble model that can build multiple decision trees sequentially and each of them will correct the errors of the previous tree.

RESULTS

After model building, we will evaluate each of the model by analyzing their:

- Mean Absolute Value (MAE)
- Mean Suqare Error (MSE)
- Root Mean Square Error (RMSE)
- R-squared Score

Model	MAE	MSE	RMSE	R^2
Linear Regression	0.108674	0.020868	0.144456	0.597870
Decision Tree Regressor	0.112816	0.030280	0.174012	0.416485
Gradient Boosting Regressor	0.089627	0.015444	0.124275	0.702378
Stacking Regressor	0.088545	0.015113	0.122934	0.708770

DISCUSSION

From the results, Gradient Boosting Regressor Model and Stacking Regressor both have better performance than the remaining model which can be evaluated through their lower MAE, MSE, RMSE and higher R^2 score. This shows that they are ensemble models as they can learn more complex relationships.