Final Project Proposal

Zhanshuo Dai MSCS Align, 2nd Year
Hla Htoo MSCS Align, 2nd Year
Danyan Liao MSCS Align, 2nd Year

Description

Setting the right price for a rental property on Airbnb can be a complex challenge for hosts, as it directly impacts the demand for their listing. Meanwhile, guests often have limited information to gauge if a given price offers fair value. Hosts aim to maximize their earnings while staying competitive, and guests seek the best value for their money—making this balance tricky to achieve. This project addresses the need for accurate Airbnb price prediction, developing a machine learning and deep learning model to help both hosts and guests evaluate prices with minimal property information available.

Summary of the data

The dataset is based on the latest data sourced from the "Inside Airbnb" platform, covering Airbnb listings in San Francisco. This tabular dataset provides detailed information on various aspects of Airbnb listings, including their geographic location, property details, host information, and reviews. (75 features and 17013 lines of data in total). LINK We will examine correlations between features like accommodates, bedrooms, and beds to ensure the independence of each feature as much as possible. We will also use algorithms like PCA to achieve this goal. When it comes to handling missing data, different models will treat these gaps in various ways. For some models, such as linear regression, we might directly remove rows with missing values to avoid introducing bias. For others, such as tree-based models, missing values can be encoded as a separate category.

Methods

In this project, we will preprocess the Airbnb dataset and then employ a range of machine learning models to predict listing prices. Linear regression will serve as a baseline to interpret direct relationships between features and price, while Support Vector Regression (SVR) and Gradient Boosting Trees will capture more complex, nonlinear interactions. Additionally, ensemble methods like Random Forests and XGBoost will enhance predictive accuracy by reducing variance and leveraging diverse perspectives within the data, with XGBoost particularly effective for high-dimensional datasets due to its efficiency and regularization techniques. However, advanced models can introduce challenges, such as overfitting with high-dimensional data. To address this, we will perform feature selection to retain only the most predictive attributes. Hyperparameter tuning—adjusting tree depth, learning rate, and regularization—will further optimize model performance.

Preliminary results

We randomly selected 1,000 data points and selected four features: beds, bedrooms, and price. We used a linear regression model, assuming that there is a linear relationship between the features and the price. MSE: 67385.22, this value is very high, indicating that the difference between the model's predicted value and the true value is large. R-squared: 0.0752, indicates that only 7.52% of the variability in house prices can be explained by these three features. In short, this means that these three features may not be able to explain the changes in house prices well, and the linear regression model is too simple to capture the complex patterns in the data. LINK