**AirBnB Listing Price Prediction**

**Vijaykumar Mehtre**

**Bellevue University**

**DSC550-T301 Data Minin (2243-1)**

**Brett Werner**

**March 1, 2024**

# Introduction

## Introduce the problem.

We will be exploring Airbnb listings from the USA; the main goal of this Kernel will be exploring the data and predicting the price of a listing given a new sample. Our methodology in this Kernel will follow a standard analysis and prediction methodology, where we will first assess the data for any missing values followed by outlier imputation. The next stage will be the EDA

# Justify why it is important/useful to solve this problem.

Solving the problem of predicting Airbnb listing prices in the USA can be valuable and useful for several reasons:

For hosts and property owners, accurately predicting listing prices can help optimize rental income. By setting the right price based on market demand, hosts can attract more guests and maximize their earnings.

For travelers and guests, knowing the expected price of an Airbnb listing can aid in planning their trip budget. Accurate price predictions can contribute to a smoother booking process and enhance the overall user experience.

Analyzing Airbnb listing prices provides valuable insights into the dynamics of the real estate market, tourism trends, and demand fluctuations in different locations. This information can be beneficial for investors, policymakers, and researchers interested in understanding market trends and making informed decisions.

# How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

Start by outlining the growing popularity of Airbnb and the significant role it plays in the hospitality and travel industry. Emphasize the vast potential for hosts and property owners to generate income through Airbnb rentals.

Explain how accurately predicting Airbnb listing prices can directly impact revenue generation for hosts. By setting optimal prices based on market demand, hosts can increase their occupancy rates and maximize their earnings.

we can improve customer satisfaction and loyalty, leading to repeat bookings and positive reviews.

# Explain where you obtained your data.

I have sourced this US AirBnb 2020 dataset from Kaggle website.

URL : <https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data>

This dataset has one file- AB\_US\_2020.csv which has columns describing features such as host id, hostname, listing id, listing name, latitude and longitude of listing, the neighborhood, price, room type, minimum number of nights, number of reviews, last review date, reviews per month, availability, host listings and city.

# EDA and DATA Prep

We have some of missing data but column is important then we have replaced numerical features with 0 and categorical feature with None , however there are some columns which has more than 50% data is missing those columns , we dropped such columns .

As part of EDA we need to see the Numerical Feature distribution to find the outliers .

A group of graphs showing different types of distribution

Description automatically generated with medium confidence

We have seen lot of outliers like price distribution , some of listing has way more outlier in price and which is not correct and similar case with minimum nights, number of reviews etc.

Once we removed such outliers, out distribution looks better. we are left we more meaningful distributions from which we can extract some insight. We can see that the number\_of\_reveiws feature, as well as the reveiws\_pre\_month feature.

A group of graphs showing different types of data

Description automatically generated with medium confidence

# Model building and evaluation

Here I have choose 2 model i.e. Decision Tree and Random Forest Regressor. We have distributed the data as 20 % test and 80% train .

We have applied both the model on training data, and we have got the RSME’s as below.

Decision Tree RMSE on Training Data: 36.72417698192171

Decision Tree RMSE on Test Data: 36.66598513226107

Decision Tree r2 on Training Data: 0.022849325707066104

Decision Tree r2 on Test Data: 0.019823422471091368

Random Forest RMSE on Training Data: 36.663043505207874

Random Forest RMSE on Test Data: 39.30737734111008

Random Forest r2 on Training Data: 0.026099876278889544

Random Forest r2 on Test Data: 0.023248793691236558

# Analysis:

Both models have relatively similar RMSE values on the training and test data, indicating that they are performing similarly in terms of error.

The R2 values for both models are very low, indicating that the models explain only a small portion of the variance in the target variable.

The random forest model has slightly lower RMSE on the training data compared to the decision tree, suggesting that it fits the training data slightly better.

However, the random forest model's RMSE on the test data is higher than that of the decision tree, indicating potential overfitting.

Both models have very low R2 values, suggesting that they may not be capturing the underlying patterns in the data effectively.

# Recommendations:

It seems that both models are not performing well in terms of explaining the variance in the target variable.

Further analysis and improvement of feature selection, model hyperparameters, or consideration of other algorithms may be necessary to improve model performance.

Additionally, cross-validation and hyperparameter tuning could be used to optimize the models and potentially mitigate overfitting in the case of the random forest model.

It's also essential to consider the context of the problem and domain-specific knowledge to identify relevant features and improve model interpretability.