Maximize earnings for NYC Airbnb Hosts

Madhava Anuraag Dasu

Data Science – MSCS2401

Problem Statement:

Airbnb hosts need to competitively price their listings to maximize earnings while ensuring high occupancy rates. The challenge is to develop a predictive model that determines the optimal listing price based on property’s characteristics and location within New York City.

Objective:

Utilize supervised learning to predict Airbnb rental prices in New York City and uncover key factors that influence pricing strategies for hosts.

Scope of Work:

Supervised Learning: We train the predictive algorithm on an initial set of data and then we test it out on a brand-new set of data. If the training we accomplished worked well, then the algorithm should be able to predict the right outcome most of the time in the test data.

Unsupervised Learning: These models are given unlabeled data and allowed to discover patterns and insights without any guidance.

We focus on developing a supervised predictive algorithm.

Methodology:

1. Data Cleaning and Preprocessing - We prepare the data by cleaning the data, removing unnecessary variables and taking care of missing and empty values.
2. Data exploratory analysis & Data Visualization - We explore the input data to understand the structure of the data, the scope of the data. We visualize the input data to understand patterns and trends in the data.
3. Data Modeling - We prepare models to train the data and use the model to test the data by making predictions.
4. Discussion of Outcomes and Findings- We compare the predictions with actual outcomes of test data by using different metrics like MAE, MSE, and RMSE to verify how accurate our model’s prediction is.

Data Cleaning:

1. We have performed data cleaning by removing fields such as “name”, “host name”, “host id”, and “last review” as these variables are not needed to maximize price.
2. I have replaced 10052 missing values in the variable “num\_reviews” by the mean of the total reviews as we cannot just ignore that variable for analysis.
3. have converted categorical variables“neighbourhood\_group”, “neighbourhood”, “room\_type” into num by using one hot encoding.

Data Exploratory analysis and Data Visualization:

Please find some data visualizations which will help us identify which variables influence the price of the listing.

1. Airbnb Listings Color Coded by Price.

A screenshot of a computer screen

Description automatically generated

1. Box Plot of Prices by Room Type.

A screenshot of a graph

Description automatically generated

1. Box Plot of Room type vs Price vs Neighborhood.

A screenshot of a graph

Description automatically generated

1. Histogram of Price Distribution.

A graph with numbers and a bar

Description automatically generated

1. Price vs Number of Reviews by Neighborhood group.

A screen shot of a graph

Description automatically generated

Data Modeling:

Data modeling as supervised learning can be categorized into regression and classification.

Regression algorithms are used to predict continuous values such as price, salary, age etc.

Classification algorithms are used to classify the discrete data values into two categories.

Our problem statement guides us to explore regression algorithms to accurately predict the price and hence to maximize earnings.

1. Linear Regression:

Basic Regression to check linear dependency between variables and prices.

1. Lasso Regression:

Lasso Regression to check any correlation between individual variables(subset) and ignore other variables.

1. XgBoost Regression:

It is a gradient boosting algorithm where each tree is dependent on the last. Each tree corrects the errors of the last one. It has both lasso and ridge regression predictions and it ensembles weaker models to predict.

Findings:

1. Linear Regression:

A screenshot of a computer code

Description automatically generated

1. Lasso Regression:

A screenshot of a computer code

Description automatically generated

1. XgBoost Regression:

A screenshot of a computer code

Description automatically generated

XgBoost model is the better predictive model as per the MAE and RMSE between predictions and test data prices.

We found that the neighbourhood\_group, room\_type, minimum nights, availability\_365 influenced the MAE (low 60's to high 60's) more in the sense that when I removed these variables, MAE and RSME increased. The number of reviews and reviews per month did not affect my model positively.

Conclusion:

XgBoost Regression model can be used as a predictive model by Airbnb Hosts to fix the prices of their listings to maximize earnings.