Predicting Airbnb Rental Prices using Statistical Learning: A Data-Driven Approach

Aanchal Dusija, Ahmed Khair, Karan Uppal

INTRODUCTION

The Airbnb dataset used in this project was compiled by collecting quarterly updates throughout the year 2022 in a seasonal manner on specific dates: March 16th, June 8th, September 12th, and December 15th. Using a quarterly update approach to compiling the Airbnb dataset provides a rich source of data for our project. It allows us to capture seasonal variations in the data, providing a more accurate picture of the Airbnb market.

DATA SCIENCE QUESTIONS

1. What machine learning models most accurately predict price? How do the models' accuracy vary between them?
2. What are the significant variables within the Airbnb Dataset in relation to price?

METHODOLOGY

Linear Regression: Identifies predictors by fitting a line to data.

Polynomial Regression: Identifies non-linear relationships between predictors and response variables.

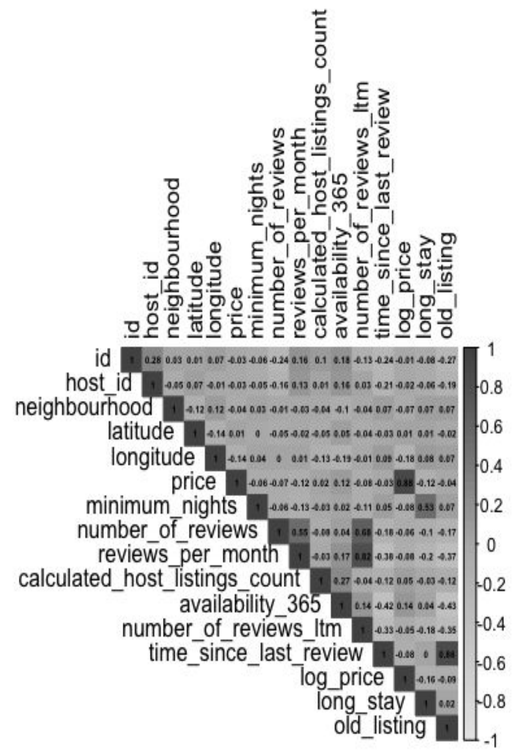
Spline Regression: Uses a set of knots to define the basis function for regression.

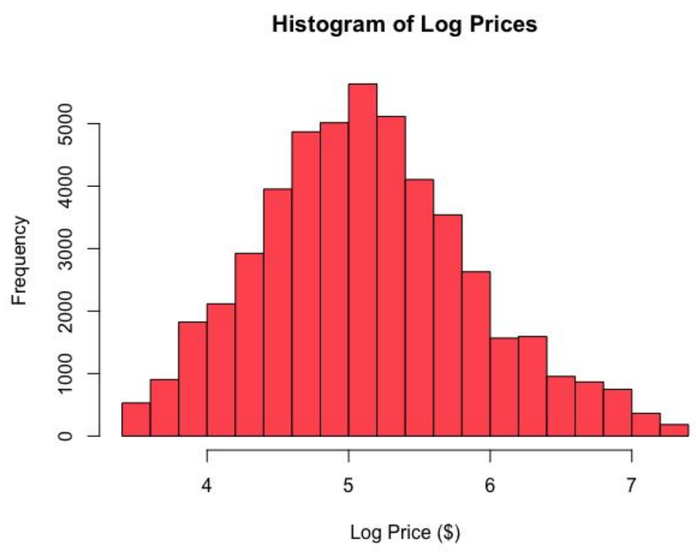
Bagging: Used to deal with bias-variance trade-offs and reduces the variance of a prediction model

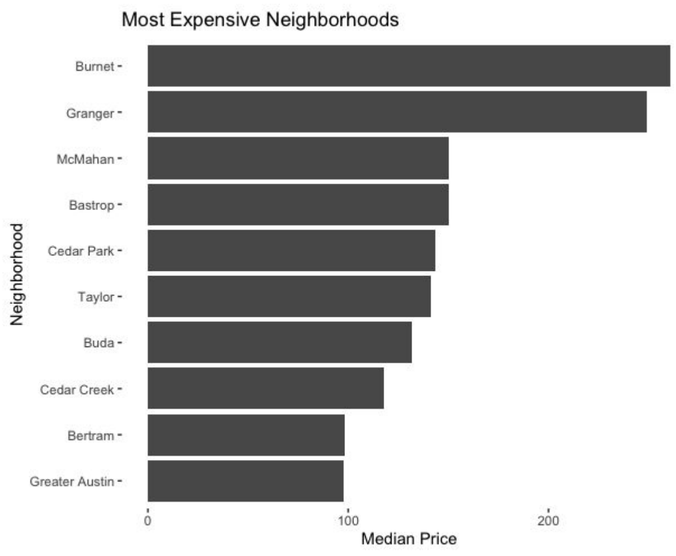
XGBoost: Uses regularization techniques to prevent overfitting and identify important features.

Random Forest: Uses decision trees to predict outcomes and takes the average of multiple trees.

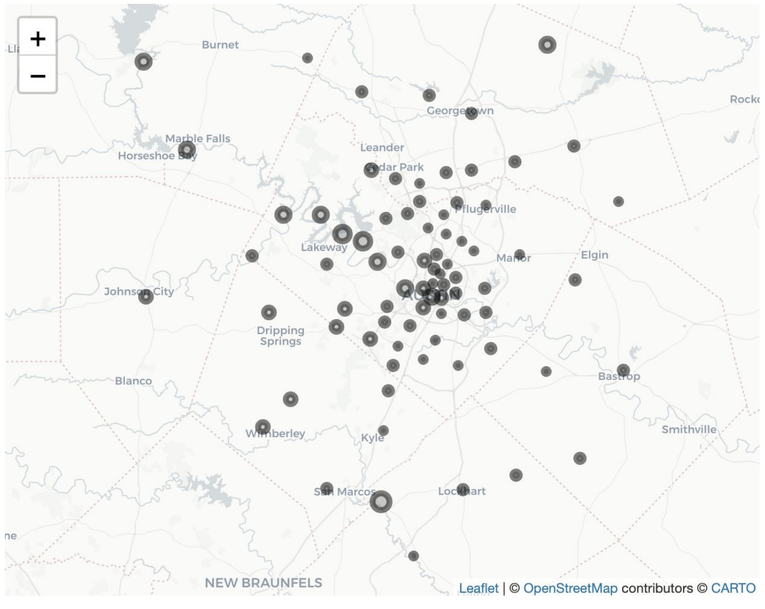
EXPLORATORY DATA ANALYSIS







AIRBNB LISTINGS FOR AUSTIN, TX



ANALYSIS & RESULTS

Statistical Model	R-square	RMSE
Linear Regression	0.31	0.63
Polynomial Regression	0.06	0.63
Spline Regression	0.09	0.72
Bagging	0.30	0.65
XGBoost	0.36	0.60
Random Forest	**0.66**	**0.52**

CONCLUSIONS

Random Forest is the best-performing model with the highest R-square and lowest RMSE values, followed by XGBoost. Linear regression and bagging models have moderate R-square but higher RMSE values, while polynomial and spline regression have lower R-square values with the same RMSE as linear regression. Therefore, Random Forest is the recommended model for the given dataset. From the Random Forest Model, we get that the Room type, Number of Reviews and Listing are the most important variables to determine the Price of the Airbnb Property.

FUTURE WORK

To enhance the accuracy of predicting Airbnb rental prices, we can incorporate additional models like neural networks, which can handle complex data relationships. Adding more data from various dates can help capture seasonal variations in prices, and including economic data, such as demand and supply, can help account for factors that affect prices. By incorporating these elements, the model's accuracy can be increased, enabling more informed pricing decisions for hosts and improved guest experience.

REFERENCES

Inside Airbnb. (n.d.). *Get the Data*. Inside Airbnb. Retrieved April 25, 2023, from http://insideairbnb.com/get-the-data/

Gibbs, C., Guttentag, D., Gretzel, U., Yao, L., & Morton, J. (2018, January 8). *Use of dynamic pricing strategies by Airbnb hosts*. International Journal of Contemporary Hospitality Management. Retrieved April 25, 2023, from https://www.emerald.com/insight/content/doi/10.1108/IJCHM-09-2016-0540/full/html?skipTracking=true

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Resources - ISLR second edition*. An Introduction to Statistical Learning. Retrieved April 25, 2023, from https://www.statlearning.com/resources-second-edition