University of Essex
MSc Artificial Intelligence
Module: ML_PCOM7E July 2025 B
Unit 6: Development Team Project – Airbnb Rental Price Model

Investigative Report

I. Executive Summary

Based on the findings from a tuned XGBoost model applied to Airbnb rental prices in the New York City area, we propose that Airbnb implement a software-based agent to alert rental hosts when their prices deviate significantly from those found most likely to encourage the highest average yearly occupancy rates.

II. Investigative Purpose

The task at hand was to evaluate the data provided by the 2019 “New York City Airbnb Open Data” collection to identify predictive trends in rental price points relevant to Airbnb’s continued growth within said market (Dgomonov, 2019; Inside Airbnb, 2024).

III. Investigative Steps

A. Dataset Retrieval & Analytical Platform

The sample dataset was retrieved from Airbnb’s open access platform (Dgomonov, 2019); Inside Airbnb, 2024) and loaded into an industry-standard Google Colab Jupyter cloud workspace (Google Colab, 2025).

B. Preliminary Data Processing

An initial scan of the data revealed several opportunities to prepare the data for analysis. In accordance with industry-standard best practices (Burkov, 2019; Côté et al, 2024), the data was processed through the following steps:

1. Initial Cleaning

a. Removing irrelevant fields (Burkov, 2019).
b. Replacing missing values with zero and removing listings when no placeholder values could be applied; for example, for properties listed as free or without a stated price (Burkov, 2019).

2. **Data Splitting**

    a. The cleaned dataset was split into a training set (80%) and a testing set (20%). The test set remained untouched during model training to provide an unbiased evaluation of performance (Burkov, 2019).

3. **Data Preprocessing**

    a. Converting categorical variables, such as room type, into numerical format using one-hot encoding (Burkov, 2019).
    b. The target variable, price, was log-transformed to normalise its distribution and to minimise skewing (Burkov, 2019).
    c. A robust scaler was fitted only on the training data's numerical fields to minimise the influence of outliers; the fitted scaler was then applied to the training and testing sets to prevent data leakage between them (Burkov, 2019; Côté et al, 2024).

C. Preliminary Regression and Correlation Analysis

The following standard regression and correlative analyses were undertaken (Burkov, 2019):

1. Overall price distribution, linear and logarithmic.
2. Price by room type, linear.
3. Average price by borough, linear.
4. Price versus number of monthly reviews, linear.
5. Rental unit availability in days per year, linear.
6. Correlative heatmaps of analyzed fields.

This process established points of comparison with the results of subsequent random forest and XGBoost modeling (Arnold et al, 2024), both of which are established methods for modeling real estate prices (Mathotaarachchi, Hasan and Mahmood, 2024).

D. Untuned & Tuned Random Forest Models

Based on the preceding, untuned and tuned (via the “RandomizedSearchCV” setting for iterative random parameter combinations) random forest models were trained. Hyperparameter tuning was undertaken as an industry-establish best practice for increasing the accuracy of findings over baseline, untuned random forest models (Arnold et al, 2024).

E. Untuned & Tuned XGBoost Models

Similarly, untuned and tuned (via the “GridSearchCV” setting for testing every combination of parameters) XGBoost models were trained. Similarly, untuned and

tuned XGBoost models were trained. For this model, “GridSearchCV” was used to exhaustively test every combination of a focused set of key parameters. The optimal hyperparameters were identified as: learning_rate: 0.05, max_depth: 7, and n_estimators: 500.

This is a computationally expensive, but worthwhile, investigatory step as the performance of XGBoost has been shown to be sensitive to the choice of its hyperparameters (Probst, Boulesteix and Bischl, 2019). To efficiently perform both hyperparameter searches and obtain less biased estimates of model performance, 3-fold cross-validation (k=3) was employed for model performance evaluation (Probst, Boulesteix and Bischl, 2019).

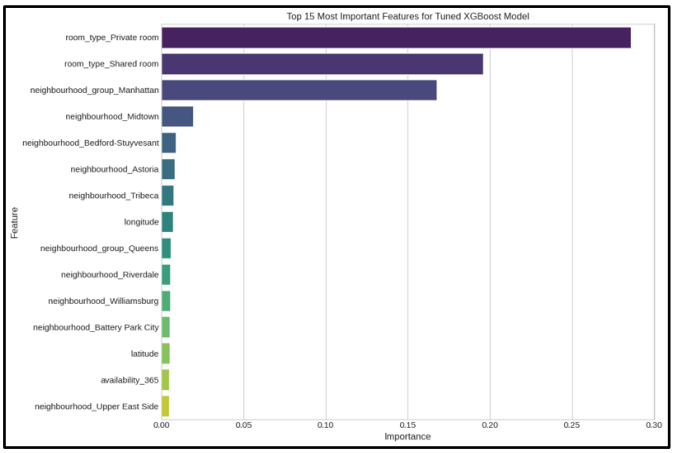
IV. Investigative Findings

Root mean squared log error (“RMSLE”) and R-squared (“R2”) scores were obtained and compared across all models evaluated:

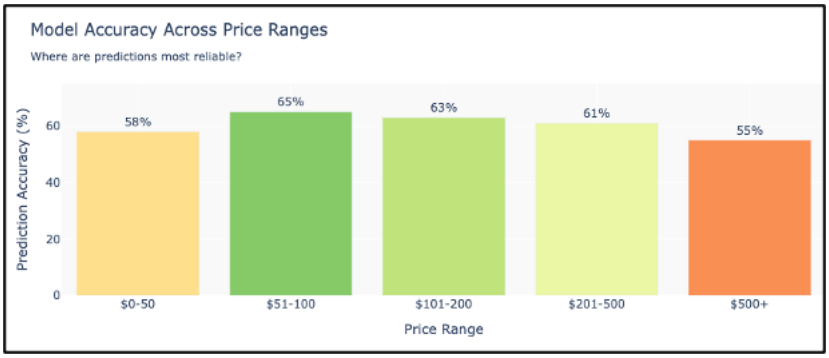
Model	RMSLE	R2 Score
Tuned XGBoost	0.4222	0.6231
Untuned XGBoost	0.4263	0.6158
Tuned Random Forest	0.4275	0.6138
Untuned Random Forest	0.4291	0.6108
Linear Regression	0.4598	0.5530

The tuned XGBoost model outperformed the baseline, yielding the lowest RMSLE and highest R-squared. The model explains 62.31% of the price variance, an approximately 7.00% percentage point increase over the 55.30% explained by the linear regression. This gain in predictive power suggests the model is a candidate for deployment (Probst, Boulesteix and Bischl, 2019).

The fifteen selected fields were then ranked by each’s influence on price:



Private rooms were found to be more attractive to consumers (40.20% relative importance) over shared rooms (27.50% relative importance) and therefore commanded higher prices. The city borough of Manhattan, in particular, exerted an outsized influence on price (23.60% relative importance). The accuracy of the tuned XGBoost model was then examined by standard price ranges per night:



The model was found to be the most reliable for low and average-priced properties (between $51.00 and $500.00 per night), but significantly less reliable for ultra-low budget properties (below $50.00 per night) and the highest-end luxury properties (above $500.00 per night).

V. Executive Recommendation: A Semi-Autonomous Pricing Agent

The use of software-based pricing agents is a long-established practice dating back at least several decades (Kephart, Hanson and Greenwald, 2000), and one that remains applicable to a wide range of industries today, including the real estate market (Mathotaarachchi, Hasan and Mahmood, 2024).

Consistent with the Industry 5.0 precept of using automation to enhance human decision-making (Sümer et al, 2025), we therefore propose that a semi-automated software agent be built to accomplish two automated tasks:

1. To monitor rental unit prices and occupancy rates in real time, and
2. To alert hosts when their rental prices deviate to a statistically-significant degree from optimal prices.

Given that room type (private versus shared) is a confounding factor, we recommend that said agent distinguish alerts based on room type for greater accuracy. This agent will also be semi-autonomous in that the majority of its work will be automated; once confirmed to work to an acceptable accuracy level, human intervention will only occur as material changes occur in the New York City rental market.

This proposal aligns the interest of Airbnb hosts maximizing their average yearly occupancy rates with Airbnb's broader, longer-term interest in growing its presence in the New York City short term rental market.

Finally, it is worthwhile to note that to accomplish said goal, additional investigation is required to identify those hyperparameters that may be further

refined to increase the accuracy of the XGBoost model used.

References

Arnold, C., Biedebach, L., Küpfer, A. and Neunhoeffer, M. (2024). The role of hyperparameters in machine learning models and how to tune them. *Political Science Research and Methods*, [online] pp.1–8. doi:https://doi.org/10.1017/psrm.2023.61.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.

Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D. and Khomh, F. (2024). Data cleaning and machine learning: a systematic literature review. *Automated software engineering*, 31(2). doi:https://doi.org/10.1007/s10515-024-00453-w.

Dgomonov (2019). *New York City Airbnb Open Data*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data.

Google Colab (2025). *Google Colab*. [online] Available at: https://colab.research.google.com/drive/1zQafvQNaXf72yH9QxcmEtKz42S5HFQ0t?usp=sharing.

Inside Airbnb (2024). *Get the Data*. [online] insideairbnb.com. Available at: https://insideairbnb.com/get-the-data.

Kephart, J.O., Hanson, J.E. and Greenwald, A.R. (2000). Dynamic pricing by software agents. *Computer Networks*, 32(6), pp.731–752. doi:https://doi.org/10.1016/s1389-1286(00)00026-8.

Mathotaarachchi, K.V., Hasan, R. and Mahmood, S. (2024). Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information*, [online] 15(6), pp.1–35. doi:https://doi.org/10.3390/info15060295.

Probst, P., Boulesteix, A.-L. and Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, [online] 20(53), pp.1–32. Available at: https://www.jmlr.org/papers/v20/18-444.html.

Sümer, M., Özsoy, T., Okay, F.Y. and Kök, I. (2025). Smart Agents as Customers: A Semi-Autonomous Digital Commerce Model for Industry 5.0. *IEEE*, [online] pp.1–7. doi:https://doi.org/10.1109/ichora65333.2025.11017052.