Predicting Airbnb Prices in New York City Using Machine Learning

ITCS 3156 Final Project Report

Nathaniel Joseph

Introduction

a. Problem Statement

This project focuses on predicting the nightly price of Airbnb listings in New York City. The goal is to see if machine learning can look at details about a listing, such as its neighborhood or room type, and make a good guess about what the price should be.

b. Motivation and Challenges

Airbnb prices change a lot in New York City. Some listings cost under 100 dollars while others cost hundreds or even thousands. It is interesting and useful to understand what affects these prices. This type of problem is challenging because the data is large and messy. Some listings have missing information, some have very high prices that do not fit the normal pattern, and many features are text and need to be converted into numbers.

c. Approach Summary

I used the New York City Airbnb Open Data. I explored the data with simple charts. Then I cleaned the data, handled missing values, and converted text features into numeric form so that machine learning models could use them. I trained two models, Linear Regression and Random Forest, and compared how well they predicted the listing prices.

Data

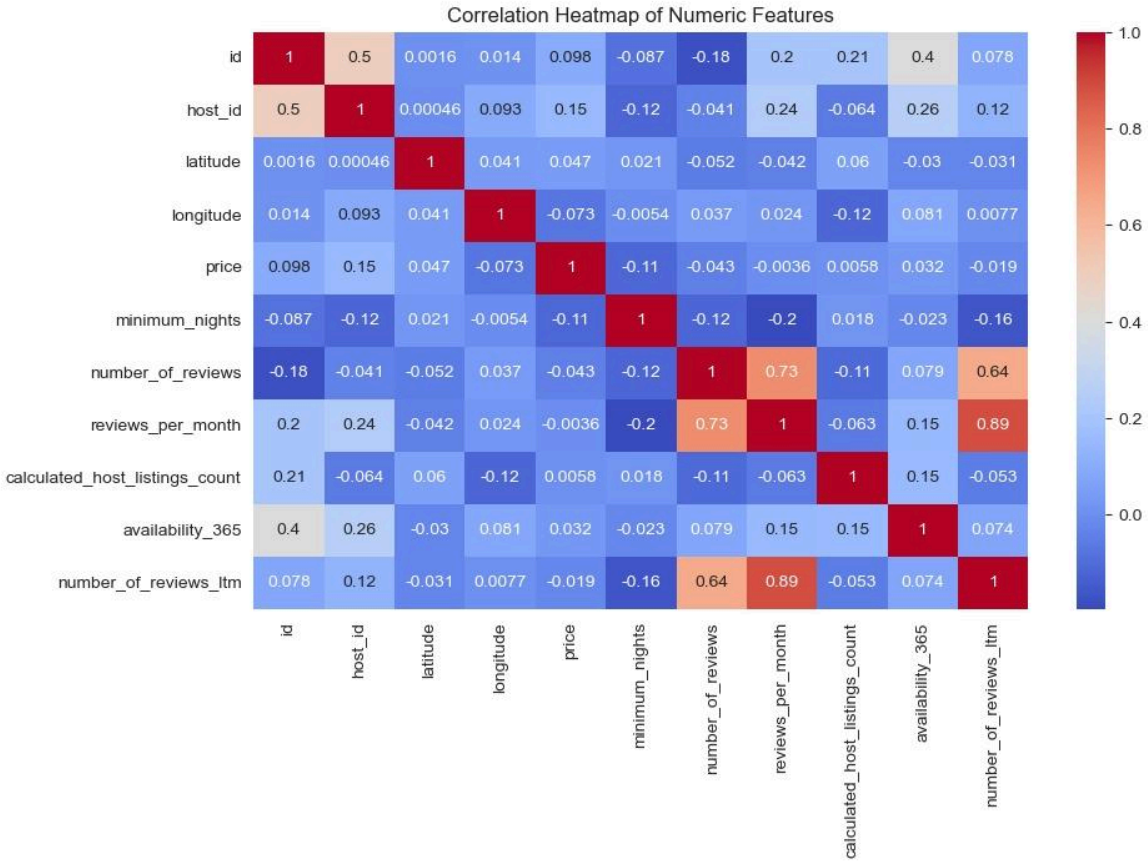
Data Introduction

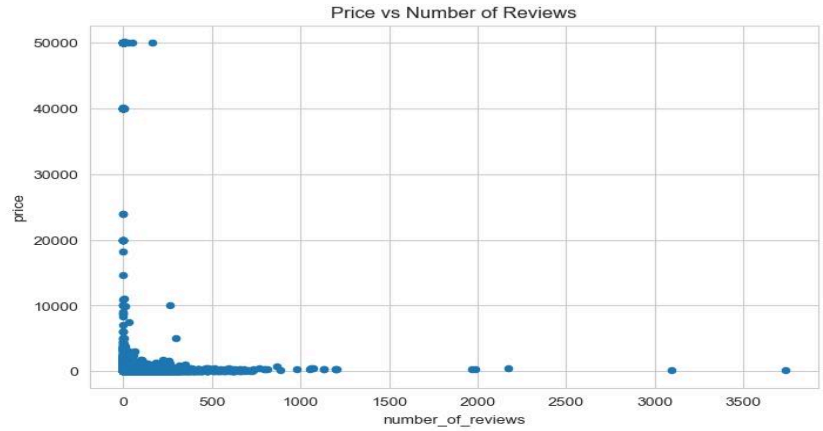
The dataset comes from the NYC Airbnb Open Data collection. It has more than forty thousand listings, which is large enough for a machine learning project. Each listing includes many features such as neighborhood, room type, minimum nights, number of reviews, reviews per month, and availability throughout the year.

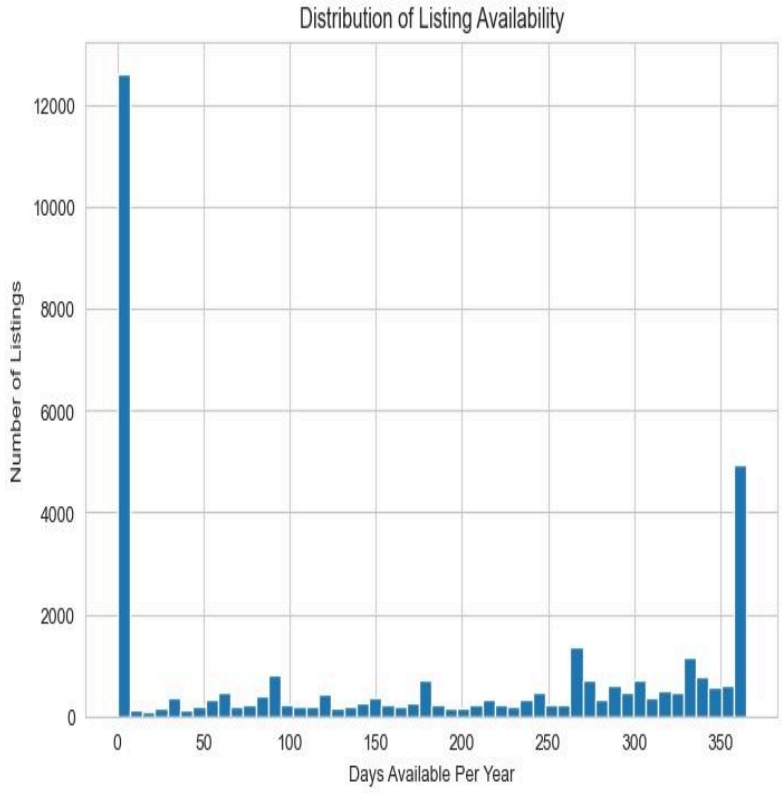
Basic Visual Analysis

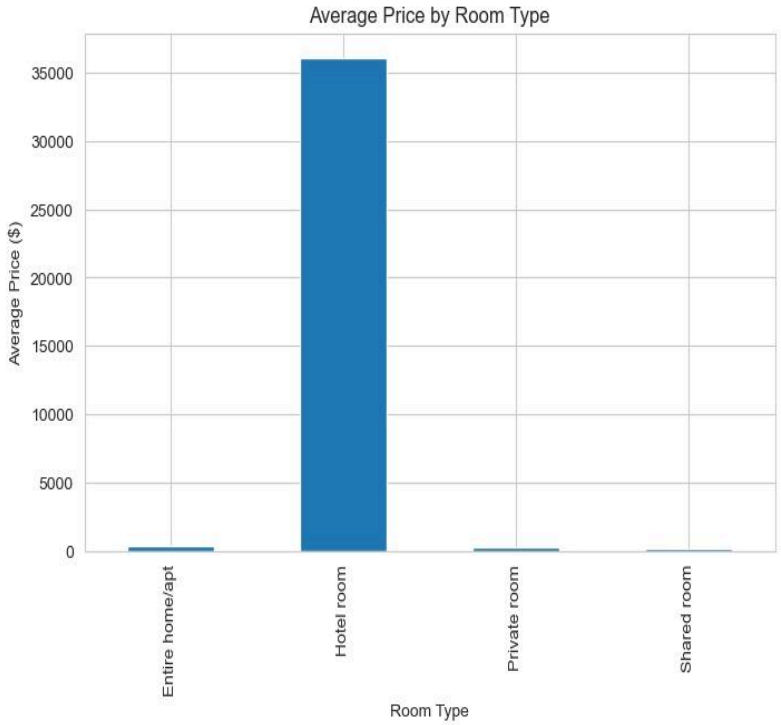
I explored the data with several simple visualizations.

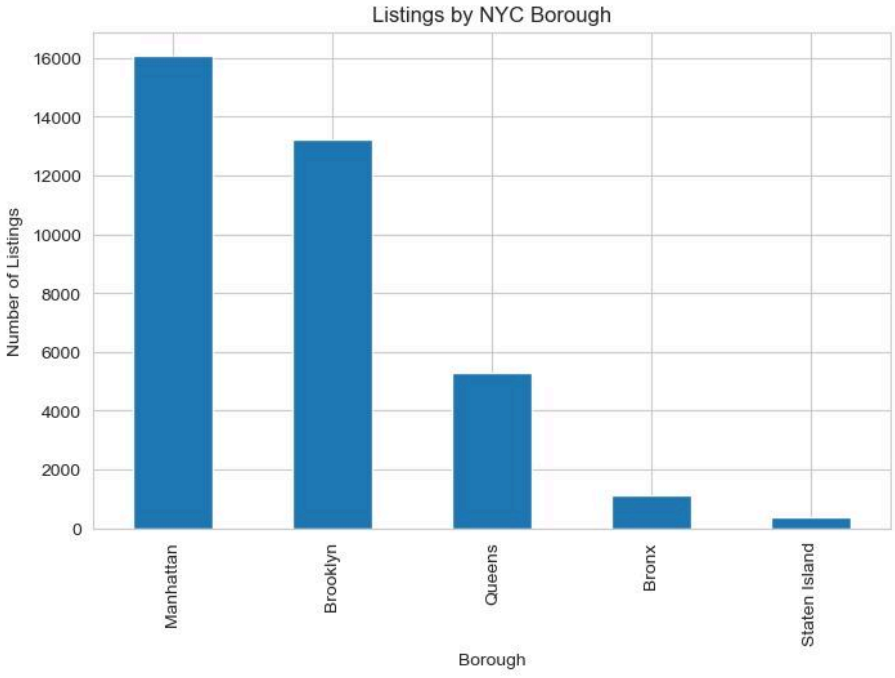

Distribution of Airbnb Prices in NYC
Number of Listings
Price ($)
0
2500
5000
7500
10000
12500
15000
17500
20000
0
10000
20000
30000
40000
50000


Correlation Heatmap of Numeric Features
id 1 0.5 0.0016 0.014 0.098 -0.087 -0.18 0.2 0.21 0.4 0.078
host_id 0.5 1 0.00046 0.093 0.15 -0.12 -0.041 0.24 -0.064 0.26 0.12
latitude 0.0016 0.00046 1 0.041 0.047 0.021 -0.052 -0.042 0.06 -0.03 -0.031
longitude 0.014 0.093 0.041 1 -0.073 -0.0054 0.037 0.024 -0.12 0.081 0.0077
price 0.098 0.15 0.047 -0.073 1 -0.11 -0.043 -0.0036 0.0058 0.032 -0.019
minimum_nights -0.087 -0.12 0.021 -0.0054 -0.11 1 -0.12 -0.2 0.018 -0.023 -0.16
number_of_reviews -0.18 -0.041 -0.052 0.037 -0.043 -0.12 1 0.73 -0.11 0.079 0.64
reviews_per_month 0.2 0.24 -0.042 0.024 -0.0036 -0.2 0.73 1 -0.063 0.15 0.89
calculated_host_listings_count 0.21 -0.064 0.06 -0.12 0.0058 0.018 -0.11 -0.063 1 0.15 -0.053
availability_365 0.4 0.26 -0.03 0.081 0.032 -0.023 0.079 0.15 0.15 1 0.074
number_of_reviews_ltm 0.078 0.12 -0.031 0.0077 -0.019 -0.16 0.64 0.89 -0.053 0.074 1
id
host_id
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
number_of_reviews_ltm
1.0
0.8
0.6
0.4
0.2
0.0


Price vs Number of Reviews
price
number_of_reviews
50000
40000
30000
20000
10000
0
0
500
1000
1500
2000
2500
3000
3500


Distribution of Listing Availability
Number of Listings
Days Available Per Year
12000
10000
8000
6000
4000
2000
0
0
50
100
150
200
250
300
350





1. **Price distribution**

A histogram showed that most Airbnb listings cost less than 300 dollars per night, but

there are some very expensive listings that create a long tail.

2. **Average price by neighborhood group**

   A bar chart showed that Manhattan and Brooklyn have the highest prices on average.

3. **Room type distribution**

   A chart showed that entire homes are the most expensive. Private rooms and shared rooms are usually cheaper.

4. **Correlation heatmap**

   This showed how different numerical features relate to each other. Some features had weak connections while others had moderate ones.

Data Analysis Observations

During analysis, I noticed that prices change a lot depending on the neighborhood. Entire homes are much more expensive than private rooms. Listings with many reviews might be more popular, but they do not always cost more. There are also some unusual listings priced extremely high. These could make the model less accurate.

Data Preprocessing

To prepare the data for the models, I did the following:

1. Removed rows with missing values.

   Removed extreme price outliers, such as listings above one thousand dollars.
2. Converted text features, like neighborhood and room type, into numeric form using one hot encoding.
3. Scaled the numeric features when needed.
4. Split the data into training and testing sets.

Methods

I used two different machine learning models.

Linear Regression

Linear Regression is a simple model that tries to find a straight line relationship between the features and the target price. It is easy to understand and makes a good starting point for comparison. The downside is that it does not work well when the data has complex or curved patterns.

Random Forest Regressor

Random Forest uses many decision trees and averages their results. It can handle more complicated relationships in the data and is more flexible than Linear Regression. It is also more resistant to noise and outliers. I chose this model because Airbnb prices do not follow a simple pattern, and a stronger model is needed to capture the variation.

Results

Experimental Setup

I split the data so that eighty percent was used for training the models and twenty percent was used for testing them. I used three evaluation metrics which were Mean Absolute Error, Mean Squared Error, and the R squared score. These metrics show how far the predictions are from the real prices.

Test Results and Observations

Linear Regression performed fairly but not great. It had trouble predicting prices in neighborhoods with very high or very low values. It also did not work as well with the categorical features even after encoding.

Random Forest performed better than Linear Regression. It predicted prices more accurately and handled the messy parts of the data more effectively. It also showed which features were the most important for price prediction, such as neighborhood group and room type.

Analysis of Results

The better performance of Random Forest makes sense because Airbnb data does not follow a straight line pattern. The price depends on many different factors that interact with each other. Random Forest was able to capture these patterns while Linear Regression was not.

Supporting Experiments

I also tried removing outliers before training. This improved both models, but Random Forest still remained the top performer. I tested the models with fewer features and found that removing key features like room type made the accuracy worse. This confirmed that these features are very important for predicting price.

Conclusion

This project showed that machine learning can be used to predict Airbnb prices in New York City. Random Forest gave the best performance and was able to handle the variety and complexity in the data. Through this project, I learned how important data cleaning and preprocessing are. I also learned how different models work and why some models fit certain problems better than others. One of the main challenges was dealing with outliers and converting text features into usable numeric data. However, following each step carefully made the process easier to manage.

References (MLA Style)

New York City Airbnb Open Data. Kaggle, www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data.

Pedregosa, Fabian, et al. Scikit Learn Machine Learning in Python. Journal of Machine Learning Research, vol. 12, 2011, pp. 2825 to 2830.

Random Forests. Scikit Learn Documentation, scikit-learn.org/stable/modules/ensemble.html.

Linear Regression. Scikit Learn Documentation, scikit-learn.org/stable/modules/linear_model.html.

Acknowledgement

I used ChatGPT as a support tool while working on the MLA formatting for this project. I did not copy anything directly from the AI. I only used it to clear up general questions I had about citations. For example, I asked if the order of entries matters in an MLA Works Cited page, and it explained that they should be listed alphabetically. This guidance helped me understand the formatting rules, but all writing and organization in the report were done by me.

Source Code

GitHub Repository Link

https://github.com/njoseph8/airbnb-nyc-ml-project.git