Predicting prices for NYC Airbnb listings

Tejal Deshmukh

Introduction

Airbnb is a company that provides an internet marketplace for short-term home and apartment rentals. It allows you to rent out your home for a period of time while you're away, or rent out your spare bedroom to travelers. They generate revenue by charging a commission % of booking value from hosts & guests. http://insideairbnb.com/get-the-data/ collects a lot of information about Airbnb property listings such as property type, number of rooms, the amenities, neighborhood, customer reviews etc. contains this information for properties across the world. New York City has an active Airbnb market, and for our project, we study a dataset that contains listings in the NYC area.

The goal - This project aims to predict the price (avg nightly price a property charges) of airbnb listings given a number of features using regression models.

The pricing of the Airbnb listings is a very important aspect for both hosts and guests. For hosts, pricing their listings correctly is very important to get the maximum occupancy and income. On the other hand, guests always look for the most competitive prices while they search for the listings. Therefore, it is important to have a model that can assist hosts price their listings competitively. Predicting prices could also help potential hosts understand how much revenue they could generate by investing in and listing a property on Airbnb. In this project, we will attempt to build a couple of models to predict the prices of Airbnb listings in New York City. We intend to train the model using data from existing Airbnb listings.

Implementation

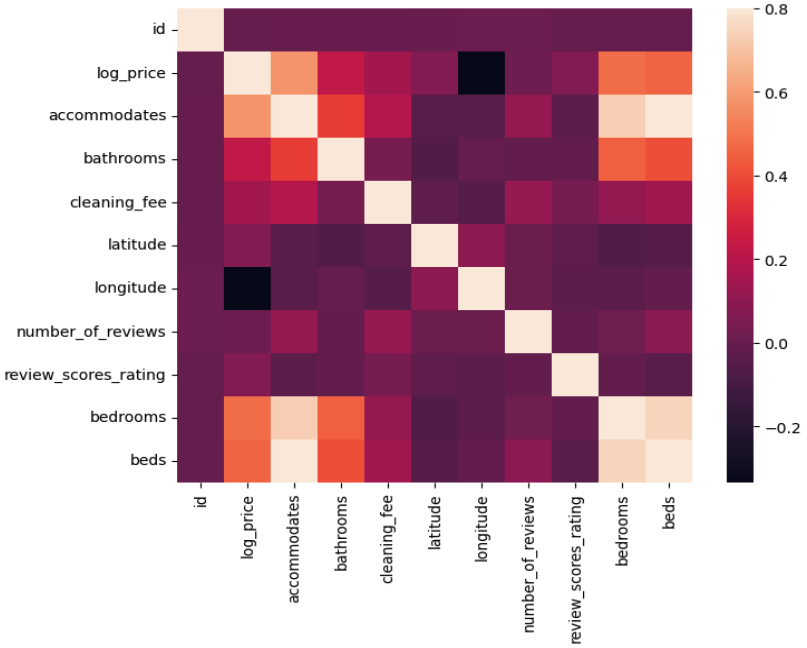
The project has been implemented by following the steps mentioned below :

1. Data Collection:
The data set has been picked from http://insideairbnb.com. The original data set includes over 40 features and 70,000 records. Features from the data set can be broadly categorized into:
1.Location related - such as neighborhood, city, latitude, longitude
2.Property related - such as property, room types, no. of people accommodates, beds, baths
3.Booking, reviews related - such as availability 365, last reviewed date, no. of reviews, score
4.Host related - such as host name, profile, image url, no. of listings, identity verified
4.Amenities - such as kitchen, dishes, parking, Wifi, TV, pets friendly, security

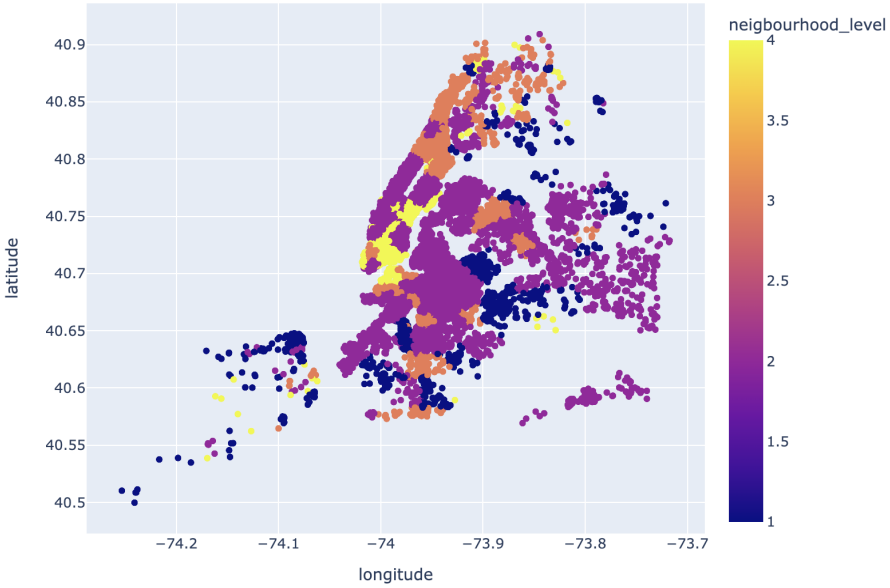
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
id	property_typ	room_type	amenities	accommodat	bathrooms	bed_type	cancellation_	cleaning_fee	description	latitude	longitude	name	neighbourho	number_of_	review_scores_rating		zipcode	bedrooms	beds
6901257	Apartment	Entire home/	{"Wireless Ir	3	1	Real Bed	strict	TRUE	Beautiful, su	40.6965236	-73.991617	Beautiful bro	Brooklyn Hei	2	100		11201	1	1
6304928	Apartment	Entire home/	{"Wireless Ir	7	1	Real Bed	strict	TRUE	Enjoy travelli	40.7661154	-73.98904	Superb 3BR /	Hell's Kitche	6	93		10019	3	3
7919400	Apartment	Entire home/	{TV,"Cable T	5	1	Real Bed	moderate	TRUE	The Oasis co	40.80811	-73.943756	The Garden (	Harlem	10	92		10027	1	3
13418779	House	Entire home/	{TV,"Cable T	4	1	Real Bed	flexible	TRUE	This light-fill	37.7720045	-122.43162	Beautiful Fla	Lower Haigh	0			94117	2	2
3808709	Apartment	Entire home/	{TV,Internet,	2	1	Real Bed	moderate	TRUE	Cool, cozy, a	38.9256269	-77.034596	Great studio	Columbia He	4	40		20009	0	1
12422935	Apartment	Private room	{TV,"Wireles	2	1	Real Bed	strict	TRUE	Beautiful pri	37.753164	-122.42953	Comfort Suit	Noe Valley	3	100		94131	1	1
11825529	Apartment	Entire home/	{TV,Internet,	3	1	Real Bed	moderate	TRUE	Warm and c	33.9804544	-118.46282	Beach Town	Studio and Pa	15	97		90292	1	1
13971273	Condominiur	Entire home/	{TV,"Cable T	2	1	Real Bed	moderate	TRUE	Arguably the	34.0467374	-118.26044	Near LA Live	Downtown	9	93		90015	1	1
180792	House	Private room	{TV,"Cable T	2	1	Real Bed	moderate	TRUE	Garden Stud	37.781128	-122.5011	Cozy Garden	Richmond Di	159	99		94121	1	1
5385260	House	Private room	{"Wireless Ir	2	1	Real Bed	moderate	TRUE	Quiet comm	33.992563	-117.896	No.7 Queen Size Cozy Roo		2	90		91748	1	1
5578513	Apartment	Private room	{Internet,"W	2	1	Real Bed	strict	TRUE	This is a brig	40.7238833	-73.98388	Large East V	Alphabet City	82	93		10009	1	1
17423675	House	Entire home/	{TV,"Cable T	4	1.5	Real Bed	strict	TRUE	A 1044 sq. ft	33.8758624	-118.40329	Sand Section	Hermosa Bea	29	97		90254	2	2
14066228	Apartment	Private room	{TV,Internet,	2	1	Real Bed	flexible	TRUE	Newly furnis	33.8132284	-118.38943	Beach Pad 1	Torrance	0			90277	1	1
2658946	Apartment	Entire home/	{TV,"Cable T	6	1.5	Real Bed	strict	TRUE	Amazing loca	38.91963	-77.031189	Charming 2 b	U Street Cor	13	89		20009	2	3
583490	Apartment	Entire home/	{Kitchen,Hea	2	1	Real Bed	strict	TRUE	This apartme	33.7785265	-118.14593	VINTAGE 1930s Mediterar		2	100		90804	1	1
6226658	Apartment	Private room	{Internet,"W	2	1.5	Real Bed	moderate	TRUE	Just west of	41.9082402	-87.695242	Sweet Home	Humboldt Pa	0			60622	1	1

2. Exploratory Data Analysis:

Although the dataset contains a large number of features, not all will be useful in predicting prices. To help us select features, we did an initial EDA. We checked the unique values of the selected features, the distributions, data types, means and value counts of these features to help us understand how to process them and how to handle categoricals that had too many categories with low items. These could be regrouped into broader ranges. For example, there were over 15 different property types, but based on the counts, they could be regrouped into 5 broad categories. We generated a heat-map of the correlation between the features and selected those which had a positive correlation with price. At the end of this step, we mostly retained location, property and amenity related features for listings in the NYC area. This reduced our dataset to about 25 features.



Correlation between features



Converting neighborhoods to groups based on avg price per room

3. Pre-processing and Data Cleaning:

Before feeding the data as input to the Machine Learning models, it needs to be pre-processed and cleaned to eliminate errors or irrelevant outliers. This step is important because it helps to ensure that the model is not biased by such erroneous data. This involved filling in some nulls and missing values with 0s or means depending on the feature. We also encoded categorical variables using get_dummies(). For some models before feeding the data, normalization of features has been performed using StandardScalar() function from Scikit learn's preprocessing package. We then split the data into training and test sets in a 80/20 ratio.

4. Feature Engineering:

What features will be useful in predicting the listing price ?

Though the dataset consists of many features, we did some feature engineering to make more meaningful use of some existing features to help in efficient price prediction. Some of the tasks we did:

1.Use the latitude & longitude to find the proximity to important locations such as times square, wall street and central station. This new feature was then appended to the original dataset.

2.Calculate avg price per room for different neighborhoods to group them into 4 classes based on how expensive this turned out to be.

3.Fetch a count of all amenities, take a count of some of most frequently sought after amenities, group them into classes such as kitchen, electric, security, etc and encode them.

4.We also had the descriptions hosts used for the listings and tried to understand if there is a correlation between this and the price.

Sentiment analysis & price prediction

Sentiment analysis is the process of detecting positives or negatives sentiment in text. We tried to understand if the property listings with positive sentiments had higher prices. We implemented sentiment analysis on the 'description' column which provides a description of Airbnb by host. The library we used here is 'nltk' which is used to manipulate and analyze linguistic data. We generated sentiment scores and categorized them into positive, negative and neutral values. We tried to encode this and use it as a feature to predict price but most of the descriptions turned out to be neutral and did not have high correlation with the prices

5. Model Training and Evaluation

For the regression task of predicting the price of property listings, we have implemented some models from the sklearn library but mainly studied these 4 - Linear Regression (with a lasso model that provided some improvement), GradientBoosting Regression (GBR) and eXtremeGradientBoosting Regression (XGBR) Model. We also tried to implement a neural network, but this did not perform better than the previously mentioned models.

Since we have implemented regression models, we used metrics such as R-squared (R2) score, Mean Absolute Error, Mean Squared Error (MSE) to evaluate our models.

Linear Regression Model

We used the basic linear regression model as a baseline and to study feature importance by looking at the resultant coefficients after fitting the model. The results were quite consistent with initial analysis, with features such as no. of people accommodated, beds, bath, room type turning out to be important. We tried to improve this with a lasso model that implements regularization for a few different values of alpha. We ran LassoCV to find the best alpha.
The best results obtained for both the models have been mentioned in below table:

	Train		Test	
	MSE	**r2**	**MSE**	**r2**
Linear Regression	0.2237	0.5745	0.2335	0.5657
Lasso Regression	0.1617	0.6447	0.1655	0.6288

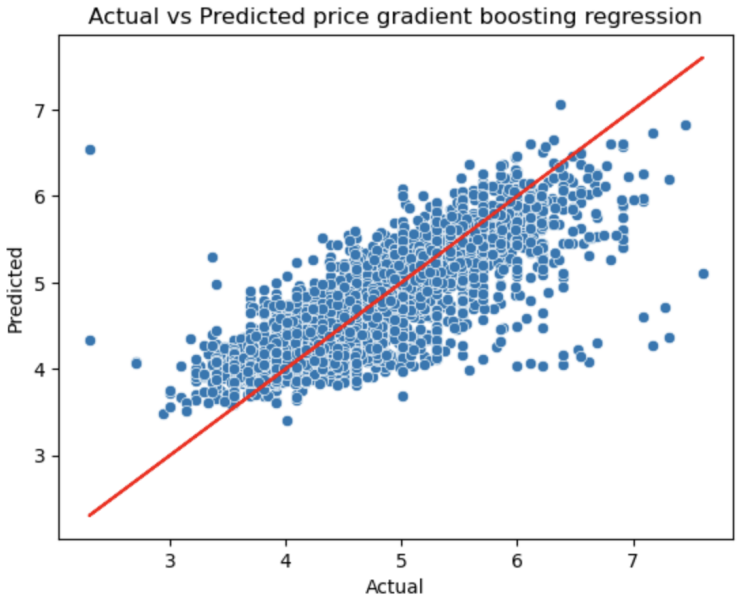
Boosting Regression Models

1) Gradient Boosting Regression (GBR) Model

Gradient boosting is an example of a boosting algorithm which is one of the variants of an ensemble method. It produces a prediction model by using multiple weak prediction models, typically decision trees and combines them to get better performance as a whole. Since we have a regression problem we have used the (GBR) Model to make predictions for our target variable price.

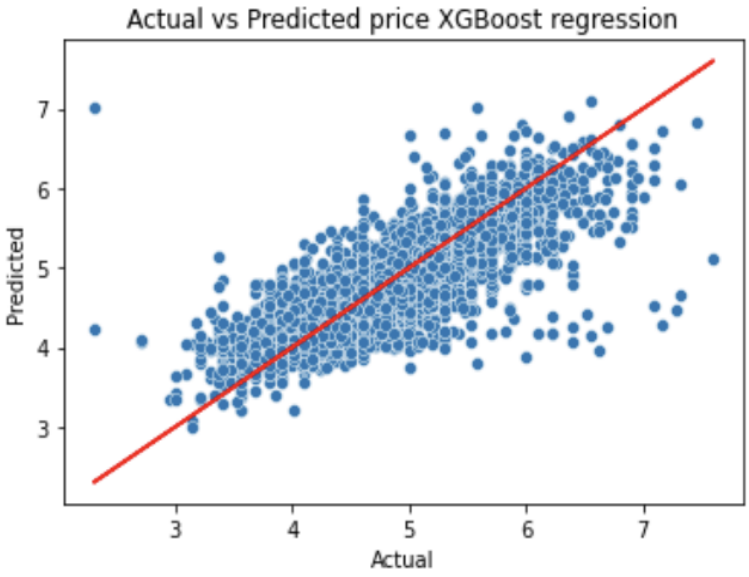
Below table highlights the evaluation results obtained by (GBR) Model

Gradient Boosting Regression (GBR)	MSE	R2
Training set	0.14	0.76
Test set	0.17	0.72



2) eXtreme Gradient Boost Regression (XGBoost) Model

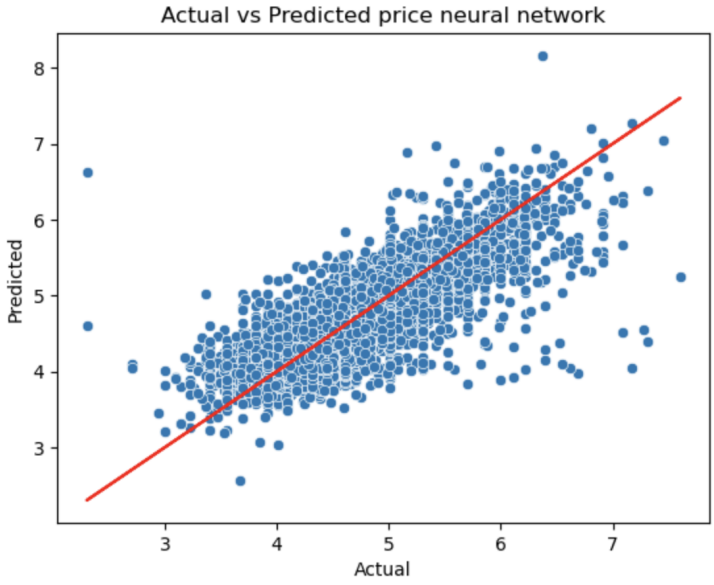
We also implemented the eXtreme Gradient Boost Regression (XGBoost) model which is an extension to the (GBR) model. XGBoost is a scalable and highly accurate implementation of gradient boosting. Since in XGBoost, trees are built in parallel, instead of sequentially like in (GBR) modeling, the speed and accuracy of our price prediction model has been improved.
K-fold cross validation has been performed on the (XGBoost) model to further tune the performance. From a range of 1-10, we decided on K=10 folds depending on the size of our dataset as well the accuracy produced by the model when K=10.



Model Results : The XGBoost Regressor model produces R2 of 0.78 and MSE of 0.20

3) Neural network

With our limited experience, we first tried a relatively shallow three layer NN with densely-connected layers, using a relu activation function for the hidden layers and a linear activation function for the output layer. The loss function was mean squared error since this is a regression task. Adding a fourth layer improved accuracy but fifth didn't. We used the keras library from tensor flow to implement this. In the end, this did not perform better than the XGBoost model. This seemed to be a situation in which such a complex model was not necessary. Regression models worked better.



Neural network	MSE	R2
Training set	0.18	0.64
Test set	0.20	0.61

Conclusion and final thoughts

Among the models that we tested XGBR was found to be the best model for predicting airbnb listing prices. XGBR had the highest R2 score and lowest MSE scores, indicating that it had the best fit and was the most accurate model. Our best performing model was able to explain approximately 75% of the variation in price. It is possible that there are other external factors, such as area, image quality and furnishing, interiors, which are not present in our dataset, that may have a strong influence on the prices of the properties since these are some important factors that a potential customer considers. It seemed there was no single feature that had a heavy influence on the price. Rather, the price was affected by a combination of features. A similar model could be used to study housing prices as well.

NOTE : Attaching python notebook (.ipynb) with our submission which has all the code for our project implementation. We also ran a few more models for our trial and testing and those are present in our python notebook (.ipynb).