Kavisha Shroff

Predicting New York Airbnb Prices

Capstone Project I - Proposal

Main Objective

The dataset contains detailed information for Airbnb listings in New York City, NY. There are about 50k listings with several different features for each listing, including the location, price, host details, etc. The main objective of the project is to build a price prediction model for NYC Airbnb listings based on its different characteristics.

Business Value or Impact

Airbnb is a company that operates an online marketplace and hospitality service which is accessible via its websites and mobile apps. Members can use the service to arrange or offer lodging, primarily homestays, or tourism experiences. This project will help analyze which characteristics contribute to higher prices and help hosts increase their revenues. This will be useful for Airbnb as a company, hosts located in NYC as well as customers who are planning on booking accomodations in the city.

Dataset Details

Available at : http://insideairbnb.com/get-the-data.html

Description Provided:

> "*Inside Airbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.*
> *By analyzing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so you can see how Airbnb is being used to compete with the residential housing market.*
> *With Inside Airbnb, you can ask fundamental questions about Airbnb in any neighbourhood, or across the city as a whole. Questions such as:*
> *How many listings are in my neighbourhood and where are they?*
> *How many houses and apartments are being rented out frequently to tourists and not to long-term residents?*
> *How much are hosts making from renting to tourists (compare that to long-term rentals)?*
> *Which hosts are running a business with multiple listings and where they?"*

Solution Approach

The dataset has 96 columns, including id, name, summary, location information, host information, prices, reviews, etc.

Firstly, the data needs to be understood and cleaned up to use for further analysis.

Once, the clean data is available, EDA would follow to understand the trends between different variables in the data.

To predict the Airbnb price amount based on different categories, I would try different regression models and compare them with each other.

1. **Understanding Data and Wrangling**
   There are a large number of columns and some of them are either unnecessary or redundant. Unnecessary columns such as ids or urls will be deleted. Other columns with the >50% null values would be inspected to see if the values can be filled in or might be deleted. By the end, I will finalize a dataset including only the columns which are useful features for modelling, and also fill in any missing values.

2. **Exploratory Data Analysis**
   First, I would look at frequency counts to find out the number of unique values in the data. Additionally, also get more insight into the data and the numerical column summary statistics. Follow up histograms, box plots and scatter plots would help visualize the data and underlying trends better.

3. **Inferential Statistics**
   Hypothesis testing to validate or invalidate specific assumptions based on data attributes and trends. Some interesting hypothesis could be - Does the average price vary depending on which neighborhood the Airbnb is located or based on the review scores and so on. Based on the data attribute distributions and adherence to the central limit theorem, we can leverage statistical tests like the t-test or the z-test to prove or disprove these assumptions. (ANOVA)

4. **Machine Learning**
   Predicting prices is essentially a regression problem. So, different regression models will be tested out on the data, starting from the most basic models, and moving on to more complex models. The regression models I plan to use with the dataset include linear regression, SVM, decision trees, random forest and gradient boosting.

Deliverables

Code for each major analysis step will be in jupyter notebooks
By the time capstone project is completed, there will be a milestone report, final report and a slide deck.
The final product can also be compiled in an article format to be shared online.