Predicting Airbnb Prices

Capstone Project
By Malak Mosly

Problem Statement

- Airbnb is a very popular American company that helps people find houses to rent.
- It is important for a company like Airbnb to offer housing with fair prices in order to remain reputable and competitive.
- Accordingly, we built a Random Forest Regression model that can predict Airbnb house prices.

Dataset Information

- The dataset was obtained from [Kaggle] and contained 226,030 rows and 17 columns of Airbnb prices in 2020.
- The columns are as follows:
    - **Id** - The listing id or house id
    - **Name** - The name of the listing or house
    - **Host_id** - the id number of the person listing the house
    - **Host_name** - the name of the person listing the house
    - **Neighbourhood_group** - the neighbourhood group the house is in
    - **Neighbourhood** - the neighbourhood the house is in
    - **Latitude** - coordinates of the house
    - **Longitude** - coordinates of the house
    - **Room_type** - the type of room being offered in the listing
    - **Price** - the price of the listing
    - **Minimum_nights** - the minimum amount of nights the house can be rented
    - **Number_of_reviews** - the number of reviews on the house from previous renters
    - **Last_review** - the date of the last review on the house
    - **Reviews_per_month** - how many reviews a particular house gets each month
    - **Calculated_host_listings_count** - how many listings the host has on Airbnb
    - **Availability_365** - the availability of a listing throughout the year
    - **City** - the city the house or listing is located in

Features Used for Prediction

- There were 7 features used by the model to predict prices:
  - room_type
  - minimum_nights
  - availability_365
  - city
  - calculated_host_listings_count
  - reviews_per_month
  - number_of_reviews

Project Steps

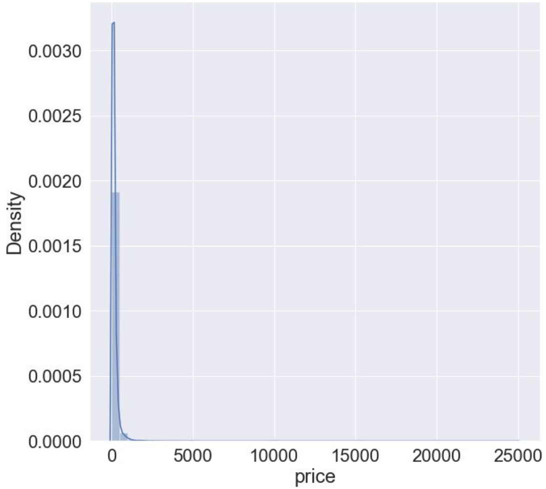
1. Data Wrangling and Cleaning
2. Exploratory Data Analysis
3. Preprocessing and Training
4. Modeling

Data Wrangling and Cleaning

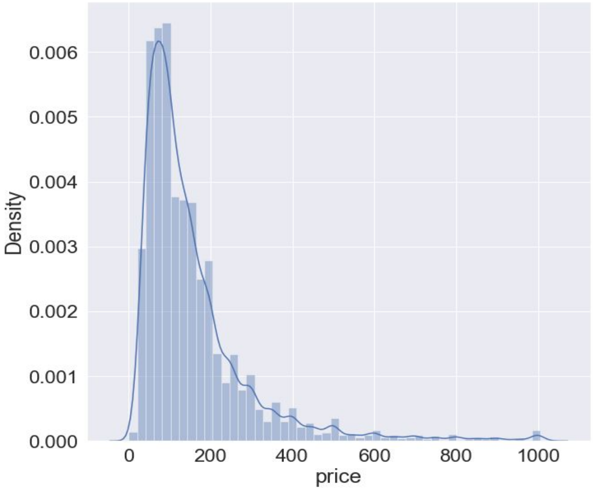
- The original dataset contained 226,030 rows and 17 columns.
- Missing values were removed, which left 85,144 rows and 17 columns of data.
- After outlier removal, there were 62,808 rows and 17 columns left.

Distribution of prices

Before outlier detection and removal, the distribution of prices looked like this:



After outlier detection and removal, the distribution of prices looks like this:



Exploratory Data Analysis (I)

- The distributions of each column was explored.
- Most columns had a skewed distribution even after outlier removal.
- Correlation analysis showed that latitude and longitude, and reviews_per_month and number of reviews were highly correlated.

Exploratory Data Analysis (II)

- Unhelpful columns like Id, host_id, name, and host_name, latitude, and longitude were dropped.

- We were left with 62,808 rows and 11 columns to work with.

Preprocessing and Training

This step of the project involved:

- Creating dummy variables for categorical variables of interest (room_type and city).
- Scaling numerical variables using sklearn's StandardScaler() function.
- Creating our X variable out of our numerical and categorical datasets.
- Splitting the data up (30% test, 70% train split)
- Our X variable had 14 features.
- Our y variable was the variable to be predicted (price).

Modeling (I)

- Three models were used in this project:
    - Linear Regression Model
    - Random Forest Model
    - K Nearest Neighbor Regression Model
- The models were cross-validated, and had hyperparameter tuning done with GridSearchCV. Performance assessment was done for each model and results were compared to determine the best performing model.

Modeling (II)

- The best performing models were the K Nearest Neighbor and Random Forest Regression models.
- Performance assessment included calculation of R2 scores and Mean Absolute Error.
- KNN and Random Forest Regression Models had the lowest Mean Absolute Errors.
- The Random Forest Regression Model had the lowest standard deviation of the MAE out of all three models.

Modeling (III)

Below is the table that assesses the mean absolute error and mean absolute error standard deviation of each model.

Model Name	Mean Absolute Error ($)	Standard Deviation
KNN Model	75.2	1.59
Random Forest Model	75.44	1.18
Linear Regression Model	77.61	1.4

The KNN and RF Models perform better than the Linear Regression Model. Both have a lower MAE by about $2.

Model Predictions Comparison

- Here is a snippet of price prediction comparisons between the RF Model and the K Neighbors Regressor Model.
- The RF Model is more consistently accurate in its price predictions.

```
ypred_rf = rf_best.predict(X_test)
ypred_rf
```

```
array([128.549      , 203.819      ,  84.44528532,
       201.865      , 213.718      ])
```

```
ypred_knn = model_best.predict(X_test)
ypred_knn
```

```
array([121.11111111, 273.11111111,  62.22222222,
       310.77777778, 183.        ])
```

```
y
```

```
0        124
1        239
2        120
```

Model Selection

- The accuracy in price prediction between the KNN and RF models were very close; only $0.22 difference.
- However, the RF model had a noticeably lower standard deviation of its MAE, as seen in the previous table.
- If no further tuning was to be done on these models, the recommendation would be to use the RF Model over the KNN Model because it has a similar level of accuracy but lower variability in its predictions.

Model Usage Suggestions

- The model can be used by Airbnb to regulate listed prices.
- The model can be used by homeowners listing their own homes, outside of Airbnb, to get an accurate valuation.
- The model can be used by developers to determine what kind of houses to build.
- The model can b e used for property valuation by mortgage companies.

Future recommendations and improvements

- The model should be expanded to include every city in the United States.
- An ensemble model could be created to expand to every city and make price predictions.
- The model can also be expanded to include more years than 2020 alone. It could be expanded to about 5 or 10 years. This would allow us to analyze the trends in rental prices over time.