**AirBnb Price Prediction Challenge**

**Problem Statement**

Given a dataset with 24 variables such as number of bedrooms and a log-price indicator (greater than 0) for each observation in the training data, the objective is to suggest the log-price of a particular listing using the 24 features provided for the test observations. The suggested prices that are closest to the true prices, as calculated by the Root Mean Squared Error, scores highest on the leaderboard.

**Pre-processing steps**

1. Zipcodes

* Filled missing zipcodes using geopy library based on longitude and latitude features.
* Used regex to remove non-numerical characters.

1. Neighbourhood

* Filled neighbourhood values based on zipcode and city features.

1. Amenities

* Used regex to remove special characters apart from ‘.’, ‘,’
* Amenities are separated and given weights to each of it based on the frequency of each amenity.

1. Host\_has\_profile\_pic

* Used mode to fill missing values

1. Host\_identity\_verified

* Used mode to fill missing values

1. Host\_response\_rate

* Used regex to remove % symbol and converted each value to integer
* Used median to fill missing values

1. Bathrooms

* Used median to fill missing values

1. Review\_scores\_rating

* Used median to fill missing values

1. Bedrooms

* Used median to fill missing values

1. Host\_since

* Used median to fill missing values
* Converted host\_since to number of days by calculating the difference between the current date to the given date in the feature.

1. Beds

* Used median to fill missing values

1. Review\_diff

* Created a new feature review\_diff, which is the difference between the days of first\_review and last\_review.
* Used median to fill missing values

1. Zipcode

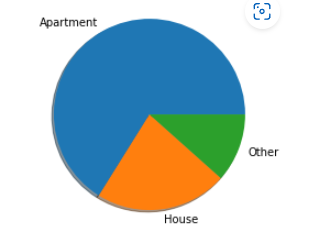
* Used median to fill missing values

1. Description

* Removed stop words using stopwords library from nltk.corpus.
* Calculated polarity scores based on the given description using sentiment analysis.

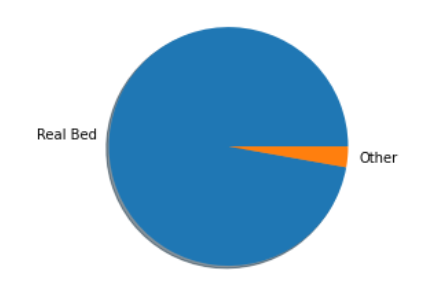
1. Property type:

* Apartment and house property types constitute 66% and 22% of the entire feature’s data respectively. Hence, the rest of the property types have been merged into others.



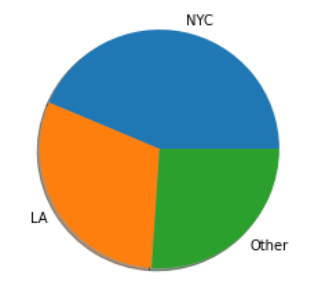
1. Bed\_type

* Real bed type constitutes 97% of the entire feature’s data. Hence, all the other bed types have been merged into others.



1. City

* NYC and LA cities constitute 43.6% and 30.3% of the entire feature’s data. Hence, all the other cities have been merged into others.



**Dropped features**

The following features have been dropped.

* + First\_review
  + Last\_review
  + Longitude
  + Latitude

**Encoding variables**

1. Dummy encoding

* Dummy encoding is performed on room\_type, bed\_type, city and property\_type.

1. Label encoding:

* Label encoding is performed on cancellation\_policy, cleaning\_fee, host\_has\_profile\_pic, host\_identity\_verified, instant\_bookable based on the priority.

**Feature scaling**

The following features have been normalized as they have higher values.

* Amenities
* Host\_since
* Accommodation
* Number\_of\_reviews
* Review\_scores\_rating
* Beds
* Review\_diff
* Host\_response\_rate

**Correlation:**

Correlation between features is calculated and the following features have values greater than 80%. Hence, they have been dropped.

* Beds
* Room\_type\_Private room
* Bed\_type\_Real Bed

**Splitting data:**

Train\_test\_split library is used to split the data into training and testing set. 50% of the data is used for training and 50% is used for testing.

**Models:**

The following are the models and their respective RMSE scores.

|  |  |
| --- | --- |
| Model Name | RMSE |
| Ridge | 0.49 |
| Linear regression | 0.48 |
| Lasso | 0.49 |
| Random forest regressor | 0.42 |
| XGB Regressor | 0.39 |

Since XGB Regressor has the lowest RMSE, it’s used in price prediction.