Airbnb is a service that allows people to list or rent their homes for the purposes of vacation, apartment rentals, home-away from homes, beds or as a hotel. Over the past few years Airbnb has seen explosive growth in popularity is one of the most popular vacation rental services because of how affordable it is. As a result many home owners want a piece of the profit and are listing their places to be rented on Airbnb’s website. However new owners may run into a roadblock when estimating the renting price of their property as there is no built in tool for this task. Using data available from Airbnb we hope to create a machine learning algorithm to predict the renting price based upon different attributes of the property.

We will be using the listings dataset which is free from www.insideairbnb.com

We will approach this problem by first wrangling and cleaning the data, then splitting up the data into train and test sets. We want to use the K-nearest neighbors algorithm for different attributes i.e. number of bedrooms, number of guests accommodated, number of beds, price, number of reviews etc. We will then calculate the similarity of each listing using a similarity metric for multivariate models then select and calculate mean of the first “K” listings. This will be our new predicted price for the listing. The models will then be evaluated for strength using R mean squared error.

The data of listings in San Francisco, was taken from [www.insideairbnb.com](http://www.insideairbnb.com). First the data was viewed through pandas, and it was found that there were 96 columns of attributes. Most of the data types were strings and were therefore unable to be used by our machine learning algorithm. The columns containing these attributes were promptly removed by selecting only those columns with numbers for use with the data frame. Using exploratory data analysis we observed that there were no outliers so that step of data wrangling could be skipped. The remaining columns that contained NaN values which were replaced by zeros since NaN cannot be utilized in machine learning. The data was then normalized using min/max methods so it would be more centralized. After these steps were taken the data was able to be used for machine learning.

Initial Statistical Analysis

Through exploratory data analysis and the usage of scatterplots we have found some relationships and correlations that are worth looking at. Bedrooms and Baths are correlated not surprisingly which could cause some bias in our model because of multi-collinearity. Another interesting pattern I saw was number of reviews is correlated to value of the scores. The more reviews an AirBNB property has the higher the reviews of the property which makes sense because the more feedback a host is receiving the more likely they are working on whatever issues they have gotten complaints for. Another correlation is between the guests included and accommodates columns, as they are positively correlated. This is also a trend that is not surprising but should be taken into account when we parse our data through the machine learning algorithms.

There are also some other interesting correlations on the heat map, there seems to be 1.0 strength correlation between number of reviews, review accuracy, review score, cleanliness score, location score, and check-in scores. The other strong correlations come from the relationships discussed previously with bedrooms, bathrooms, beds, accommodation number and guests included. The other data seems to have very weak correlations with each other at .25 and below. It is peculiar because one would think that review score or cleanliness score would be a driving factor of price. If the property is well-maintained and cleaned that should be reflective on the price. The largest driving factor of price seems to be accommodates, bedrooms and bathrooms but are there other factors at play here? I think this data set would benefit geo-spatial dictionaries for the latitude and longitude columns (For example, where in relation is this property to things like public transportation, airport, landmarks, restaurants etc.)

Machine Learning

We are going to be fitting the K-nearest neighbors algorithm to our data for this analysis which finds listings similar to ours and takes the average price.

**K Neighbors Algorithm Walkthrough**

First the number of similar listings we want to compare ours to (k) is selected.

Then we need to calculate using the Euclidean distance metric how similar each listing is to ours.

The listings are then ranked and the first k number of listings are selected.

Finally we calculate the mean price for the selected k listings and set that as our “list price”

**Evaluation**

To evaluate our model we split the data into 75%/25% train/test sets and use the data entries in the

training set to predict the value of the entries in the test set.

The values are then compared with the actual prices to see how accurate our predictions are.

We will be using root mean squared error (RMSE) to measure accuracy where a lower value of RMSE

Translates into a better model.

**Univariate Model**

The first model we will be utilizing is a univariate model meaning it will only take into account one feature at a time. The second model we will evaluate will be a bit more complex utilizing Scikit learn the ability to select multiple features will be available and thus our model will be clearly be improved.

As we predicted using only review scores to predict the price the RMSE is very high at 568.82$

**Multivariate Model**

The second group of models evaluated will be multivariate models. Since they are able to effectively capture more attributes, the multivariate models should be stronger tools of prediction.

Model One RMSE = .1165

The first model will use accommodated guests, number of bedrooms, number of bathrooms and number of beds as attributes to predict price.

Model One RMSE = .1165

The second model uses accommodated guests, number of bedrooms and number of bathrooms.

Model Two RMSE = .1117

The third model uses accommodated guests, number of bedrooms, number of bathrooms, number of beds, and review scores.

Model Three RMSE = .1166

The fourth model uses only review scores, review score accuracy and guests included.

Model Four RMSE = .1172

**Conclusions**

As assumed using only one attribute to predict price yielded very high RMSE’s consequently adding more attributes didn’t necessarily produce the best model either. The best model is the one that uses accommodated guests, number of bedrooms and number of bathrooms as predictors of price. The RMSE of this model was .1117 and the variables are not likely to be correlated.

AirBNB should use machine learning models to improve the customer experience for guests and hosts alike. Both client and business side will benefit from AirBNB providing a tool to appraise new property rental value. If hosts are able to fairly appraise their own properties, they will get more customers. Consequently if customers feel that a host’s price is fair, they will provide more business either by returning, referral or high reviews.

**NOTES ON PROJECT PRESENTATION**

**Write out ten most important features in the project report. Use random forest to find the most important features.**

**Steps to clean data: remove variables which are highly correlated, it’s going to over fit training will be ok but testing will not be. Can change routine convention.**

**R Squared Pearson test between actual and predicted prices. Check over fitting**

**Report as to what I’m doing in terms of data processing.**