Machine Learning: Airbnb, San Francisco listings

**I. Background**

When you want to go on a vacation or you have to move to a different city, you need to plan many things. One of the most important things you will need is a place to stay. There are houses with unused rooms and other unoccupied housings that people ignore. Airbnb is allows people to lease or short term rent such lodgings.

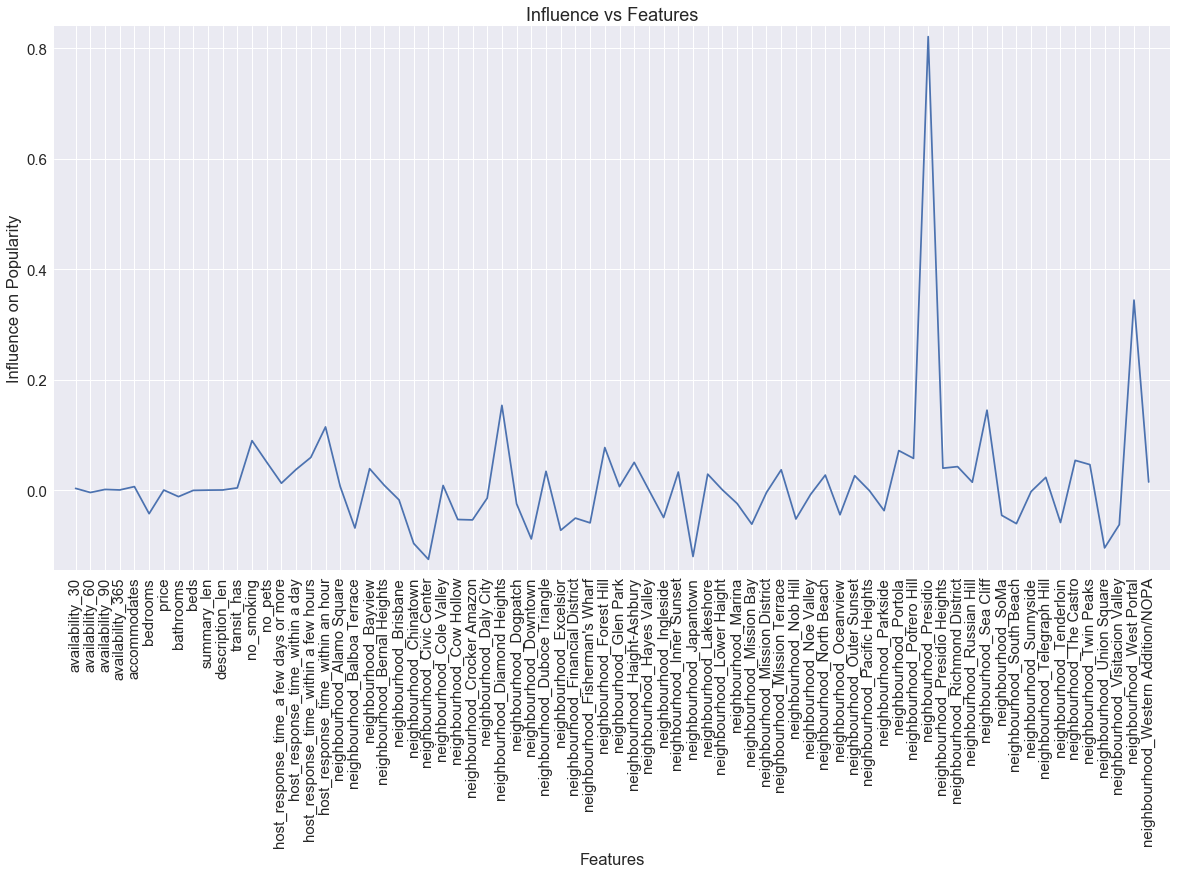
**II. Problem and Customer**

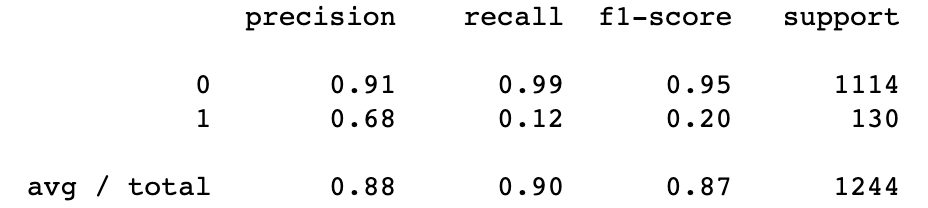
There are listings in airbnb, especially in a big city such as San Fransisco. But how will we know which listing to choose? To figure this out, my project will focus on which neighborhoods and listings are the most popular, and what factors are associated with them. I use machine learning techniques to learn more about the listings.

**III. Machine Learning process**

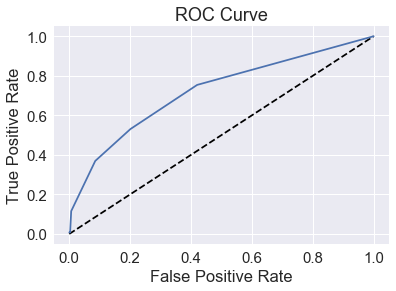
The first thing I did was clean up my data so that there were no missing values in the features I’m looking at. My goals in this section are as follows:

1. Train a KNN Classifier with high accuracy so we can accurately predict future trends.
2. Use lasso to find out which factors are associated with the most popular listings.
3. Make a classification matrix and find out the precision, recall and the f1-score
4. Find out the AUC scores using cross validation

First I trained a KNN Classifier on the following features: availability\_30, availability\_60, availability\_90, availability\_365, accommodates, bedrooms, neighbourhood, price, bathrooms, beds, summary\_len, description\_len, transit\_has, no\_smoking, no\_pets, host\_response\_time\_a few days or more, host\_response\_time\_within a day, host\_response\_time\_within a few hours, and host\_response\_time\_within an hour. I left out number\_of\_reviews since the target of the classifier is popularity, which is derived from the number\_of\_reviews. This would give reviews an unfair advantage in accuracy. Unless I am given the popular listings or I use a different predictor for popularity, I cannot use the number\_of\_reviews as an indicator for popularity. After creating a classifier with all of the above features with a train split of 80% and a test split of 20%, I got a .9019 for the knn score. This means a classifier with all of the features is about 90% accurate. This made me think how accurate the classifier would be if I only used the neighborhood, or only used the price, or different combinations of the features. However, all of these classifiers resulted in an accuracy of less than 90%. 

Second, I wanted to find out which of the above features affect the listings’ popularity the most. To do this, I used lasso regression analysis to see how much each feature influences the popularity. The results are shown in the figure above. The highest indicator of popularity is the neighborhood, Presidio, and the second highest indicator is the neighborhood, West Portal. The graph shows that most of the indicators that influence popularity are neighborhoods followed by pets and smoking policies.

Third, I wanted to know how much precision and recall the KNN Classifier has. To do this I used a classification report and a confusion matrix. The results showed that the TN = 1107, TP = 15, FN = 115, and the FP = 15. The classification report is shown above. The precision, recall, and the f1-score are above 90% so the KNN Classifier is very accurate in predicting popularity.



Fourth, I plotted the ROC Curve and calculated the AUC scores. The AUC scores computed using 5-fold cross-validation = [ 0.64186922 0.68629678 0.70105743 0.68076923 0.71992729]. The results of the cross validation are acceptable since they are mostly around 70%. The AUC is also closer to 1 than 0 so the model is is better at calculating True and False Positives.

**IV. Conclusion**

I learned many things while working on the Machine Learning section of this project. I learned that the KNN Classifier was the most accurate when I used all of the relevant features in the classifier. The Lasso analysis told me the most about the factors that affect popularity the most. In this case neighborhoods affected the popularity the most, especially Presidio which influenced the popularity over 80% of the time. This shows that the most popular listings are usually dependent on the neighborhood. This makes sense because a neighborhood closer to popular tourist attractions would tend to host more popular listings. The confusion matrix and classification report showed that the results were over 90% accurate and mostly True Negatives.