AirBnB Pricing Prediction (Milestone #2)

Objective

To develop a model to predict the price of a property on a given date using features/amenities of comparable properties.

Proposal

Airbnb enables a marketplace where people can rent short-term lodging in residential properties. The offers range from shared rooms to full condos/houses and can offer unique and differentiated options as compared to chain-hotels. The pricing is set by the property owner and this can be a daunting task given the variety and unique features of the property – no two properties are the same and the relative attractiveness can depend on a number of factors not limited to location, amenities, area, transit options, season, host reviews, etc.

We would like to analyze the historical prices of rentals in New York City and extract features that could be relevant for setting a data-informed price by the owner.

We referred to a couple of readings summarized below to learn about a couple of different approaches to modelling.

Paper #1: "Predicting AirBnb listing prices with Scikit-Learn and Apache Spark

Intelligent pricing predictions are the result of a variety of machine learning techniques, and this paper details the particular process taken to provide useful results for Airbnb consumers. The paper employs methods of variable selection and estimation seen before (Lasso, OLS, Ridge), as well as more customized and precise tools such as GridSearch and Ensemble Methods. This gives a good idea of how to begin our approach to this data problem, in addition to adding our own adjustments, such as images, KNN, sentiment analysis, etc., pending their appropriateness during the d[1]ata discovery process of this project.

Additionally, this paper introduces Apache Spark, as a map-reduce solution, shifting the approach to a distributive framework. This has obvious advantages, such as lessening individual CPU processing time for such a large dataset, as well as scalability, i.e. being able to aptly fine-tune regression parameters even as the dataset increases by n orders of magnitude. Ideally, this would be the case for any large dataset, but as we do not have the resources necessary to do this for this project, this procedure serves as an idealized method for when datasets become very large.

1

Group: Airbnb Pricing Prediction - 4
TA: Christine Hwang
Team Members: Paul Harris, Sanjeev Bagaria, Samuel Heller, Amay Umradia

AirBnB Pricing Prediction (Milestone #2)

Paper #2: Neighborhood and Price Prediction for San Francisco Airbnb Listings

This paper builds on analysis of Airbnb listings by using text and image features to predict neighborhood and price of listings. The text analysis portion of the paper employs multiple methods, looking first at the 1000 most commonly-used word stems, then looking at words classes, which the researchers chose manually and then sorted words into. They also find sentiment features using the TextBlob package. These methods can help guide how we do our own textual analysis, showing how we can extract and sort words that can then be used for prediction.

Another important aspect of this paper is how it deals with high dimensionality and overfitting. Because text analysis involves a large number of features, the paper used Recursive Feature Elimination (RFE) to find the *k* most important features. The best value of *k* was found by testing a range of *k*s against a dev set, separate from the test and train set. After a value of *k* was found for each model, the paper used a Support Vector Machine (SVM) to determine listing price. This may be out of the range of this course, but it is an interesting approach to consider beyond more standard regression techniques that we have covered in class. The RFE method also provides an interesting approach to feature selection that we may want to consider instead or or in addition to PCA.

Datasets

1. Listing Data
    - http://data.beta.nyc/dataset/inside-airbnb-data/resource/9d64399b-36d6-40a9-b0bb-f26ae0d9c53f?view_id=33b9a800-4ed6-4d41-8f87-494c6c8582eb
2. Seasonality Data
    - http://data.beta.nyc/dataset/inside-airbnb-data/resource/ce0cbf46-83f9-414a-8a1d-7fd5321d83ca
3. For text analysis:
    - http://data.beta.nyc/dataset/inside-airbnb-data/resource/8115833e-8a0e-4af6-8aed-4d96a0ae0b73

Group: Airbnb Pricing Prediction - 4
TA: Christine Hwang
Team Members: Paul Harris, Sanjeev Bagaria, Samuel Heller, Amay Umradia

AirBnB Pricing Prediction (Milestone #2)

Approach

We will likely take a two-step approach:

1) Classify the subject property to a neighborhood using a Classifier
2) Use property specific features within the neighborhood to predict the price using a Regression model

Feature Extraction

For designing our features, we will use the Listing and Seasonality data mentioned above. Some features based on initial observation could be:

Use a classifier method to group:

1. Neighborhood
2. Listing price based on Neighborhood

Determining Listing price:

Basic features from the neighborhood:

1. Prices
2. Amenities
3. Area of the house with # of beds/baths
4. Local Transit nearby
5. Host biography

Advanced features:

1. Number of reviews
2. Host response rate
3. Number of references

Neighborhood Classification:

1. Crime rate
2. Population
3. Culture, night-life
4. Facebook/Uber check-ins to see

Dynamic pricing tool consideration → Advanced idea

1. Already a tool called Aerosolve.Can we come up with a substitute and not a complicated one

Feature Extraction:

For pricing prediction, can we use only #1 or #2 can be helpful
For neighborhood prediction, we might use 2,3 below:

Group: Airbnb Pricing Prediction - 4
TA: Christine Hwang
Team Members: Paul Harris, Sanjeev Bagaria, Samuel Heller, Amay Umradia

AirBnB Pricing Prediction (Milestone #2)

1. **Listing info** Take the data and fill the missing values if needed. This info could be whether the house is apartment, condo, dorm, # of bed, charge/guest/night

2. **Bag of words** Info like Summary of listing, space, description, experiences offered Which technique can be used here?
Word-class: In paper, 9-word class were chosen people, nightlife, activities, style, accessibility, culture, nature, amenities, and comfort – might not be needed

3. **Text sentiment features** à TextBlob package, which calculates the polarity of a segment of text by averaging the polarity of each word in the text included in the package's lexicon

Analysis and Prediction:

1. KNN classifier can be used to neighborhood detection. We can remove few neighborhoods with fewer listings. Determining optimal value of K will be important. We have always run K-classfier and checked its R2 value with various value of K. According to paper, sklearn's Recursive Feature Elimination (RFE) was used
2. Regression Technique like Lasso or Ridge will be useful to find significant features while also on bag of words to see which are good enough for deductions.
3. Train, Test using any of the CV method à Bootstrapping for better estimation of BETAs. Also is R2, AIC, BIC only the measure for good fit?
4. As of now, polynomial or in general Linear regression technique but if any better algorithm is taught in the class then we can use that.
5. Correlation between neighborhood and pricing will depend on the area of the house located. Same characteristics of houses but better area will have higher pricing.

References:

Tang, Sangani. "Neighborhood and Price Prediction for San Francisco Airbnb Listings"
- http://cs229.stanford.edu/proj2015/236_report.pdf

A previous project on this topic:
- https://www.mapr.com/blog/predictingairbnblistingpricesscikitlearnandapachespark