CMPE 256 - Large Scale Analytics

Airbnb Data Analysis

Predictions and Recommendations

Team 2

SAN JOSÉ STATE
UNIVERSITY

Submitted to

Dr. Gheorghi Guzun

on

05/01/2019

by

Abhishek Konduri	013710645	abhishek.konduri@sjsu.edu
Nitish Joshi	013736320	nitish.joshi@sjsu.edu
Rohan Kamat	013759252	rohansantosh.kamat@sjsu.edu
Varun Jain	013719108	varun.jain@sjsu.edu

TABLE OF CONTENTS

Chapter 1: Introduction

1.1 Motivation

In today's world, we are constantly surrounded by social media. We share, read and post a lot of things on the internet. In this digital world, there is always a place for analysis, predictions, recommendations as well as suggestions. Keeping this in mind, the main idea behind our project was to use this platform to perform analysis and obtain conclusions by processing large amounts of data with the help of various data analysis and visualization techniques and data mining algorithms in the domain of Large Scale Analytics.

One such use case that intrigued us the most was to perform analysis and predictions on the Airbnb dataset. Airbnb is an online marketplace and hospitality service which is accessible via its websites and mobile apps. Airbnb assists people all around the world to find a place to stay of their choice. In addition, Airbnb also allows hosting their apartments and villas. Having a huge web presence and popularity all over the world propelled us to work and have a closer look at the various functionalities of Airbnb.

1.2 Objectives

Our project objectives are focused on the analysis and predictions of some of the important aspects of the Airbnb features. Below is the list of our objectives that we hope to achieve as part of our project implementation:

- Causality for different price listings.
- Location-specific investment to get maximum profit/returns.
- Prediction of prices for new property listings.
- Sentiment Analysis of reviews.
- Host Analysis.
- Analysis of the seasonal pattern of prices
- Recommender System

Chapter 2: System Design and Architecture

2.1 Algorithms Selected

2.1.1 Prediction of prices for new property listings:

1. Vanilla Linear Regression
2. Ridge Linear Regression
3. Lasso Linear Regression
4. Elastic Net Linear Regression
5. Bayesian Ridge Linear Regression

6. Orthogonal Matching Pursuit (OMP) Linear Regression

2.1.2 Location specific investment to get maximum profit:

1. Orthogonal Matching Pursuit (OMP) Linear Regression
2. Bayesian Ridge Linear Regression
3. Lasso Linear Regression
4. Linear Estimator
5. Ridge Linear Regression
6. Elastic Net Linear Regression
7. LARS (Least Angle Regression model)

2.1.3 Sentiment Analysis of reviews:

1. SentimentIntensityAnalyzer

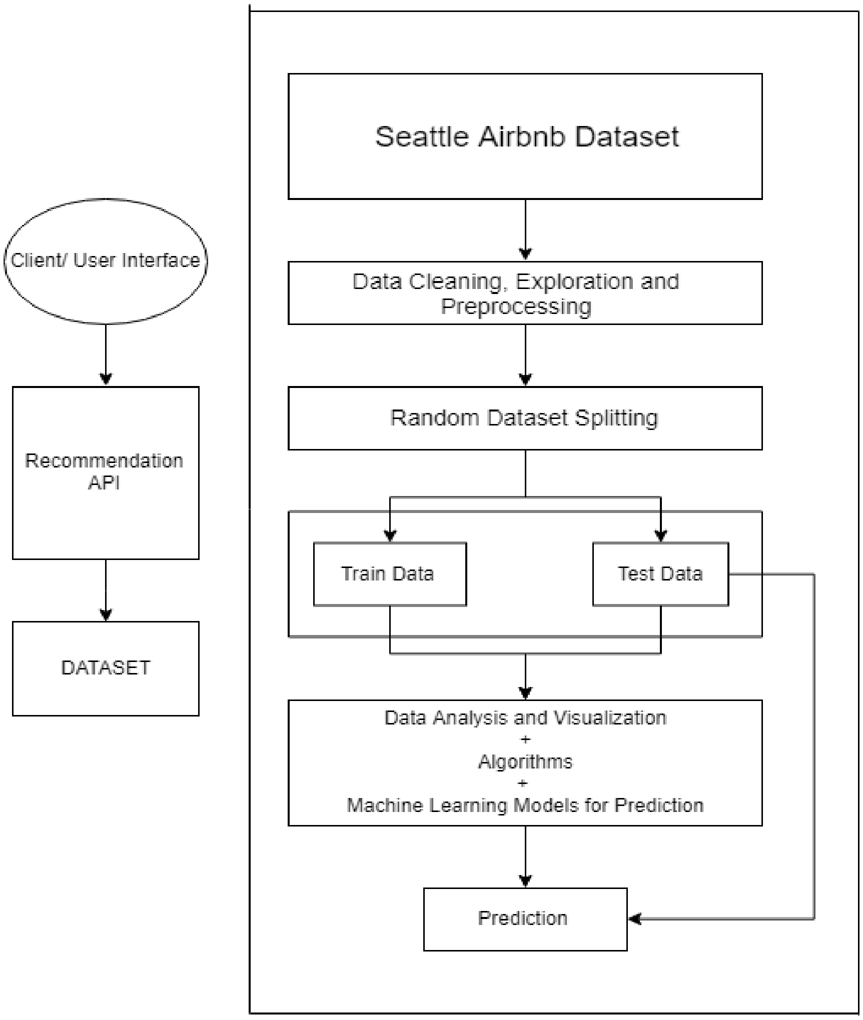
2.1.4 Host Analysis:

1. Logistic Regression
2. Random Forest Regressor
3. Random Forest Classifier
4. Support Vector Classifier
5. K Neighbors Classifier
6. Gradient Boost & Ada Boost
7. Decision Tree

2.2 Technologies and Tools Used

1. Python 3
2. Jupyter Notebooks
3. Scikit-learn

2.3 System Design and Architecture



Chapter 3: Experiments

3.1 Dataset

Seattle Airbnb Open Data

The dataset includes 3 CSV files (86 MB)

Calendar.csv - 35 MB
Listings.csv - 16 MB
Reviews.csv - 35 MB

Description

1. Listings.csv - Consists of details of all the listings in Seattle including their price, accommodations, ratings, number of reviews, summary, name, owner name, description, host Id and many other columns describing details of listings.

2. Calendar.csv - Consists of details of listings and its date, availability and its price.
3. Reviews.csv - Consists of reviews for each listing in Seattle.

3.2 Data Preprocessing

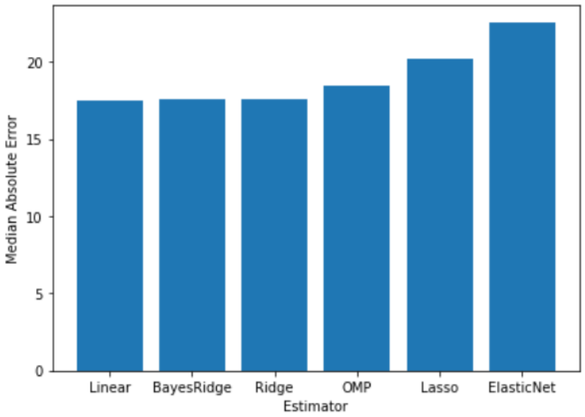
The dataset available was not ready to be used for our analysis and prediction purposes. The dataset we had had to be cleaned and transformed into the best suitable format for our use. We had to use a few preprocessing steps in order to make the dataset more user-friendly. Below are the steps/techniques we used to preprocess the data:

- Leveraging Pandas and Numpy for Data Wrangling.
- Replacing NaN values with 0 and cleaning the data to make it float.
- Selection of only the necessary columns for further processing and dropping the remaining columns from the data frame.
- Excluding the listings with null values/zero values for prices, bedrooms, beds.
- Added dummy values for some columns which had non-quantifiable values.
- Split the data for each record as data, day and month for better analysis of seasonal variation of prices.
- Grouping of the data to get a mean price for each listing for each day.

3.3 Methodology

3.3.1 Prediction of prices for new property listings

Prices for a property is dependent on various factors like location, type of property (Condo, bungalow, flat), various provided amenities like security, Wi-Fi, hot water, reachability to the city sites and hot spots, and the time is it for booked and time it is booked on. Once the data was cleaned, and the final data frame was ready to use for further analysis, we observed that 80% of apartments host only single bedroom and rest 20% host 2, 3, 4 bedrooms. To obtain the accuracy we zoomed into data for only 1-bedroom apartments. Once the model is built for a 1-bedroom apartment, same can be applied for other 20% apartments as well. The data was divided into 4:1 ratio as training data and testing data respectively. We can try for multiple algorithms in one go and comparing the result and further we can decide what all tuning can be done. For this particular problem, we have observed that the Bayesian Ridge Linear Regression model gave us the best result with the best absolute model and maximum accuracy. We have used median absolute error for the data accuracy.



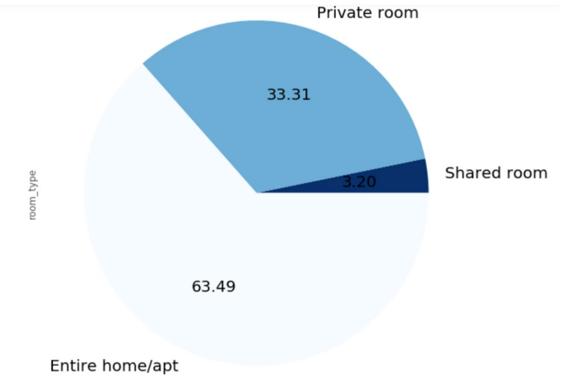
3.3.2 Sentiment Analysis of reviews

There are so many factors which contribute towards the price of a listing on AirBnB.

The main task is to do the sentiment analysis of the reviews and see if it has any relationship with the price. Initially, we decide to drop all the columns that we did not find important according to our use case and selected only the ones which were required and could be impactful to deliver the result. In the revews.csv there are reviews which are not I English, so we'll remove all those entries which are not in English. We built a new column called comments with all the comments on the English language. We are using built-in analyzer in the NLTK Python library to assign a polarity score to each comment. We have assigned Polarity score to every comment which includes detail like Positivity, negativity neutral and compound. After getting the comments column we used SentimentIntensityAnalyzer package on all of these comments. This particular algorithm breaks down all the works and classifies if that word is a positive word/ negative word or a neutral word. We used a polarity-based approach for sentiment analysis where pieces of texts are classified as either positive or negative. A SentimentAnalyzer is a tool to implement and facilitate Sentiment Analysis tasks using NLTK features and classifiers, especially for teaching and demonstrative purposes.

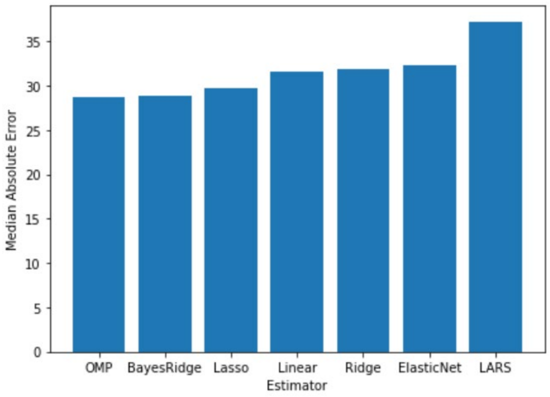
3.3.3 Causality for different price listings

In this analysis, we identified different factors which lead to different prices of listings. We collected, cleaned and transformed the listings.csv data by leveraging pandas and numpy. We categorized different listings based upon their room type and then categorized different listings based upon their property type. As a result, we could conclude that people are more inclined towards listing their entire property than that of private rooms or shared rooms. We also analyzed that for almost all property types, prices for the entire home/apartment were maximum which suggested that Property type and Room type play a very important role in deciding the price of a listing. After further analysis, we could observe that with the increase in the number of bedrooms, the price of listing increases. Finally, we investigated the summary and price for all the listings. We also chose top 100 listings on the basis of price to find out what all common words are used by hosts while posting a listing on Airbnb. We created a word cloud for the same and in order to plot it, we cleaned the data by removing punctuation, unwanted characters, and numbers. We also removed all the stopwords and lemmatized it using WordNetLemmatizer from NTLK and plotted a word cloud for the same to visualize the most common words these hosts utilized to describe their listings on Airbnb. We also analyzed how the amenities provided by the listing is related to the price of the same and plotted a word cloud for the same.



3.3.4 Location specific investment to get maximum profit/returns

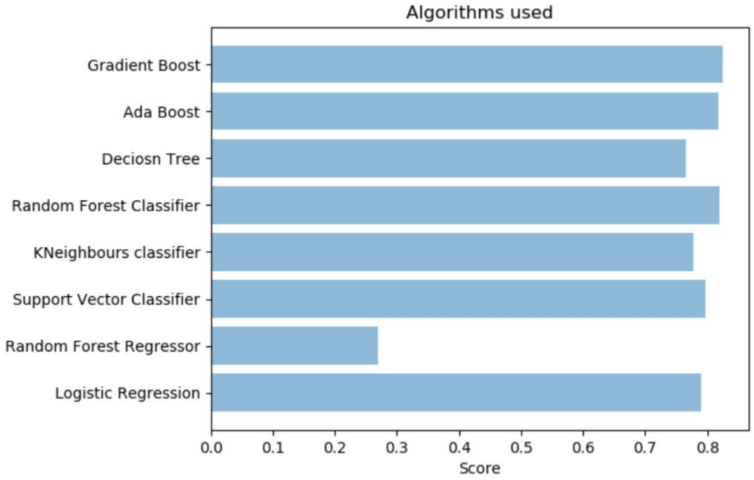
This task required to make use of multiple attributes independent of each other which affect the prices. To predict the best locations, we have used scikit-learn and different linear regressions models which can return a maximum profit.



We performed tests on many linear regression models like Bayesian ridge, ridge, lasso, OMP, linear estimator, elastic net, LARS and compared their results to choose the best one. LARS gave the best result.

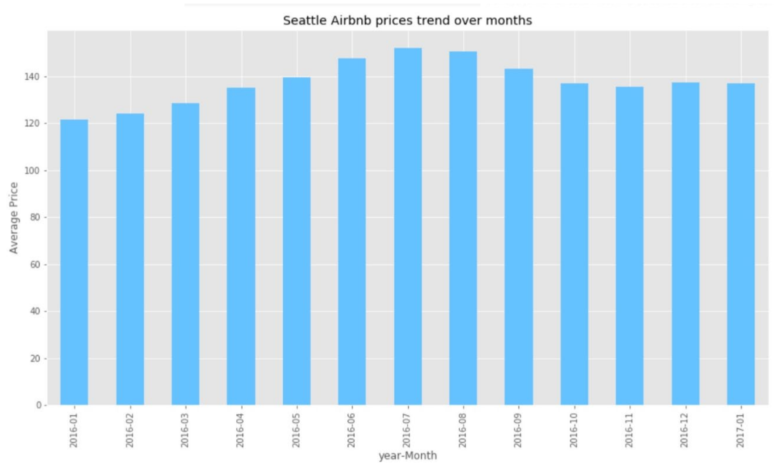
3.3.5 Host Analysis

The data is imported into data frames and columns are inspected. A new data frame was created by calculating the mean per day, and all the missing values were dropped. The mode of the list was taken to check the most number of listing per day and plotted. The motive is to understand the behavior of hosts, super hosts, and delta hosts.The listings data is first imported into a dataframe and columns are inspected, collecting important columns which are important determining the superhost criteria. A heatmap is plotted for super hosts and hosts against these selected columns to understand the correlation between them on the basis of the given attributes(kendall). Multiple visualizations are made to determine the attributes of accounting qualification as a super host. Models used are Logistic regression, Random Forest Classifier, Random Forest Regressor, KNN, SVC out of which Random Forest Classifier gave the best result.



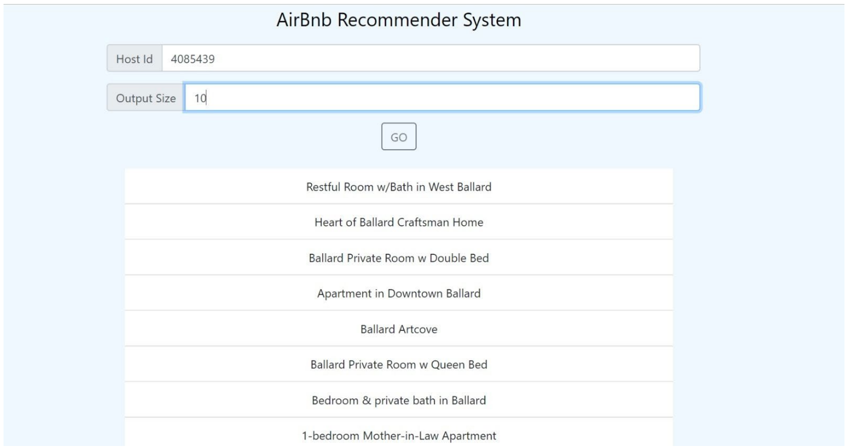
3.3.6 Analysis of the seasonal pattern of prices

The main task is to analyze the given data and see which days and months of the year, the prices of hotels are the highest so that the customer can decide when to visit Seattle. Initially, we decide to drop all the columns that we did not find important according to our use case and selected only the ones which were required and could be impactful to deliver the result. The columns in listing.csv consisted of listing_id, date, availability, price. For cleaning the data, we initially replaced all the NaN values with 0 in the prices column. Then we separated the date column into day month and year columns as we needed to see which month of the year had the highest average price. So, by now the cleaning of data has been done and new columns namely Year, Month and Day by splitting date has been added. Then we extracted the prices column from the table and cleaned it by removing $ symbols and kept just the int value of the prices. Then we replaced the price column with the new prices column with price values in the required format. Now we analyzed the data by a group it on the basis of Year and Month to see the trend of prices and also plotted the respective graphs for it for the better understanding of the user.



3.3.7 Recommender System

A recommender system was modeled using a correlation between rows of a given matrix. The code was written in a class and a method structure to generate a model which could be loaded in a pickle file. The flask api ensured a route, provided the hostid and the number of similar records to be displayed, return the most similar apartments. The front end was developed using react and backend using a flask. The issue faced was CORS(Cross-Origin Resource Sharing), which was allowed so that react could access the routes.


AirBnb Recommender System

Host Id: 4085439

Output Size: 10

Restful Room w/Bath in West Ballard

Heart of Ballard Craftsman Home

Ballard Private Room w Double Bed

Apartment in Downtown Ballard

Ballard Artcove

Ballard Private Room w Queen Bed

Bedroom & private bath in Ballard

1-bedroom Mother-in-Law Apartment

Chapter 4: Discussions and Conclusions

4.1. Decisions Made

- The team discussed and researched about selecting the dataset that met all the criteria necessary for the project implementation.
- After finalizing the dataset, the team decided and narrowed down various tasks that were to be implemented during the project lifecycle.

- The team also made decisions regarding various preprocessing steps and data cleaning strategies to be used such as different data wrangling techniques, using stopwords, etc. in order to achieve the best possible results.
- Finally, the team discussed and made decisions about various algorithms to be used from logistic regression, linear regression to decision trees and random forest algorithms to obtain the best possible predictions on the dataset.

4.2. Difficulties Faced

- The dataset that the team selected was very big in size and hence, needed a lot of time and effort to perform data cleaning, transformation and preprocessing steps on it.
- Since the number of rows was very large, the amount of memory required for computation was huge and hence, the data had to be processed step by step in limited chunks.
- The team also faced difficulties in implementing algorithms and obtain the best possible results considering all the factors such as the complexity of data, the volume of data, a large number of features, etc.

4.3. Things that worked well

- The dataset selection was ideal for us in a way that since it was huge, we got to work and experience the complexity of handling such a huge dataset. Although initially, it looked like a daunting task, the team did their research and found out ways to handle and work with it.
- Selection of data cleaning and preprocessing tasks worked really well as a result of which we got the best out of our algorithms.

4.4 Things that didn’t work well

- Use of algorithms like Support Vector Machines (SVM) took a good amount of time to run and did not produce results that were good enough. As a result, the team faced some problems while selecting the right algorithms that would produce satisfactory results.

4.5 Conclusion

- While implementing this project, we learned about a wide array of techniques, algorithms, and different preprocessing tasks involved in data analysis and prediction and how they affect the performance of the algorithms as a whole.
- We learned how to handle a large imbalance dataset.
- We learned how to apply various algorithms and use predictions models.
- We learned that the output of a model varies based on the requirement of the predictions.

Chapter 5: Project Plan/Task Distribution

Task	Responsibility
Dataset Selection	All
Data Exploration and Cleaning	All
Data Preprocessing	All
Research on Algorithms	All
Prediction of prices for new property listings	Varun Jain
Sentiment Analysis of reviews	Varun Jain
Causality for different price listings	Nitish Joshi
Location-specific investment to get maximum profit/returns	Nitish Joshi
Predicting the best time to get cheap and best deals.	Abhishek Konduri
Analyzing the seasonal pattern of prices.	Abhishek Konduri
Host Analysis	Rohan Kamat
Recommendation API	Rohan Kamat
Documentation and Report	All
PPT	All

References:

[1] https://www.kaggle.com/airbnb/seattle
[2] https://scikit-learn.org
[3] https://scikit-learn.org/stable/tutorial/index.html
[4] https://matplotlib.org/
[5] https://www.airbnb.com/
[6] https://en.wikipedia.org/wiki/Airbnb
[7] https://www.datacamp.com/community/tutorials/matplotlib-tutorial-python