ANALYSIS AND PREDICTION OF PRICE OF AIRBNB LISTINGS

**PROJECT REPORT**

***Submitted by***

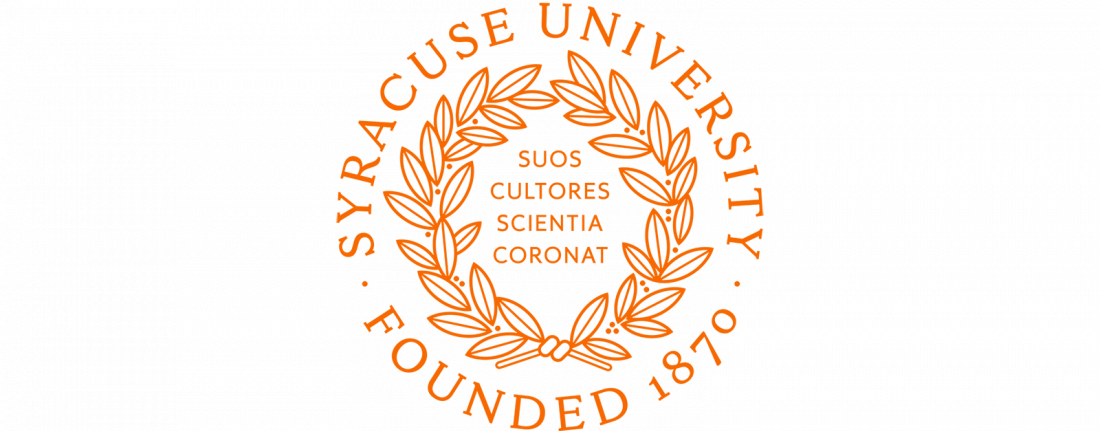
**ASHUTOSH JHA**

**NIDHI JOHNSON**

**IST 707 Data Analytics**

**Fall 2020**





# 

Table of Contents

[1) INTRODUCTION 2](#_Toc57887074)

[2) AIM OF THE PROJECT 3](#_Toc57887075)

[3) DATASETS OVERVIEW 3](#_Toc57887076)

[4) DATA CLEANING AND PRE-PROCESSING 4](#_Toc57887077)

[4.1) Firstly – Removal of NA Values and Blank Spaces 4](#_Toc57887078)

[4.2) Secondly – Binning of Price Variable 4](#_Toc57887079)

[4.3) Thirdly – Converting the Variables to Numeric values 4](#_Toc57887080)

[4.4) Fourth – Converting variables to Factors 4](#_Toc57887081)

[5) FEATURES SET SELECTION 5](#_Toc57887082)

[6) DECISION TREE 5](#_Toc57887083)

[Figure 1.1 Decision Tree (Jersey City) 6](#_Toc57887084)

[Figure 1.2 Decision Tree (Boston) 7](#_Toc57887085)

[7) NAÏVE BAYES 8](#_Toc57887086)

[8) ENSEMBLE METHOD 9](#_Toc57887087)

[9) SUPPORT VECTOR MACHINES (LINEAR) 9](#_Toc57887088)

[10) Brief Comparison of all algorithms 10](#_Toc57887089)

[11) Conclusion 10](#_Toc57887090)

[Reference: 12](#_Toc57887091)

# **INTRODUCTION**

Airbnb is becoming increasingly popular amongst tourists around the world. Airbnb has its own pricing system unlike other hotels. The pricing of the Airbnb listings is mostly decided by the host but there are other factors also which influence the price of the listings.

For the final project

* We have created an algorithm which will help us to analyze the factors that influence the price of the Airbnb listings across the world.
* Our algorithm works for any Airbnb Listings Dataset available on Airbnb Insider.com.
* To demonstrate our algorithm, we choose to analyze the Airbnb listings of Boston and Jersey City, the two of the most populous cities on the east coast.
* Our algorithm with help us to understand how factors like neighborhood, property type, number of rooms, minimum and maximum stay affect the prices of the Airbnb listings across the world.
* Our end goal is to help host and consumer as host can set the correct price for their property/listings and it will help consumers to finds the right place according to their needs.

# **AIM OF THE PROJECT**

* Our initial hypothesis is that prices of the Airbnb rooms are co related to various factors like Neighborhood, Room type, Property type etc.
* Our aim is to validate our hypothesis by using different classification algorithms to see what are the main factors that influence the Price of the Airbnb listing
* Using classification models Decision Tree, Naïve Bayes, Ensemble and SVM we are going to predict the price of the Airbnb listings

# **DATASETS OVERVIEW**

Since our Algorithm works with any listing’s dataset available on the insideairbnb.com which is sourced for public use from official Airbnb Website.

The dataset which we are using here to showcase our analysis:

* Boston
* Jersey City

**Initial Dataset** (Raw Data Obtained from Airbnb official source)

|  |  |  |  |
| --- | --- | --- | --- |
| **Datasets** | **Columns** | **Rows** | **Source** |
| Boston | 74 | 3254 | Insideairbnb.com |
| Jersey City | 74 | 1428 | Insideairbnb.com |

Initially the dataset has 74 observation but for analysis we decided to select only the variables which are ideal for our classifications.

We did our analysis on two cities to validate our results.

**Columns Present in the Final Dataset**

|  |  |
| --- | --- |
| **Column Id** | **Name of the Columns** |
| 1 | Host\_id |
| 2 | Host\_response\_time |
| 3 | Host\_acceptance\_rate |
| 4 | Host\_total\_listings\_count |
| 5 | Neighbourhood\_cleansed |
| 6 | Property\_type |
| 7 | Room\_type |
| 8 | Accommodates |
| 9 | Bedrooms |
| 10 | Beds |
| 11 | Price |
| 12 | Maximum\_nights |
| 13 | Number\_of\_reviews |
| 14 | Reviews\_scores\_rating |
| 15 | Review\_scores\_accuracy |
| 16 | Review\_score\_cleanliness |
| 17 | Review\_score\_checkin |
| 18 | Review\_scores\_communication |
| 19 | Review\_score\_location |
| 20 | Review\_Score\_Value |
| 21 | Instant bookable |
| 22 | Has\_availability |

# **DATA CLEANING AND PRE-PROCESSING**

To run our classification algorithms, we need to Perform Data Cleaning and Data Pre- Processing

## **4.1) Firstly – Removal of NA Values and Blank Spaces**

* We have created a user-defined function for fast cleaning of the dataset. The Function removes all the NA values and blank spaces from the dataset.
* We have decided to remove all NA values in the dataset as substituting it with mean or median affect the integrity of the dataset.

*removeRowsWithNA <- function(data, desiredCols) {*

*completeVec <- complete.cases(data[, desiredCols])*

*return(data[completeVec, ])*

## **4.2) Secondly – Binning of Price Variable**

Since the price data is continuous, in order to make meaningful predictions, we divided all prices into three bins and gave them the labels, “**Low**”, “**Medium**”, “**High**”.

## **4.3) Thirdly – Converting the Variables to Numeric values**

Converted the variables: **host\_response\_rate, host\_acceptance\_rate, price** from factor to numeric values

## **4.4) Fourth – Converting variables to Factors**

Converted the variables: **host\_response\_time, instant\_bookable, room\_type, property\_type, price\_level** to factors

# **FEATURES SET SELECTION**

* We have decided to include only 10 features in our analysis. The reason for selecting only 10 features was to avoid losing of power of the explanatory models
* Splitting the dataset into training and testing. We have generated the testing accuracy by training on 80% of the dataset and testing on 20% of the dataset

**Feature Set**

|  |  |
| --- | --- |
| 1. Neighborhood | 1. Room\_type |
| 1. Host\_response\_time | 1. Bedrooms |
| 1. Host\_response\_rate | 1. Availability\_365 |
| 1. Minimum\_nights | 1. Number\_of\_reviews |
| 1. Instant\_bookable | 1. Price\_level |

# **DECISION TREE**

**Def**: Decision tree as the name suggests it is a flow like a tree structure that works on the principle of conditions. It is efficient and has strong algorithms used for predictive analysis.

**Feature set used**: (Neighborhood, Host\_response\_time, Host\_response\_rate, Minimum\_nights, Instant\_bookable, Room\_type, Bedrooms, Availability\_365, Number\_of\_reviews, Price\_level)

**Methods**: For decision tree we have defined the controls as follows: **Method = repeatedcv, number = 10, repeats = 5**. Then with the help of caret package we call **method rpart for the decision tree**

For our decision tree model, we have selected 10 features from the features set.

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| Jersey City | 75.8% |
| Boston | 70% |

**Room Type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Feature set** | **Feature Detail** | **Prediction** | **Price Range** |
| Jersey City | Room\_type | as entire room/apt or hotel room | 86% | Low-Price |
| Boston | Room\_type | as entire room/apt or hotel room | 81% | Low-Price |

**Bedroom Count**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Feature set** | **Feature Detail** | **Prediction** | **Price Range** |
| Jersey City | Bedroom\_count | more than 2 | 58% | High -Price |
| Jersey City | Bedroom\_count | Less than 2 | 85% | Medium -Price |
| Boston | Bedroom\_count | more than 2 | 71% | High-Price |

**Neighborhood**: Another decisive factor our decision tree shows us

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Feature set** | **Feature Detail** | **Prediction** | **Price Range** |
| Jersey City | Neighborhood | WardE | 76% | High -Price |
| Boston | Neighborhood | Neigbourhood\_listed(dataset  Village, becon, hill, Charlestown, downtown, leather district, Fenway | 80% | High-Price |

**Minimum\_nights:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Feature set** | **Feature Detail** | **Prediction** | **Price Range** |
| Jersey City | Minimum\_nights | Less than 4 nights | 60% | High -Price |
| Boston | Minimum\_nights | Less than 23 nights | 68% | High-Price |

**Accuracy of the Decision Tree model:** The decision tree model gave us the accuracy of the 75.8%.

**Jersey City - Accuracy – 75.8%**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7826 | 0.8103 | 0.6792 |
| **Recall** | 0.6923 | 0.8868 | 0.6923 |
| **F Measure** | 0.7347 | 0.8468 | 0.6857 |

High Precision values for low class is pretty good as high precision rate mean low false positives (high number of relevant instances). Further high F Score for low price class also collaborates the balance between the positive and negative predictive value.

**Boston – Accuracy - 70%**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7111 | 0.7090 | 0.5773 |
| **Recall** | 0.6038 | 0.8879 | 0.5185 |
| **F Measure** | 0.6531 | 0.7884 | 0.5463 |

High Price level has the highest value for Precision which mean low false positive.

Since our decision tree for both Boston and Jersey City gave us similar results so it validates our algorithm.

## **Figure 1.1 Decision Tree (Jersey City)**

Timeline

Description automatically generated

## **Figure 1.2 Decision Tree (Boston)**

Timeline

Description automatically generated

# **NAÏVE BAYES**

**Def**: Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of [feature](https://en.wikipedia.org/wiki/Feature_vector) values, where the class labels are drawn from some finite set

**Feature set used**: (Neighborhood, Host\_response\_time, Host\_response\_rate, Minimum\_nights, Instant\_bookable, Room\_type, Bedrooms, Availability\_365, Number\_of\_reviews, Price\_level)

**Methods**: For decision tree we have defined the controls as follows: **Method = repeatedcv, number = 10, repeats = 5**. Then with the help of caret package we call **method nb for the naïve bayes.**

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| Jersey City | 73.25% |
| Boston | 65.11% |

**Jersey City - Accuracy – 73.25%**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.8919 | 0.7679 | 0.6094 |
| **Recall** | 0.6346 | 0.8113 | 0.7500 |
| **F** **Measure** | 0.7416 | 0.7890 | 0.6724 |

High values for Precision tell us that our prediction is correct and less false positive are present.

**Boston – Accuracy - 65.11%**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price** **Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7253 | 0.6923 | 0.5300 |
| **Recall** | 0.6226 | 0.8411 | 0.4907 |
| **F Measure** | 0.6701 | 0.7595 | 0.5096 |

The accuracy of naïve bayes model for Jersey City and Boston is similar to the accuracy obtained in the decision tree which validates our algorithm.

# **ENSEMBLE METHOD**

**Def**: For Ensemble we have decided to use two different method for both the Jersey City and Boston and then compare the results

**Feature set used**: (Neighborhood, Host\_response\_time, Host\_response\_rate, Minimum\_nights, Instant\_bookable, Room\_type, Bedrooms, Availability\_365, Number\_of\_reviews, Price\_level)

**Methods**: For decision tree we have defined the controls as follows: **Method = repeatedcv, number = 10, repeats = 5**. Method called **Adabag** and **Ranger**

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| Jersey City | 73.25% |
| Boston | 65.11% |

The accuracy obtained here is similar to ones obtained in Decision Tree and Naïve Bayes.

Now lets us compare the Precision recall F Measure values obtained

**Jersey City**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7925 | 0.8462 | 0.6731 |
| **Recall** | 0.8077 | 0.8302 | 0.6731 |
| **F Measure** | 0.8000 | 0.8381 | 0.6731 |

**Boston**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7374 | 0.7917 | 0.6176 |
| **Recall** | 0.6887 | 0.8879 | 0.5833 |
| **F Measure** | 0.7122 | 0.8370 | 0.6000 |

# **SUPPORT VECTOR MACHINES (LINEAR)**

For SVM we have converted attributes some of the attributes to dummy columns

**Columns converted to Dummy**: (neighborhood, room type, instant bookable, host\_response\_time)

**Feature set used**: (Neighborhood, Host\_response\_time, Host\_response\_rate, Minimum\_nights, Instant\_bookable, Room\_type, Bedrooms, Availability\_365, Number\_of\_reviews, Price\_level)

**TrControls : Method :** CV, number 10 ; **TuneGrid Parameters** : ( seq(0,2) length :11 )

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Method** | **Accuracy** |
| Jersey City | SVM | 80.77% |
| Boston | SVM | 73.12% |

The accuracy obtained here is similar to ones obtained in Decision Tree, Naïve Bayes. But Jersey City and Boston SVM has the highest accuracy among all other models

**Jersey City**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.7857 | 0.9167 | 0.7308 |
| **Recall** | 0.8462 | 0.8462 | 0.7308 |
| **F Measure** | 0.8148 | 0.8800 | 0.7308 |

**Boston**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values** | **Price Class** | | |
|  | **High** | **Low** | **Medium** |
| **Precision** | 0.8261 | 0.7500 | 0.6379 |
| **Recall** | 0.7170 | 0.7925 | 0.6852 |
| **F Measure** | 0.7677 | 0.7706 | 0.6607 |

High values of prediction explain us that the result obtained has less false positive are present

# **Brief Comparison of all algorithms**

|  |  |  |
| --- | --- | --- |
| **Models** | **Dataset** | **Accuracy** |
| **Decision Tree** | Jersey City | 75.8% |
| Boston | 70% |
| **Naïve Bayes** | Jersey City | 73.25% |
| Boston | 65.11% |
| **Ensemble** | Jersey City | 77.07% |
| Boston | 71.96% |
| **SVM(Highest)** | Jersey City | 80.77% |
| Boston | 73.12% |

# **Conclusion**

* After running the 4 different classification methods, we achieved highest accuracy with linear SVM method for Jersey City and Treebag ensemble method for Boston.
* For both cities we observe that he attributes that affect the price the most are room\_type, neighbourhood, number\_of\_reviews, bedrooms and minimum\_nights
* For **Jersey City**, High priced listings have the following factors: room\_type-entire home/apt/hotel room, Neighbourhood – ward E
* Medium price: bedrooms>=2, minimum\_nights<4
* Low price: room\_type is not entire home/apt
* For **Boston**, High priced listings have the following factors: room\_type- entire home/apt, bedrooms>=2,neighboorhood-BackBay,Bay Village,Charlestown,Chinatown,Downtown,Fenway,Jamaica Plain,Mission Hill,North End,Roxbury,South Boston,South End,West End
* Medium price: minimum nights>23, bedrooms<2,
* Low price: room\_type is not entire apt/house, neighboorhood: Allston,Bay Village,Brighton,Charlestown,Dorchester,Downtown,East Boston,Hyde Park,Jamaica Plain,Longwood Medical Area,Mattapan,Roslindale,Roxbury,South End,West R

# **Reference:**

Insideairbnb.com

Lecture Notes