Airbnb in New York City

A Big Data Analysis for Competitive Advantage

University Of North Carolina at Charlotte

Alekhya Akkinepally
Student ID: 801006369
aakkinepl@uncc.edu

Naga Poorna Pujitha Perakalapudi
Student ID: 801039609
nperakal@uncc.edu

Mounika Yendluri
Student ID: 801039130
myendlur@uncc.edu

Lalith Sandeep Bhavineni
Student ID: 800987686
lbhavine@uncc.edu

Akshay Karai
Student ID: 801038933
akarai@uncc.edu

Alay Chaniyara
Student ID: 801039243
achaniya@uncc.edu

ABSTRACT

Running a business is always about making profits. Today's world is all about Big Data. Business owners are using technologies to gather relevant information which can be used to expand their businesses. This paper is about one such business which happens to grow very rapidly in today's world. It is Airbnb. So, we chose the Airbnb data on one of the most visited city in the world i.e. New York City. We performed analysis on Airbnb data for NYC we got from Kaggle. This paper gives us insights on the data processed and models built on this data and some of variables that can be predicted along with suggestions on how to improve the Airbnb business in NYC along with maintaining a great customer satisfaction. A comprehensive analysis on the dataset, most of the variables, constraints and all attributes that are reliable and useful is performed. Further there are some ideal models found for hypothesis presented by us. Along with that are challenges faced to find the best fit model. The process consisted of hypothesis identification from observing the data initially, get baseline description, identify constraints, make model selection, perform data mining task, implementation of model, results and reasoning, tuning of parameters, find best fits and provide outcomes.

1. INTRODUCTION

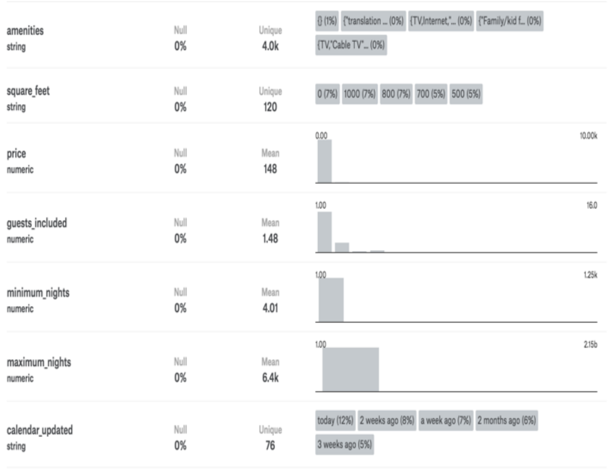
Airbnb is a commercial center where visitors can book housing from a rundown of verified hosts. Enrollment to the site is totally free and there is no cost to post a posting. Utilizing a user UI intended to limit voyaging inclinations, Airbnb offers an appealing, cost-sparing option to hotel appointments & vacation house rentals.

After finding a coveted posting, visitors are incited to agree to accept participation, which gives access to contact the host straightforwardly and also give installment data to a demand. Just once the host acknowledges the exchange and the visitor checks in is the Visa charged, alongside a 6-12% exchange expense from Airbnb.

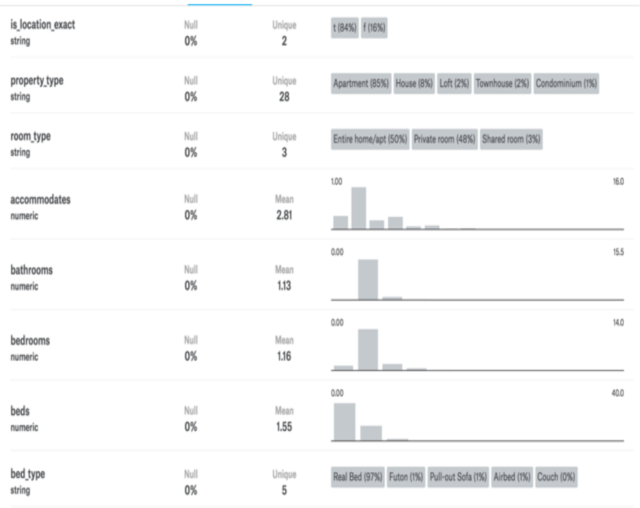
The procedure is comparably straightforward to hosts, where 3 get a notice once a visitor demonstrates enthusiasm for a specific posting and have the choice to endorse or deny the exchange. Once the posting is reserved, the host gets the installment and Airbnb takes a 3% exchange charge. Outline a model that will foresee where will another visitor book their first travel. By precisely anticipating where another client will book their experience of first travel, Airbnb can impart more customized substance to their group, diminish the normal time to first reserving, and better estimate request. We think it would be nice for making price predictor to develop a fair price for its reference.

For example, features such as geographical information, number of bedrooms & bedroom types are considered to have a significant influence on the price. Now in making model, we needed to consider about:

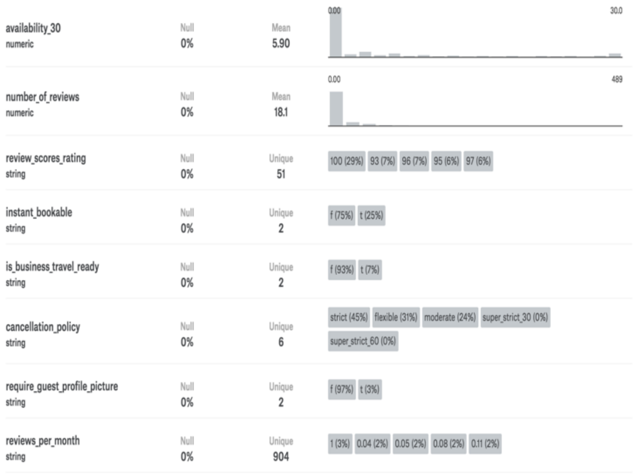
● Size of Data: Some datasets are huge (in excess of 1M factors). Consider diminishing the dataset to an irregular example of 10% to run tests. Moreover, you should join datasets utilizing "client id".
● Quality of Data: Some numerical and downright esteems are absent. You can supplant missing esteems with the media or mean. Elective, you can dispense with the missing esteems by sub setting your dataset. Both methodologies have tradeoff.
● Model of Multiclass: To understand this, we have to manufacture a "multi-class classification" model. This implies you will require a few classifications to predict nations. Some methodologies might be:
1.Break the problem into single classifier to get new reservation which will be processed by classifiers by every destination.
2.Use Modelling Techniques (classification, regression, clustering).



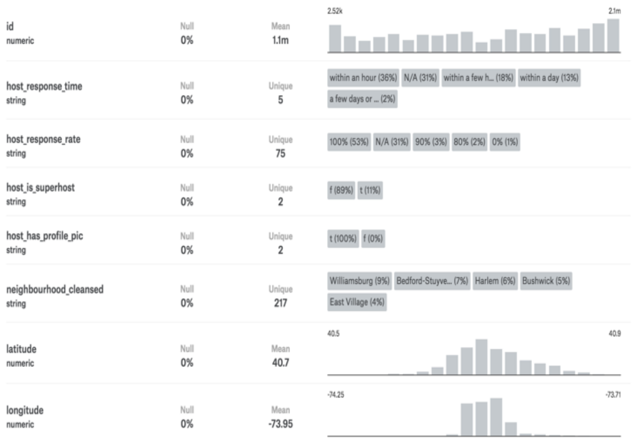
amenities string	Null 0%	Unique 4.0k	{} (1%) {"translation ... (0%) {TV,Internet,"... (0%) {"Family/kid f... (0%) {TV,"Cable TV"... (0%)
square_feet string	Null 0%	Unique 120	0 (7%) 1000 (7%) 800 (7%) 700 (5%) 500 (5%)
price numeric	Null 0%	Mean 148	0.00 10.00k
guests_included numeric	Null 0%	Mean 1.48	1.00 16.0
minimum_nights numeric	Null 0%	Mean 4.01	1.00 1.25k
maximum_nights numeric	Null 0%	Mean 6.4k	1.00 2.15b
calendar_updated string	Null 0%	Unique 76	today (12%) 2 weeks ago (8%) a week ago (7%) 2 months ago (6%) 3 weeks ago (5%)



is_location_exact string	Null 0%	Unique 2	t (84%) f (16%)
property_type string	Null 0%	Unique 28	Apartment (85%) House (8%) Loft (2%) Townhouse (2%) Condominium (1%)
room_type string	Null 0%	Unique 3	Entire home/apt (50%) Private room (48%) Shared room (3%)
accommodates numeric	Null 0%	Mean 2.81	1.00 16.0
bathrooms numeric	Null 0%	Mean 1.13	0.00 15.5
bedrooms numeric	Null 0%	Mean 1.16	0.00 14.0
beds numeric	Null 0%	Mean 1.55	0.00 40.0
bed_type string	Null 0%	Unique 5	Real Bed (97%) Futon (1%) Pull-out Sofa (1%) Airbed (1%) Couch (0%)



availability_30 numeric	Null 0%	Mean 5.90	0.00 30.0
number_of_reviews numeric	Null 0%	Mean 18.1	0.00 489
review_scores_rating string	Null 0%	Unique 51	100 (29%) 93 (7%) 96 (7%) 95 (6%) 97 (6%)
instant_bookable string	Null 0%	Unique 2	f (75%) t (25%)
is_business_travel_ready string	Null 0%	Unique 2	f (93%) t (7%)
cancellation_policy string	Null 0%	Unique 6	strict (45%) flexible (31%) moderate (24%) super_strict_30 (0%) super_strict_60 (0%)
require_guest_profile_picture string	Null 0%	Unique 2	f (97%) t (3%)
reviews_per_month string	Null 0%	Unique 904	1 (3%) 0.04 (2%) 0.05 (2%) 0.08 (2%) 0.11 (2%)



id numeric	Null 0%	Mean 1.1m	2.52k 2.1m
host_response_time string	Null 0%	Unique 5	within an hour (36%) N/A (31%) within a few h... (18%) within a day (13%) a few days or ... (2%)
host_response_rate string	Null 0%	Unique 75	100% (53%) N/A (31%) 90% (3%) 80% (2%) 0% (1%)
host_is_superhost string	Null 0%	Unique 2	f (89%) t (11%)
host_has_profile_pic string	Null 0%	Unique 2	t (100%) f (0%)
neighbourhood_cleansed string	Null 0%	Unique 217	Williamsburg (9%) Bedford-Stuyve... (7%) Harlem (6%) Bushwick (5%) East Village (4%)
latitude numeric	Null 0%	Mean 40.7	40.5 40.9
longitude numeric	Null 0%	Mean -73.95	-74.25 -73.71

Table 1 Columns Variables of NYC Airbnb dataset

2. DATA SET

We have chosen Airbnb data set for New York city and the dataset was taken from "Kaggle" website, it's a non-commercial, independent set of tools and here data which allows you to find how Airbnb really is being used in places of cities around world. The dataset is in csv file where each single file consists of survey or scrape Airbnb website for that city. We've used the listings in NYC dataset, which can be downloaded. Below are some examples of variables used in building models.

host_response_time	string
price	integer price per day
numeber_of_views	integer reviews given by customers
available_30	number of days room is available
instant_bookable	string if we can book immediately
cancellation_policy	string describes levels of cancellations

2.1 Hypothesis

Data set explains about listings house rentals in NYC. It contains 49,348 rows in total. Every row is the full descriptions of a booking and has columns in all, includes availability info, date info, location info, review info, and goes on. The full details about contents of dataset are shown in Table 1.

Hypothesis 1:
To predict the price of room per day using prevailing Airbnb data set. One important thing every host expect is to get best hospitality service for short term accommodation at low price. To make it easy for guest in selection we provide model that predicts price per day using Airbnb data set.

Hypothesis 2: [2]
To analyze whether the cancellation policy effects the reviews given by the user. We try to identify the proportionality between cancellation policy and reviews because using this we can predict the user approach for last minute booking and cancellations, so that the availabilities can also be altered for future use.

Hypothesis 3:
To predict which room types are preferred in different neighborhoods. While searching for places or neighborhoods customer prefer rooms that are reliable, convenient. So, using this we predict different types of rooms available near those neighborhoods.

Objective of the Dataset:

The objective of the dataset is to help customers to choose a standardized place with great hospitality during stay of their time (hotels, houses etc.) at feasible price rates. Also using this data, we wanted to help the hosts and Airbnb to increase their ratings precisely along with their rentals. This paper depends on different attributes such as location, price distribution, bed room variations & other features. Now, we use our analysis to understand previous patterns/trends & predict the future trends in pricing.

2.2 Price Constraints [1]

The main motto is that we tried to know how the price is distributed in the dataset. Figure 1 shows exactly the distribution of the price in our dataset. This one shows the price distribution among the whole dataset, from which we can see that most prices lied in the range (0-50) and then we got peak value at 60(approx.). Later, the graph declines as per distribution within that range, which shows that people seem to have a tendency to reduce the price. In order, to verify our assumption we made 2 bins i.e. price (0-200) range and (200-345) range. We can clearly see that it has very high frequency in the data and it means our hypothesis is plausible.

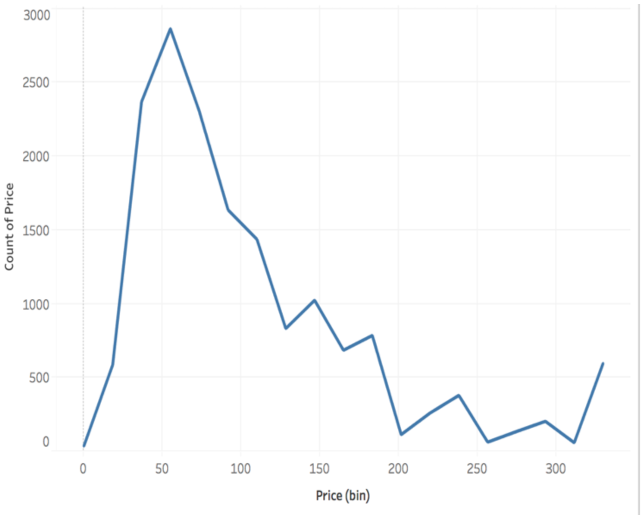


Figure 1: Distribution of Price in Dataset

2.3 Constraints of Location [1]

Discussing about the prices of home then definitely we consider the geographical features at first place, because it varies from urban areas to suburban areas. Now the importance of location is identified from given variables like “latitude”, “longitude”, “is_location_exact”, “neighbourhood_cleansed” from the given dataset. We can observe the distribution of areas average maximum price is seen is mainly located in the places around the Battery Park City which is operated by urban development corporation as it has many casual eateries and famous bars of NYC. We had done exploration on the neighborhood information. This dataset has been classified into 72 different neighborhoods

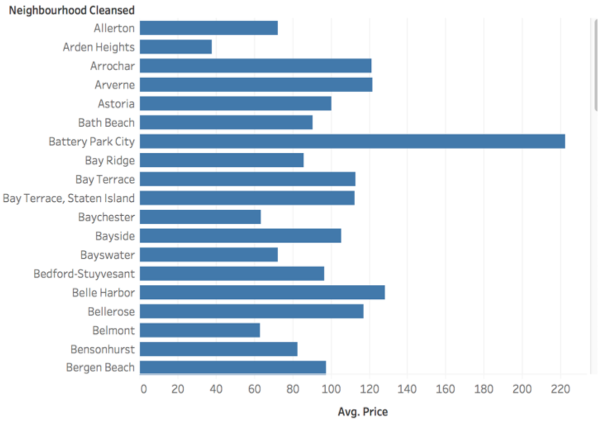


Figure 3. Neighborhoods Vs Average Price.

in NYC like ‘Chelsea’, ‘DUMBO’, ‘Castel Hill’ etc. Calculated the average prices for every neighborhood and plotted those features in Figure 3. It gives us information about the neighborhood which booking place is located will affect the price a lot.

2.4 Constraints of Bedroom [1]

The next most radiant feature which effects the price drastically is bedroom, type of the bedroom and the number of rooms. The multiple bedrooms will respectively have the higher than the single bedroom. To verify this on our dataset, we’ve plotted the price versus bed type, the number of bedrooms, in Figure 4. The plot is as expected as below.

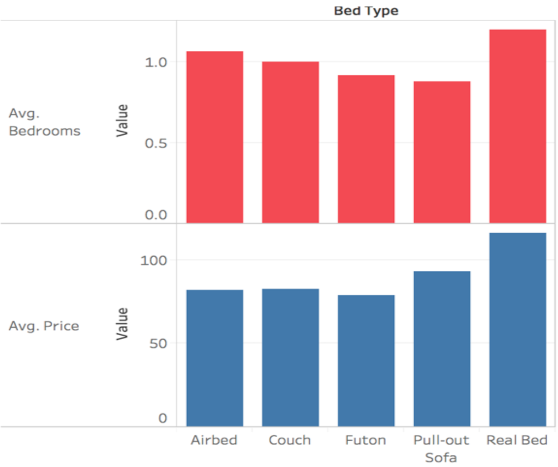


Figure 4. Price Vs Bedrooms/Bed Type

These rooms with many types are likely to be having variant prices also. In this dataset, we have three columns related to types: ‘property type’, “bed type” , ”room type”, also in Figure 5 we observe that rooms with type ”Entire home/apt”, beds with type ”Real bed” and property type ”apartment” tend to have more prices. On the other side, rooms with type “private room”, “Tent” are very cheap.

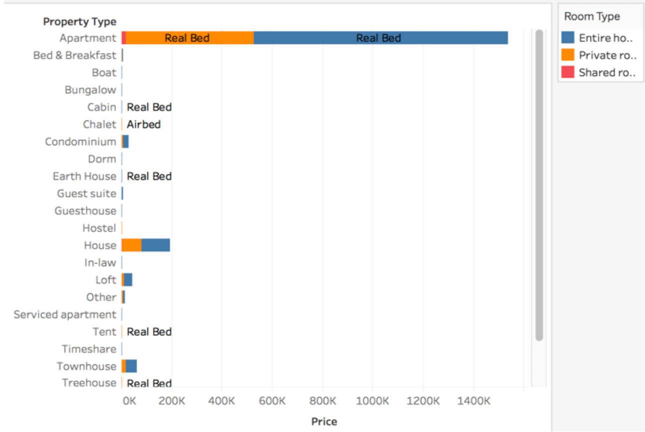


Figure 5. Price vs Property type/Room type/Bed type

2.5 Other Constraints

There are so many features are available other than bedroom and location features in dataset. We go with flow then further decide judge whether those features must be included in the final features by some evaluation results.

3. PROCESS OF BUILDING MODEL

3.1 Preprocessing of Data

At first glance of the dataset, we observed that there are many redundant and irrelevant columns that we will not use in our model. Some columns such as ‘host picture URL’ and ‘host name’ will not make any effect on our predictions of ‘price’. But, columns like ‘square feet’ and ‘bedrooms’ are very likely to influence prices. It implies that it must be a very important thing to do is perform some Exploratory Analysis on the dataset given with so many columns, it must first should help us to finish some filter and clean works, further we can select the best features for our model.

Preprocessing work is divided into different ways: [11]

(1) Removed outliers that are more than 3 standard deviations from the mean-

R Code:

```
#outliers (price)
boxplot(data$price)
bench ← 175 + 1.5*IQR(data$price)
data$price[data$price > bench] ←bench
summary(data$price)
```

(2) Now we have removed all the missing values, the next step in our preprocessing was to check the outliers and normalize the data by removing the outliers.

R Code:

```
#outliers (price)
boxplot(data$price)
```

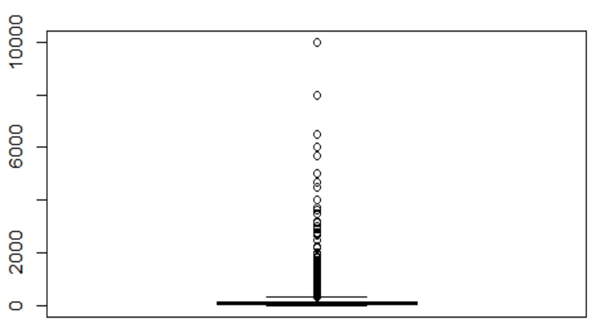


Figure 6. Boxplot before removing outliers

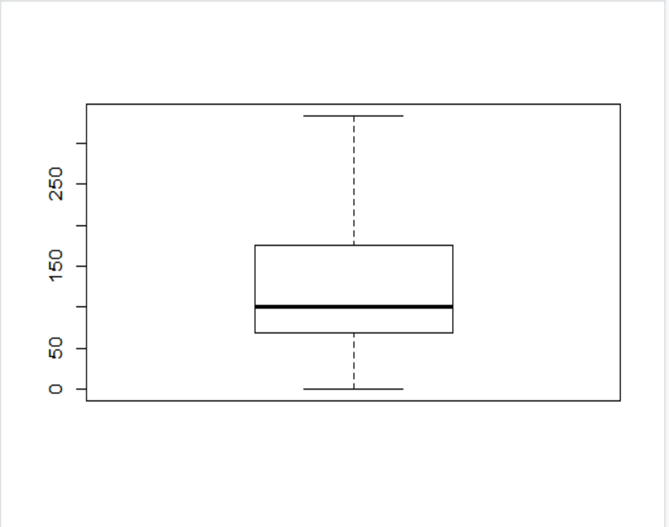


Figure 7. Boxplot after removing outliers

(3) Removed the missing data inputs with percentage of missing values greater than 10% and drop these columns. In this step the no. of inputs reduced from 60 to 38. This reduction implies to be very biased, but we used R programming in this process.

(4) Replaced the rows which has missing values just to make sure there are no NAN values in data. In this step. we replaced approx. 42000 records of data with median.

R Code:

```
summary(data$review_scores_rating)
data$review_scores_rating[is.na
(data$review_scores_rating)] ← 96
 summary(data$reviews_per_month)
data$reviews_per_month[is.na
(data$reviews_per_month)] ←1
```

(5) We removed columns that doesn't contain too much information to impact prediction. For example, Columns which are not used for huge number of missing variables are below: "review_scores_accuracy", "review_scores_cleanliness", "review_scores_checkin", "review_scores_communication", "review_scores_location", "review_scores_value".

(6) Here, after this step, the original number of inputs is reduced from 95 to 60.

(7) Next, we performed correlation detection and only one among them is sustained which has high correlation.

(8) Further, the data is divide it into 2 halves and the first 70% of data is the training set and next 30% of data is test set.

Sample Business Use Case:

Airbnb needs help in providing the best suitable accommodations to customers thus leading to increase in its customer base. It will also help his team to provide better services to its customers.

New York City is the most populous city and is described as the cultural, financial, and media capital of the world. The city has heavy influx of visitors and tourists. Airbnb is looking at different options to increase its customer base in the city. The company is looking for details regarding the following criteria –

1. The neighborhoods in which number of accommodations needs to be increased.
2. Provide better pricing options to attract customers.
3. Number of persons allowed to share a room.
4. Customers preference to room types.
5. Find the hosts who have which get highest rating with more no. of reviews.

3.2 VISUALIZATION

From the below Figure 8 visualization we can clearly say that Brooklyn is the neighborhood which received more number of reviews

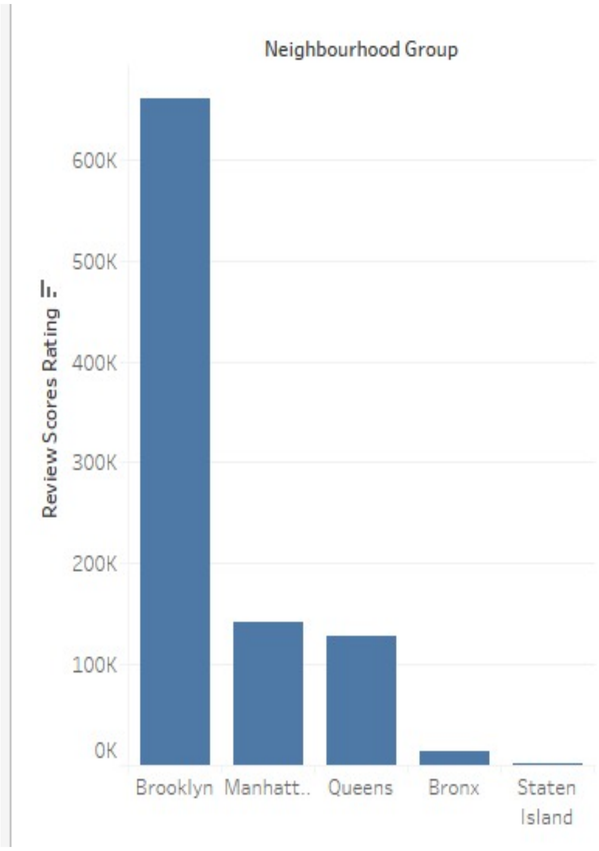


Figure 8 Neighborhood group vs Reviews score rating

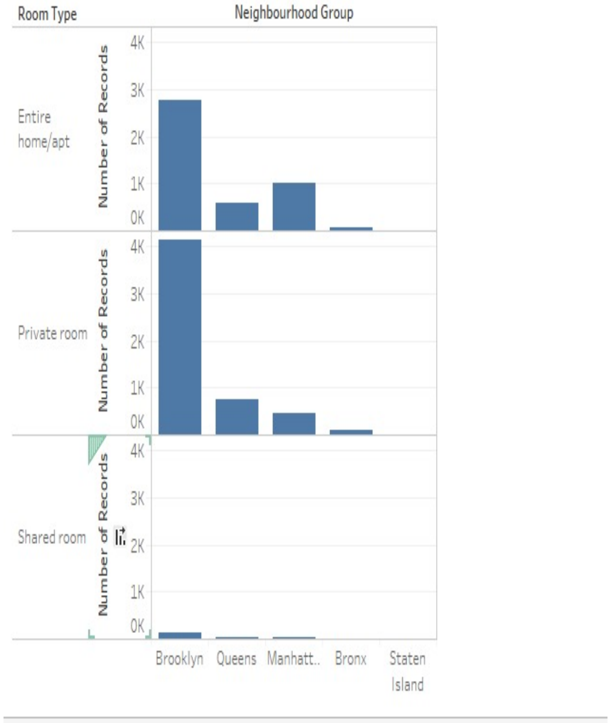


Figure 9. Neighborhood group vs Room type

This visualization Figure 9. shows that the private rooms are the one which are booked by the customers

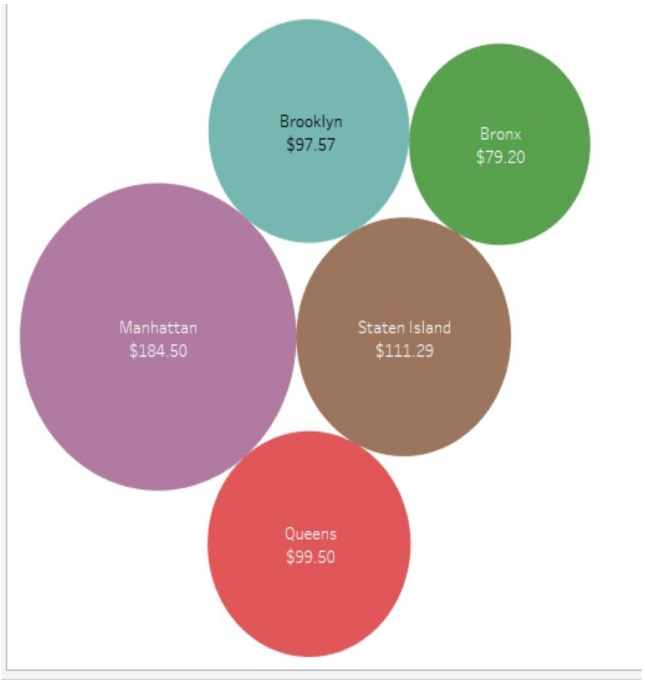


Figure 10. Bubble chart for AVG price in neighborhood

This bubble chart in Figure 10. shows clearly the average price for different neighborhoods.

The scatter plot Figure 11. shows the trend of the calendar updates done by the host

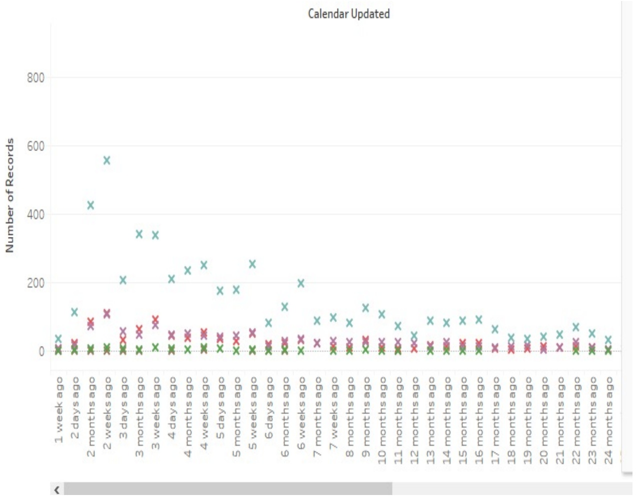


Figure 11 Calendar update vs host

From analysis we have made on the Airbnb data set, we can say that in Brooklyn neighborhood people preferred private room rather than shared room or entire apartment and the review score rating is high in Brooklyn neighborhood. In Brooklyn neighborhood the average price of a room is $97 which is second lowest when compared to other neighborhood areas. So, we can say that the Brooklyn is the neighborhood where more number of booking of room are done. Therefore, Airbnb should increase the number of host in the Brooklyn neighborhood since the bookings are more. If we consider Bronx neighborhood the average price of a room is $ 79 which is very less compared to other neighborhood areas but the review score rating and booking of a room is very less when compared to another neighborhood. So, Airbnb should look in to the problem and rectify the problem so that it results in good customer review score rating which improves the booking of the room in the neighborhood area.

3.3 DATA MINING TASK SELECTION

We have so many different modelling techniques which can be applied on the data set which was pre-processed for analyzing and predicting the data. Now we select the best modelling techniques to prove that our hypothesis is really best for that applicable model.

Classification

Grouping is a directed learning procedure that allots things in an accumulation to target classifications or classes. A model or classifier is developed to foresee all out marks (the class name traits). The objective of characterization is to precisely anticipate the objective class for each case in the information.

We utilized the standard from Airbnb. Since the NDF

US check more than 80 percent of the train set, the standard just predicts these two nations. It predicts the NDF and US then again, as NDF, US, NDF, US, ...and so on. The expectation score on the validate set was 0.78640. This gives benchmark appears to be to some degree trifling, however it got a decent score looked at to other pattern show we made independent from anyone else like direct relapse, which just got a low score of 0.49859. In this way,
it was a baseline to enhance the classification model
Classification can be implemented using the following algorithms

1. Classification tree
2. Logistic Regression
3. Neural Networks
4. Naive Bayes

Naive Bayes [9]
Naive Bayes is from Bayesian theorem. It is best suited for classification problems. Naive Bayes classifiers are much faster than any other sophisticated methods. This require small amount of the training data in order to get trained. It is very simple and easy to implement.

Neural Networks [12]
Neural networks have a huge potential. Neural network provides the best solution to many problems which enables the computer to learn from the observational data. Neural networks process information in a similar way the human brain does. Neural network has the capability of working even after a part of networks is damaged and also itself organization and fault tolerance.

Association
Association is an information mining capacity that finds the likelihood of the co-event of things in a set of collection. The connections between co-occurring things are communicated as association rules. The use of association rule algorithms compared to decision tree is it has standard choice which can exist between any of the variables, but any other tree can't build rules for more attributes as they are restricted to single variable. So, here association rules have different variables which give different results and conclusions. The objective is to discover relationship of things that happen together more frequently than you would anticipate from an arbitrary inspecting of all potential outcomes. Affiliation principles can be accomplished utilizing Apriori Algorithm.

Prediction
Predictive model is utilized to predict the reaction variable esteem based on predictor indicator variable. Continuous models function like predicting the missing or unknown values. There are many different predictive methodologies like: regression tree, neural network, k- nearest neighbors, multiple linear regression etc. Every technique has its own special highlights and the determination of one is ordinarily controlled by the idea of the factors included.

Clustering
Cluster analysis is the undertaking of collection an arrangement of articles such that items in a similar trend are more comparable to each other than to those in different groups. It is a primary errand of exploratory mining of data, and a typical procedure for data analysis, utilized as a part of numerous fields, including machine learning, analysis of picture, data recovery, bioinformatics, information pressure, and PC illustrations.
Now in first step we do data partition on data set and further it groups the data based on it similarity terms, as it is flexible enough to changes and could be able to give support in getting required features from different groups we use this clustering.
Clustering can be performed in many types:

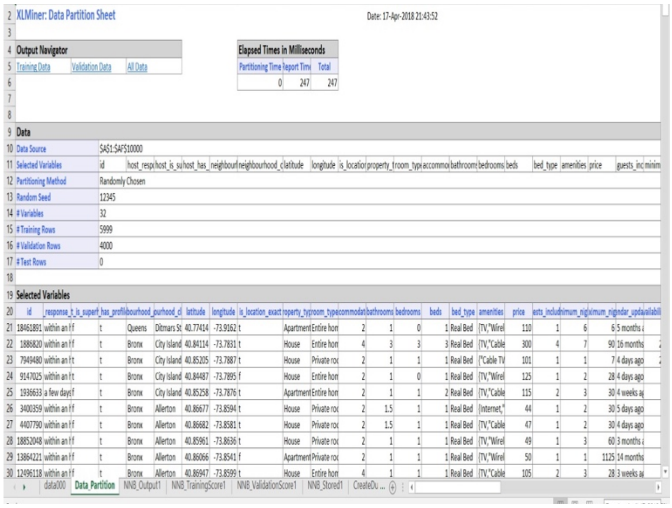
1. K – means clustering
2. Hierarchical clustering

3.4 DATA PARTITION

Data Partitioning is the formal procedure of figuring out which information subjects, information event gatherings, and information attributes are required at every datum site. It is a deliberate procedure for assigning information to information destinations that is done inside a similar basic information engineering. It is also means sensibly and additionally physically dividing information into fragments that are all the more effectively kept up or got to Dividing of information helps in execution and utility preparing.

Steps in Data Partitioning:

- The data partitioning can be dealt with in 2 different ways
- There can be a partition variable that can divide data into validate and training sets.
- The data partition should be possibly random then there will be choice of mentioning seed of randomization.
- The training set utilized to assemble the model, and the approval segment is utilized to perceive how best the model does when connected to new information.
- The validation set is utilized for important choice i.e. selection of model and the test for evaluating the last model error prediction.



2 XLMiner: Data Partition Sheet — Date: 17-Apr-2018 21:43:52

4 Output Navigator

5 Training Data | Validation Data | All Data

Elapsed Times in Milliseconds		
Partitioning Time	Report Time	Total
0	247	247

9 Data

Data Source	$A$1:$AF$10000
Selected Variables	id, host_resp, host_is_su, host_has_, neighbourl, neighbourhood_c, latitude, longitude, is_locatior, property_t, room_typ, accommo, bathroom, bedrooms, beds, bed_type, amenities, price, guests_inc, minim
Partitioning Method	Randomly Chosen
Random Seed	12345
# Variables	32
# Training Rows	5999
# Validation Rows	4000
# Test Rows	0

19 Selected Variables

20	id	_response_	t_is_supert	has_profil	bourhood_	ourhood_cl	latitude	longitude	is_location_exact	roperty_ty	room_type	commodat	bathrooms	bedrooms	beds	bed_type	amenities	price	ests_includ	nimum_nig	ximum_nig	ndar_upda	vailabili
21	18461891	within an h	f	t	Queens	Ditmars St	40.77414	-73.9162	t	Apartment	Entire hom	2	1	0	1	Real Bed	{TV,"Wirel	110	1	6	6	5 months a	
22	1886820	within an h	f	t	Bronx	City Island	40.84114	-73.7831	t	House	Entire hom	4	3	3	3	Real Bed	{TV,"Cable	300	4	7	90	16 months	2
23	7949480	within an h	t	t	Bronx	City Island	40.85205	-73.7887	t	House	Private roo	2	1	1	1	Real Bed	{"Cable TV	101	1	1	7	4 days ago	2
24	9147025	within an h	t	t	Bronx	City Island	40.84487	-73.7895	f	House	Entire hom	2	1	0	1	Real Bed	{TV,"Wirel	125	1	2	28	4 days ago	
25	1936633	a few days	f	t	Bronx	City Island	40.85258	-73.7876	t	Apartment	Entire hom	2	1	1	2	Real Bed	{TV,"Cable	115	2	3	30	4 weeks ag	
26	3400359	within an h	f	t	Bronx	Allerton	40.86677	-73.8594	t	House	Private roo	2	1.5	1	1	Real Bed	{Internet,"	44	1	2	30	5 days ago	
27	4407790	within an h	f	t	Bronx	Allerton	40.86682	-73.8581	t	House	Private roo	2	1.5	1	1	Real Bed	{TV,"Cable	47	1	2	30	4 days ago	
28	18852048	within an h	f	t	Bronx	Allerton	40.85961	-73.8636	t	House	Private roo	2	1	1	1	Real Bed	{TV,"Wirel	49	1	3	60	3 months a	
29	13864221	within an h	f	t	Bronx	Allerton	40.86066	-73.8541	f	Apartment	Private roo	2	1	1	1	Real Bed	{TV,"Wirel	50	1	1	1125	14 months	
30	12496118	within an h	f	t	Bronx	Allerton	40.86947	-73.8599	t	House	Entire hom	4	1	1	1	Real Bed	{TV,"Cable	105	2	3	28	3 weeks ag	

Figure 12. Data Partition.

4. IMPLEMENTATION OF MODELS

Any kind of models can be implemented on every hypothesis but only the most relevant model with the optimized solution should be given high priority. Variant variables are selected as inputs and we also get different output variables which are really important as they play key role in decision making for each hypothesis

Below are the different modelling techniques that are considered for each hypothesis.

Hypothesis 1:

For feasibility of customer in selection we provide this model which predicts price per day in New York city using Airbnb data set.

Dependent variables:

latitude, longitude, accommodates, bathrooms, bedrooms, bed, guest_included, minimum _nights, maximum_nights, availability_30, number_of_reviews, review_scores_rating, reviews_per_month, room_type.

Independent variable:

Price(bin)

Rationale:

As we want to predict the price we explore the dependent variables. All input variables mentioned above are those variables that effect the price instantly. So, we consider the target variable as **Price bin**.

Algorithms used:

Clustering, Neural Network, Logistic Regression

Limitations of the modelling techniques:

Neural Network is a good model because it gives the same result always for output variable. Clustering we cannot predict the price in which there are so many neighborhoods, also it is difficult to predict the number of clusters (K-Value). But more data can give better result.

Best model:

Out of the models we looked at Neural Network is best modelling technique that can be applied to this hypothesis as we use this for predicting future trends using present trend basing on many dependent variables. Also, we get a good accuracy in Logistic Regression technique. But as we use only 10K data inputs we may have different results when we use the whole 500K data. [6]

Hypothesis: 2

Just to know the last minute cancellations and bookings of user we can use this hypothesis to do analysis for cancellation policy and the reviews given by the user.

By this prediction the availabilities of room can also be altered for future use.

Dependent variables:

latitude, longitude, accommodates, bathrooms, bedrooms, bed, guest_included, minimum _nights, maximum_nights, availability_30, review_scores_rating, reviews_per_month, cancellation_policy_flexible, cancellation_policy_moderate, cancellation_policy_strict, cancellation_policy_super_strict_30

Independent variable:

Number_of_reviews

Rationale:

Here our predictor variables are cancellation policy along with all the input variables. Such that we get accurate output of **number_of_reviews**

Algorithms used:

Multiple Linear Regression, Neural Network, Regression Tree

Limitations of the modelling techniques:

Neural Network is not best as it does not always result in number_of_reviews as output variable.

Best model:

Multiple Linear Regression is the good technique which we can apply to this hypothesis to find the number of reviews per room by user by using Multiple linear regression as there are different dependent variables on one independent variable (cancellation_policy). Here, we can find the effect on reviews by cancellation_policy.

Hypothesis 3:

If customer is in search for the places or neighborhoods they prefer to identify the rooms that are having high flexible, reliable, convenient.

By implementing this we predict different types of rooms available near those neighborhoods

Dependent variables

latitude, longitude, accommodates, bathrooms, bedrooms, beds, review_scores_rating, reviews_per_month, number_of_reviews, neighbourhood_group_Bronx, neighbourhood_group_Brooklyn, eighbourhood_group_Manhattan, neighbourhood_group_Queens, neighbourhood_group_Staten Island

Independent variable:

room type

Rationale:

Here our predictor variables are neighbourhood_group_* along with all the input variables. Such that we get accurate output of **type of rooms**

Algorithms used:

Neural network, Naïve Bayes, Clustering

Limitations of the modelling techniques:

Naïve Bayes cannot be performed on this hypothesis as the features are not independent which given in the class label which can't make right decisions.

Clustering is not good for predicting the value of a dependent attribute. Because we can't group the areas of neighborhood as the categories are 72 where it wasn't supported in tool.

Best model:

Neural Network is the best technique that can be applied to the given hypothesis to find the room type with highest number of accommodates, bedrooms, beds. etc.

5. RESULTS

As we used XL miner we have certain constraints, so we have taken only 10,000 data entries for evaluating the model which we have built. We also grouped certain other variables/ parameters while building the model as it is difficult to run on so many categories such as Price(bin), review, neighbourhood(group).

5.1 Hypothesis-1 Result

After implementing Logistic regression, Neural networks, KNN on the data set we observed that ROC curve has 0.90 accuracy i.e. 90% by performing it on train and validation sets of Neural Networks. Also the error rate is appreciably low which is a good sign to accept that model is best fit for data set. Also, we get a good accuracy in Logistic Regression technique. But as we use only 10K data inputs we may have different results when we use the whole 500K data.

The figure 13 and 14 below displayed represent Receiver Operating Characteristic curve(ROC) on Training, Validation and Test datasets using Logistic Regression.

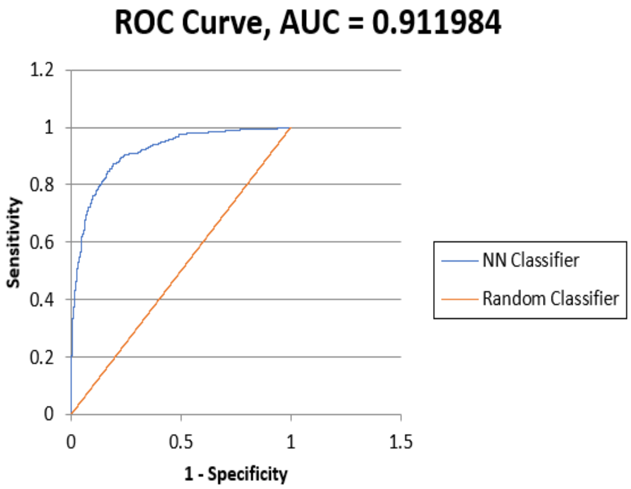
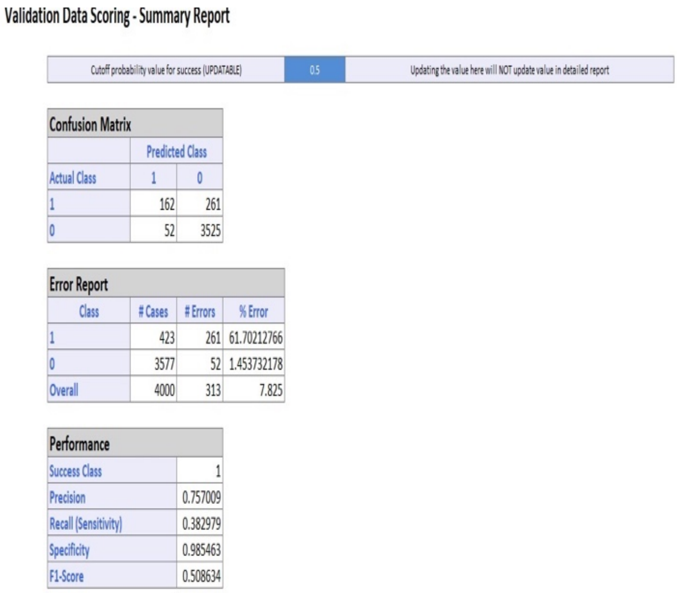


Figure 13. ROC curve on Validation Data



Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.5	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	162	261
0	52	3525

Error Report			
Class	# Cases	# Errors	% Error
1	423	261	61.70212766
0	3577	52	1.453732178
Overall	4000	313	7.825

Performance	
Success Class	1
Precision	0.757009
Recall (Sensitivity)	0.382979
Specificity	0.985463
F1-Score	0.508634

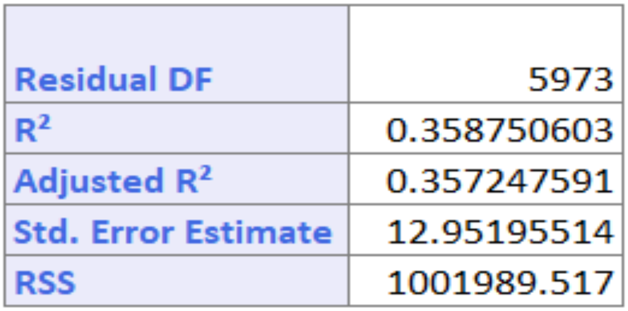
Figure 14. Error % output

5.2 Hypothesis-2 Result

We evaluated the results from model that was implemented on second hypothesis using Multiple linear regression. Now we will consider the universal parameters to evaluate this model they are R^2 and Mean Absolute Error (MAE) values, both of them are most common statistical measurements of multiple linear regression model.

We just compared the models on both MAE and R^2 score on the validation set.

The result is shown in Figure15 and 16, from which we can see that the Multiple Linear Regression model gives where the ROC curve has very low accuracy. This represents that cancellation policy doesn't affect the number_of_reviews of the user. It means last minute bookings and cancellations are highly possible which will never be a problem for the Airbnb to estimate their bookings based on user reviews.



Residual DF	5973
R^2	0.358750603
Adjusted R^2	0.357247591
Std. Error Estimate	12.95195514
RSS	1001989.517

Figure 15. R^2 and Error.

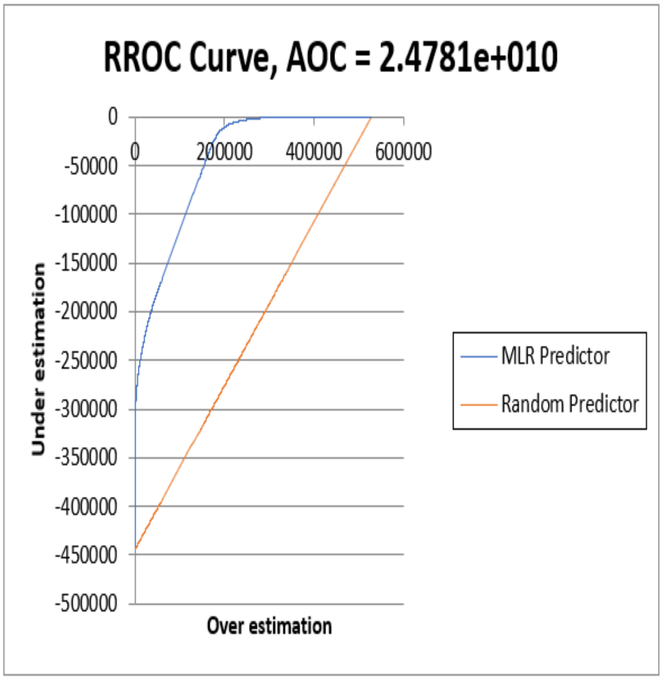
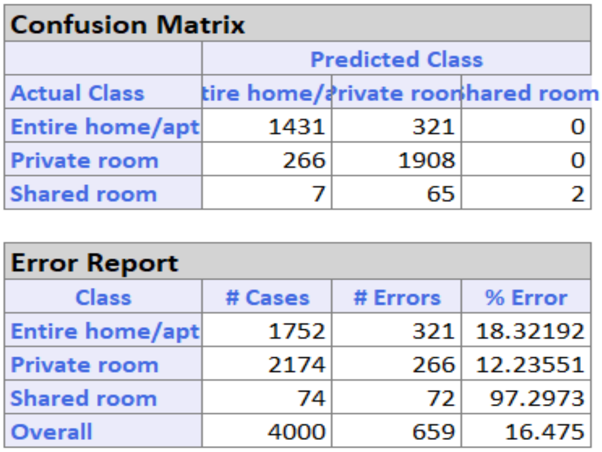


Figure 16. ROC curve on Validation Data

5.3 Hypothesis-3 Result

As we have implemented three models on these hypothesis, out of all these models Neural network is the best algorithm classification tree is the best model since the overall rate is 16.475%

From the figure 17we get that error rate in the shared room is very high compared to other room because we have less number of values in the shared room in the Airbnb dataset**.**



Confusion Matrix			
	Predicted Class		
Actual Class	tire home/a	rivate roon	hared room
Entire home/apt	1431	321	0
Private room	266	1908	0
Shared room	7	65	2

Error Report			
Class	# Cases	# Errors	% Error
Entire home/apt	1752	321	18.32192
Private room	2174	266	12.23551
Shared room	74	72	97.2973
Overall	4000	659	16.475

Figure 17 Error report on Validation Data Set.

6. EXECUTIVE SUMMARY

1. We built a Logistic Regression model for New York dataset with price range as dependent variable. It is a categorical variable and other numerical variables have been used as independent variables. This model predicts the price range in the future.
2. We tried to analyze whether the cancellation policy effects the reviews given by the user. Multiple linear regression has been utilized to build the model. It is found that Cancellation policies are fairly spread out but, it doesn't make a big difference to most people.
3. In this hypothesis we tried to predict which room-types are preferred in different neighborhoods. The modelling technique used here is Neural Networks. Room type is the dependent variable and neighborhood, number of accommodates, bedrooms are independent variables.

7. RECOMMENDATIONS

1. We have used XL Miner for the analysis, as it is still being used in the IT companies, but a limitation of the XL Miner tool is that it could only take 10,000 observations and in real world the data would have a large number of observations. So, it is resourceful to use other software's like SaS, Spark.
2. With sufficient data we can work on one of our hypothesis mentioned in the initial phase, which is

to find the seasonal trends in the data. Better analysis will be possible if there is large enough data. The user can make better predictions if there are large number of records. Different metrics like specificity, sensitivity should also be considered along with accuracy and ROC to evaluate models.

3. Advanced planning is required for Big data analysis. The organization needs to plan well in advance regarding its hardware requirements and business needs when performing this kind of analysis. Developing a business strategy for the entire organization is much useful rather than starting out with smaller independent units.
4. An organization becomes an Analytics 3.0 competitor if it makes business decisions based on analytics and Big data. The company should have a Chief Analytics Officer who is focused on analytics driven approach of the organization.
5. Airbnb must analyze and investigate how to keep prices under check in certain boroughs like Manhattan and Staten Island.
6. Increase number of accommodations in certain neighborhoods that attract more visitors.
7. Check the market price and the features being provided by rival companies.
8. Provide a year wise data visualization for the data set to predict the seasonal and event-based predictions.

8.REFERENCES

[1] Predicting Airbnb Listing Prices with Scikit-Learn and Apache Spark, Nick Amato. 2016

[2] Predicting Purchased Policy for Customers in Allstate Purchase Prediction Challenge on Kaggle, Saba Arslan Shah, Mehreen Saeed

[3] Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery, pp 55-86, Jiawei Han, H. C. (2007)

[4] The Secret of Airbnb's Pricing Algorithm. IEEE Spectrum, DAN HILL (August 2015)

[5] Approximating a collection of frequent sets. In: Proceedings of the 2004 ACM SIGKDD international conference knowledge discovery in databases (KDD'04), Seattle, WA, pp 12–19, Afrati FN, Gionis A, Mannila H (2004)

[6] AY Ng, MI Jordan "On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes

[7] A Krogh, J Vedelsby "Neural network ensembles, cross validation, and active learning"

[8] CV Subbulakshmi, SN Deepa, "Comparative analysis of XLMiner and WEKA for pattern classification"

[9] Outline Systems, Solver Naive Bayes

[10] Selva Prabhakaran, Missing Value Treatments

[11] Gourab Nath, Outlier Treatment in R - Part 1 - Discarding Outliers

[12] Outline Systems, Neural Networks Solver