**IDS 575 Project Final Report**

Prathyusha Dasari

**UIN:** 667122494

**Appendix:**

|  |  |
| --- | --- |
| Pg. No | Content and Figures |
| 2. | Fig.1 |
| 3. | Fig.2 |
| 4. | Fig.3 |
| 5. | Fig.4 |

**Title:**

Overall Satisfaction level of Airbnb Chicago customers.

**Objective:**

To predict the Overall Satisfaction level of the customers in Airbnb Chicago.

**Airbnb:**

Airbnb is an online hospitality service, that lets people lease or rent their houses. It has over 3,000,000 lodging listings in 65,000 cities and 191 countries. It does not own any lodging, it is merely a [broker](https://en.wikipedia.org/wiki/Broker) and receives percentage service [fees](https://en.wikipedia.org/wiki/Fee) ([commissions](https://en.wikipedia.org/wiki/Commission_(remuneration))) from both guests and hosts in conjunction with every booking. Hosts make money by renting out their homes, and Airbnb takes a cut too. The cost of lodging is set by the host.

**Hypothesis:**

This dataset has 13 independent variables. Using these independent variables our aim is to predict the response variable – Overall satisfaction level.

**Data Summary:**

Number of Independent variables – 13

Number of Rows – 4989

Response Variable – Overall Satisfaction

**Data Source:**

Data has been taken from the tomslee.net website.

Below is the link for the data source.

**Link:** http://tomslee.net/airbnb-data-collection-get-the-data

**Variable Description:**

Response Variable: overall satisfaction

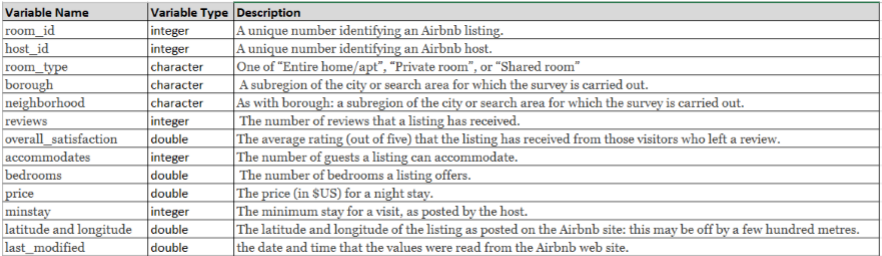


Fig.1

**Method:**

Below are steps that needs to be followed to predict the response variable.

1. Data Cleaning:

We detect and remove inaccurate records or null values from the database and then modify or delete the records as needed.

2. Data Splitting:

We then split the dataset into Training and Test data for cross validation purposes. The training data is used to develop the predictive model. After building the model we then use test data to evaluate the performance of the data.

We usually split the data in the ratio of 60:40 or 70:30.

3. Build a Model:

Since the response variable is a categorical variable, we build three different models.

They are:

1. Multiple Linear Regression
2. Random Forest
3. Support Vector Method

4. Predicting Response variable:

After building the regression model we predict the response variable - Overall satisfaction level by using test data.

**Data Cleaning:**

First I have imported the Chicago AirBnB dataset into R and explored the data. The data had missing values in the form of blank spaces, so I have replaced them with NA in excel.

The total number of missing values in the dataset are 5234. Therefore, removed all the NA values and replaced them with mean of the predictor variable.

**Data Splitting:**

Then, I have Split the dataset into Training and Test data, in the ratio of 70:30.

**Build a Model:**

Since the Response variable is numeric, I have chosen three different models, to predict the response variable.

**Model 1:**

**Multiple Linear Regression**

[I have chosen Multiple Linear Regression model, because it allows us to model the relationship between a scalar dependent variable and multiple explanatory variables.]

**Variable Selection:**

If prediction performance is the main goal we can perform variable Selection. In variable selection redundant predictors can be removed, because unnecessary predictors will add noise. Cost is also a factor that needs to be considered. If the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Below are the Variable Selection methods used :

1. Forward Selection
2. Backward Selection

Below table shows the count of the variables that are selected as highly significant, using the above two methods.

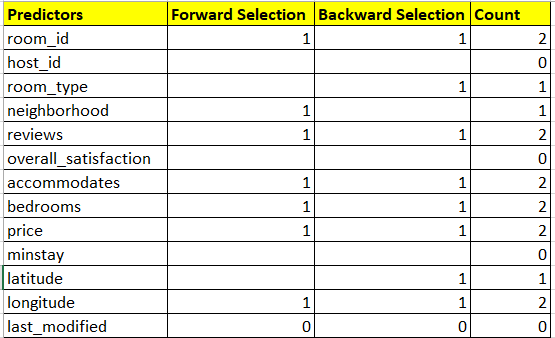


Fig.2

The common variables from both methods are chosen. They are room\_id, reviews, accommodates, bedrooms, price, longitude.

The below multiple linear regression model is built:

lin\_reg <- lm(overall\_satisfaction ~ room\_id +reviews + accommodates + bedrooms + price + longitude, data = train\_data)

Later, I have tested this model on the test data and found RMSE and accuracy.

**The RMSE value for Linear Regression is 6.630567.**

**The Accuracy of the model is 93.36943**

**Model 2:**

**Random Forest**

[I have chosen Random Forest as my next model to predict the satisfaction level, because Random decision forests, can be used for classification and regression. And they help in correcting overfitting in training dataset.]

Built the below model, using all variables and removing room\_type and neighborhood, as they are categorical variables.

rf\_fit <- randomForest(overall\_satisfaction ~ . -room\_type -neighborhood, data = train\_data, ntree = 100, proximity = TRUE, replace= TRUE, sampsize = nrow(train\_data), importance = TRUE)

Upon testing the model with test data, I found the RMSE value and accuracy.

**The RMSE value for Random Forest model is 6.207239**

**The Accuracy of the model is 93.79276**

**Model 3:**

**Support Vector Method**

[I have chosen Support Vector Method as my next model to predict the satisfaction level, because SVMs are also one of the methods that can be used to perform regression.]

I have then built an SVM model, using all the predictors.

svm\_fit <- svm(overall\_satisfaction ~ ., data = train\_data, cost = 100, gamma = 1, kernel = "linear", type = "eps-regression")

Upon testing the model with test data, I found the RMSE value and accuracy.

**The RMSE value for Support Vector Method model is 7.064051**

**The Accuracy of the model is 92.93595**

**Comparing the models and choosing best model**

After performing regression on all the three models, in order to choose the best model, I have taken into consideration RMSE.

Root Mean Square Deviation or Root Mean Square Error (RMSE) is the measure of accuracy. It is a difference between predicted values of the model and actual values.

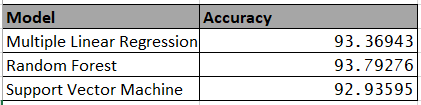
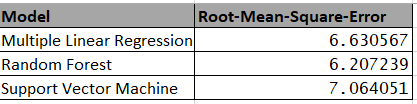


Fig.3

The RMSE for Random Forest is least and Accuracy is high, compared to other models.

**Conclusion:** We chose **“Random Forest”** as the best model

**Random Forest - Output Interpretation:**

By analyzing the importance function, reviews predictor variables are having more significance in predicting the Overall satisfaction level of the customers in the AirBnB dataset.

Random Forest is chosen as the best model. However, Random Forest models need more trees, to give accurate predictions, which make the model slower.

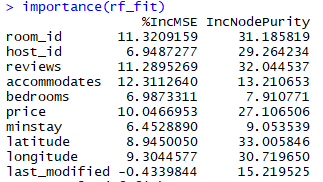


Fig 4

**Key challenges and Learnings**

* The predictor variable borough had no data. Therefore, I have eliminated the variable borough.
* The Dataset has two categorical predictor variables, room\_type and neighborhood. Performing multiple linear regression on a data which has at least one categorical variable is not possible. These two categorical variables were however eliminated in the variable selection method, because they had least significance. However, it would have gone wrong, i.e. the model wouldn’t have been run, if the categorical predictor variables were significant.
* Gained hands on experience in building models and testing them in R.
* Learnt Model selection
* Using variable selection methods in multiple linear regression has improved the accuracy, which worked well.