Report for Assignment 1 Machine learning

**Introduction**

This assignment is to predict a certain attribute from data that has been provided using Machine Learning. Machine Learning (ML) uses programming languages used in computers to learn from data and predict either values or functions depending on the given target that has been chosen.

**What is Machine Learning and why use it?**

Machine Learning is used for issues or problems that are too complex for normal or traditional approaches to predicting, or these issues have no known algorithm for tackling such a task. With newer technology like speech recognition (which is very hard to write programs for) used in our smart phones or other various software's around the globe, machine learning can recognise thousands of words in different languages spoken in multiple dialects… instantly. By writing an algorithm that learns by itself, it is a faster and more accurate way of completing objectives, and regarding speech recognition, the data provided for the algorithm is thousands if not millions of recorded words/syllables/patterns.

Machine learning can also help humans learn, by being studied to find out what these algorithms have learned, revealing mistakes or room for improvement in certain objectives, leading to better understanding.

**The Chosen Data**

Airbnb is a privately held global company, offering services in an online marketplace. Airbnb offers hospitality for travellers to stay in a home in their spare room, or an accommodation owned by someone. Think of Airbnb as a more personal and homelier hostel, offering rooms at prices that vary from cheap to expensive and exclusive, the cheap rooms aren't anything like hostels as they are usually rooms in homes where families might stay, bringing the feeling of home to people traveling. It also gives travellers an insight to homes in other countries, experiencing the cultural difference in the way people live abroad, a different experience from corporate hotels whom are practically the same in every city they present themselves. This data set was collected from the English city of Bristol, and has recorded values under the attributes of:

***Id, name, host_id, host_name, neighbourhood, postcode, latitude , longitude, property_type, room_type, accommodates, bathrooms, bedrooms, beds, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, calculated_host_listings_count, availability_365,***

**Objective**

The objective is to predict an attribute value, using chosen attributes from which research will find useful or a good relationship to the chosen factor. Predicting an attribute that is of use could potentially help or be used by the company or software used to find cheap rooms based on certain factors, such as Skyscanner uses algorithms to find cheap flights from data collected by commercial travel companies.

## Problem

Taking into consideration the users, attributes that are most important to them will be closely studied so that this code is useful regarding customers of the company or their main objective, and that is finding somewhere to stay on their travels, both safe and satisfactory.

Before starting this assignment, certain issues will have to be considered, which could prove to be difficult in finding an accurate prediction, such as; the data having missing values, changing data from numerical to non-numerical values, and vies-versa, working with both numerical and non-numerical values, combining them for the final result, using a number of methods to see which ones are better than other, looking further into these methods to see if they perform well in the future with unseen data, and to see if their results are accurate when doing calculations on positive results such as rout means square error (RSME).

This project has been coded in Python programming language and submitted in a Jupiter Notebook file.

## Data summaries and Visualization - Summarising and visualising your data set (5)

Firstly, the data set must be visualized and summarized. This starting point highlights and dissects the data set so it can be better understood. It is important to know data very well before working with it, to save any issues presenting themselves at times that effect the algorithm.

Using histograms, a general idea of the values for some of the attributes in the data set. Only selecting a few, it is a good starting point for seeing what attributes hold the highest/lowest values. Then the whole data set is studied in regard to how many nulls are present, as this is one of the issues discussed. Because the of number of attributes in the data, it is a good idea to see what each of them individually are full of. This is a very in-depth look to the data set right down to each value in each attribute.

Going with the target of prediction price, certain attributes must be dropped because they contain non-numerical values. Location, Bedroom sizes, and review are what customers look at when deciding where they would to stay, other than the photographs (which are generally viewed just as much as these attributes) most customers care about the price of their stay, and also question this based on reviews or location. Locations and reviews help decide if somewhere is worth the price they are choosing to stay, meaning if somewhere is expensive, with bad review and is quite far away from their tourist destinations they will not want to reserve the room. A room with good reviews however but, high in price might seem adequate, but what happens when they check attributes like location, or see that there isn't enough space for people traveling or space for essential things the customer is carrying.

**Preparing your dataset for analysis (data cleansing, choice of features) etc. (5)**

After deciding which attribute will be used to predict price, the attributes being used must be cleaned so that the null objects discovered are changed to something that can be worked with. Also, because the data contains objects inside numerical columns (as found before from data visualization), a fair amount of cleaning must be executed. Changing null objects to averages, and objects to nulls which are then changed to said averages, fills the data with numerical values. Averages are used because we do not want to affect the values in the rest of the column, due to un accurate results being presented, averages will not increase or decrease the average of the column as a whole. This method has been done on only the columns still present in the data set after the others were dropped due to no relevance for the intended outcome.

After these methods the code still recognizing columns with objects in them as 'object' columns, in order to secure that the rest of the assignment goes well, these have to be changed to a numerical label so that there are no disruptions in the coming methods that will be carried out, as said before, issues have to be avoided for accuracy and reliability. It is important to check that this has been successful, so there is no room for failure.

In order to assure these attributes chosen will be successful in predicting the price of a room, we must establish what kind of relationship they have with the target price. Correlations are a reliable method, using scatter plots to show relationships and how variables interact and bond with each other. Doing so, a clear indication of whether the right attributes have been selected will present itself, if not the accuracy will not be successful, and the assignment will not produce good results for work put in.

The first step is the matrix, which results high light the relationships between variables. With price having a very successful relationship with itself, a maximum fitness value for the correlations is equal to 1. The results show that there are 4 main attributes that are closely related to price; Beds, accommodates, bedrooms & bathrooms. The reaming attributes are close, with a few acceptations. Because these attributes all consist of numerical value, they might appear to have a bad relationship or not as effective as the before mentioned, they still provide a better result for accuracy, the more used the better, and the results will be proven to be officiant as more data has been used to predict.

Using pandas' package for Python, a visualization of these correlations is in front of us. The data in numerical script helps understand which attributes are important, but by being studied visually as well, it helps dig deep into the surrounding factors playing a part in the prediction and helping to execute the objective well.

Following the results of the correlations, more in-depth studying has taken place to see the relationships under a microscope, to see the bond between price and each individual attribute shown. Represented visually since the count value has already been shown the scatter plots show with detail the relationship between each attribute and the intended target. It is important to research into this well, and as much as possible for the results to be positive.

Given that beds, bathrooms, accommodates and bedrooms were top of the list regarding relationship, the results show that even though the counts for the remaining attributes weren't as successful, the scatter plots show detail we haven't yet noticed, and could prove to be usual, despite the low value for relationship.

**Review cleanliness** shows a that most prices that are lower have a high cleanliness value

**Review scores for checking in**, are mostly in the range of 9-10, showing really good review scores with a lot of prices.

**Review scores rating** highlights a large amount of data from price is extremely high

**Latitude** highlights most of the data collected is between 51.48 and 51.44

The in-depth look at the data has provided a better valuation of these remaining attributes, even though their relationship isn't close, the information's collected has helped with the whereabout or ratings of the rooms to stay in the Airbnb data frame.

**Splitting the Data**

Using the stratified shuffle, which creates a training data set worth 80% and a testing data set from the whole of the column for price, a new column is created, and the data is hence forth split. The reason for doing this is so the methods tested on the training data, can then be tested on the remaining 20% of test data. This is for future data that will be given, to show the success of the methods used against unseen data.

Stratified shuffle, ensures that both parts of the split may differ in size, but their calculations are the same regarding means, mediums and other important results. This means both training and testing sets are equal, and the bigger portion has more to work with since the use of this code in the future is less important than it is now. The results show a successful split in regards to ratio.

It is important now to reflect on the methods we have used, since they played an important part in visualizing the data at the start, they are now equally important as the data will be studied again since it has been split. Errors cannot be present, effecting the results. Histograms are used again as they prove to be accurate when trying to understand the data being used. The results show similar patterns to the chosen attribute histograms as done before, evidently on a smaller scale since the data has been split. The data is also removed from the new column created 'price_cat' as it was used just for splitting, that empty column will also have to dropped.

**Choice of models, application, evaluation and validation (8)**

**Models for application, evaluation and validation**

Firstly, before selecting models, the 'no room for errors' trend is again brought into action even this late on in the process. The variable for the training and test data must be changed to avoid confusion, and the labels for the actual values we are predicting are called and a variable name change has taken place.

Linear Regression

Linear regression is a common method to predict a certain size or value of a new piece of data based on where the regression line is, if a new variable was added, and we know the x value, the regression line based on the rest of the variables calculations is used to find out what the value of y is using the data found before. Linear Regression in statistics is a linear approach to modelling the relationship

between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of explanatory variable is called simple linear regression.

Linear regression tells us that the linear line has an equation of the from Y = a + bX, where C is the explanatory variable and Y is the dependent variables. The slope of the line is B, and A is the intercept (the value of y when x = 0). Statistical researchers often use a linear relationship to predict the (average) numerical value of Y for a given value of X using a straight line (called the regression line) … before moving forward to find he question for your regression line, you must identify which of your two variables is X and which is Y. In regression if both variables increase the same time the relationship is known as a positive relationship and if one increases but the other decreases or doesn't move then it Is known as negative relationship. The variable distance from the line is known as errors, it basically says that there are said amount of errors between the estimated value (the line) and the location of where the variable is in relation to where the line is. A common and perfect method to start with.

Support vector

Support vector machines help decided what the difference between two variables is, to uncover what said variable it is. It segregates two classes to define what it is. The algorithm uses an unoptimized decision boundary to work out the attributes that make something a class, positioning it is fine, but you need to consider the mistake of classifying the wrong variables as something else. The support vectors take the nearest variables from each side of the chart closet to the line and draw the support vector line at either side to mark the radio of how far away each line goes. This is free space and is known as the margin. If a line is not possible to make, then you make the 2d graph 3d and it becomes easier to separate the variables.

**Random Forests**

Random forests are when a decision tree has predicted something, but because of the size of the data, the decision tree may only predict a small part. So, they are brought together into a random forest with loads of predictions for certain aspects, some are right, and some are wrong, some can be close to either. For instance if we were looking at an image of a hand, and trying to predict or see that an image has in fact a hand in it, in computer vision, then the decision tree has too much data to run on, so we use loads of trees that form the forest where they are all looking for multiple things such as fingers, thumbs, palms, wrists, people and nails etc, this means the image is being monitored and searched from all angles, and will result in a more accurate prediction.

The technique uses boosting and bagging for separating things that are stronger than the rest, so if you had 1000 decision trees in your forest, that is a lot of trees, so boosting takes the tress that are the strongest predictions and bags them so they are used more when prediction for the result, instead of waiting for 1000 trees and some of them will predict nothing so they are used over and over again but have the same results, a waste of time.

Bagging

Bootstap aggregating or commonly known as 'bagging' is an ensemble meta-algorithm. Bagging is designed to improve and give higher results for accuracy or stability in machine learning. Algorithms in machine learning are improved using this method especially in statistical classification and regression. Using the regression method, bagging reduces variance and helps when trying to avoid overfitting when working with data.

Taking the best predictions and then move them forward and take bad predictions and make them better with accuracy. It is used for the outcomes and data made from all algorithms such as regression or decision tress after forests, it is basically the last step in making sure accurate data is much more accurate.

Decision trees

Decision trees are used to predict an outcome based on given data. A tree will predict using data if something is related or not. An attribute is split up and the close relationships are chosen and then compared to others once they have reached the next branch or second stage. Using attributes to measure how far the relationship goes, the decision tree does as many as possible, and predicts using ratios which are basically yes or no based on the next attribute being assessed has a close relationship with the first, and so on and so forth until a reliable ratio is found, and a prediction presents itself.

## Interpretation and explanation of the results of your models and implications of these for your initial problem (4)

### Results

The results from the chosen methods highlight that Support vector and Random Forrest have the best accuracy based on the calculation of Root Means Square Error (RMSE). RMSE is used because it is the appropriate way of finding errors.

Decision tree has a RMSE of 0.0 which cannot be accurate since it is much better than other methods. Decision tree will be tested on the test data to see if the results are accurate or not, along with bagging since it is a less commonly used method and has an average RMSE as the others.

The results show that both methods have a very high RMSE when the testing data is used. This is not accurate, and these methods have proved to be useful with the training data, but for the future and any unseen data, they are not reliable. To further evaluate predictions were based on the data before the split, and the RMSE was performed on it, which have proved to be lower the test data, but still high in comparison to the training.

**Conclusion**

To conclude, the methods were accurate for data that has been presented but doesn't work and is inaccurate for future data which or other data collected from other cities. To view the predictions for the training data, and the data a whole, two visuals are present to highlight how close they were but the larger data (adding on the 20%) has caused more errors.