Title

Submitted by

13020363 Yudi Chen

12978452 Kai Chen

Table of Contents

[1. Introduction 3](#_Toc20390007)

[Business understanding 3](#_Toc20390008)

[2. Identify Challenges 4](#_Toc20390009)

[3. Project Plan 4](#_Toc20390010)

[4. Data Exploration 5](#_Toc20390011)

[Data understanding 5](#_Toc20390012)

[Pie chart 6](#_Toc20390013)

[Box plot 7](#_Toc20390014)

[Correlation matrix 9](#_Toc20390015)

[Preprocessing 11](#_Toc20390016)

[Convert categorical variable to numeric 11](#_Toc20390017)

[Bin price 12](#_Toc20390018)

[Features selection 12](#_Toc20390019)

[5. Methodology 14](#_Toc20390020)

[5.1 Decision Tree Classifier 14](#_Toc20390021)

[Parameter Optimization for DT 14](#_Toc20390022)

[Visualization of Decision Tree 17](#_Toc20390023)

[5.2 Support Vector Machine 18](#_Toc20390024)

[Parameter Optimization for SVM 19](#_Toc20390025)

[5.3 Nearest Neighbors Classifier 20](#_Toc20390026)

[Parameter Optimization for NN 21](#_Toc20390027)

[5.4 Random Forest Classifier 21](#_Toc20390028)

[Parameter Optimization for RF 22](#_Toc20390029)

[6. Evaluation 23](#_Toc20390030)

[Ethical issues 24](#_Toc20390031)

[Conclusion 24](#_Toc20390032)

# Introduction

With the development of the Internet and transportation, people are more likely to find and book accommodation through the Internet when they are traveling. Airbnb, as a platform for online arrangements or accommodation, is one of the representatives. This report uses the decision tree model to analyze and predict the price of home based on Airbnb data in Barcelona, ​​Spain. First, the data in each column of the dataset needs to be preprocessed and explored, including defining, classifying and analyzing the type of data through tools, such as Pie chart, Boxplot and Correlation matrix, and removing irrelevant data. Then, building a decision tree, nearest neighbors, random forest and SVM model by Python to analyze and predict the house price according to the other data information. In addition, improve the accuracy by adjusting the parameters of different models. Finally, to get the accuracy rate results and come up with conclusions.

The purpose of researching and predicting home price accuracy is to help homebuyers to give a more reasonable price based on their own homes condition. In addition, Airbnb can help Airbnb regulate the housing market so that home prices are within reasonable price ranges. However, in the process of analyzing data, there are also ethical issues, including the leakage of user personal information data, the discrimination of algorithmic systems and the responsibility and security issues as well as the analysis of these problems through utilitarianism and conclusion.

## Business understanding

With the development of society and the improvement of people's living standards, people's demand for housing is growing, and house prices have become the focus of attention. Since house prices are affected by many factors, such as transportation, geographical location, living environment, etc., people are eager to know the reasonable price of the house. The accuracy of the housing price predication can not only help the customer make the right choice, but also help the seller to make a reasonable price. In addition, it helps the platform standardize housing prices and prevent price irrationality.

# Identify Challenges

Different problems and challenges may be encountered in the process of data exploration and mining.

Firstly, a large amount of data and complex data types are among the top challenges. Since researching the price of the house, there are many different types of attributes involved, including digital information about customer satisfaction with the home and the number of bedrooms as well as the character information for the type of housing.

Then, choosing the model is one challenge should be solved. Each model has its own advantages and disadvantages. We need to select the model according to the type of the data set, the data information that needs to be predicted, and the reasonable selection of the model, which is conducive to improving the accuracy and efficiency of the prediction.

The last one is the adjustment and modification of the model parameters. For an example, in the decision tree model, we need to adjust some default parameters in the process to control the complexity and length of the tree, aiming to reduce the memory loss and get higher accuracy.

# Project Plan

The planned steps and processes of this project refer to the CRISP-DM model. In order to achieve business needs and goals, we need a complete project plan. The data used for research will be the training model for different models. Of course, we need to build these models in Python. In the process of training data, the highest value of the predicted accuracy is achieved by adjusting the parameters. Finally, the accuracy rate and related information are analyzed to select the optimal model.

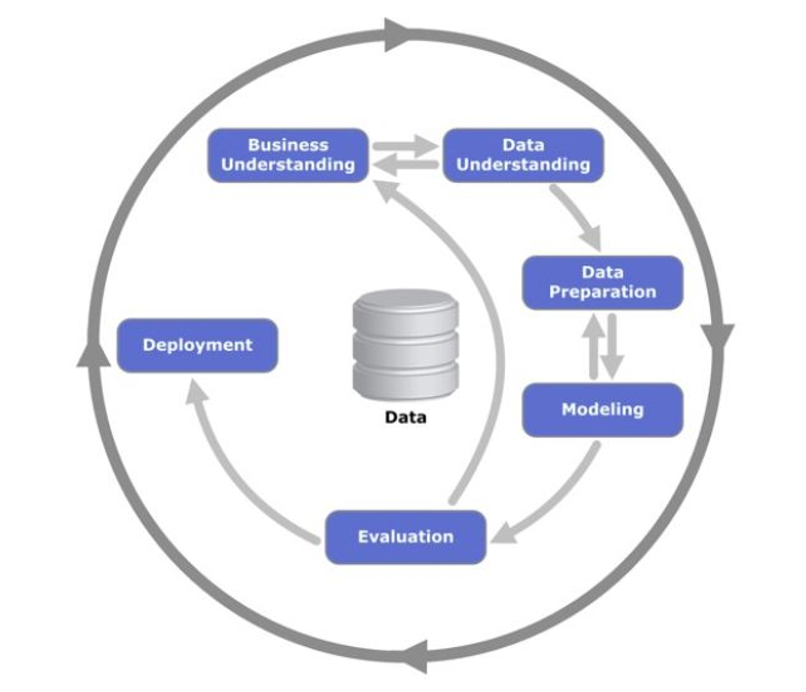


Figure 3-1

# Data Exploration

## Data understanding

Our dataset is sourced from the Kaggle website which is a platform for providing dataset resources and learning machine language. This data set consists of 9 variables and 12633 rows of data that describe housing price data on Airbnb in the urban area of ​​Barcelona, ​​Spain. These data types include the type of room, neighborhood area, reviews, customers’ satisfaction, accommodates, the number of bedrooms, price, latitude and longitude. Next we will explore and analyze these data types.

There are 9 attributes in testing dataset, the description of these attribute are shown as the table below：

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Type** | **Comment** | **Range** |
| room\_type | Integer | Types of homes | “private room” or “shared room” or “whole room” |
| neighborhood | String | Names of different urban areas | “Eixample” and other 10 regions |
| reviews | Integer | Number of visits to housing information | The number must be positive |
| overall\_satisfaction | Floating point | Customer satisfaction with the houses | From “1” to “5” |
| accommodates | Integer | The number of people a house can hold | The number must be positive |
| bedrooms | Integer | The number of bedrooms in this house | The number must be positive |
| price | Floating point | The price of the houses | The number must be positive |
| latitude | Floating point | The latitude of the houses | According to the location of the house shown by the tool |
| longitude | Floating point | The longitude of the houses | According to the location of the house shown by the tool |

Table 4-1

### Pie chart

The pie chart is a tool to divide the data into the number ratios of different blocks.

The pie charts in this section are used to show the distributions of the categorial features: “room\_type” and neighborhood. Here is the code that implements it.

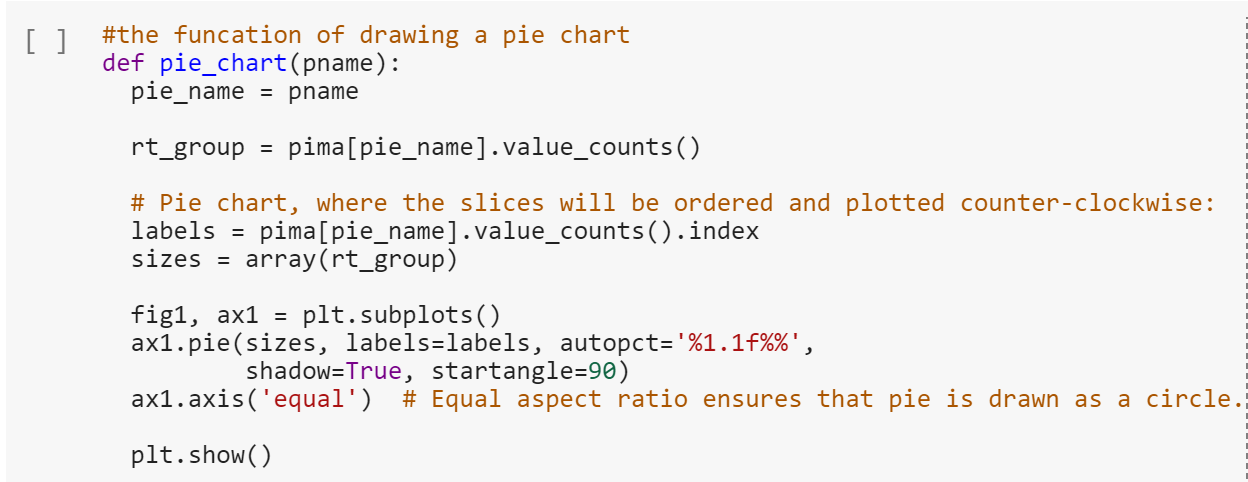


Figure 4-1

Figure 4-2 shows that “room\_type” has 3 values: entire home/apt, private room and shared room, and private room occupies more than half (51.7%), followed by entire home/ apt(47.7%). The last part is shared room taking up 0.7%.

The figure 4-3 indicates the distribution of neighborhood. The “Eixample” is the most name taking up 33.9%. The value of the second name Ciutat Vella is 12.3% less than the first one. The following names, “Sants-Montjuïc”, “Sant Marti” and “Les Gràcia”, have the similar percentages around 10%. The left space is separated by “Sarrià-Sant Gervasi”, “Horta-Guinardó”, “Les Corts”, “Sant Andreu” and “Nou Barris”.

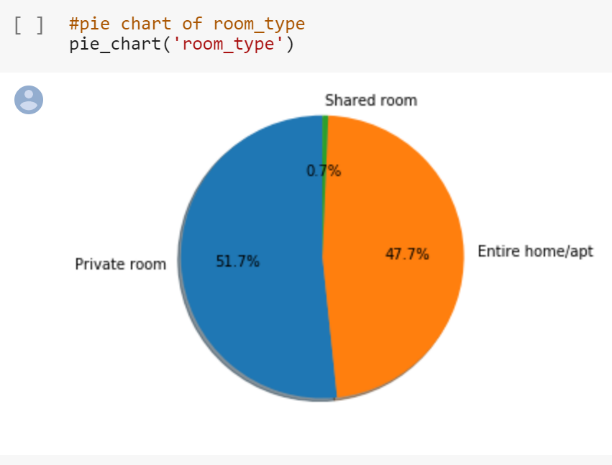
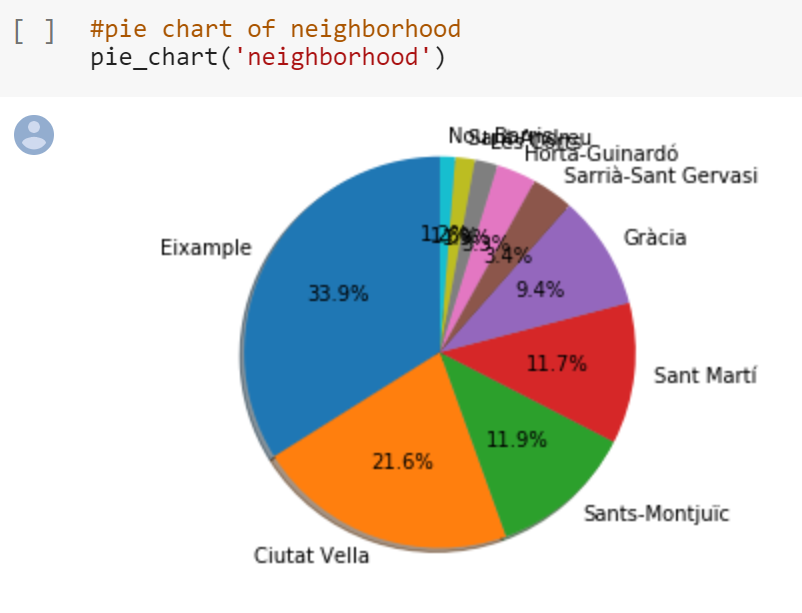
 

Figure 4-3

Figure 4-2

### Box plot

Box plot is an important tool to analyze the distribution of digital information. In this part, we use the box plot to show display and analysis the distribution of the “reviews”, “overall\_satisfaction”, “accommodates”, “bedrooms”, “price”, “latitude” and “longitude”. The following is the code that implements this.



The graph could show the maximum, minimum and medium of the attributes, it is helpful for us to analysis the distribution of the attributes. Figure 4-3 to figure 4-9 show the distribution of every attributions in detail.

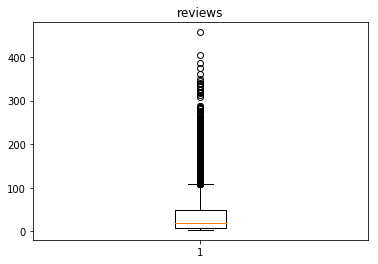


Figure 4-3

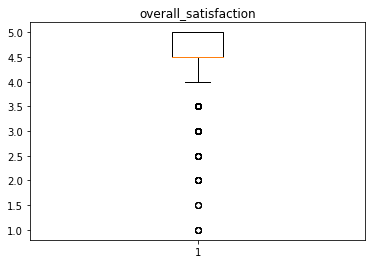


Figure 4-4

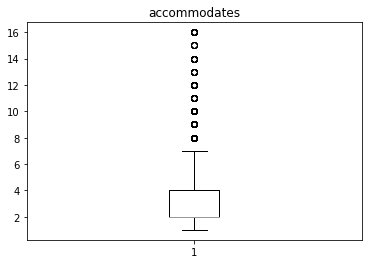


Figure 4-5

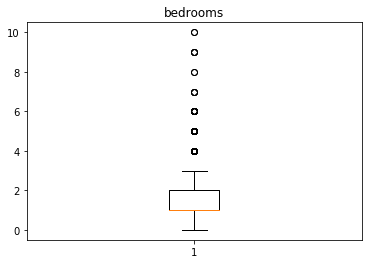


Figure 4-6

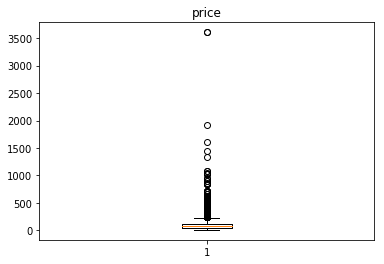


Figure 4-7

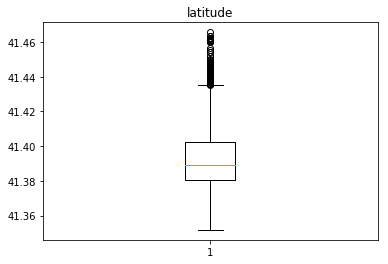


Figure 4-8

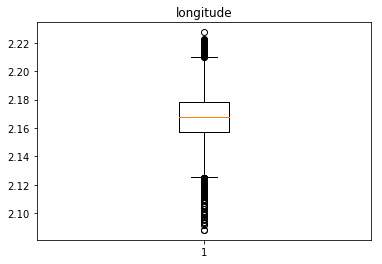


Figure 4-9

### Correlation matrix

Correlation matrix is a table representing the correlation of variables and an important tool to study the correlation between variables. The following is a correlation matrix between the seven variables.



Figure 4-10

Here is a more concise expression of correlation matrix and the associated code. See figure 4-11 and 4-12.



Figure 4-11

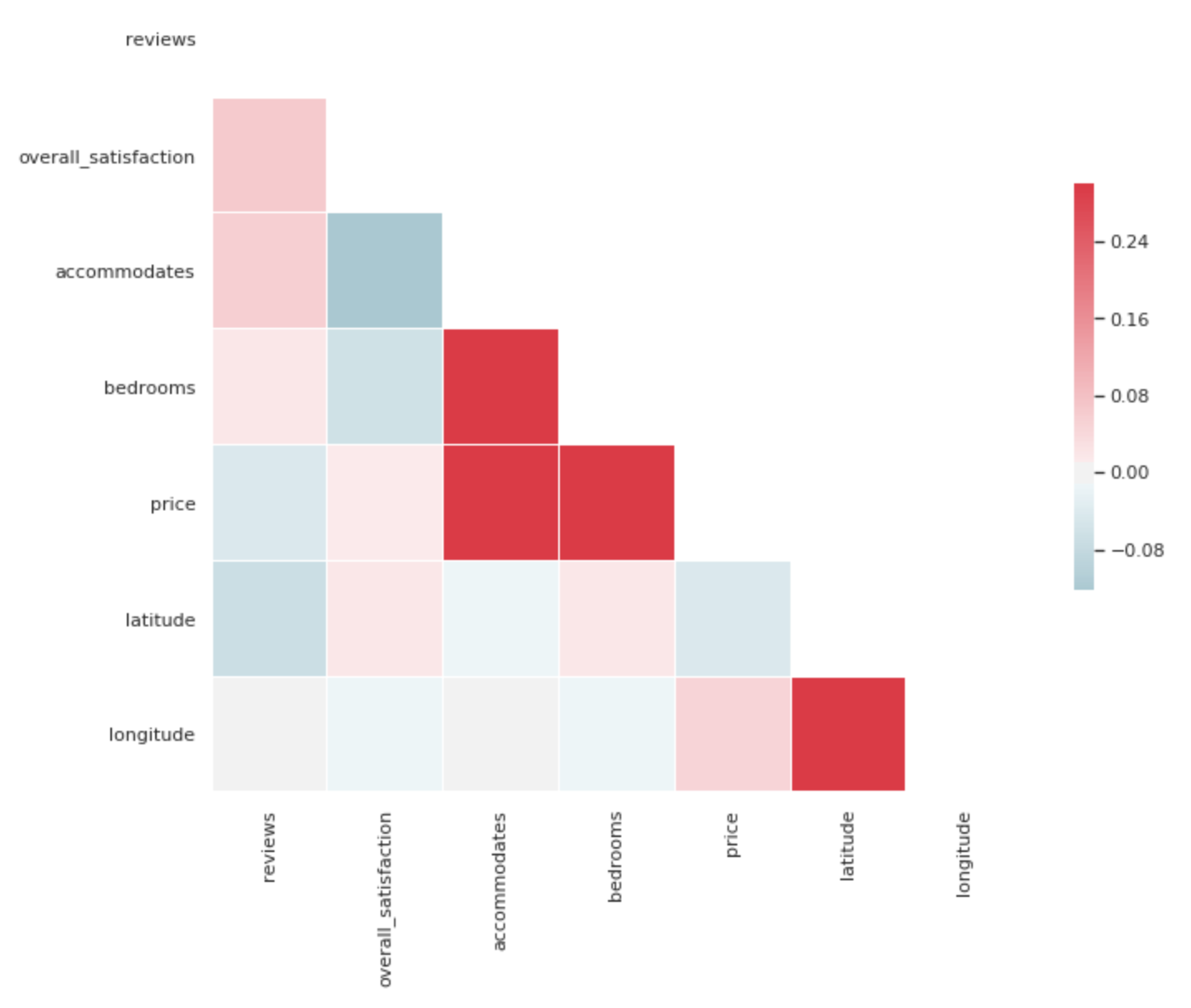


Figure 4-12

From the above colored correlation matrix chart, we can clearly see that the house price has the greatest correlation with the number of people available and the number of bedrooms.

## Preprocessing

Through the previous steps, something interesting are found and the methods of preprocessing can be considered. In this section, three methods are used to preprocess the dataset, so that the classifiers can predict the result more accurately. The first one is changing the type of categorical variables to numeric, because the classifiers used in the study only accept numeric features. The next one divides the data points of price into 3 categories( 0, 1, 2) so it can increase the speed by reducing the number of tree levels. The last one decides the features which will be used in classifiers.

### Convert categorical variable to numeric

There totally are two categorical features, room\_type and neighborhood respectively, needing to convert. The way of converting is replacing the distinct values of each feature by integers which start from 1. As is shown in Figure n+2, room\_typecategory and neighborhoodcategory replace the old relevant features.

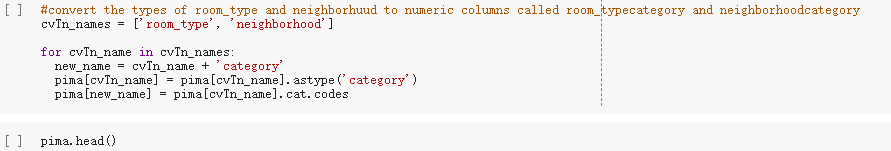


Figure 4-13



Figure 4-14

### Bin price

The values of price are divided into 3 bins and the bins are called 0, 1 and 2 respectively, based on the same width of each feature.

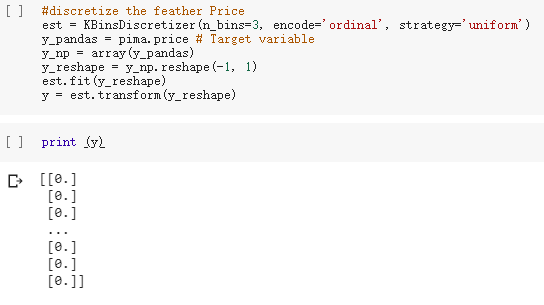


Figure 4-15

### Features selection

All the features can be used after reprocessing; thus, no features are filtered. Finally, split dataset into training set and test set as X\_train, X\_test, y\_train and y\_test (Figure .3-2). Due to the series of preprocessing, the accuracy increases to 0.9998021369212505 from 0.84.

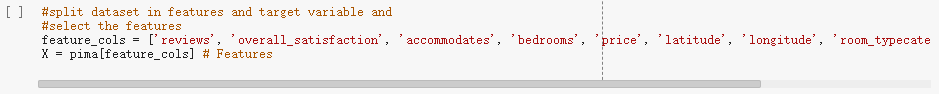


Figure 4-16

Split dataset:



Figure 4-17

# Methodology

In this section, 4 classifiers, Decision Tree, Support Vector Machine, Nearest neighborhood and Random Forest, will be compared and optimized to find one which has the best performance. The method has two parts. Firstly, using the default parameters of each classifier first then tuning its performance by adjusting the values of some parameters, at last, scoring its performance. Next, evaluating the performance of each classifier based on the score and select the best one as the final proposal.

## Decision Tree Classifier

Build Decision Tree model and predict using default parameters as a reference standard. The score is 0.9998416970080735.

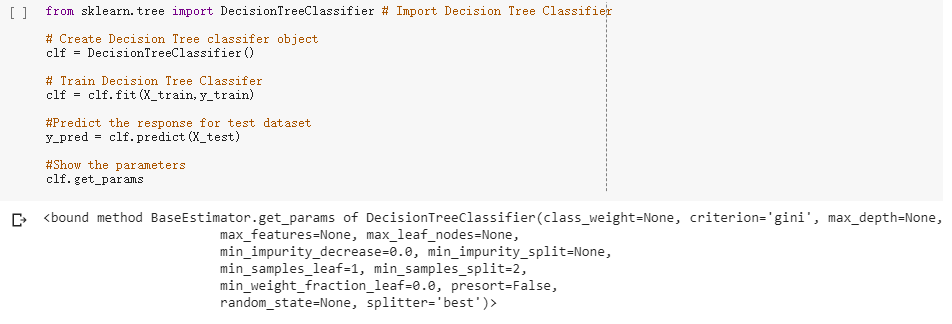


Figure 5.1-1

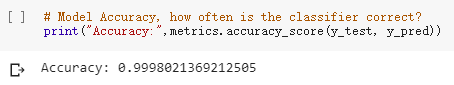


Figure 5.1-2

### Parameter Optimization for DT

This section mainly tests two main parameters: min\_samples\_split and max\_depth, because the key point of optimizing Decision Tree is controlling the size of trees. To reduce the size of trees, both parameters need optimization. The first part is tuning the min\_spales\_split and its range is between 2 to 50. The second part is tuning the max\_depth and its range is between 2 to 50. From the accuracies of both, the performance increases although the original accuracy was very high. The best value of accuracy is 0.9998680564718301.

The value of min\_samples\_split

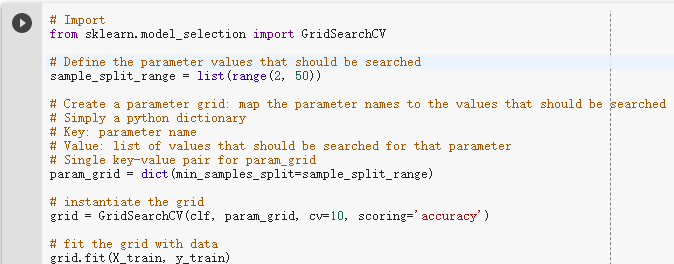


Figure 5.1-3

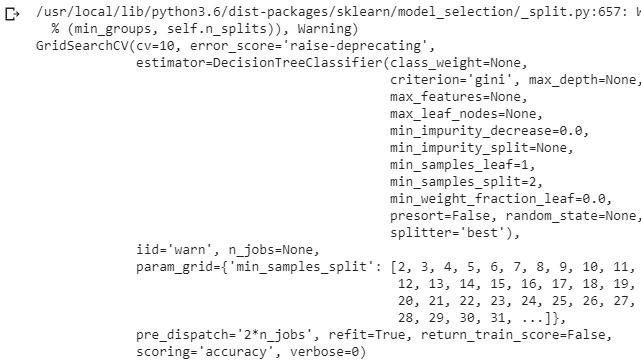


Figure 5.1-4

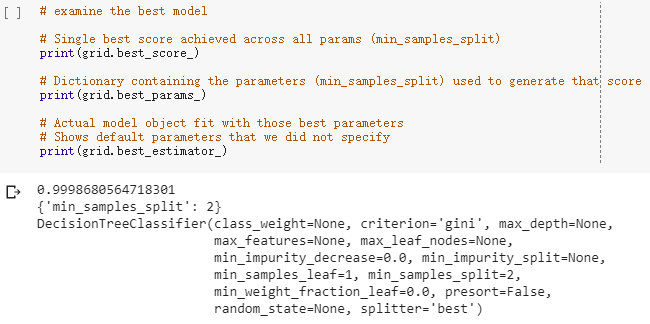


Figure 5.1-5

The value of max\_depth

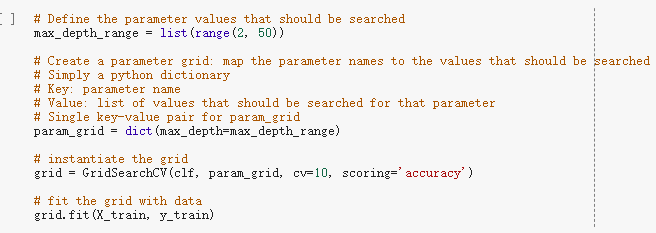


Figure 5.1-6

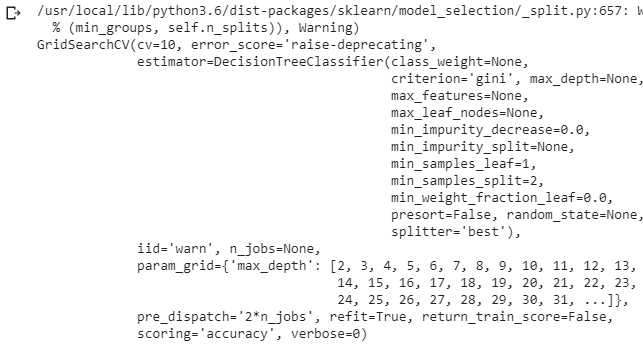


Figure 5.1-7

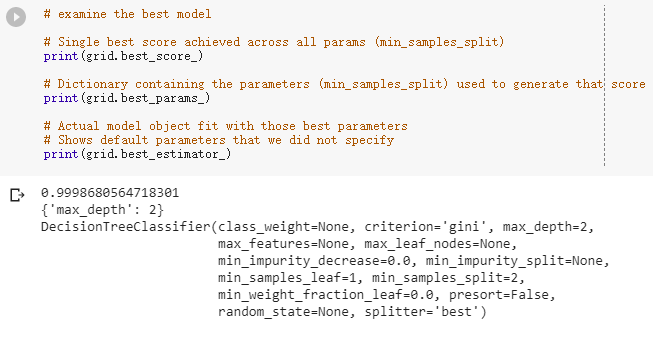


Figure 5.1-8

### Visualization of Decision Tree

Through the visualization of Decision Tree, the process of decision is very clear.

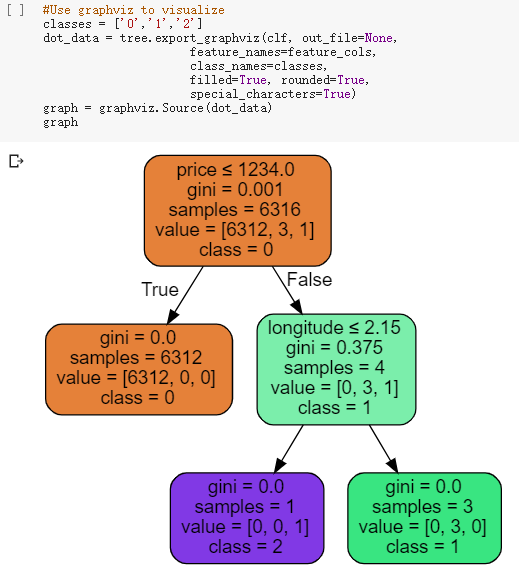


Figure 5.1-9

## Support Vector Machine

The accuracy of Support Vector Machine is 0.999604273842501.

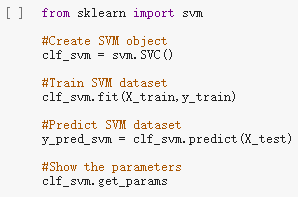


Figure 5.2-1

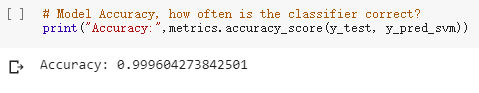


Figure 5.2-2

### Parameter Optimization for SVM

The key point is controlling the C and kernel, because both parameters can affect the performance directly. The process of optimization is calculating the highest accuracy in the range of C ( 0.001, 0.01, 0.1, 1, 10.). The best value of C is 0.001 and the related accuracy is 0.9994722258873202, while the best kernel is linear and the accuracy is 0.9998680564718301. The best accuracy is 9998680564718301.

The value of C

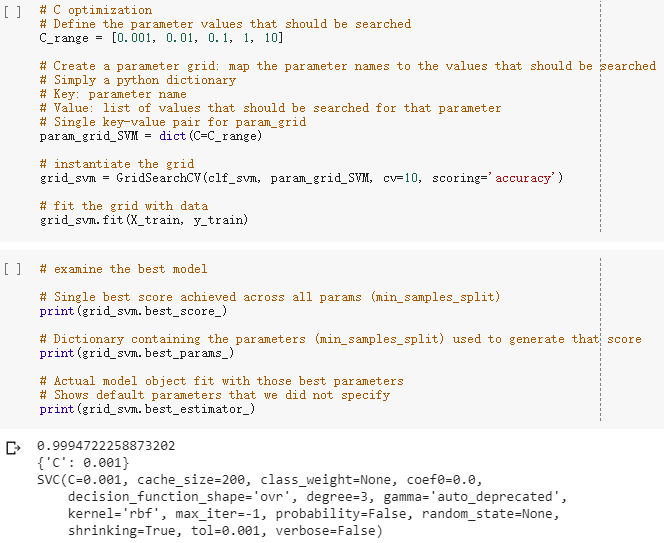


Figure 5.2-3

The value of kernel

The method is similar with the optimization of C, but the range of kernel is in linear, rbf and poly. The best value of kernel is linear and the accuracy is 0.9998680564718301.

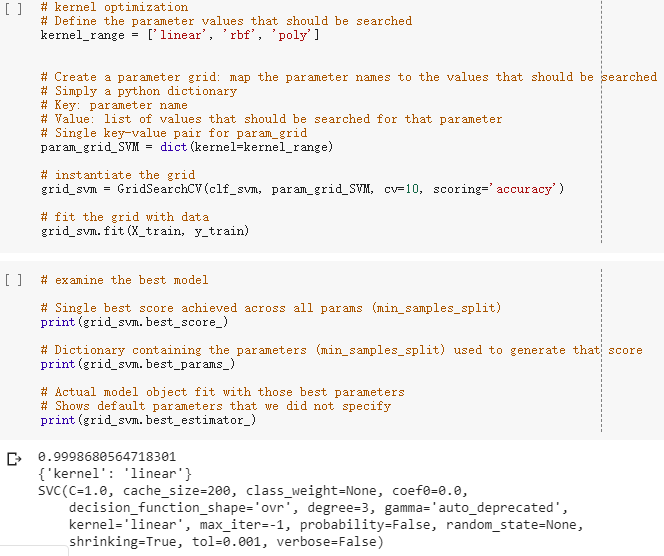


Figure 5.2-4

## Nearest Neighbors Classifier

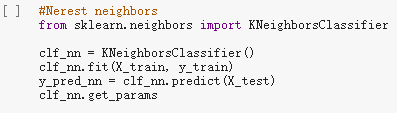


Figure 5.3-1

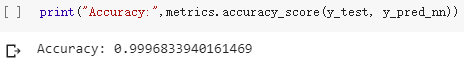


Figure 5.3-2

### Parameter Optimization for NN

The value of n\_neighbor

The key parameter which affects the performance of Nearest Neighbors Classifier is n\_neighbors. The range is defined in 1, 2, 3, 4 and 5. After the optimization, the best value is 1 and the accuracy is 0.9998416719442685.

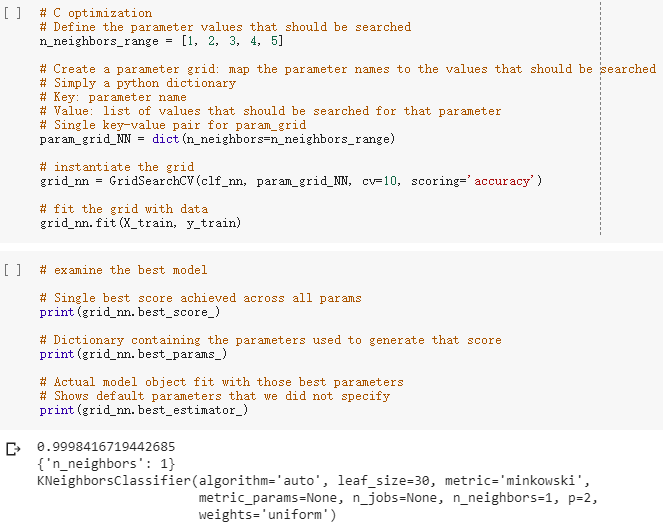


Figure 5.3-3

## Random Forest Classifier

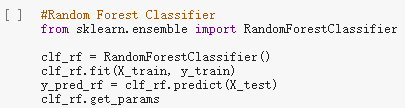


Figure 5.4-1

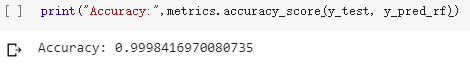


Figure 5.4-2

### Parameter Optimization for RF

The value of n\_estimators

According to the websit Scikit Learn, the main parameters are n\_estimaters and max\_features. The parameter n\_estimaters represents the number of trees in the forest. The higher value of this parameter takes more samples, but it will spend more time.

Another one max\_features means the size of the subset of features. One should be careful when tuning this parameter. Because the higher value can increase the bias, while the lower value can decline the variance. Thus, it recommands that the empirical default value is None for regression or sqtr for classification, meaning considering all the features instead of random subsets. Except for the two parameters, one parameter to consider is max\_depth. It is usually None along with min\_samples\_split=2. After the optimization, the best value of n\_estimaters is 40 and the best accuracy is 0.9996833438885371.

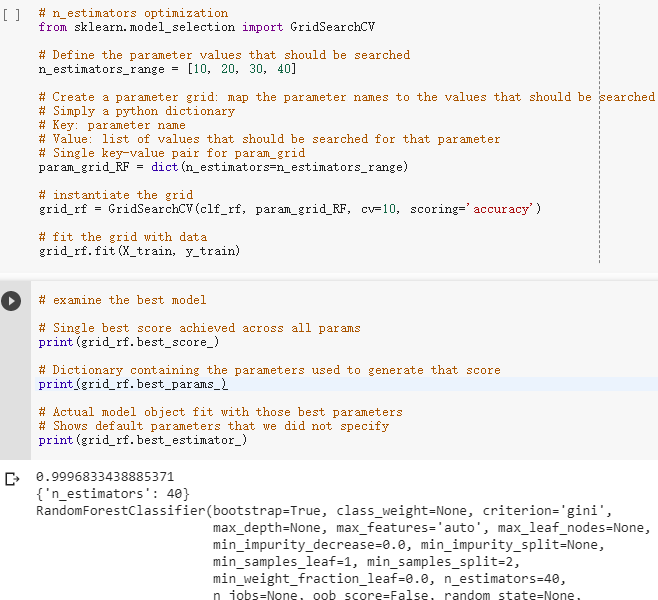


Figure 5.4-3

# Evaluation

At last, the accuracies of the classifiers are listed in Table n. In the table, the highest value is 0.9998680564718301 resulting from Decision Tree and Support Vector Machine. The accuracies of Nearest Neighborhood and Random Forest are 0.9998416719442685 and 0.9996833438885371 respectively, ranking the second and the third. The simplist way is comparing the values of the accuraies of each classifier, thus, the better proposals are Decision Tree and Support Vector Machine. Compared with Decision Tree, Support Vector Machine is not good at processing a huge amount of dataset, so the best classifier may be Decision Tree. Except for the accuracies, however, the characteristic of each classifier should be considered more, because the charactors can be suitable to different datasets. Next part analyses the characteristics of the classifiers. The process of Decision Tree is based on a lot of decisions, yes or no, and then it goes one direction after answer the question and forms a branch. This way is the closest to human's way of thingking. Support Vecter Machine maps the data in the lower dimensional space to the higher dimensional space and separate them by using the hyperplane. The method of Nearest Neighbors is that a point selects some points nearest to it and confirms its value based on the value which most points are. Simply,Random Forest is a way consisting of many Decision Trees and result the final model based on the Decision Trees. On one hand, in contrast, the ways of both Desicion Tree and Support Vector Machine are reasonable and accurate, while Nearest Neighbors may be affected by probability. On the other hand, however, the Decision Tree and Support Vector Machine consume more system resources or spend more time. In summary, for this dataset, the best way is Decision Tree or Support Vector Machine.

# Ethical issues

However, the analysis of customer data will also have ethical issues and social impact. Next, we will analyze this issue from both utilitarian and Kantian perspectives.

From the perspective of utilitarianism, the analysis of data is in line with the needs of the society. Because the housing market and prices are not as good as the predicted values ​​without reference, there will be market disorder, which will affect the macroeconomics of the market. Although obtaining user privacy information violates the privacy of customers, analyzing data information can predict market trends and have a positive effect on social development.

From the perspective of Kantism, it is wrong to obtain customer information and carry out research with commercial value, because Kantism is the origin of the right and wrong of things, in other words, it means using good duty to establish its own principle foundation. Therefore, stealing customer information is a violation of the privacy of the users of the website, and it is a behavior that violates social ethics[1].

# Conclusion

This article is a forecast model of the housing market price based on the CRISP-DM model. Its purpose is to help buyers to give a more reasonable price based on their housing situation. In addition, Airbnb can help Airbnb regulate the housing market and keep prices at reasonable price ranges. From the perspective of utilitarianism, the analysis of data is in line with the needs of the society. But from the perspective of Kantianism, stealing customer information is a violation of the privacy of the users of the website, and it is a behavior that violates social ethics.

In addition, our prediction model has some areas for improvement. From a technical point of view, in the design process of the model, due to time and technical reasons, we only use a part of the parameters to adjust. In fact, we can use more parameters to adjust, which can reduce system loss and improve work efficiency, thereby improving the accuracy of the prediction. From a non-technical point of view, because the size of the dataset we selected is not enough, some variables are not classified, at the same time, the size of the data will also affect the accuracy of the prediction.

# References

[1] Stanford Encyclopedia of Philosophy 2004, *Kant’s Moral Philosophy*, viewed 20 September 2019, <https://plato.stanford.edu/entries/kant-moral/>