Tampere University of Applied Sciences

Applying Statistical Learning Methods to Predicting Real Estate Value

Son Nguyen

BACHELOR'S THESIS
November 2019

Degree Programme in International Business

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree Programme in International Business

Son Nguyen
Applying Statistical Learning Methods to Predicting Real Estate Value.

Bachelor's thesis 45 pages, appendices 0 pages
November 2019

This thesis was conducted as a data mining project that was inspired by the recent advances in the fields of statistics and computer science as well as their applications in the business field. The main objective of the thesis was to predict future house price using a publicly available data set containing observed information about real estate value in Sindian District, Taiwan. First, the data was examined for useful information by computing basic descriptive statistics as well as plotting graphs for visualization of the distribution and relationships of the variables. Subsequently, the data was divided into a training set and a test set, then linear regression and random forests models were built and tested. These models used statistics to identify the pattern as well as the relationships between the predictors and the response in the training data which would then be used to predict future values of the response on the basis of the predictors. The linear model selection was done by the best subset method while the random forests models were compared using test MSE and the model with the lowest test MSE was chosen. The results showed that random forests models had significantly lower MSE and thus proved to be more suitable for the predicting purpose.

Key words: statistical learning, data mining, real estate, business analytics

Table of Contents

GLOSSARY

MSE	Mean Squared Error
RSS	Residual Sum of Squares
TSS	Total Sum of Squares

1 INTRODUCTION

Data mining refers to applying statistical and machine learning methods to extract meaningful information from data repositories. Thanks to the development of state-of-the-art technologies, there is a massive amount of data generated which opens the doors for data mining techniques to bring considerable value. (SIGKDD 2018.) According to Columbus (2017), the amount of companies using big data analytics rose sharply from 17% in 2015 to 53% in 2017 (Columbus 2017). Indeed, data mining as well as statistical learning methods are being applied in a variety of fields such as medical, social sciences or business (Mangasarian, Street & Wolberg 1995, 570-577; Buza 2014; Metzger, Leitner, Ivanovic, Schmieders, Franklin, Carro, Dustdar & Pohl 2015, 276-290). Inspired by the value that data mining and statistical learning could bring, this thesis was conducted as a data mining project with the main objective of using statistical learning methods to predict future house price in Sindian District, Taiwan and a publicly available data set from Yeh & Hsu (2018) containing recorded data about real estate valuation in Sindian District was used for this purpose (Yeh & Hsu 2018, 260-271). The statistical learning methods would identify the patterns and the relationship between the variables in the data set, use this information to build models that could predict the house price based on the other variables. The information of the house used for leaning and predicting consisted of its transaction date, age, distance from the nearest metro, the number of convenience stores nearby and geographic coordinates. After the statistical models that can learn by identifying the relationships between the house price and the aforementioned data were built, they would take new data of the latter as inputs and come up with predictions for the former as output.

The outline of this thesis is structured as follows. Section 2 contains the theoretical framework needed for this project which mainly involves the theory behind the statistical learning methods. Section 3 describes specifically the methodology of this thesis. This includes data gathering, exploring, preprocessing and model building. Additionally, the methods as well as criteria for evaluating and choosing models are also presented. Section 4 is mainly used to describe the

data exploration process since this is a crucial step in a common data mining project. Section 5 describes the results of applying the methods in section 3 to the data. The last section is used to analyze and discuss the results from the previous section, suggest potential application and further development and lastly make final conclusions for the thesis.

2 THEORETICAL FRAMEWORK

2.1 Statistical learning

2.1.1 Definition

Generally, statistical learning means a set of approaches used for estimating some assumed relationship between variables of a data set. Normally, there would be some input variables $X_{1,}X_{2,}...,X_p$ and a corresponding output variable Y. There are several names for these variables. The input variables can also be called predictors, independent variables, features or just variables while the output variable can be called response or dependent variable. The relationship between Y and $X=(X_{1,}X_{2,}...,X_p)$ can be generalized as:

$$Y=f(X)+\epsilon.$$

In this formula, f is a fixed but unknown function of X while ϵ is a random error term having mean zero and independent of X. Statistical learning methods aim at giving a good estimation of f. (James et al. 2013, 15-17.)

2.1.2 Importance of estimating f

The estimate of f is important because of two main reasons, one of which is when X or a set of input variables are available but Y or output variable is unknown, a sufficiently accurate estimate of f allows prediction of Y using the formula:

$$\hat{Y}=\hat{f}(X)$$

with $\hat{f}$ is the estimate of f and $\hat{Y}$ is the prediction of Y. For this objective, f can be treated as a black box in the sense that predicting Y accurately is more important than knowing its true form. The accuracy of the prediction depends on two quantities known as reducible error and irreducible error. If $\hat{f}$ is not an exact estimate of f, which happens most of the time, then the error caused by this is called reducible error since it is possible to reduce this quantity by choosing a more suitable statistical method. On the other hand, even if $\hat{f}$ is exactly equal to f, the prediction would still have some error due to the fact that Y is also a function of ϵ which could not be predicted using X. In other words, regardless of the accuracy of $\hat{f}$, the error ϵ is irreducible. The other reason for estimating f is to properly have an insight on the dynamics of the variables, how they affect each other. f is now used to understand how Y changes as a function of X. Therefore, contrary to prediction, the true form of f must be known. (James et al. 2013, 17-19.)

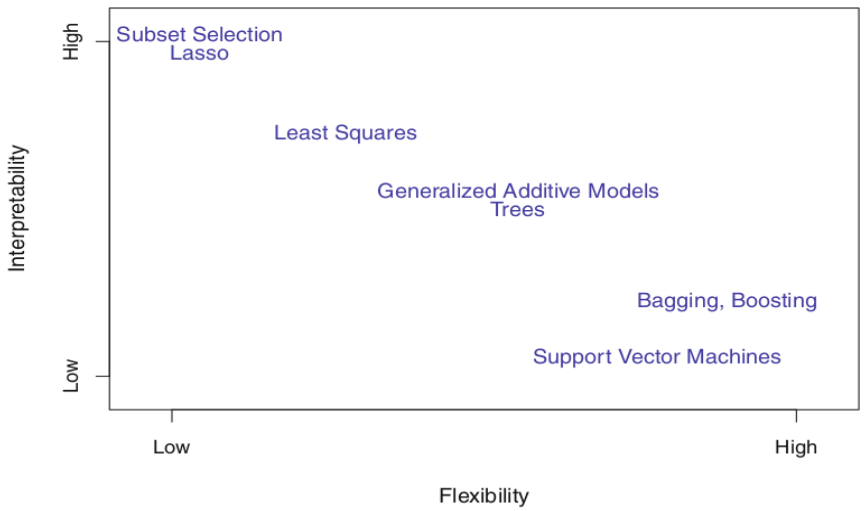

FIGURE 1. Flexibility and interpretability trade-off of different statistical methods (James et al. 2013, 25).

2.1.3 Flexibility and interpretability trade-off

There is a significant trade-off between flexibility and interpretability of the various statistical methods. The flexibility of a method could be understood as the ability to give many shapes and forms to the estimate of f. For instance, linear regression is seen as inflexible since it can only create linear functions. Figure 1 depicts the trade-off between flexibility and interpretability of some well-known methods. Inflexible methods are easy to interpret, making it easier to understand the relationship between the predictors and response, however, these methods might be biased for estimating a much more complicated problem by a too simple model. These methods are suitable in situations where the main objective is inference since they are capable of describing the relationship of the variables in an understandable way. On the contrary, flexible methods can fit data with complex relationship, however, they usually result in complicated estimates of f that make it difficult to understand how variables interact with each other. These methods potentially work better when the task of prediction is interested since the estimated models can match the data better. Unfortunately, highly flexible methods could be at risk of overfitting which essentially means that the model fits the training data too well and consequently follows some patterns randomly caused in the training data only, therefore, the model would have substandard performance facing new data since the random patterns found in the training data would not exist in the new data. (James et al. 2013, 25-26.)

2.1.4 Supervised learning

Most statistical learning problems could be divided into two categories: supervised and unsupervised. This thesis focuses on the supervised learning task. In essence, supervised learning means that the statistical method is trained using a data set that has each of its observation tagged with an answer that the method should come up with. In other words, the data is fully labeled. (Salian 2018.) For example, for every observation of $X_1, X_2, \ldots, X_p$ there is a corresponding value of Y (James et al. 2013, 26). Supervised learning mainly aims at relating the response to the predictors so as to predict the response accurately with

future predictors and this objective matches regression problem well, therefore, supervised learning is commonly used in regression problems. The statistical methods that are used in this thesis are linear regression and random forests, both of which fall into regression and supervised learning category.

2.1.5 The mean squared error

The mean squared error (MSE) computed by the formula:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2,$$

is commonly used to evaluate the performance of a statistical method by quantifying how close the predicted responses are to the actual responses. In the formula, $\hat{f}(x_i)$ stands for the predicted response of the ith observation. The MSE would be small if the predicted responses are close to the observed responses and vice versa. The MSE obtained from using the training data is called train MSE while the MSE obtained from applying the statistical method to new data is called test MSE. Usually, the test MSE is more important since the train MSE only shows the statistical method's performance on old data while the test MSE shows the method's performance on unseen data which can be a valid measure of how well it would predict future data, which is also the main objective of the prediction task. (James et al. 2013, 29-30.)

2.2 Linear regression

2.2.1 Simple linear regression

Linear regression is a simple approach for supervised learning as well as predicting quantitative value. Simple linear regression predicts one variable based on another variable (Lane, Scott, Hebl, Guerra, Osherson & Zimmer n.d., 462). The method involves only one independent variable and one dependent variable, assuming there is a linear relationship between two variables (James et al. 2013, 61). Therefore, the predicted values of the response plotted as a function of the predictor would form a straight line (Lane et al. n.d., 462). Figure 2 shows an example of body weight plotted as a function of height (Sullivan & LaMorte 2016). The stars represent the actual data points while the straight line is the prediction of body weight and the linear relationship is quite visible.

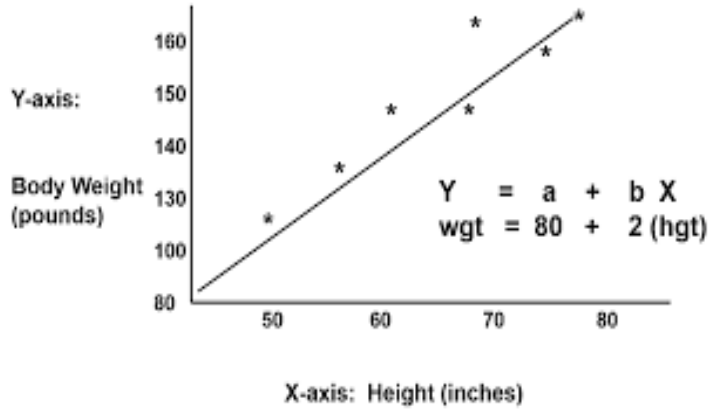

FIGURE 2. Body weight plotted as a function of height (Sullivan, LaMorte 2016).

The simple linear relationship could be expressed mathematically as:

$$Y \approx \beta_0 + \beta_1 X.$$

In this equation, β_0 and β_1 are two unknown constant, in which β_0 is the intercept parameter while β_1 is the slope parameter. They are also known as the model coefficients. By estimating the coefficients, future values of Y could be predicted using this formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

In the formula, $\hat{y}$ is a single prediction of Y based on a value x of X, while $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of the coefficients β_0 and β_1 respectively. (James et al. 2013, 61.)

2.2.2 Estimating β_0 and β_1

Normally, the real values of β_0 and β_1 are unknown, therefore, the data are used to make estimates of them. The objective is to come up with estimates that make the model fit the data well and while there are a number of measures for evaluating this quality, the most commonly used is called least squares which aims at minimizing the residual sum of squares (RSS):

$$RSS = e_1^2 + e_1^2 + ... + e_n^2$$

where $e_i = y_i - \hat{y}_i$ known as the ith residual is the difference between the ith observed value y_i and predicted value $\hat{y}_i$ of Y. Figure 3 shows an example of a linear model fitted using least squares for a data set in which Sales is the response and TV, which stands for TV advertising cost, is the predictor. Each gray line represents an error and the fit is calculated by minimizing the sum of these errors. $\hat{\beta}_0$ and $\hat{\beta}_1$ can be found using the following formulae:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

in which $\bar{y}$ and $\bar{x}$ are the sample means. (James et al. 2013, 62.)

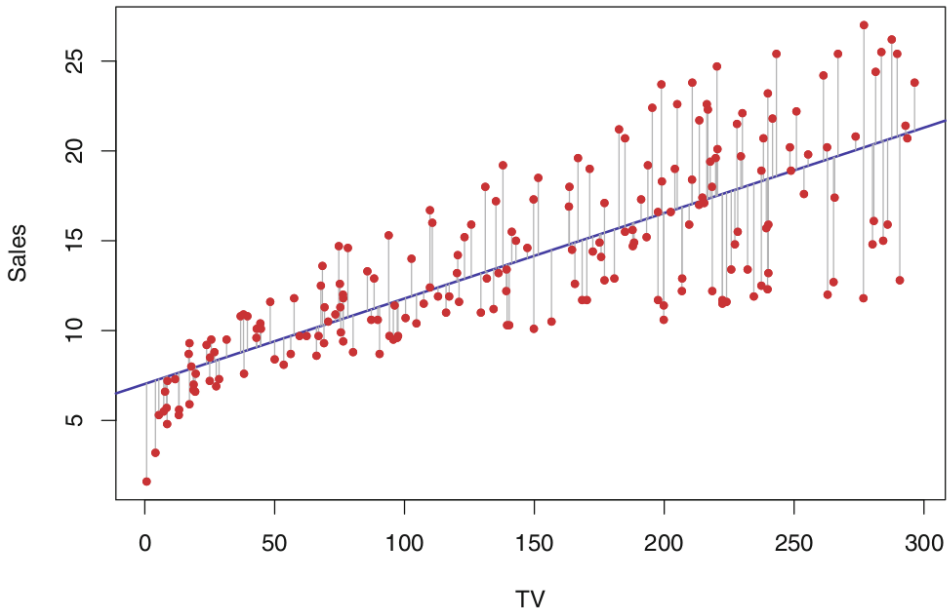

FIGURE 3. Example of a linear model fitted using least squares (James et al. 2013, 62).

2.2.3 Multiple linear regression

In reality, there are usually more than one predictors, therefore, simple linear regression could be insufficient to compute a reasonable prediction. Multiple linear regression alleviates this problem by extending the simple linear model to fit multiple predictors, giving each predictor a slope coefficient:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon.$$

The coefficients could be interpreted as the average effect on Y of a unit increase in X_j while the other predictors remain the same. Similar to simple linear regression, multiple linear regression predicts future values of the response by the following formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p.$$

The coefficients are also estimated by minimizing the $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. (James et al. 2013, 71-72.)

2.2.4 Best subset selection

One of the main concerns when multiple linear regression is applied is choosing the predictors to include in the model because of several reasons. First, redundant predictors that do not improve the regression performance should be removed so that the model becomes simpler and easier to interpret. Second, most of the times, not all of the predictors have meaningful relationship with the response. If all of the predictors are used to train the model, the ones that do not have strong correlation with the response would actually add noise and negatively affect the performance of the model. Additionally, collinearity would be caused if too many variables are working the same task. Finally, using only important variables helps reducing the cost of measuring the redundant variables. (Faraway 2002, 124.)

One possible method for identifying the best predictors for multiple linear regression is called best subset selection. Basically, this method fits separate least squares linear models for each combination of the predictors, then the resulting models would be reviewed to come up with the best model. This algorithm describes the steps involved in using best subset selection:

- Let M_0 be a null model which contains no predictors and only has the sample mean as the intercept
- For k = 1, 2, …, p:
  - Fit $\binom{p}{k}$ linear models that have k predictors
  - Choose the best among these by smallest RSS and call it M_k
- Select the best model from $\mathrm{M}_0, \ldots, \mathrm{M}_p$. (James et al. 2013, 205.)

2.3 Random forests

2.3.1 Decision trees

Essentially, decision trees solve regression task by split the data set into different regions, calculate the mean of the responses in each region then simply use that as the prediction for every observation in that region (James et al. 2013, 306). Figure 4 and 5 illustrate an example of a decision tree with two predictors $X1$ and $X2$ (Drakos 2019). First, the data set is split into two regions based on the split value 0.302548 of predictor $X2$. The region that has $X2 < 0.302548$ is then divided again using split value 0.800113 of predictor $X1$ and for simplicity, these two regions are called $R1$ and $R2$ (Figure 5). At this point, the splitting stops and the means of the responses of $R1$ and $R2$ are calculated and used for prediction. In other words, every observation that has $X2 < 0.302548$ and $X1 < 0.800113$ would have a predicted response value of 0.807 (mean of responses of $R1$) and, on the other hand, every observation that has $X2 < 0.302548$ and $X1 \geq 0.800113$ would have a predicted response value of 0.5 (mean of responses of $R2$).

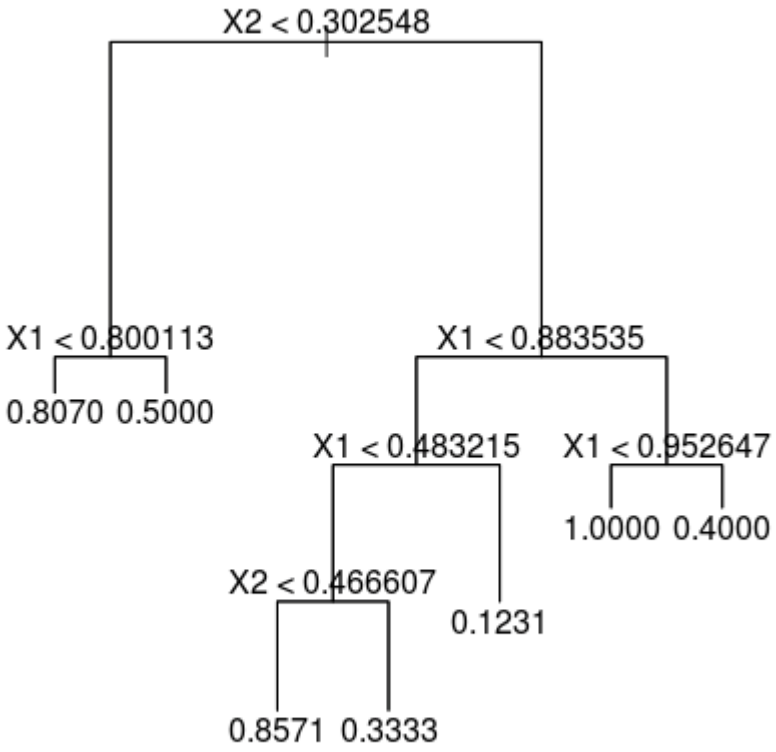

FIGURE 4. Example of a decision tree (Drakos 2019).

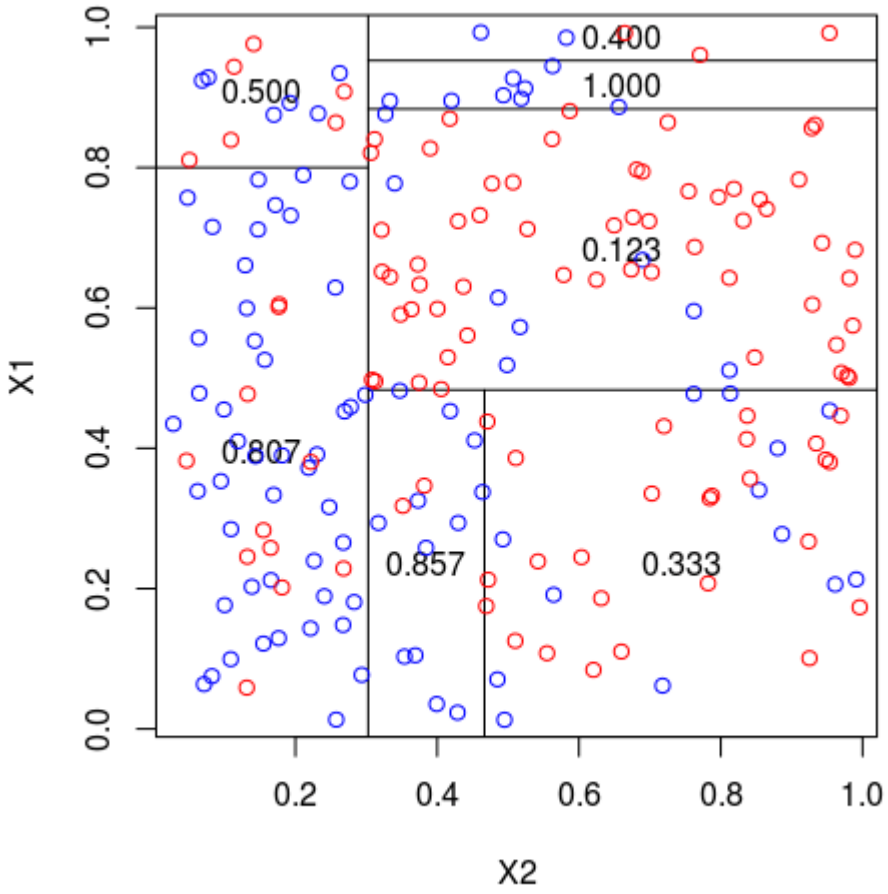

FIGURE 5. Example of a decision tree (Drakos 2019). $R1$ is the bottom left region and $R2$ is the top left region.

According to James et al. (2013, 306), the decision on how to split the data focuses on minimizing the RSS:

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

in which $\hat{y}_{R_j}$ is the mean response of the R_j region. Starting with the whole training data set, all predictors $X_1, X_2, \ldots, X_p$ and all possible cutpoints s are considered then the ones that yield the tree with the lowest RSS are picked. The cut-

point s is the value that is used to split the data into two regions $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$. This process is applied recursively to split the data further until a certain criterion is met, such as until no region has more than five observations. Apparently, decision trees offer several advantages as a statistical method. To begin with, trees are intuitive, therefore, easy to be explained. Additionally, they could be easily visualized and interpreted. (James et al. 2013, 315.) Unfortunately, trees also have a major drawback known as high variance which means even a small change in the data could affect the prediction negatively (Hastie, Tibshirani & Friedman 2009, 312). Usually, models that overfit would have high variance. A method called random forests which is an extension of decision trees can help alleviate these problem.

2.3.2 Random forests

According to Ruozzi (2016), it is possible to reduce the variance of a variable by averaging the whole set. Suppose there are $Z_1, Z_2, \ldots, Z_p$ random variables, then the variance reduction could be mathematically expressed as follows:

$$Var\left(\frac{1}{p}\sum_{i=1}^{p} Z_i\right) = \frac{1}{p}Var(Z_i).$$

Evidently, the variance of the variables is reduced by p times by taking average. Therefore, it would be intuitive to reduce the variance of a statistical method by building several distinct models from different training data sets then averaging those models. However, in reality, there would not be multiple different training data sets, therefore, a brilliant resampling technique called bootstrap is used to generate more data sets. Essentially, bootstrapping means randomly taking n observations from a size n data set with replacement which means that a same observation can occur in the new data (James et al. 2013, 189). This process can be applied repeatedly to create different data sets. Figure 6 illustrates the bootstrap process as described (Yen 2019).

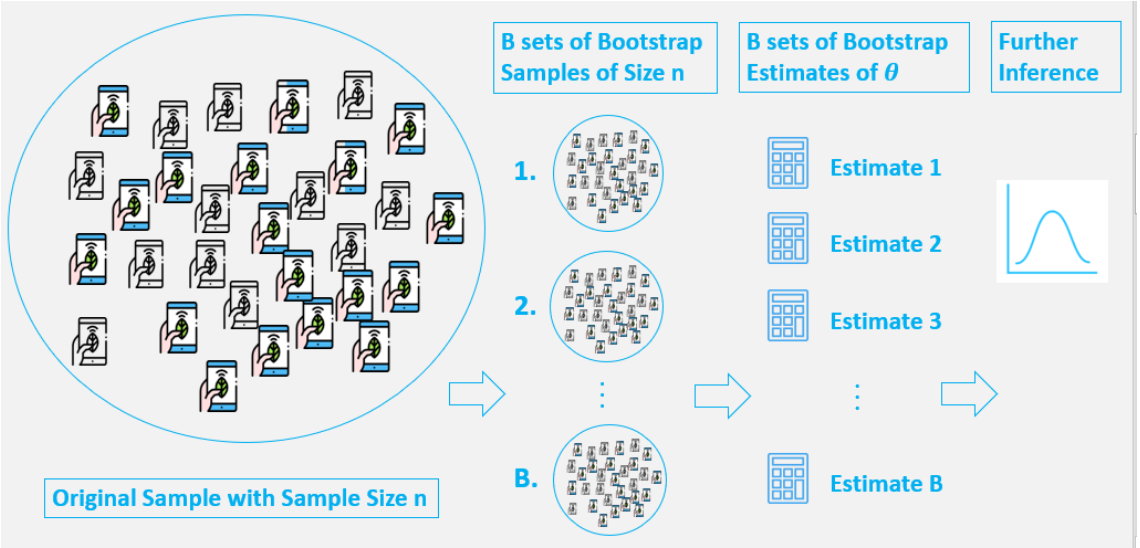

FIGURE 6. The bootstrap process (Yen 2019).

Basically, the random forests method uses bootstrapping to generate a number of different data sets, builds tree models from those data then averages the models. However, at the model building process, whenever a split decision is made, only a random sample of the predictors are taken into consideration. For example, suppose there are $X_{1,}\ X_{2,}\ \ldots,\ X_{p}$ predictors, at each split a new random sample of m predictors is made and usually $m \approx \sqrt{p}$. This is because if all of the predictors are considered at each split, there might be some prominent predictors that the trees would end up choosing, thus creating a number of similar trees and averaging similar trees would not significantly reduce the variance. Therefore, by allowing only a sample of predictors at each split, random forests reduce the significance of the outstanding predictors and give other predictors more chance, making the result trees less similar and correlated and thus more reliable. (James et al. 2013, 320.)

3 METHODOLOGY

This thesis was carried out as a data mining project that focused on making future predictions for the house price in Sindian District, New Taipei City, Taiwan. According to Shmueli, Bruce, Yahav, Patel & Lichtendahl Jr. (2018), a typical data mining process includes the following steps:

- Understand the purpose of the project
- Obtain the data
- Explore, clean, preprocess data
- Reduce data dimension
- Determine the mining task
- Divide the data
- Choose the techniques
- Use algorithm
- Interpret the results
- Deploy.

As shown in Figure 7, Walker (2016) also came up with a similar process for data mining, although this was divided into four main steps with smaller substeps rather than a detailed list of ten steps by Shmueli et al. (2018).

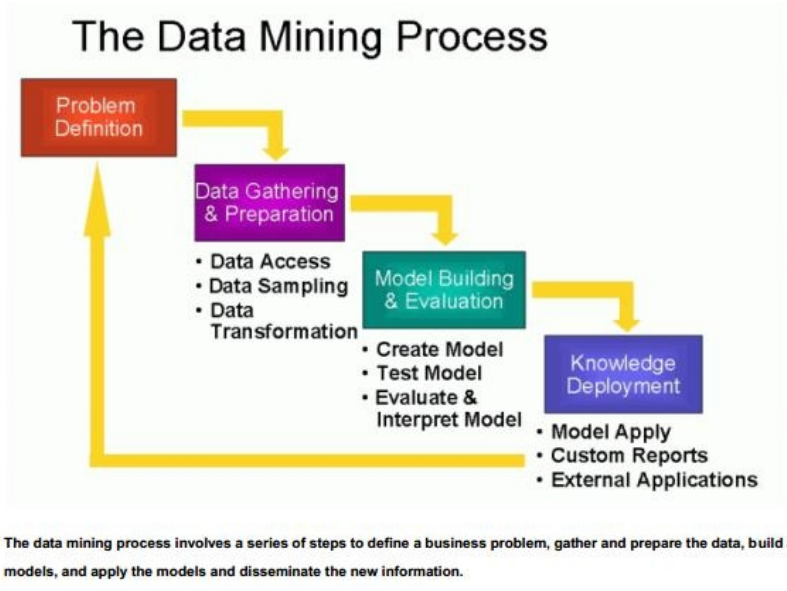

FIGURE 7. The data mining process (Walker 2016).

The process starts with defining the main purpose of the project. This is important since having a good understanding of the purpose of the project would lead to a clear overview, better planning and resource investment.

Once the purpose is properly identified, the next step would be data acquisition, which potentially involves taking sample from a larger database or even from multiple data sources. (Shmueli et al. 2018.) Therefore, this step requires a number of important skills such as sampling, extracting, handling or merging data. Indeed, when it comes to data, there is a wide variety of data kinds that require different treatments such as database data, data warehouse and transactional data. Relational databases consist of tables, each table has a set of attributes and tuples which represent objects that are identified by unique keys and attribute values. Normally, this kind of data is accessed by database queries which would be transformed into relational operations to allow retrieval of specified subset of data. Data warehouses are repositories of information from various sources and are modeled by a multidimensional data structure known as data cube. Even though data warehouse tools could help in data analysis, additional tools are still needed for deeper analysis. Transactional data store transactions as records, for examples, purchases, bookings and clicks. Additionally, there are other kinds of data with different forms and structures such as sequence data, data streams, spatial data, design data, multimedia data and networked data which offer various kinds of information. (Han, Kamber & Pei 2012, 9-14.) While data mining could involve massive database, the actual analysis might need only a much smaller amount and moreover, using smaller data if possible also significantly reduces time and resource cost, thus knowing what kinds of data, what portion of the whole data and how much is needed is crucial. (Shmueli et al. 2018.)

After the data are properly gathered, they need to be thoroughly explored and preprocessed to ensure usable condition. This step usually consists of handling missing values, identifying anomalies such as outliers, visualizing data. (Shmueli et al. 2018.) Data exploration aims at having a good insight on the data. There are many kinds of data attribute such as nominal, ordinal, binary and numeric and this information matters because data attribute decides the

kind of values that data have. For examples, binary attribute has only two possible values that are 0 and 1 which mean absence and presence respectively while numeric attribute refers to quantitative values usually measured in real or integer numbers. Another typical exploratory step would be examining basic statistical descriptions of the data since these quantitative values reveal a lot about the data's properties. Usually, these statistical descriptions include values that measure central tendency such as mean, median and mode, or measure the dispersion such as range, variance and standard deviation. Data exploration also heavily relies on data visualization with some common approaches such as scatter plot, histogram and boxplot since visualization makes it easier to notice trends and relationships. (Han et al. 2012, 41-56.)

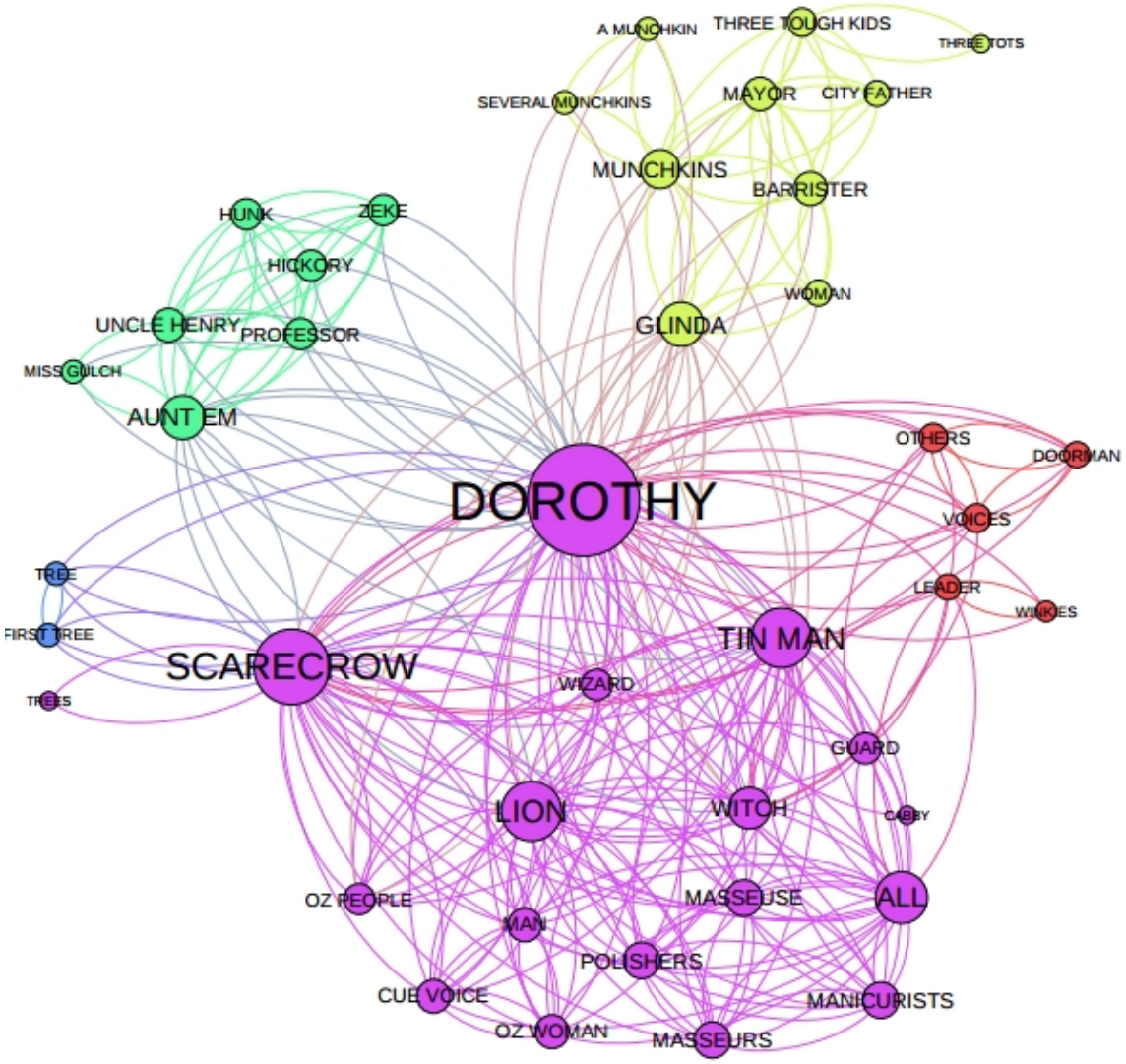

FIGURE 8. Network graph for The Wizard of Oz (Blue, 2015).

Figure 8 is an example for data visualization using network graph to illustrate the interactions of the characters in The Wizard of Oz. Each node is a character while the lines represent the interactions between them and the bigger the node, the more interactions that character has. Additionally, the characters are also classified according to their community and coded by color. (Blue, 2015.) Moving on to data preprocessing, this usually has four main forms: data cleaning, data integration, data reduction and data transformation. Data cleaning involves handling missing values, outliers and inconsistencies. Next, data integration essentially means integrating multiple data files together. When the data is too large and difficult to handle, data reduction could be used to reduce the dimensions resulting in a new data set with fewer variables. Lastly, data transformation alters the data so that mining method would have better performance. (Han et al. 2012, 85-87.)

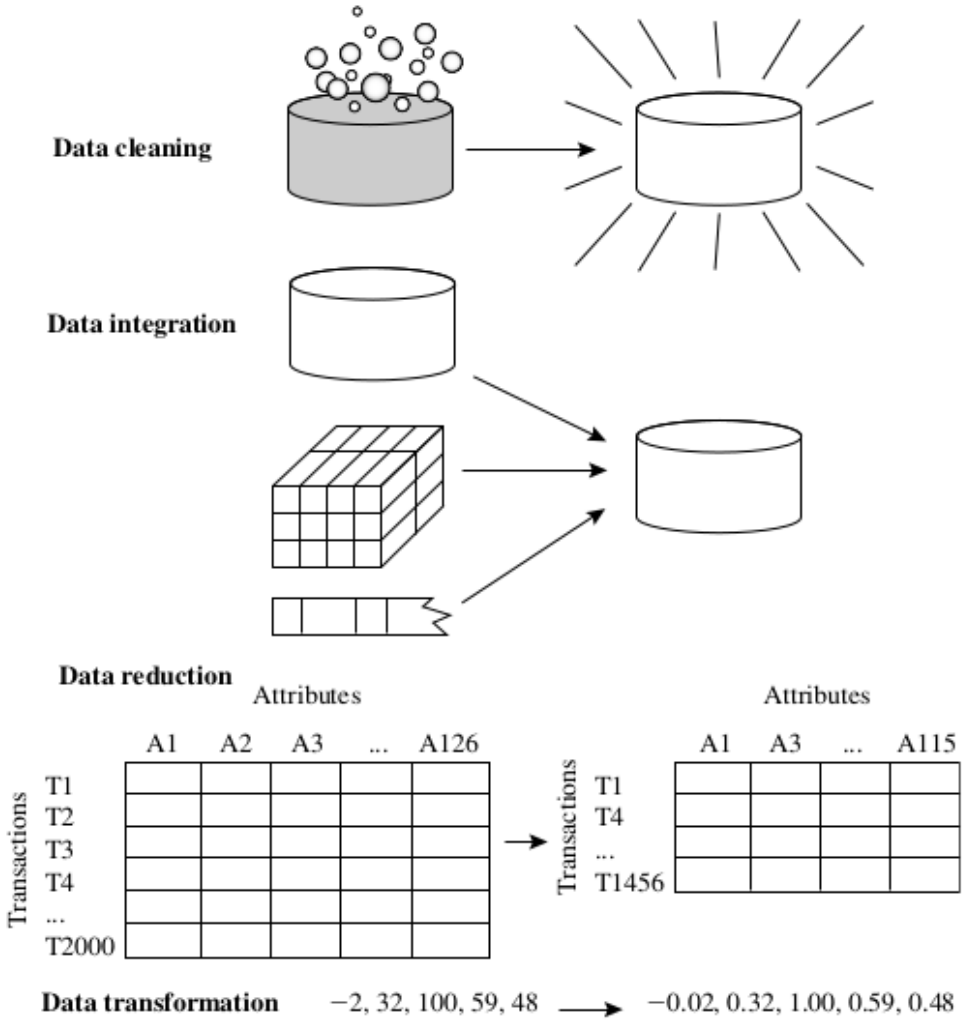

FIGURE 9. Forms of data preprocess (Han et al. 2012, 87).

The next step in the mining process would be deciding on the suitable mining techniques, depending on the goal of the project and information from the exploration step. Typically, this step involves trying different statistical learning methods or even different settings of the same method. Then, the models are tested to identify the optimal choice and the results are interpreted to extract meaningful insights. (Shmueli et al. 2018.)

The purpose of this thesis was defined to be predicting house price in the Sindian District, New Taipei City, Taiwan by applying statistical learning methods on recorded data about real estate to build regression models. The data set was originally from Yeh & Hsu's research paper (2018) and was made publicly available on UCI Machine Learning Repository (Yeh & Hsu 2018, 260-271). After downloaded, the data were handled using the software R. First, the data were examined for missing values, distributions of the variables and basic descriptive statistics. Then a correlation test was performed to check the relationship between each of the predictors and the response. After the exploration was done, the data were split into a training set and a test set, with the training set took two-thirds observations of the original data set. The training set was then used to fit regression models. The first model was multiple linear regression. Since there were six predictors, best subset selection method was used to identify the best linear model by fitting the training data set $\binom{6}{k}$ times, each time with a different set of k predictors for $k = 1, 2, \ldots, 6$. The models were fit using least squares method that focuses on minimizing:

$$\sum_{i=1}^{n} \epsilon^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta),$$

in which:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix},$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}.$$

Using linear algebra, it can be shown that the least squares estimates for β is the vector:

$$\hat{\beta} = (X'X)^{-1}X'y.$$ (Bremer 2012.)

In R, there is a built-in package that automatically computes the estimated coefficients using least squares. For each value of k, the best model was identified using RSS or R^2 and named M_k. R^2 is a quantitative value that measures the proportion of variability in the response that can be explained using the chosen predictors and is defined by:

$$R^2 = \frac{(TSS - RSS)}{TSS} = 1 - \frac{RSS}{TSS},$$

with $TSS = \sum (y_i - \bar{y})^2$ known as total sum of squares which measures the total variance in the response. The value of R^2 is between 0 and 1, being close to 1 means a large proportion of variability of the response is explained and close to 0 means the opposite. Therefore, M_k was supposed to be the model with the highest R^2 or lowest RSS. Then the best linear model was chosen from $\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_6$ using adjusted R^2:

$$Adjusted\ R^2 = 1 - \frac{RSS/(n - d - 1)}{TSS/(n - 1)},$$

where n is the number of observations and d is the number of variables. A model with a large value of adjusted R^2 would have small test error. Adjusted R^2 had to be used in this step because $\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_6$ are models with different number of predictors and R^2 could not evaluate their performance fairly due to the fact that R^2 just increases as the number of predictors increases. (James et al. 2013, 205-210.) Subsequently, the training data set was used to fit random forests models by bootstrapping 500 different data sets, building different trees from these sets then averaging to get the result. This process was also conducted using the programming language R. Additionally, since the number of predictors was rather small, random forests models were fit with every possible value for the number of predictors considered at each split. The resulting random forests models as well as the linear models were then used to make predictions using the test data set and the performance was evaluated using test MSE.

4 DATA EXPLORATION

4.1 Basic descriptive statistics and correlation

There are 414 observations, 6 independent variables and 1 dependent variable, in which the house price is the dependent variable that needs to be predicted and the others are independent variables that are used to make predictions. The variables' names are as follows:

- X1: transaction date
- X2: house age (unit: year)
- X3: distance to the nearest MRT station (unit: meter)
- X4: number of convenience stores in the living circle
- X5: latitude (unit: degree)
- X6: longitude (unit: degree)
- Y: house price of unit area (10,000 New Taiwan Dollar/Ping, 1 Ping = 3.3 meter squared)

Below is a table containing some basic descriptive statistics of the variables. The variable X1 was not included in this table because its values represent the month and year of transaction, for which computing these numerical statistics did not provide much information. Additionally, there were no missing values in the data.

Table 1. Basic descriptive statistics of the data

	X2	X3	X4	X5	X6	Y
Min	0.000	23.38	0.000	24.93	121.5	7.60
1st Quartile	9.025	289.32	1.000	24.96	121.5	27.70
Median	16.100	492.23	4.000	24.97	121.5	38.45
Mean	17.713	1083.89	4.094	24.97	121.5	37.98
3rd Quartile	28.150	1454.28	6.000	24.98	121.5	46.60
Max	43.800	6488.02	10.000	25.01	121.6	117.50

Subsequently, the correlation of the variables were investigated. Figure 10 shows a scatter plot matrix of the data which gives a general overview of the relationship between the variables by putting multiple plots together.

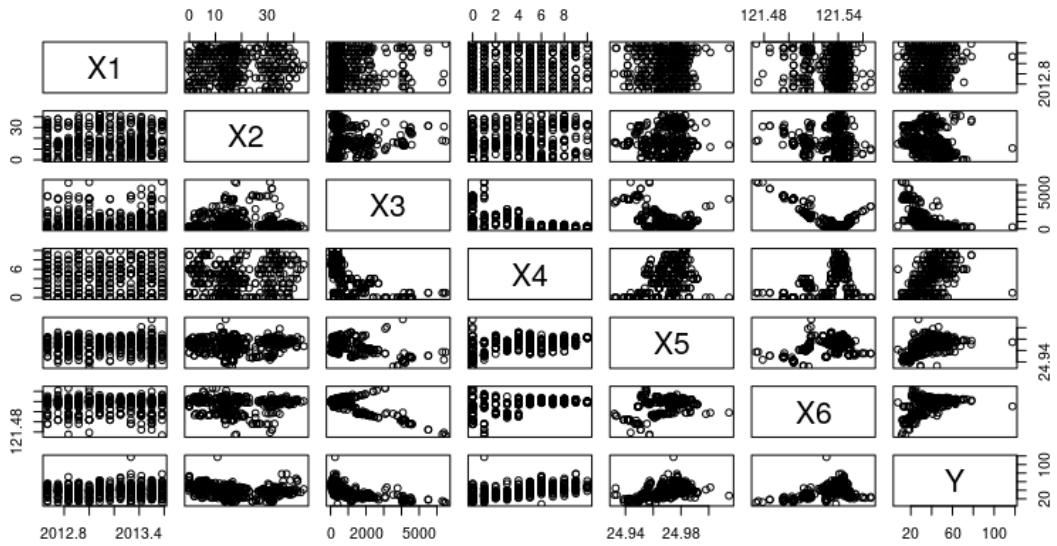

FIGURE 10. Scatter plot matrix of the data

It appeared that the response Y had some visible relationships with predictors X3, X4, X5 and X6. On the other hand, there were no likely relationships between Y and X1 or X2. Moreover, X3 and X6 also seemed to correlate. This could be confirmed by computing the correlations between these variables.

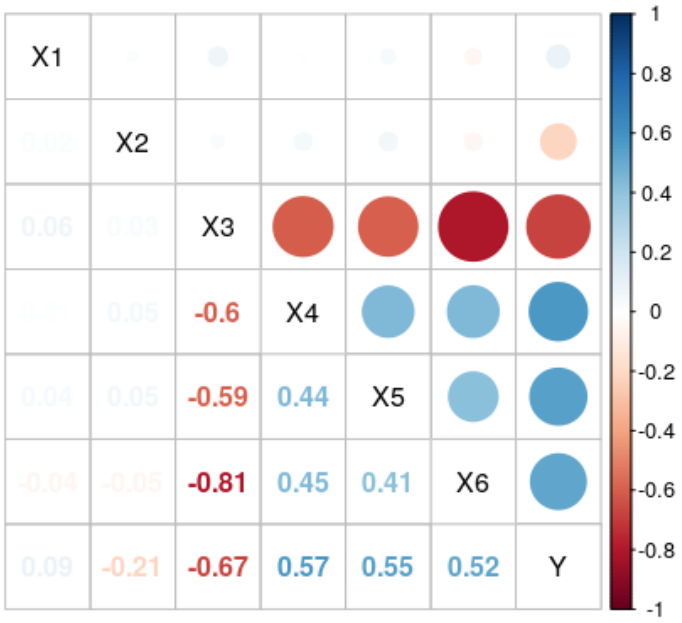

FIGURE 11. Correlations of the variables

Figure 11 shows the computed correlation of the variables whose values range from -1 to 1, with being close to 1 means a strong positive relationship and being close to 1 means a strong negative relationship while being close to 0 means the relationship is weak. Evidently, Y had considerable relationships with X3, X4, X5 and X6. However, X3 and X6 also strongly correlated. Judge, Hill, Griffiths, Lutkepohl & Lee (1988, 882) stated that the existence of near linear relationships among the explanatory variables was called multicollinearity. This relationship would make the estimated coefficient have large variance and thus unstable from sample to sample. This instability renders the estimate unreliable. (Judge et al. 1988, 882.) According to Goldberger (1991, 245), the variance of an estimated coefficient $\hat{\beta}_j$ for variable x_j is given by:

$$\sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{[(1 - R_j^2)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2]},$$

where $\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the variation of x_j and R_j^2 known as the coefficient of determination in the auxiliary regression of x_j on other x's would be close to 1 if there

is high multicollinearity. It is obvious that if everything else is kept the same, a large R_j^2 means a large variance $\sigma_{\hat{\beta}_j}^2$, which would make the estimate $\hat{\beta}_j$ unreliable because the sample value could be significantly different from the true β_j. (Goldberger 1991, 245.)

After correlation investigation, each variable was examined more carefully. Figure 12 shows the distribution of the response. The values appeared to have normal distribution with a slight positive skewness of 0.59. This could show that most of the houses in the Sindian District had prices close to the average price and there were fewer expensive than average or cheap houses, which seemed intuitively correct.

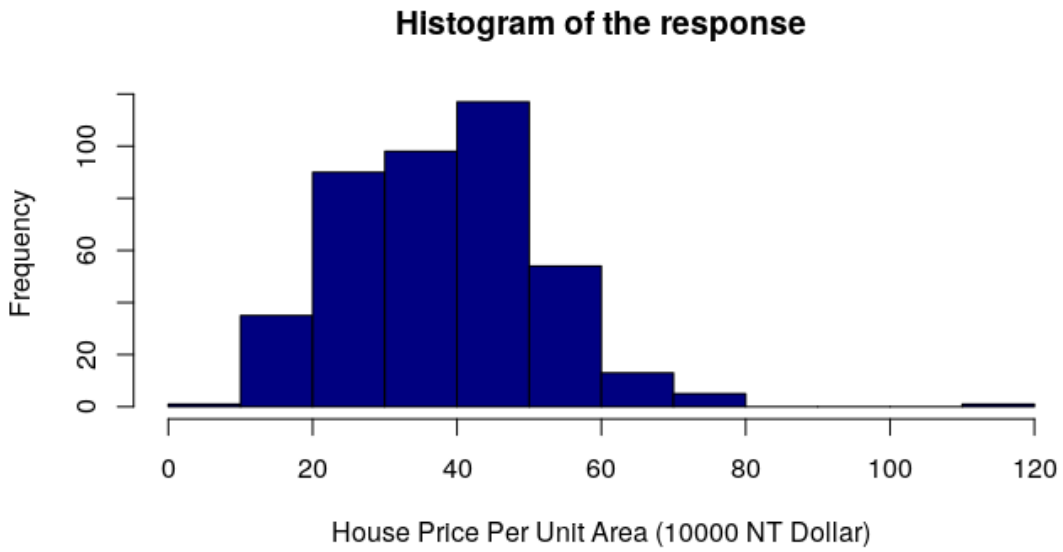

FIGURE 12. Histogram of the response.

4.2 X1 – Transaction date

The transaction date (X1) is formatted as real numbers, consisting of an integer part and a fractional part in which the former indicates the year of the transaction while the latter could be used to interpret the month of the transaction using this formula:

$$(X1 - Year) \times 12 = Month.$$

For example, 2013.500 would be interpreted as June 2013. By following this formula, it could be shown that the data set contained values from August 2012 to July 2013. A boxplot of the response divided into different transaction months is shown in Figure 13. It could be seen that there was no significant difference in the house price among the months, which agreed with the correlation check in Figure 11.

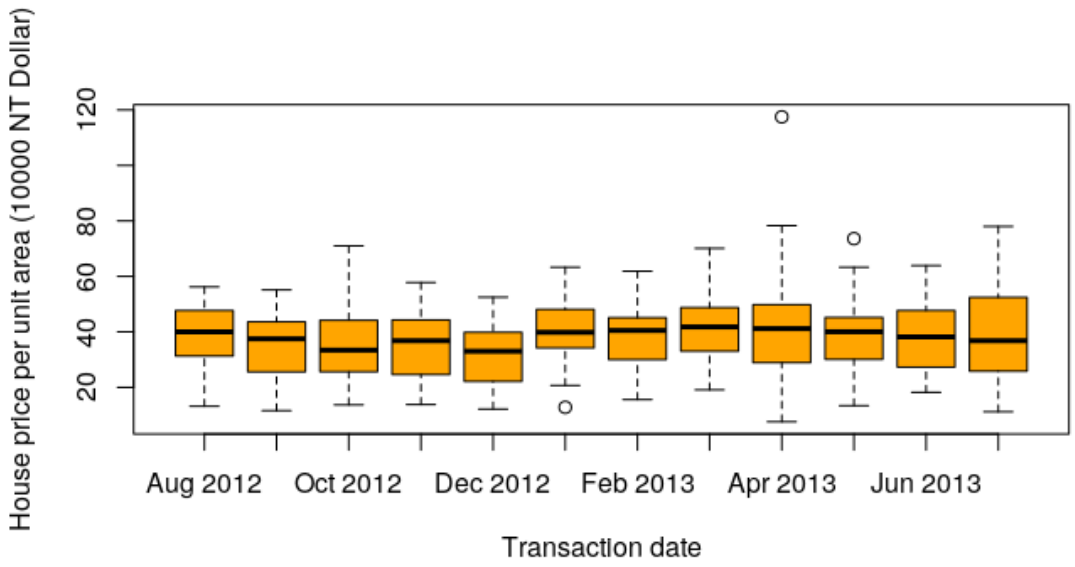

FIGURE 13. Boxplot of house price based on transaction date.

4.3 X2 – House age

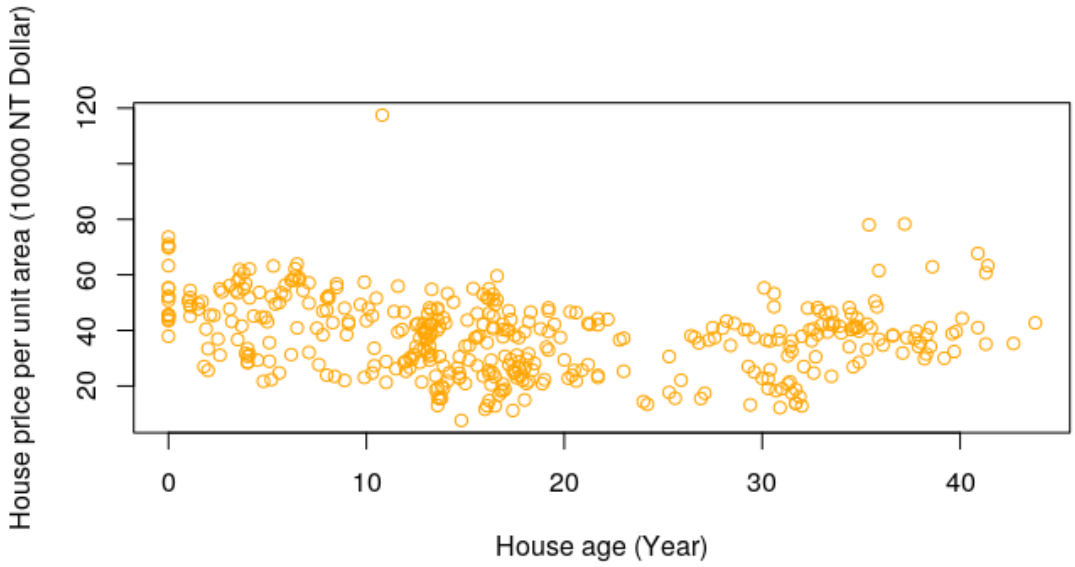

FIGURE 14. Plot of house price and house age.

The data contained houses ranging from newly built to more than 40 years old. There was a decreasing trend in the price for houses that were built longer ago, however, this trend was insignificant (Figure 14). Some older houses that were more than 35 years old even had higher prices than newer houses, but these were only a few exceptional cases.

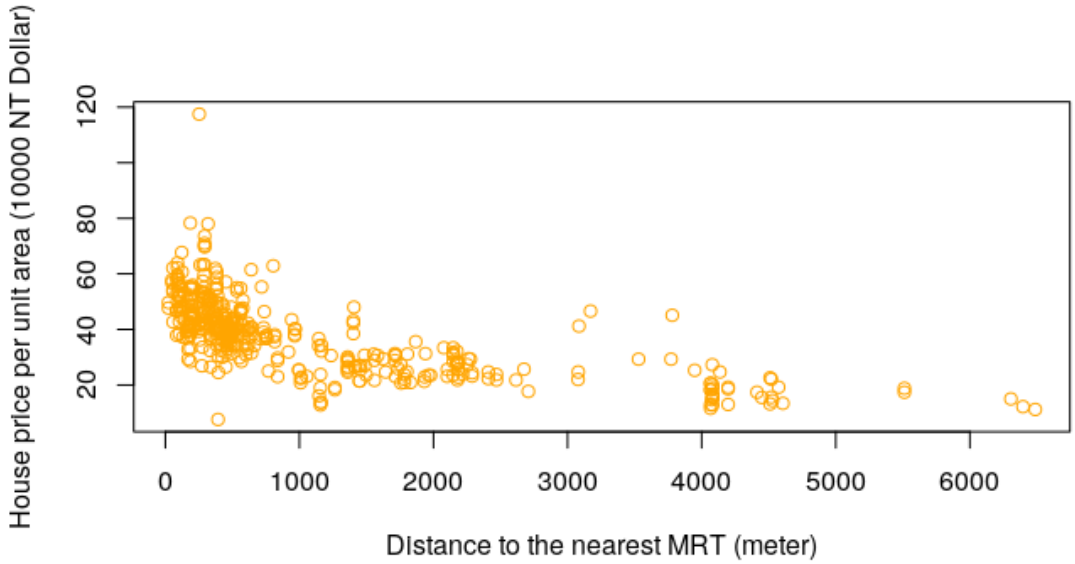

FIGURE 15. Plot of house price and distance to the nearest MRT.

4.4 X3 – Distance to the nearest MRT

MRT stands for Taipei Mass Rapid Transit is a metro system in Taiwan. It was clear that houses that were nearer to the metro were more expensive (Figure 15). Additionally, most of the houses were within half a kilometer distance from the nearest metro. This could be confirmed by the histogram of X3 in Figure 16.

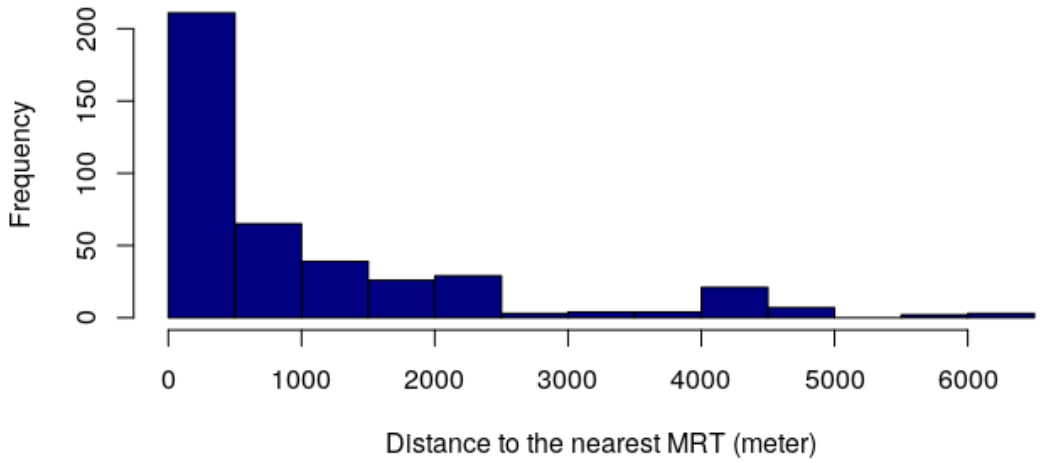

FIGURE 16. Histogram of distance to the nearest MRT.

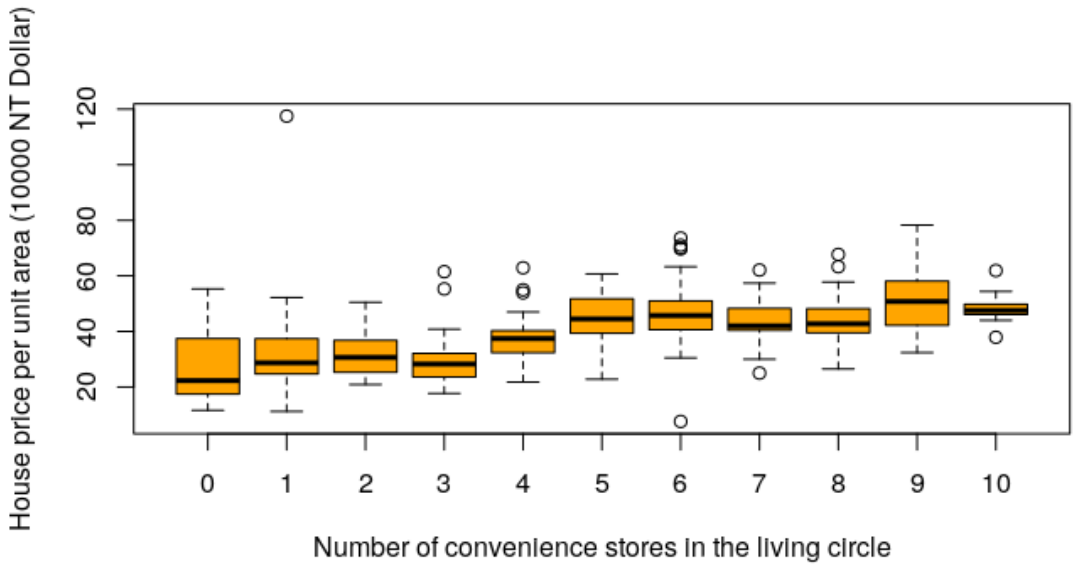

FIGURE 17. Plot of house price and number of convenience stores in the living circle.

4.5 X4 – Number of convenience stores in the living circle

The variable X4 recorded the number of convenience stores in the living circle, which was defined to be within 500 meters of the house (Yeh & Hsu 2018, 260-217). As expected, houses that were near more convenience stores would have higher value (Figure 17).

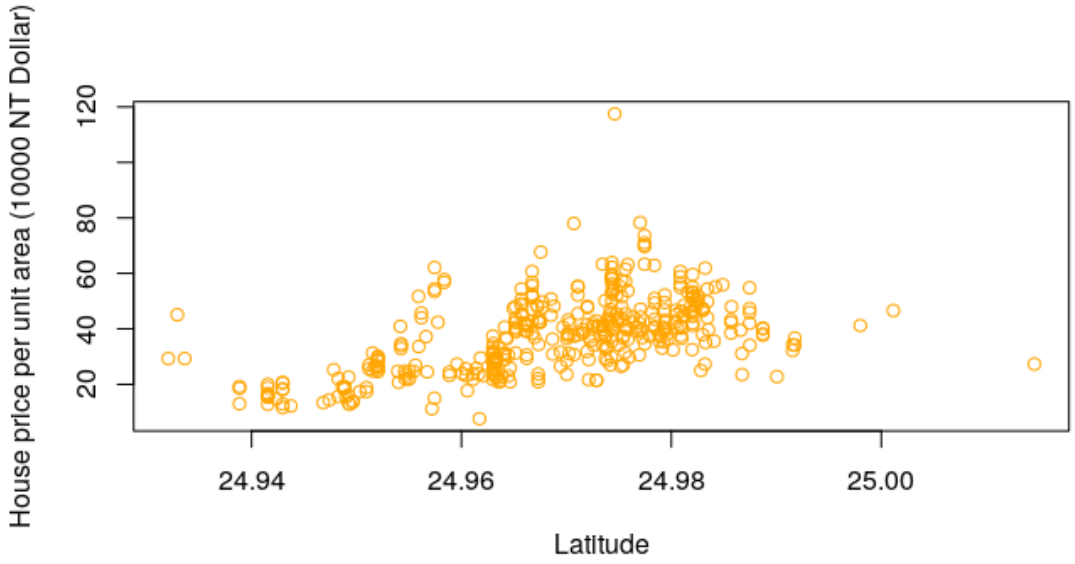

FIGURE 18. Plot of house price and latitude.

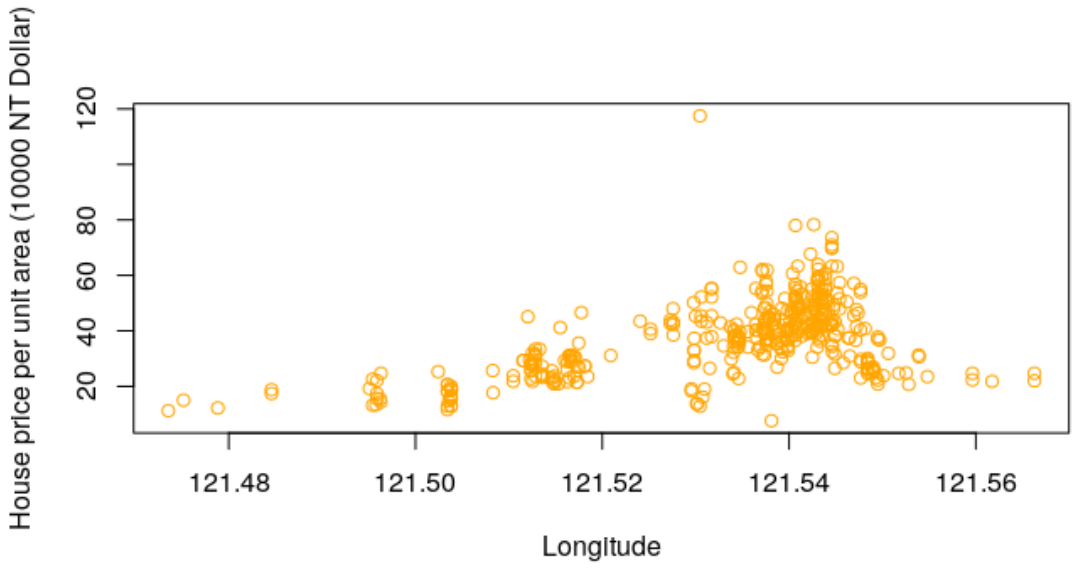

FIGURE 19. Plot of house price and longitude.

4.6 X5 & X6 – Geographic coordinates

Finally, the plots between the house price and the houses' geographic coordinates shown in Figure 18 and 19 were studied. The location of the house could be considered important factor in predicting the price. For example, houses that are near the downtown would help save time traveling to the office or shopping places (Yeh & Hsu 2018, 260-271). The plots showed that there were a slight positive correlation between the price and the location.

5 RESULTS

5.1 Multiple linear regression

The results of applying best subset selection is shown in the table below, in which each row represents the number of explanatory variables included in the model and each column is the variable name. For example, in row 3, X2, X3 and X4 are marked which means the best linear model that used 3 predictors is the one that used X2, X3 and X4. Surprisingly, even though X1 and X2 did not have strong correlation with Y individually, they were still included in models that used more than 3 variables.

Table 2. Best subset selection.

	X1	X2	X3	X4	X5	X6
1			x			
2		x	x			
3		x	x	x		
4	x	x	x	x		
5	x	x	x	x	x	
6	x	x	x	x	x	x

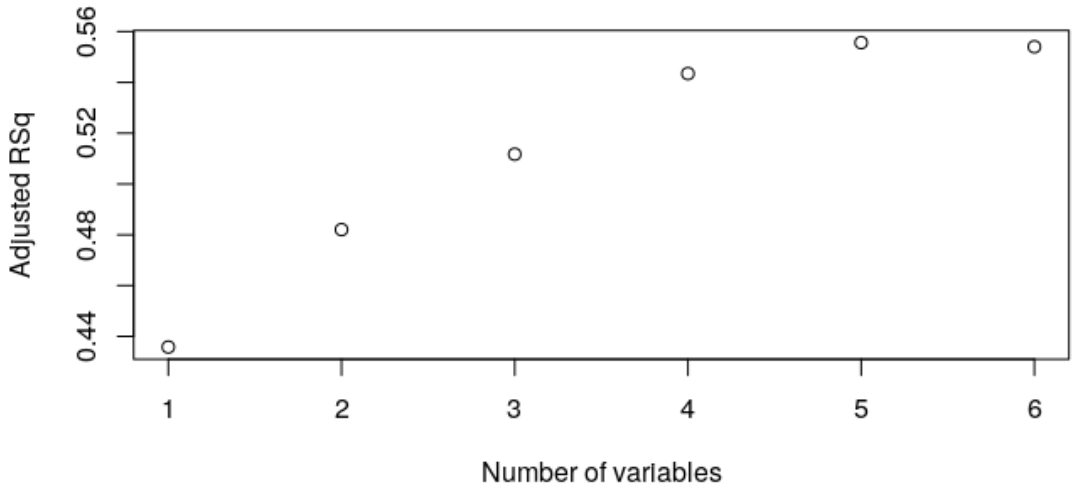

FIGURE 20. The adjusted R^2 of linear models with different number of variables.

Figure 20 shows a plot of the adjusted R^2 of the models, indicating that models that used 5 and 6 variables performed the best and therefore were used to predict the house price using the test data set. Below are the summary of these models, which contain the estimates of the coefficients as well as their standard error, t value and p value. In this case, the standard error of an estimated coefficient is an estimate of its standard deviation, the t value is the estimated coefficient divided by its standard error and the p value is the probability of getting a similar result in a different data set where the variable has no predictive power. (Princeton University Library 2007.) The stars next to the p values represent the significance of the estimated coefficient based on its p value, with 3 stars meaning the most significant. For example, in Figure 21, the p value of the estimated coefficient of X2 was 0.000482, which meant the probability of having this same result for X2 in a random data set was only 0.0482%, therefore, X2 truly contributed to predicting Y in this data set. Additionally, the values of the estimated coefficients could indicate how the response would change if a predictor increase or decrease one unit while other predictors stay the same. In the model with 5 predictors, the p-values indicated that all of the predictors' coefficients were significant and this model had a test MSE value of 61.185. The summary of the 6 predictor model showed that X6 was not a significant predictor and this model had a test MSE value of 61.164.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.942e+04  4.243e+03  -4.576 7.23e-06 ***
X1           7.148e+00  2.023e+00   3.534 0.000482 ***
X2          -3.035e-01  5.056e-02  -6.004 6.20e-09 ***
X3          -4.474e-03  6.874e-04  -6.509 3.66e-10 ***
X4           1.252e+00  2.558e-01   4.896 1.69e-06 ***
X5           2.031e+02  6.136e+01   3.309 0.001062 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 21. Summary of linear model with 5 predictors.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.884e+04  9.015e+03  -2.090 0.037542 *
X1           7.148e+00  2.027e+00   3.527 0.000494 ***
X2          -3.037e-01  5.068e-02  -5.992 6.65e-09 ***
X3          -4.526e-03  9.883e-04  -4.579 7.15e-06 ***
X4           1.251e+00  2.569e-01   4.870 1.91e-06 ***
X5           2.024e+02  6.218e+01   3.255 0.001278 **
X6          -4.585e+00  6.348e+01  -0.072 0.942469
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 22. Summary of linear model with 6 predictors.

5.2 Random forests

The test *MSE* of the random forests models with different number of variables considered at each split are shown in Figure 23.

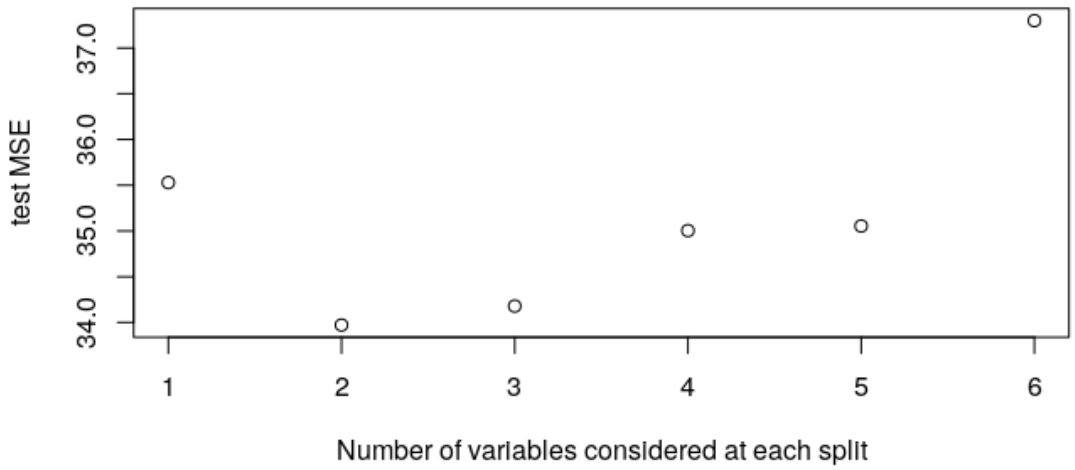

FIGURE 23. Test MSE of random forests models.

It appeared that the models that used only 2 or 3 random variables at each split had the lowest test *MSE*. The detailed test *MSE* values are shown below.

Table 3. Test MSE of random forests models.

Number of variables at each split	1	2	3	4	5	6
Test *MSE*	35.52947	33.97212	34.17915	35.00408	35.05431	37.29897

Additionally, the permutation importance of the variables in the random forests models in which the number of variables considered at each split was 2 (m = 2) and 3 (m = 3) were computed and are shown below in Figure 24 and Figure 25. This quantitative value measured by the increase in *MSE* represents the drop in the prediction accuracy of the model by permuting the values of a predictor. The essential idea is that if a predictor is important for the prediction then permuting

its values would considerably reduce the accuracy or the performance of the model.

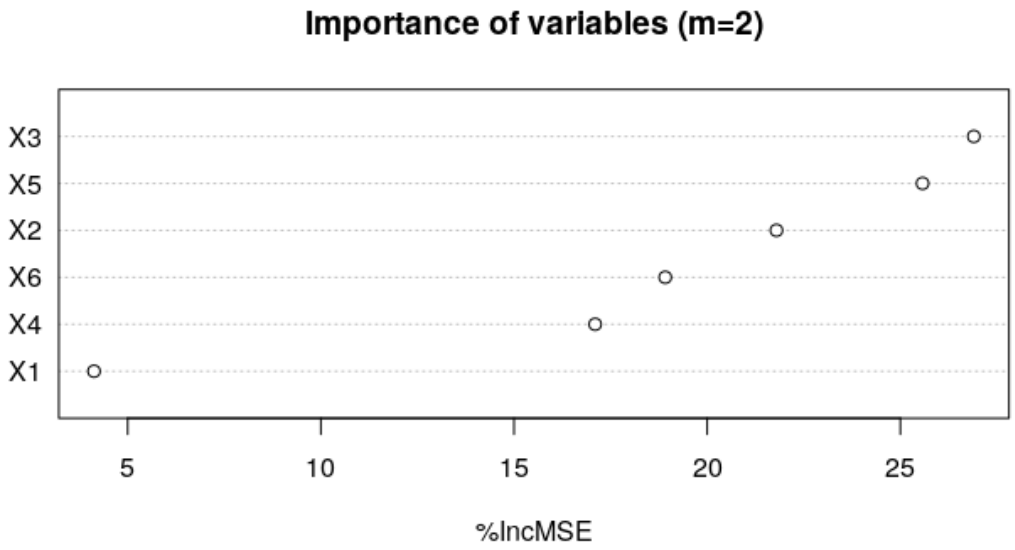

FIGURE 24. Variable permutation importance of random forests model (m = 2).

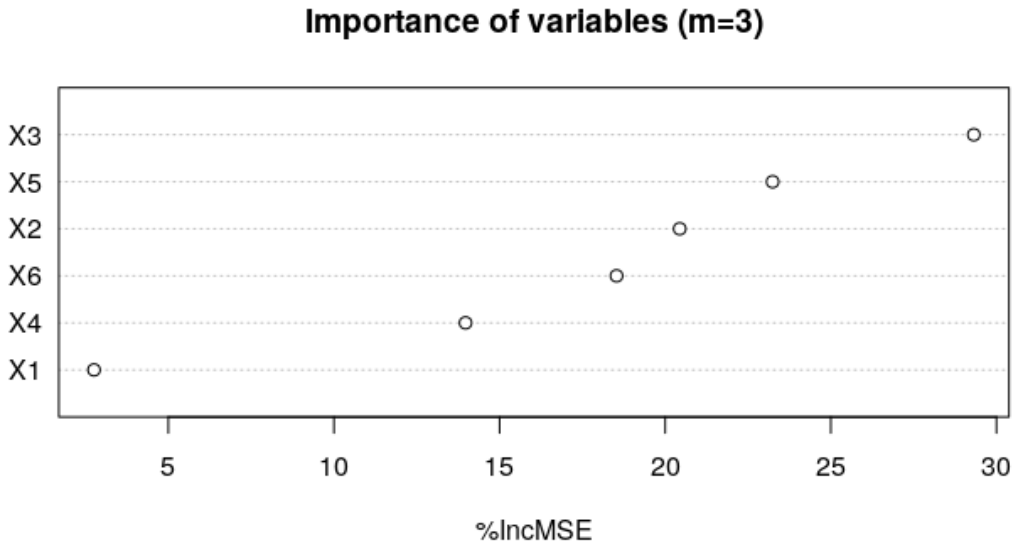

FIGURE 25. Variable permutation importance of random forests model (m = 3).

6 DISCUSSION & CONCLUSION

6.1 The models

For the multiple linear regression method, the performance and accuracy of a model is usually measured by the MSE (James et al. 2013, 29). Even though the model that used all of the predictors yielded the best test MSE, the other model that did not include predictor X5 should be a better choice due to several reasons. First, the model with 5 predictors had a test MSE that closely matched that of the full model but the reduction of one predictor made the former easier to interpret and explain. Second, it is also recommended that predictor X6 not included in the model to make prediction because of the high correlation it had with predictor X3, making the estimated coefficient of X6 unstable as well as unreliable (James et al. 2013, 99-101). Using the estimated coefficients in Figure 21, it is clear that the transaction date and the number of convenience stores in the living area had a positive impact on the house price. Indeed, according to Sinyi Realty, Global Property Guide (2019), there was an increase in house price in Taiwan from 2012 to 2013, which is shown in Figure 26.

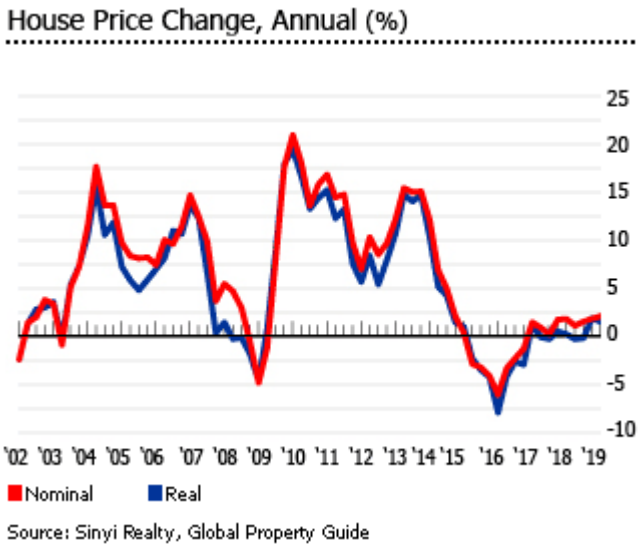

FIGURE 26: House price change (Sinyi Realty, Global Property Guide).

On the other hand, the house age as well as the distance from the nearest MRT had some negative relationship with the house price, whose estimated coefficients were -3.035×10^{-1} and -4.474×10^{-3} respectively. Since the estimated coefficients were quite small, it seemed that the effect of these predictors on the response were not significant.

Among random forests models, the ones that considered only 2 and 3 random variables at each split performed the best with the lowest test MSE. Through the variable permutation importance plots, it seemed that predictor X3 which is the distance to the nearest MRT played the most important part in making the predictions since permuting its value increased the MSE by approximately 30%. The same conclusion could be made from the multiple linear regression model, since in that model X3 had the smallest p value, indicating that it was the most significant predictor in the model. The latitude and the house age also had considerable importance in the random forests models. According to James et al. (2013, 25, 329), bagging is considered a statistical learning method with high flexibility and since bagging a special case of random forests, random forests also has high flexibility. It is widely acknowledged that in a flexible method, the model performs similar to a black box which is nearly unable to interpret and thus the interaction between the predictors and the response is not as straightforward as in a linear regression model (Plate 1999). Even when the importance of the variables is known, it is still unclear how changes in a predictor would affect the response.

Based on the test MSE, apparently random forests models performed better than linear regression models with the largest MSE of random forests is still significantly smaller than that of linear regression. Figure 27 and 28 visualize the observed values and the predicted values of the response using random forest model and linear regression model respectively. More specifically, the random forests model used considered 2 predictors at each split and the linear model used 5 predictors to make predictions. The x-axis shows the index number of the observation in the test data set and the y-axis shows the values of the response. It is visible that the predictions made by the random forests model matched the observed values better. The random forests method outperformed

the linear regression method in terms of predicting the future values of the response which is the house price based on the information of its transaction date, house age, distance from MRT, the number of convenience stores and geographic coordinates. However, the random forests method did not provide as much information about the interaction of the variables as the linear regression method. Nevertheless, the goal of this thesis was to build a model that maximized predictive power and the random forests model was able to satisfy this criterion at the cost of interpretability.

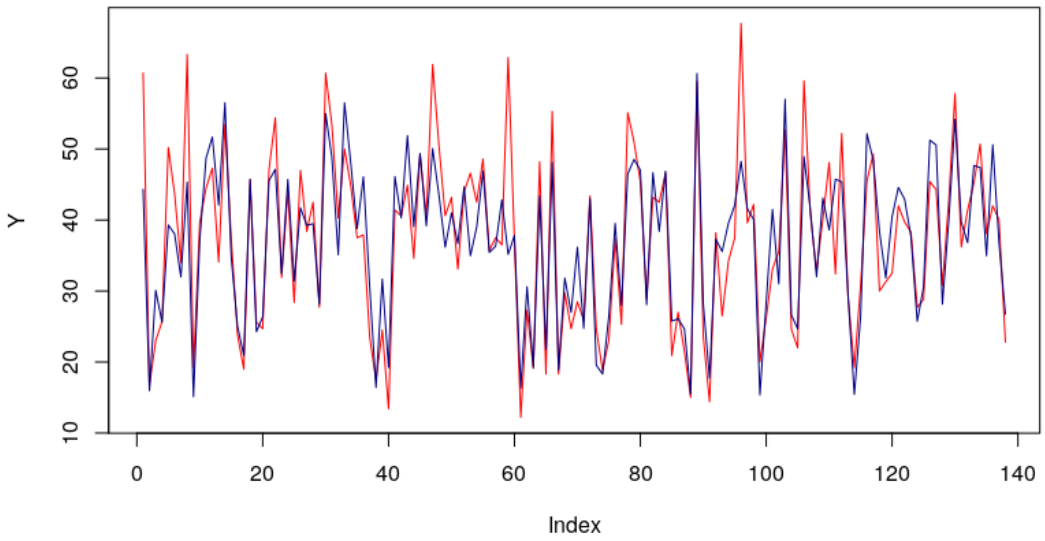

FIGURE 27. Observed values versus predicted values of the response made by random forests model. The red line is the observed values and the blue line is the predicted values.

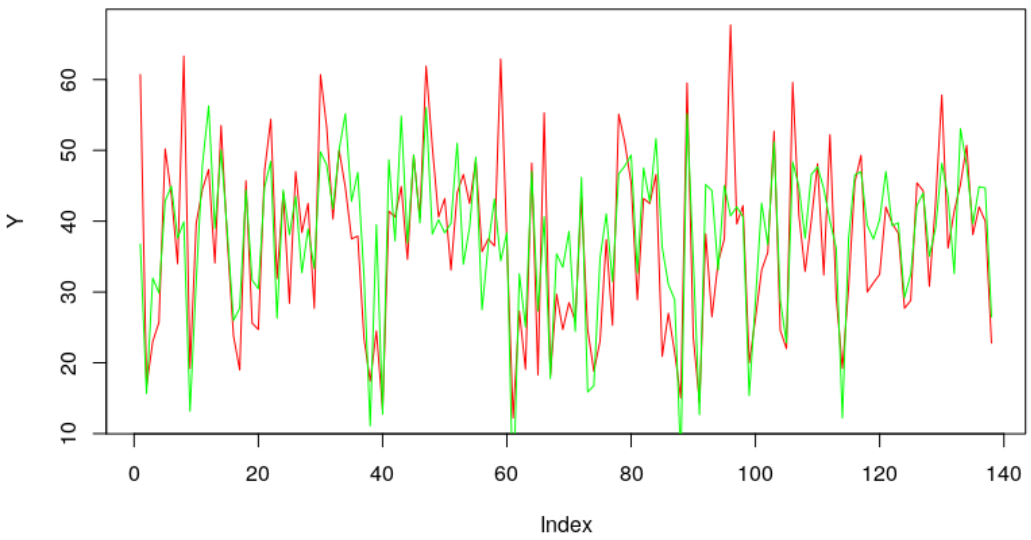

FIGURE 28. Observed values versus predicted values of the response made by linear regression model. The red line is the observed values and the green line is the predicted values.

6.2 Utilization & further development

This approach of estimating house price using statistical learning could be utilized in several ways. Admittedly, businesses such as real estate developers, real estate brokers, investors and banks rely a lot on the ability to predict the future housing market to make strategies and decisions. For examples, Park and Bae (2014) proposed machine learning algorithms for predicting house price that could offer mortgage lenders and financial institutions better appraisal, risk analysis and lending decisions whose advantages were believed to consist of analysis cost reduction as well as faster loan decisions. Construction companies could also use statistical models to estimate house price before a new construction to decide whether it should be built or not (Rafiei & Adeli 2016). Real estate economics researchers could use predictions of future house price to formulate theories, study or analyze the real estate market, for examples, the relationship between supply and demand (Koskinen 2019). The government regulators also require to be able to predict future price as well as analyze the effects of other factors on the price for several important tasks such as urban planning

and development, making laws that regulate the market or estimating taxes from real estate.

In order to develop and improve the performance as well as effect of the statistical learning methods in predicting house price, it could be suggested that a larger and more detailed data set be used. The data set could include more variables that describe the house more thoroughly, for examples, the number of rooms, house area, building type, number of stories or tax amount. Even though a larger data set would make the mining task more challenging, it would also be more rewarding. More sophisticated and advanced methods such as support vector machine or neural network would improve the predicting ability, however, these methods require a deeper theory understanding.

6.3 Conclusions

In this thesis, two statistical learning methods known as multiple linear regression and random forests were implemented on a publicly available real estate data set to build models that could predict the house price in Sindian District, Taiwan (Yeh & Hsu 2018). The theoretical framework was presented to serve as theoretical foundations for other parts of the thesis. First, it introduced and defined statistical learning as well as the multiple linear regression and random forests methods. This part also provided the theory of interpretability and flexibility trade-off which showed that linear regression was easier to interpret while random forests was more flexible (James et al. 2013, 25). The MSE was also explained and later used as the measure for performance. Subsequently, the methodology section was used to describe how this thesis was conducted as a data mining project by listing the steps and how the theory as well as techniques in the theoretical framework would be utilized. The data would be examined, then split into a training set to train the models and a test set to assess the models' performance using MSE. From there, it was shown that the two main steps for this project were data exploration and model building, whose results were reported respectively. The data exploration step provided better understanding of the data and could be used for model selection and interpretation

(Han et al. 2012, 39). The results from training and testing models showed that random forests model managed to predict house price with considerably higher accuracy than linear regression but also was almost impossible to interpret. Applications and further developments were also discussed.

REFERENCES

Blue, J. 2015. Driving Insights with Network Graphs. Released on 30.03.2015. Read on 27.10.2019. https://mapr.com/blog/driving-insights-network-graphs/

Bremer, M. 2012. Multiple Linear Regression. Read on 26.10.2019. http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf

Buza, K. Feedback Prediction for Blogs. Data Analysis, Machine Learning and Knowledge Discovery, 145-152.

Columbus, L. 2017. 53% of Companies are Adopting Big Data Analytics. Released on 24.12.2017. Read on 25.10.2019. https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#4b5d331e39a1

Drakos, G. 2019. Decision Tree Regressor explained in depth. Released on 23.05.2019. Read on 25.10.2019. https://gdcoder.com/decision-tree-regressor-explained-in-depth/

Delmendo, L. C. 2019. Taiwan's housing market – overvalued, but rising. Released on 08.08.2019. Read on 25.10.2019. https://www.globalpropertyguide.com/Asia/Taiwan/Price-History

Faraway, J. J. 2002. Practical Regression and Anova using R. Released on 07.2002. Read on 26.10.2019. https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Goldberger, A. S. 1991. A Course in Econometrics. Cambridge: Harvard University Press.

Han, J., Kamber, M. & Pei, J. 2012. Data Mining: Concepts and Techniques. 3rd edition. Burlington: Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd edition. New York: Springer Science+Business Media.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. 1st edition. New York: Springer Science+Business Media.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lutkepohl, H. & Lee, T. C. 1988. Introduction to the Theory and Practice of Econometrics. 2nd edition. New York: John Wiley & Sons.

Koskinen, J. 2019. Hedonic Models and Prediction Accuracy: Using Machine Learning to Predict House Prices. University of Helsinki. Department of Political and Economic Studies. Master's thesis.

Lane, D. M., Hebl, M., Guerra, R., Osherson, D. & Zimmer, H. n.d. Introduction to Statistics. http://onlinestatbook.com/Online_Statistics_Education.pdf

Mangasarian, O. L., Street, W. N. & Wolberg, W. H. 1995. Breast Cancer Diagnosis and Prognosis via Linear Programming. Operations Research 43 (4), 570-577.

Metzger, A., Leitner, P., Ivanovic, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S. & Pohl, K. 2015. Comparing and Combining Predictive Business Process Monitoring Techniques. IEEE Transactions on Systems, Man, and Cybernetics: Systems 45 (2), 276-290.

Park, B. & Bae, J. K. 2014. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. Expert Systems with Applications 42 (6), 2928-2934.

Plate, T. 1999. Accuracy versus Interpretability in Flexible Modeling: Implementing a Tradeoff Using Gaussian Process Models. Behaviormetrika 26 (1), 29-50.

Princeton University Library. 2007. Interpreting Regression Output. Read on 28.10.2019. https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

Rafiei, M. H. & Adeli, H. 2016. A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units. Journal of Construction Engineering and Management 142 (2).

Ruozzi, N. 2016. Variance Reduction and Ensemble Methods. Read on 25.10.2019. https://personal.utdallas.edu/~nrr150130/cs7301/2016fa/lects/Lecture_10_Ensemble.pdf

Salian, I. 2018. SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning? Released on 02.08.2019. Read on 25.10.2019. https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R. & Lichtendahl, K. C. Jr. 2018. Data Mining For Business Analytics: Concepts, Technique, and Application in R. 1st edition. Hoboken: John Wiley & Sons, Inc.

Sullivan, L. & LaMorte, W. W. 2016. Multivariable Methods. Modified on 31.05.2016. Read on 25.10.2019. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/

SIGKDD. 2018. Data Mining Curriculum: A Proposal. Read on 25.10.2019. https://www.kdd.org/curriculum/index.html

Walker, M. 2016. The Data Mining Process. Released on 09.01.2016. Read on 26.10.2019. https://www.datascienceassn.org/content/data-mining-process

Yeh, I. C., & Hsu, T. K. 2018. Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

Yen, L. 2019. An Introduction to the Bootstrap Method. Released on 26.01.2019. Read on 26.10.2019. https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60