STAT 3011 Project 1

Group 5

Table of Content

[Introduction 2](#_Toc36231341)

[Description of Data 2](#_Toc36231342)

[Motivation 1: Predicting the price of the house by past data 3](#_Toc36231343)

[Data Preprocessing 3](#_Toc36231344)

[Model selection 3](#_Toc36231345)

[A) Linear regression 4](#_Toc36231346)

[B) GBM 5](#_Toc36231347)

[Comparison with linear regression 7](#_Toc36231348)

[Inflation rate 7](#_Toc36231349)

[Advantages and Disadvantages of Different Models 8](#_Toc36231350)

[Limitation 8](#_Toc36231351)

[Motivation 2: Characteristics of Best-selling apartment 9](#_Toc36231352)

[Assumption and limitation 9](#_Toc36231353)

[Data preprocessing 9](#_Toc36231354)

[Model selection 10](#_Toc36231355)

[A) Random forest 10](#_Toc36231356)

[B) Support Vector Machine (SVM) 14](#_Toc36231357)

[Advantages and disadvantages of different models used 18](#_Toc36231358)

[C) Linear Discriminant Analysis (LDA) 19](#_Toc36231359)

[D) K-nearest neighbor (KNN) 19](#_Toc36231360)

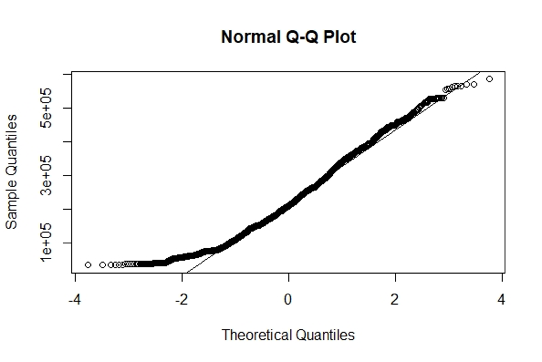
[Executive summary 20](#_Toc36231361)

# Introduction

Given the dataset of housing transaction records of an Asian city, our group has two directions to conduct the analysis. In the first direction, we are focusing the selling price of the apartment and we would like to predict the price by using the past data. Hence, we wish to calculate the housing inflation rate in that Asian city. For the second direction, we are investigating the properties of different level of apartments in terms of degree of popularity. Machine learning algorithms and regression models are both included in our methodology.

# **Description of Data**

1. The study is observational
2. No NAs is found in the dataset
3. There are 3 types of variables:
   1. Continuous: *Price, Size, Year\_Bulit, Year\_Sold, Month\_Sold*
   2. Discrete: *Floor, all “N\_” variables*
   3. Categorical: *TimeToSubway, TimeToBusStop*
4. The Price of housing is approximately normal with heavy tails



# Motivation 1: Predicting the price of the house by past data

We would like to predict the future house price by investigating the most valuable variables as well as interpret the inflation rate of the house price in the Asian city.

## Data Preprocessing

We changed the categorical variables (*TimeToSubway, TimeToBusStop*) into dummy variables. And we partitioned the data that 80% is training data, while 20% is test data.

## Model selection

We used two criteria to determine which model is chosen to present in our report. The first criteria are the prediction ability and second criteria are the interpretation ability. Thus, Principle component analysis will not be considered as it is hard for us to understand which the meaning of each component.

Please note that any type of Neural Network will not be considered as we do not have that much of the computation power. The below table shows the RMSE of each algorithm. All RMSE are based on 10-folds cross validation in order to receive a fair result.

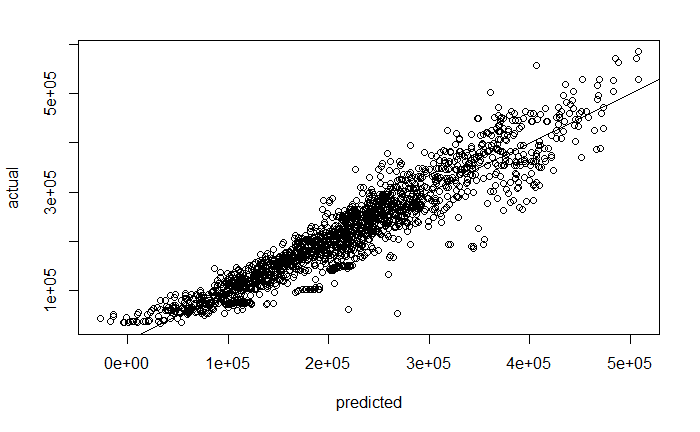
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Gradient Boosting Machine** | **Random forest** | **Linear**  **Regression**  **(with tuning)** | **Lasso Regression** | **Rigid Regression** |
| **RMSE** | 15181.34 | 15409.05 | 37026.77 | 38216.18 | 39211.74 |

From the result of table, we decided to use GBM to do the prediction *since we are sta*nding in house agent’s point of view use linear regression to support GBM in terms of explanation.

## A) Linear regression

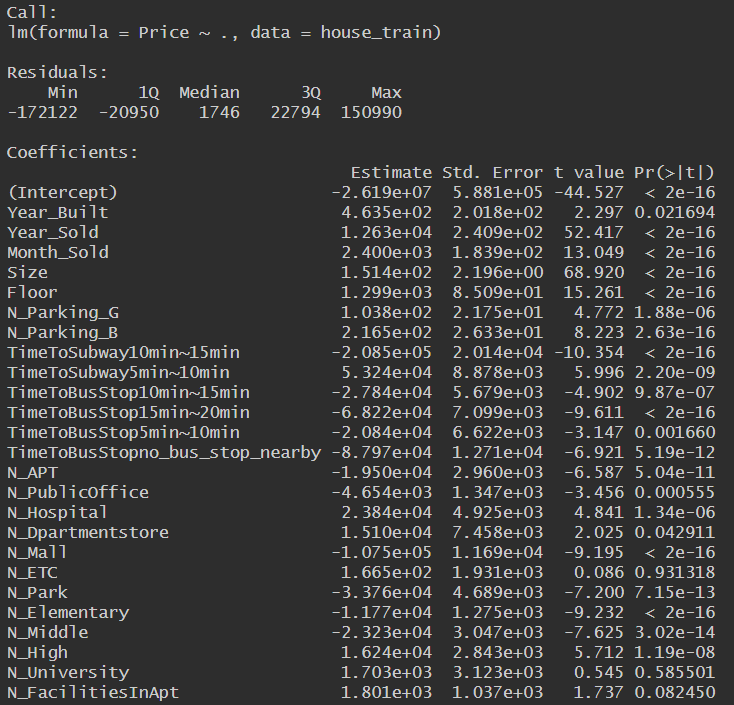
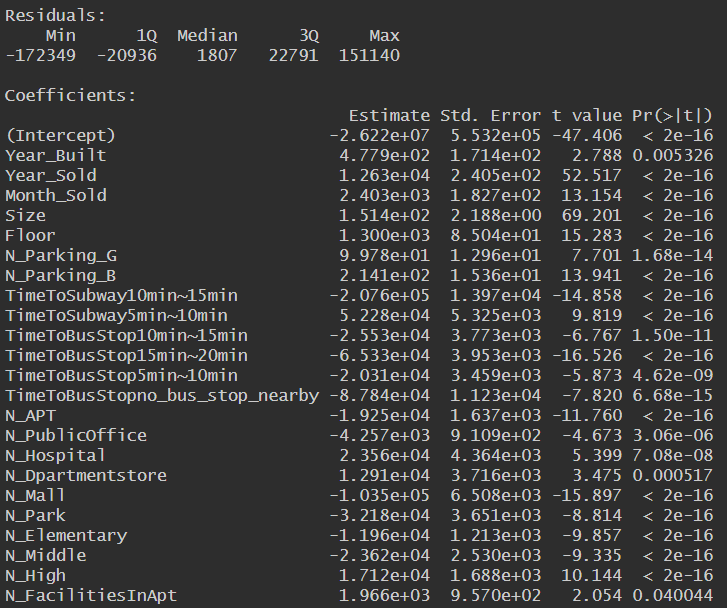
**Model parameters**

A linear regression is applied to see the relationship between price and other variables. The result is as follow,



The model is already fitted quite well, but we tried to drop some variables to prevent overfitting and to see any variable are being unimportant. Adjusted R2 = 0.8767

**Variables selection**

AIC is introduced in order to drop the insignificant variables; we then choose the model with the least AIC:

As we can see, few variables have been dropped and we receive a ‘lighter’ model. It is obvious that both time factors are still contributing to our mode. Adjusted R2 still retain 0.8767.

Although linear regression is good enough to provide the intuition of how the variables contribute to the model, linear regression is considered as a primary step to explore the dataset which drives us to explore in another model. After exploration we figured out that Gradient Boosting Machine (GBM) performs the best.

## B) GBM

Similar to random forest, Gradient Boosting Machine (GBM) is also an ensemble learning algorithm. However, rather than using bagging to improve the model, GMB use boosting to improve the model.

Its idea is to train multiple of weak models and finally combine them together to output a strong model. The base learner of GBM is still decision tree but rather than using a part of data, GBM use the entire set of data. It first trains a decision tree, then fit the data into the decision tree and find the RMSE of that particular decision tree. And then train another decision tree to find the characteristics from the RMSE. Repeat this process for n times and add those n trees up, I.e. sum of the decision trees.

Mathematically, we can represent it as:

Fit a tree:

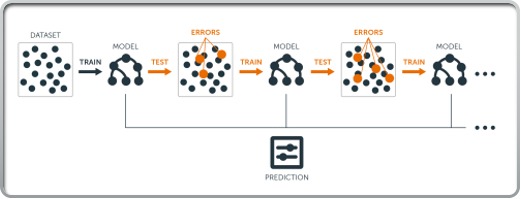
Train another tree model prior to the previous residual*s*

Add a new tree

Train another tree model prior to the previous residual*s*

Add a new tree

Repeat the above algorithm until cross validation tells it to stop.



**Tuning**

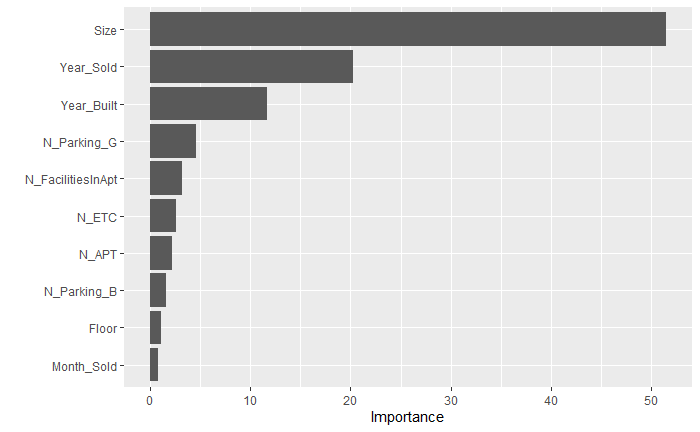
In order to increase the efficiency, we used grid search to do parameter tuning. The following criteria are the learning rate, number of terminal nodes, minimum number of observations of each terminal node. I.e. min (n of min nodes, n of nodes), percentage of data apply to stochastic gradient descent (SGD). And the optimal model is a model with 0.01 learning rate, 5 terminal nodes at the end, minimum of 5 observations of each terminal nodes, 80 percentage of the data applying SGD method with 50000 optimal trees.

**Optimal Solution**

In this model we are using stochastic gradient descent (SGD). Every time GBM picks a mini-batch sample and then update the tree. The advantage of it is to avoid only finding the local minimum instead of the global minimum of the loss function.

**Model explanation**

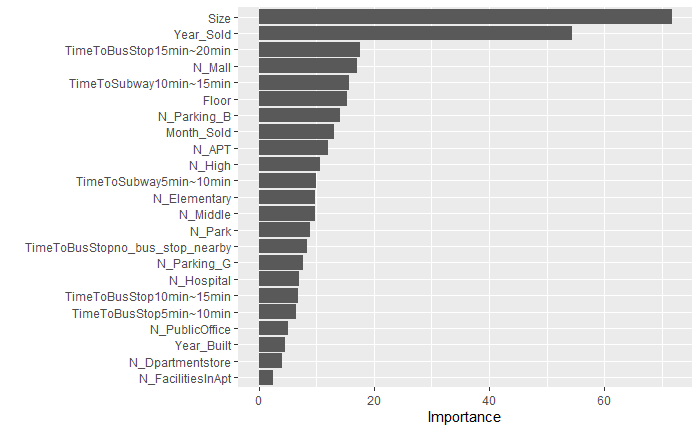
Similar as random forest, GBM is known as a black-box algorithm. We cannot explain anything from it and since it is a weak-learner algorithm, we cannot have the same approach with random forest to explain the flow of the model. However, we can still figure out which variables contributes the most to our model. The plot below is the variable importance plot indicating the importance of each variables.



As we can see, Size, Year Sold and Year Built are the 3 most contributing factors to the model.

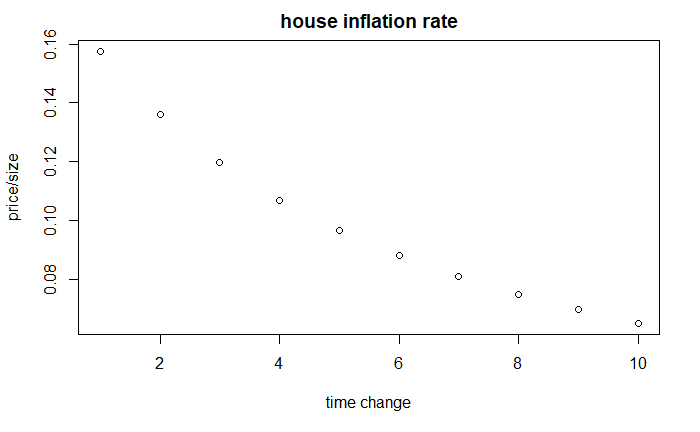
## Comparison with linear regression

Now let us compare with linear regression. The plot below is the variable importance plot indicating the importance of each variables of linear regression model.



As we can see, Size, Year Sold and Time to bus stop of 15mins to 20 mins are the 3 most contributing factors to the model. Which somehow agrees with the GBM model. It also indicates us we may use the linear regression model as a reference of GBM model to gain a better understanding.

## Inflation rate



We have further investigated on the inflation rate on house price and the above figure shows that the price of the house increases in a smaller rate than before, thus we suspect that the country has a *disinflation* inhouse price and it has the trend to become deflation in the near future.

## Advantages and Disadvantages of Different Models

**Advantages of Linear Regression**

There are several advantages bring out by linear regression method such as it is intuitive and even a layman can easily understand. Besides, it is easy to visualize and has a decent accuracy if the parameter tuning methods used in a right way.

**Disadvantages of Linear Regression**

However, linear regression required some assumptions before-hand. For example, it assumes the data are independent, the relationship between the explanatory variable and the response variable is linear and sensitive to outliers in the data.

**Advantages of GBM**

Based on the discussion of GBM, it is known for its high accuracy in terms of prediction as well as applicable on different loss functions. Moreover, various methods such as SGD and cross validation can be applied on the model to find the optimal solution. Other advantages are not requiring data preprocessing such as scaling and able to deal with missing values.

**Disadvantages of GBM**

On the other hand, disadvantages of GBM can also be found, for instance it will be overfitted easily if we do not control the learning rate properly and it is costly in terms of computation as a good computer is vital. Also, it is too flexible that sometimes grid search has to applied which requires more computation power. Last but not least, GBM is treated as a black-box model. Another supporting method are needed to visualize the model better.

## Limitation

As the number of observations per year sold is relatively small and fluctuates every year. For example, 2007, 2008, 2014 have number of observations below 500 (merely 104, 390, 456 observations respectively). Therefore, the variance of the price could be large. We believe our analysis could be more predictive if more data is given.

# Motivation 2: Characteristics of Best-selling apartment

We would like to seek the relationship between the year of transaction and other variables. More exactly, we wish to investigate the features of the best-selling, intermediate, and worst-selling apartments.

## Assumption and limitation

In the study, apartments are classified into three types in term of popularity in the dataset: Best-selling apartments (year of transaction is below 4 years), Intermediate apartments (year of transaction is below 10 years and above 4 years), Worst-selling apartments (year of transaction is above 10 years).

For the portion of these three types is about 1:1:1 ratio (2094:1886:1898).

Moreover, in the research, we assume all houses included in the dataset are first transaction, which means no second-hand apartment exist in the dataset, house with same characteristics will be regarded as distinct cases.

The reasons of the assumption made are: First, it is hard to define a second-hand apartment, due to the possibility of same room type was designed massively, as a result it is not rigorous to determine whether a house is first-hand just base on its years of transaction. Secondly, buying intention of first-hand and second-hand apartment is disparate, it will affect the fairness of classification if biased observations were included. As a result, assumptions limit the research to be a classification problem that classification of variables will be considered as the most important purpose.

## Data preprocessing

We factorized the categorical variables and use the same partition ratio in Motivation 1.

Moreover, we transform the *Year\_Built*, *Year\_Sold* into a new variable *Year\_Trans.*

*Year\_Trans = Year\_Sold – Year\_Built*

While *Month\_Sold* is removed as there is no information about the month of build to be compared with and the effect of removing should not be considerable.

## Model selection

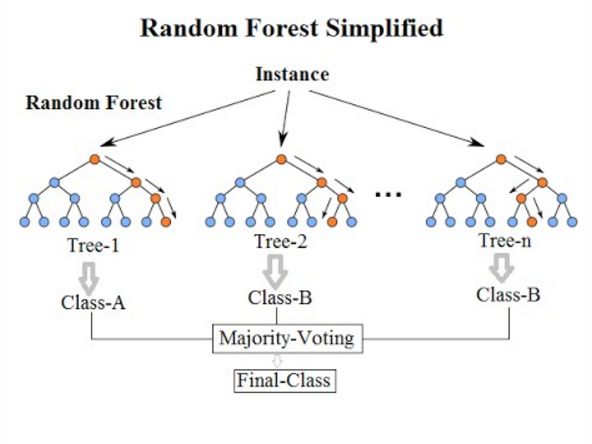
Similar with the previous part, we choose the models in terms of power of explanation and prediction ability. Below shows the accuracy of each model.

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy Table |  |  |  |
| Random Forest | SVM | LDA | KNN |
| 0.9132161 | 0.90935 | 0.8228 | 0.8993129\* |

Finally, we chose random forest and SVM as they produce the highest accuracy result.

Please note the reasons of not using PCA and neural network are same as the previous part. However, we will not consider using logistic regression for several reasons. The first reason is logistic regression usually not performing well in terms of more than 2 variables. And the second reason is tuning the decision boundary of logistic regression maybe to a severe overfitting problem. As a result, logistic regression will not be done, we will rather use decision tree to explain our result.

## Random forest

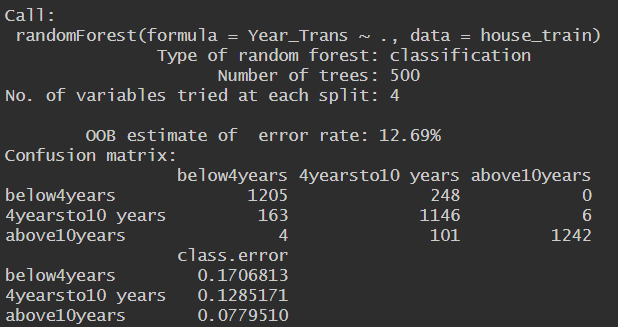


Random forest is a method that constructing multiple decision trees during the training phrase, I.e. an ensemble learning method. We first randomly select n subset from our training dataset, then we train exactly n decision trees. Each individual tree then predicts the result by the test data from the training data. Finally, a prediction is made by majority voting.

The random forest classifiers split the whole dataset and gather the data that share the same characteristics.

Since random forest uses multiple trees, it can reduce the risk of overfitting and the training time is less. Moreover, it runs better on small amount of data (compare to neural network) and produce highly accurate predictions.

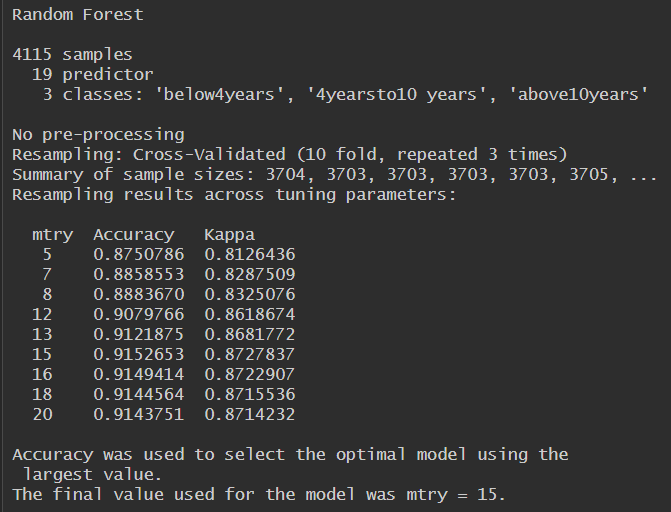
**Initial model**



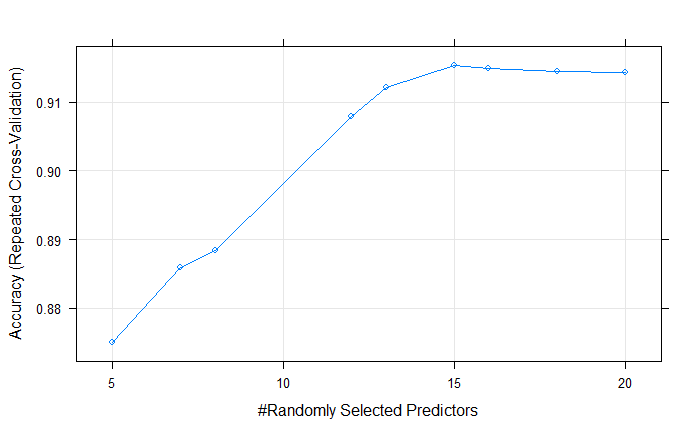
Accuracy = 0.8684061

**Parameter tuning**

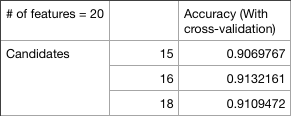
As we can see, random forest seems doing quite well already without any parameter tuning, now let’s improve the model and see how far random forest can reach.



Here is the model with 10-folds cross validations and the result shows that when number of features are greater, the better the prediction result.



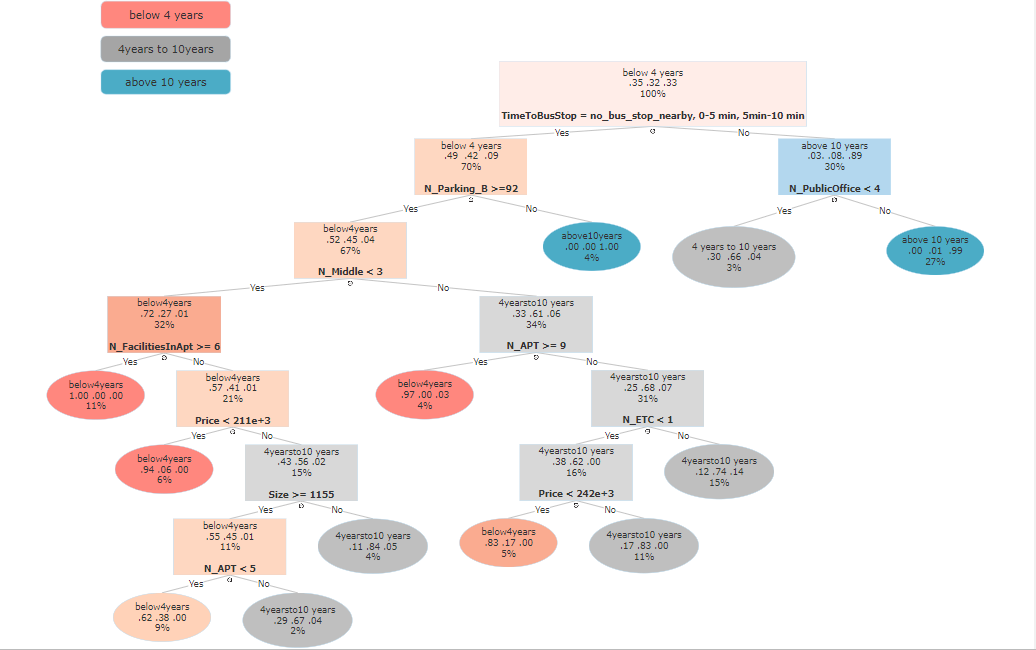
The confusion matrix shows the best performance among different numbers of predictors which means in each split, random forest with 15, 16 and 18 independent variables on each tree make the models perform the best. Three parameters will be tested since number of features = 15, 16 and 18 give a very similar accuracy of models.



**Model explanation**

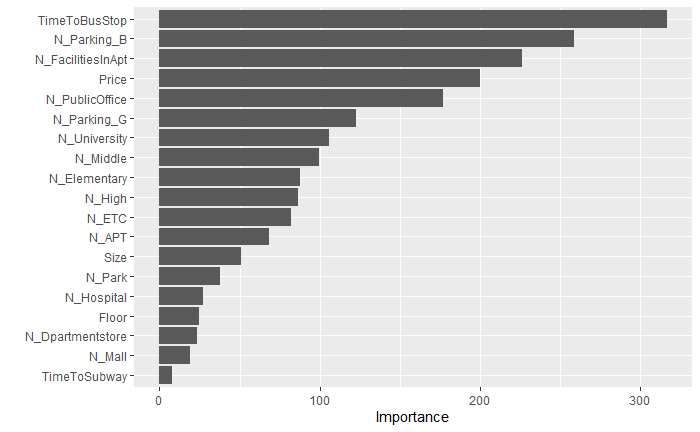
Since random forest constructs many classification trees (in this case the tree size is 500), it can consider as a black-box model and we cannot really visualize it. Alternatively, we can pick one of the classification trees and see the flow of the classification tree. This classification tree contains all variables in order to see the complete flow of the random forest.

**Classification Tree**



**Variable Importance Plot**

The figure below shows how the variables weight in our model.



## B) Support Vector Machine (SVM)

A support vector machine (SVM) is a classification method. Especially, it is a binary classifier. It works on the principle of fitting a boundary to “best” separate two classes.

To obtain the “best” boundary, the key is to find the boundary that can maximize the gap between two classes. The rationale behind this is that the larger the gap, the fewer chance the new points would fall in the wrong class region.

Literally, the “support vector” in an SVM means a core set of points that can help identify and fix the boundary, that is, these data points “support” the boundary. One of the interesting characteristics is that once the boundary is established nicely, data points that are not belong in the “support vector” become unessential. Figure (4b.1) can clearly illustrate that.

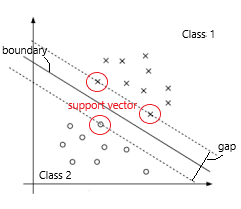
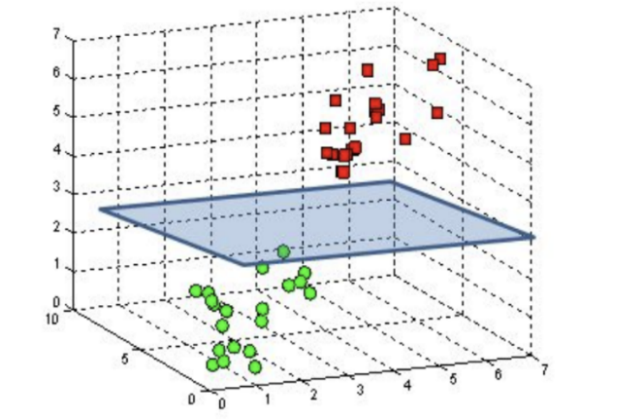
 

Figure 4b.1 Figure 4b.2

The boundary is traditionally called a hyperplane. In a 2 dimensions (attributes) case, hyperplane is a straight line or a curve. In a 3 dimensions case, it is a plane or an irregular complex surface (Figure 4b.2). Higher dimensions are apparently impossible to visualize and a hyperplane is thus a generic name for a boundary in more than three dimensions.

***Binarizing approach***

As SVM is binary classifier, we have two basic approaches to classify data into binary manner: Assume there are n classes.

· One-versus-One scheme is regarding to select a pair of classes from a set of n classes and do binary classification for each pair. There will be n(n+1)/2 pairs in total

· One-versus-All scheme (or called One-against-Rest) means that we select one class to be “positive”, then all the rest would take “negative” aspect. There will be n pairs in total

Different schemes have its own benefits and drawbacks. In our implementation, we chose One-versus-All as we would like to emphasize on the features of each kind of apartments against all other apartments, and considering One-versus-All scheme is more memory and time efficient while the loss is almost as same as One-versus-One scheme, so we chose it.

***Tuning***

* **Class probability plug in**

As classical SVMs merely output which side of boundary that the new observations fall in (predicted class labels). Consequently, they do not automatically output the predicted class probabilities. However, predicted class labels give little information for the later interpretation, so our SVMs have been tuned to be capable in producing predict class probabilities.

* **Kernel function**

Mainly, there are three types of kernel functions that are popular in constructing SVMs:

Linear kernel, Polynomial kernel, Radial basis function (RBF) kernel.

In general, RBF kernel would be the prior choice as it can manage the situation which the relation between class labels and attributes is non-linear. And it has less hyperparameters, and hence less complex compared with polynomial kernel.

* **Hyperparameters**

There are 2 hyperparameter in the SVM with radial basis kernel --- Cost(C) and Sigma

With the 10-fold cross validation, the final values used for the model were sigma = 0.03371159

and C = 128.

***Optimal solution***

We used One-versus-All SVM with radial basis kernel as our final model. The hyper-parameters sigma and cost equal 0.03371159 and 128 respectively.

***Model Result and Interpretation***

Year\_Trans ≤ 4 vs the rest: acc=0.92548

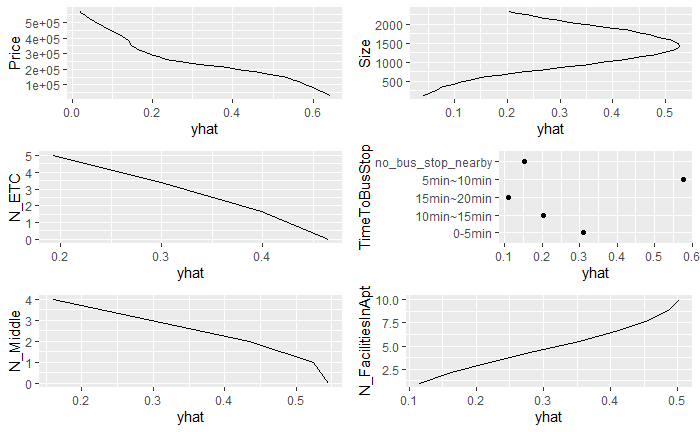
4 < Year\_Trans ≤ 10 vs the rest: acc = 0.89061

Year\_Trans > 10 vs the rest: acc = 0.91196

**Overall average = 0. 90935**

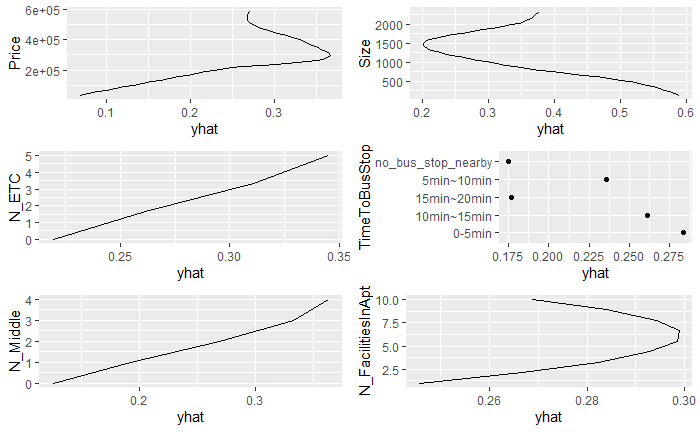
The figures below show the **most 6 important variables** in the SVM model.

*“***Best-selling**” ( *Year\_Trans 4*) vs the rest



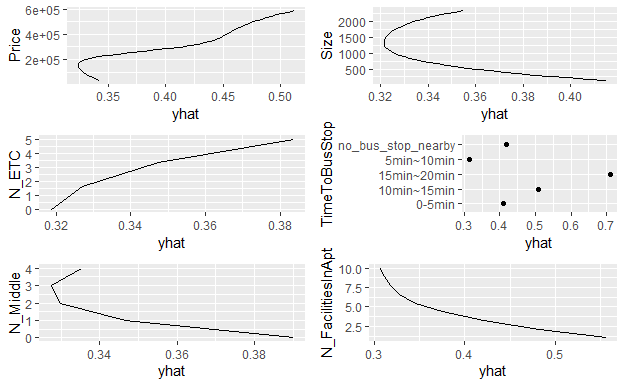
The **“Best-selling”** apartment tends to be lower in price. While in size aspect, apartment with 1500sqft. is the most appealing. The time taken from the apartment to bus stop is also as minimum as possible. Lower number of middle schools and the other facilities such as hotels and special school nearby the apartment is preferred. Finally, it tends to have more in-apartment facilities for residential recreation.

*“***Intermediate**” ( 4 < *Year\_Trans < 10*) vs the rest



The **“intermediate”** apartment tends to have a moderate price. The size of apartment is likely to be either lower than 750sqft. or higher than 2000sqft.The travelling time to bus stop is relatively short. The number of other facilities around the apartment and middle school nearby is likely to be higher. While the number of in-apartment facilities for resident also tends to be in average.

**“Worst-selling”** ( Year\_Trans ≥ 10) vs the rest



The **“Worst-selling”** apartment is basically the opposite of “Best-Selling” apartment, except the number of Middle School nearby. Both results suggest the number of Middle School nearby the apartments is as smaller as possible.

## Advantages and disadvantages of different models used

**Advantages of random forest**

First of all, the greatest advantage of random forest is the reduction of overfitting problem in decision trees as well as the variance by basing on bagging algorithm and using Ensemble Learning technique. The way it works can create as many trees on the subset of the data and combines the output of all the trees. Thus, the accuracy is improved. Moreover, the performance of random forest will not be affected by non-linear parameters unlike curve- based algorithms do. Therefore, it shows the high efficiency when dealing with high non-linearity between the independent variables. Last but not least, random forest can be used to solve both classifications and regression problems, works well with both categorical and continuous variables and handle missing values automatically.

**Disadvantages of random forest**

Random forest is relatively more complex than the case of one decision tree as it combines the outputs of many trees created during the computation and make decision on the majority of votes. It means that random forest requires more computational power and resources than others algorithm. Therefore, a longer training period is required.

**Advantages of Support Vector Machines**

SVM can easily handle high dimensional data and works well with even unstructured and semi-structured data. Due to its kernel trick, we can solve high dimensional or complex data with ease. After the addition of the class probabilities plug in, the result would be highly interpretable. Compared to other supervised learning models and neural networks models, it is relatively memory efficient.

**Disadvantages of Support Vector Machines**

Choosing a good kernel for SVM is not easy and the hyperparameters of it is hard to fine tune. Class probabilities plug in is a must for interpreting the result of SVM, otherwise, SVM would directly output the predicted class thus the interpretation and understanding of the final model would be difficult. If the dataset is large, the training time for the SVM (especially under one-versus-one scheme)

*(Because the random forest and SVM can interpret the result well and have relatively high accuracy. Therefore, we would mainly focus on these two models. For LDA and KNN, only brief introduction would be included in the content.)*

## C) Linear Discriminant Analysis (LDA)

LDA applies dimension reduction techniques to reduce the number of dimensions in a dataset without losing too much information. It attempts to express one dependent variable as a linear combination of other variables. We can classify the data into several groups according to their features (i.e. variables). After defining the LDA model we can try to predict which group does a new datum belongs to.

**Advantages of LDA**

Firstly, the result is easy to be interpreted. Many people would like to start the analysis by using LDA, as a motivation to other model types. Also, the computation for data-training is fast.

**Disadvantages of LDA**

Several assumptions are needed for the model. For instance, the distribution is assumed to be a normal distribution. Also, the relationship between the explanatory variable and the response variable is assumed to be linear.

## D) K-nearest neighbor (KNN)

KNN is a non-parametric and supervised algorithm for classification and regression, where the result of new instance query is classified based on majority of K-nearest neighbor category. The purpose of the method is to classify a new object based on attributes and training samples. Moreover, this algorithm captures the idea of similarity or closeness with calculating the distance between data points, hence find the nearest K neighbors of the prediction and their labels. Finally, we can classify the prediction into majority labels of the k nearest neighbor.

**Advantages of KNN**

KNN is a simple algorithm to explain, understand and interpret the data with relatively high accuracy, also it does not require any assumption on the dataset.

**Disadvantages of KNN**

However, KNN is not competitive with better supervised learning models, since it is computationally expensive as the algorithm stores all of the training data, so it requires high memory. Moreover, the prediction stage might be slow with big N as well as the algorithm is sensitive to irrelevant features and the scale of the data.

# Executive summary

* **Description of data**

The data is observational, NAs-free with 3 types of variables. The “Price” in the dataset is approximately normal and yet with heavy tail.

* **Motivation1--- Predicting the price of the house by past data**
  + **Data Preprocessing**

Changed categorical variables into dummy variables. 80:20 training-testing data ratio.

* + **Model selection**
    - **Linear regression, GBM**
  + **Major Findings**

Size and Year Sold are the two most importance variables in the dataset.

The inflation rate of house price in this Asian city is deflating.

An apartment with a lower price would be more attractive.

* + **Limitation:** observations per year is relatively small, may result in high variance.
* **Motivation2--- Characteristics of Best-selling apartment**
  + **Assumption**

Assume all transactions were first-hand, all observations are divided into 3 groups in terms of popularity. (Best-selling | intermediate | worst-selling)

* + **Data Preprocessing**

Changed categorical variables into factor variables. 80:20 training-testing data ratio. (new variable) *Year\_Trans = Year\_Sold – Year\_Built*. Month\_Sold is neglected.

* + **Model selection:**
    - **Random forest, Support Vector Machines**
  + **Major findings**

Both models agree that the best-selling apartment tends to have the following major characteristics:

Short travelling time to bus stop, low selling Price, a greater number of in-apartment facilities for residential recreation, a smaller number of middle schools nearby, etc.