Data mining for Business Analytics

PROJECT REPORT

Melbourne Housing Market

Prepare by:

Đỗ Anh Luyện

Hanoi, 2021

TABLE OF CONTENTS

Contents

I. Introduction.

In the real world, data mining is widely used in many fields, it helps the company have a deeper insight knowledge into the market, make more accurate decisions that bring more profit. In this project, we will use our knowledge in data mining to solve the business problem in Melbourne housing market.

Business problem:

The main business problem is to help real estate companies most accurately predict house prices so that they can maximize the revenue from the sale of houses in Melbourne.

After answering this question, companies will have methods to increase sales and profits from house sales.

Dataset:

We will use Melbourne Housing Market dataset to handle this problem. This data includes 20 input variables and 1 output variable. From the data, we can do analysis to have the insight into Melbourne housing market and predict house prices in here.

Independent variables:

Suburb: Suburb
Address: Address
Rooms: Number of rooms
Price: Price in Australian dollars
Method:
S - property sold;
SP - property sold prior;
PI - property passed in;
PN - sold prior not disclosed;
SN - sold not disclosed;
NB - no bid;
VB - vendor bid;
W - withdrawn prior to auction;
SA - sold after auction;
SS - sold after auction price not disclosed.
Type:
h - house,cottage,villa, semi,terrace;
u - unit, duplex;
t - townhouse;
SellerG: Real Estate Agent

Date: Date sold
Distance: Distance from Central Business District in Kilometres
Regionname: General Region (West, North West, North, North east …etc)
Propertycount: Number of properties that exist in the suburb.
Bedroom2: Scraped # of Bedrooms (from different source)
Bathroom: Number of Bathrooms
Car: Number of carspots
Landsize: Land Size in square meters
BuildingArea: Building Size in square meters
YearBuilt: Year the house was built
CouncilArea: Governing council for the area
Lattitude: Self explanatory
Longtitude: Self explanatory
Postcode: post code number.

Dependent variable:

Price: Price in Australian dollars

Some questions to ask using the dataset:

1. In overall, house prices in Melbourne tend to increase or decrease in the period 2016-2018 and in the future? Should the company promote the sale of the house at this time (2018)?
2. What is the trend that most customers will favor houses with characteristics?
3. In overall, what factors affect house prices? (The relevance of each independent variable for the prediction of the dependent variable).
4. In which region is the house price per square meter high or low? and explain why it has this price? Advantages and disadvantages if customers buy house in this region?
5. The most accurate prediction of house prices based on given data.

Method in used:

For analysis, we use both descriptive statistic and inference statistic to have the insight of the data. In preprocessing-data, we will do some works such as handling missing values, adding or deleting columns, encoding categorical variables, splitting train and test set and scaling them. For building model, there are four models in used linear regression, K-nearest neighbor, decision tree and random forest. For fine-tuning, Cross-validation, Gridseachcv and RandomizedSearchCV were in used. In evaluation, because the output is continuous variable so using coefficient of determination (R^2) assesses how strong the linear relationship is between the model and the dependent variable, mean squared error (MSE) that is the average squared difference between the estimated values and the actual value, mean absolute error (MAE) is the absolute value of the difference between the forecasted value and the actual value. In visualization, we use bar chart, line chart, scatter plot, mix of bar and line chart, heatmap, boxplot, table.

II. Exploratory Data Analysis.

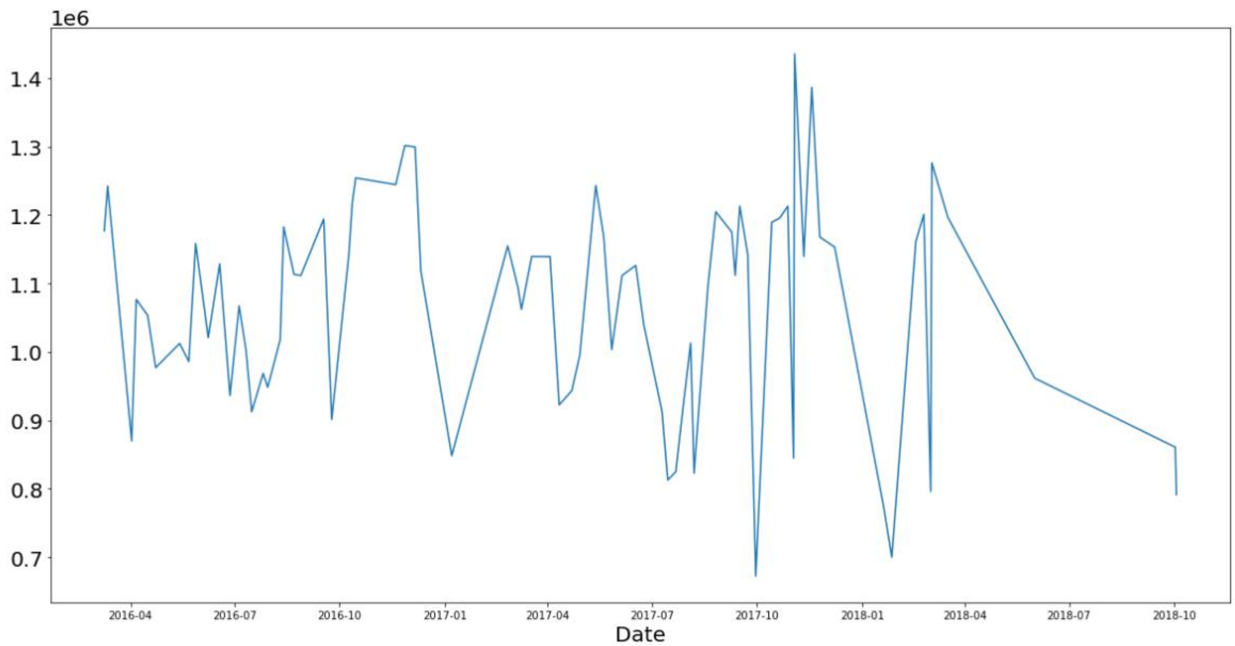
In this part, we will use both descriptive statistic and inference statistic technique to understand the insight of data, combined with visualization to answer four questions mentioned above.

First of all, we determine datatype of each feature and convert all ‘object’ type to ‘category’ type and Date to ‘datetime’ type for easy to manipulate.

```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         9863 non-null   object
 1   Address        9863 non-null   object
 2   Rooms          9853 non-null   float64
 3   Type           9838 non-null   object
 4   Price          9672 non-null   float64
 5   Method         9863 non-null   object
 6   SellerG        9863 non-null   object
 7   Date           9863 non-null   object
 8   Distance       9863 non-null   float64
 9   Postcode       9863 non-null   int64
 10  Bedroom2       9845 non-null   float64
 11  Bathroom       9850 non-null   float64
 12  Car            9839 non-null   float64
 13  Landsize       9776 non-null   float64
 14  BuildingArea   9765 non-null   float64
 15  YearBuilt      9839 non-null   float64
 16  CouncilArea    9863 non-null   object
 17  Lattitude      9863 non-null   float64
 18  Longtitude     9863 non-null   float64
 19  Regionname     9863 non-null   object
 20  Propertycount  9863 non-null   int64
```

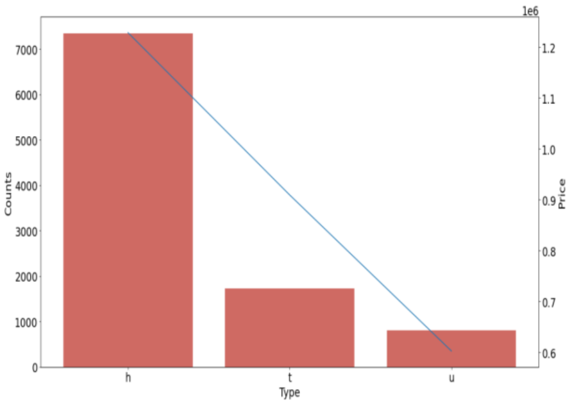
```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         9863 non-null   category
 1   Address        9863 non-null   category
 2   Rooms          9863 non-null   float64
 3   Type           9863 non-null   category
 4   Price          9672 non-null   float64
 5   Method         9863 non-null   category
 6   SellerG        9863 non-null   category
 7   Date           9863 non-null   datetime64[ns]
 8   Distance       9863 non-null   float64
 9   Postcode       9863 non-null   int64
 10  Bedroom2       9863 non-null   float64
 11  Bathroom       9863 non-null   float64
 12  Car            9863 non-null   float64
 13  Landsize       9776 non-null   float64
 14  BuildingArea   9765 non-null   float64
 15  YearBuilt      9839 non-null   category
 16  CouncilArea    9863 non-null   category
 17  Lattitude      9863 non-null   float64
 18  Longtitude     9863 non-null   float64
 19  Regionname     9863 non-null   category
 20  Propertycount  9863 non-null   int64
```

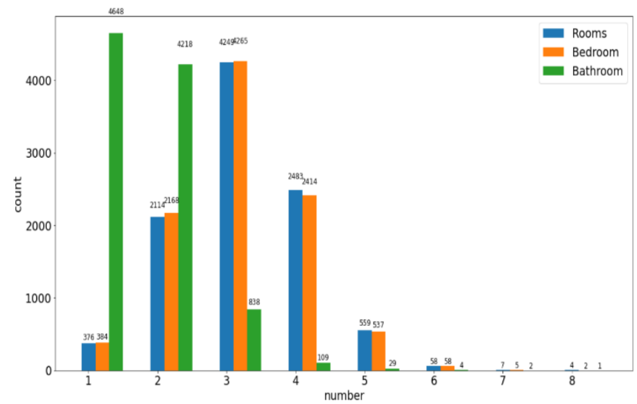
In the first question, we need to get a comprehensive view of house prices from 2016 to 2018.



From the chart, the house price fluctuated significantly from April-2016 to October-2018, by the eye, we can see that there is strongly decrease in housing price from April-2018 to October-2018 and it doesn't seem to be showing any signs of stopping. To understand why is there this trend, we need to understand the economic situation in Australia at that time, since 2015, Australian authorities have tightened controls on risky lending by banks, in the tightened lending conditions make it difficult for buyers and especially investors to borrow money to buy a home, low interest rate from the bank. By the business knowledge, when the price decreased quickly like this, the government will offer incentives to help growth again in next quarter (2-3 months), make the demand curve shift to the right, both quantity and price will increase. The advice is that real estate company should start promoting home sales in the next quarter, at this time house price will increase quickly.

In question two:





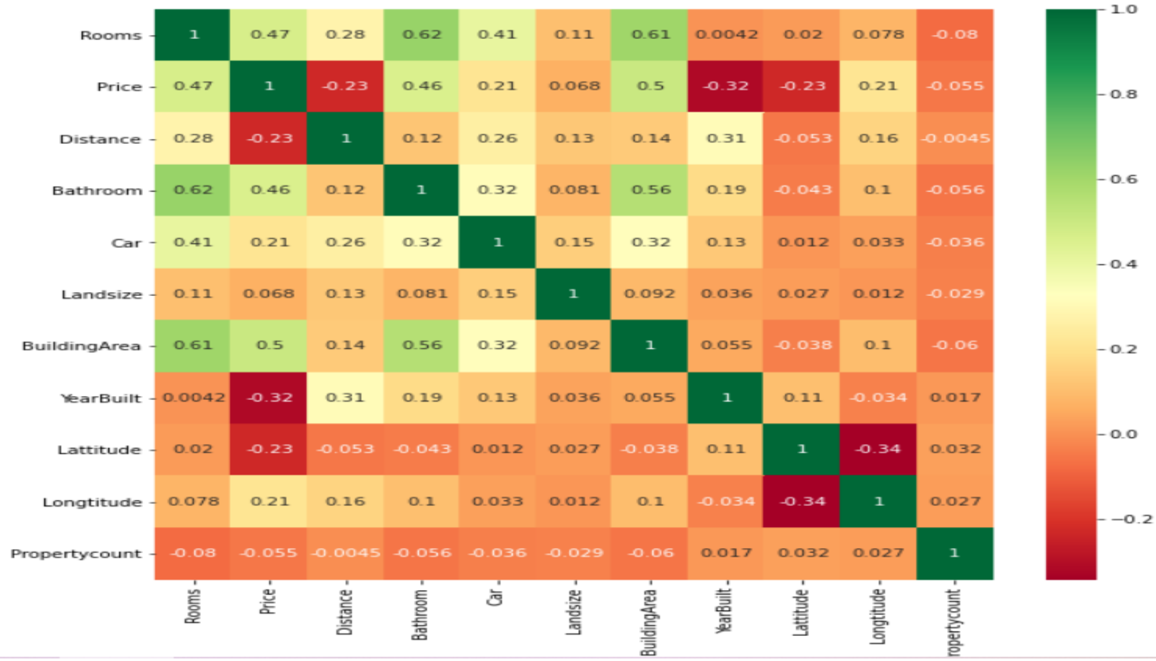
	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	L
count	9853.000000	9.672000e+03	9863.000000	9863.000000	9845.000000	9850.000000	9839.000000	9776.000000	9765.000000	9839.000000	9863.000000	98
mean	3.099462	1.091555e+06	11.199169	3111.513333	3.079126	1.648122	1.693465	517.106894	149.489755	1965.957008	-37.804699	1
std	0.963847	6.822714e+05	6.774656	111.076502	0.966598	0.723985	0.972455	929.834824	86.975142	36.846954	0.090368	
min	1.000000	1.310000e+05	0.000000	3000.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1196.000000	-38.184150	1
25%	2.000000	6.400000e+05	6.400000	3044.000000	2.000000	1.000000	1.000000	212.000000	100.000000	1948.000000	-37.859110	1
50%	3.000000	8.950000e+05	10.300000	3084.000000	3.000000	2.000000	2.000000	479.000000	132.000000	1970.000000	-37.799020	1
75%	4.000000	1.345000e+06	13.900000	3150.000000	4.000000	2.000000	2.000000	653.000000	180.000000	2000.000000	-37.748665	1
max	12.000000	9.000000e+06	47.400000	3977.000000	12.000000	9.000000	10.000000	40469.000000	3112.000000	2019.000000	-37.407200	1

Type of house is mostly house, cottage, villa, semi, terrace, they also have the highest average prices. Houses with 1 or 2 bathrooms, 3 rooms and bedrooms, 2 car parking spaces with building area from 132 to 180 square meters, year built from 1970 to 2000 are the most common.

However, we see that in summary of room and bed room, it’s quite similar so we will drop 1 of those 2 columns later.

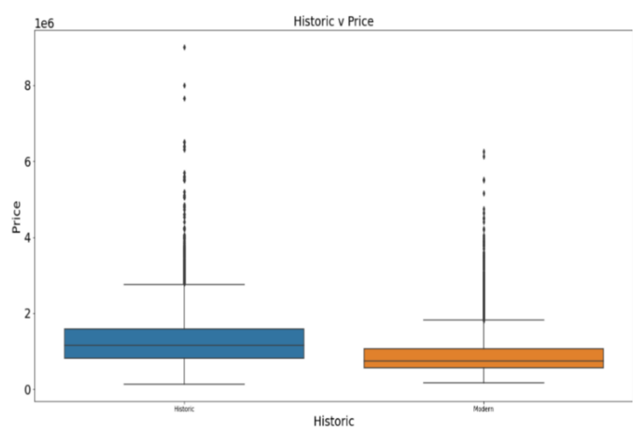
For easy to answer question 3 and 4, we create new feature namely ‘Historic’, this feature has two values ‘historic’ and ‘modern’, ‘historic’ means house ages older than 50 years and ‘modern’ means house ages less than 50 year. The second new column is ‘P/m2’, this is the house price per squared.

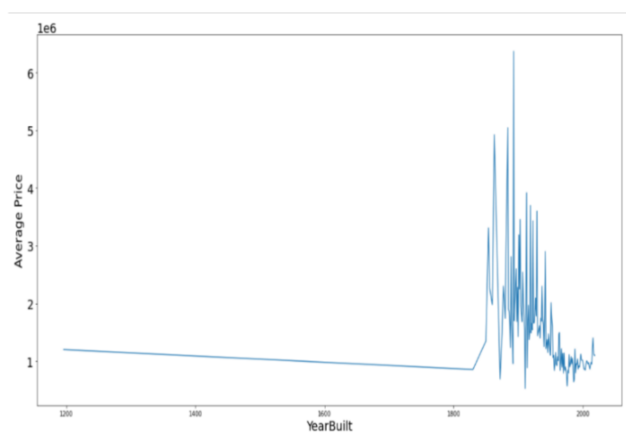
In question three, we will draw correlation plot to see what factors affect house prices.



Light to dark blue means positive correlation from little to more, light to dark red means negative correlation from little to more.

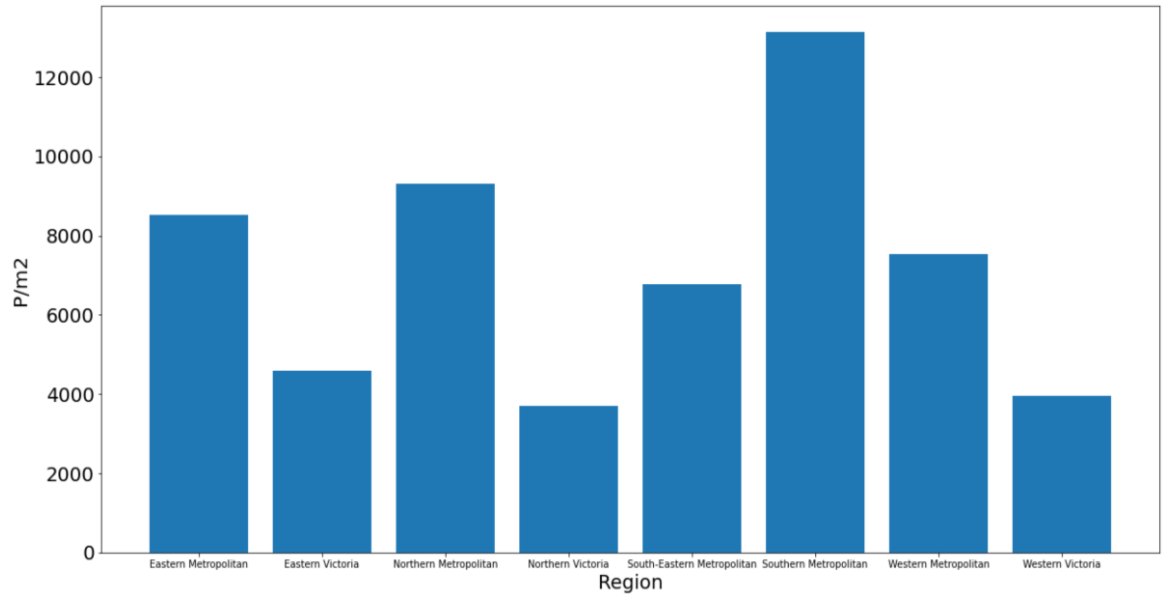
From the matrix, the number of rooms and bathrooms, building area have strongly positive correlation with dependent variable (Price) which means as the number of rooms and bathrooms, building area increase, house prices also increase. Distance to the central business district, year build and latitude have strongly negative correlation with dependent variable which means as the distance to the central business district, year build and latitude increase, the house price will decrease. Other features have not much effect to the dependent variable (Price).





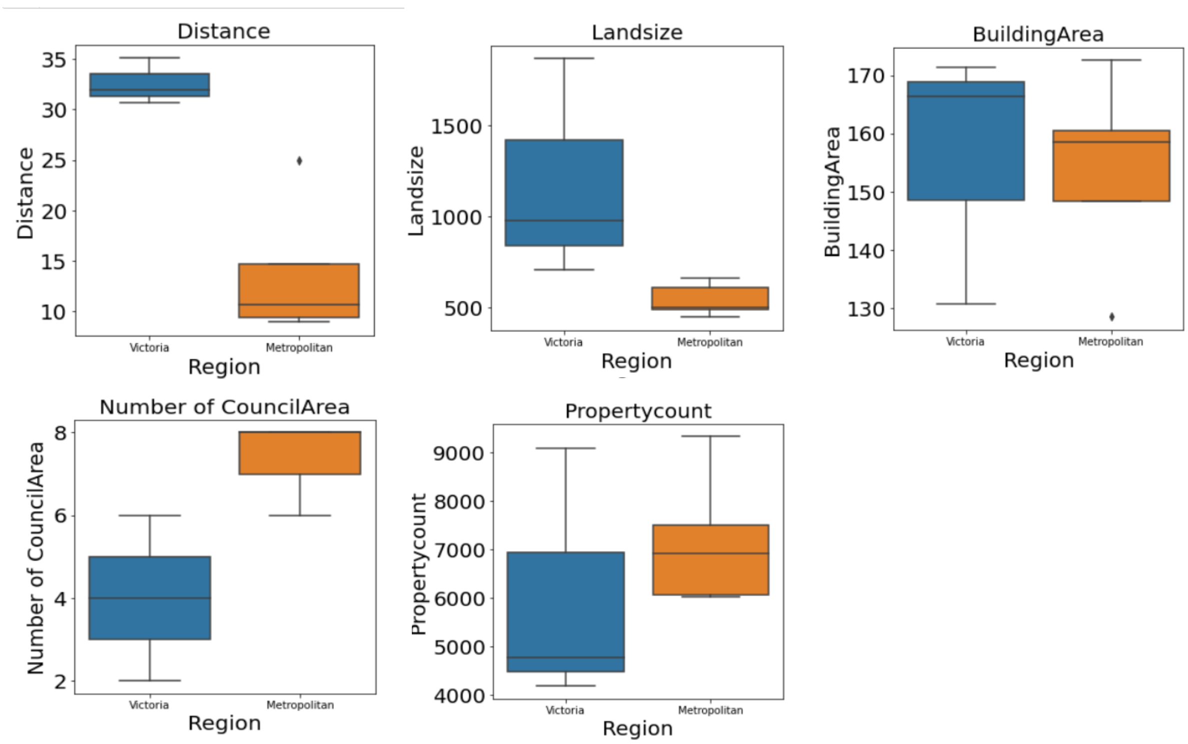
It seems to be that houses older than 50 years have higher price than modern houses, specifically house built between 1810 and 1950 have highest price in overall.

In question four, we will use prices per square meter instead of the price of the whole house.



As we can see, the highest price per square meter is Southern Metropolitan followed by Northern Metropolitan, Eastern Metropolitan, Western Metropolitan and South-Eastern Metropolitan… In general, house price per square meter in Metropolitan much higher than Victoria. Suburb Boronia and Balwyn have highest price per meter, they all belong to the Metropolitan.

The reason why price in Metropolitan is high?



From these boxplots, we can go to conclude that the house price per square in Metropolitan higher than Victoria because:

- In Metropolitan, distance to central business district shorter than Victoria
- Land size in Victoria is bigger than Metropolitan
- The number of council area in Metropolitan is 2 times higher than Victoria
- Property in Victoria is smaller than Metropolitan

In conclusion, house price in Victoria is lower than Metropolitan. In Victoria, it has more land size with less property that means the space here will be more spacious and airier, but less council area that means this area will be less secure, Victoria will be suitable for customer who want to live in less noisy areas, more free space and less price and don't mind going as far as central business district. In Metropolitan, it will be suitable for customer with good finance, usually have job at central business district and council area, prefer to live in areas with good security.

III. Data Pre-processing.

1.Data cleaning:

Firstly, finding all missing values and handling it. It has 472 missing values in overall, details are as follows:

```
Rooms             10
Type              25
Price            191
Bathroom          13
Car               24
Landsize          87
BuildingArea      98
YearBuilt         24
```

For categorical variables:

'Rooms', 'Type', 'Bathroom', 'Car' are categorical variable so the best choice for filling the missing value is using the most frequent. In 'YearBuilt', we handled missing values based on the most frequent building year in each region

For continuous variables:

We fill the missing value of 'Price', 'Landsize', 'BuildingArea' based on the average price of selling house in each region.

2. Drop columns:

We drop 'Bedroom2', 'Postcode', 'Address', 'YearBuilt'(use Age column instead), 'Suburb', 'Historic', 'P/m^2' because these features are no longer meaningful and effect to our model.

3. Encoding categorical variables:

Encoding 'Regionname', 'CouncilArea', 'Type' and 'Method' features into numbers because we can't build the model if there's still text data.

4. Train-test split:

In training set, we will take randomly 70% of data, the rest is for testing set.

5. Scaling data:

Because the range of values in each column has a huge difference, it will affect very much to our model. Using min-max scaler to scale data into range between 0 and 1**,** to use this method, we will call MinMaxScaler package, and fit it with training set and after that transform both training and testing set. Another reason to use MinMaxScaler because the distribution in our data is not Gaussian.

IV. Build and train model.

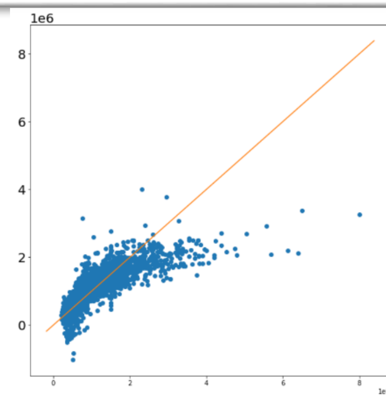
1. Build and train model:

Because the dependent variable is continuous variable, so we will use regression model for handling this. Three models are regression model, decision tree and random forest. We will call 'LinearRegression', 'DecisionTreeRegressor' and 'RandomForestRegressor' package from scikit-learn library. Moreover, we will try KNN (k-nearest neighbor) that an algorithm is often used in classification problems.

- ***Linear Regression:***

From the training, we will have coefficient and intercept, from that we can write down the equation for linear regression.

```
coeficient [ 1.18467552e+06 -3.12100626e+05  1.12166500e+03 -1.77323358e+06
  1.01147393e+06  4.97768669e+05  1.25637534e+06  5.02150956e+06
 -7.44177690e+04 -6.98080376e+05  1.45924773e+06  2.87229405e+05
  5.50329403e+04  2.62865282e+06]
intercept 151550.4800679027
```



- Evaluation:

We will calculate the coefficient of determination R2 of the prediction, mean squared error and mean absolute error:

R_squared	mean squared error	Mean Absolute error
0.620299	1.893229e+11	278415.0

R^2 is 0.62 and the chart show that how well model fits with the dependent variable, we can see that there are still some points that stay outside the line. MSE and MAE both for calculate error of the model but different ways in formular.

- Fine-tuning:

Linear regression has not any parameter for searching so we will use cross-validation method. In Cross-validation, we use k-fold that means we separate original data into K subsets, in each time, one subset uses for testing and other K-1 subsets use for training.

In this case, we split data into 10 subsets.

- Evaluation after fine-tuning:

R_squared	mean squared error	Mean Absolute error
0.62512958	1.39066723e+11	247890.61326346

- ***KNN (K-nearest Neighbor):***

K-nearest Neighbor is a simple supervised learning algorithm. Its work based on the distance between each point, the predicted value is calculated based on the number of nearest values (n_neighbors). We will choose 5 neighbors for starting.

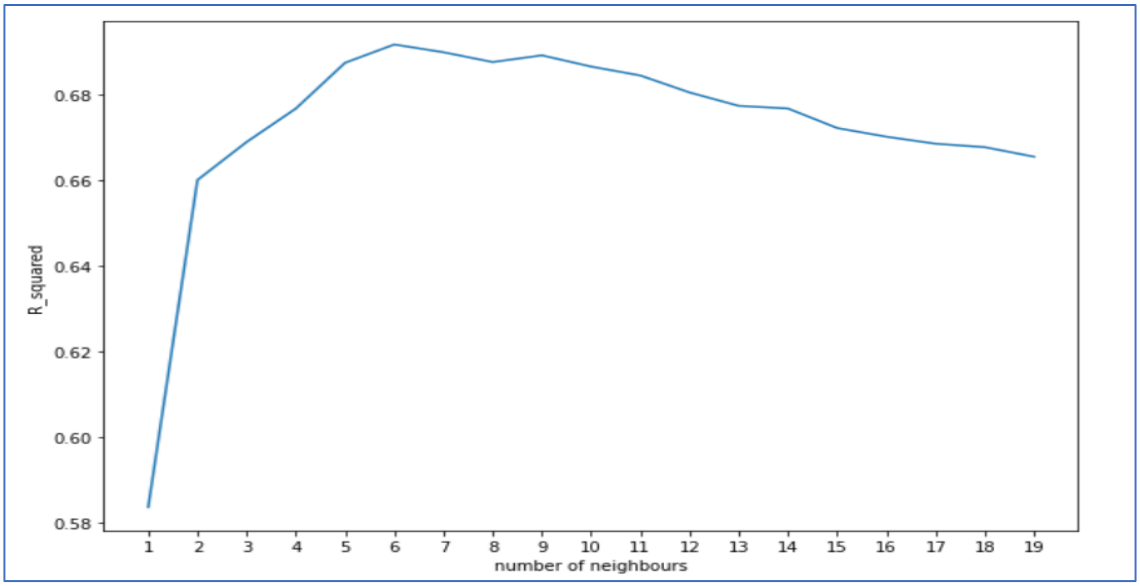
- Evaluation:

R_squared	mean squared error	Mean Absolute error
0.687561	1.557856e+11	233290.248023

- Fine-tuning:

We will select the best number of neighbors:

```
'Best parameters':{n_neighbors: 6}
```



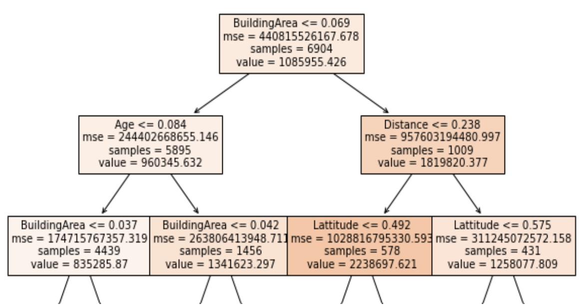
- Evaluation after fine-tuning:

R_squared	mean squared error	Mean Absolute error
0.691846	1.536487e+11	232317.085942

R^2 is almost 70%, and error decreased, so we can make sure that KNN has better performance than Liner Regression.

- ***Decision Tree:***

The number of depths we will select max depths equal to 5, criterion is mse (mean squared error) which is equal to variance reduction as feature selection criterion and minimizes the L2 loss (least Square Errors) using the mean of each terminal node, min sample split is 10.



The root note is building area which has the smallest L2 loss, the internal nodes are age and distance which has the second smallest L2 loss, in the lower branches are higher least square error.

- Evaluation:

The model fits with dependent in test set up to 64%. Both R^2 and MSE and MAE seem to be decrease comparing to Linear Regression.

```
Score R^2 for training : 0.6509760235163139
Score R^2 for testing : 0.6374349359051753
mean_squared_error: 180778931021.56656
mean_absolute_error: 272699.89262451977
```

- Fine-tuning:

We will use two methods GridSearchCv and RandomizedSearchCV

+ GridSearchCv works like a for loop function, we will set the range of loop in each parameter of the model and it will take the best prediction score in this range.

```
Best parameters Decision Tree: {'max_depth': 15, 'min_samples_split': 50}
```

Max depths: 15 and min sample split: 50 are the best parameter if we use grid search

+ RandomizedSearchCv is very easy, it will select random value of each parameter that we set up before.

```
Best parameters: {'min_samples_split': 60, 'max_features': 'sqrt', 'max_depth': 23, 'criterion': 'friedman_mse'}
```

Min sample split: 60, max features: 'sqrt', max depth: 23, criterion: 'friedman_mse' are parameters if we use RandomizedSearchCv.

- Evaluation after Fine-tuning:

+ Gridsearchcv:

```
Test score Decision Tree: 0.715670658376762
mean_squared_error: 141769738805.47818
mean_absolute_error: 218770.32145871074
```

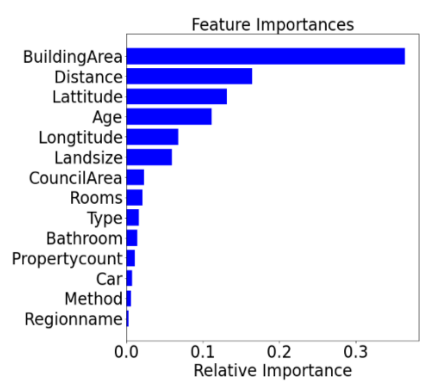
+ RandomizedSearchCv:

```
Test set score: 0.7200306226309185
mean_squared_error: 139595812646.524
mean_absolute_error: 224257.26568981988
```

Two methods that give us almost the same results.

- ***Random forest:***

Random forest is a supervised learning algorithm. As the name implies, Random Forest uses trees as a base. Random forest is a collection of Decision Trees, each of which is selected by a random-based algorithm. We will set max_depth to 10, n_estimators (the number of trees in the forest) to 100. From random forest, we can see building area have most effect to the price. Followed by distance, latitude, age, longtitude, landsize, coucilArea, rooms, type, bathroom and propertycount. Others features have not much effect with the model.



- Evaluation:

```
Score R^2 for training: 0.9063901103128891
Score R^2 for testing: 0.8043380065915144
mean_squared_error: 97559223192.78891
mean_absolute_error: 182319.41470858632
```

Surprisingly, model fits pretty well with dependent variable, up to 80% in testing
And error decrease quite a lot comparing to other models.

- Fine-tuning:

+ GridSearchCv:

```
Best parameters Random Forest: {'max depth': 10, 'n estimators': 150}
```

The number of depths is 10 and the number of trees in the forest is 150, we will have the best parameters for random forest model.

+ RandomizedSearchCv:

```
Best parameters: {'n_estimators': 200, 'min_samples_split': 5, 'max_features': 'sqrt', 'max_depth': 45}
```

The number of trees in the forest is 200, at least 5 samples split, calculate max feature based on square root of training data with 45 depths as a result of random search cv.

- Evaluation after Fine-tuning:

+ GridSearchCv:

```
Score R^2 for testing: 0.8106803914167178
mean_squared_error: 94396840320.37901
mean_absolute_error: 180633.12290855122
```

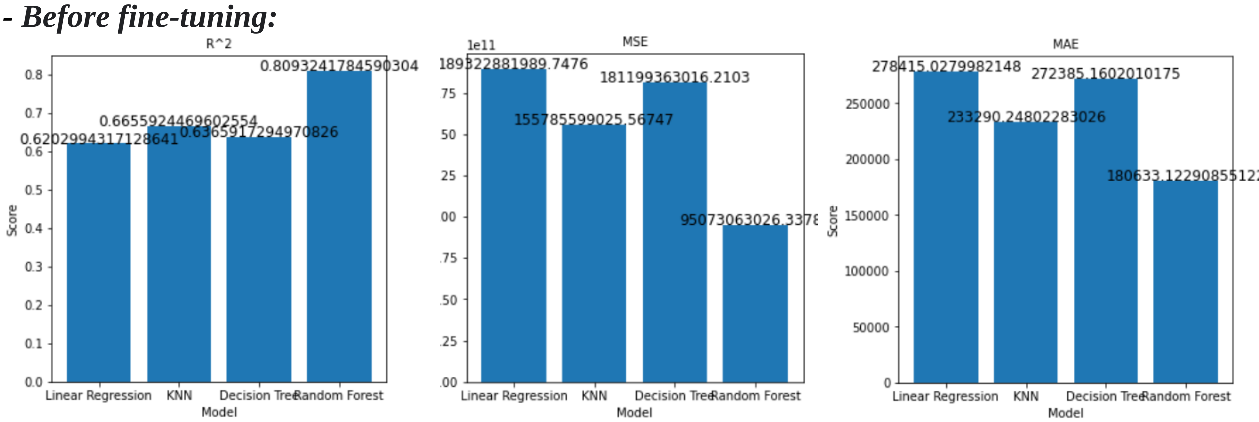
+ RandomizedSearchCv:

```
Test set score: 0.8337524185291422
mean_squared_error: 82892873692.21591
mean_absolute_error: 161372.96089680205
```

RandomizedSearchCv gave us better score for random forest.

2. Comparison between models:

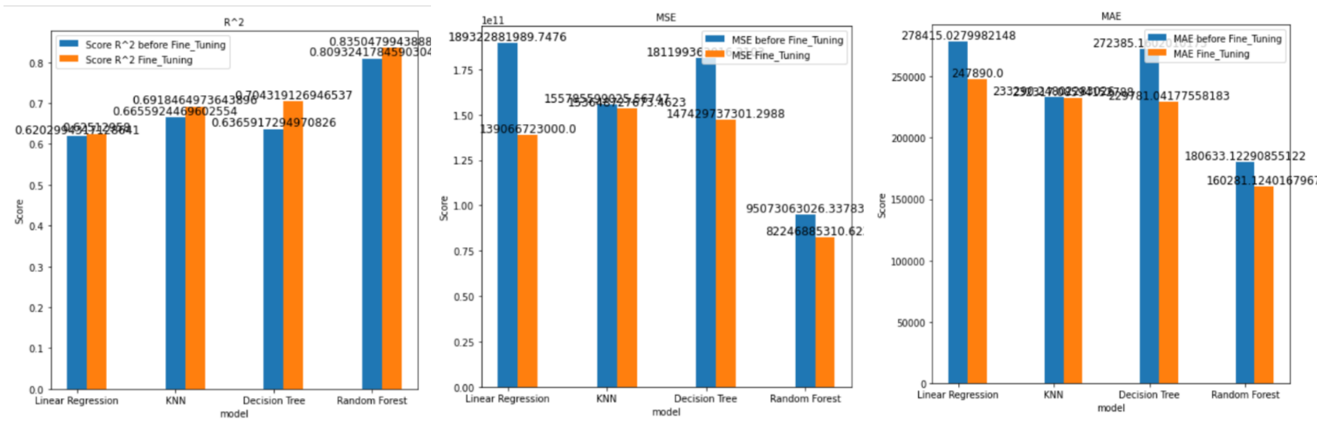
- Before fine-tuning:



Because the score in each function has a huge different, so we decided to draw three bar charts.

As we can see, random forest is most suitable algorithm in this problem, has the most accurate predictions, with highest coefficient of determination score and lowest mean squared error and mean absolute error; followed by KNN. Decision tree and linear regression, we can see that these two algorithms have almost the same score.

- After fine-tuning:



Orange columns are the evaluation after fine-tuning, blue columns are the evaluation before fine-tuning. Random forest still gives the best results. The coefficient of determination score of random forest increases from 0.80 to 0.83 (by 3%), mean squared error decreases from 9.755922e+10 to 8.289287e+10. Decision tree and KNN also have better result, linear regression doesn't seem to have much change. We can be seen figures below:

Model	Score R^2 before Fine_Tuning	Score R^2 Fine_Tuning	MSE before Fine_Tuning	MSE Fine_Tuning	MAE before Fine_Tuning	MAE Fine_Tuning
Linear Regression	0.620299	0.625130	1.893229e+11	1.390667e+11	278415.027998	247890.000000
KNN	0.665592	0.691846	1.557856e+11	1.536487e+11	233290.248023	232317.085942
Decision Tree	0.636592	0.704319	1.811994e+11	1.474297e+11	272385.160201	229781.041776
Random Forest	0.809324	0.835048	9.507306e+10	8.224689e+10	180633.122909	160281.124017

In conclusion, random forest with the number of trees in the forest is 200, at least 5 samples split, calculate max feature based on square root of training data with 45 depths is the best choice with the goodness of fit up to 83% and lowest error.

V. Answering for the business problem.

From the analysis and prediction method above, we can answer the question: "How can real estate companies maximize their profits from selling the house Melbourne". Real estate companies will have two main activities: buying and selling.

In purchasing activities, before buying any house, they can use machine learning algorithm that we have already trained by their data (random forest) with the fitting rate up to 83% to predict the price they will sell the house comparing with the price they have to pay to get that house, so they can minimize buying more expensive than selling, risk mitigation and lose.

In selling activities, they should focus on selling houses with types are house, cottage, villa, semi, terrace, with 1 or 2 bathrooms, 3 rooms and bedrooms, 2 car parking spaces with building area from 132 to 180 square meters, year built from 1810 to 1950 and 1970 to 2000; located at suburb Metropolitan. Moreover, they can cluster customers into 2 group, introduce houses in Victoria for who want to live in less noisy areas, more free space and less price and don't mind going as far as central business district; in Metropolitan for customer with good finance, usually have job at central business district and council area, prefer to live in areas with good security; by that, they can increase the level of customer satisfaction and sales figures. In overall, companies will sell more houses, the number of houses sold will increase, more revenue.

In addition, predicting housing prices also helps stakeholders such as people who want to buy a house in Melbourne can estimate the price of the house that is suitable for them, in addition, they can avoid being scammed into buying a house at sky-high prices.

VI. Project expansion.

If the company gives us more information about houses such as how many floors, traffic situation in there, information about buyer (have family or not, how many children…). The more information we have, the higher the correct prediction rate.

Moreover, we have just only predicted house price in Victoria and Metropolitan state in general and in Melbourne in particular. What happens if we want to predict the house price in other states or other cities. I make sure that the result will be incorrect if we still use this already built model.