ECA5372 Big Data Analytics and Technologies

Housing Price Model in Melbourne

--Based on Multiple Linear Regression Model And

Random Forest Regression Model

Presentation Link(https://youtu.be/defBN2zG1dg)

Supervisor: Prof. Wee Kiang Yeo

Group 2

Yang Huan A0224968N

Shi Yishu A0224975U

Wu Feihan A0226293A

Liang Yue A0225081N

Fu Zhehao A0226162M

Department of Economics

National University of Singapore

1. Problem statement

We mainly studied a static model of Melbourne house prices in 2017. We take the price of Melbourne as the target column, find the factors that influence the price from the number of rooms, geographical location, distance and other known factors, and build the multiple linear regression model and random forest regression model respectively.

Since the data is from 2017, we assume that the market environment and other external factors of all houses are the same. So we look for influencing factors based on the characteristics of houses themselves, without considering policies and market environment.

2. Dataset

Source of dataset : Link

2.1 General description of dataset

This dataset has 13580 rows and 21 columns. Data variables are shown in the following table. It was created in September 2017 by Dan B. Additionally, homes with no Price have been removed.

Table 1 Variable description

Column Name	Meaning	Data Type	Feature Type
Suburb	Location of suburb	String	Categorical
Address	Premises location	String	Categorical
Rooms	Number of rooms	Integer	Numerical
Type	Premises type	String	Categorical
Price	Price in dollars	Double	Numerical
Method	Way to sold	String	Categorical
SellerG	Real Estate Agent	String	Categorical
Date	Date sold	String	Categorical
Distance	Distance from CBD	Double	Numerical
Postcode	Premises location	Double	Categorical
Bedroom2	Number of Bedrooms	Double	Numerical
Bathroom	Number of Bathrooms	Double	Numerical
Car	Number of carspots	Double	Numerical
Landsize	Land Size	Double	Numerical
BuildingArea	Building Size	Double	Numerical
YearBuilt	Built year	Double	Numerical
CouncilArea	Governing council for the area	String	Categorical
Lattitude	Precise location	Double	Categorical
Longtitude	Precise location	Double	Categorical
Regionname	General Region (West, North West, North, North east …etc)	String	Categorical
Propertycount	Number of properties that exist in the suburb.	String	Categorical

2.2 Why this dataset is fit with problem

It includes lots of useful information about the Melbourne's real estate. We could come up lots of insights and problem base on this dataset. Our problem will not limit to the dataset. And this dataset has more than 13,000 rows, for one city's real estate, it is a great volume for us to investigate.

3. Data preparation

3.1 Dataset cleaning

Firstly, when we browse the dataset, we find that three columns 'BuildingArea', 'YearBuilt' and 'Car' have a lot of empty cells. These three columns are related to the house price. If we just delete all rows which have even one empty cell, the dataset will delete too many data and be biased. Thus, we decide to use mean value of these three columns to fill empty cells, respectively.

Secondly, after filling empty cells in that three columns, we drop rows that have duplicate values in all columns, that means exact duplicates. The result shows there is no duplicated rows in this dataset.

Finally, we remove all the rows that have empty cells in any of the columns. Thereafter, we re-assign the results back to the another dataframe. There are 12211 rows remaining.

3.2 Column selection

Firstly, convert all String-type columns to Double-type and Vector-type columns, respectively. Select some features to calculate their correlation coefficient and draw the relationship picture between every two columns.

According to their correlation coefficient, Rooms is highly correlated with Bedroom2 and we keep Rooms and remove Bedroom2. Suburb is highly correlated with Propertycount and we keep Propertycount and remove Suburb. SellerG, Date and CouncilArea have low correlation with Price and we delete all of these three columns.

For precise location, we are worried about overfitting because house prices of nearby locations are very close. That is, when we input the precise location, the price will be locked and frozened. Even if it will enhance our predictions, it is not the nature of the property or economic variables that contributes to price like Size and Location to CBD etc. It is enough for us to just add one column that can represent a fuzzy location of houses. Thus, we remove Longtitude, Lattitude, Postcode, and Address.

Finally, we select 'Rooms', 'TypeclassVec', 'MethodclassVec', 'Distance', 'Car', 'Bathroom', 'Landsize', 'BuildingArea', 'YearBuilt', 'RegionnameclassVec', 'Propertycount' as feature columns and 'Price' as the target label.

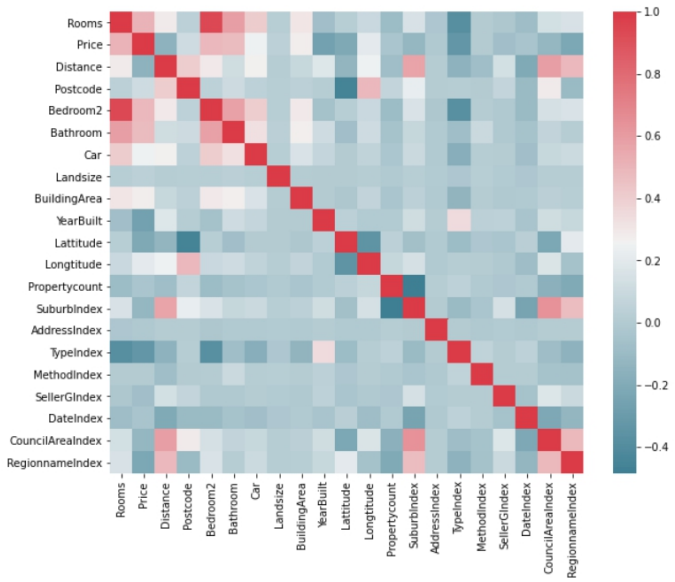

Figure 1 Correlation coefficient

3.3 Categorical columns

Convert all String-type columns to Vector-type columns and assemble all selected columns to one column named ‘features’.

4. Machine Learning Models

4.1 Models selection

(1) Multiple Linear Regression Model

Multiple Linear Regression Analysis is a statistical Analysis method that takes one variable as dependent variable and more variables as independent variables, establishes the quantitative relationship of linear mathematical models among multiple variables.

The multiple linear regression analysis method is simple and convenient when analyzing multi-factor models. It is the most basic and simplest kind of multiple regression analysis. Using the multiple linear regression model, as long as the model and data used are the same, the only result can be calculated through standard statistical methods. Besides, regression analysis can accurately measure the degree of correlation between various factors, and improve the effect of the prediction equation. The multiple linear regression analysis method is more suitable for actual economic problems, and is used when there are multiple influencing factors.

(2) Random Forest Regression Model

It's typically generated from the top down. Each decision or event may lead to two or more events, leading to different results. Such decision branches are drawn like the branches of a

tree, so they are called decision trees. Regression tree splits effectively partition the covariate space into a set of rectangles and then fit a simple model in each one. Random forest is a classifier containing multiple decision trees. It forms multiple decision trees by sampling the original data with put back, and the category of its output is determined by the mode of the category output by the individual tree.

Random Forest Model can solve both type of problems that is classification and regression. For many kinds of data, it can produce high accuracy classifier and have power of handling large data sets with higher dimensionality. Even if a large part of the data is missing, it can still maintain a certain accuracy in the estimation. Besides, this method evaluates the importance of variables when determining categories and it can produce an internal estimate of the error after generalization when constructing the forest. The data selected by our group is characterized by a large number of independent variables and a large number of missing values of some independent variables. We believe that the advantages of random forest are in line with the data characteristics.

4.2 Model Established

We first load and parse the data file, converting it to a DataFrame. Then, Split the data into training and test sets (30% held out for testing). We train Random Forest Regression Model and Multiple Linear Regression Model respectively, and then make the predictions for both training set and test set. After that, we could select predictions and compute test error. This step is preparing for the validation.

4.3 Model Validation

Because we want to predict the house price, we focus on the metrics that reflect the price prediction's deviations. As a result, we choose the mean absolute error (MAE), root mean squared error (RMSE) and R-squared (R2). Since average housing price is around 1,000,000 dollars, we expected that absolute may be more intuitive. But other two is also valid.

For Random Forest Regression Model, we check the prediction and then calculate the test error. The result shows below.

Table 2 Result of Random Forest Regression Model

Random Forest Regression	Training set	Test set
MAE	238525	249385
RMSE	360488	405662
R2	0.668156	0.630271

We find that in Random Forest Regression Model, the metrics value of error between Training set and Test set is close, we may conclude that this model is appropriate. For the absolute value, we could infer that this amount error is reasonable because the error is just one-fifth of the average price.

For Multiple Linear Regression Model, the result shows below.

Table 3 Result of Multiple Linear Regression Model

Multiple Linear Regression	Training set	Test set
MAE	268293	273998
RMSE	393699	425818
R2	0.604195	0.592618

We find that in Multiple Linear Regression Model, the metrics value of error between Training set and Test set is close and Test set is a little higher than Training set.

5. Conclusion and discussion

5.1 Results

Both models in our study seem to be stable and reliable because the error value is not so big and the outcome between training set and test set is very close.

If we should choose one, we may prefer Random Forest Regression Model because of validation part, the error is such smaller than Multiple Linear Regression Model.

Random Forest Regression Model seems better in our cases, because a Random Forest's nonlinear nature may give it a leg up over linear algorithms, making it a great option. But it may be over predicted sometimes.

5.2 Challenges

During the project, we also encountered some challenges and problems. In the process of collecting data, due to issues such as the availability of data, we cannot cover all the factors that affect Melbourne's housing prices, so the model is not fully explanatory. Because the data is only from 2017, our results may not be representative.

In the process of establishing the model, columns cannot be completely selected according to the correlation of the independent variables, so it is necessary to repeatedly modify the variables to achieve the best feature combination.

In addition, in the process of writing the code, many unexpected bugs appeared. We have to find out the causes of all the bugs, which costs us a lot of time and energy.

5.3 Insights

Through the project work, all of us have gained a lot of insights. Because of the different hardware environments of team members' computers, everyone's code cannot be put together in a short period of time, and it takes a certain amount of time to modify. Therefore, before writing code, we should standardize the language and structure of the code, and improve the versatility and reusability of our own code.