Housing Price Model In Melbourne

--Based On Multiple Linear Regression Model And Random Forest Regression Model

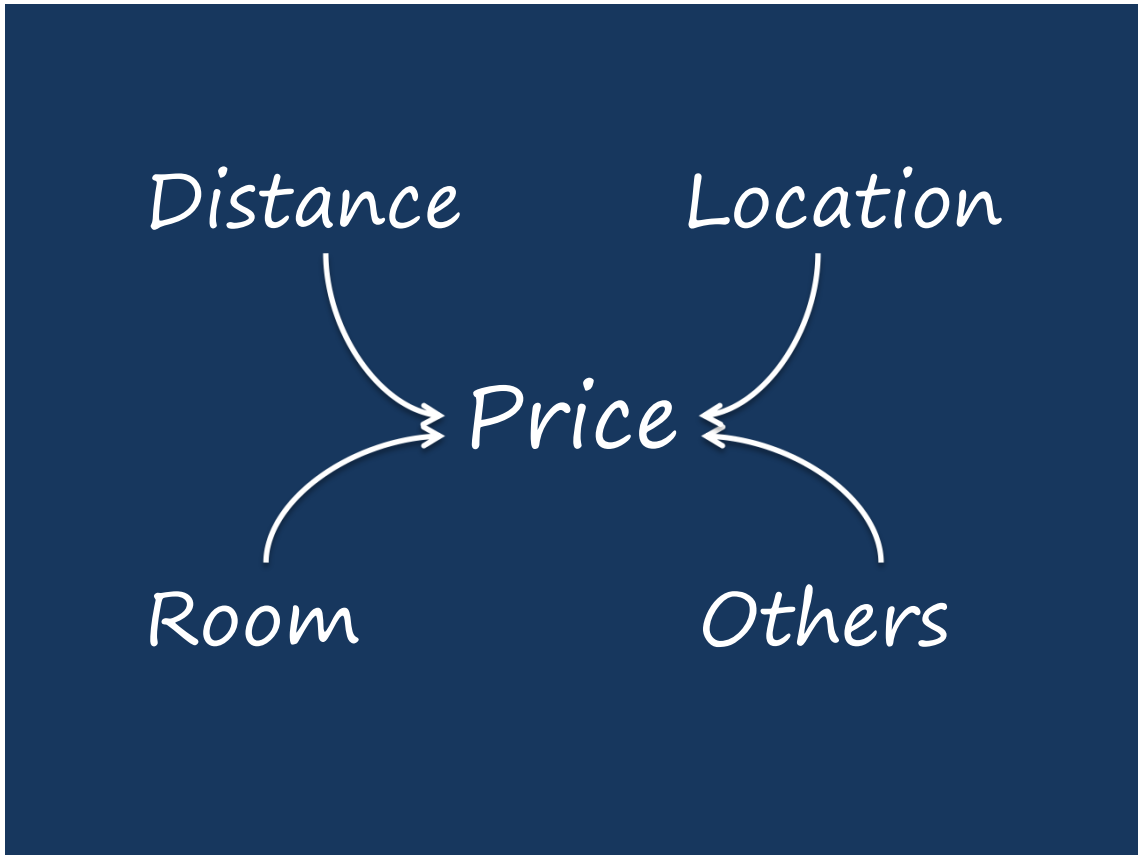
Group2

Yang Huan A0224968N
Shi Yishu A0224975U
Wu Feihan A0226293A
Liang Yue A0225081N
Fu Zhehao A0226162M

CATALOGUE

Part 1 Problem statement

Problem statement



Assumption

- Market environment
- External factors

Data: House Prices in Melbourne (2017)

Part 2 Dataset

General description of dataset

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

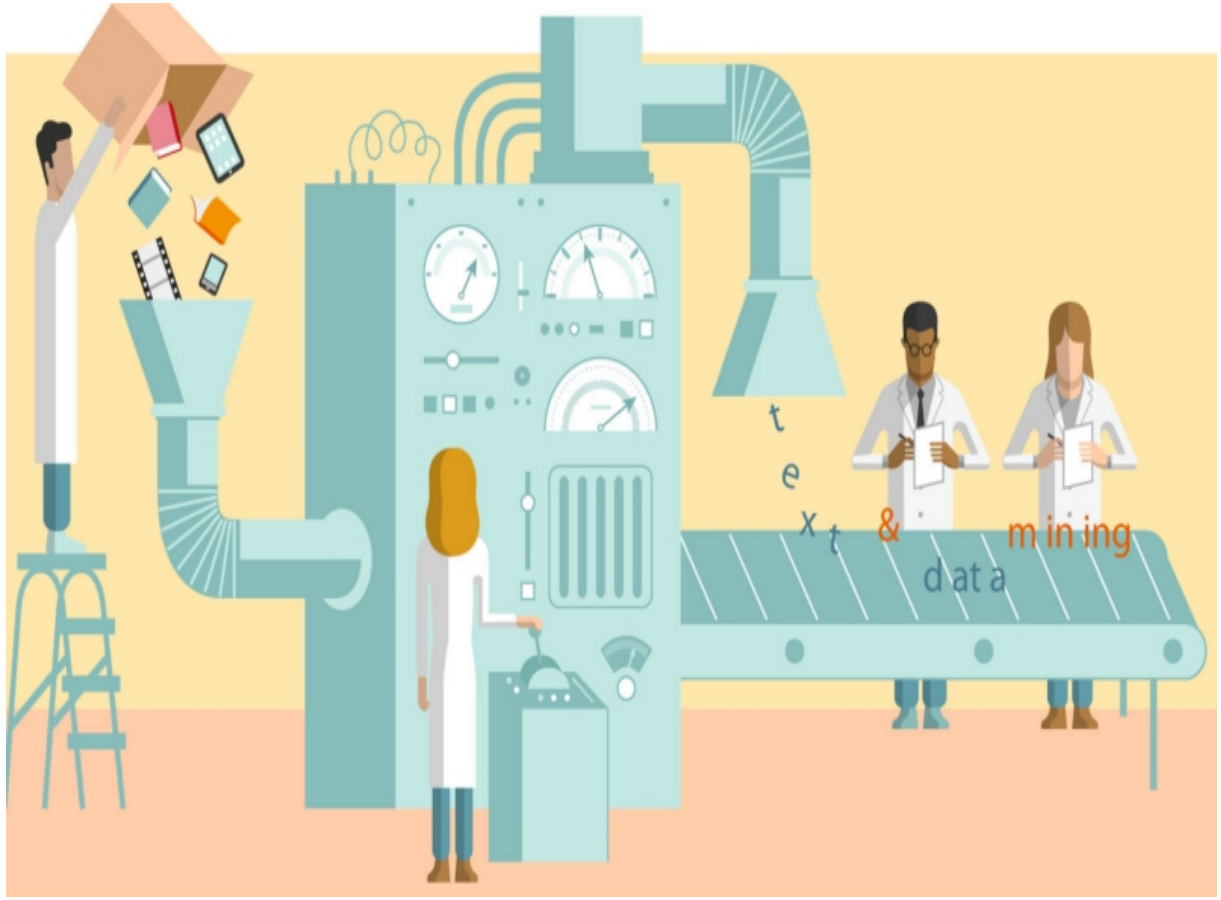
- Rows: 13580
- Columns: 21

- Numerical: 9
- Categorical: 12

Part 3 Data Preparation

Dataset Cleaning

1. Filling empty cells with mean of the corresponding columns
2. Deleting duplicate rows
3. Removing all rows that have empty cells in any of columns



Columns Selected

Step1

Converting all String-type columns to Double-type and Vector-type columns

MethodclassVec	SellerGIndex	SellerGclassVec	DateIndex	DateclassVec	CouncilAreaIndex	CouncilAreaclassVec	RegionnameIndex	RegionnameclassV
(1.0, 0.0, 0.0, 0.0)	15.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	14.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	3.0	(0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	1.0	(0.0, 1.0, 0.0, 0.0, 0.0, 0
(0.0, 1.0, 0.0, 0.0)	52.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	45.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	12.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0
(1.0, 0.0, 0.0, 0.0)	0.0	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	16.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0
(1.0, 0.0, 0.0, 0.0)	1.0	(0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	39.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	1.0	(0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	0.0	(1.0, 0.0, 0.0, 0.0, 0.0, 0
(1.0, 0.0, 0.0, 0.0)	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	41.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	4.0	(0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, ...	0.0	(1.0, 0.0, 0.0, 0.0, 0.0, 0
(0.0, 0.0, 1.0, 0.0)	3.0	(0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	9.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	15.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	3.0	(0.0, 0.0, 0.0, 1.0, 0.0, 0
(0.0, 0.0, 1.0, 0.0)	1.0	(0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	9.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	1.0	(0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	0.0	(1.0, 0.0, 0.0, 0.0, 0.0, 0
(1.0, 0.0, 0.0, 0.0)	0.0	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	48.0	(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	2.0	(0.0, 0.0, 1.0, 0.0, 0.0, 0

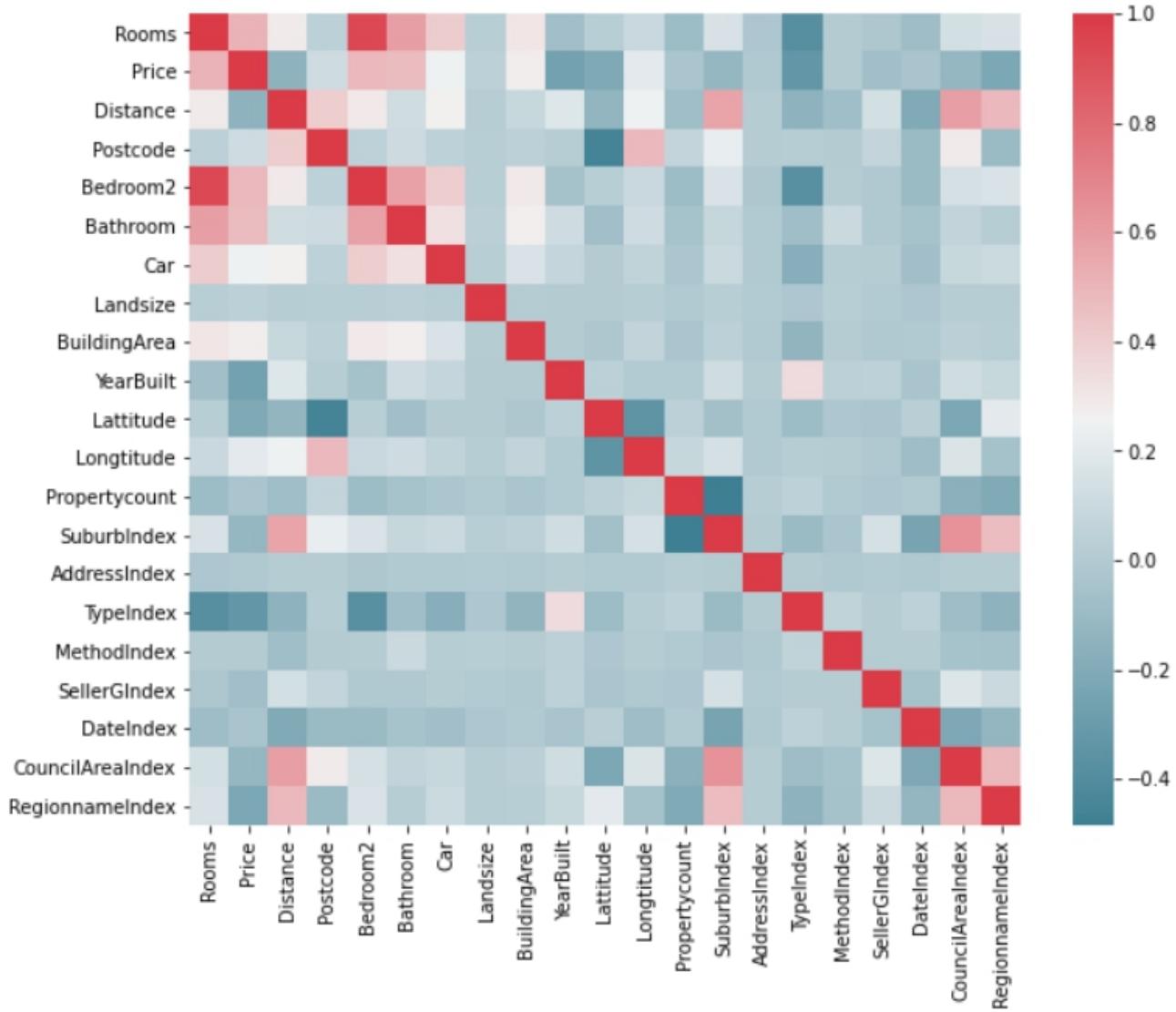
Columns Selected

Step2

calculate their correlation coefficient

Step3

select 11 columns as features and ‘Price’ as the target label



Part 4 Machine Learning Models

Multiple Linear Regression Model

Advantages

The regression analysis method is simpler and more convenient when analyzing multi-factor models; It is the most basic and simplest kind of multiple regression analysis. 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 the degree of regression fitting, and improve the effect of the prediction equation; The multiple regression analysis method is more suitable for actual economic problems, and is used when it is affected by multiple factor

Outcome

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

Random Forest Regression Model

Advantages

Random Forest Regress Model can solve both type of problems that is classification and regression.

For many kinds of data, it can produce high accuracy classifier; it can produce high accuracy classifier and have power of handle large data sets with higher dimensionality.

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.

Outcome

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

Part 5 Conclusion and discussion

Conclusion & Discussion

Results

- Both models in our study may seems like stable and reliable because the error value is not so big and the outcome between training and test is very close;
- We may prefer Random Forest Regression Model as from validation part, the error is such smaller than Multiple Linear Regression Model;
- Random Forest Model seems better, it is may because a Random Forest's nonlinear nature can give it a leg up over linear algorithms, making it a great option. But it may be over predicted sometimes.

Conclusion & Discussion

Challenges

- 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;
- 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 the process of writing the code, many unexpected bugs appeared. We had to find out the causes of all the bugs, which cost us a lot of time and energy.

Conclusion & Discussion

Insights

- Due to the different hardware environments of the 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 at the same time improve the versatility and reusability of our own code.

NUS National University of Singapore

THANK YOU !