Assignment 2 Report

Student name: Jielin Feng, Zedong Wen

SID: 480019484, 500334863

1 Set Up

1.1 Problem

The development brought about by globalization has profoundly affected our lives in many ways. Many studies have shown that the changes brought about by globalization have a significant impact on housing prices (Moallemi, M. et., al, 2021). More and more research is aimed at developed economies, especially Britain and the United States, while little is known about the housing price dynamics in Australia, so the housing price in Australia has become the focus of this study (Awaworyi Churchill, S. et., al, 2018). Our research mainly focuses on the following issues:

1. Find and simplify the key attributes related to the changes in housing prices.
2. Build models that can effectively forecast housing prices based on selected attributes.
3. Effectively improve the accuracy of prediction models and compare the effectiveness and significance of different models.

1.2 Hypotheses

H_0: There is no relationship between Price and "Bathroom, Rooms, Bedroom2, Longtitude, Type, Distance, Car, Lattitude".

H_1: There is relationship between Price and "Bathroom, Rooms, Bedroom2, Longtitude, Type, Distance, Car, Lattitude".

1.3 Reliability and effectiveness

In order to best predict housing price, Multiple linear regression, regression tree, random forest, KNN and SVR will be selected as the evaluation models. Test the quantitative reliability by calculating the significance testing. R^2 will be calculated and compared to measure effectiveness. MSE/MAE will be compared and to evaluate errors of building models.

1.4 Dataset

The data selected was retrieved from "https://www.kaggle.com/dansbecker/melbourne-housing-snapshot" it was created by Tony Pino (Kaggle, 2016), then processed further by DanB (Kaggle, 2018). The actual dataset contains 13580 instances described by 21 attributes. Further information and explanation of each attribute are provided in supporting notes (Kaggle, 2016). After initial data preprocessing in previous assignment 1, 8 key attributes that has top correlation to Price was selected.

2 Approach

2.1 Data preprocessing

A) Data cleaning

Firstly, as what was done in project 1, data was cleaned up in R with no missing values.

B) Transfer Categorical data into numerical data

Secondly, in order to better discover correlation between each attribute, categorical data is transformed into numerical data.

C) Feature selection

Thirdly, the correlation coefficient obtained by Pearson correlation analysis of the processed data set is shown in Figure 1. It can be seen that "Bathroom, Rooms, Bedroom2, Longtitude, Type, Distance, Car, Lattitude" has the highest correlation with "Price". Thus these 8 features are selected to build the model.

D) Log transformation

Fourthly, it is found that the distribution of price does not conform to normal distribution, what's more, assumption of linear regression does not meet as shown in Figure 2. Thus, we carried out a log transformation to the "Price", since the value cannot be negative, we performed log transformation on the basis of "original data plus one", which confirms all data are positive. The assumption of linear regression was met after log transformation as can be seen from Figure 3. From the residual plot, homoskedasticity and linearity assumption can be proven by the evidence of randomly spread out and symmetric across y axis. Also, from the QQ plot, apart from several points in the lower and upper tails, the majority of the points lie close to the diagonal line, hence, the normality and independence assumption is satisfied.

E) Load data from R to Python

Fifthly, the final data set in R was divided into two separate data sets, the data set "data_price_log" contains all prices after the log, and the data set "data_3" contains the other 8 attributes. These two datasets were saved and imported into Jupyter Notebook for further processing and analysis.

F) Normalisation

Normalisation is used to normalize the features of data by the "MinMaxScaler" method. The aim of normalisation is to prevent the weight of some features from being too high which might influence the result in the model. Final data after preprocessing contains 13518 rows, 8 attributes and Price.

2.2 Classifier method algorithms

5 algorithms are selected as list below, Random Forest was selected as a benchmark model, while the others were chosen to be comparable models. The reason for choosing specific algorithm and the choice of parameter tuning will be analysis as following:

A) Random Forest Regression

Random Forest is an ensemble classifier of the base classifier decision tree. It contains predictions from several regression trees and produces the output of mean of predictions which may achieve higher accuracy and prevent from overfitting. Thus, Random Forest Regression was chosen as benchmark model that is expected to have the best performance

Parameter tuning

[{'n_estimators':[100,150,200],'max_depth': [5, 10, 50, 100],'criterion':
['mse','absolute_error'],'max_features': ['auto','sqrt','log2'],'random_state':[88]}]

Random forest is the ensemble classifier of a regression tree. Thus, the selection of parameters of random forest are mostly consistent with the regression tree.

B) Multiple Linear regression

Multiple linear regression is a statistical technique that uses 8 attributes to predict the outcome of price, in which, these 8 features are independent variables, while Price is dependent variable. It is aiming to summarise a final formula using these 8 attributes to predict Price.

C) Regression tree

Regression tree is a model that combines the advantages of linear regression and decision tree to work on a regression task.

Parameter tuning

[{'max_depth': [5, 10, 50, 100],'max_features': [5,6,7],
'splitter': ['best', 'random'], 'random_state':[88]}]

'max_depth' gives the maximum depth, this parameter is added to avoid overfitting. 'max_features" represents the number of features that tree uses to make the split decision each time, as using all features may cause overfitting.'splitter' allows adopting the most optimal method to improve accuracy, "best" is chosen.

D) KNN regression

K- nearest neighbour regression is a non-parametric classification method. The input is decided by k closest neighbours of dataset, while the output value is the average of k closest neighbours' value

Parameter tuning

[{'n_neighbors': [6, 8, 12, 14],'p': [1,2]}]

'n_neighbors' determines the selection of k. The selection of k is very sensitive since a small k might lead to overfitting while a large k might decrease the performance effect. 'p' is used to determine the distance measurement.

E) SVR

SVR is known as supported vector regression, which applies mostly the same principle with SVM. In SVR, a margin of tolerance is set in approximation, which is different to SVM. Except for that. SVR is aiming to minimize the error, maximising the margin.

Parameter tuning

[{'C': [1, 10, 20],'epsilon': [1,0.1,0.5]}]

'C' (cost) decides the penalty value for the error/division data. 'Epsilon' represents margin of tolerance, the larger the epsilon, the less likely the penalty is applied, in contrast, small epsilon represents low tolerance.

3 Result

3.1 Results

Results for R^2, MSE and MAE are rounded into 3 decimals.

Time for each model is rounded into 3 decimals, then displayed in second(sec).

Table 1 Performance Comparison

	Random forest	Linear regression	Regression tree	KNN	SVR
R^2	0.814	0.673	0.753	0.785	0.782
MSE	0.053	0.093	0.071	0.062	0.062
MAE	0.168	0.233	0.195	0.185	0.184
Time	2.468	0.013	0.027	0.193	11.985

As can be seen from table1, compared with the R^2 of the model, the random forest is the largest in all the models. It is 0.814. The smallest is linear regression. is 0.673. Regression tree, and the fitting effects of KNN and SVR are similar, being 0.753 respectively. 0.785 and 0.782. It shows that the random forest model is the model with the best fitting effect among all models. See MSE and MAE again. Contrary to the R-square distribution, the numerical value with the largest is linear regression and that with the smallest is random forest. In MSE and MAE, the numerical value with the larger is, it means that the larger the interpretation deviation of the model from the prediction is, the worse the fitting effect of the model is. This is also similar to the conclusion we get in r square. For the model calculation time, the calculation times of linear regression, regression tree and KNN are all within 0.2. The fastest one is linear regression, with 0.013. The run time of random forest with the best model fitting effect is 2.468. The longest one is SVR, with 11.985.

3.2 Discussion

A) Random Forest Regression

Random Forest Regression has the best performance of R^2 is 0.814, MSE is 0.053 and MAE is 0.168. This is due to random forest being the ensemble classifiers of the base decision tree. The biggest advantage of random forest is that it can effectively avoid overfitting. Also, by combining the results of several decision trees, the result of Random Forest is more accurate and reliable. The only issue is that ensemble classifiers have lower computational time compared to single classifiers. Though in our case, the computational time is acceptable, if we use a larger dataset in future works, the computational time might be very slow.

B) Multiple linear regression

The R^2 of Multiple linear regression is 0.673, MSE is 0.093 and MAE is 0.233. This is because Multiple linear regression performs better in a small dataset, in contrast, our data set is relatively large, which might lead to overfitting and not ideal performance. In addition, the feature selected may not fit linearity very well, which can be improved in future works. However, Linear regression only takes 0.013 sec to run. The modelling and computational time is very fast even under a relatively large dataset. According to Figure 4, the final formula is:

$$\log(Price + 1) = -204.804 + 0.123 * Bathroom + 0.149 * Rooms + 0.027 * Bedroom2 + 1.074 * Longtitude - 0.275 * Type - 0.042 * Distance + 0.035 * Car + 1.664 * Lattitude.$$

C) Regression tree

Regression tree has a relatively good performance of R^2 is 0.753, MSE is 0.071 and MAE is 0.195. This may be due to the dataset being large, and the dimension of the dataset being relatively low, as only 8 features were selected as final. Although there is restriction of "max-depth", the biggest problem of the Regression tree is still likely to have overfitting. Overfitting might be the reason that Regression tree does not perform as good as Random Forest regression.

D) KNN Regression

KNN Regression has a relatively good performance of R^2 is 0.785, MSE is 0.062 and MAE is 0.185. The advantage of KNN regression is that it is simple and robust to outliers. KNN is good at dealing with low dimensional data, in which our dataset only contains 8 features. In addition, the selection of k is sensitive.

E) SVR

The R^2 of SVR is 0.782, MSE is 0.062 and MAE is 0.184. SVM does not perform as well as expected. This might be due to the low dimensional property of the dataset. Since if the dimension is low, information contained when projected to higher dimensional space will be very restricted and is very possible to cause overfitting.

In short, there are some limitations not only in the method, but also in the selection of attributes. In our research, we only selected 8 key attributes. It is not excluded that other attributes also have significant influence on the price. In the future, we should implement more accurate feature selection and do more preprocessing to make the model more complete and reliable.

4 Conclusion

In conclusion, Random Forest regression has the overall best performance, its r square is 0.814, MSE is 0.053, MAE is 0.168 and Time is 2.468. And due to the performance of models (from table 1 and figure 4, P-value and R^2), it can be concluded that we reject H_0 and there is sufficient evidence to show that a relationship between Price and "Bathroom, Rooms, Bedroom2, Longtitude, Type, Distance, Car, Lattitude". From this research, we realized the usefulness of machine learning, so as to improve our sources as data analysts. Machine learning is the core of artificial intelligence. In the future, we will apply machine learning to future data work and wider work.

References

Awaworyi Churchill, S., Inekwe, J., & Ivanovski, K. (2018). House price convergence: Evidence from Australian cities. *Economics Letters, 170*, 88–90. https://doi.org/10.1016/j.econlet.2018.06.004

Kaggle 2016, *Melbourne Housing Market*, Retrieved 14 September 2021, from https://www.kaggle.com/anthonypino/melbourne-housing-market

Kaggle 2018, *Melbourne Housing Snapshot*, Retrieved 14 September 2021, from https://www.kaggle.com/dansbecker/melbourne-housing-snapshot

Moallemi, M., Melser, D., Chen, X., & De Silva, A. (2021). The Globalization of Local Housing Markets: Immigrants, the Motherland and Housing Prices in Australia. *The Journal of Real Estate Finance and Economics*. https://doi.org/10.1007/s11146-021-09828-2

Appendix

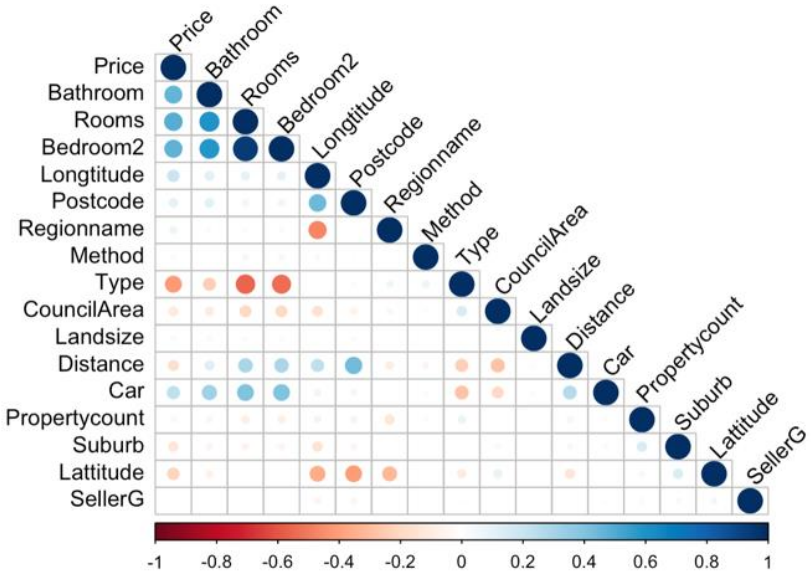

Figure 1 Numeric variable correlation (from R)

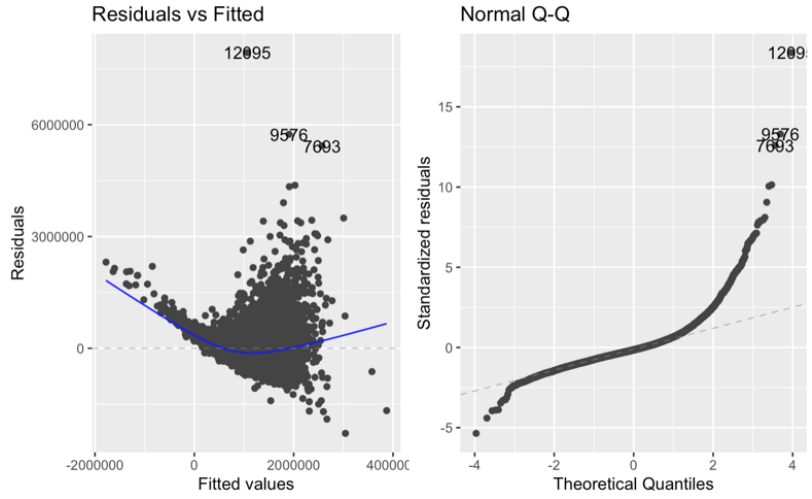

Figure 2 Price residual plot and Q-Q plot (from R)

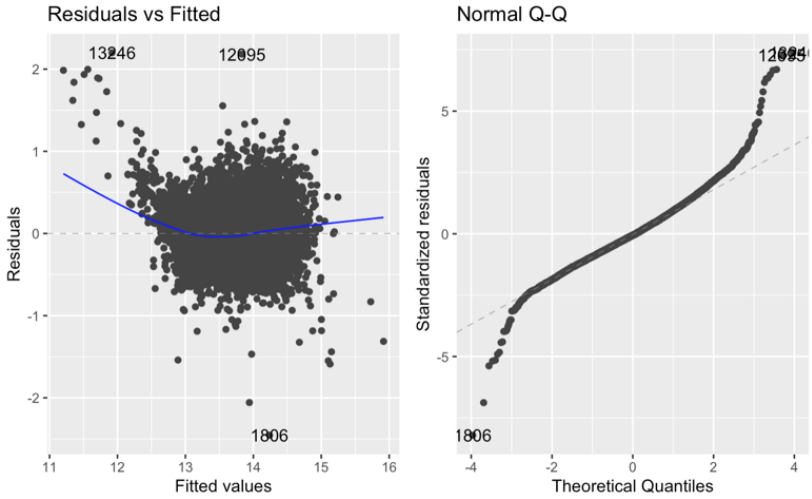

Figure 3 log(Price+1) residual plot and Q-Q plot (from R)

```
Call:
lm(formula = price ~ Bathroom + Rooms + Bedroom2 + Longtitude +
    Type + Distance + Car + Lattitude, data = data_all)

Residuals:
     Min       1Q   Median       3Q      Max
-2.45559 -0.19296 -0.01832  0.17584  2.19988

Coefficients:
                Estimate   Std. Error t value             Pr(>|t|)
(Intercept) -204.8036180    3.7251869 -54.978 < 0.0000000000000002 ***
Bathroom       0.1227807    0.0047607  25.790 < 0.0000000000000002 ***
Rooms          0.1494898    0.0084449  17.702 < 0.0000000000000002 ***
Bedroom2       0.0273924    0.0081415   3.365                0.000769 ***
Longtitude     1.0739445    0.0272116  39.466 < 0.0000000000000002 ***
Type          -0.2748544    0.0038098 -72.144 < 0.0000000000000002 ***
Distance      -0.0423508    0.0004842 -87.465 < 0.0000000000000002 ***
Car            0.0353411    0.0030010  11.777 < 0.0000000000000002 ***
Lattitude      1.6640539    0.0351636  47.323 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2992 on 13509 degrees of freedom
Multiple R-squared:  0.6779,    Adjusted R-squared:  0.6777
F-statistic:  3554 on 8 and 13509 DF,  p-value: < 0.00000000000000022
```

Figure 4 Multiple linear regression summary (from R)

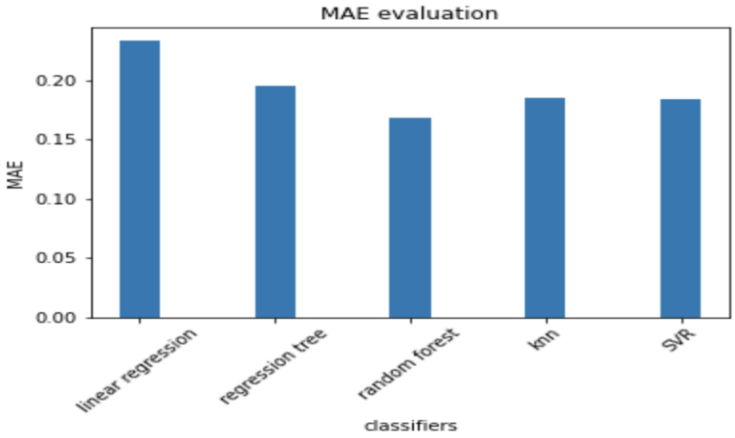

Figure 5 MAE evaluation (from Python)

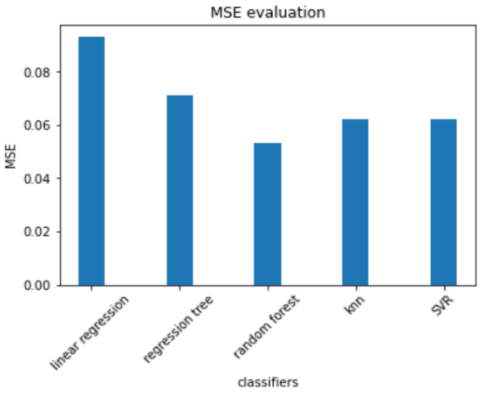

Figure 6 MSE evaluation (from Python)

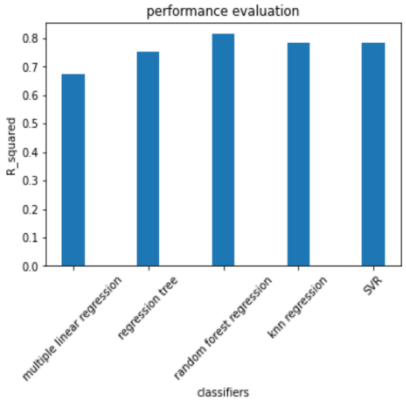

Figure 7 Performance evaluation (from Python)