INFOSYS 722 Research Report

Factors Affecting Sydney House Prices

Research Paper (Details of Steps 1 – 9)

Shiyu Lin

UPI: slin648

ID: 502522556

Department of Information Systems and Operations Management

(ISOM)

University of Auckland, New Zealand

Contents

(a) The practical problem

Sustainable cities and human settlements is one of the 17 Sustainable Development Goals of the UN.[1] Cities are centers of business, culture, science, productivity, social, human and economic development. Urban planning, transport systems, water, sanitation, waste management, disaster risk reduction, access to information, education and capacity building are all issues relevant to sustainable urban development.

In 2008, the global urban population exceeded the rural population for the first time in history.[1] This milestone marks the arrival of a new “urban millennium”, where by 2050 two-thirds of the world’s population is expected to live in urban areas.

In some cities, housing prices have experienced severe housing bubbles, such as Tokyo, where house prices were as high as $220,000 per square meter, and now are far below this value. Today, housing prices in some cities are seriously too high for local income levels, such as Shenzhen and Shanghai in China. All this poses a challenge to achieve the goal of sustainable cities and human settlements.

"Promoting sustainable human settlements development" is the subject of Chapter 7 of Agenda 21, which calls for providing adequate shelter for all, and providing adequate environment and support.

We know that it is very difficult to provide shelter for everyone while ensuring that the environment of the shelter is good. The price of a house is affected by many factors. This includes internal factors, such as the size of the house, the size of the house, etc.; it also includes external factors, such as the location of the house, the crime rate in the neighborhood where the house is located, etc.[2] A report by Etch Real Estate states that one of the most important short-term factors affecting house prices is interest rates. When the cost of borrowing money decreases, the number of qualified buyers increases. The price of a house is also related to the macroeconomic situation in the area.

Our practical problem here is to understand what factors affect house prices, and finally provide real estate market advice from the perspective of a data scientist.

(b) The research problem

Our research problems include collecting data, data transformation, choosing appropriate algorithms and drawing conclusions. This includes all steps in CRISP-DM.

First, we need to collect data and analyze the data to find the factors affecting house prices. The data we are looking for should contain one or more response variables to record house prices; it should also contain multiple explanatory variables that explain what factors affect house prices. Each explanatory variable can be a potentially significant factor, but it can also be a trivial factor. Importantly, the data needs to be relevant: we need each row of data to describe a particular property transaction, or the price of a particular property.

The data we collected may contain some quality issues, such as outliers, or incomplete data. We need to examine and process the data to make sure it is reliable. This allows us to build reliable models and draw meaningful conclusions.

Data sometimes needs to be transformed to meet some model assumptions. A frequently performed transformation is the Box-Cox transformation, which is used to transform data into normal equal variance data. Sometimes log transformations are also done, especially when applying linear models.

Next, we should choose an appropriate algorithm according to our data mining goals. In general, there are two types of algorithms: supervised learning and unsupervised learning, and each type contains many different algorithms.

There are many factors that affect house prices, some of which may not be described in the explanatory variables of the dataset. We assume that factors not described in the dataset are trivial, but in fact it may not be the case. Therefore, we need to select some important variables to do the modeling.

Finally, we need to draw conclusions based on the results of our model(s).

(c) The research objectives

Our objectives are Explanation and Prediction.

Explanation: I'll make assertions such as “a 1% increase in neighborhood crime, a 2% decrease in home prices, ceteris paribus”. I will also build models to find out which factors have the greatest impact on house prices.

Prediction: I will split the dataset into training and test sets, model on the training set, and test the accuracy of the model on the test set.

Desired Outcome:

1. Get to know what factors affect house prices the most, and what's trivial.
2. Make the model obtain a high prediction accuracy in the test set.
3. Based on the actual situation, I will explain my conclusions and put forward some logical conjectures.

(d) The literature that explores potential solutions and methodologies that addresses your objectives

An article published in the International Journal of Science and Research (IJSR) (Machine Learning Algorithms - A Review by Batta Mahesh) describes some machine learning algorithms that I might use. The article says that Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise.

As mentioned in the article, in general, machine learning algorithms can be divided into two categories: supervised learning and unsupervised learning.

Supervised learning is a machine learning task that infers a function from labeled training data consisting of a set of training examples. Supervised machine learning algorithms are those that require outside help. The input dataset is divided into training dataset and test dataset. The training dataset has output variables that need to be predicted or classified.

Unsupervised learning is different from supervised learning where there is no right answer and the algorithms are left to their own design to discover and present interesting structures in the data. It is mainly used for clustering and feature reduction.

Another IEEE paper (A review of supervised machine learning algorithms) details machine learning algorithms we might use in this project. A simple example is a (generalized) multivariate linear model. This is a supervised learning model. In this model, the response variable is continuous, while the explanatory variables can be continuous or categorical. During the training process, the explanatory variables and response variables on the training set are input; on the test set, the explanatory variables are input and the response variables are output. Another model is Bagging. Bagging or bootstrap aggregating is applied where the accuracy and stability of a machine learning algorithm needs to be increased. It is applicable in classification and regression. Bagging also decreases variance and helps in handling overfitting. Random Forest is a kind of Bagging algorithm.

(e) The research methodology adopted

The goal of this research project is to study the factors that influence Sydney property prices and to predict property prices based on different explanatory variables in the dataset.

The data mining goals we focus on are:
Explanation: Identify important explanatory variables that affect property prices and explain the relationship between these variables and property prices.
Prediction: Predict the house price given the explanatory variables.

Desired Outcome: Get to know what factors affect house prices the most, and what's trivial, and make the model obtain a high prediction accuracy in the test set.

There are three main categories of data mining methods: Classification, Clustering and Regression.[7] we can implement all these algorithms in Python, and there are packages for these algorithms. The three types of data mining methods will be explained below and combined with our data mining goals to see why only regression is the appropriate data mining method for this project.

Classification is a supervised learning method. Classification models are suitable for datasets where the response variable is discrete.[7]

Clustering is an unsupervised learning method. Unsupervised learning is a training method of machine learning, which is essentially a statistical method, a training method in which some potential structures can be found in unlabeled data.[7] Clustering focuses on identifying similar data and labeling the data according to the group to which the data corresponds.

Regression is a common data analysis method. Regression models produce continuous response variables.[7] Common regression models include linear regression, generalized linear regression, random forest regression, and lasso regression.

In this project, we will use a variety of models and data mining methods and come up with the most suitable model or models for prediction.

Based on the objectives of data mining,

1. For **classification**, the response variable we want to model and predict is house price (or its logarithm), which is a continuous variable rather than a categorical variable, and if it were a categorical variable, there would be close to the number of rows of data levels, thus making it impossible to interpret the modeling results and make predictions. Therefore, **classification is not a method that meets our data mining goals.**

2. **Clustering** is an unsupervised learning method with no predefined outputs, is suitable when the original dataset does not contain a response variable. We were also unable to verify the predictive accuracy of the clustering, as there was no established criterion. This does not meet our data mining goals, so **clustering is not the data mining method we should choose here.**

3. For **regression**, the response variable here is house price (or its logarithm), which is a continuous variable, regression methods are suitable in this case. So, **I decided to use regression as the data mining method we will use.**

(f) The design of the processes that converts data into insights

For this research, the data came from the R package [3]HRW: Datasets, Functions and Scripts for Semiparametric Regression Supporting Harezlak, Ruppert & Wand (2018). This is an open-source R package that contains a large number of datasets.

The data has 37676 rows and 39 columns. Each row is a record of a property transaction, and each column represents a factor that may affect the price of a property. Having 39 columns means we have 38 explanatory variables and one response variable.

First, let's initially select some data.
Select rows. Each row of data contains a real estate transaction information. Our study is about all Sydney property transactions, and more data leads to a better fit, so I decided to keep data for all rows.

Select columns. All the columns here are potentially valuable variables that may be helpful in predicting real estate prices. Since the data does not contain sensitive customer information, all columns can be preserved here.

In addition, there are some variables that are more suitable to be treated as categorical variables, such as zip code and transaction date. However, when used as categorical variables, they will have too many levels, which is not conducive to us building a stable model, and will also bring trouble to interpret the statistical significance of the model, so I choose to discard them.

There are two columns in the dataset contain missing data. Unfortunately, based on my intuition, they are in two columns that are important for predicting property prices, and simply dropping those two columns is not a good idea.

We could delete all rows with data, but I'd like to try to avoid doing that. Doing so will lose a lot of data, potentially making our model less robust; one of the columns is missing non-randomly (MNAR), and deleting the corresponding row may lead to bias in our final model predictions.[5]

Now, we need to apply an appropriate imputation method for missing values. Using a measure of central tendency for the attribute seems to be a good way, but whether it is mean, median or most frequent has certain limitations. Here, I used IterativeImputer in sklearn to deal with missing values. It estimates each feature from all the others. In the first iteration, I chose the "median" strategy. Run checking again, all missing values have been filled.

To remove outliers, we only kept data within plus or minus 3 standard deviations of the mean. Within the normal distribution assumptions, this would contain 99.73% of the data. Any data outside this is likely to be outliers or extreme values. After removing possible outliers for all variables, the number of rows of the data is reduced.

Then we need to construct the data. The real estate prices in the original data are on the log-scale. This can be useful when modeling, but is less intuitive when interpreting. So, I'm adding a new variable here: SalePrice(thousands of AUD). This variable is obtained by taking the exponent of “logSalePrice” and dividing it by 1000.

Then, the data needs to be formatted. Financial quarter is more suitable to be considered a categorical variable because it is discrete and has only 4 levels. In addition, input and target should also be specified. input refers to the explanatory variable and target refers to the response variable. In this case, Variables except logSalePrice and SalePrice are explanatory variables, logSalePrice and SalePrice are response variables.

Including too many explanatory variables in the model can overfit the model. We need to reduce the dimensionality of the dataset by variable importance.

Feature selection helps identify the most important fields when modeling and predicting outcomes. This is useful when there are many explanatory variables, because not every explanatory variable will be helpful for modeling and prediction, and useless explanatory variables will increase the computational cost and affect the stability of the model.

Here, I used Feature selection in sklearn. According to the method mentioned in Univariate feature selection, SelectKBest removes all features except the K features with the highest score. Since here it is a regression problem, we can use f_regression for Feature selection, which is a Feature selection method based on ANOVA F-value. This method calculates the ANOVA F-value of each feature. The larger the F-value, the greater the relationship between the feature and the prediction, so the feature is more "important". Here, I keep the top 10 important variables.

The relationship between property prices and most of the explanatory variables is the inverse of the log (and its multiples). Also, when plotting against log-scaled price, we find that the data has a near-constant variance, whereas when plotting against price, the variance varies with level. Many statistical models require the assumption of constant variance, so it would be a better idea to use log-scale explanatory variables.[6]

Log-transformed data generally better fit the model assumptions. Reciprocal-shaped distributions are difficult to interpret, whereas linear-like distributions are easier to fit and interpret.[6]

(g) The description of the implementation using various algorithms and enabling technologies

As we discussed earlier, we will use regression as the data mining method. There are many data mining methods for regression, such as regression trees, random forests, Bayesian regression, etc.

Regression trees are basically decision trees for regression tasks that can be used to predict continuous valued outputs instead of discrete outputs.[12] The basic idea behind the algorithm is to find the point in the independent variable, split the dataset into 2 parts so that the mean squared error is minimum at that point. The algorithm does this in an iterative manner, forming a tree-like structure.[12]
Random forests grow many classification trees. To classify new objects in the input vector, place the input vector under each tree in the forest. Each tree is given a classification, and we say the tree "votes" for that class. The forest selects the class with the most votes (amongst all trees in the forest).[13]

There are two types of machine learning: supervised learning and unsupervised learning.
Supervised learning is a machine learning method defined by the use of labeled datasets. These datasets are designed to train or "supervised" algorithms to classify data or accurately predict outcomes. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.[14]
Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns in data without human intervention (thus, they are "unsupervised").[14]

The goal of this research is to study the factors that affect the price of real estate, build a model and use these factors to predict the price of real estate, so the input of the model is some explanatory variables, and the output is the house price. This is a scenario with a clear output target, so we should choose a supervised model.

Specifically, there are some models that are more traditional and straightforward, such as multivariate regression and linear models, which are helpful for our later explanations. There are models that use more modern statistical theory, which may perform better, but also incur greater computational overhead. Here, we will choose some models and proceed to the subsequent modeling steps.

In fact, all models available in Python are our potential models. This includes Regression, Generalized Linear, KNN, SVM, etc. We can fit all the models and compare their predictive performance on the input data, but fitting all the models is computationally expensive.

I decide to fit three models: **Linear model, Multivariate Regression model** and **Random Forest**.

Linear regression is the most traditional one of regression methods, which assumes a linear relationship between the explanatory variable and the response variable. Here, I will choose a single variable as the explanatory variable.

Multivariate Regression can tell the relationship between multiple explanatory variables and a single response variable. Here, I will put all explanatory variables (after the feature selection) into the model.

Random forest belongs to the Bagging algorithm. The general idea is to train multiple weak models and package them to form a strong model. The performance of the strong model is much better than that of a single weak model.[8] In the training phase, random forest uses bootstrap sampling to collect multiple different sub-training datasets from the input training dataset to train multiple different decision trees in turn; in the prediction phase, random forest averages the prediction results of multiple internal decision trees to obtain final result.[8]

For the **linear model**, the response variable is house price (in log-scale) and the explanatory variable will be one of the explanatory variables of the dataset. We already know that the most important variable affecting house prices is longitude. Therefore, I will use longitude as the only explanatory variable in the linear regression model. For linear regression, I used the LinearRegression method from sklearn.linear_model. It is a function in the sklearn package. I wrote a function to do all the steps of the linear model. For the univariate linear model here, the parameter passed in is longitude, not the entire dataset.

For the **Multivariate Regression model**, I used all the explanatory variables previously selected as input, and the output (response) variable being house price (AUD) on log scale. This is obviously a more complex model, with a high probability that it will perform better than the linear model. Compared with the previous univariate linear regression, the only difference is that the parameter X passed into the linear regression function is all explanatory variables here.

Similar to the multivariate regression model, our input variables are all explanatory variables (selected earlier), and the output explanatory variable is the log-scale house price (AUD). We also need to set the number of trees used in the bagging process. I set it to 20 here. Larger values will give more stable results, but the runtime will be greatly increased. In order to get reproducible results, for algorithms with randomness we need to set a seed. Here I set it to 648, which is the last three digits of my UPI.

In statistical modeling practice, some methods of data partitioning are often used. This is based on the Golden rule of machine learning -- the test data cannot influence training the model in any way. [9] Therefore, we need to separate the training and test sets. Models are trained on the training set and evaluated on the test set. Usually we only care about the performance of the model on the test set, because over-optimizing on the training set can lead to overfitting, which can lead to large variance. In supervised learning, overfitting happens when our model captures the noise along with the underlying pattern in data. [10] To a large extent, the process of machine learning is a Bias-Variance Tradeoff process.

Specifically, we can use a 70/30 split, that is, use 70% of the data as the training set and the remaining 30% as the test set. This method is also known as Hold-out. Another popular solution is cross-validation. With higher accuracy and stability, it has a higher computational cost. Cross-validation randomly splits the dataset into "k" groups. One of the groups is used as the test set, and the remaining groups are used as the training set. The model is trained on the training set and scored on the test set. The process is then repeated until each unique group is used as a test set.[11]

Here, I decided to use 5-fold cross-validation, which is to use 20% of the data as the test set each time, and the rest of the data as the training set, and repeat it five times. This is because cross-validation has higher accuracy and stability; with higher folds it will perform better, but the run time will be beyond my tolerance.

For each algorithm, I run it 5 times, using 5-fold cross-validation each time. Both n_splits and n_repeats are 5 here. For each model, I keep 4 performance measures, that is the mean and standard deviation of **R2** and **RMSE**.

R2 is a measure that tells us the proportion of variance in the response variable of the regression model that can be explained by the predictor variables. This value ranges from 0 to 1. The higher the R2 value, the better the model fits the dataset.[15]

RMSE is a metric that tells us the average distance of the predicted value from the observed value in the dataset. The lower the RMSE, the better the model fits the dataset.[15]

The mean value of R2 and RMSE reflects the average performance of the model fitting the test data, and the standard deviation of R2 and RMSE reflects the stability of the model. The smaller the standard deviation, the more stable the model is.

By calling the functions defined earlier, I ran the three algorithms mentioned earlier on the dataset and got some results. Here just to proof that I ran it successfully, and I'll explore these results in the next section.

(h) The interpretation of the patterns and results

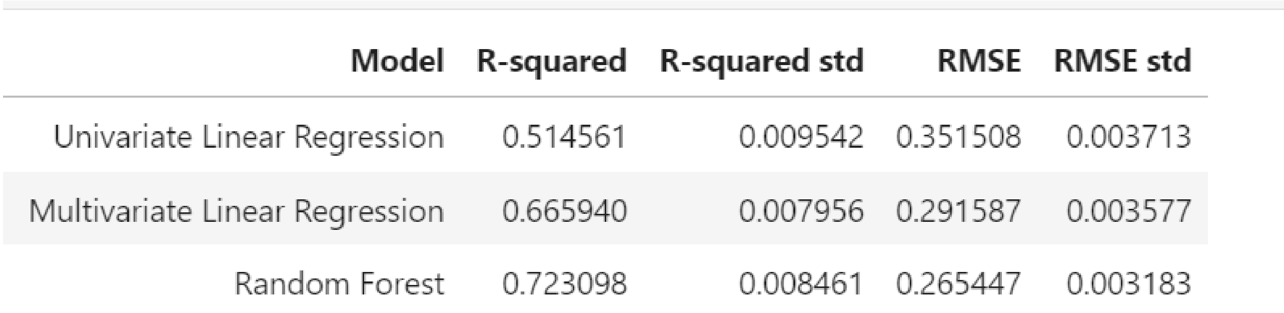
In the previous steps, we identified a supervised method rather than an unsupervised method to achieve our data mining goals, and further confirmed that we used a regression method rather than a classification method.

Random forests have the best predictive performance, while multivariate linear models also perform well and are more interpretable than random forests.

Before fitting a model, it is necessary to determine whether the data conform to the statistical assumptions of the model. Linear regression models assume that the relationship between the response variable and the explanatory variable is linear, and the residuals are independent, with a constant variance. We observe that the relationship between property prices and some important explanatory variables is not linear, and the variance is obviously not constant, but increasing with the increase of levels. This does not meet the assumptions of linear regression, so we need to convert prices to log scale.

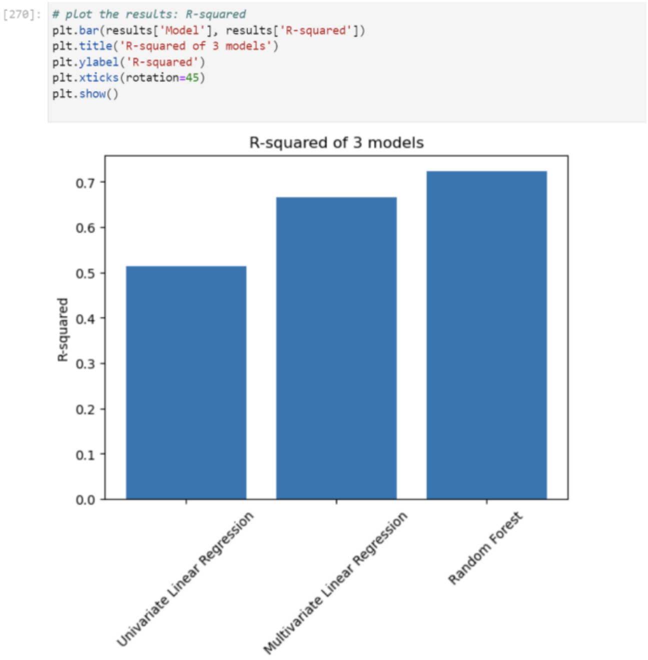
Finally, we used the mean of R2 and RMSE to estimate the average predictive performance of the models, and the standard deviation of R2 and RMSE to estimate the stability of the models. The final conclusion is that univariate linear regression has the worst performance and random forest has the best performance.

Below are the performance of the three models.



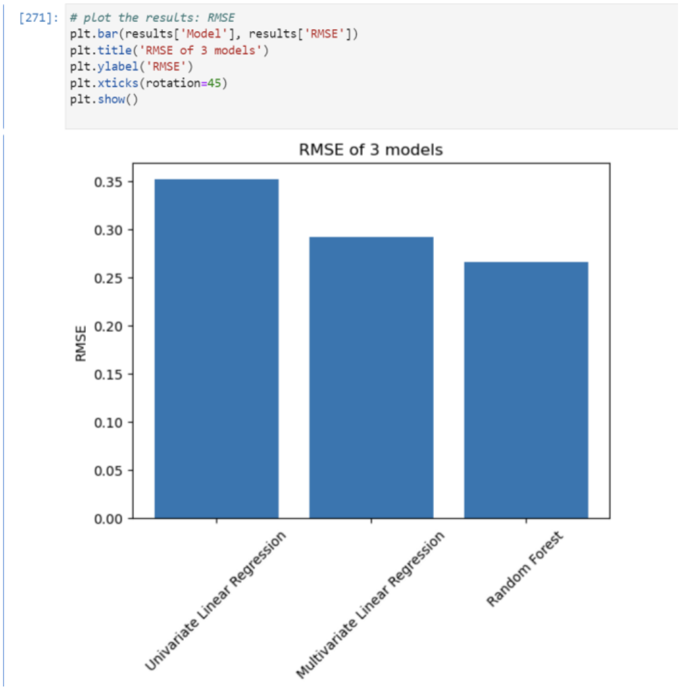
Model	R-squared	R-squared std	RMSE	RMSE std
Univariate Linear Regression	0.514561	0.009542	0.351508	0.003713
Multivariate Linear Regression	0.665940	0.007956	0.291587	0.003577
Random Forest	0.723098	0.008461	0.265447	0.003183

Figure1: Model performance



```
[270]: # plot the results: R-squared
       plt.bar(results['Model'], results['R-squared'])
       plt.title('R-squared of 3 models')
       plt.ylabel('R-squared')
       plt.xticks(rotation=45)
       plt.show()
```

Figure2: R2 of 3 models



```
[271]: # plot the results: RMSE
       plt.bar(results['Model'], results['RMSE'])
       plt.title('RMSE of 3 models')
       plt.ylabel('RMSE')
       plt.xticks(rotation=45)
       plt.show()
```

Figure3: RMSE of 3 models

From the results, in terms of model stability, the standard deviation of the indicators of the univariate linear regression model is slightly larger, so its stability is slightly worse than the other two models. In terms of average prediction performance, the univariate linear regression model has the smallest R2 and the largest RMSE, which means its performance is the worst. The random forest has the largest R2 and the smallest RMSE, which means that its performance is the best among the three models. The performance of multivariate linear models is somewhere in between.

These results are not surprising: only one explanatory variable is used in the linear model, resulting in a model that is too simplistic and has a large bias. This is a typical underfitting model, which will not perform well on either the training set or the test set. Both multivariate regression models and random forests use the full dataset as explanatory variables and thus have much better predictive performance. Being a more complex algorithm, random forest has higher prediction performance than the other two, but also has the longest running time.

In the predicted value -- actual value plot of the linear model, the linear relationship between the two is not very obvious. The scattered distribution of data points is shown in the plot.

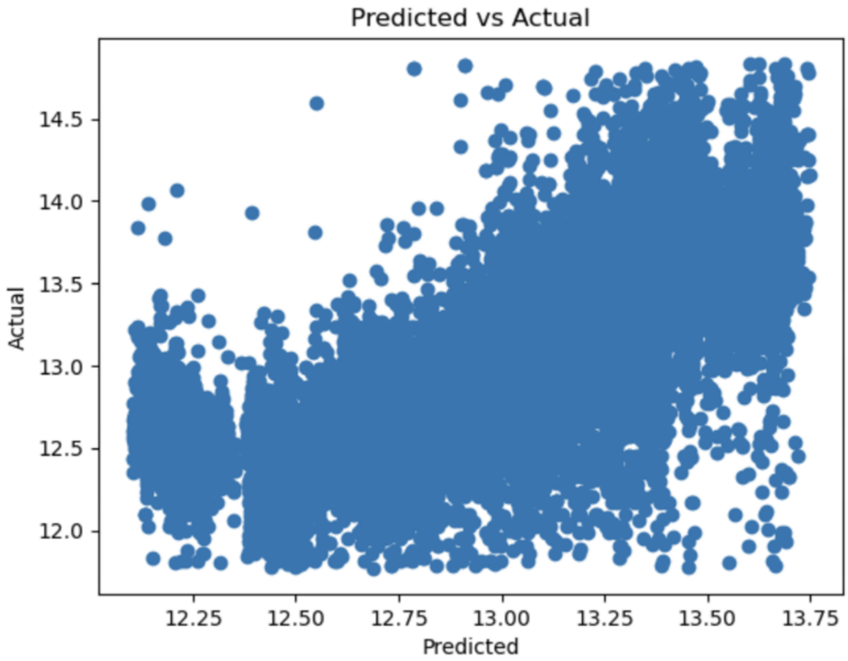


Figure4: Performance of Linear Model

In the plot of the multivariate regression model, a clear linear relationship can be observed.

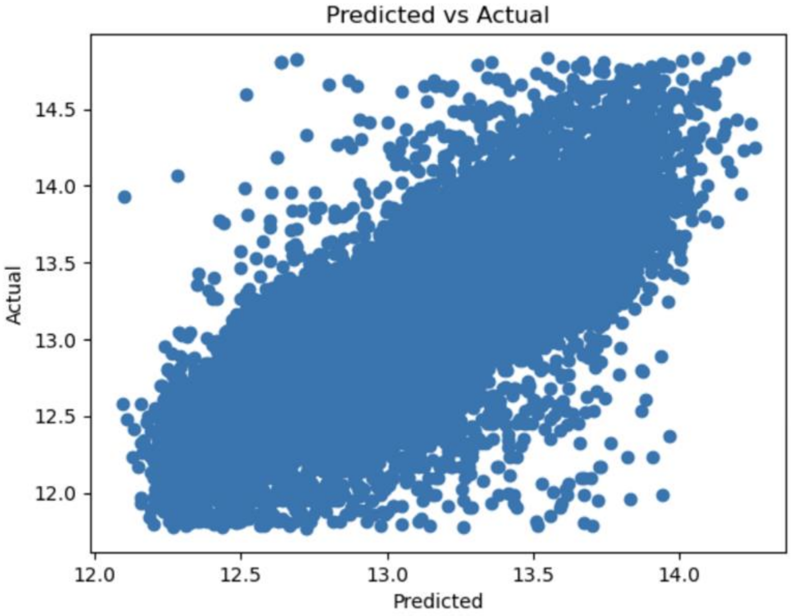


Figure5: Performance of Multi-regression Model

For random forest model, the linear relationship between the actual value and the predicted value is the most obvious among the three models.

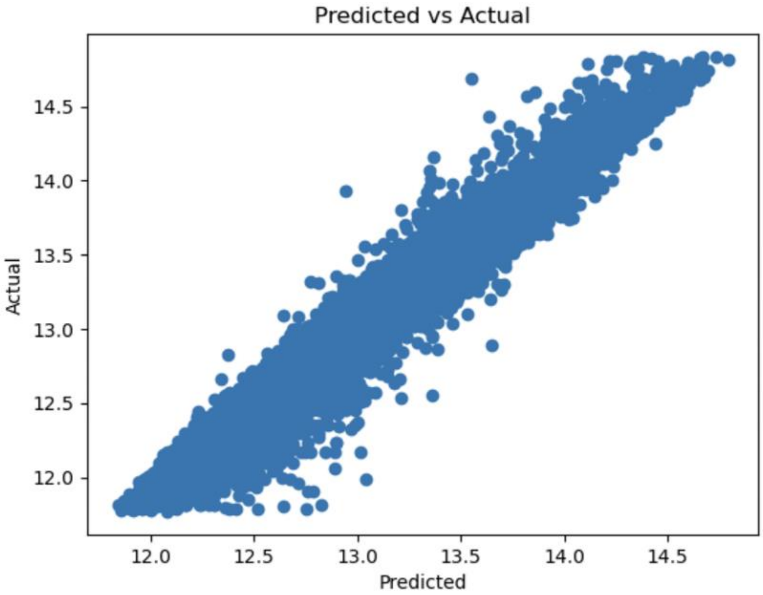


Figure6: Performance of RF Model

Random forest gives the importance ranking of explanatory variables.

For housing prices, the top four most important influencing factors are longitude, income, lotSize and distToCoastline. This is slightly different from the previous results get using SelectKBest.

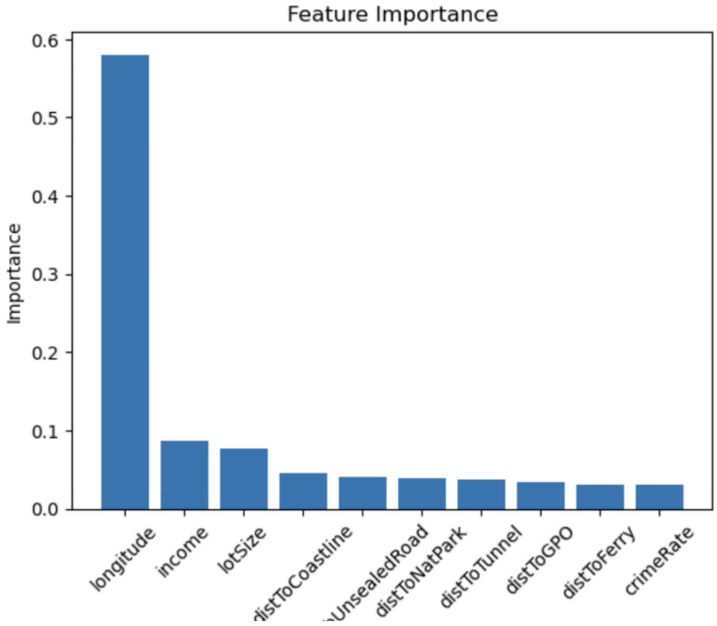


Figure7: Feature Importance from RF

Some variables are positively correlated with the logarithm of property prices, such as income level; others are negatively correlated with the logarithm of property prices, such as the distance from the property to the coastline.

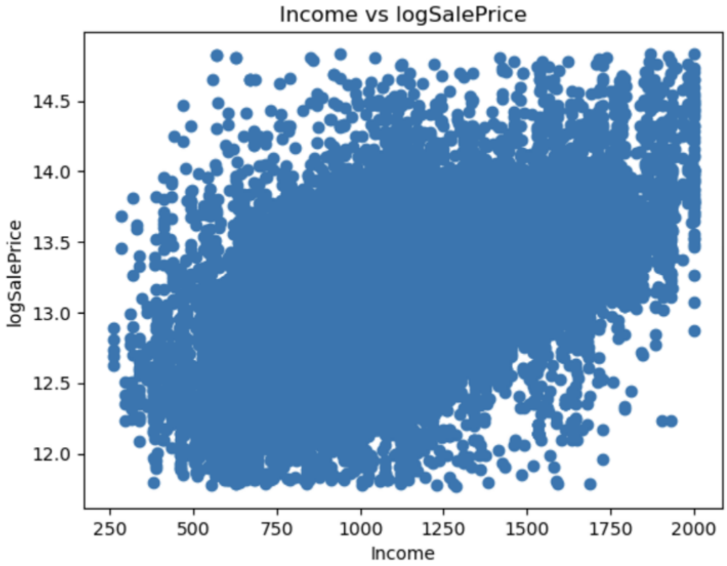


Figure8: Positive Correlation

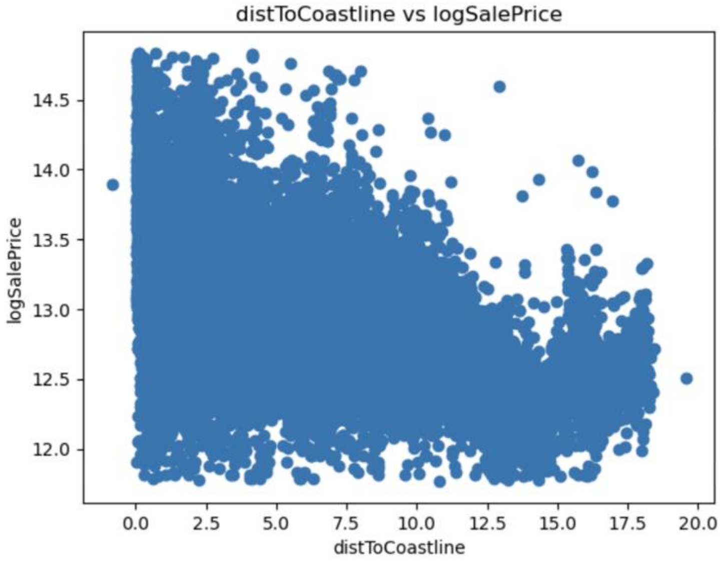


Figure9: Negative Correlation

I would expect a very clear positive correlation between longitude and property prices as it is considered the most important influencer on property prices in Sydney.

The result surprised me a little, it seems that the relationship between longitude and property prices is not monotonically positive: properties with a longitude of 150.75 appear to have the lowest prices.

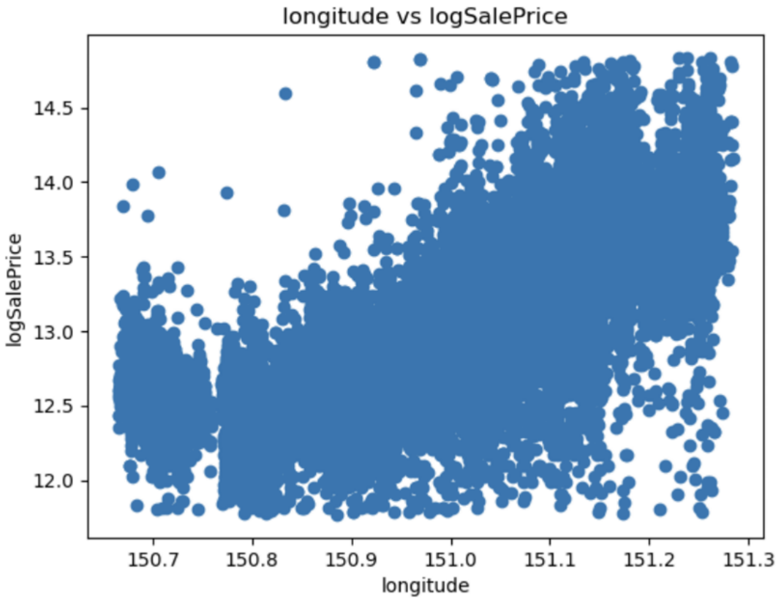
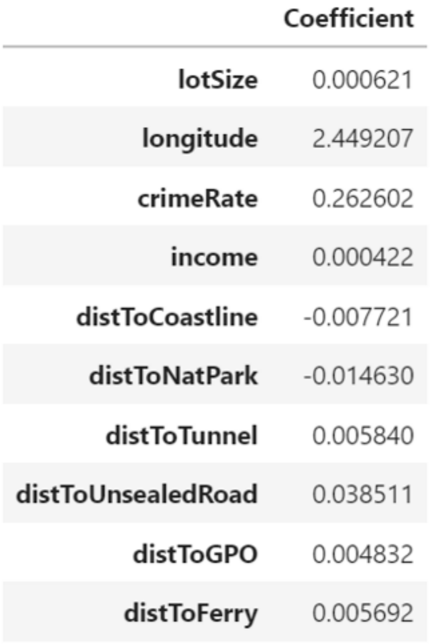


Figure10: Longitude vs logSalePrice

Specifically, when the longitude is less than 150.75, the more eastward the house price is lower; when the longitude is greater than 150.75, the more eastward the house price is higher.

The model parameters of the **Multivariate Regression model** are as follows:



	Coefficient
lotSize	0.000621
longitude	2.449207
crimeRate	0.262602
income	0.000422
distToCoastline	-0.007721
distToNatPark	-0.014630
distToTunnel	0.005840
distToUnsealedRoad	0.038511
distToGPO	0.004832
distToFerry	0.005692

Figure11: Multivariate Regression Model Parameters

The parameters of longitude have changed compared to previous model, because here we take all explanatory variables into the model. Some parameters can be difficult to interpret because not all parameters fit the assumption of a linear relationship with property prices on log-scale. For example, if you want to explain the relationship between lot size and house price, you can say that log(Price) increases by 0.000621 for every 1 square meter increase in lot size, that is, property price multiplied by exp(0.000621) = 1.00062119. This is obviously problematic because the relationship between property price and lot size can't be exponential. After transforming the response variable to log-scale, the model is more predictive, but less interpretable.

However, we can still see some trends from the parameter list. When the coefficient of an explanatory variable is greater than 0, it can be said to be positively correlated with the response variable, such as lot size and income; when the coefficient of an explanatory variable is less than 0, it can be said to be negatively correlated with the response variable, such as the distance to the coastline. This is not surprising, intuitively, houses with larger lot sizes are more expensive; people with higher incomes tend to buy more expensive houses; people are more Prefer to live closer to the coastline, so the farther from the coastline, the cheaper the property. The relationship between longitude and house prices may be explained by the location and coastline of Sydney city. The larger the longitude, the further east. Since Sydney is on the east coast of Australia, so the larger the longitude, the closer it is to the sea.

We fit a total of three models: linear model, multiple regression model and random forest. The performance of the linear model is the worst, with severe underfitting. In actual forecasting, we cannot use such a model. Multiple regression and random forests perform similarly, with random forests slightly better than multiple regression models. However, the running time of the random forest model is significantly longer, and the cost of slightly better prediction performance is obvious. It is reasonable to believe that random forests will achieve better predictive performance if the number of bagging trees is increased. In actual data mining practice, data mining goals should be fully considered, and whether more complex, longer-running, and harder-to-interpret models should be used for small improvements. My answer tends to be no.

From the parameter coefficient of the multivariate linear model, the factors that lead to higher real estate prices are:
--larger lot size
-- Higher average weekly income of the suburb in which the house is located
-- larger longitude
. . . . . .
Factors that lead to lower property prices include:
-- farther from the coastline
-- farther from the national parks
. . . . . .

The multiple linear model also gave us some incredible conclusions, such as the higher the crime rate around the property, the higher the property price. A possible explanation is that the more expensive homes are located in the city center, where has a higher crime rate.

(i) The proposed actions based on the discovered knowledge

In general, Sydney real estate is of high quality, covers a large area and is relatively expensive.

According to the conclusions mentioned earlier, the most important factor affecting the price of a property is the longitude: the more east the property is, the more expensive it is. This essentially means that properties by the ocean will be more expensive because people prefer ocean views. For developers, a good strategy would be to find a way to acquire land on Sydney's east coast close to the sea, so that the built houses can easily be sold at high prices. Of course, the price of the land also needs to be considered. For home buyers, they can decide what location to buy based on their economic strength and preference for the sea, if he is ready to buy a property in Sydney.

Income is also an important factor affecting property prices, if a home buyer has a high income, he tends to buy a home in a wealthy neighborhood. Therefore, for real estate agents and developers, focusing on those high-income groups will bring them more revenue per customer.

Contrary to people's intuition, the relationship between floor space and property prices is not as obvious as one might think. People sometimes sacrifice size in order to buy a house in a better location.

From the government's point of view, we need to do some considerations in the public interest. For example, if the City of Sydney or the Australian government needs to arrange low-rent housing for the homeless living in Sydney, it should not consider houses on the east coast, because the funds are limited. In addition, for houses near the sea and in wealthy neighborhoods, a higher rate of property tax can be considered. For those neighborhoods in the city center with small footprint and expensive, high crime rate, consider increasing the police force to curb crime.

Sustainable cities are our common goal. With the continuous advancement of urbanization today, it is particularly important to solve the housing problem of urban residents. I hope this paper of mine can shed some light on this topic.

References

[Aut], J. D. H. R. (2021a, November 23). *HRW: Datasets, Functions and Scripts for Semiparametric Regression Supporting Harezlak, Ruppert & Wand (2018)*. R Package Documentation. https://rdrr.io/cran/HRW/

[Aut], J. D. H. R. (2021b, November 23). *Sydney real estate*. R Package Documentation. https://rdrr.io/cran/HRW/man/SydneyRealEstate.html

Realestate, E. (2021, October 20). *Factors That Affect Housing Prices in Sydney*. Etch Real Estate. https://www.etchrealestate.com.au/factors-that-affect-housing-prices-in-sydney/

Sustainable cities and human settlements | Department of Economic and Social Affairs. (2018). United Nations. https://sdgs.un.org/topics/sustainable-cities-and-human-settlements

Yee, T. (Ed.). (2022). Introduction to Data Mining. In *STATS 784: Statistical Data Mining* (pp. 60–67).

Cialdella, L. (2020, August 30). *When do we log transform the response variable? Model assumptions, multiplicative combinations and log-linear models*. Casual Inference. https://lmc2179.github.io/posts/multiplicative.html

Agrawal, P., Gupta, C., Sharma, A., Madaan, V., & Joshi, N. (2022). *Machine Learning and Data Science: Fundamentals and Applications* (1st ed.). Wiley-Scrivener.

Education, I. C. (2021, January 26). *Random Forest*. IBM. https://www.ibm.com/cloud/learn/random-forest

Machine Learning and Data Mining. (2016). In *CPSC 340 UBC* (p. 19).

Singh, S. (2022, February 18). *Understanding the Bias-Variance Tradeoff - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Allibhai, E. (2022, June 21). *Hold-out vs. Cross-validation in Machine Learning - Eijaz Allibhai*. Medium. https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

Prasad, A. (2022, January 6). Regression Trees | Decision Tree for Regression | Machine Learning. Medium. Retrieved September 23, 2022, from https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047

Random forests - classification description. (n.d.). Retrieved September 23, 2022, from https://www.stat.berkeley.edu/%7Ebreiman/RandomForests/cc_home.htm

Supervised vs. Unsupervised Learning: What’s the Difference? (2021, March 12). IBM. Retrieved September 23, 2022, from https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Gradient boosting. (2022, September 1). Wikipedia. Retrieved September 23, 2022, from https://en.wikipedia.org/wiki/Gradient_boosting

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright.

(See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."