Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia

The Danh Phan
Macquarie University
Sydney, Australia
danh.phan-the@students.mq.edu.au

***Abstract*—House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in Australia to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Melbourne. Moreover, experiments demonstrate that the combination of Stepwise and Support Vector Machine that is based on mean squared error measurement is a competitive approach.**

Keywords—House price prediction, Regression Trees, Neural Network, Support Vector Machine, Stepwise, Principal Component Analysis

I. Introduction

Buying a house is undoubtedly one of the most important decisions one makes in his life. The price of a house may depend on a wide variety of factors ranging from the house's location, its features, as well as the property demand and supply in the real estate market. The housing market is also one crucial element of the national economy. Therefore, forecasting housing values is not only beneficial for buyers, but also for real estate agents and economic professionals.

Studies on housing market forecasting investigate the house values, growth trend, and its relationships with various factors. The improvement of machine learning techniques and the proliferation of data or big data available have paved the way for real estate studies in recent years. There is a variety of research leveraging statistical learning methods to investigate the housing market. In these studies, the most popular investigated locations are the United States [1], [2], [3], [4], [5], [6]; Europe [7], [8], [9]; as well as China [10], [11], [12], [13]; and Taiwan [14], [15]. However, research on the housing market by applying data analytics with machine learning algorithms in Australia is rare, or elusive to find.

The goal of this study is through analyzing a real historical transactional dataset to derive valuable insight into the housing market in Melbourne city. It seeks useful models to predict the value of a house given a set of its characteristics. Effective models could allow home buyers, or real estate agents to make better decisions. Moreover, it could benefit the projection of future house prices and the policymaking process for the real estate market.

The study follows the cross-industry standard process for data mining, known as CRISP-DM [16]. This is a commonly used approach for tackling data analytics problems. The next parts of the paper are constructed as follows: Section 2 reviews previous work on housing market forecasting applying different machine learning techniques. Section 3 explains the dataset and how to transform it into cleaned data. In section 4, various machine learning methodologies are proposed. Model implementation and evaluation will be discussed in Section 5, and the conclusion is deduced in Section 6.

II. Related Work

Previous studies on the real estate market using machine learning approaches can be categorized into two groups: the trend forecasting of house price index, and house price valuation. Literature review indicates that studies in the former category deem predominant.

In the house growth forecasting, researchers try to find optimal solutions to predict the movement of housing market using historical growth rates or house price indices, which are often calculated from a national average house price [3], [5], [6], or the median house price [4]. [1] contends that house growth forecasting could act as a leading indicator for policymakers to assess the overall economy. Factors that affect house price growth tend to be macroeconomic features such as income, credit, interest rates, and construction costs. In these papers, Vector Auto regression (VAR) model was commonly applied in earlier periods [10], [11], while Dynamic Model Averaging (DMA) has become more popular in recent years [3], [13], [14].

On the other hand, house price valuation studies focus on the estimation of house values [2], [12], [15]. These studies seek useful models to predict the house price given its characteristics like location, land size, and the number of rooms. Support Vector Machine (SVM) and its combination with other techniques have been commonly adopted for house value prediction. For instance, [12] integrates SVM with Genetic Algorithm to improve accuracy, while [15] combines SVM and Stepwise to effectively project house prices. Furthermore, other methods such as Neural Network and Partial Least Squares (PLS) are also employed for house values prediction [2].

It is underscored that Neural Network (NN) and SVM has recently been applied in a wide variety of applications across numerous industries. Neural Networks has been further developed to become deep networks or Deep learning method. Besides, the advance of SVM deems to achieve by integrating it with other algorithms. For example, Principal component

978-1-7281-0404-1/19/$31.00 ©2019 IEEE
DOI 10.1109/iCMLDE.2018.00017

analysis (PCA) is combined with SVM to address prediction issues in different industries such as Information security [17], Stock Market [18], or Industrial processes [19]. In addition, the combination of Stepwise and SVM is widely used for Credit Scoring [20], Faulty detection of Semiconductor Producing [21], or Dimension Reduction of High-Dimensional Datasets [22]. Both NN and SVM methods will be implemented and discussed in this paper.

III. Data Preparation and Exploration

1. *Original Data*

The data implemented in this study is the Melbourne Housing Market dataset downloaded from Kaggle website [23]. The original dataset has 34,857 observations and 21 variables. Each observation presents a real sold house transaction in the city of Melbourne from 2016 to 2018. These variables can be categorized into 3 groups:

- Transactional variables include Price, Date, Method, Seller, Property count.
- Related location predictors which contain Address, Suburbs, Distance to CBD, Postcode, Building Area, Council Area, Region name, Longitude, Latitude
- Other house features such as House Type, Number of Bedrooms, Number of Bathrooms, Number of Car slots, and Land size

The outcome of house value prediction is the price which is a continuous value, and predictors consist of other features with both numeric and categorical types.

2. *Data preparation*

Before applying models for house price prediction, the dataset needs to be pre-processed. The investigation of missing data is at first performed. Several missing patterns are assessed rigorously since they play an important role in deciding suitable methods for handling missing data [24].

Columns with more than 55% values missing are removed from the original dataset since it is difficult to impute these missing values with an acceptable level of accuracy. In addition, there are many rows with missing values of the outcome variable (Price). Since the imputation of these values could increase bias in input data, observations with missing values of the Price column are deleted.

In addition, imputation is performed for other predictors with a small portion of missing values. Longitude and Latitude, for instance, are imputed from house addresses using a Google map's Application Programming Interface (API). Another example is the imputation of Land size values by using its median values group by house types and suburbs.

Furthermore, outliers are also discovered and addressed. Outliers are defined as an observation which seems to be inconsistent with the remainder of the dataset [25]. Outliers may stem from factors such as human errors, relationship with probability models, and even structured situations [25]. For instance, land sizes of less than 10 square meters are removed.

As a result, the cleaned data, which is used to build and evaluate models, has 11 variables (in Table I) and more than 20 thousand observations.

Table I. Features description

Name	Type	Description
Price	Numerical	House price (prediction outcome)
Year	Numerical	Sold year: 2016-2018
Property Count	Numerical	Number of properties
Distance	Numerical	Distance to CBD
Longitude	Numerical	House's Longitude
Latitude	Numerical	House's Latitude
Rooms	Numerical	Number of bedrooms
Bathrooms	Numerical	Number of bathrooms
Car	Numerical	Number of car spots
Land size	Numerical	House's land size
Type	Categorical	House's type: u-unit, h-house, t-townhouse

3. *Descriptive exploration*

This section only presents the most important findings. The data summary information and other informative figures are allocated in the Appendix.

Descriptive analysis indicates that a median house has three bedrooms, one bathroom, with land size above 500 square meters. Its median price is roughly 900 thousand dollars. Fig. 1 and 2 indicate the histograms of Price and log(Price). While the range of Price values varies widely with a long tail, log(Price) deems to have a normal distribution. Thus, log(Price) will be used as the output in model building and evaluation phases.

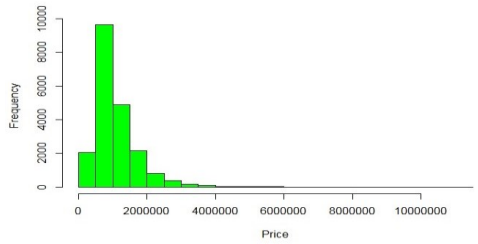

Fig. 1. Histogram of Price

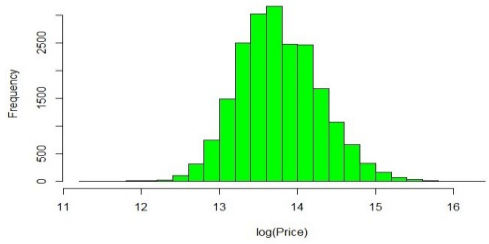

Fig. 2. Histogram of log(Price)

The list of suburbs with the most expensive and cheapest median house prices are demonstrated in Figs 3 and 4, respectively.

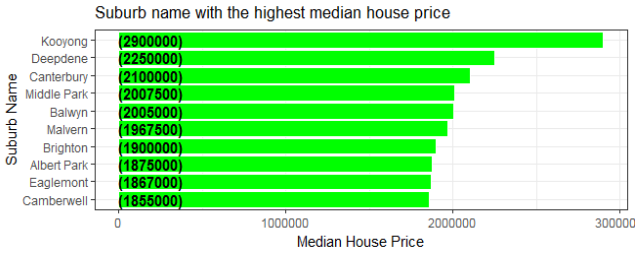

Fig. 3. Suburbs with most expensive houses

Kooyong, in Fig. 3, is the suburb with the highest median price at nearly three million dollars, and one may spend around two million dollars to buy a house in other expensive suburbs.

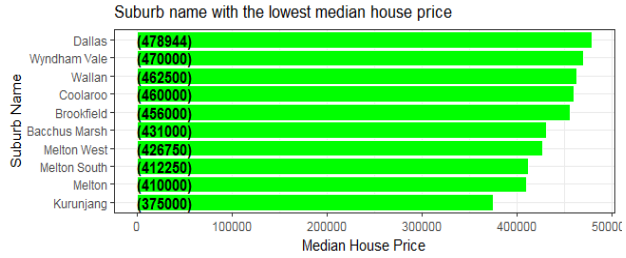

Fig. 4. Suburbs with cheapest houses

On the other hand, in Fig. 4 with cheap suburbs such as Kurunjang, Melton, and Melton South, one can own a property with about 400 thousand dollars. Other affordable suburbs have the median house price of fewer than 500 thousand dollars. Hence, the difference in median house prices among low and high price suburbs is significant, which varies from around four to six folds.

IV. Methodologies

Data reduction and transformation

In order to improve the interpretability and enhance the performance of prediction models, data reduction techniques like Stepwise and Boosting are exploited to derive the most important predictors. Moreover, PCA, a data transformation technique, is applied to get significant components in order to integrate with SVM.

Model selection and evaluation

The paper implements different regression models to find the useful ones.

An attribute subset from Stepwise will be inputted in Linear Regression, Polynomial Regression, Regression Tree, as well as Neural Network, and SVM. In addition, SVM is also integrated with PCA to compare the accuracy of its integration with Stepwise.

Linear regression will be used as a baseline for model evaluation, which based on Mean Squared Error (MSE) measured on an evaluation dataset. MSE is the most popular tool to measure the quality of fit [26]. It is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \quad (1)$$

In this formulation, n is the number of observations, while $\hat{f}(x_i)$ is the prediction of i^{th} observation.

Before fitting data into models, the cleaned dataset is divided into train and evaluation data. The evaluation set will be kept isolated from model building, and only used for model evaluation. The model fitting process utilizes train data using ten folds cross-validation. It is noted that cross-validation is applied in both data reduction and model construction stages.

The next subsections will introduce several important machine learning techniques utilized in this study.

1. Stepwise

Stepwise is one common-used method for subset selection. It is an improved technique of Best Subset Selection [26] which trains a least squares regression for 2p possible models of p predictors. In this study, we use forward stepwise selection [26] which only involves fitting $(1+p(p+1)/2)$ models. Fig. 5 indicates the important level of predictors for the outcome log(price).

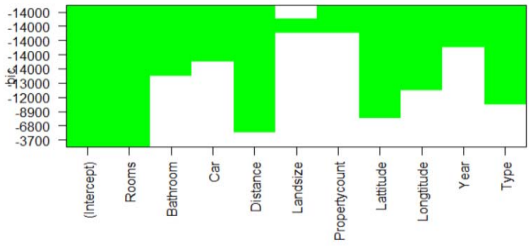

Fig. 5. Feature importance scores

Five most important variables related to the outcome variable are derived from cross-validation results. These predictors comprise of Rooms, Distance, Latitude, Longitude, and Type of houses. Interestingly, the land size contributes an insignificant portion of house prices. There is also an insignificant influence of the number of car spots and the year when the house was sold. Moreover, the Boosting method produces similar feature importance list which confirms the reliability of these five predictors. The detailed importance scores of predictors in Boosting are shown in Table II.

Table II. Feature Importance Scores in Boosting

Predictor	Importance
Type	27.3600
Latitude	20.7183
Distance	17.8677
Rooms	15.0577
Longitude	14.3742
Bathrooms	4.5816
Land size	0.0405

2. *Principal component analysis*

Principal component analysis (PCA), which is an unsupervised approach, can be utilized for data reduction. PCA allows us to create a low-dimensional representation of data that captures as much of the feature variation as possible [26]. It can assist in the improvement of SVM performance.

After implementing PCA for the train data using cross-validation, the first six components are extracted for further analysis since they account for nearly 80% of all predictors' variance. Fig. 6 demonstrates the scree plot of the cumulative proportion of explained variance along with the number of principal components.

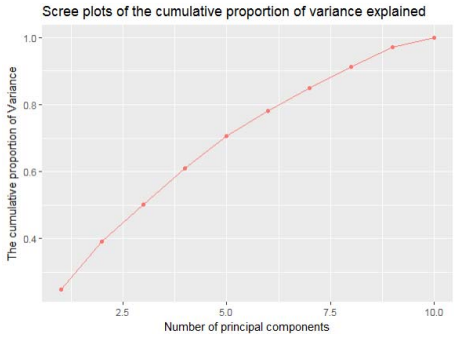

Fig. 6. The cumulative proportion of variance

3. *Polynomial Regression*

Polynomial regression is a standard extension of linear regression [26].

From a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

To a polynomial formation with *d* degree:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d + \epsilon_i \tag{3}$$

The degree *d* in Polynomial regression is often less than five since when degree becomes larger, the polynomial model tends to be over-flexible [26]. Therefore, Polynomial Regression models are implemented using cross-validation with the degree *d* varies from one to five. Fig. 7 indicates that three is the optimal degree for Polynomial Regression.

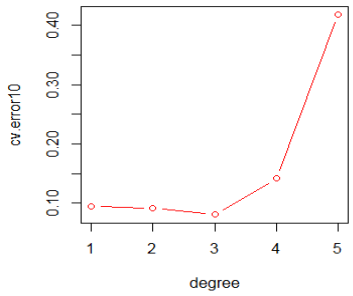

Fig. 7. The plot of polynomial degree and MSE

4. *Regression Trees*

Decision Trees is a widely known methodology for classification; and Regression Trees, which use for continuous outcome prediction, is a special case of Decision Trees. Each leaf contains the prediction value which is the mean of prices of all observations in that leaf. The feature selection as a node in a Regression Tree will be based on the goal of minimizing the Residual Sum of Squares (RSS) [26].

$$\text{RSS} = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}(R_j))^2 \tag{4}$$

The tree induction is implemented with ten folds cross-validation to get minimum RSS with a tree size of twelve. It then is pruned to derive an optimal tree, as shown in Fig. 8.

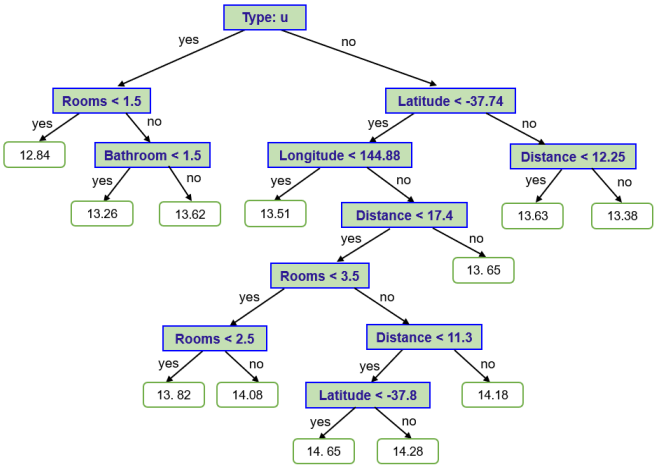

Fig. 8. A pruned regression tree

5. *Neural Network*

Neural Network is the methodology which has widely deployed in many real-world systems. The idea of a neural network is a connected network of nodes or units with related weights and bias [27]. These units are confined into different layers. A neural network normally has one input layer, one output layer, and one or more hidden layers. The complexity arises when the number of hidden layers and/or the numbers of units in each layer increases.

The network learns by adjusting the weights to reduce the prediction error [27]. Initially, all weights and bias are allocated randomly. The algorithm then runs iteratively, and each iteration comprises two steps: forward feeding and backpropagation.

In the forward feeding phase, the output of each unit is calculated from outputs of nodes from the previous layer, as depicted in Fig. 9. The prediction of the output layer is then compared to the observed outcome to derive the learning rate and errors.

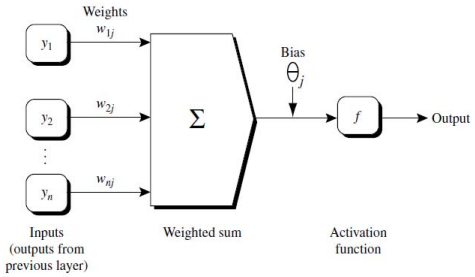

Fig. 9. A neural network unit [27]

In backpropagation, given the learning rate and errors, it recalculates the weights and bias in hidden layers and makes appropriate changes to reduce prediction errors.

In this research, different neural networks are tested with one to three hidden layers. Results demonstrate that the neural network with two hidden layers indicated in Fig. 10, has the smallest Mean Squared Error.

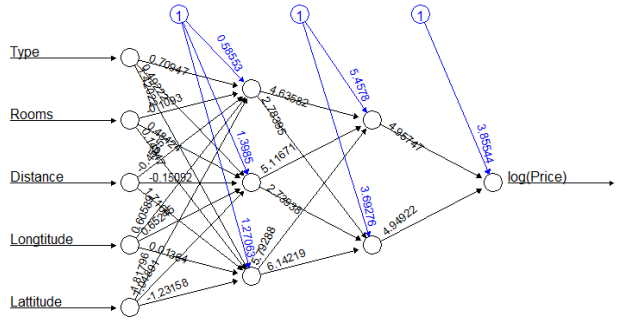

Fig. 10. A 2-hidden-layer neural network.

6. *Support Vector Machine*

Support Vector Machine (SVM) is a powerful technique for supervised learning. SVM algorithm transforms the original data into a high dimension to seek a hyperplane for data segregation [27]. The hyperplane is established by "essential training tuples" which are called support vectors. In comparison with other models, SVM tends to deliver better accuracy due to its ability of fitting nonlinear boundary [27].

SVM models are implemented with two different sets of input variables. The first one stems from five most important features of Stepwise subset selection. The second input set is the six components from PCA transformation.

These are four basic kernels in SVM including linear, polynomial, radial basis function (RBF), and sigmoid. RBF kernel is selected since the number of variables is not large, and RBF deems to suitable with regression problems [28].

The selection of other parameters at first is arbitrary, with Cost of 10 and Gamma of 0.1. Tuned functions are then performed to get the best parameters. TABLE III and TABLE IV show the detailed information of these parameters.

TABLE III. SVM WITH STEPWISE PARAMETERS

Parameters	SVM	Tuned SVM
Cost	10	1
Gamma	0.1	1
The number of support vectors	12842	12599

TABLE IV. SVM WITH PCA PARAMETERS

Parameters	SVM	Tuned SVM
Cost	10	1
Gamma	0.1	1
The number of support vectors	13402	13032

V. RESULTS

The experiments have been deployed in R language on a Window system. The Mean Squared Error (MSE) of both train and evaluation datasets are presented in TABLE V. As in the previous discussion, linear regression will act as the baseline for model comparison. The evaluation ratio of each model is equal to its evaluation MSE divides to the evaluation MSE of Linear regression. The smaller evaluation ratio, the higher accuracy of the model's prediction.

TABLE V. PREDICTION RESULTS

Model	Train MSE	Eval. MSE	Eval. Ratio
Linear regression	0.0948	0.0994	**1.00**
Polynomial regression	0.0773	0.0832	**0.84**
Regression tree	0.0925	0.0985	**0.99**
Neural Network	0.2657	0.2749	**2.77**
Stepwise & SVM	0.0558	0.0615	**0.62**
Stepwise & tuned SVM	0.0480	0.0561	**0.56**
PCA & SVM	0.0721	0.0810	**0.82**
PCA & tuned SVM	0.0474	0.0728	**0.74**

It can be seen from TABLE V that Regression tree delivers a prediction result as good as linear regression, while Polynomial regression results in lower errors which is acceptable. Furthermore, Neural Network seems not to work effectively with this dataset. This may not represent the efficacy of modern deep learning methods.

In addition, PCA and tuned SVM deliver a relatively higher accuracy. However, there is an over-fitting issue in PCA and tuned SVM case, since its evaluation MSE increases significantly in compared with train MSE. The combination of Stepwise and tuned SVM, which produces the lowest error on this dataset, is the most competitive models.

Regarding the model's performance, when the complexity of models increases, the model fitting time also goes up. While Linear and Polynomial regression deliver results instantly, other models could take longer durations, which are indicated in TABLE VI.

TABLE VI. FITTING MODEL RUNTIME

Model	Time (min.)
Regression tree	0.033
Neural Network	0.033
Stepwise & SVM	1.583
Stepwise & tuned SVM	1.400
PCA & SVM	2.317
PCA & tuned SVM	2.733

In comparison with SVM, Regression tree and Neural Network are relatively faster. Therefore, there is a trade-off between modes' runtime and prediction accuracy. It is also underlined that PCA and SVM spend more training time than Stepwise and SVM. Thus, in this case, Stepwise seems more efficient when combining with SVM than PCA.

In terms of interpretability, it is easy to explain the prediction results in simple models such as Linear regression, Polynomial regression, and Decision tree. For instance, we can get the coefficients of related features in Polynomial function, while using decision tree for explanation is simple. In contrast, it will be more difficult to interpret the prediction outcome in Neural Network and SVM. These models run like "black boxes", and we do not know the relationship among predictors and the price prediction.

For further investigation, it is suggested to deploy two models: Stepwise - SVM, and Polynomial regression to predict observations with no outcome values. Polynomial regression can act as a new baseline for comparing prediction results. This implementation should be rigorously tested on historical datasets from different cities in Australia. The results could help to improve the performance and accuracy of these models.

VI. CONCLUSION

In summary, this paper seeks useful models for house price prediction. It also provides insights into the Melbourne Housing Market. Firstly, the original data is prepared and transformed into a cleaned dataset ready for analysis. Data reduction and transformation are then applied by using Stepwise and PCA techniques. Different methods are then implemented and evaluated to achieve an optimal solution. The evaluation phase indicates that the combination of Step-wise and SVM model is a competitive approach. Therefore, it could be used for further deployment. This research can also be applied for transactional datasets of the housing market from different locations across Australia.

References

[1] Gupta, R., Kabundi, A., & Miller, S. M. (2011). Forecasting the US real house price index: Structural and non-structural models with and without fundamentals. Economic Modelling, 28(4), 2013-2021.

[2] Mu, J., Wu, F. & Zhang, A., 2014. Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, 2014(2014), p.7.

[3] Bork L. & Moller S., 2015. Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. International Journal of Forecasting, 31(1), pp.63–78.

[4] Balcilar, M., Gupta, R., & Miller, S. M. (2015). The out-of-sample forecasting performance of nonlinear models of regional housing prices in the US. Applied Economics, 47(22), 2259-2277.

[5] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.

[6] Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the US real house price index. Economic Modelling, 45, 259-267.

[7] Ng, A., & Deisenroth, M. (2015). Machine learning for a London housing price prediction mobile application. Technical Report, June 2015, Imperial College, London, UK.

[8] Rahal, C. (2015). House Price Forecasts with Factor Combinations (No. 15-05).

[9] Risse M. & Kern M., 2016. Forecasting house-price growth in the Euro area with dynamic model averaging. North American Journal of Economics and Finance, 38, pp.70–85.

[10] Jie, T. J. Z. (2005). What pushes up the house price: Evidence from Shanghai [J]. World Economy, 5, 005.

[11] Changrong, X. K. M. Y. D. (2010). Volatility Clustering and Short-term Forecast of China House Price [J]. Chinese Journal of Management, 6, 024.

[12] Gu J., Zhu M. & Jiang L., 2011. Housing price forecasting based on genetic algorithm and support vector machine. Expert Systems with Applications, 38(4), pp.3383–3386.

[13] Wei Y. & Cao Y., 2017. Forecasting house prices using dynamic model averaging approach: Evidence from China. Economic Modelling, 61, pp.147–155.

[14] Chen, P. F., Chien, M. S., & Lee, C. C. (2011). Dynamic modeling of regional house price diffusion in Taiwan. Journal of Housing Economics, 20(4), 315-332.

[15] Chen, J.-H. et al., 2017. Forecasting spatial dynamics of the housing market using Support Vector Machine. International Journal of Strategic Property Management, 21(3), pp.273–283.

[16] Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39).

[17] Ahmad, I., Hussain, M., Alghamdi, A., & Alelaiwi, A. (2014). Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components. Neural computing and applications, 24(7-8), 1671-1682.

[18] Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. Procedia computer science, 31, 406-412.

[19] Jing, C., & Hou, J. (2015). SVM and PCA based fault classification approaches for complicated industrial process. Neurocomputing, 167, 636-642.

[20] Yao, P. (2009, June). Feature selection based on SVM for credit scoring. In Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference on (Vol. 2, pp. 44-47). IEEE.

[21] An, D., Ko, H. H., Gulambar, T., Kim, J., Baek, J. G., & Kim, S. S. (2009, November). A semiconductor yields prediction using stepwise support vector machine. In Assembly and Manufacturing, 2009. ISAM 2009. IEEE International Symposium on (pp. 130-136). IEEE.

[22] Chou, E. P., & Ko, T. W. (2017). Dimension Reduction of High-Dimensional Datasets Based on Stepwise SVM. arXiv preprint arXiv:1711.03346.

[23] Pino A 2018. Melbourne Housing Market data. Kaggle. https://www.kaggle.com/anthonypino/melbourne-housing-market.

[24] Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data (Vol. 333). John Wiley & Sons.

[25] Barnett, V., & Lewis, T. (1974). Outliers in statistical data. Wiley.

[26] James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112). New York: springer.

[27] Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

[28] Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

APPENDIX

Additional informative figures in data preparation and descriptive exploration processes.

Rooms	Price	Distance	Postcode	Bathroom
Min. : 1.000000	Min. : 85000	Min. : 0.00000	Min. :3000.000	Min. :0.000000
1st Qu.: 2.000000	1st Qu.: 657000	1st Qu.: 6.60000	1st Qu.:3046.000	1st Qu.:1.000000
Median : 3.000000	Median : 910000	Median :10.50000	Median :3087.000	Median :1.000000
Mean : 3.067743	Mean : 1089994	Mean :11.41328	Mean :3114.166	Mean :1.596642
3rd Qu.: 4.000000	3rd Qu.: 1333250	3rd Qu.:14.20000	3rd Qu.:3152.000	3rd Qu.:2.000000
Max. :16.000000	Max. :11200000	Max. :48.10000	Max. :3977.000	Max. :9.000000

Landsize	Lattitude	Longtitude	Propertycount	Car
Min. : 14.0000	Min. :-38.19043	Min. :144.4238	Min. : 121.000	Min. : 0.000000
1st Qu.: 308.0000	1st Qu.:-37.86130	1st Qu.:144.9250	1st Qu.: 4385.000	1st Qu.: 1.000000
Median : 567.0000	Median :-37.79970	Median :145.0045	Median : 6567.000	Median : 2.000000
Mean : 597.8127	Mean :-37.80689	Mean :144.9973	Mean : 7521.028	Mean : 1.715242
3rd Qu.: 696.0000	3rd Qu.:-37.74800	3rd Qu.:145.0695	3rd Qu.:10331.000	3rd Qu.: 2.000000
Max. :9838.0000	Max. :-37.39780	Max. :145.5264	Max. :21650.000	Max. :18.000000

Fig. 11 Data summary of numeric predictors in data preparation

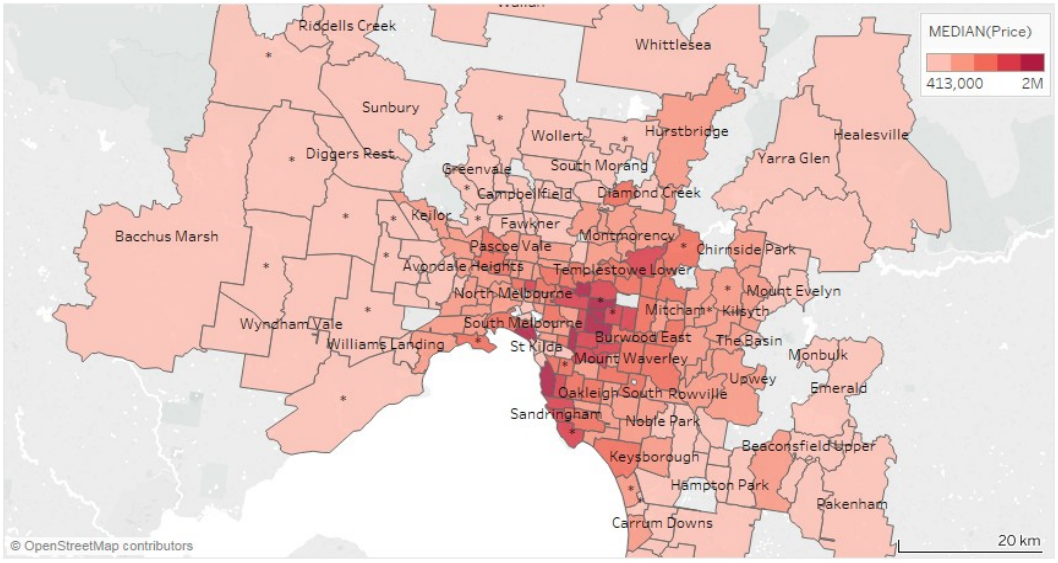

Fig. 12. Median House Price by Suburbs in the City of Melbourne

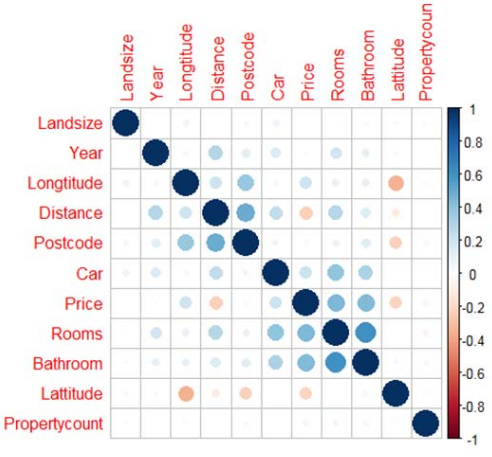

Fig. 13. Numeric variable correlation in data preparation

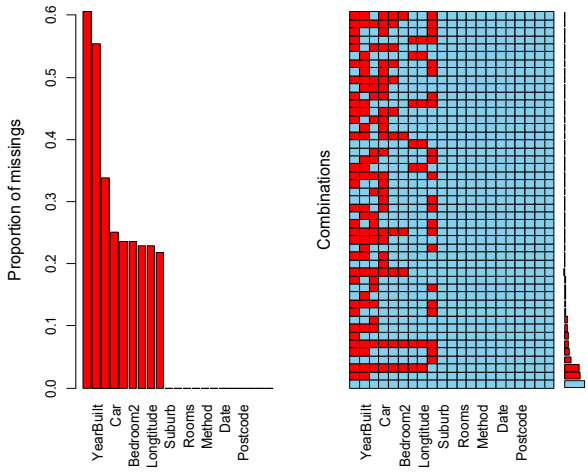

Fig. 14. Missing data patterns in data preparation