Machine Learning Algorithms for Predicting Housing Prices: The Case of Melbourne City, Australia

***Abstract*—A crucial area of real estate is house price forecasting. The literature tries to extract practical knowledge from historical real estate market data. In Australia, historical real estate transactions are analyzed using machine learning techniques to find models that can be helpful to home buyers and sellers. The significant price disparity between homes in Melbourne's most costly and least expensive suburbs is made clear. Additionally, tests show that the Stepwise and PCA, which is based on mean squared error assessment root mean square error, is a competitive strategy.**

Keywords— House price prediction, Random Forest Regressor, Random Forest Classifier, Stepwise, Principal Component Analysis, Neural network.

# Introduction

The goal of the project is to examine previous transactional data to learn important information about Melbourne's property market. This research has the potential to assist homebuyers and real estate professionals in making wise selections by creating precise models to anticipate house values based on their attributes.

Key requirements are very standard, a reliable historical transactional dataset, Data prepossessing for EDA, implementing machine learning algorithms that are suitable for prediction, and lastly, model evaluation and validation. It will be helpful to the public for better price transparency, guiding first-time home buyers to make informed decisions and increase house affordability. Ultimately, fostering transparency and efficiency in the real estate market will help real estate buyers and sellers in Melbourne.

The analytical business problem of this study will be accurately predicting house prices in Melbourne based on market data. Numerous studies have investigated the housing market utilising statistical learning techniques. 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] are the most often studied regions in these studies. However, it is difficult or uncommon to locate studies on the housing market in Australia employing data analytics and machine learning techniques [16].

# Data Collection and cleaning

## Data Collection

There is a slowdown in the housing market of Melbourne City, Australia. We have certain hypotheses, why there is a decline in the buying of houses? or what factors could have led to such a condition? or what to expect from the coming quarters?

To answer such questions, we have a dataset with us have various columns such as Suburb, Address, Rooms, and so on. We found the link to this dataset from Kaggle and below is the link to the dataset. After going through the dataset, we have come across various columns that will help to run different machine-learning algorithms and make predictions from it.

Price, Date, Method, Seller, and Property count are examples of transactional variables. Related location predictors that include an address, a suburb's distance from the city centre, a postcode, a building's area, a council area, a region's name, a longitude, and a latitude [16]

Other elements of a property, such as the type of house, the number of bedrooms, bathrooms, parking spaces, and land size

## Data Cleaning

To assure the integrity and dependability of the results, a comprehensive procedure of data pretreatment and cleaning was implemented in the quest of rigorous data analysis. The actions taken to produce a well-refined dataset for further analysis are described in this section.[16]

Identifying and handling missing values in the dataset was one of the main considerations. A deliberate approach was taken to address this. A wise choice was taken to substitute a value of 0 for missing values in numerical columns when they were present. This decision was made because the numeric properties of the column were naturally suitable for such imputation [19].

A careful selection of essential characteristics became necessary in the pursuit of analytical efficiency. To select necessary variables for our analysis, *Stepwise* - commonly used for subset selection and Principal Component Analysis – usually used for data reduction. In light of this, some columns that were deemed to be unrelated to the analytical focus were purposefully left out of the dataset [19].

In the Type Column, we have the following three keywords: Apartment, Unit, House. After performing one-hot encoding on the "Type" column, we have expanded the categorical data into separate binary columns for each unique category. Each data point is now represented by a combination of 0s and 1s in these new columns, depending on the category it belongs to.

* After applying one-hot encoding, we would get the following expanded representation:

|  |  |  |
| --- | --- | --- |
| **Type\_A** | **Type\_H** | **Type\_U** |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

* In this encoding, a value of 1 in a particular category column indicates that the data point belongs to that category, and all other category columns for that data point will have a value of 0. With the widespread use of analytical techniques like Neural Networks, Decision Trees, and Principal Component Analysis, dealing with outliers has become an urgent issue. In particular, neural networks are sensitive to outliers, especially those that are extreme. An intelligent use of Data Normalisation was chosen to address this problem. The detrimental effect of outliers on model performance was reduced by harmonising the scale of numerical characteristics, promoting improved resilience and accuracy in the subsequent studies [19].

A key stage in our data pre-processing effort was the strategic normalisation of numerical features. The potential disturbances brought on by outliers were successfully minimised by bringing these characteristics to a uniform scale. This normalisation improves the model's capacity to produce insightful and precise findings while also preserving the integrity of the analytical procedure [19].

# Data Prepation and exploration

## Data Prepation

The dataset needs to be pre-processed before using models to predict home prices. First, a missing data investigation is carried out. Numerous missing patterns are carefully evaluated since they are crucial in determining the best handling strategies for missing data [20][16].

Since it is challenging to impute these missing values with an acceptable level of accuracy, columns with more than 55% of their values missing are eliminated from the original dataset. Additionally, the result variable (Price) has several rows with missing values. The observations with missing values in the Price column are eliminated since the imputation of these values can enhance bias in the input data [19][16].

|  |  |  |
| --- | --- | --- |
| **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 |

## Data Exploration

Only the most significant findings are presented in this section. This section contains the data summary information as well as additional educational graphics.

The dataset is erratic, with erroneous data points being collected for each date. The inadequate distribution presents a big barrier for ML. Let's nevertheless not lose heart and persevere [17].

A graph of blue lines

Description automatically generated

Fig.1 Depicting Date-Related Poor Price Distribution

Beyond the 75th quartile figure of $1.33 million, it appears that there are a few outliers. The starting price for a property is $85,000, and the median cost of a home is $903,000. I have to concede that $85,000 is a remarkably low price for a home in Melbourne. Permit me to look for homes in these three categories [17].

A graph with a line and a blue bar

Description automatically generated with medium confidence

Fig.2 Box Plot Analysis for Detecting Price Outliers

A graph of regions vs home prices is shown in the figure. The Y-axis displays housing prices, and the X-axis lists the names of the regions. Despite having the highest total cost of housing, the Southern Metropolitan region does not necessarily have the costliest homes. The Southern Metropolitan has the largest density, according to the region names' density. The Southern Metropolitan region may not actually be the most expensive due to its highest pricing [18].

A graph of different colored rectangular bars

Description automatically generated

Fig.3 Exploring the Relationship Between Region and Price

The Y-axis in the scatter plot above reflects house prices, and the X-axis the distance to the Central Business Sub District. The information demonstrates that lower distances are where most house prices and distances are concentrated. This pattern makes sense because individuals want homes near the Central Business Sub District, which increases demand and drives up housing costs [18].

A graph of a graph

Description automatically generated with medium confidence

Fig.4 Analyzing the Distance-Price Relationship Through a Scatter Plot

The X-axis of the scatter plot shows the names of the regions, and the Y-axis shows house prices. The South-Eastern Metropolitan area is home to the most costly residence, which costs $9 million [18].

The Southern Metropolitan region has the highest total housing values, according to an analysis of the data using a bar plot. It is important to keep in mind, though, that just because one area has the highest total prices doesn't indicate it's where the most costly homes are. The Southern Metropolitan is the region with the highest cost, according to the scatter plot [18].

On the other side, Western Victoria is the most affordable region for homes, and the least costly home, which costs $85K, is situated in the Western Metropolitan area [18].

A graph with blue lines

Description automatically generated

Fig.5 Exploring the Relationship: Region and Price Scatter Plot Analysis

The Melbourne House Prices are displayed on the heatmap. Zoom as needed for a more thorough examination. The highest values, which are reasonable for this investigation, are found in South-Eastern Metropolitan [18].

A map of a heatmap

Description automatically generated

Fig.6 Melbourne House Price Heatmap

The Melbourne Dataset's Heatmap clearly shows that the Rooms feature has a significant positive link with house prices. This conclusion is supported by the scatter plot of these features. On the other hand, in line with their scatter plot, House Price and Distance show a negative association. It suggests that the price of homes tends to rise as the distance from the CBD decreases [18].

A strong correlation between house prices and the features of the bedroom and bathroom has also been discovered. The scatter plot of these factors and this observation line up, proving the importance of these variables. I could choose the best characteristics for the machine learning model by using this technique [18].

A screenshot of a graph

Description automatically generated

Fig.7 House Price Correlation Heatmap: Exploring Distance-Price Relationship and Feature Correlations

It is best to remove the variable from the dataset in light of the mistakes discovered in the recorded land size data[17]. The following factor formed the basis for this choice:

One of the dataset's properties with the largest land size (389 Gore St., Fitzroy) contained errors. Because of point 1, there is little trust in the precision of all other data values. A more efficient model development procedure will result from the removal of erroneous data [17].

A graph with numbers and lines

Description automatically generated

Fig.8 Land Size Distribution: Box Plot of Distance vs. Price

# Methodologies

***Model Selection and Evaluation***

A key aspect of this study is the incorporation of attribute subsets obtained using Stepwise feature selection into various models. Random Forest Regressor, Random Forest Classifier for Price Prediction Class, and Neural Network are the models that were selected. Notably, Stepwise is used in conjunction with Principal Component Analysis (PCA) to determine how it affects model correctness [22][16].

A baseline is created using linear regression, and assessments are based on mean absolute percentage error (MAPE) and root mean square error (RMSE), both of which are quantified using data from a separate evaluation dataset. The prepared dataset is split into training and evaluation subsets; the latter is kept clean throughout the model construction process and is only used for evaluation. Ten-fold cross-validation is crucial to the model fitting process and is skillfully used during the data reduction and model creation phases, highlighting the robustness of the methodology used [22][16].

## Stepwise

One popular technique for choosing a subset is stepwise. A least squares regression is trained for 2p potential models of p predictors using a modified version of the Best Subset Selection approach [26]. We employ forward stepwise selection in this investigation [26] requires just fitting models with (1+p(p+1)/2) parameters [22][16].

We take the crucial factors that are significant for the result from the cross-validation results. Room parameters, land size, longitude, latitude distance, and room type are all included in this exclusive group of predictors. The next figure precisely explains the varied emphasis given to each of these variables and captures how each one contributed to the overall analysis [22][16].

A graph of a bar graph

Description automatically generated with medium confidence

Fig.9 Feature Importance Highlighted in Bar Plot

## Principal Component Analysis

Data reduction techniques like Principal Component Analysis (PCA) are unsupervised methods. It enables us to represent data in a low-dimensional manner while retaining as much feature diversity as feasible. The first six components, which make up almost 80% of the variance of all predictors, are extracted for further analysis after applying PCA to the train data using cross-validation. demonstrates the screen plot of the total percentage of variance across the primary components that can be explained [22][16].

The scree plot, shown in Fig. below, which displays the cumulative fraction of variance expounded along with the matching count of principle components, validates this well-informed choice [22][16].

A graph with a line

Description automatically generated

Fig.10 PCA Scree plot.

##### 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. T.D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.
17. Daverattan, “Melbourne house prices EDA and Deep Learning,” Kaggle, https://www.kaggle.com/code/daverattan/melbourne-house-prices-eda-and-deep-learning (accessed Aug. 7, 2023).
18. Atasaygin, “Melbourne Housing Randomforestregression and Eda,” Kaggle, https://www.kaggle.com/code/atasaygin/melbourne-housing-randomforestregression-and-eda (accessed Aug. 7, 2023).
19. Barnett, V., & Lewis, T. (1974). Outliers in statistical data. Wiley.
20. R. J. Little and R. D. Dublin, Statistical Analysis with Missing Data, John Wiley & Sons.R., vol. 333.
21. T. Pino, “Melbourne housing market,” Kaggle, https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market (accessed Aug. 7, 2023).
22. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: With Applications in R. Springer, 2022.