Analysis of Factors that Affect Housing Prices in Melbourne

Team Information

Team #: 106

Team Members (Name, last 3 digits of GTID, email):

1. Darian Lam; 537; dlam40@gatech.edu
2. Haotian Yu; 644; hyu447@gatech.edu
3. Hye Bae; 164; hbae35@gatech.edu
4. Van Dang 756 vdang35@gatech.edu

Course Name: ISyE 7406

Submission Date: Nov. 25th, 2024

Abstract

The housing market has long been a vital sector of Australia's national economy, exerting a profound influence on economic growth, household wealth, and investment. As a stable yet substantial investment, and a frequently purchased asset, the analysis of key drivers of housing prices is a widely researched and scrutinized topic. This paper aims to identify the primary factors influencing housing prices in Melbourne, Australia, and develop a data-driven model to predict housing prices accurately. To achieve this objective, various machine learning models are explored, utilizing a public dataset that has been thoroughly cleansed and analyzed prior to training to enhance accuracy and interpretability of features. An initial exploratory data analysis is conducted to inform training methodologies, followed by training with potential models to evaluate model performance. Most importantly, the study seeks to identify key features and factors that can reliably predict housing prices. The results of this study provide valuable insights into the most influential factors affecting housing prices and identify the most accurate models for predicting future prices. The findings have significant implications for prospective homebuyers, real estate agents, and property investors, enabling them to make informed decisions in the dynamic Australian housing market.

Introduction

Motivation

Australia’s housing market was valued at nearly 11 billion dollars in Q2 [1]. This large sector of the Australian economy significantly influences national economic growth, household wealth, and economic investment. In the 2010s, Melbourne went through a regional housing boom where housing prices increased by 50% from 2006 to 2012 [2]. This shift in the housing market will directly impact financial stability, mortgage holders, investors, and renters alike.

These changes have reinforced the importance of accurate market predictors as industries from banks to real estate agencies, seek to anticipate future trends. Homeowners with mortgages can better predict interest rate changes, while investors and developers can use insights to make better decisions about property acquisitions. Even renters are impacted as changes to the costs incurred by a landlord will trickle down into rental fees. Improved predictive models could serve as important tools for managing these market shifts for all parties involved.

Problem Statement

This project aims to find the key factors that influence housing prices in Melbourne, Australia allowing us to create an accurate, data-driven model that can predict future housing prices. This model aims to assist prospective homebuyers, real estate agents, and property investors to make informed decisions by providing reliable price estimates.

Research Questions

- What are the most influential factors affecting the housing prices in Melbourne?
- What are the most effective models that can predict the value of a house given a set of predictors?

Methodology

Upon completing the pre-processing of the data, we will explore various predictive models. The dataset will be split into a 70-30 split for training and testing sets. 70-30 was chosen to match the industry standard. The accuracy of the models will be evaluated on the testing set using their RMSE and R-squared values. When possible, 5-fold cross validation will be used to improve model accuracy and prevent overfitting.

The models created with the process outlined above are grouped into three different categories: regularization models, baseline models, and ensemble models.

The use of regularization models was applied first in an attempt to perform dimensionality reduction or variable selection. The large number of predictors in the dataset can potentially lead to overfitting or reduced interpretability in the final model. There are also various forms of spatial data that have varying degrees of predictive power and the most effective predictor will be identified through these models. The specific models used for this purpose are:

- Ridge Regression
- LASSO Regression
- Ridge and LASSO with bagging

Baseline models were then created to compare these results with more easily interpretable models. These models can also give insight into the initial relationship between predictors. The specific models used for these baseline models are:

- Linear regression with all predictors
- Linear regression with Akaike Information Criterion (AIC)

The regularization and baseline models do not capture the complex relationship between the predictors and response. Due to their non-linear nature, more powerful ensemble methods are then used to capture this relationship while maintaining some degree of interpretability and efficiency. These models should also improve accuracy and mitigate any weaknesses a singular model may have. The specific models used for these baseline models are:

- Random forest (RF)
- RF with feature selection
- XGBoosting
- XGB with feature selection
- Adaboost

- Adaboost with feature selection
- Stacking
- Blending

Dataset

The dataset includes information on actual housing transactions that occurred in Melbourne, Australia, during the period from 2016 to 2018. This data was taken from Kaggle (link) where the data itself was scraped from publicly available results from Domain.com.au. The full dataset has 34,857 observations and 21 predictors including the response variable (price) with both discrete and continuous data types.

A data pre-processing step will be performed where rows or columns with a significant number of missing values will be removed, and columns with minor missing values will be imputed. Therefore, not all attributes shown here are used in the modelling process. During the pre-processing step, time series analysis will also be performed to determine if the data has seasonality or remains stationary.

All attributes are listed below:

- Suburb: Suburb area of the house
- Address: Address of the house
- Rooms: Number of rooms
- Price: Price in Australian dollars
- Method: How the property was sold
- Type: Type of house
- SellerG: Real Estate Agent
- Date: Date sold
- Distance: Distance from CBD in Kilometres
- Regionname: General Region (West, North West, North, North east …etc)
- Propertycount: Number of properties that exist in the suburb.
- Bedroom2: Scraped # of Bedrooms
- Bathroom: Number of Bathrooms
- Car: Number of carspots
- Landsize: Land Size in Metres
- BuildingArea: Building Size in Metres
- YearBuilt: Year the house was built
- CouncilArea: Governing council for the area
- Latitude: Latitude of the house
- Longitude: Longitude of the house

Data Cleaning and Transformation

Before implementing models for prediction and time series analysis, some preprocessing was necessary. First, we examined the missing data and identified several patterns. Specifically, the columns “YearBuilt” and “BuildingArea” had more than 55% of their observations missing. Additionally, 22% of the total observations were missing values for the response variable (price). Other columns with missing data included Bedroom2, Bathroom, Car, Landsize, Latitude, and Longitude. Figure 1 illustrates the missing data by column, highlighting rows with missing data in navy.

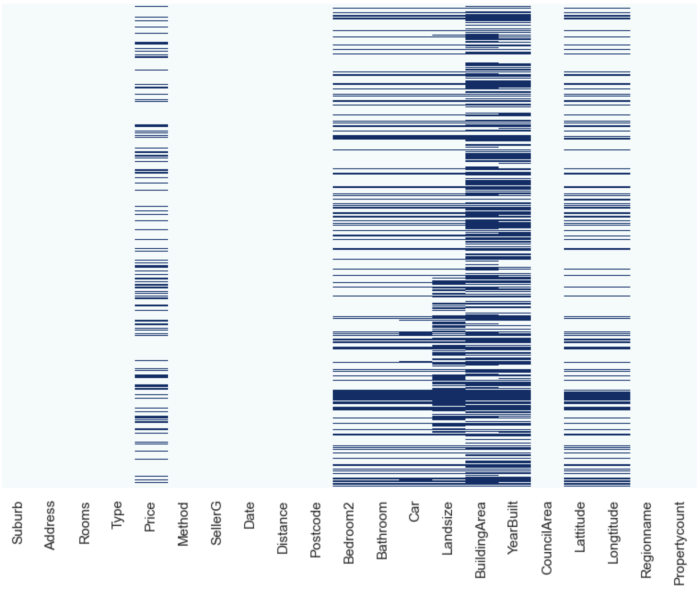


Figure 1. Missing Data by Columns

Columns with more than 55% of observations missing were dropped, as accurately imputing these values would be challenging. Observations missing values for the response variable (price) were also removed, since imputing these values could introduce additional bias into the dataset. Instead of using a single imputation method for all missing values, we applied techniques tailored to produce the best estimates for each specific case. To impute missing latitude and longitude data, we used the Google Maps API, referencing other location-related columns like address and postcode. Landsize was imputed based on median values grouped by house type. Similarly, columns like Bathroom and Car were imputed using the median values grouped by Bedroom2. After imputing missing values, we identified outliers as data points with a z-score greater than 3 or less than -3. Approximately 1% of data points were flagged as outliers and removed to maintain dataset integrity. The final dataset included 23,070 observations.

Exploratory Data Analysis (EDA)

The median house has 3 bedrooms, 1 bathroom, 2 parking spaces, and a land size of 448 square meters, with a median price of $888,000 AUD. A detailed descriptive analysis of the dataset of all variables are provided in Appendix A. The distribution of the response variable, price, shows a wide range with a long right tail (Appendix B). To normalize the distribution, we transformed the data using log (price), which will be used as the output variable in modeling. A map of Melbourne, using the latitude and longitude columns, revealed that housing prices are highest near the city center and tend to decrease as the distance from the city center increases, shown in Figure 2. Lastly, a correlation analysis among the predictor variables and the response variable showed that Rooms and Bathroom are strongly correlated with price, with correlation coefficients of 0.47 and 0.39, respectively (Appendix C). However, it is unlikely that the response variable has a simple linear relationship with the predictors. Due to the complexity of predicting housing prices, the relationship between the predictors and the response variable is likely more intricate and interconnected.

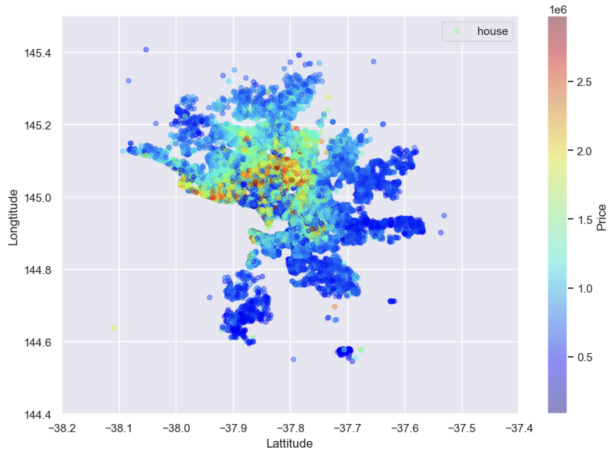


Figure 2. Scatterplot of Longitude and Latitude against Price

Time Series Analysis

We conducted a time series analysis on the Australian housing data to determine if the data is stationary and to understand its predictive capabilities. Time series analysis is a useful technique for modeling data that exhibits temporal dependencies. We began by visualizing the data to identify any trends, seasonality, or irregular patterns.

In addition to visualizing the original time series data, we applied both additive and multiplicative decomposition techniques to further decompose the data. Additive decomposition assumes that the trend, seasonality, and residuals are added together to form the original time series, while multiplicative decomposition assumes that these components are multiplied together. However, even after applying these decomposition techniques, we did not observe any significant trends or patterns in the data. The trend component was relatively flat, and the seasonality component did not exhibit any regular or predictable patterns.

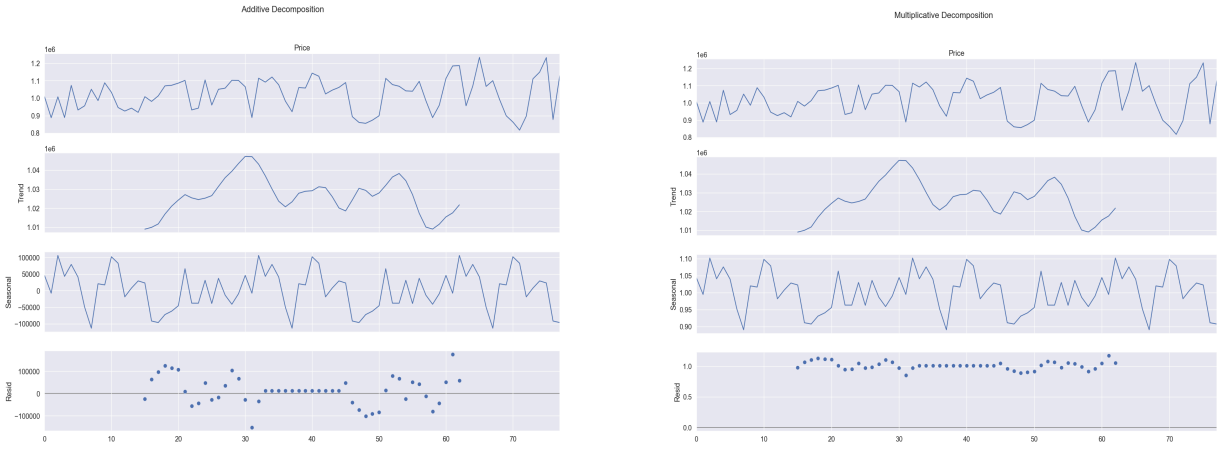


Figure 3. Additive (left) and Multiplicative Decomposition (right)

We also used the Auto ARIMA function to automatically select the best ARIMA model for our data. After evaluating various models, the Auto ARIMA function selected an ARIMA(0,0,2) model as the best fit for the data. This model suggests that the data does not exhibit any autoregressive behavior (AR term is 0), does not require differencing (d term is 0), and exhibits moving average behavior of order 2 (MA term is 2). However, based on results from the ARIMA model, we ultimately decided to forgo the time series analysis due to the sparse nature of the data. The data was considered sparse because not many data points were provided with specific suburbs, at a given year.

In conclusion, while the time series analysis provided some insights into the data, we believe that a more comprehensive and accurate understanding of the Australian housing market can be achieved through other modeling techniques, such as regularization models and regression models with some feature engineering.

Model Training & Optimization

Regularization Models

Two models were selected for regularization: LASSO and Ridge. LASSO was chosen because it uses a penalty or regularization term, lambda, to the sum of the absolute value of the coefficients. This shrinks the coefficients and can even perform feature selection by reducing the coefficients to zero. This is shown in the equation 1:

$$\hat{\beta}_{lasso} = min\left\|Y_{nx1} - X_{nxp}\beta_{px1}\right\|^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

Equation 1. LASSO

The regularization term is tuned through cross validation to find the model that has the best trade off of prediction accuracy and model complexity. By using cross validation and finding the optimal tuning parameter, we aim to achieve a careful balance between bias and variance.

Ridge regression follows the same approach as LASSO but instead applies the penalty to the sum of square values as shown in the equation below. This only shrinks the coefficients and does not eliminate them, preventing ridge from performing any variable selection. This however, does allow ridge to account for multicollinearity.

$$\hat{\beta}_{ridge} = min\left\|Y_{nx1} - X_{nxp}\beta_{px1}\right\|^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Equation 2. Ridge regression

In either case, both of these models will require scaling of the data to ensure that all features have the same impact on the model. In addition, both models are still linear regression models that rely on using ordinary least squares (OLS). This means that complex relationships in the data may not be adequately captured. In addition, bagging was implemented for both LASSO and RIdge. This bootstrap aggregating technique is used to improve the accuracy of both models through multiple training sets. Each model is fit on each bootstrapped sample and all the predictions are aggregated for one final prediction. This should ideally reduce variance and improve the consistency of feature selection at some computational cost.

Baseline Models

For the baseline model, we chose the simple multiple linear regression for reasons below:

1. Linear regression is a simple model that helps us understand the direct relationships between housing prices and individual predictors, such as the number of rooms, land size, etc. This gives us a straightforward initial view of which factors might be associated with housing prices in a simple linear framework.

2. MLR with stepwise feature selection using the AIC criterion allows us to identify a subset of the most relevant predictors. This could help us simplify our model and remove less relevant features.
3. MLR models establish a baseline for prediction performance. By comparing more complex models to these baselines, we can assess whether advanced techniques like ensemble methods meaningfully improve predictions.

Ensemble Models

Other than the models above, we chose Random Forest (RF), XGBoosting (XGB), Adaboost, and Stacking/Blending as ensemble methods to test due to some factors below:

1. Our EDA indicates that housing prices are likely influenced by non-linear relationships and complex interactions among variables. Ensemble methods, especially tree-based models, are well-suited for capturing these relationships because they do not assume linearity.
2. Ensemble methods, which combine predictions from multiple models, often achieve higher accuracy by reducing variance (through bagging methods like Random Forest) or bias (through boosting methods like XGBoost and Adaboost). These techniques can lead to better generalization on the test data.
3. Random Forest and XGBoost provide built-in measures of feature importance, helping us identify which predictors have the strongest impact on housing prices. This aligns with our project goal of understanding the most influential factors.
4. Combining models through stacking or blending allows us to leverage the strengths of each model type. For instance, blending predictions from a Random Forest with those from XGBoost can capture different aspects of the data and potentially improve performance by balancing bias and variance.

Results & Analysis

Regularization Models

LASSO and Ridge were both created with five fold cross validation and used longitude and latitude as spatial data but the results for both models were poor. The models produced an R-squared of 0.5041 as shown in Table 1. The penalty term, lambda, is tuned and another set of models were created with bagging. The use of bagging indicates that the results should be consistent as it accounts for multiple samples and iterations. This indicates that the model is likely non-linear and more complex than what can be captured with this regularization model.

	Lambda	R^2	RMSE
LASSO	0.0001	0.5041	0.147
LASSO W/Bagging	0.0001	0.5041	0.147
Ridge	10	0.5042	0.147
Ridge W/Bagging	10	0.5041	0.147

Table 1. Results of Regularization Models using long and lat as spatial data

In addition, the regularization models performed feature selection. It eliminated the features longitude, latitude, landsize, and propertycount. While the regularization models did not perform well in terms of the accuracy, their feature importance is useful. In order to further understand the predictive power of spatial variables such as longitude and latitude, we looked into interaction variables.

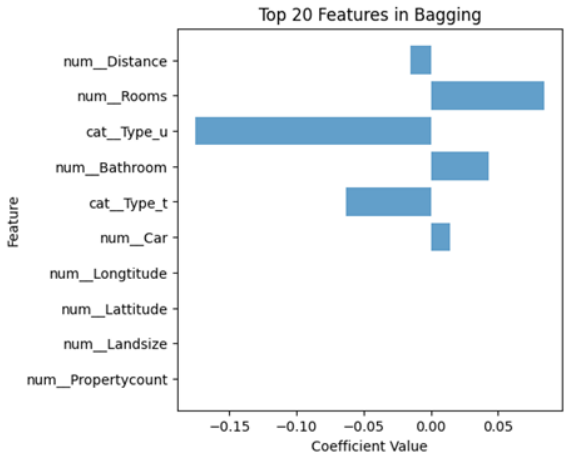


Figure 4: Top 20 Features in Bagging

LASSO and Ridge treat each predictor independently, which might overlook relationships between predictors. For example, latitude and longitude may not have been properly evaluated together in the previous model. To address this, we introduced an interaction variable to capture the combined effect of latitude and longitude, adding some multicollinearity to potentially boost the predictive power of these two variables. However, the model did not perform well and had an even lower R^2, as shown in the Tabel 2. Figure 5 also shows the interaction variable was reduced to zero. Therefore, we decided to remove longitude and latitude from the list of predictors.

	Lambda	R^2	RMSE
LASSO W/Interaction Variable	0.001	0.5036	0.1471

Table 2. Results of Regularization Models using lat and long as spatial data and adding an interaction variable

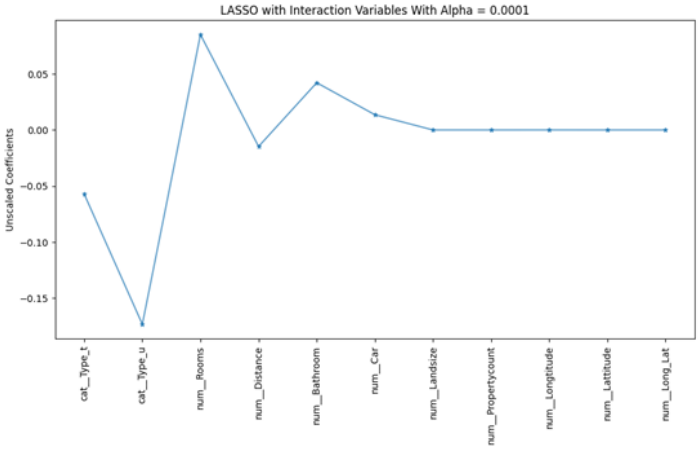


Figure 5. Coefficient of Predictors for LASSO with Bagging

Instead of using latitude and longitude, , we re-ran the model with suburb as the location variable. As shown in Table 3, the model including suburb has a much higher predictive power as many of the suburbs are statistically significant as shown in Figure 6 However, Figure 6 also shows the negative impact to interpretability for the model. This gives us critical information that spatial data like suburb and similarly, postal code, will be important for housing prices at the cost of some interpretability.

	Lambda	R^2	RMSE
LASSO W/Suburb	0.0001	0.7695	0.1002
LASSO W/Suburb and Bagging	0.0001	0.7695	0.1002

Table 3. Results of Regularization Models using suburb as spatial data

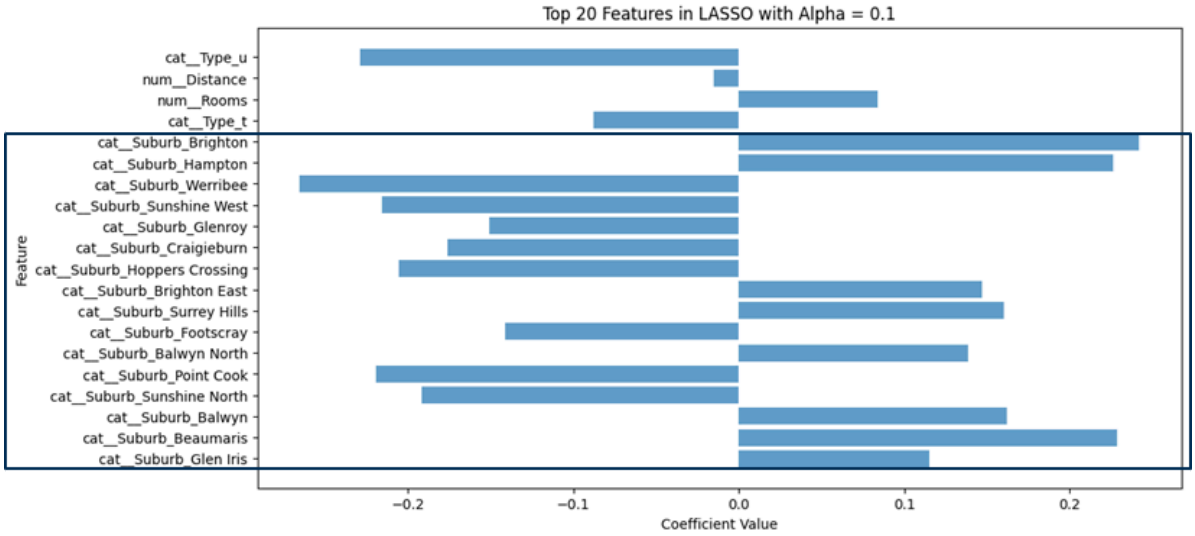


Figure 6. Coefficient values by suburbs in LASSO

Baseline Models

Next, we further look into which of the different location variables (postcode, suburb, and longitude/latitude) provided the best predictive power and model stability. This was particularly important because of multicollinearity concerns - in regression models, this could distort the relationship between each predictor and the target leading to unreliable estimates. Since we are using MLR models as the baseline, an optimal location variable in the MLR models was needed to establish a robust baseline.

	Model	Description	MSE	MAE	R2
0	Multiple linear regression	MLR w/ postcode and other variables	1677.244257	30.882203	0.761515
1	Multiple linear regression	MLR w/ suburb and other variables	1691.569706	30.460650	0.759478
2	Multiple linear regression	MLR w/ longtitude+lattitude and other variables	3727.213105	48.228645	0.470032

Table 4. Results of the MLR models with different location variables

Table 4 shows that using postcode or suburb as the location variable led to substantial improvements in model accuracy (lower MSE and MAE, and higher R^2 values) compared to using longitude and latitude. This is consistent with the results from the regularization models. Postcode achieved the lowest Mean Squared Error (MSE) and highest R^2, indicating it best captures regional price variations. Based on this performance, we selected postcode as the primary location variable for all subsequent models. Postcode provides a balance of geographic specificity and categorical structure, which is well-suited for regression and ensemble models alike. After establishing the final dataset that we would adopt for future models, we built our first baseline simple MLR model using postcode with all other variables.

Beyond the baseline model, we also tried stepwise feature selection using the AIC criterion to aid us in selecting only features of importance. However, similar to what was found above, all features were deemed to be important as shown in Figure 7, where in all 10 folds, all features were chosen.

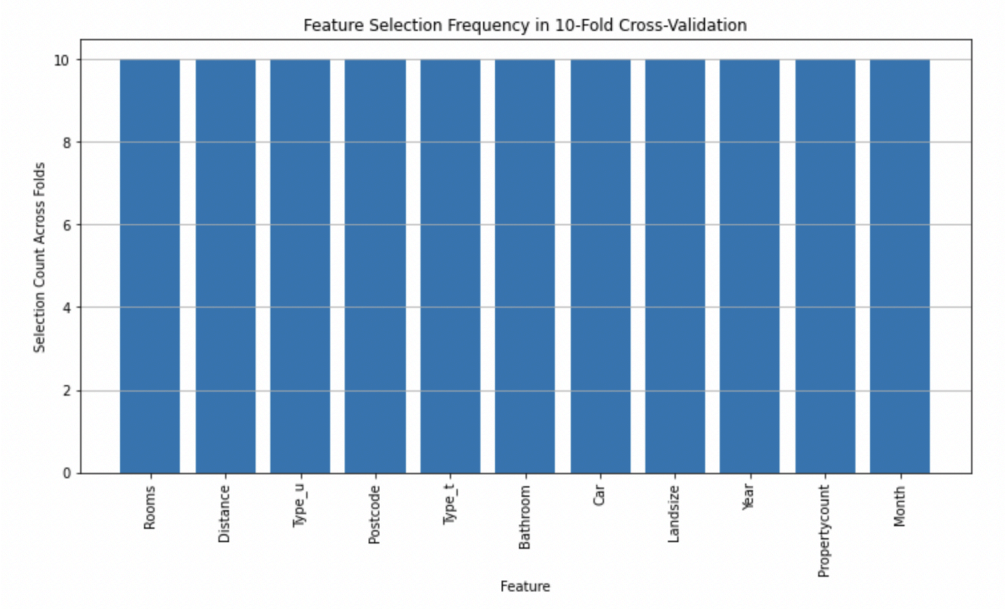


Figure 7. Feature Selection Frequency in 10-Fold Cross-Validation

This further justifies that all features are important predictors of price and more complex models should be tried in order to capture nonlinear more complex relationships.

Ensemble Models

Base models

Since tree models have their own feature selection method by importance(RF model ranks by variance reduction, XGBoost scores by number of times a feature was used to split data), we tried the respective feature selection methods for the random forest model and the XGBoost model as well. Another important step was hyperparameter tuning, each model requires a set of hyperparameters to be tuned for optimal performance. Using GridSearchCV, we tested various combinations of parameters to identify the configuration that produced the best results. Here's a breakdown of the specific parameters we tuned for each model:

Models	Parameters tested in GridSearchCV (Optimal parameters in red)
Random Forest	n_estimators: [50, 100, 200], max_depth: [10, 20, None], min_samples_split: [2, 10], Feature selection with RandomForestRegressor
XGBoost	n_estimators: [50, 100, 200], max_depth: [3, 6, 10], learning_rate: [0.01, 0.1, 0.2], Feature selection with XGBRegressor
AdaBoost	n_estimators: [50, 100, 200], learning_rate: [0.01, 0.1, 1]

Table 5. Parameters tested in GridSearchCV

Stacked/ Blended models

After obtaining the optimal parameters for these three base models, we combined all their predictions and used a linear regression model as the meta model to learn from these predictions. This approach allows the meta-model to capture any residual errors from the base models, improving accuracy.In addition, we tried the blended method where we averaged the predictions from the three base models, reducing variance and balancing out any biases. By adjusting the weighting of each model's predictions, we fine-tuned the ensemble's performance to optimize the final results.

Finally, through feature selection and GridSearchCV hyperparameter tuning, we optimized each tree-based and boosting model, leveraging their unique strengths while minimizing weaknesses. The stacking and blending approaches further combine these optimized models to produce more robust and accurate predictions.

Summarized results

	Model	Description	MSE	MAE	R2
8	Optimized Stacking	RF + XGBoost + Adaboost	1134.294939	24.435030	0.838716
3	Optimized XGBoost	XGBoost with GridSearchCV	1143.479379	24.612207	0.837410
7	Optimized AdaBoost	AdaBoost with GridSearchCV	1143.479379	24.612207	0.837410
4	XGBoost with Feature Selection	XGBoost + SelectFromModel + GridSearchCV	1185.253730	25.385373	0.831470
1	Optimized Random Forest	RF with GridSearchCV	1226.135536	25.365913	0.825657
2	RF with Feature Selection	RF + SelectFromModel + GridSearchCV	1230.842883	25.496688	0.824988
9	Optimized Blending	(RF + XGBoost + Adaboost) Predictions / 3	1349.645939	26.919003	0.808096
5	LASSO	LASSO with all features	1621.799699	30.476579	0.769543
6	Ridge Regression	Ridge regression with all features	1622.052260	30.502302	0.769264
0	Baseline multiple linear regression	MLR with all features	1677.244257	30.882203	0.761515

Table 6. Summarized results of all models tested

From the summarized results ranked by MSE, we can see that the Optimized Stacking model (RF + XGBoost + AdaBoost) achieved the lowest Mean Squared Error (MSE) and highest R^2 of 0.8387, demonstrating that combining models through stacking can significantly improve prediction accuracy. The meta-model in stacking leverages the strengths of each base model, capturing residual patterns that individual models might miss.

Another interesting point to note is that models with feature selection (RF & XGBoost) actually performed worse than their counterparts (without feature selection). While feature selection can simplify models and reduce dimensionality, the inherent feature importance mechanisms in RF and XGBoost may already manage less important predictors effectively. Coupling with what was found from earlier feature importance models (stepwise elimination), where all features were deemed important in the final refined dataset, it is clear why the base model (slightly) outperformed their respective models with feature selection.

Lessons Learned

The project was valuable in consolidating everything I learned throughout the course and producing results with real-world applications. Overall, I enjoyed the class and appreciated that the assignments emphasized writing and presentation skills, not just the technical aspects. I learned that to deliver a clear and coherent story that anyone can understand, I need to be thoroughly versed in the technical details. We also came into this project expecting to be able to use different forms of regularized regression for variable selection. When we actually applied these models to the project, this did not prove as effective as we hoped. We had to adapt to these changes and pursue new models with different variables as we interpreted our results.

Conclusion

In summary, this project analyzes transactional housing data in Melbourne, with a focus on data preparation to create a clean dataset. Various methods were tested with iterative variable selection, and the best-performing model, based on MSE values, was the stacked ensemble model, which explained 83.9% of the variability in the response variable. All features were considered important, with postcode emerging as the key location variable for modeling purposes. The findings highlight the value of machine learning in predicting housing prices in Melbourne more consistently. Future work could involve expanding the dataset or testing the models on data from other regions to further improve the models' robustness.

Bibliography

[1] Australian Bureau of Statistics. “Total Value of Dwellings, March Quarter 2022 | Australian Bureau of Statistics.” *Www.abs.gov.au*, Australian Bureau of Statistics, 14 June 2022, www.abs.gov.au/statistics/economy/price-indexes-and-inflation/total-value-dwellings/latest-release.

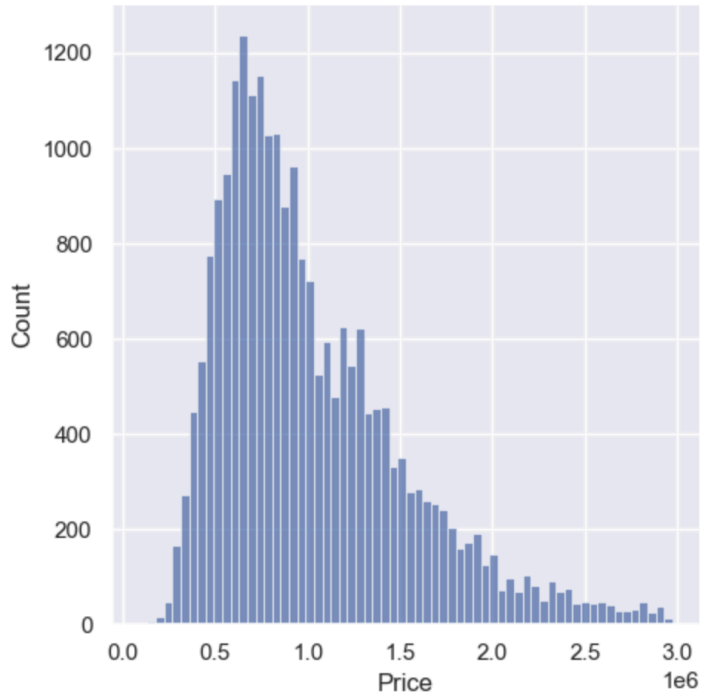
[2] Dept, International Monetary Fund Asia and Pacific. “Australia: Selected Issues.” *IMF Staff Country Reports*, vol. 2018, no. 045, 20 Feb. 2018, www.elibrary.imf.org/view/journals/002/2018/045/article-A003-en.xml, https://doi.org/10.5089/9781484341872.002.A003.

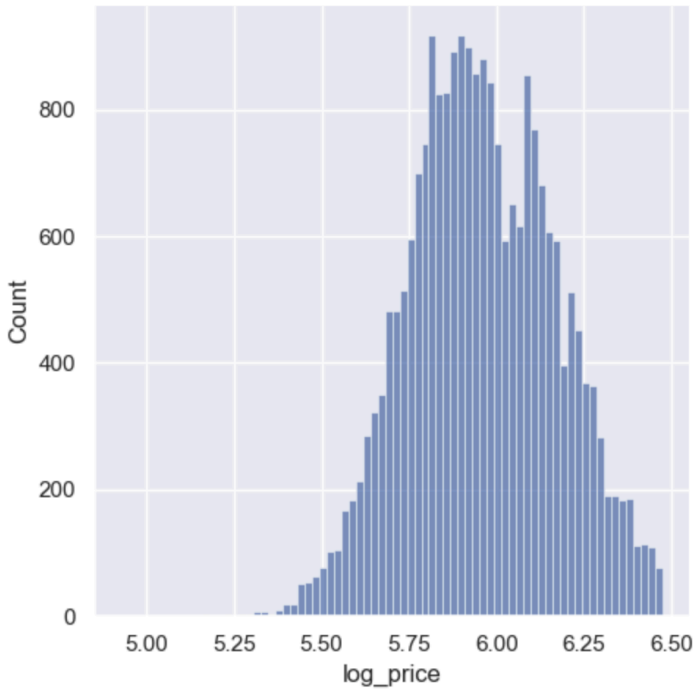
Appendices

Appendix A. Descriptive Analysis of all variables

	count	mean	std	min	25%	50%	75%	max
Rooms	23070.0	2.941439e+00	0.895170	1.000000	2.000000	3.000000	4.000000e+00	5.000000e+00
Price	23070.0	1.017984e+06	507927.799496	85000.000000	646000.000000	888000.000000	1.291000e+06	2.975000e+06
Distance	23070.0	1.068929e+01	5.757407	0.000000	6.300000	10.100000	1.390000e+01	3.160000e+01
Bathroom	23070.0	1.479931e+00	0.607016	0.000000	1.000000	1.000000	2.000000e+00	3.000000e+00
Car	23070.0	1.700000e+00	0.775248	0.000000	1.000000	2.000000	2.000000e+00	4.000000e+00
Landsize	23070.0	4.422721e+02	475.557696	0.000000	191.000000	448.000000	6.210000e+02	9.838000e+03
Lattitude	23070.0	-3.308820e+01	18.690407	-45.898862	-37.857800	-37.796100	-3.773976e+01	5.811726e+01
Longtitude	23070.0	1.327348e+02	50.556977	-157.863295	144.907581	144.997683	1.450673e+02	1.769154e+02
Propertycount	23070.0	7.129758e+03	3840.710876	121.000000	4217.000000	6482.000000	9.758000e+03	1.749600e+04

Appendix B. Distribution of the response variable





Appendix C. Correlation analysis between variables