Fuentes, Paolo

Pecundo, Allan Magno

Villegas, Mylene

CSCI 271-A

FINAL REPORT

I. Introduction and Significance of the Study

Background of the Study

Australians traditionally place a high degree of importance on homeownership as owning one is a sign of having achieved the "Great Australian Dream". This ideology is described as the pathway to securing success and security financially. Around 70% of Australian households live in houses that they purchased themselves. Of these, 50% own their dwellings outright (without a mortgage loan), and 50% have mortgage loans. As such, the burden of government aged care welfare payments is reduced because of the high level of home ownership in Australia. (Eslake, 2017) [1].

However, this ideology has become unsustainable for some echelons of society as Australian housing is becoming increasingly expensive as people are spending more of their income on housing. This growth has been observed to rise faster in major capital cities such as Melbourne. This has been the case in suburban areas as well but at a slightly slower rate(Eslake, 2017) [1].

This crisis in affordability can be attributed to two issues. First, the shift from high to low interest rates has boosted borrowing ability and hence buying power. Second, there has been an inadequate supply response to demand. Starting in the mid-2000's annual population growth surged by around 150,000 people per year and this was not matched by an increase in the supply of houses resulting in a chronic shortage. Therefore, these factors explain why Australian housing is expensive compared to many other countries that have low or even lower interest rates (Oliver, 2021) [2].

One segment of the market that has suffered the disparity between house and unit prices are Australians who are looking to sell their apartment to move into a free-standing home. This is making it far harder for first home buyers to get into the market – it now takes 8 years to save for a deposit in Sydney and nearly 7 years in Melbourne (Marsh,2021) [3].

In fact, the expensive housing rates have only exacerbated after the country has eased lockdown measures during the ongoing COVID-19 pandemic. Melbourne's housing market has boomed post-lockdown as listings and auction clearance rates surge. The average house in Melbourne is now selling for $973,000 and units for $622,000 (The Urban Developer, 2021) [4]. According to a study by KPMG (2021), The post-lockdown increase in housing prices is attributed to the re-entry of vendors in the housing market, as onsite work is gradually resuming [5].

Stakeholders of the Study

Certain demographics of residents living in Melbourne can be considered as stakeholders who will benefit from this study. This study is meant for first time buyers who would want to transition to living independently. Second, it can be for families who would want to relocate depending on their lifestyle. Lastly, it can also be for businessmen who want to engage in buying properties then having them rented to gain profit from it.

Significance of the Study

As the rates of houses in Melbourne are rapidly increasing and have no signs of slowing down in the near future, it is essential and practical for house buyers to derive utmost value from their purchases as they spend large amounts of money for these houses. This is because buyers consider various factors before shelling out hundreds of thousands of Australian dollars on houses. According to an article by Farina (2017), a property intelligence company, when home buyers are considering a purchase, they are buying into a lifestyle [6]. As such, their type of lifestyle plays a role in the decision making process, particularly choosing the features that the house possesses such as its quantity of rooms, location, and land area, among others. With the aid of regression analysis, they will be able to determine which characteristics of the house contribute the most to the overall price, which can then serve as a decision point of evaluating whether their preferred lifestyle is in accordance with their budget.

Moreover, the ongoing pandemic has had negative effects on the economy and employment, thereby limiting the purchasing power of buyers on the market. Through data-driven prediction, home buyers will be able to benefit from the study as price predictions will be able to help them in regulating their expectations with regards to practical options in terms of house features that they can actually afford.

With all these facts in mind, the study will aim to address the following questions:

- Knowing the various characteristics of a house listing, can we predict the house's price?
- Which characteristics of a house's listing contribute the most to the house's pricing?

II. Methodology

To address the following questions, a Machine Learning model will be used to come up with predictions of a house's price while only considering the other characteristics of a house listing that is provided. The same model will also be used to analyze the importance of each of the house's characteristics in determining the price of a house.

The dataset used is the Melbourne Housing dataset publicly posted on Kaggle [6]. This dataset contains house listings of Melbourne found on domain.au, a real estate website

for listings in Australia along with characteristics describing the house such as location, room count, number of bathrooms, and more.

Pre-processing

First, the dataset was inspected for any missing transactions. Upon checking the dataset, it was observed that the "Price" column had missing observations. Given that the price column is the target variable that will be predicted by the model, the observations missing this variable will be removed as these cannot be used to either train or evaluate the model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34857 entries, 0 to 34856
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         34857 non-null  object
 1   Address        34857 non-null  object
 2   Rooms          34857 non-null  int64
 3   Type           34857 non-null  object
 4   Price          27247 non-null  float64
 5   Method         34857 non-null  object
 6   SellerG        34857 non-null  object
 7   Date           34857 non-null  object
 8   Distance       34856 non-null  float64
 9   Postcode       34856 non-null  float64
 10  Bedroom2       26640 non-null  float64
 11  Bathroom       26631 non-null  float64
 12  Car            26129 non-null  float64
 13  Landsize       23047 non-null  float64
 14  BuildingArea   13742 non-null  float64
 15  YearBuilt      15551 non-null  float64
 16  CouncilArea    34854 non-null  object
 17  Lattitude      26881 non-null  float64
 18  Longtitude     26881 non-null  float64
 19  Regionname     34854 non-null  object
 20  Propertycount  34854 non-null  float64
dtypes: float64(12), int64(1), object(8)
memory usage: 5.6+ MB
```

Figure 1. Summary Statistics of Melbourne Housing from domain.au

After the observations with missing data from "Price" were removed, the "Bathroom" (# of bathrooms), "Car" (# of car capacity), and "Landsize" columns were then inspected as missing data was still observed in these columns after the prior cleaning. Given that the 3 columns mentioned had relatively larger count of data missing, KNN Imputation was performed considering the following features for determining the neighbors: "Rooms" (# of rooms), Type (type of house encoded as "house", "townhouse", or "unit"), and Land size. After the imputation was done, atypical transactions wherein "Bathroom" exceeded "Rooms" (# of rooms in the house) were removed. This was done separately for the training and test set to avoid any data leakage.

Finally, the last set of transactions that were found to have missing data were those that had missing "Distance" (distance to the city center) and "Regionname" (area in Melbourne where the house is found). Overall, around 28% of the data was reduced (from

34,857 to 27,208) after the cleaning; with most of the data removed being due to missing "Price" which is the target feature.

After the features were cleaned, the categorical features that will be used in the model were one-hot encoded to allow for model usage during the training and prediction; these categorical features were "Type" and "Regionname". The final features to be used in building the model were identified to be features that were either:

- Features that are found in the house listing that the home buyers may research and identify before purchasing or making an offer on the property.
- Features that could be pre-processed for model building within the time constraint of the exercise.

The following were the final set of features that will be used for the modelling:

Input Features: Room, Bathroom, Type (encoded as unit, townhouse; housing is dropped as per encoding standard), Distance, Bathroom, Car, Landsize, Regionname, Propertycount (count of properties present within the area)

Target Features: Price

After the features were selected, one last step of minmax scaling was applied to the input features to reduce any influence of magnitude on the importance of the features for predicting the price.

Model Building

Prior to building the model, the data was first split into training and test set following an 80% and 20% split, respectively. It was also after the splitting process that the imputation of missing values for "Bathroom", "Car", and "Landsize" was applied. The model was trained on the training set, and model performance was measured on the test set.

To be able to predict the house price given other known features of a house, the ElasticNet regression model was used, and tuned to different alpha values; with 0 alpha representing no penalization of coefficients (e.g., Linear Regression).

A. ElasticNet Regression

ElasticNet model fits the similar to a simple linear regression model which takes the form,

$$\hat{\mathbf{y}} = \boldsymbol{wX} + \boldsymbol{b}$$

And tries to identify the best set of weights (coefficients) w and intercept b wherein,

$\hat{y}$ = predicted target variable X = vector of input features

w = vector of coefficients b = y-intercept

Unlike simple linear regression, the model identifies the w and b by minimizing the cost function:

$$C(\vec{w}) = \sum_{i=1}^{n} (\hat{y}_i - \vec{y_i})^2 + \alpha_1 \sum_{j=1}^{m} |w_j| + \alpha_2 \sum_{j=1}^{m} w_j^2.$$

Figure 2. Cost Function of ElasticNet Regression [8]

While Linear Regression only tries to minimize the difference between predicted $\hat{y}$ and observed y (i.e., cost function), ElasticNet regression's cost function also considers a regularization parameter, α, which acts as a penalty to the coefficients w identified by the model. The larger the coefficients computed by the model for prediction, the larger corresponding cost function. The ElasticNet model does this to prevent possible cases wherein the coefficients calculated by the model during training become too large such that they influence the model more than the change in the input features, leading to overfitting on the training data while not performing as well on any new (e.g., test data) dataset.

The data was fitted using the following α values: [0.00001, 0.001, 0, 1, 100, 100000], with the 0 value representing the standard Linear Regression and its corresponding cost function. Based on the results, the best model was generated with α set at 0.00001 with the following performance metrics:

Model	Training R^2	Testing R^2
ElasticNet (alpha = 0)	0.5642	0.5535
ElasticNet (alpha = 0.00001)	0.5642	0.5533
ElasticNet (alpha = 0.001)	0.5622	0.5508
ElasticNet (alpha = 0.1)	0.3852	0.3699

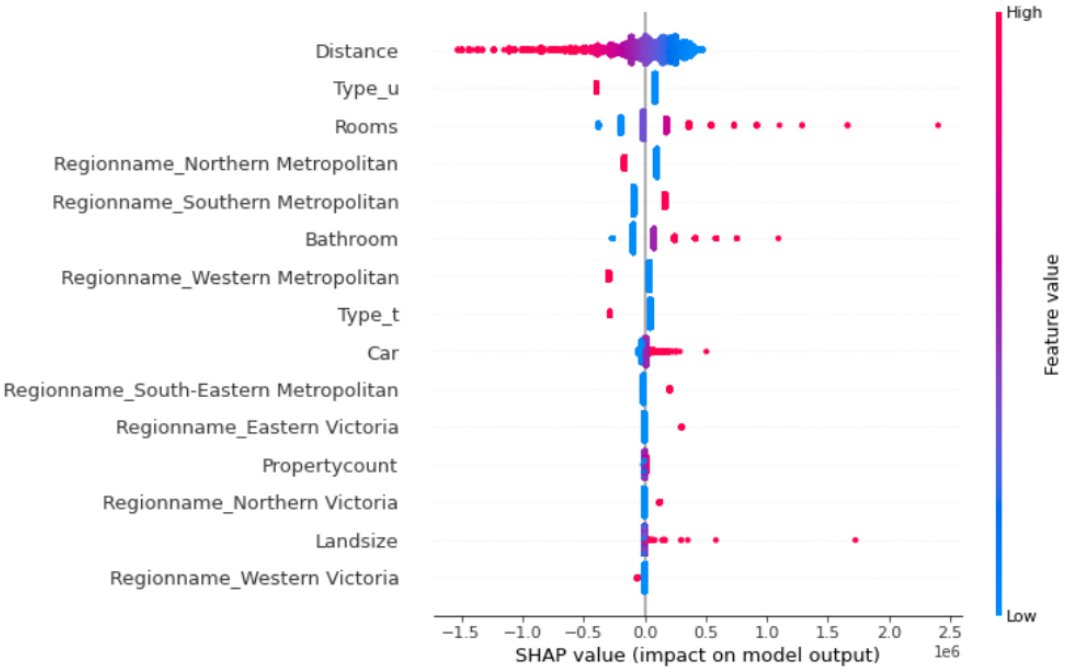
ElasticNet (alpha = 1)	0.1629	0.1548
ElasticNet (alpha = 100)	0.0026	0.0025
ElasticNet (alpha = 10000)	0.0000	0.0000

Figure 3. Summary of model performance

Based on the results, the best performing model was the ElasticNet regression with the α value of 0.

III. Insights

To further understand the final model, the SHAP values of the features used in the trained model were derived and analyzed. As mentioned in the SHAP Documentation [9], SHAP values or SHapley Additive exPlanations mainly interpret the impact of having a certain value for a given feature in comparison to the prediction we would make if that feature took some baseline value. The plot below is made of all the dots in the train data. It demonstrates the feature importance, impact, and correlation.



- Feature Importance: The average of the absolute Shapley values was derived per feature across the data. These values were used to rank the variables in descending order.

- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Correlation: To interpret the correlation as represented in the plot, it is important to take note of the colors that were used as legends. Color Red represents high feature value and Color Blue represents low feature value. If a high feature value is associated with a lower prediction, then that means that the feature is negatively correlated with the price. If a high feature value is associated with a higher prediction, then that means that the feature is positively correlated with the price.

From the initial checking, it was observed that the distance feature is the most important among the features used to train the model. This particular feature refers to the distance of the house from the Central Business District. The Central Business District is Melbourne's business and financial centre and as a house owner, it might be important to consider the easy access to this location. As can be seen in the plot, the high feature values of distance are associated with lower predictions. This means that the distance feature is negatively correlated to the house price. Hence, the farther the house is from the central business district, the cheaper the pricing is.

Additionally, the features type_u and type_t were also observed to be important. These features are categorical variables that are one-hot encoded. As can be observed in the plot, the high feature values for each are associated with lower predictions, which means that these two features are negatively correlated with the price. Hence Units and Townhouses tend to have lower prices.

For the rooms feature, the plot indicates that higher feature values are associated with higher predictions which means that the rooms feature is positively correlated with the price. Hence, houses with more rooms tend to have higher prices as well. This just makes sense, since more rooms indicate that more materials and space are needed for the construction. Similarly, Houses with higher bathroom count, higher car parking spot count, and larger land size tend to have higher prices.

Finally, the SHAP values for features Region Name Northern Metropolitan, Western Metropolitan, and Western Victoria were also analyzed. All these features are negatively correlated with the price. Hence, houses located in these regions tend to have lower prices.

Based on these feature importance values, we can say that generally a family that would require a lot of rooms, such as a large family, are at a disadvantage in terms of house prices as they will likely require more money to invest in a house that will have more rooms. However, other things can be considered to try and offset the likelihood of having a relatively higher price.

A family can consider living farther from the city center as the model shows that houses outside of the city center tend to be cheaper. With this trade off, the family must consider that at the cost of housing price they might require more time in their travels to activities at the city center and also consider the possible cost of travel before considering how much overall cost is offset by this aspect of the house. Another consideration the family

can have is deciding to live in a townhouse or unit type building, if they do decide that living in the city center is essential. Another consideration that the family can have is choosing to live in the mentioned cheaper regions, regardless of location relative to city center, given how it is identified that houses are cheaper there. The family needs to consider the possible impact of this to their lifestyle if they decide on this to offset the price. Lastly, the family can consider looking for housing units that do not have garages to keep cars as this house feature is also seen to increase the house's price; with this the family does need to consider their form of transportation in their daily life and the corresponding cost that comes with this compromise.

IV. Conclusion

Prediction of House Prices can be helpful especially for new house owners looking to maximize their budget. After the modeling experiments conducted during study, a machine learning model that could predict the house price given the characteristics of a house listing was created. The most important features from the trained Elastic Net model are Distance from CBD, Type, and Count of Rooms. The insights that were generated from analyzing the SHAP values of the model can be considered in the decision making of the homeowners to maximize the value of the house to be purchased. For instance, since large families have a disadvantage due to the room count, trade-offs with the other features can be considered to have a lower house price. Small families on the other hand have more freedom on the other aspects due to the low room count requirement.

V. References

[1] Eslake, Saul. 2007. 'An Introduction to the Australian economy', Working Paper, ANZ Bank Limited.

[2] Oliver, Shane. "Why Is Australian Housing so Expensive and What Can Be Done to Improve Housing Affordability?" AMP Capital, 29 Sept. 2021, https://www.ampcapital.com/au/en/insights-hub/articles/2021/september/wh-is-australian-housing-so-expensive-and-what-can-be-done-to-improve-housing-affordability.

[3] Marsh, Stuart. "Sydney Property Prices Rise by $620 A Day." 9News Breaking News, 9News, 14 Oct. 2021, https://www.9news.com.au/national/australia-property-affordability-crisis-sydney-prices-grow-620-a-day/c346591c-fa00-43c8-a832-e6827e704ae5.

[4] "Melbourne Housing Market Insights: November 2021." The Urban Developer, The Urban Developer, 25 Nov. 2021, https://www.theurbandeveloper.com/articles/melbourne-housing-market-update.

[5] “Property Prices Higher than If Covid-19 Had Not Happened.” KPMG, KPMG, 12 July 2021, https://home.kpmg/au/en/home/media/press-releases/2021/07/property-prices-higher-covid19-had-not-happened-12-july-2021.html.

[6] Farina, Gino. “7 Factors That Influence a Home Buyer's Decision by CoreLogic.” LinkedIn, LinkedIn, 10 Mar. 2018, https://www.linkedin.com/pulse/7-factors-influence-home-buyers-decision-corelogic-gino-farina/.

[7] Pino, Tony. “Melbourne Housing Market.” Kaggle.com, 2018, www.kaggle.com/anthonypino/melbourne-housing-market.

[8] Hastie, Trevor, et al. *The Elements of Statistical Learning, Second Edition : Data Mining, Inference, and Prediction*. New York, Springer, 2009.

[9] Lundberg, Scott. “Shap”. PyPI, 21 Oct. 2021, https://pypi.org/project/shap/