**Predicting Real Estate Prices in Melbourne**

Lincoln Brown and James Mulvihill

Bellevue University

Predictive Analytics - DSC630-T302

Professor Hua

**Introduction**

One key question that realtors need to ask is how much can I market this home for? Getting the price right could determine how fast it sells and how much commission the realtor gets. Since we have property taxes in the United States, we also have an assessed value of each home. This can often be used as a benchmark to pick a competitive price. However, when marketing homes in other countries that do not have an assessed value, determining price can become trickier. This is where predictive analytics can shine by looking at comparable homes and their features that have sold recently to determine what price is reasonable for the next home.

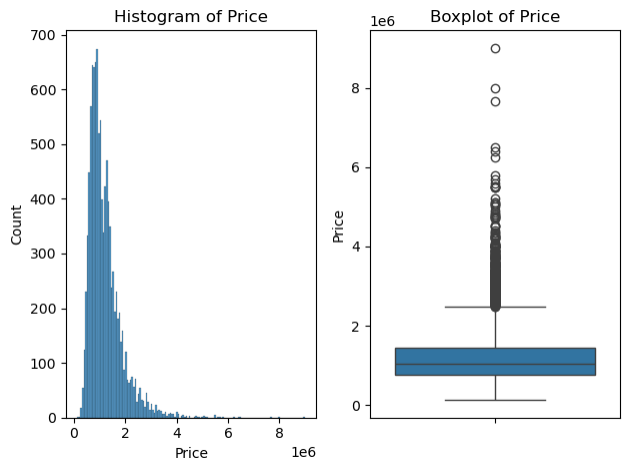
The customer that stands to gain from our project is Expanse Realty: a realty firm that takes a data-centric approach to selling property and is considering expanding into the housing market in Melbourne, Australia. Melbourne is a rapidly growing city that for many years was recognized as the most liveable city in the world (Wahlquist, 2020), and has consequently had a huge amount of real estate transactions. We will be marketing our data science services to Expanse Realty in our PowerPoint presentation.

**Data Selection**

We will be looking at data from the Melbourne housing market as it was in 2017. The data we will be using comes from Kaggle (*Melbourne Housing snapshot*, 2018) uploaded by user Tony Pino. The dataset has over 13,580 records and 21 features that include the suburb the house is in, number of rooms, size, year built and the all-important target variable - price. Part of the reason we chose this dataset is because it had more records and was more complete than most other real estate datasets on Kaggle. Also, doing this project for a different country that does not have assessed value meant we had more of a challenge doing the modeling without this strongly correlated feature.

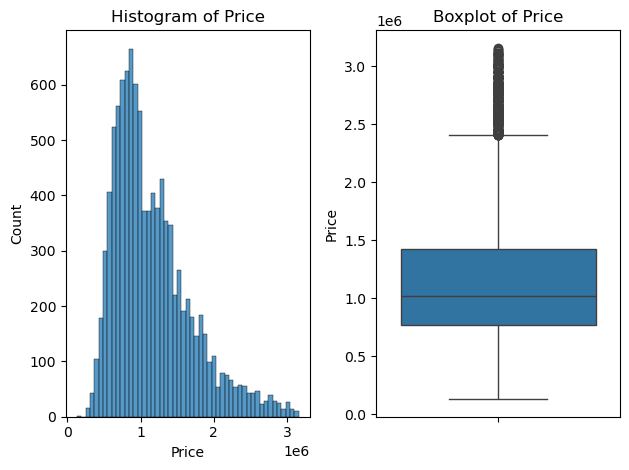
There were multiple steps we needed to take to get the data fit for modeling. First, we were only interested in houses and townhouses, so we needed to remove all the apartments from the dataset. Since the data came from Australia, the dates were listed in day/month/year format, so we converted that column to datetime. Since there were a significant number of missing values in BuildingArea, YearBuilt, and CouncilArea columns we chose to remove them. Additionally, since CouncilArea is mostly determined by Suburb, we dropped that column. After removing these columns our dataset had 10,563 records.

Our dependent variable ‘Price’ was the focus of a lot of our efforts. Unsurprisingly, the price had a positive skew due to the presence of outliers in the right tail of the dataset. We can see the evidence of this skew and the presence of outliers in the histogram and boxplot below.



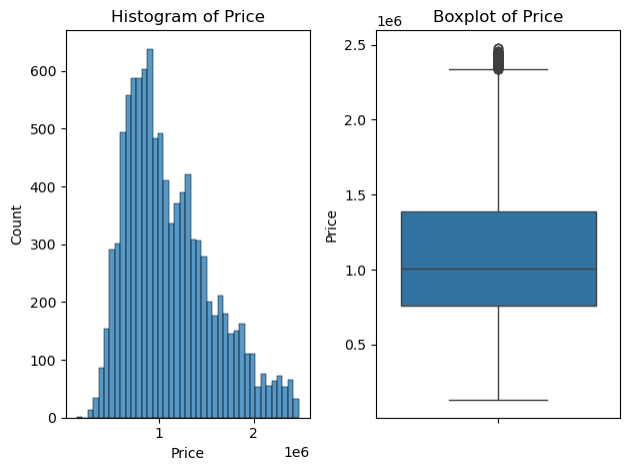
From the histogram above we can see that the dataset has a positive skew. Also, from the boxplot, we can determine there are many outliers in the higher price range. To correct this, we knew we needed to eliminate some of the outliers in our dataset to the best of our ability without affecting the quality of the data. We tried two different methods for outlier removal z-score and interquartile range.

The first method we tried was the z-score method, which measures how many standard deviations a data point is away from the mean. We calculated a z-score value to each record in the dataset and then dropped any outliers that were greater than or less than 3 standard deviations of the mean. After the outliers were identified, we were left with a distribution that looked like this.



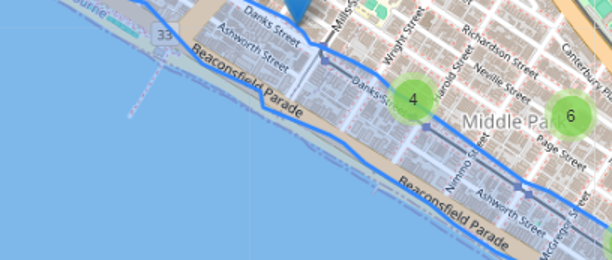
We can see there are many outliers still present in the data. For this reason, we decided to see the results of our second method for outlier removal, the interquartile range.

The process we used to employ the IQR method for identifying and removing outliers was as follows. Initially, we found the difference between the 75th percentile (Q3) and the 25th percentile (Q1). From this interquartile range, we decided to eliminate outliers that exceeded a threshold of 1.5 times the IQR. Thus, any data points that fall outside below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR were considered outliers and were removed from the dataset. After eliminating the outliers using the IQR method, the distribution of Price can be seen in the graphs below.

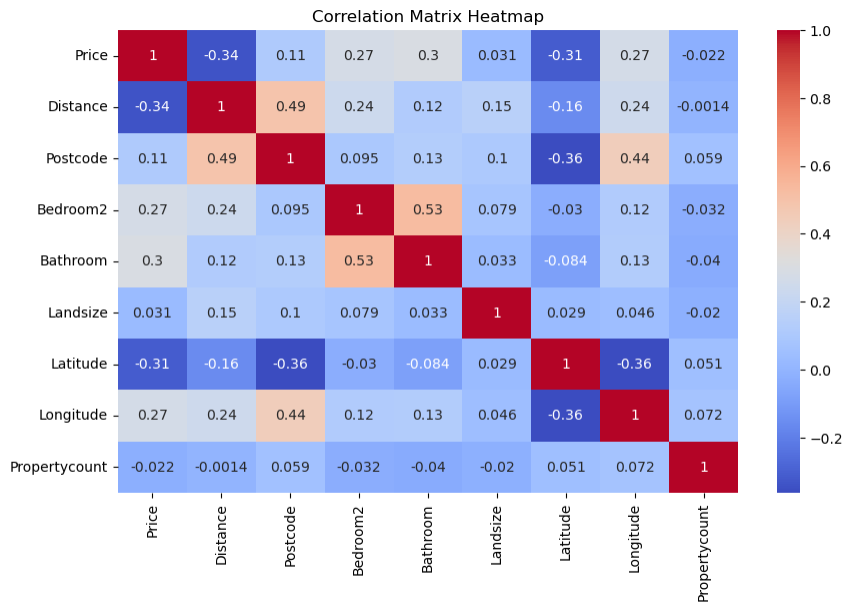


From the boxplot above, we can see that there are still some outliers in the dataset. However, we elected to keep these data points because they were still within the 1.5 threshold that we established earlier. Although we still do not have a normal distribution for the Price feature, we are comfortable enough with the improvement made by dropping the outliers to continue with the analysis of the dataset.

With Melbourne being a coastal city, we thought it could make a significant difference to the home price if the property is located next to the coast. Adding a Boolean column to the dataset for this required getting shapefiles from the United States Geological Survey website to get the water boundaries for the area near Melbourne. We then had to stitch the shapefiles together and select only the boundaries for Port Phillip Bay, our target water feature. After this was accomplished, we needed to make a buffer zone around the water feature to determine which properties fell within approximately 200 meters of Port Phillip Bay. At the end of our efforts, we created a feature titled “IsWaterfront,” that had 49 houses listed as waterfront properties.



Since we wanted to establish a linear regression model as our baseline, we decided that it would be beneficial to understand the correlation between the individual features in the dataset. As part of the visualization process, we created a correlation matrix to generate a heatmap to observe the correlation between the individual features visually.



From the heatmap above, we can see that Distance has a moderately negative correlation with Price with a coefficient of -0.34. In other words, the farther the distance from the City Center, the lower the price becomes. The next strongest correlation we see is latitude, also with a negative correlation of -0.31. This indicates that the higher the latitude, the lower the sale price of the house will be generally. Additionally, there is a moderate positive correlation with the price of a property and the number of bathrooms. Indicating that houses that have more bathrooms tend to have higher sale prices.

**Modeling and Methods**

For model creation, we chose to create three models, linear regression, decision tree, and a random forest. Since we knew that there were not a lot of highly correlated features with the sale price, we did not expect impressive performance out of the linear regression model. However, we created this model so that we had an effective baseline to compare with the performance of the other two models.

For model evaluation, we decided to use several evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2 score.

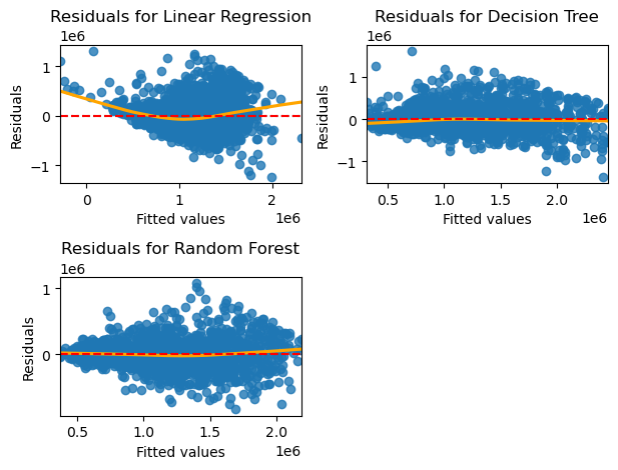
* Root Mean Squared Error (RMSE) - Mean Squared Error (MSE) is the metric that measures the average of the squared differences between the predicted and actual sale price (Brownlee, 2021). RMSE is the square root of MSE. RMSE penalizes large errors, which makes it sensitive to outliers in the data. Lower RMSE values indicate better performance. RMSE values are interpreted in the context of the dependent variable.
* Mean Absolute Error (MAE) - This metric is a measure of the average absolute difference between the predicted values and the actual values in a dataset (Zach, 2021). MAE values are measured in the same context as the dependent variable, making them easy to interpret. In the case of sale prices, the MAE represents the average absolute difference between the predicted sale price and the actual sale price.
* R2 Score - The R2 score, or Coefficient of Determination is a measure of goodness of fit. The R2 score explains how much of the proportion of variance in the dependent variable is predictable from the independent variables Rowe, n.d.). R2 scores are measured between 0 and 1, with 1 indicating a perfect fit. Perfect R2 scores are a strong indication of overfitting.

Below are the performance metrics of each model.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | MAE | R2 |
| Linear Regression | 301296.443979 | 225158.913033 | 0.550277 |
| Decision Tree | 302648.094235 | 211754.803775 | 0.546233 |
| Random Forest | 202831.118469 | 144277.653145 | 0.796190 |

In addition to evaluating the performance of these metrics, we also evaluated the model’s performance by plotting the residuals. Residuals are the difference between the actual values found in the dataset and the predicted values of the model.

Residual Plots



**Interpretation of the Results**

From the table of each model’s metrics, we can see that the random forest model performed the best. It achieved an R2 score of 79% and had an RMSE of $202,831.12 and a MAE of $144,277.65. An R2 score of 79% means that 79% of the variance in the dependent variable can be explained by the independent variables. In addition, the price feature from the dataset has a mean of $1,111,655.61 and a median of $1,005,000.00. To understand the scale of the errors a little better, the RMSE divided by the mean is .18 and the MAE divided by the median is .13. Therefore, we can conclude that our model’s errors can deviate by up to 18% of the mean and 13% of the median. Knowing these numbers does not help prove the predictive capabilities of the model, but it is useful to understand the scale of deviations in predicted prices.

Regarding the residual plots above, we can see several different things. First, the blue dots represent the residuals of the specific data points. These help us to understand where the model’s errors are occurring regarding the predicted sale price (x-axis). Next, we see a red dotted line across the center of the graph. This is known as the zero residual line and all data points would fall on this line in a perfect prediction scenario. Ideally, the points should fall as close as possible to this line. The last aspect of our graph is the solid orange line. This is a line generated by Locally Weighted Scatterplot Smoothing (LOWESS). LOWESS is a non-parametric strategy to fit a smoothed curve to capture the trend between the fitted values and the residuals. This allows us to view the trend of the residuals more easily.

Now that we have described the graph's components, let us begin our analysis. In the linear regression residual plot, we see that the smoothed curve bows upward at the lower and higher ends of the predicted price. This indicates that the linear regression's performance suffers at the graph's left and right sides. When looking at residual plots, patterns in the errors can indicate problems with the model. Clearly, the linear regression model is not our best predictor.

In the context of the decision tree residual plot, the smoothed curve performs well, with a small sag in the lower end of the predicted prices. Outside of this, the smoothed curve does not reveal any noticeable patterns. Incidentally, we see some residuals plotted higher towards the left side of the graph but tend to cluster towards the middle and then are plotted lower on the right side of the graph. This indicates that the model may suffer when predicting low and high prices.

Next, we will analyze the residual plot for the random forest model. This model’s smoothed curve also looks fairly straight with a slight sag in the middle of the graph and then tipping upwards on the right-hand side. Interestingly, this indicates higher residuals near the center of the graph, which is confirmed with the plotted residuals. Furthermore, we also see residuals plotted near the bottom of the graph on the right-hand side, indicating issues with the predictions on the higher prices as well.

**Conclusion and Recommendation**

Overall, we can see that the performance metrics for the random forest model indicate that it has the best performance. This is reinforced with a fairly straight smoothed curve on the residual plot. Looking at the residual plots of the decision tree and random forest models, we do not see noticeable differences in the model’s capabilities. However, when we account for the performance metrics, we can see that the random forest model is a better predictor. Given that the performance metrics of the random forest model are good, but not great, it is evident that there is room for improvement with this model. However, for this project's purposes, we are satisfied with its performance.

We believe that this model will benefit Expanse Realty’s efforts to break into the Melbourne housing market. The model’s performance is best for the Metropolitan area in the average house price range, which means that they should primarily focus on this area initially. Specifically, the Southern Metropolitan region had the highest average sale price, so focusing on this area could bring the highest sale prices. Incorporating more data points for the outer regions would be the best way to improve the model’s performance in these regions. However, these regions tend to have lower sale prices, so any expansion into the regions would need to be carefully considered. Overall, we are happy with the performance of the model and believe it will be a beneficial asset for Expanse Realty.

**References:**

Brownlee, J. (2021, February 15). *Regression metrics for machine learning*. MachineLearningMastery.com. <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>

*Melbourne Housing snapshot*. (2018, June 5). Kaggle. <https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot>

Wahlquist, C. (2020, February 3). Melbourne “world’s most liveable city” for seventh year running. *The Guardian*. <https://www.theguardian.com/australia-news/2017/aug/16/melbourne-worlds-most-liveable-city-for-seventh-year-running>

Zach. (2021, October 4). *MAE vs. RMSE: Which Metric Should You Use?* Statology. <https://www.statology.org/mae-vs-rmse/>