Domain

HOUSE PRICE IN NSW

Data Analysis & Prediction

Domain Real Estate
By: Brian Ong

BACKGROUND & GOAL

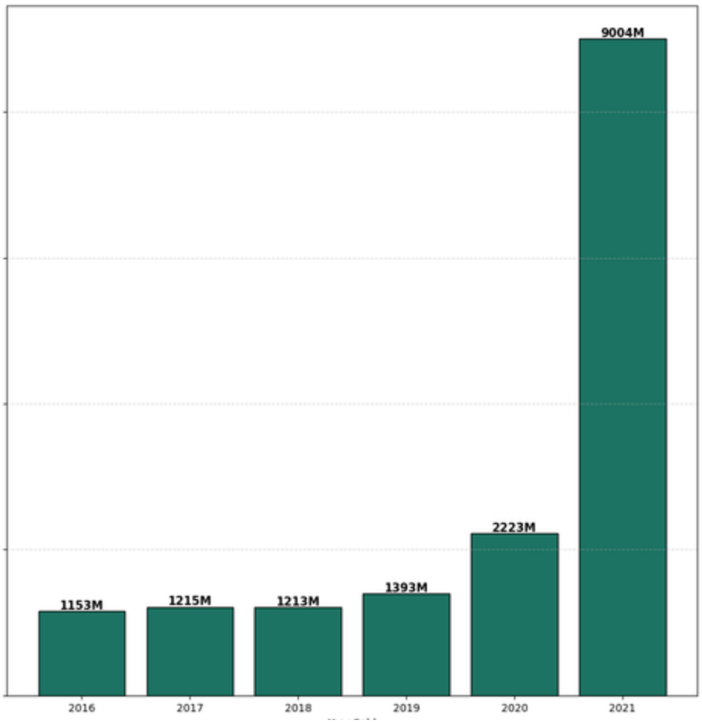
- The real estate industry remains one of the most attractive sectors for both investors and homebuyers, as owning property - whether as an investment or a family home—is a widely shared aspiration. Domain, one of the leading companies in the real estate market, has experienced significant growth, particularly with property prices climbing dramatically in recent years. Real estate encompasses various property types, including houses, apartments, and vacant land, and its trends often reflect the broader economic conditions of a country. Understanding these shifts is key, and my house price prediction project, focusing on Domain's data, aims to shed light on these developments and forecast future market behavior.

SUMMARY

- House pricing is typically influenced by several factors, but in this dataset, the focus is on external factors like suburb characteristics and internal house features such as the number of bedrooms, bathrooms, and parking spaces. This approach helps reflect how the surrounding environment influences buyer trends in the housing market.
- After cleaning the data and removing outliers, the project performed EDA (Exploratory Data Analysis) based on two key insights. It also visualized the relationship between population, elevation, suburb income, and other features with price per square meter, providing a clearer picture of how these variables impact property value.
- Additionally, the project investigated a business question of Domain Dataset: Does increasing the number of rooms always lead to higher prices? Using statistical evidence, the analysis revealed that—even though the dataset is imbalanced for homes with a high number of rooms—there is over 15% support for this relationship.
- Finally, four models were applied: two time series models and two tree-based models, with and without feature engineering. This comparison highlighted the importance of AI and machine learning in price prediction and demonstrates how feature engineering significantly boosts model performance.

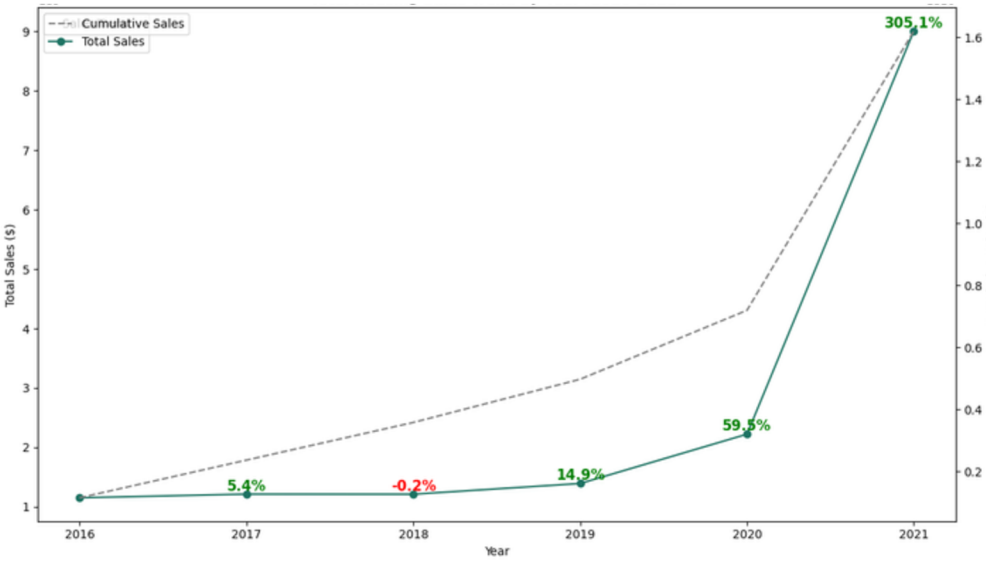
How fast is the market growing?

- Revenue of real estate properties increased constantly from 2016 and dramatically rise up since 2020, reached **$9 Billion** in 2021
  - **Figure 1** - Revenue of property over years



305.1%

was **the increase** in Domain's revenue in 2021 compared to 2020



- The rise up highlighted a significant shift likely driven by COVID-19's impact on the housing market. This surge suggested a potential house pricing crisis, fueled by post-pandemic demand, changing buyer preferences.
- **Figure 2** -Total Housing Sales by Years (2016-2021)

88.4% of real estate revenue came from **houses**, making it the **dominant segment** in the market

Apartments contributed only a **4.6%** of total revenue, which suggests investigating whether **it could potentially increase**.

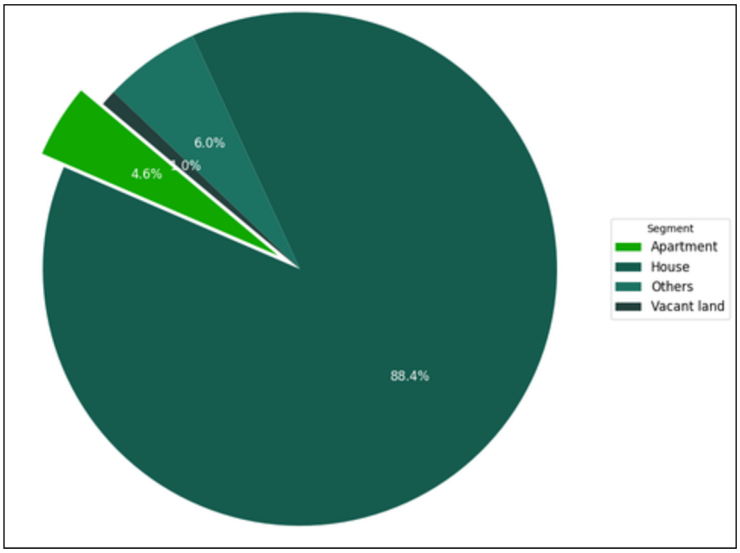


Figure 3 - Real Estate Revenue Distribution

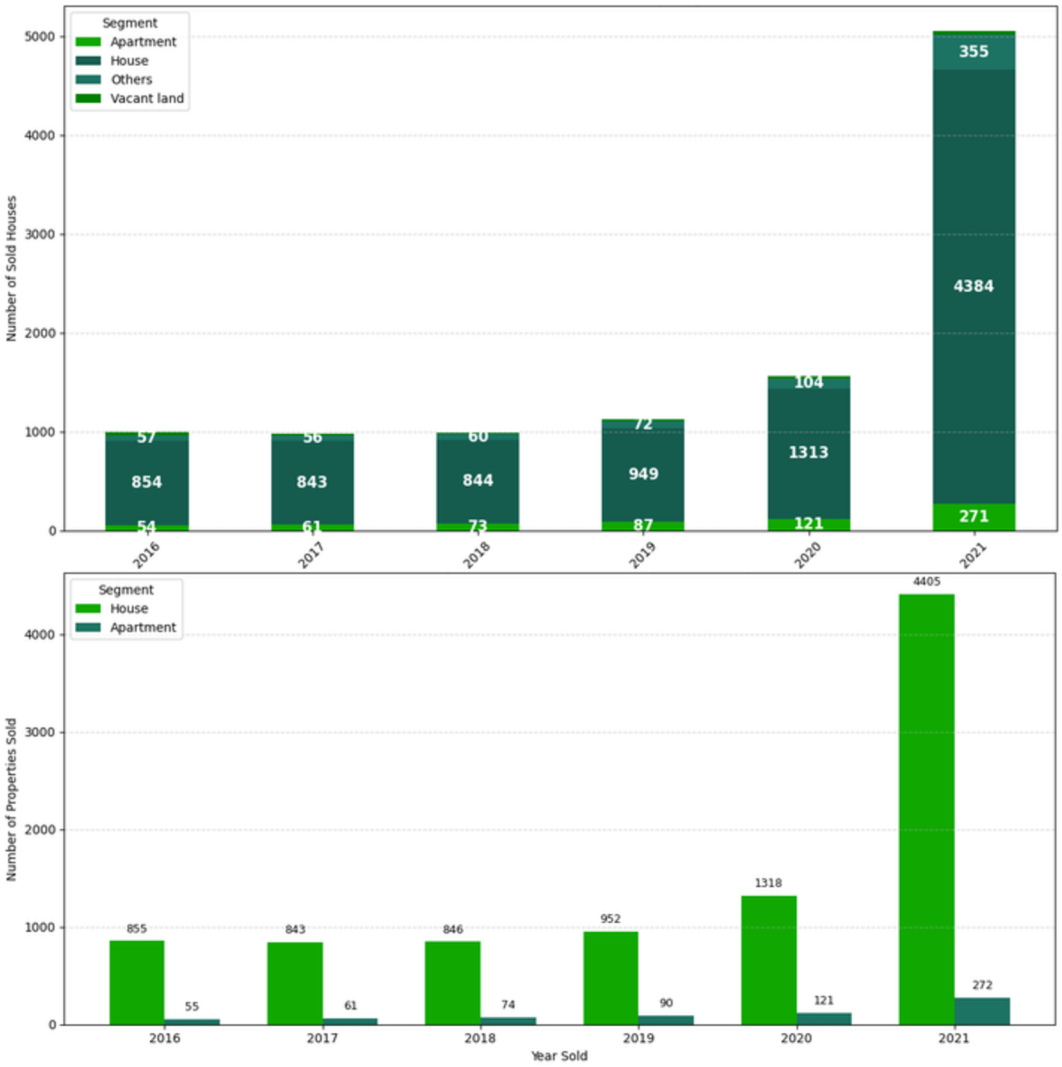


Figure 4 - Sold Properties Distribution over years

3.3x

Number of **houses sold** in 2021 was 3.3 times higher than in 2020.

2.2x

Number of **apartments sold** in 2021 was 2.2 times higher than in 2020.

1.63x

- The **average apartment price increased dramatically** in 2021, reaching 1.63 times the average price in 2020.
- Before this, apartment prices had shown a declining trend, but 2021 marked a strong rebound, accompanied by a **rise in the number of transactions**.

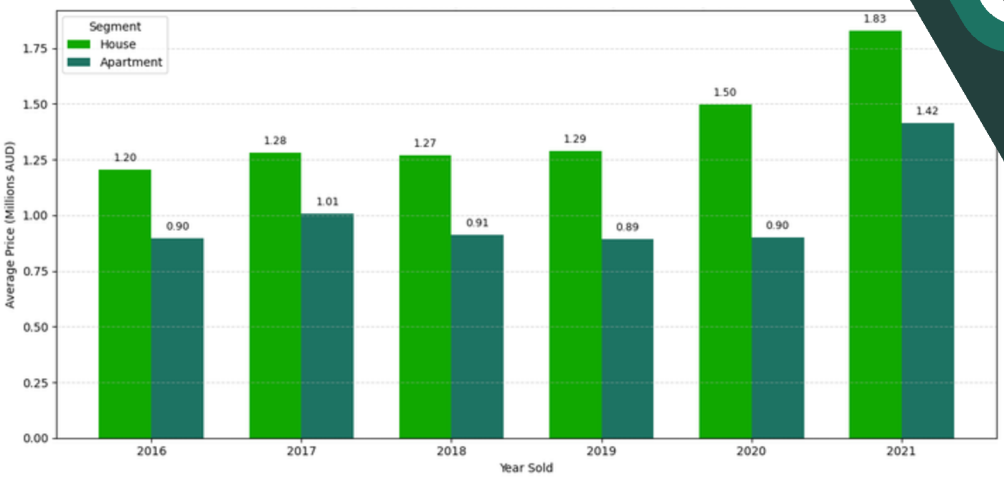


Figure 4 - Average Price of House and Apartment (2016-2021)

Which suburb offers the best investment potential?

Top 10

- Suburbs with **the highest price per square meter in NSW** include Surry Hills, Paddington, Darlinghurst, and others.
- Notably, the first two suburbs had average price exceeding **$20,000** per square meter, highlighting their premium market positioning and **high demand for inner-city living**.

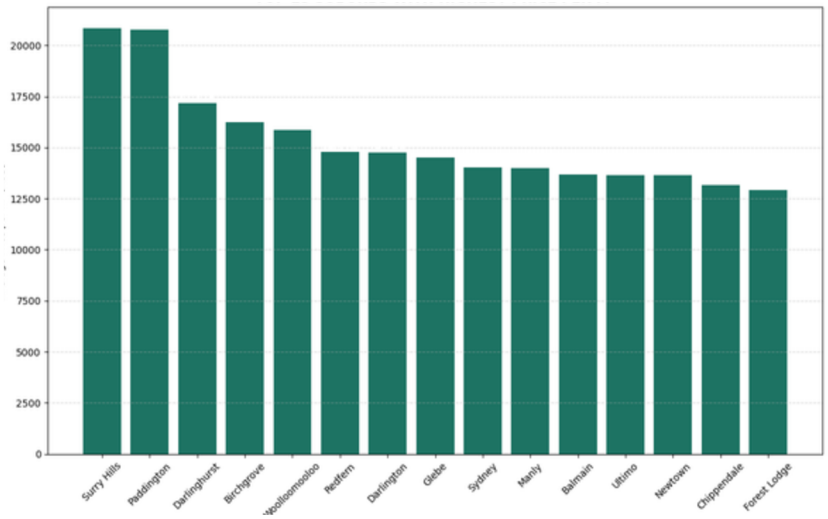
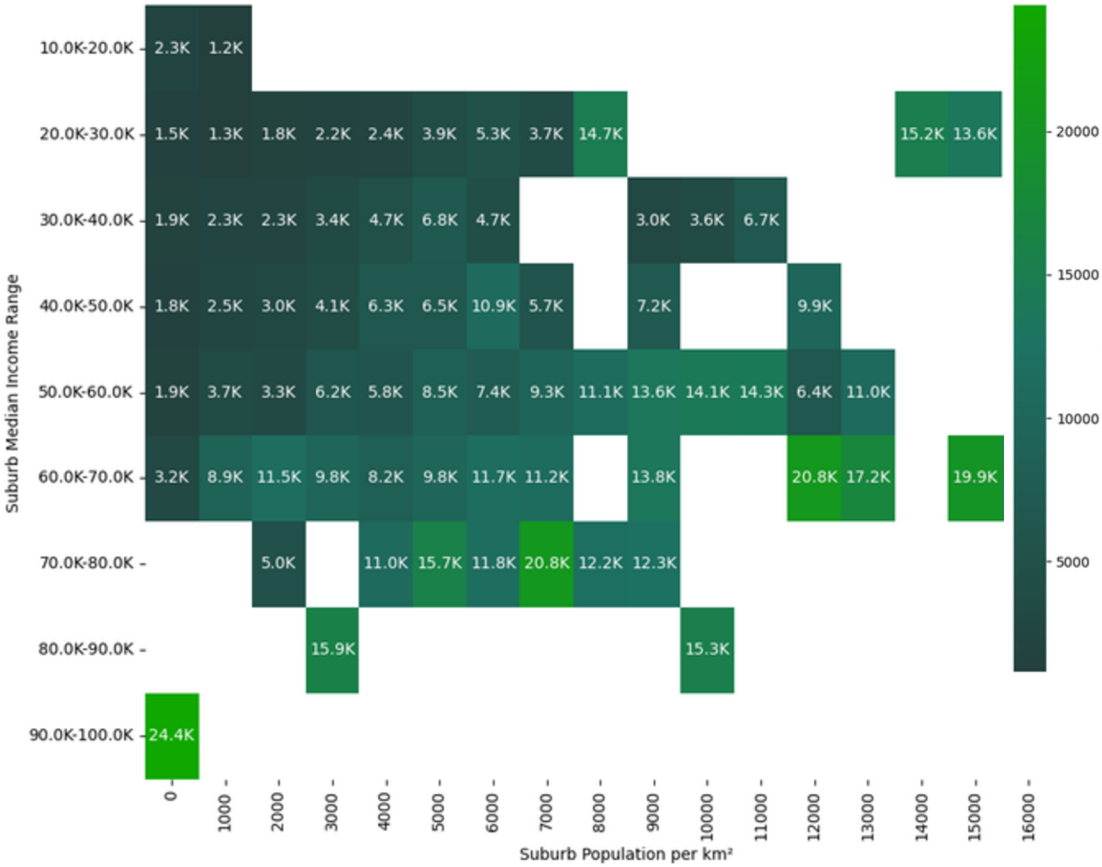


Figure 4 -Top 10 "expensive" suburbs

Which factors significantly impact in house price ?



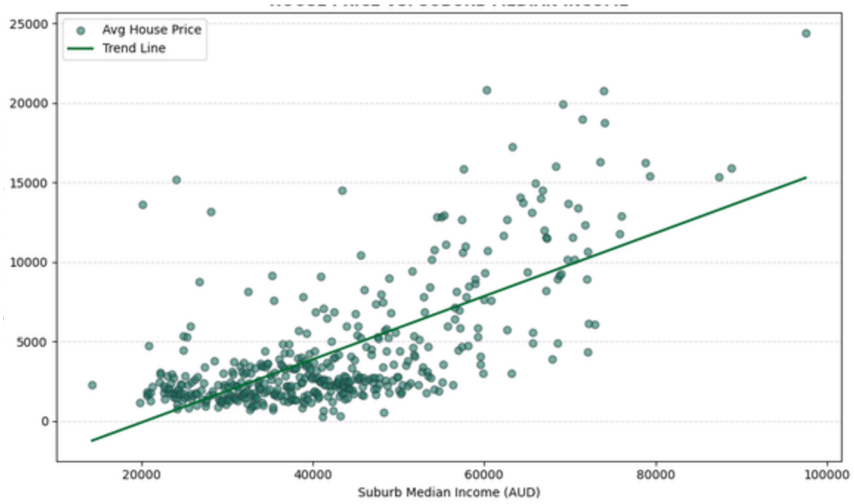
$24.4K

is the **average amount of money** to own the property in the **most expensive suburb** in NSW, despite having a population density of less than 1,000 people per km².

Additionally, **suburbs with a median income above $50,000** tend to exhibit **higher house prices**, suggesting that affluence plays a strong role in property valuation.

- Moreover, while **it's often assumed that more crowded areas may lead to higher prices**, our data suggests **otherwise** — prompting a deeper analysis into population density and its true impact on housing value

Suburb median Income



0.447

is **correlation coefficient** between suburb median income and price per square.
Suburbs with **higher median income** tend to have significantly **higher property prices per square** meter, indicating a strong positive correlation.

Figure 4 - Price per square & Suburb Median Income Relationship

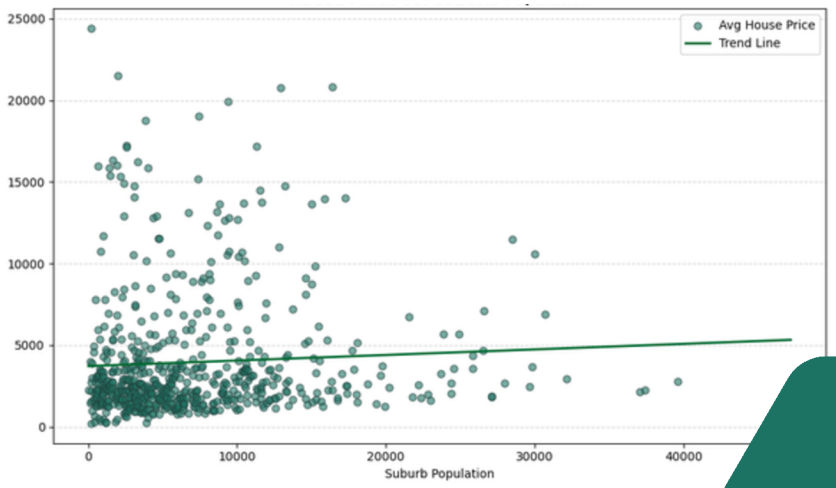
---

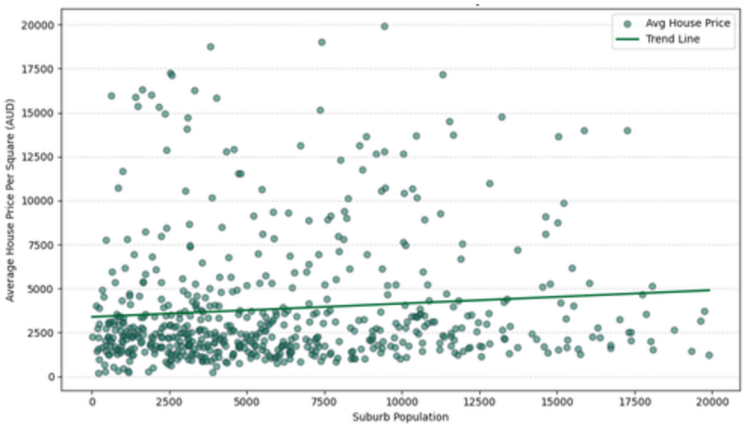
0.043

Indicating **small impact** between suburb population and property price.
This suggests that **population size** alone **does not significantly influence housing prices**, and other factors like income, location, or land availability likely play a more critical role.

Figure 4 - Price per square & Suburb Population Relationship

Suburb Population

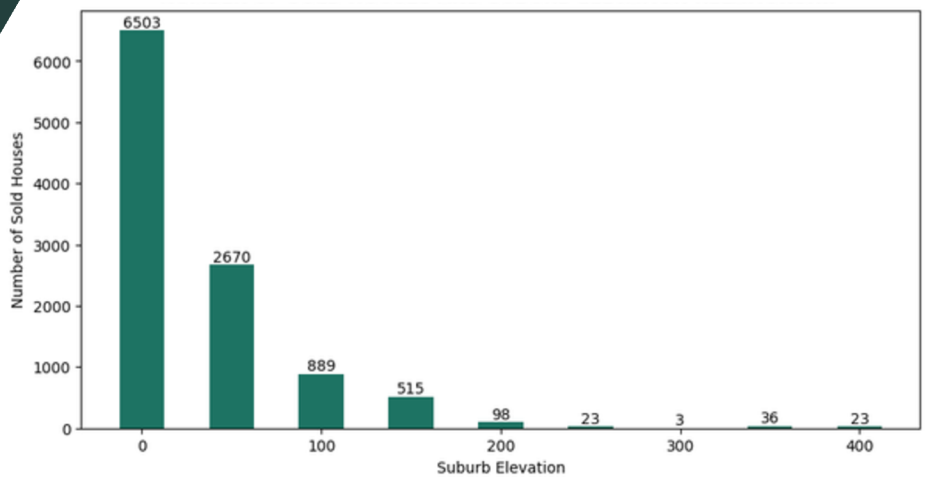


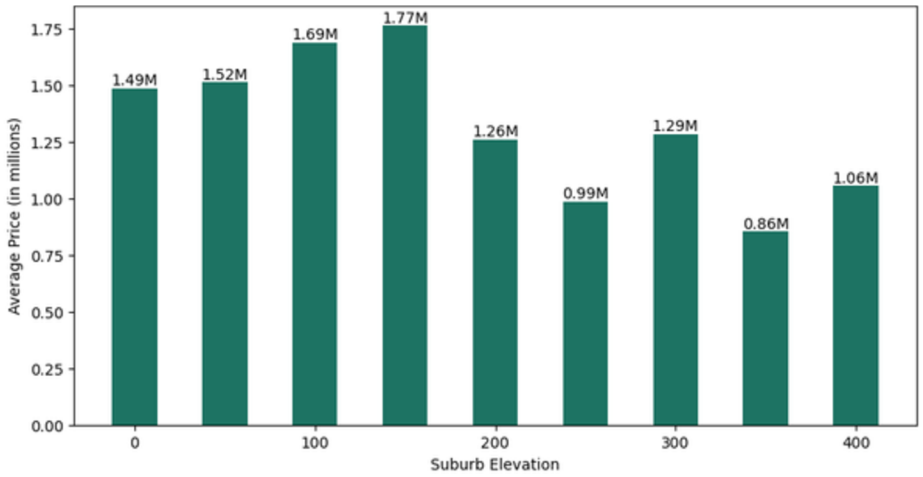


Similar observation with remove extreme data point over 20.000

Figure 4 - Price per square & Suburb Population Relationship (Extreme data point remove)

Suburb Elevation

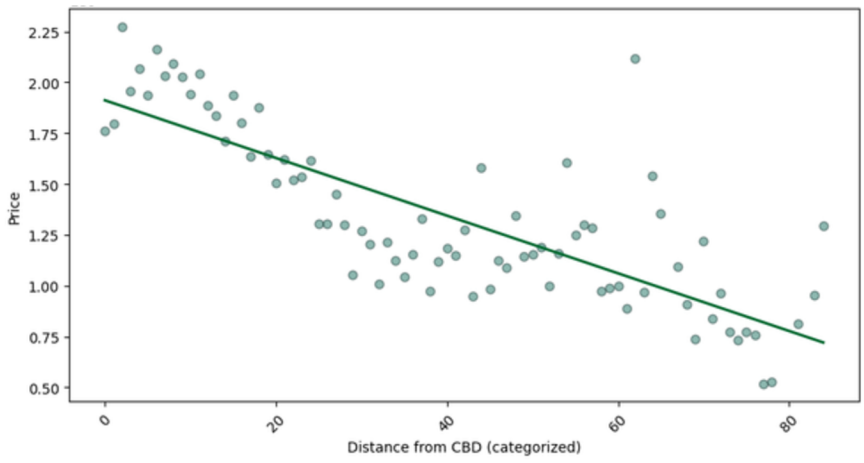




- **The number of houses sold decreases as elevation increases.**
- Additionally, the **average price of houses at elevations of 150 meters and lower** seems **higher** than those at elevations over 200 meters.

Figure 4 - Average price and elevation relationship bar chart

CBD Distance



-0.464

The trend line indicates a **negative correlation**, suggesting that **properties closer to the CBD** tend to have **higher prices**

Figure 4 - Price and Distance from CBD relationship

The more bathroom and bathroom, the more expensive property is?

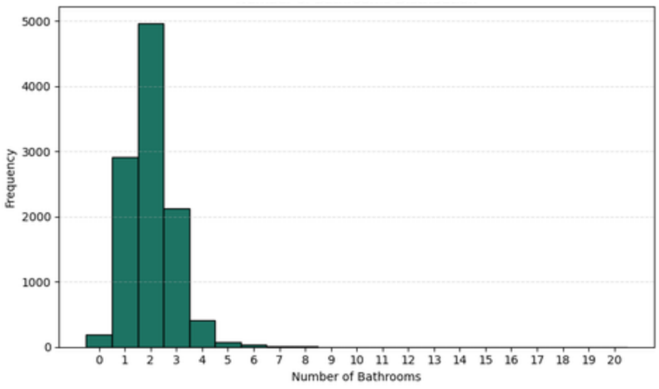


Figure 4 - Bathroom distribution

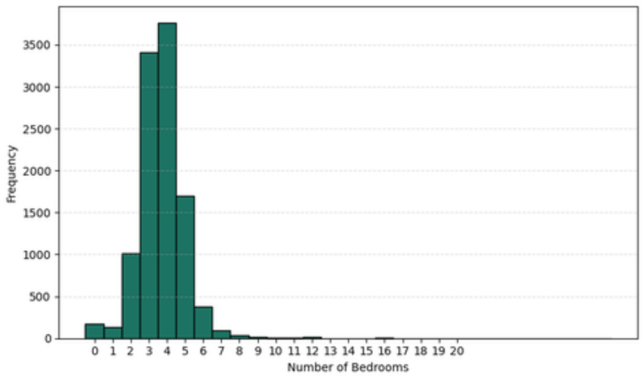


Figure 4 - Bedroom distribution

2-5

This is the most popular range for the number of bedrooms. The heatmap also shows that properties in this range tend to have higher price per square meter compared to others.

1-3

This is the most common range for the number of bathrooms. The highest price per square meter (20K) is observed in properties with 1 bedroom and 2 bathrooms. However, since there are fewer than 100 such properties, it is considered a special case.

95.89%

This is the percentage of properties with 2-5 bedrooms and 1-3 bathrooms in the Domain Company's historical records

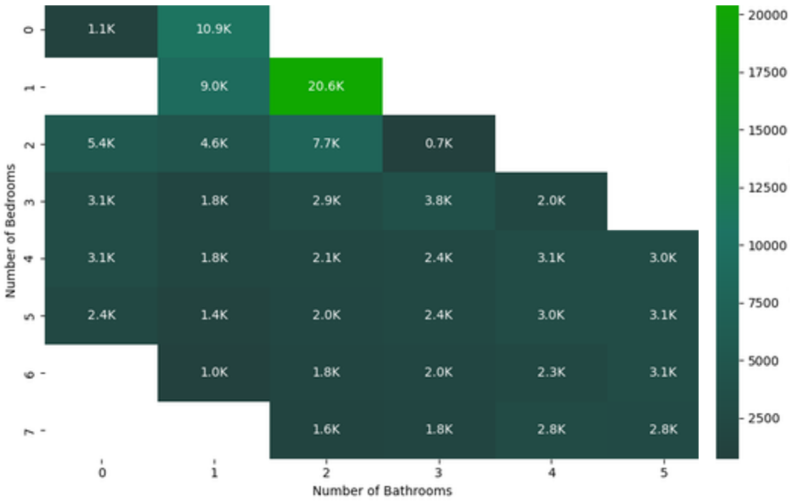


Figure 4 - Bedroom and Bathroom Heatmap (PPS)

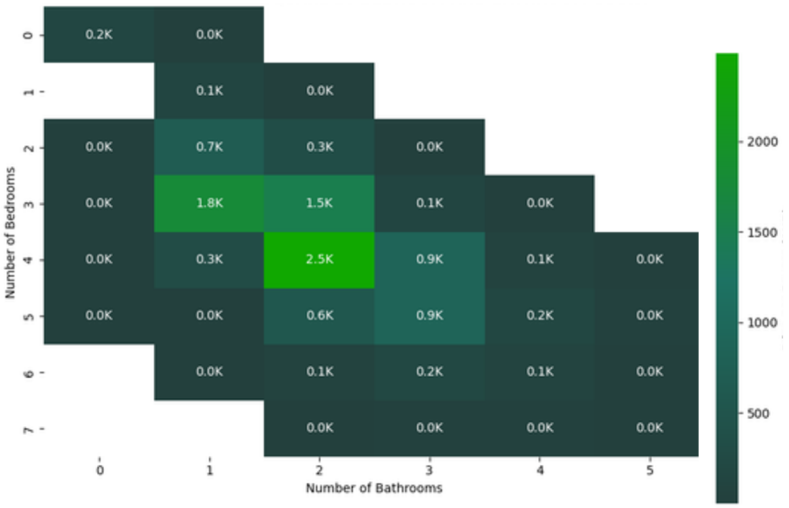
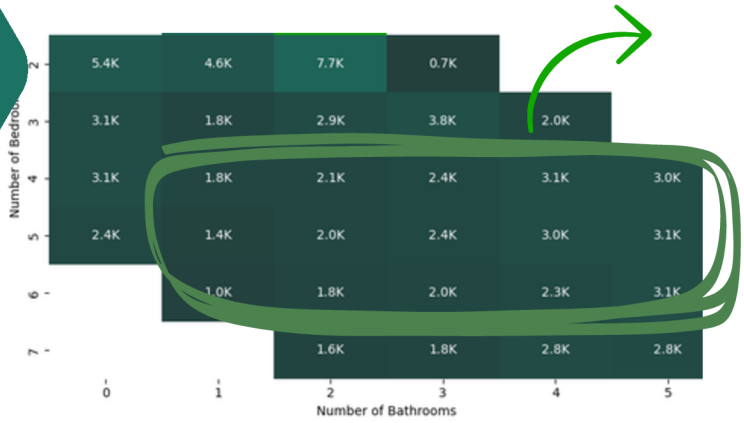


Figure 4 - Bedroom and Bathroom Distribution

*PPS: Price Per Square

Does Having More Rooms Really Increase Property Value?



As observed from the heatmap, **properties with 4–5 bedrooms** tend to have a **higher price per square meter** (PPS) as the number of bathrooms increases. However, this **trend is not consistent** across all bedroom ranges

Statistic Evidences

Conducted OLS Regression

- **Each additional bathroom increases the log of property price by 0.2178**, a **stronger** effect than an extra bedroom, which adds 0.0684.
- However, **the negative interaction coefficient of −0.0116** suggests **diminishing returns** when both bedrooms and bathrooms **increase together**.

Variable	Coefficient	Std. Error	t-Statistic	P-value	95% Confidence Interval
Intercept	13.5123	0.017	785.741	0.000	[13.479, 13.546]
Number of Bedrooms	0.0684	0.006	12.199	0.000	[0.057, 0.079]
Number of Bathrooms	0.2178	0.008	27.488	0.000	[0.202, 0.233]
Bed-Bath Interaction	−0.0116	0.001	−13.836	0.000	[−0.013, −0.010]

- While the model explains **15.4% of the variance in property prices** (R^2 = 0.154), all variables are **statistically significant (p < 0.001)**, confirming that **room count influences property value,** though other unaccounted factors likely play a role

MODEL TRAINING

Feature Engineering before training

- The [**suburb_mean_price]** feature captures the **overall pricing trend of each suburb**, providing valuable local market context. It helps smooth out anomalies from individual property listings and reduces noise in the data. This engineered feature **enhances model performance** by incorporating location-based economic signals directly into the prediction process.
- **Merge Suburb Mean House Price dataset** into dataset before splitting

Suburb	Year Sold	Suburb Mean Price
0	2021	1,507,833.00
1	2017	2,265,000.00
1	2021	1,870,250.00
2	2016	940,000.00
2	2021	1,254,556.00

The Evaluation Metric

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad \text{Explained Variance} = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}$$

- y_i: actual value
- $\hat{y}_i$: predicted value
- n: total number of observations

- **Mean Absolute Error (MAE)** calculates Average absolute difference between true and predicted values.
- **Mean Absolute Percentage Error (MAPE)** shows us how far predictions are from actual values, in percentage terms.
- **R^2 Score** (Coefficient of Determination) is the proportion of variance in the target explained by the model.
- **Explained Variance Score:** Similar to R^2, it shows how well the model captures variance patterns, regardless of a consistent offset (bias)

ARIMA Model (Time-series)

- Conducted an ARIMAX model that extends the traditional ARIMA by including **exogenous variables** like property features and suburb statistics. This allows the model to **update the ARIMA intercept (p, d, q)** at each time step based on external inputs, helping it **adapt more accurately** to real-world factors affecting house prices.

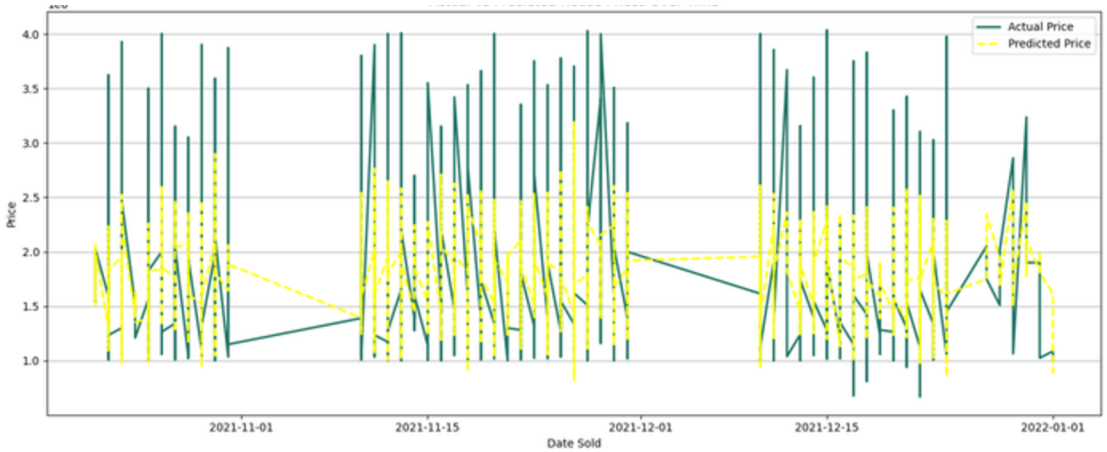


Figure 4 - Actual vs Predicted House Prices Over Without FE (*)

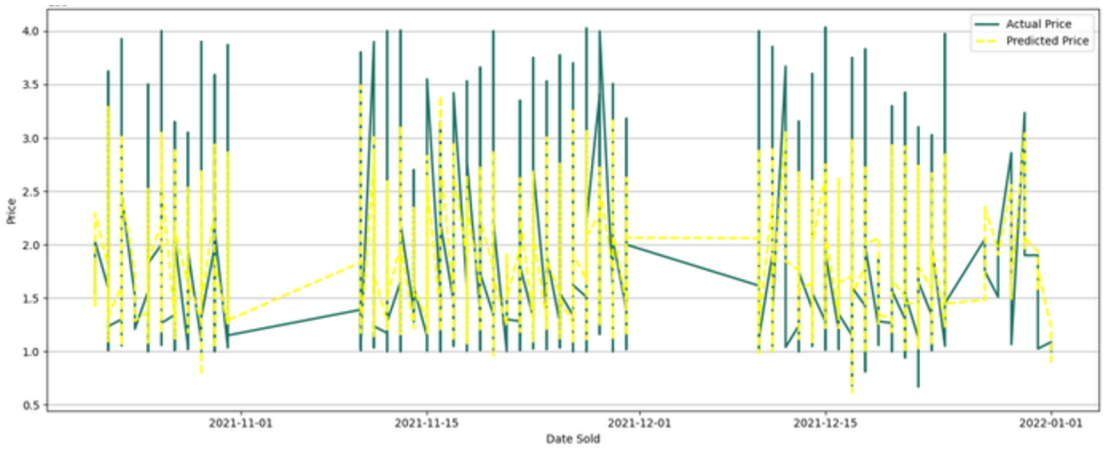


Figure 4 - Actual vs Predicted House Prices Over With FE (*)

Metric	With feature Engineering	Without Feature Engineering
R2 Score	0.2971	0.5994
MAE	443,456.30	309,810.06
MAPE	26.82%	18.2503%
Explained Variance Score	0.2997	0.6006

LSTM Model (Long Short-Term Memory)

- Building LSTM model, which is a special type of Recurrent Neural Network (RNN). LSTM can be used to **learn from historical data sequences**, such as price trends over time. So it can predict future values by **recognizing both short-term and long-term patterns** in the data

Metric	With feature Engineering	Without Feature Engineering
R2 Score	0.6868	0.7835
MAE	298,004.08	235,246.50
MAPE	21.01%	16.11
Explained Variance Score	0.6891	0.7846

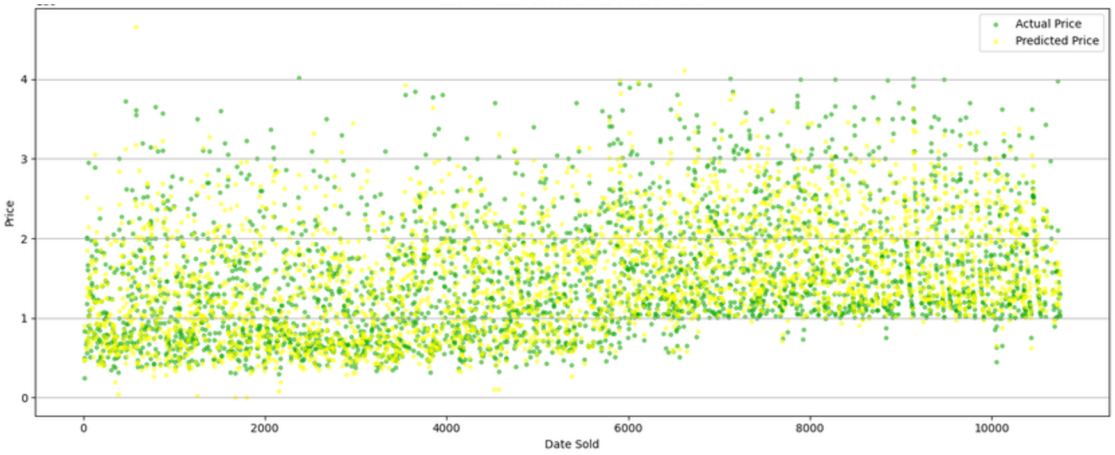


Figure 4 - Actual vs Predicted House Prices Over Without FE (*)

Random Forest Regression Model

- With a **Decision Tree model**, the **Random Forest** works as an ensemble method, combining multiple decision trees, **multiple random subsets** of the original training data. At each split, only a random subset of features is considered (**average target value is returned after traversing down**). Finally, it takes the **average of all tree predictions** to reduce overfitting and increase generalization performance, ultimately leading to a more accurate and stable prediction.

Metric	With feature Engineering	Without Feature Engineering
R2 Score	0.7549	0.8001
MAE	248,362.378	219,163.308
MAPE	16.63%	14.61%
Explained Variance Score	0.7553	0.8003

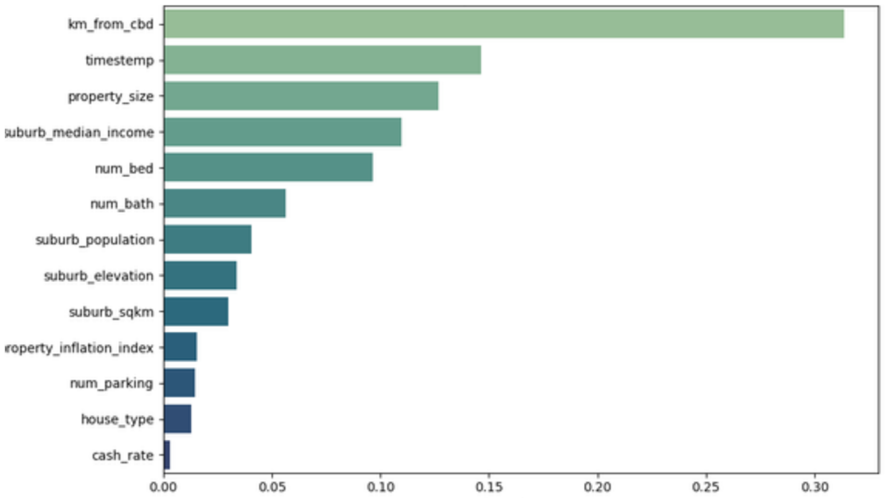


Figure 4 - Feature Importance Without FE

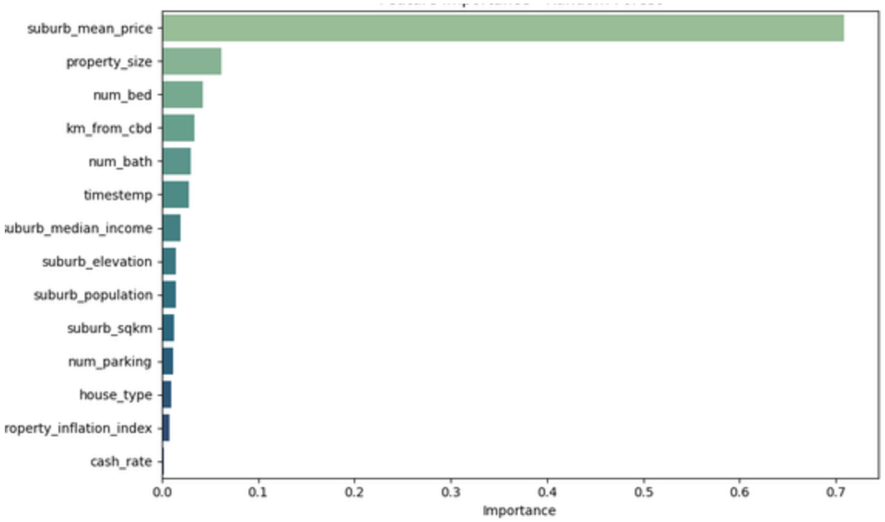


Figure 4 - Feature Importance With FE

Gradient Boosting Regression Model

- In this project, Gradient Boosting Regression model predicts house prices by **sequentially training decision trees** that **correct the errors** of previous ones. Unlike models that rely on averaging like Random Forest, Gradient Boosting builds **trees in a chain**, where each tree focuses on **minimizing the residuals** from the last. This allows the model to **capture complex patterns** in housing features and **improve prediction accuracy** over time.

Metric	With feature Engineering	Without Feature Engineering
R2 Score	0.7582	0.8157
MAE	252,412.351	210,774.812
MAPE	17.11%	13.97%
Explained Variance Score	0.7584	0.8158

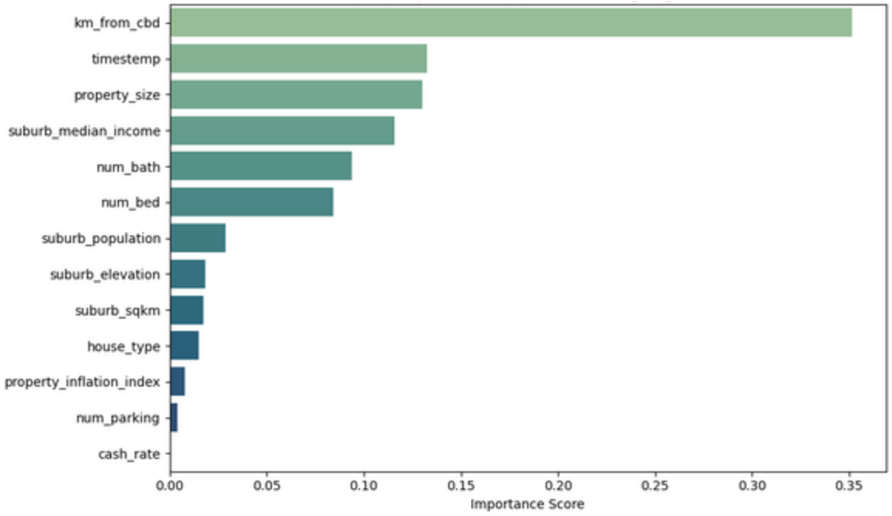


Figure 4 - Feature Importance Without FE

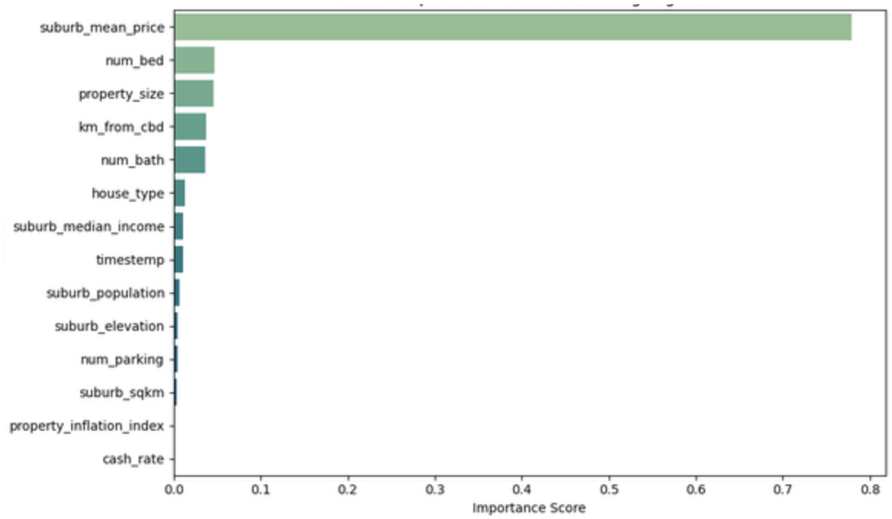


Figure 4 - Feature Importance With FE

MODEL SELECTION

- Based on the evaluation metrics, choosing Gradient Boosting Regression is the most appropriate decision for this project. It achieved the **highest R^2** score of nearly 0.82, which means that **the model is able to explain approximately 82% of the variability** in house prices — indicating strong predictive power.

- It also recorded the **lowest MAE of 210,000**, meaning that, on average, the predicted **house prices differ from the actual prices by only $210K** — a relatively low error given the scale of property prices. Furthermore, the model delivered the lowest **MAPE of 13%**, showing that the **average prediction error is only 13%** relative to actual values, making it a highly reliable model for estimating property prices.

INVESTMENT STRATEGIES

- These days, investing in property isn't easy. The market's competitive, unpredictable, and shaped by factors far beyond our control. But whether you're buying for your family or thinking long-term about investment, the key is to choose wisely based on what really matters: income levels in the area, elevation, and the overall environment — all of which we explored earlier in the analysis.

- When it comes to house features, 3–4 bedrooms and 2–3 bathrooms are the sweet spot, they show up most often in the data, making them less competitive and easier to find. Oddly enough, 2-bed, 2-bath homes top the list in price, likely because of their compact convenience, while options with 1 bathroom, 2–4 bedrooms tend to be more affordable.

- Of course, the future is hard to predict. Economic shocks — like the pandemic — triggered housing spikes, with prices soaring up to 300% in some areas. So, whether you're buying your dream home or looking for a solid investment, the smartest thing you can do is understand the trends, know your priorities, and trust the data.