House Price Regression and House Type Classification Report

Yi Yang

1 Challenges

1.1 Predicting property prices is inherently difficult due to several factors:

- **Non-linear relationships between features and property price.**
  Initial exploration suggested that features like property size or commuting time to CBD exhibit non-linear impacts on property price. For instance, a large number of properties are concentrated in the 0–5000 range, with prices varying significantly—some very high, some very low—indicating that there is no clear direct relationship with property size. Same situation applies in driving time to CBD features.

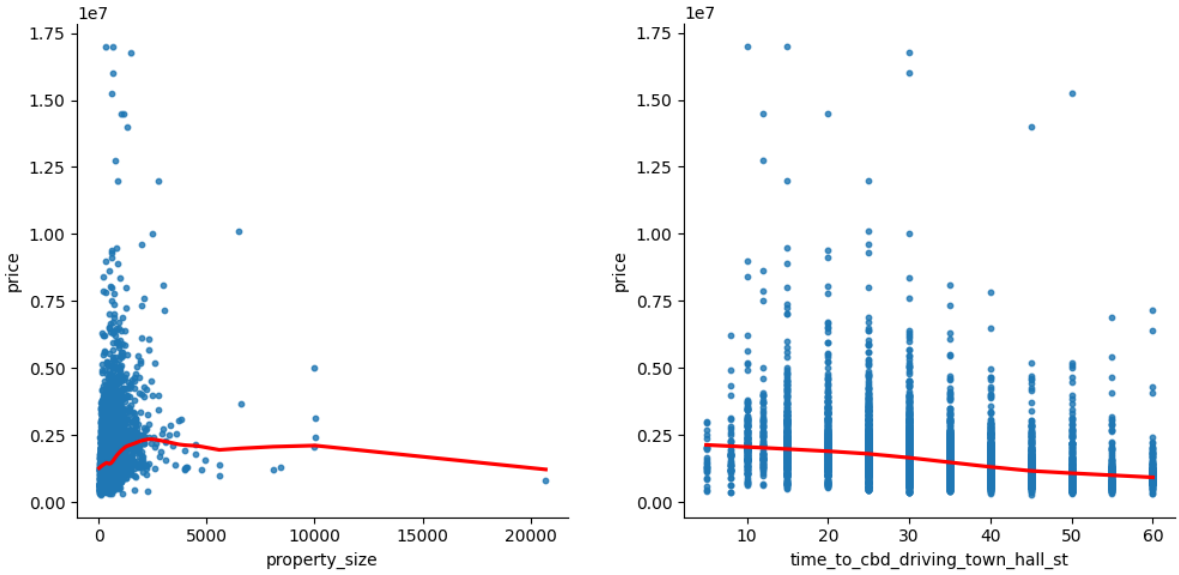

Figure 1: Examples of Non-linear Feature Relationships

- **High variance in the suburbs.**
  Even with small geographic distances, neighboring suburbs can show significant differences in property prices. I took some surrounding suburbs for example.

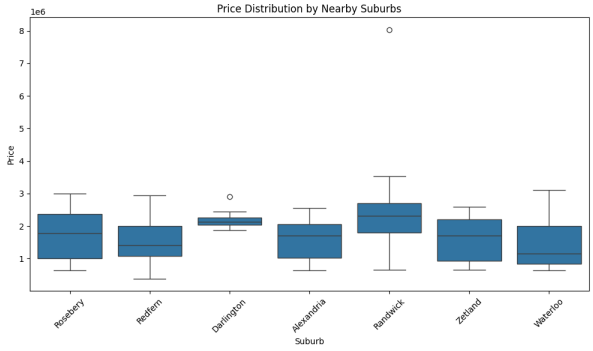

Figure 2: Price Distribution in Neighboring Suburbs

- **Influence of external factors**
  There are a lot of external factors such as cash rate, noises, safety, pet-friendly, inflation index. They can also determine the price.

- **Presence of outliers**
  There are also some special cases that would lead to extreme high price of properties such as luxury properties or newly renovated homes.

1.2 There are several strategies can be employed to address these challenges:

- Do corr-matrix analysis and filter non-related or less-related features
- Apply log transformation to handle skewed data
- Conduct feature engineering
- Cap or remove outliers based on IQR
- Employ advanced models like XGBoost to capture non-linear

1.3 There are some External Data Sources can be used to improve the prediction results:

- **Economic Indicators:**
  Interest Rate, Unemployment Rates, Inflation Rate
- **Conduct feature engineering:**
  Crime Rates, School Zones, Amenities
- **Market Trends:**
  Real Estate Market Data, Rental Yields, Supply and Demand

2 Incorporating Temporal Features

2.1 The temporal nature of house prices is so important because:

- **Seasonality**
  House prices often fluctuate due to seasonal demand.
- **Market Trends**
  Over time, the economy, interest rates, and other factors may cause an overall increase or decrease in house prices.
- **Inflation**
  Prices generally increase over time due to inflation, which may need to be accounted for in price prediction.

2.2 Techniques to Integrate Date-Related Features Effectively:

- **Date Breakdown into Components:**
  Year, Month, Day: Extract components like year, month, and day from the date of sale. These components can help capture seasonality and long-term trends.

2.3 How temporal features effects the price:

- **How price fluctuate with different year, quarter and month**
  The average house price by year shows that there was little fluctuation in prices from 2016 to 2019, but a significant increase occurred from 2019 to 2021, followed by a sharp decrease in 2022. In the third and fourth quarters, house prices were generally higher than in the first and second quarters. This may be because the third and fourth quarters coincide with Australia's spring and summer, a time when people may be more inclined to move.

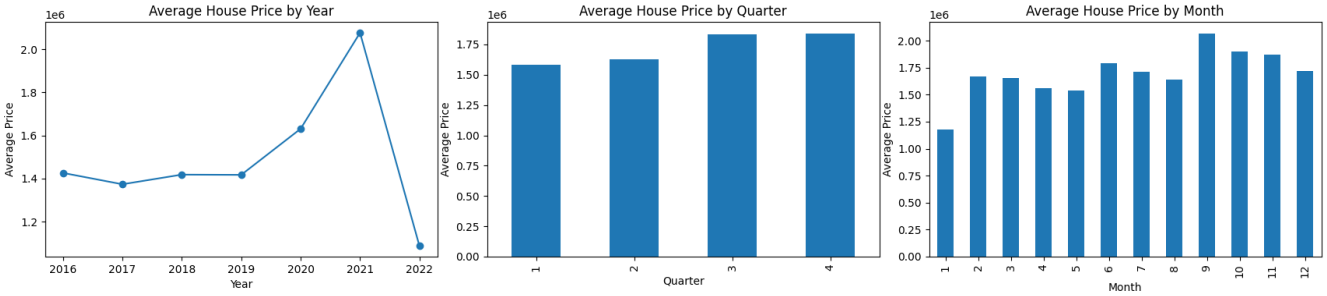

Figure 3: Examples of Non-linear Feature Relationships

- **The correlation matrix illustrates the relationships between different variables**
  A value closer to 1 indicates a stronger positive correlation. From the matrix, we observe a strong positive correlation between the year and index inflation, and a strong negative correlation between inflation and the cash rate. Additionally, there is a slight correlation between house prices and the year, as well as between the cash rate and index inflation.

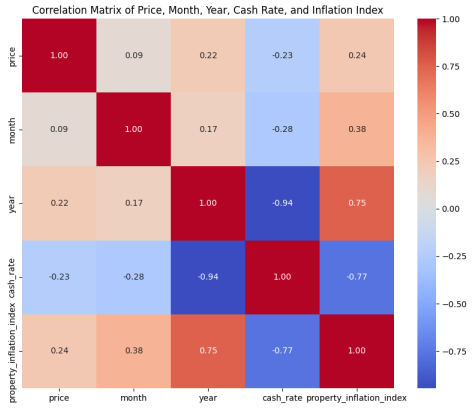

Figure 4: Examples of Non-linear Feature Relationships

3 Evaluation Metrics

3.1 Regression

- **Mean Absolute Error (MAE)**
  MAE calculates the average absolute difference between the predicted and actual values, offering a clear interpretation of the model's accuracy in terms of the average prediction error. Given that house prices can vary significantly, MAE provides a simple and interpretable way to assess how close the predictions are to the actual prices. The MAE in this case is **309k**, meaning there is, on average, a 309k difference between the predicted and actual values. While this difference may appear substantial, it's important to consider that predicting property prices is inherently challenging. Therefore, this gap is acceptable within the context of such a complex task.

3.2 Classification:

- **F1-Score**
  F1-score, provides a balanced metric when there is an imbalance between precision and recall. Precision, recall, and the F1-score are particularly useful when the dataset is imbalanced or when misclassifying certain types of houses has more significant consequences than others.

I achieved an F1-score of **91%**. In the training dataset, the samples of different house types are imbalanced, with houses accounting for over 80% of the data. The F1-score effectively addresses the potential issue of relying solely on accuracy. An F1-score of 91% suggests that the model performs well in identifying all house types while avoiding the bias of favoring the majority class.

3.3 Machine Learning models Comparison:

After trying Linear Regression, Random Forest, and XGBoost, the final result shows that XGBoost performs the best. Here's a justification for why XGBoost might have outperformed the other models:

- **XGBoost**
  XGBoost's ability to capture complex, non-linear relationships in the data allowed it to outperform both Linear Regression and Random Forest. The non-linear nature of housing prices (affected by various interacting factors) requires a model that can adapt and capture these intricate patterns, which XGBoost does very well. XGBoost has powerful hyperparameter tuning capabilities, which, when optimized correctly, can provide better performance compared to Random Forest or Linear Regression. It also provides built-in cross-validation and early stopping mechanisms to prevent overfitting and to improve model generalization.

- **Random Forest**
  It does not perform as well as XGBoost for certain tasks due to a lack of fine-tuning during training. Random Forest can overfit if not properly regularized or tuned, especially in the property price predication cases.

- **Linear Regression**
  Linear Regression may fail to capture interactions between features, which can significantly reduce its predictive power for complex tasks like house price prediction, where factors are often interdependent and nonlinear such as property size, distance to CBD, etc.