Final Project – Data Analytics: Machine Learning

Jack Cleary

Introduction

Ireland, renowned for its vibrant culture, stunning landscapes, and welcoming communities, is undeniably a great place to live. However, it is impossible to overlook the frequently discussed housing crisis that grips the nation. This issue, marked by soaring property prices, a severe shortage of affordable homes, and escalating rents, casts a shadow over the otherwise high quality of life enjoyed by many. Considering this, it was decided that an interesting approach for this project may be to leverage machine learning to predict house prices based on a variety factors.

Predicting house prices is crucial for buyers, sellers, investors, and real estate professionals, aiding informed decision-making, optimizing investments, and ensuring fair pricing. This project utilizes machine learning models to predict house prices using features like date of sale, location, price, year built, size, bedrooms, bathrooms, house type, and house style.

The report structure is as follows: First, we describe the dataset and preprocessing steps. Next, we perform an exploratory data analysis (EDA) to uncover initial insights and trends. We then discuss the selection and implementation of machine learning models, presenting their performance metrics and interpreting the results. Finally, we conclude with a discussion of the findings, limitations, and potential areas for future work. This structured approach ensures a comprehensive understanding of the predictive modelling process.

The Dataset – Importing & Preparing

Unfortunately, a comprehensive dataset of authentic Irish housing sale prices was not readily available, and hence it was necessary to look elsewhere for a data on which to train my model. For simplicity, a mock dataset available to UCD students was used to train and test the model, however, further work can be carried out to apply these techniques to a real dataset.

Data Import & Initial Exploration

The data was retrieved from the following public URL: http://mlg.ucd.ie/modules/python/housing/, by parsing the HTML using a 'Beautiful Soup' parser and appending each returned row of data into a dictionary. This dictionary was subsequently reformatted into a pandas DataFrame for ease of manipulation and analysis. The dataset used in this project consists of various features pertinent to house sales, and my aim was to predict house prices based on these features. Figure 2.1 below an extract of the DataFrame after I had successfully imported the data. Each column provides crucial information about the house and the sale, forming the foundation for the predictive models.

Date of Sale	Location	Price	Year Built	Size	Bedrooms	Bathrooms	House Type	House Style
06 Jan 2020	West End	€732986.00	2010	1696 sq ft	3	2	Detached	2-Storey
06 Jan 2020	West End	€985889.00	2004	2355 sq ft	4	2	Detached	2-Storey
07 Jan 20	Brookville	€1047124.00	2013	1836 sq ft	3	2	Detached	2-Storey
07 Jan 2020	Brookville	€516439.00	2000	1000 sq ft	3	1	Detached	1-Storey
16 Jan 2020	Brookville	€890423.00	2011	1536 sq ft	3	2	Detached	1-Storey

Figure 2.1

The initial exploration of the data highlighted several inconsistencies and issues that required preprocessing and cleaning to ensure the dataset was suitable for machine learning models:

Standardizing Dates

The "Date of Sale" column contained dates in two different formats. To standardize this column, all dates were converted to a uniform format and then transformed into datetime objects using the ‘pd.to_datetime()’ function. This standardization was crucial for any time-series analysis and ensured consistency across the dataset.

Correcting Categorical Variables

The House Style and House Type columns exhibited irregular cardinality, spelling mistakes and inconsistencies in the naming conventions. For example, all instances of 'Bungalow' were misspelled, and both ‘1-Storey’ and ‘One-Storey’ were present in the House Style column. These errors were corrected to maintain uniformity across the categorical variables.

Converting Data Types

The Price, Year Built, Size, Bedrooms, and Bathrooms columns were all converted to integers to facilitate numerical operations and modeling. For Price and Size, this meant reformatting and removing the string elements of each entry (‘€’ and ‘sq ft’). For the Year Built column, many entries contained a string ‘????’ for unknown, so these were converted to NaN.

DropNa

All missing values were dropped from the dataset. Other data imputation solutions were considered to avoid this, such as using average values or average values grouped by location etc., however there were such a small number of missing variables that it was decided instead to remove these to avoid any potentially incorrect data.

One-Hot Encoding for Categorical Variables

As a final step in the data preprocessing phase, the categorical variables ("Location", "House Type", and "House Style") were transformed into numerical data using one-hot encoding. This method creates binary columns for each category, allowing the machine learning models to process the categorical data effectively. This method was chosen as there was no ordinal relationship between the categories and the number of unique categories was not too large.

It should be noted that to make the Exploratory Data Analysis simpler and clearer, the One-Hot Encoding was not applied to the dataset until after this phase. Figure 2.2 below shows and extract of the DataFrame after all preprocessing changes had been applied:

Date of Sale	Location	Price	Year Built	Size	Bedrooms	Bathrooms	House Type	House Style
2020-01-06	West End	732986	2010	1696	3	2	Detached	2-Storey
2020-01-06	West End	985889	2004	2355	4	2	Detached	2-Storey
2020-01-07	Brookville	1047124	2013	1836	3	2	Detached	2-Storey
2020-01-07	Brookville	516439	2000	1000	3	1	Detached	1-Storey
2020-01-16	Brookville	890423	2011	1536	3	2	Detached	1-Storey

Figure 2.2

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns, relationships, and distributions within the dataset. For this project, various visualizations were created using matplotlib and seaborn to gain insights into the house price data. The following sections describe each visualization and the key findings from the EDA process.

Figure 3.1

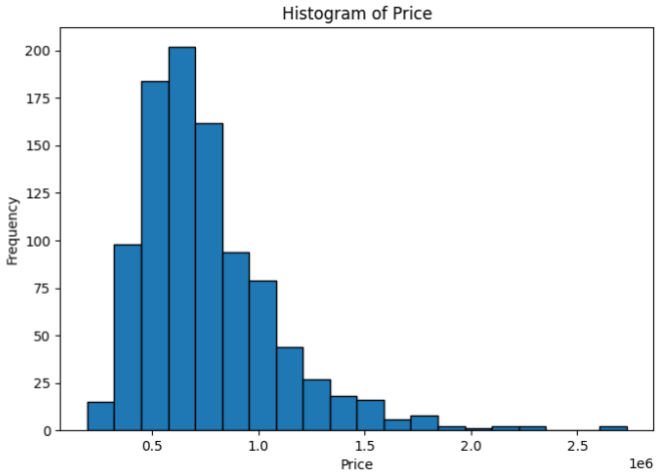

Figure 3.2

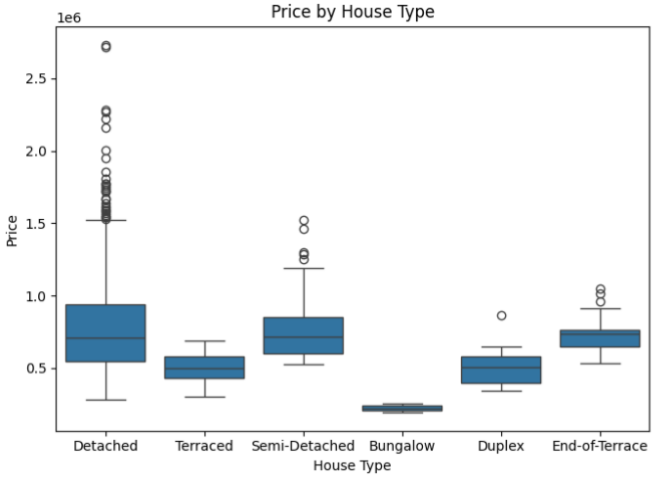

Figure 3.3

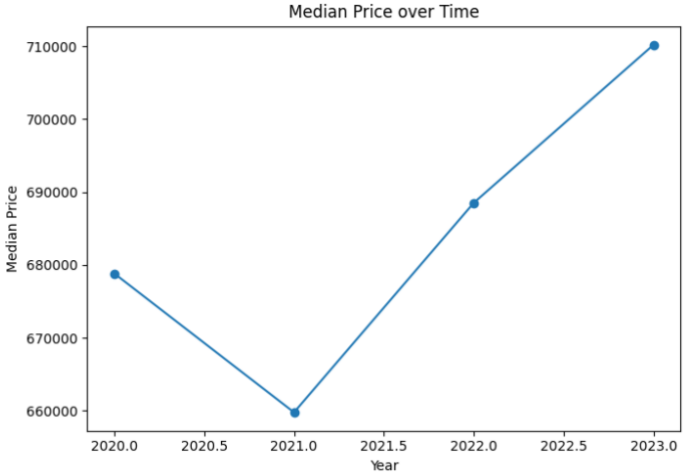

Figure 3.4

A scatter plot (Figure 3.1 above) of house price versus house size using matplotlib showed a positive correlation, indicating that larger houses tend to have higher prices. The histogram of house prices (Figure 3.2) revealed a right-skewed distribution, with most houses priced below the mean and a few significantly higher.

In the boxplot of house prices by House Type, plotted using matplotlib and seaborn, several high-price outliers were observed, particularly among Detached houses. The line plot of median house prices over time (Figure 3.4) indicated a price decrease from 2020 to 2021, followed by an increase from 2021 to 2023, perhaps reflecting the impact of the Covid-19 pandemic and subsequent market recovery.

Figure 3.5

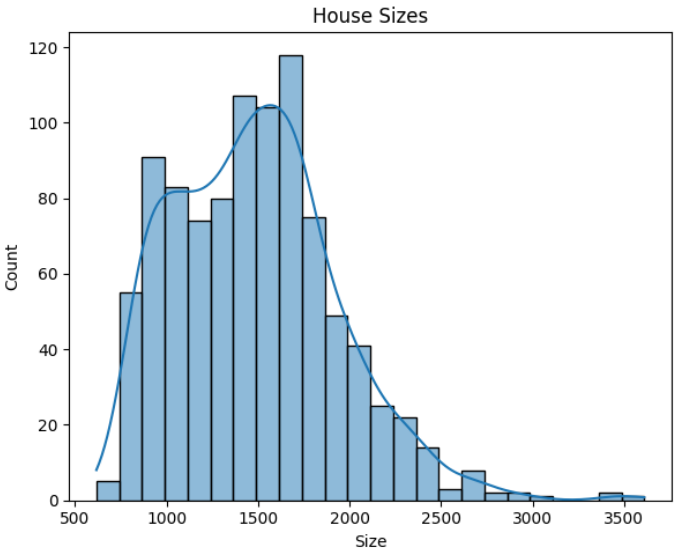

Figure 3.6

A histogram of house prices was plotted again, this time using seaborn and including a Kernel Density Estimation (KDE) line, confirmed the right-skewed nature of the data (Figure 3.15). Another histogram of house sizes (Figure 3.6) showed a bimodal distribution with peaks around 1000 and 1600 square feet, suggesting two distinct groups of houses, possibly indicating different housing types or market segments.

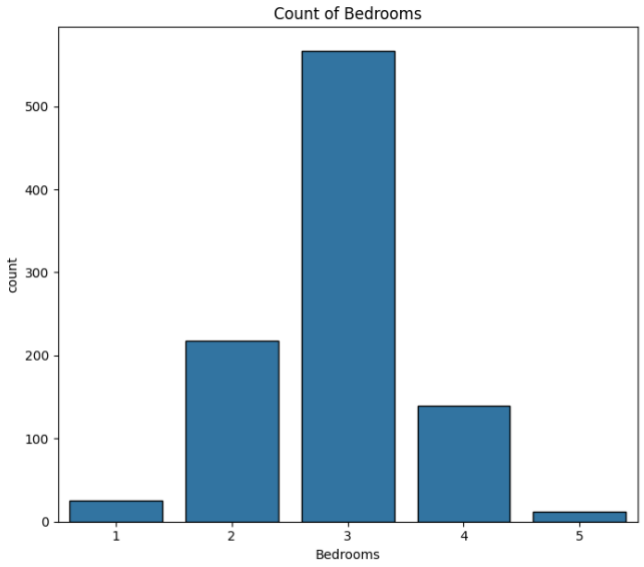

Figure 3.7

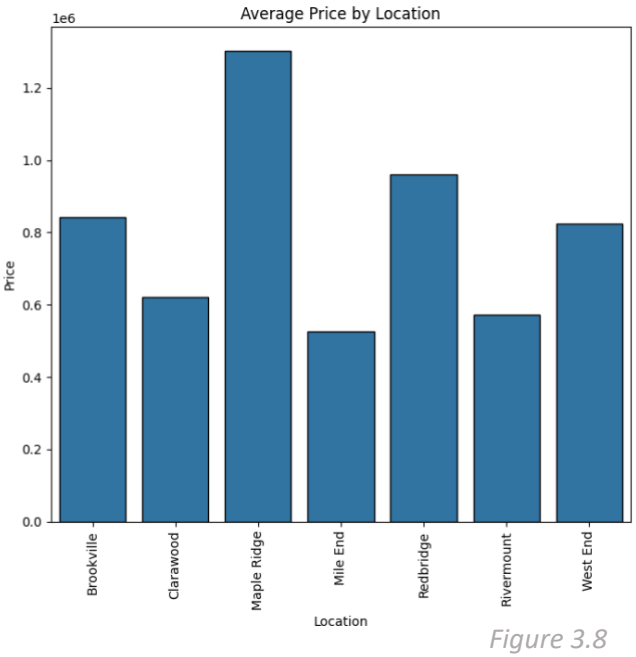

Figure 3.8

Figures 3.7 and 3.8 above are a count plot and bar plot respectively, plotted using seaborn. The count plot visualizes the frequency distribution of bedroom counts and shows that 3-bedroom houses were by far the most common. The bar plot outlines which areas have the highest and lowest mean prices, with Maple Ridge appearing the most high-end location, and Mile End the cheapest location in which to buy a house.

Next, a heatmap of correlations between different variables was plotted using seaborn, as shown in Figure 3.9 below. Specifically, the variables Price, Size, Bedrooms, Bathrooms, and Year Built were chosen as these were the numerical data points of interest. The heatmap showed strong positive correlations between Price and Size, Price and Bathrooms, and Price and Year Built, indicating that larger, newer houses with more bathrooms tend to be more expensive. Interestingly, the correlation between price and bedrooms was not as high as expected, however this is likely due to the non-uniform distribution of the number bedrooms shown in Figure 3.7 above. The high number of 3-bedroom houses in this dataset may be the cause of this slightly lower, yet still positive correlation.

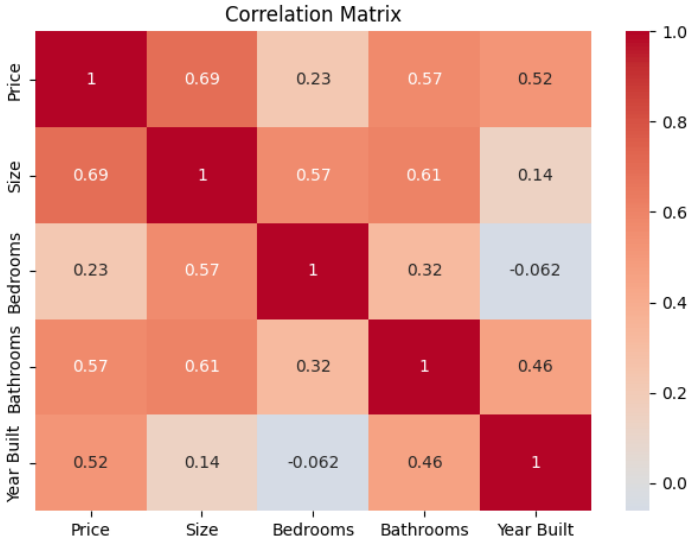

Figure 3.9

Lastly, facet grids (Figure 3.10 and 3.11) plotted price versus size, faceted by the number of bedrooms and bathrooms respectively, revealing that houses with more bedrooms and bathrooms tend to have higher prices for a given size. However, the limited data for 1 and 5-bedroom houses may reduce the reliability of those specific graphs compared to the others.

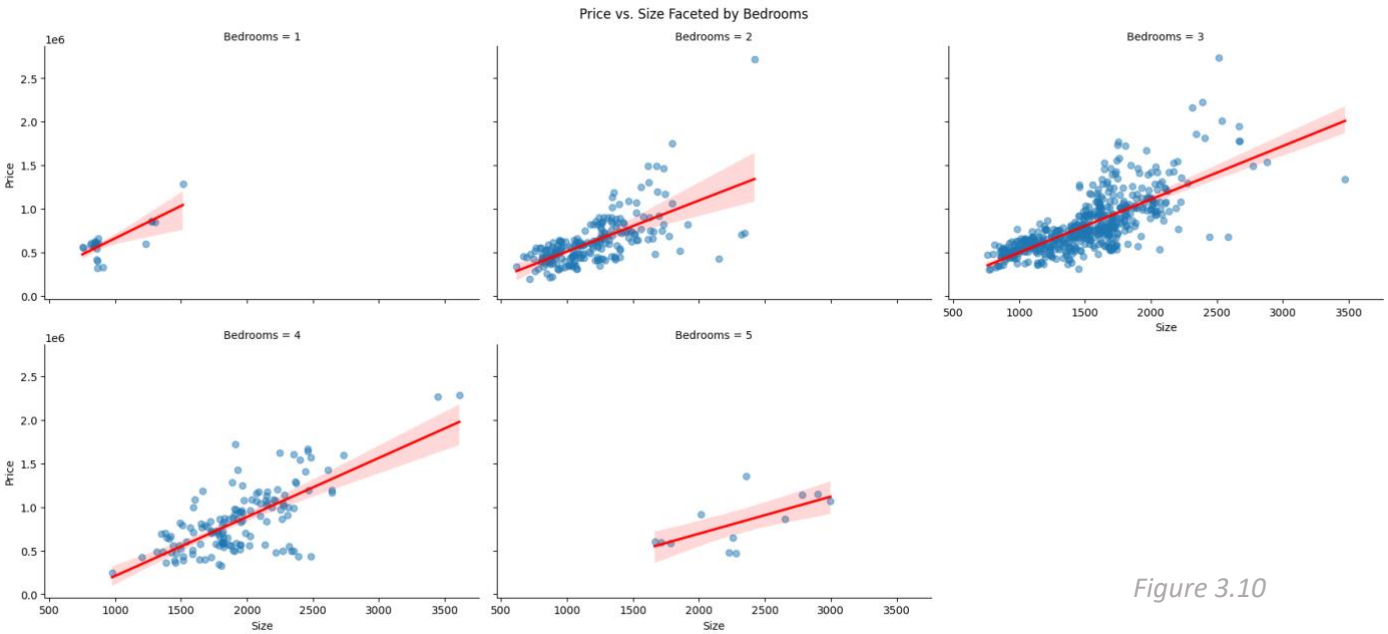

Figure 3.10

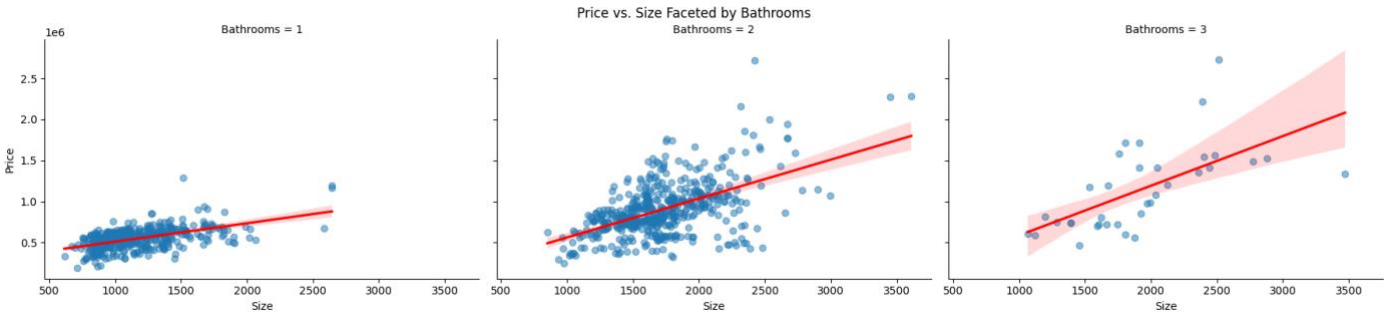

Figure 3.11

Machine Learning Models – Implementation & Findings

In this project, supervised and unsupervised machine learning models were employed to predict house prices based on various features.

Unsupervised Learning Model

The sole unsupervised learning method employed in this project was K-means clustering, a technique for unearthing hidden patterns in data, which sheds light on house price trends here. By grouping houses based on size and price, it reveals three distinct clusters.

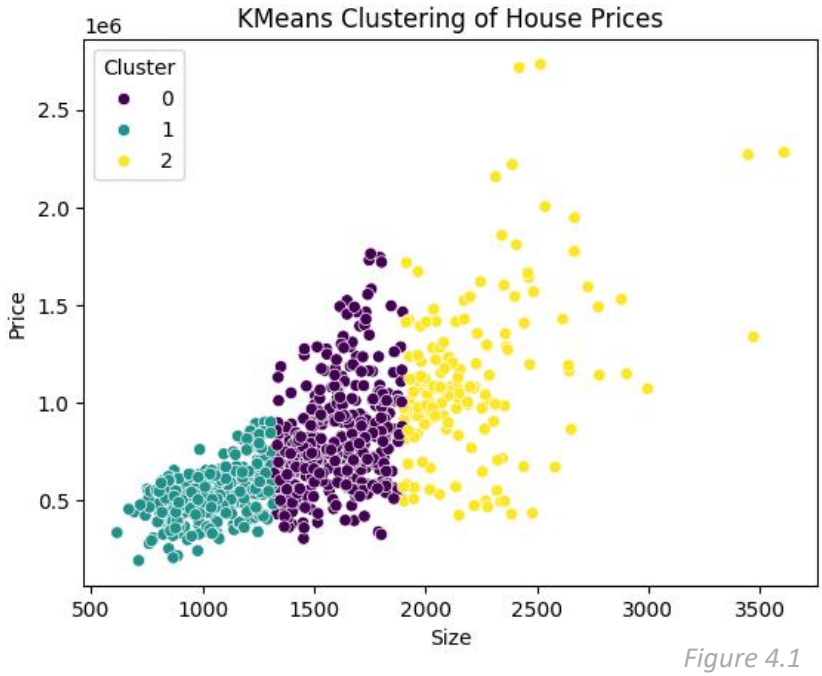

Figure 4.1

The green cluster represents smaller, budget-friendly homes. The blue cluster encompass mid-sized houses with mid-range prices. Finally, the orange cluster signifies larger, luxury homes. This is a preliminary analysis, but it provides valuable insights into potential housing market segments. It's important to remember that k-means clustering has limitations, and the groupings can be influenced by the number of clusters chosen. However, this initial look offers a promising starting point for understanding the housing market landscape.

Supervised Learning Models

The supervised models used were Linear Regression, Polynomial Regression, Random Forest, and a tuned Random Forest model. Due to device constraints, it was unfortunately not possible to implement TensorFlow and deep learning models. For each model, a 70/30 split was used for creating training and test data. Below is a comparison of each model's performance, as well as discussion and justification for each chosen model.

Model Comparison

The table below shows a comparison between the different models used. The Hyper-tuned Random Forest was the most accurate at predicting house prices, as shown by the lower RMSE value on the test set, and the higher R^2 value. The higher RMSE on the training set in comparison with the default Random Forest indicates reduced overfitting.

Model	RMSE (Train)	RMSE (Test)	R^2 (Train)	R^2 (Test)
Linear Regression	144,997.89	135,086.02	0.8094	0.7898
Polynomial Regression	106,265.64	1.235e17	0.9004	-1.757e23
Random Forest	48,851.54	116,535.15	0.9789	0.8436
Hyper-tuned Random Forest	72,693.28	114,628.02	0.9534	0.8486

Linear Regression

Linear Regression was selected as a baseline model due to its simplicity and interpretability. It provides a straightforward approach to understanding the relationship between features and the target variable. After standardizing the features, the model was trained on the training set and evaluated on the testing set.

These results indicate that Linear Regression provides a reasonable fit to the training data but shows a slight decrease in performance on the test data, suggesting some overfitting. The relatively large RMSE values also indicate inaccuracy, both for training and test data, however it is likely that the outliers identified in the Exploratory Data Analysis may account for an increased average error.

Polynomial Regression

Polynomial Regression was then used to attempt to capture potential non-linear relationships between features and the target variable, which linear models might miss. After standardizing the features, the model produced the above results.

The Polynomial Regression model drastically overfits the training data and performs poorly on the test set, indicated by the extreme RMSE and negative R^2 values. This suggests that the model captured noise as opposed to the underlying trend. Hence, this model was a poor predictor of house prices.

Random Forest

Random Forest was selected for its ability to handle non-linearities and interactions between features, providing a more robust prediction than linear models.

The Random Forest model performed better than Linear Regression, showing high accuracy on both training and test sets. However, the slight difference between train and test performance suggests mild overfitting.

Hyperparameter Tuning for Random Forest

Hyperparameter tuning was performed on the Random Forest model to optimize its performance. Grid search with cross-validation was used to find the best parameters.

The tuned Random Forest model shows improved test set performance and reduced overfitting, indicated by the slightly closer train and test metrics, however there is still a degree of inaccuracy here which could be improved upon.

Feature Importance

After electing the Hyper-tuned model as the best-performing, an investigation into feature importance was carried out on the model to provide an inherent measure of which features had the most impact on house price predictions. The top features are shown in Figure 4.1 below:

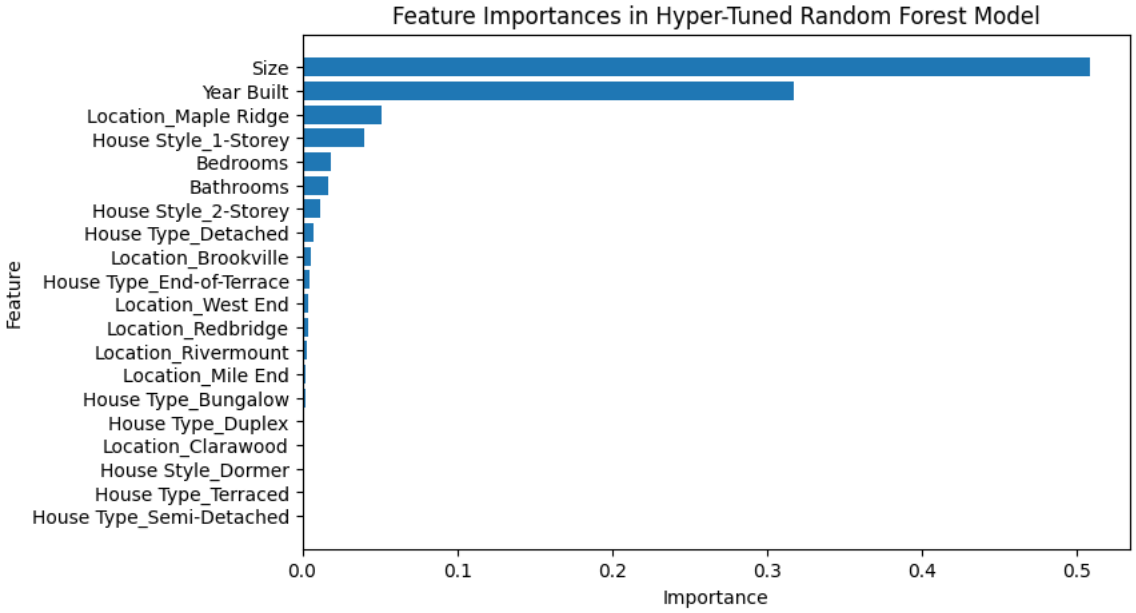

Figure 4.2

Size was significantly the most important feature in price prediction, followed by age and whether the house was located in Maple Ridge (identified earlier as the most 'high-end' location). In comparison, whether a house was terraced or semi-detached appears to have had little to no impact on price.

Summary

The results from this project highlight several important findings and limitations:

Implications

The tuned Random Forest model outperformed other models, demonstrating the importance of hyperparameter tuning and the effectiveness of ensemble methods in capturing complex relationships in the data.

Key features like size, year built and location were crucial in predicting house prices, aligning with domain knowledge.

Limitations

Both the Random Forest and Polynomial Regression models exhibited signs of overfitting, particularly the latter. It is also worth noting that the quality and quantity of data can significantly impact model performance, and the fact that this is not authentic house price data will likely have had a major impact on model performance in this case.

Future Work

Acquiring real-world data would be extremely beneficial to this project, as well as expanding the dataset with more features and examples to improve model robustness. Exploring other advanced models like Gradient Boosting, XGBoost, or neural networks would likely further enhance model performance.

Conclusion

This purpose of this project was to predict house prices using various machine learning models. Accurate house price prediction is vital for stakeholders in the real estate market, and the insights gained from this project provide a solid foundation for developing robust predictive models in this domain. The key takeaways were that the tuned Random Forest model outperforming others, emphasizing ensemble methods and hyperparameter tuning, and identifying features like size, location, and year built as critical for accurate predictions. Future improvements involve incorporating more data, exploring different models, and refining feature engineering.

I remain hopeful that with the application of intelligent solutions such as these, and of course given time, the seemingly unfixable housing crisis in Ireland will end, and there will be affordable housing for all once again!