IS6052: Predictive Analytics

Decoding Dublin's Housing Market: Predictive Insights into Irish Homebuyers

Omkar Nandkumar Phadtare – 124109697

Submission Date: 2nd December, 2024

Introduction

There are several economic, cultural and other factors that determine the housing market situation in Dublin. It is therefore important to identify characteristics which can aid stakeholders ranging from real estate developers, policy makers, and prospective residental home owners gain deeper insight on the Irish home buyer. This report examines key factors influencing property purchases in Dublin, including price, location, size, and insulation quality, with varying patterns across the city's four local authorities: Dublin City, Fingal, Dun Laoghaire Rathdown and South Dublin.

Current trends in the Irish property market have been illustrated, in some regard by situations such as those arising from Britain exit and other Central Bank lending policies, in relation to price and demand changes in various regions. It is even more complicated due to Dublin's presence of a strong tech industry, employment, multicultural diverseness, and stability of weather, for which MA predictive analytics is quite helpful to determine market tendencies.

In this report, a dataset of property transactions in Dublin city is used to estimate the probability of a property in Dublin being bought, based on a range of buyer characteristics and market trends.

Key steps include:

- Getting insights in the data for decision making.
- Data cleaning and feature extraction to enhance the values capable to give out by the model.
- Predictive models and how to apply and compare them to measuring performance.
- Presenting the results in a form of visualizations and getting support from the material found in literature.

Finally, this report highlights the factors affecting buying decisions and provide suggestions to the developers to enhance their products. It also presents key findings of the role of the predictive analytics in managing with uncertainties, and offers the guideline for other comparative studies of market uncertainty in different areas or fields.

Dataset Description

The dataset used in this study offers a comprehensive view of Dublin's housing market, with 13,320 entries and 12 features. It includes key factors that influence homebuyers' decisions, such as size, location, price, and buyer intent, providing a solid foundation for predictive analytics. Here's an overview of the dataset:

1. **Key Features**:
    - **ID**: Unique identifier for each property.
    - **Property Scope**: Describes property type, such as "Extended Coverage" or "Land Parcel."
    - **Availability**: Indicates when the property will be available, such as "Ready to Move."
    - **Location**: Specifies the local authority (e.g., Fingal, DCC, Dun Laoghaire, South Dublin).
    - **Size**: Number of bedrooms (e.g., "2 BED," "4 Bedroom").
    - **Total Sqft**: Total square footage of the property (600–4689 sqft).
    - **Bath**: Number of bathrooms, often correlating with property size.
    - **Balcony**: Number of balconies (0–3).
    - **Buying Intent**: The target variable, indicating whether the property was purchased ("Yes" or "No").
    - **BER**: Building Energy Rating (A–G), reflecting energy efficiency.
    - **Renovation Needed**: Whether the property requires renovation ("Yes," "No," "Maybe").
    - **Price-per-sqft-$**: Price per square foot, a key metric for assessing affordability.

	ID	property_scope	availability	location	size	total_sqft	bath	balcony	buying or not buying	BER	Renovation needed	price-per-sqft-$
0	0	Extended Coverage	17-Oct	Fingal	2 BED	1056	2.0	1.0	No	A	No	419.928030
1	1	Land Parcel	Ready To Move	South Dublin	4 Bedroom	2600	5.0	3.0	No	D	Yes	523.846154
2	2	Constructed Space	Ready To Move	Dun Laoghaire	3 BED	1440	2.0	3.0	No	G	Yes	488.680556
3	3	Extended Coverage	Ready To Move	South Dublin	3 BED	1521	3.0	1.0	No	G	Yes	708.908613
4	4	Extended Coverage	Ready To Move	DCC	2 BED	1200	2.0	1.0	No	F	Yes	482.375000

Fig 1: Raw Data

2. Data Quality and Missing Values:

While the dataset is well-structured, some features exhibit missing or inconsistent values:

- Location: Contains one missing value.
- Size: Features 16 missing values and various inconsistent formats (e.g., "2 BED" vs. "2 Bedroom").
- Bath: 73 missing values, possibly due to misreporting.
- Balcony: 609 missing values, with entries ranging from 0 to 3 balconies.
- Price-per-sqft-$: A significant number of missing values (246), requiring imputation or removal.

Handling these missing and inconsistent values is critical for ensuring the reliability of predictive models.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13320 entries, 0 to 13319
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ID                    13320 non-null  int64
 1   property_scope        13320 non-null  object
 2   availability          13320 non-null  object
 3   location              13319 non-null  object
 4   size                  13304 non-null  object
 5   total_sqft            13320 non-null  object
 6   bath                  13247 non-null  float64
 7   balcony               12711 non-null  float64
 8   buying or not buying  13320 non-null  object
 9   BER                   13320 non-null  object
 10  Renovation needed     13320 non-null  object
 11  price-per-sqft-$      13074 non-null  float64
dtypes: float64(3), int64(1), object(8)
memory usage: 1.2+ MB
```

Fig 2: Overview of raw data

```
location
Fingal           4875
DCC              3030
South Dublin     2610
Dun Laoghaire    2324
Other             480
NaN                 1
Name: count, dtype: int64
```

```
balcony
2.0    5113
1.0    4897
3.0    1672
0.0    1029
NaN     609
Name: count, dtype: int64
```

Fig 3: Frequency of distinct rows

3. Data Distribution and Insights:

A closer look at the distribution of key features reveals important trends:

a. Location:
   - Most properties are in Fingal (4875) and Dublin City Council (3030).
   - Few are listed as "Other" (480), indicating niche or misclassified entries.
   - One missing value highlights the need for cleaning.

b. Size:
   - Most properties have 1–4 bedrooms, with outliers like "43 Bedroom" suggesting errors or special cases.

c. Buying Intent:
   - 68% were not purchased, while 32% were.
   - Cross-analysis with price and BER can reveal buyer trends.

d. BER:
   - Distribution is fairly balanced, with more properties rated C and B, possibly reflecting newer or more energy-efficient properties.

e. Renovation Needed:
   - 56% require renovation, which may affect buying intent or price.

f. Availability:
   - "Ready to Move" is most common (10,581 entries), likely influencing purchase decisions.

4. Visualization of Property Data:

Property Distribution Across Locations:

a. Visualization Summary:
   - A bar plot was created to visualize the distribution of properties across different locations in Dublin.
   - X-Axis: Locations (e.g., Fingal, DCC, Dun Laoghaire, South Dublin, Other).

- Y-Axis: Number of properties.

b. Insights:

- Fingal has the most properties, followed by Dublin City Council and South Dublin.
- The “Other” category is small, suggesting niche or misclassified entries.
- High counts in Fingal and DCC may indicate active residential areas.

c. Actionable Steps:

- Investigate why Dun Laoghaire has fewer properties.
- Correlate location with price, BER ratings, and renovation needs for trends.

Fig 4: Property Distribution Across Locations

Property Distribution by BER Rating:

- Visualization Summary: A bar plot shows the distribution of properties by Building Energy Rating (BER).
  - X-Axis: BER ratings (A–G).
  - Y-Axis: Number of properties per rating.
- **Insights**:
  - The distribution of BER ratings is relatively balanced, with **C** and **B** ratings slightly more prevalent.
  - Lower-rated properties (F, G) are as common as higher-rated ones (A, B), which may reflect a mix of new and older housing stock in the market.
  - BER ratings could be a critical factor for buyer preferences, as higher-rated properties are likely more energy-efficient and desirable.
  - **Actionable Steps**:
    - Cross-analyze BER ratings with price-per-sqft, location, and buying intent to understand their influence on property desirability.
    - Explore if specific locations or property types are overrepresented in certain BER categories.

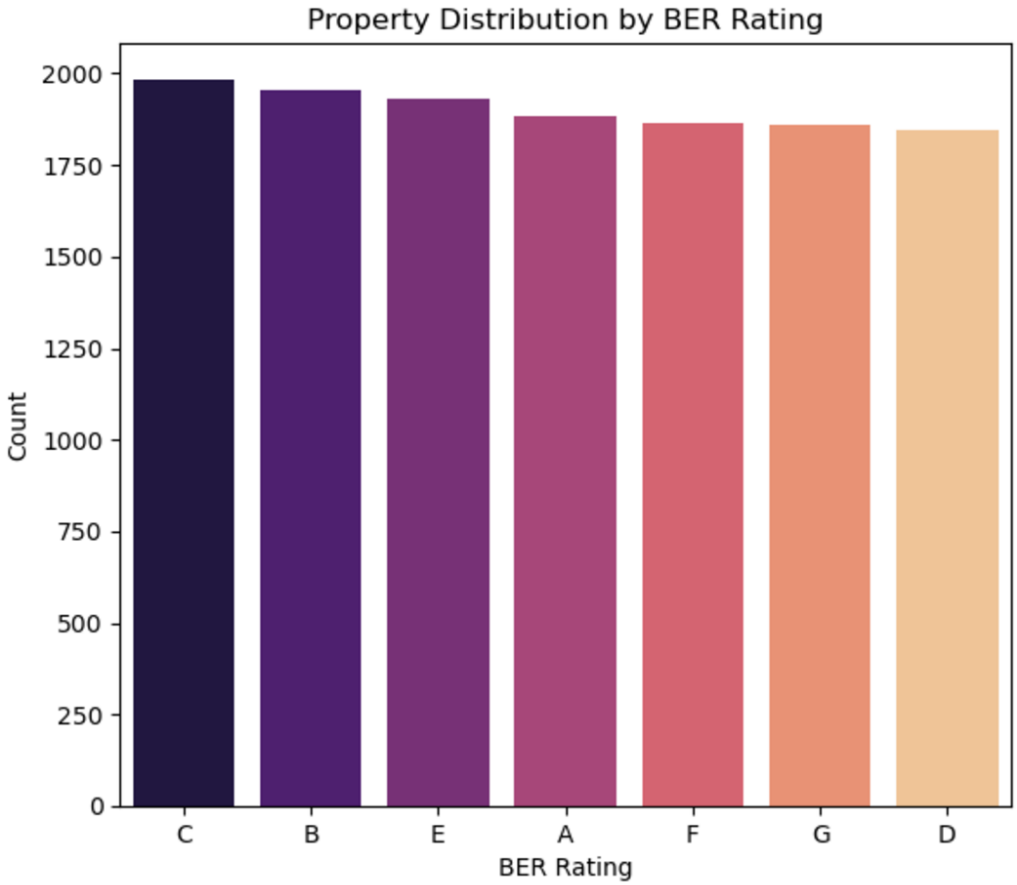

Fig 5: Property Distribution by BER Rating

Data Preparation

This phase focused on cleaning, transforming, and standardizing the dataset for analysis, making features consistent, interpretable, and ready for modeling.

1. Renaming Columns for Clarity:
   Column names were standardized for readability, e.g., “property_scope” became “Property_Scope,” and “price-per-sqft-$” changed to “Price_per_sqft($),” reducing errors in later steps.

	ID	Property_Scope	Availability	Location	Size	Total_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)
0	0	Extended Coverage	17-Oct	Fingal	2 BED	1056	2.0	1.0	No	A	No	419.928030
1	1	Land Parcel	Ready To Move	South Dublin	4 Bedroom	2600	5.0	3.0	No	D	Yes	523.846154

Fig 6: Renamed Columns

2. Handling the Size Column:
   Entries like “3 BED” and “4 Bedroom” were simplified by extracting the numeric values, converting them to integers. This enabled easy categorization and analysis of how size impacts price or buying intent.

	ID	Property_Scope	Availability	Location	Size	Total_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)
0	0	Extended Coverage	17-Oct	Fingal	2	1056	2.0	1.0	No	A	No	419.928030
1	1	Land Parcel	Ready To Move	South Dublin	4	2600	5.0	3.0	No	D	Yes	523.846154

Fig 7: Cleaned Size Column

3. Standardizing Total Area (Total_sqft):
   The Total_sqft column, with inconsistencies like ranges (e.g., “1000-2000”) and various units, was standardized:
   - Ranges were averaged to single values.
   - Units (e.g., acres, square meters) were converted to square feet.

- Invalid entries were set to NaN to maintain data quality.

This ensured all values were numeric for meaningful analysis.

	ID	Property_Scope	Availability	Location	Size	Total_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)
0	0	Extended Coverage	17-Oct	Fingal	2	1056.0	2.0	1.0	No	A	No	419.928030
1	1	Land Parcel	Ready To Move	South Dublin	4	2600.0	5.0	3.0	No	D	Yes	523.846154
2	2	Constructed Space	Ready To Move	Dun Laoghaire	3	1440.0	2.0	3.0	No	G	Yes	488.680556
3	3	Extended Coverage	Ready To Move	South Dublin	3	1521.0	3.0	1.0	No	G	Yes	708.908613
4	4	Extended Coverage	Ready To Move	DCC	2	1200.0	2.0	1.0	No	F	Yes	482.375000

Fig 8: Cleaned Total Sq. Ft. Column

4. Date Cleaning for Availability:

   The Availability column, containing mixed formats (e.g., "18-Dec" and "Ready To Move"), was cleaned:

   - Dates were standardized to DD-MM-YYYY, assuming 2024 for missing years.
   - "Ready To Move" was replaced with today's date to indicate immediate availability.

   This allowed for time-based analysis.

	ID	Property_Scope	Availability	Location	Size	Total_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)
0	0	Extended Coverage	17-10-2024	Fingal	2	1056.0	2.0	1.0	No	A	No	419.928030
1	1	Land Parcel	02-12-2024	South Dublin	4	2600.0	5.0	3.0	No	D	Yes	523.846154

Fig 9: Cleaned Availability Column

Outcome of Data Preparation:

The dataset was reshaped for better structure and to remove any unnecessary noise. Size, Total_sqft and Availability were normalized and all the improper entries were corrected and missing data dealt in a proper manner. This preparation keeps it credible, usable and prepared for analysis and modeling.

Outlier Detection and Handling

In this step, the focus was on addressing missing values and outliers in key features like Balcony, Bath, Size, Total_sqft, and Price_per_sqft($) that could affect the predictive models.

Step 1: Imputing Missing Values

Missing values in categorical features (e.g., Balcony, Bath) were filled with the most frequent value (mode). For example, Balcony's missing values were replaced with 2, the most common value, ensuring completeness without bias.

Step 2: Outlier Detection

Outliers were identified using the Interquartile Range (IQR) method. Any values outside the range defined by:

- Lower bound = Q1 - 1.5 * IQR
- Upper bound = Q3 + 1.5 * IQR

Outliers included:

- 40 baths in the Bath column.
- Over $4.9 million per square foot in Price_per_sqft($).

Step 3: Handling Outliers

Outliers were replaced with NaN, then imputed to maintain consistency:

- Continuous features (Size, Total_sqft, Price_per_sqft) were filled with the median.
- Categorical features (Balcony, Bath) were filled with the mode. For example, Bath outliers were replaced with 2, and Total_sqft values were filled with the median.

Step 4: Data Visualization

Histograms and boxplots showed the distribution of features before and after outlier treatment. This highlighted improved normalization, especially in Price_per_sqft($) and Total_sqft.

Results After Handling Outliers:

- Baths: Maximum reduced from 40 to 4; mean dropped from 2.69 to 2.37.
- Total_sqft: Maximum dropped from 1.3M to 2,550 sqft; mean stabilized at ~1,315 sqft.
- Price_per_sqft($): Maximum reduced from $4.9M to $1,350 per sqft.

These steps refined the dataset, ensuring accuracy and better predictive insights.

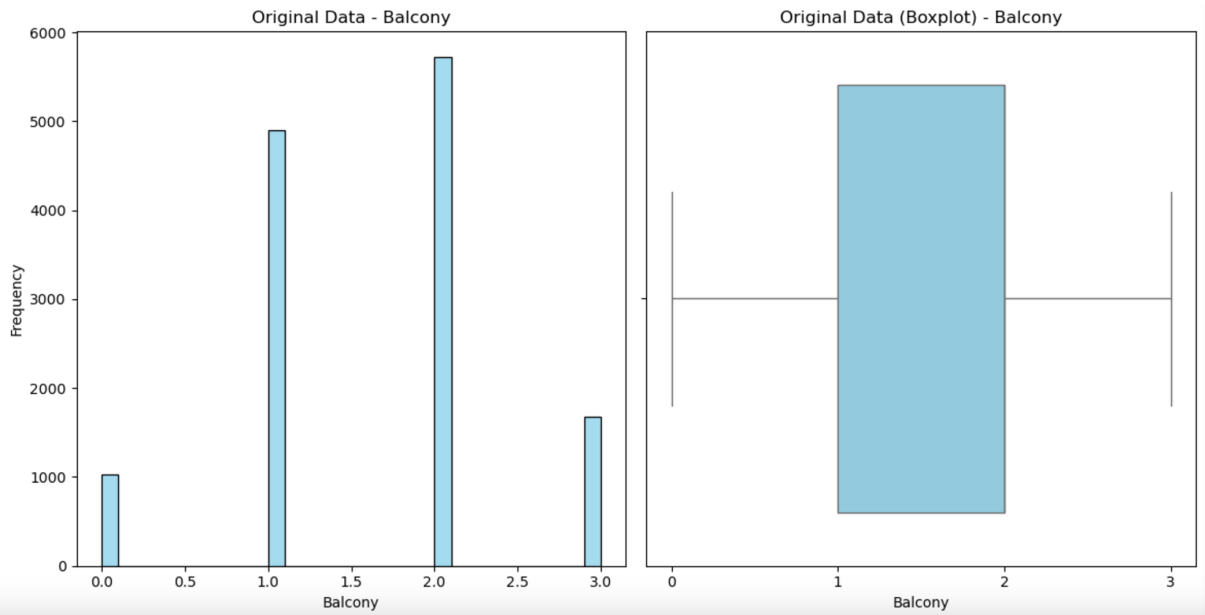

Fig 10: Balcony data plot before handling

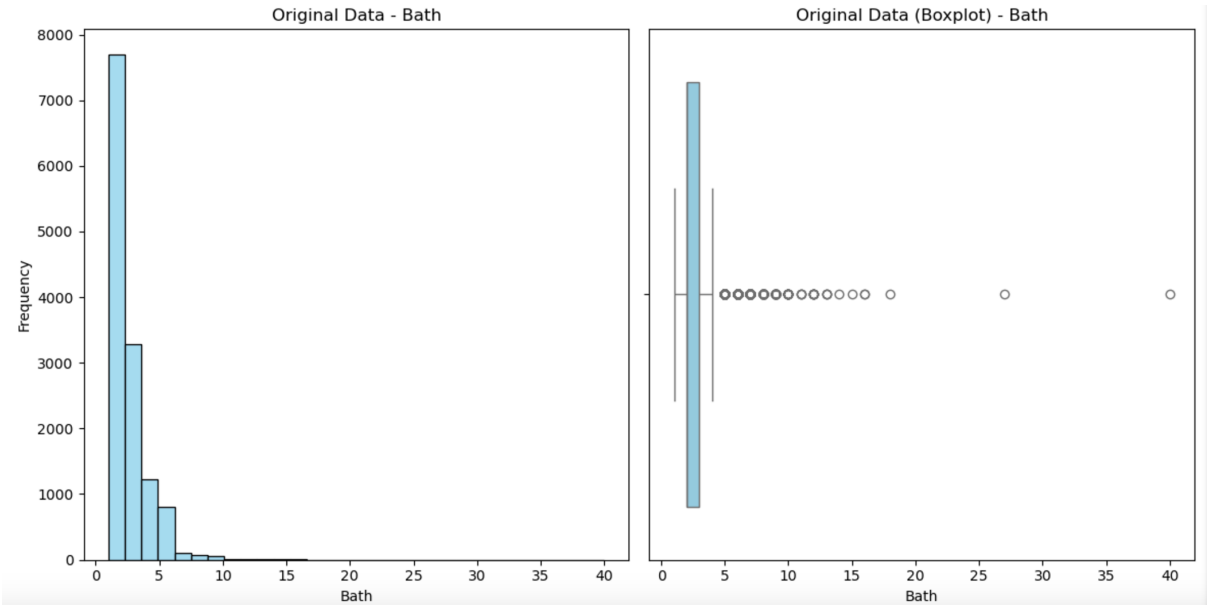

Fig 11: Bath data plot before handling

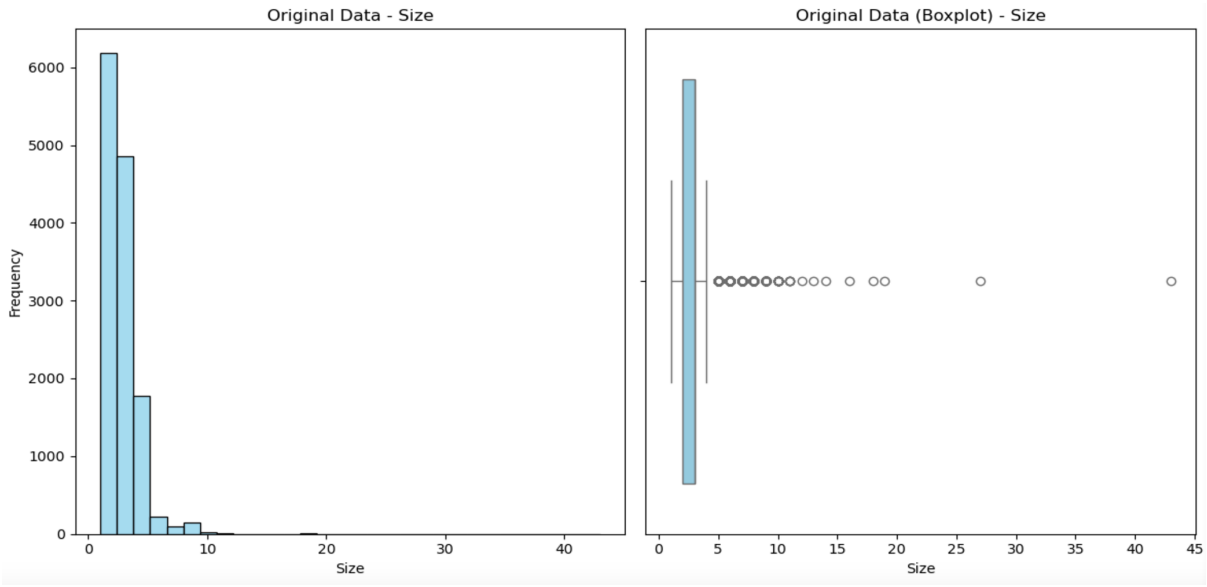

Fig 12: Size data plot before handling

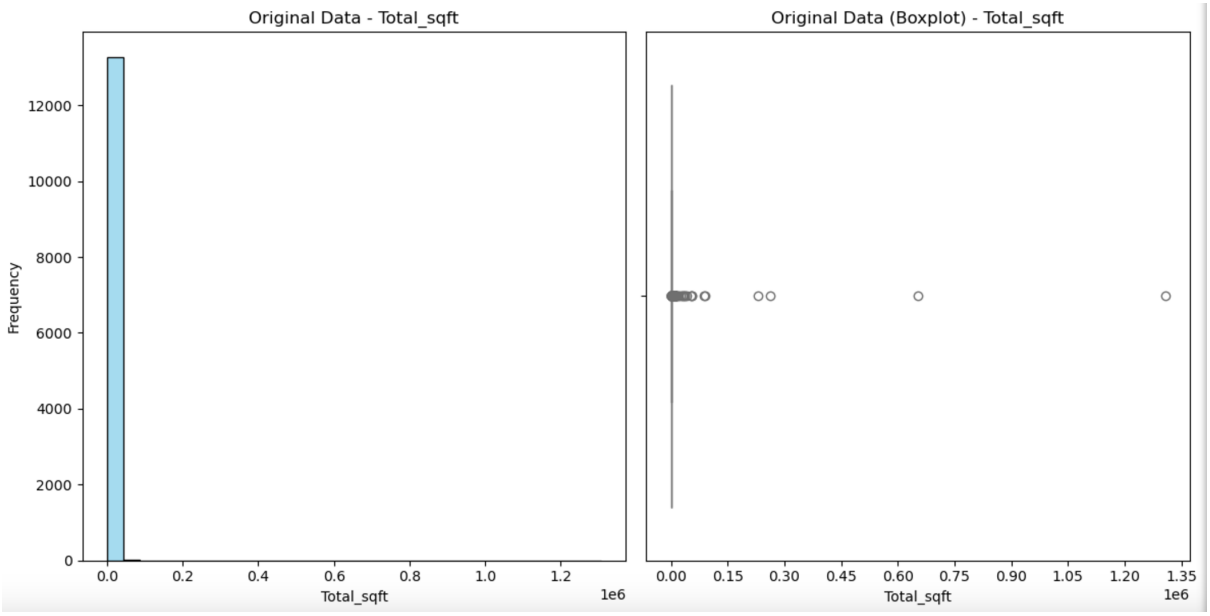

Fig 13: Total Sq. Ft. data plot before handling

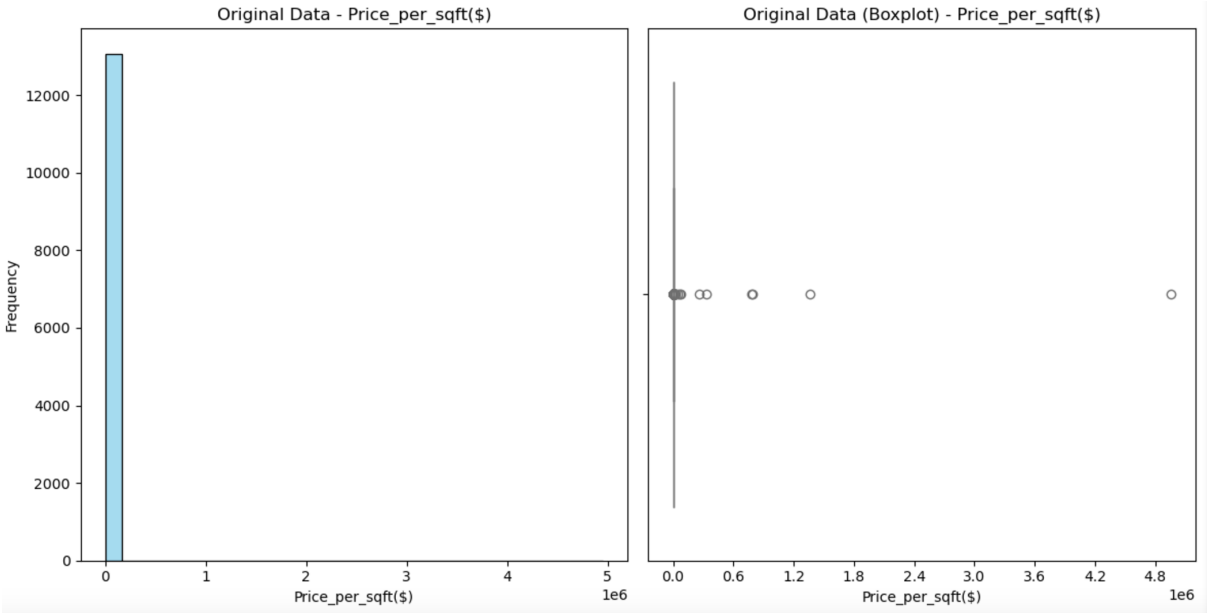

Fig 14: Price per sqft data plot before handling

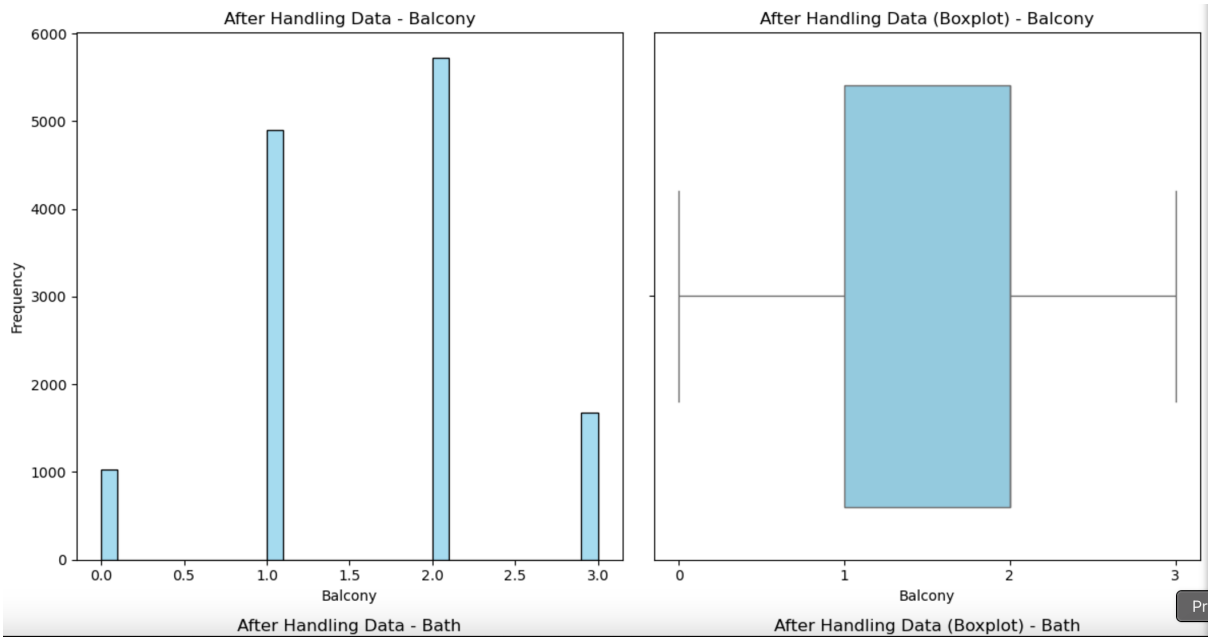

Fig 15: Balcony data plot post handling

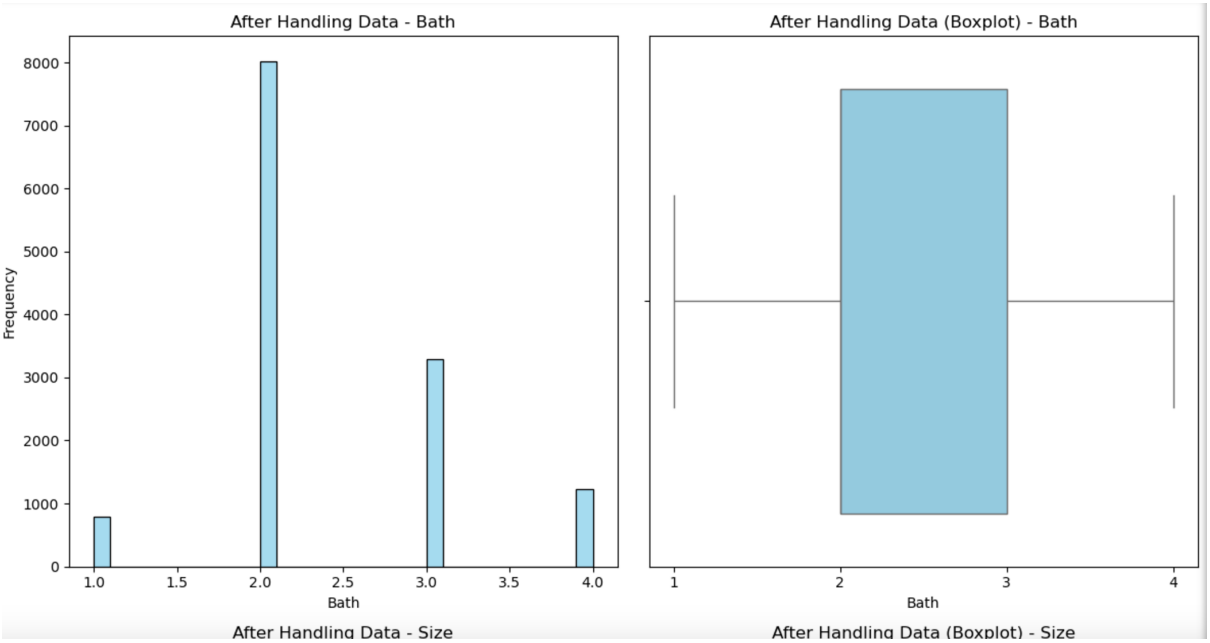

Fig 16: Bath data plot post handling

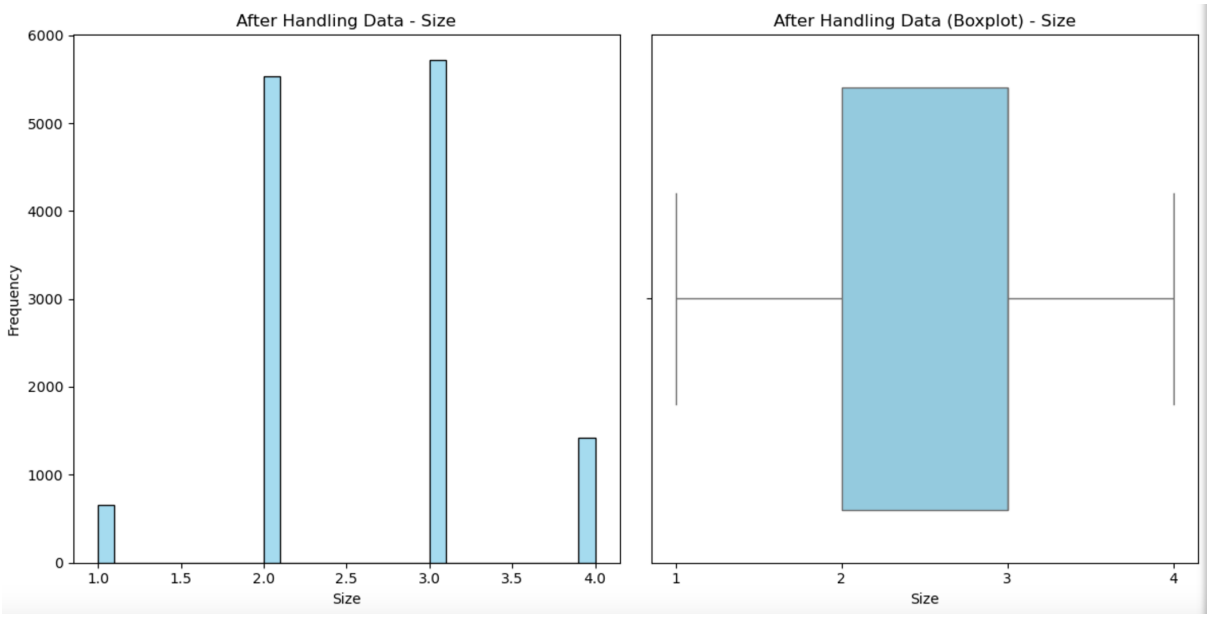

Fig 17: Size data plot post handling

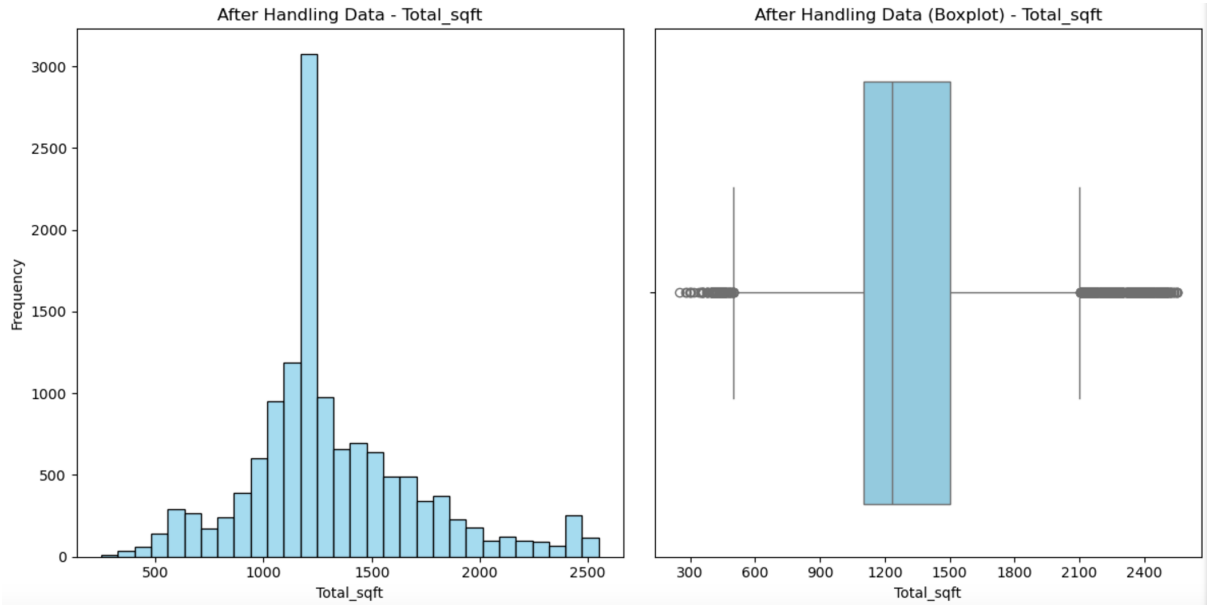

Fig 18: Total sq. ft. data plot post handling

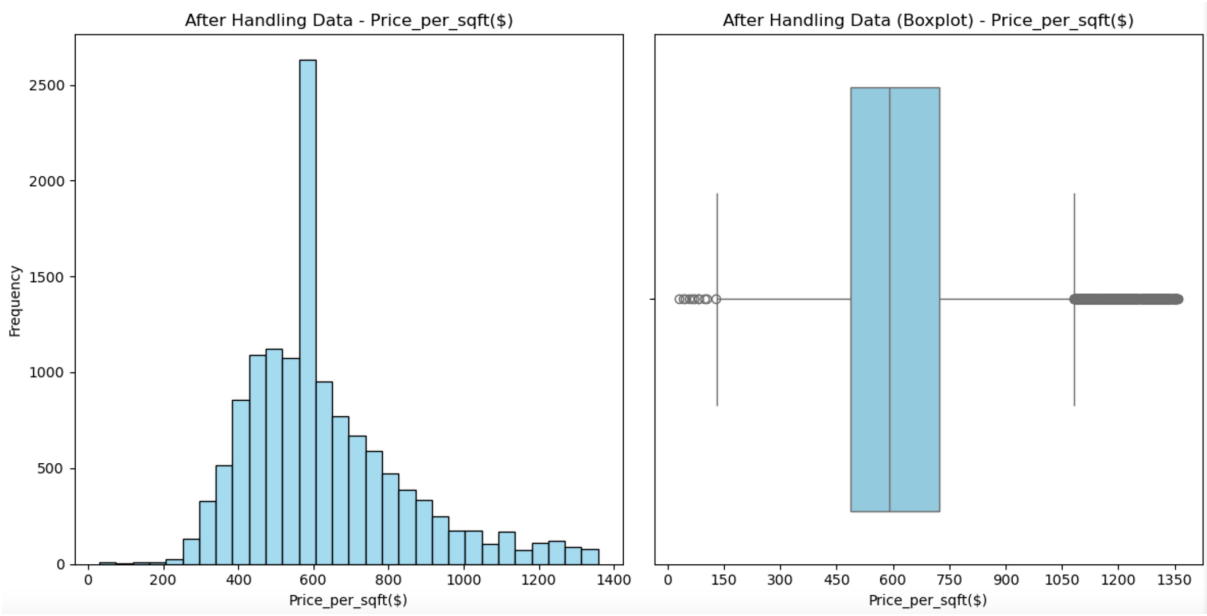

Fig 19: Price per sqft. data plot post handling

Feature Engineering

To improve insights and predictive accuracy, new features were added to the dataset, focusing on temporal, calculated, and scaled variables:

1. Extracting Month and Season:
   - Availability Month: The month was extracted from the Availability column to analyze its effect on pricing and demand.
   - Availability Season: Each property was categorized into seasons (Winter, Spring, Summer, Fall) based on the month, helping to study seasonal trends in availability.

_Scope	Availability	Location	Size	Total_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)	Availability_month	Availability_season
tended verage	17-10-2024	Fingal	2	1056.0	2	1	No	A	No	419.928030	10	Fall
Parcel	02-12-2024	South Dublin	4	1236.0	2	3	No	D	Yes	523.846154	12	Winter
tructed Space	02-12-2024	Dun Laoghaire	3	1440.0	2	3	No	G	Yes	488.680556	12	Winter
tended verage	02-12-2024	South Dublin	3	1521.0	3	1	No	G	Yes	708.908613	12	Winter
tended verage	02-12-2024	DCC	2	1200.0	2	1	No	F	Yes	482.375000	12	Winter

Fig 20: Extracted Availability Month and Season Columns

2. Total Price Calculation:
   - Total_price($): Created by multiplying Total_sqft with Price_per_sqft($) to represent the overall property value.
   - Total_price_in_million_$: Scaled the total price for easier interpretation.

al_sqft	Bath	Balcony	Buying_Intent	BER	Renovation_needed	Price_per_sqft($)	Availability_month	Availability_season	Total_price($)	Total_price_in_million_$
1056.0	2	1	No	A	No	419.928030	10	Fall	4.434440e+05	0.443444
1236.0	2	3	No	D	Yes	523.846154	12	Winter	6.474738e+05	0.647474
1440.0	2	3	No	G	Yes	488.680556	12	Winter	7.037000e+05	0.703700
1521.0	3	1	No	G	Yes	708.908613	12	Winter	1.078250e+06	1.078250
1200.0	2	1	No	F	Yes	482.375000	12	Winter	5.788500e+05	0.578850

Fig 21: Calculated Total Price and Total in Million USD

3. Calculating Affordability Ratio:
    - Affordability_ratio: A measure of affordability, calculated as Total_sqft / Total_price($).
    - Affordability_ratio_scaled: Scaled the ratio for better comparison across properties.

enovation_needed	Price_per_sqft($)	Availability_month	Availability_season	Total_price($)	Total_price_in_million_$	Affordability_ratio	Affordability_ratio_scaled
No	419.928030	10	Fall	4.434440e+05	0.443444	0.002381	2.381360
Yes	523.846154	12	Winter	6.474738e+05	0.647474	0.001909	1.908957
Yes	488.680556	12	Winter	7.037000e+05	0.703700	0.002046	2.046327
Yes	708.908613	12	Winter	1.078250e+06	1.078250	0.001411	1.410619
Yes	482.375000	12	Winter	5.788500e+05	0.578850	0.002073	2.073076

Fig 22: Calculated Affordability ratio and Scaled Affordability ratio

4. Categorical Encoding
    - Applied **One-Hot Encoding** to categorical variables like **Property_Scope**, **Location**, **BER**, **Renovation_needed**, and **Availability_season**, turning them into binary columns for machine learning compatibility.
    - For **Buying_Intent**, only the "Yes" column was retained after encoding.

Property_Scope_Constructed Space	Property_Scope_Extended Coverage	Property_Scope_Land Parcel	Property_Scope_Usable Interior	Location_DCC	...	BER_E	BER_F	BER_G
0	1	0	0	0	...	0	0	0
0	0	1	0	0	...	0	0	0
1	0	0	0	0	...	0	0	1
0	1	0	0	0	...	0	0	1
0	1	0	0	1	...	0	1	0
...	...	...	...	...	...	...	...	...
1	0	0	0	0	...	0	0	0
0	1	0	0	0	...	0	0	0
1	0	0	0	0	...	0	0	0
0	1	0	0	0	...	0	0	0
0	1	0	0	0	...	1	0	0

Fig 23.1: Categorically encoded columns

Renovation_needed_Maybe	Renovation_needed_No	Renovation_needed_Yes	Availability_season_Fall	Availability_season_Spring	Availability_season_Summer
0	1	0	1	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
...	...	...	...	...	...
1	0	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
1	0	0	0	0	1
0	0	1	0	0	0

Fig 23.2: Categorically encoded columns

Final Dataset Features:

- Numerical: Size, Bath, Balcony, Total_price_in_million_$, Affordability_ratio_scaled.
- Encoded Categorical: Binary columns for Property_Scope, Location, BER, Renovation_needed, and Availability_season.

Key Additions and Rationale:

1. Availability_season: Tracks Seasonal Trends

Categorizing properties into seasons (Winter, Spring, Summer, Fall) reveals how market demand and availability fluctuate throughout the year. For example, Spring and Summer often attract more buyers. This feature helps uncover temporal patterns and informs seasonally adjusted predictions.

2. Total_price($): Simplifies Property Value Analysis

By multiplying Total_sqft and Price_per_sqft($), this feature provides a clear measure of property value. It enables intuitive comparisons, highlights pricing trends, and supports segmentation of high-value and affordable properties. Scaling it into millions further simplifies interpretation.

3. Affordability_ratio_scaled: Highlights Property Affordability

The affordability ratio (Total_sqft / Total_price($)) indicates the amount of space buyers receive for each dollar spent. Scaling makes it easier to compare affordability across properties, helping identify good value options and target buyer segments effectively.

4. One-Hot Encoded Features: Prepares Categorical Data for Models

Encoding variables like Location, BER, and Availability_season into binary columns enhances model compatibility and predictive power. It clarifies how categories like energy-efficient ratings (BER: A, B) or specific locations influence outcomes, improving insights and decision-making.

These engineered features make it easier to explore the dataset and build models that can accurately predict property prices and trends based on availability, size, and other important factors.

Predictive Analysis

This phase focuses on building, evaluating, and comparing regression and classification models to predict property prices and buying intent using performance metrics tailored to each task.

1. Feature and Target Separation:

To prepare the data for predictive modeling, features and target variables were separated for both regression and classification tasks.

a. **For Regression**:
   - Features (X_regression): All columns except Total_price_in_million_$ and Buying_Intent.
   - Target (y_regression): Total_price_in_million_$.

b. **For Classification**:
   - Features (X_classification): Same as X_regression.
   - Target (y_classification): Buying_Intent.

2. Data Splitting:

The data was split into training and testing sets using a 75%-25% split for both regression and classification tasks:

- Training: 75% of the data.
- Testing: 25% of the data.

The random_state=42 parameter ensures reproducibility.

3. Model Definition:

The following models were selected for evaluation,

Regression Models:

- Linear Regression
- Gradient Boosting Regressor

Classification Models:

- Decision Tree Classifier
- Random Forest Classifier
- XGBoost Classifier
- AdaBoost Classifier

4. Model Evaluation Metrics

Each model was evaluated using relevant metrics:

- Regression Metrics: R^2 (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE).
- Classification Metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Confusion Matrices and ROC Curves were visualized for detailed performance insights.

Regression Metrics -

1. Linear Regression:
    - R^2: 0.5363
    - MAE: 0.1905
    - RMSE: 0.2792
2. Gradient Boosting Regressor:
    - R^2: 0.7249
    - MAE: 0.1326
    - RMSE: 0.2151

Classification Metrics -

1. Decision Tree Classifier:
    - Accuracy: 63.42%
    - Precision: 43.81%
    - Recall: 43.73%

- F1-Score: 43.77%
- ROC-AUC: 0.5821

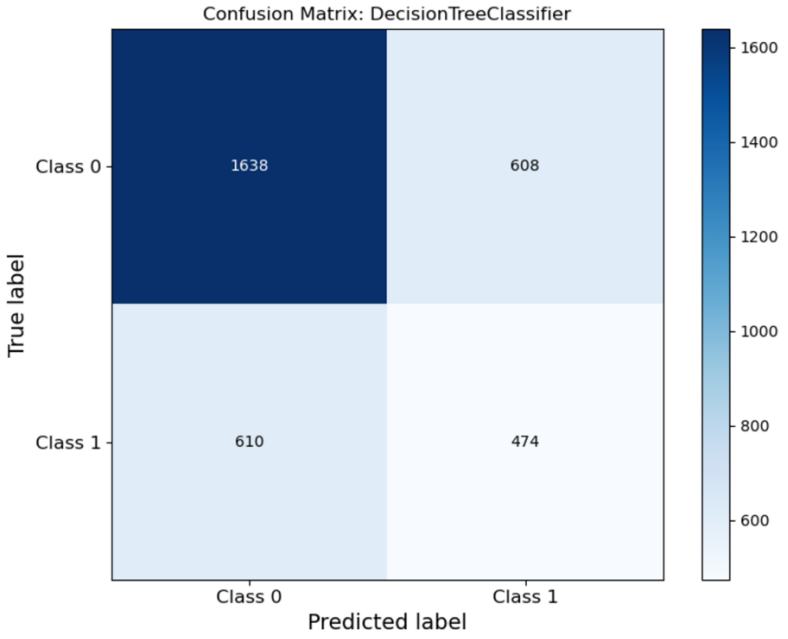

Fig 24: Confusion Matrix for Decision Tree Classifier

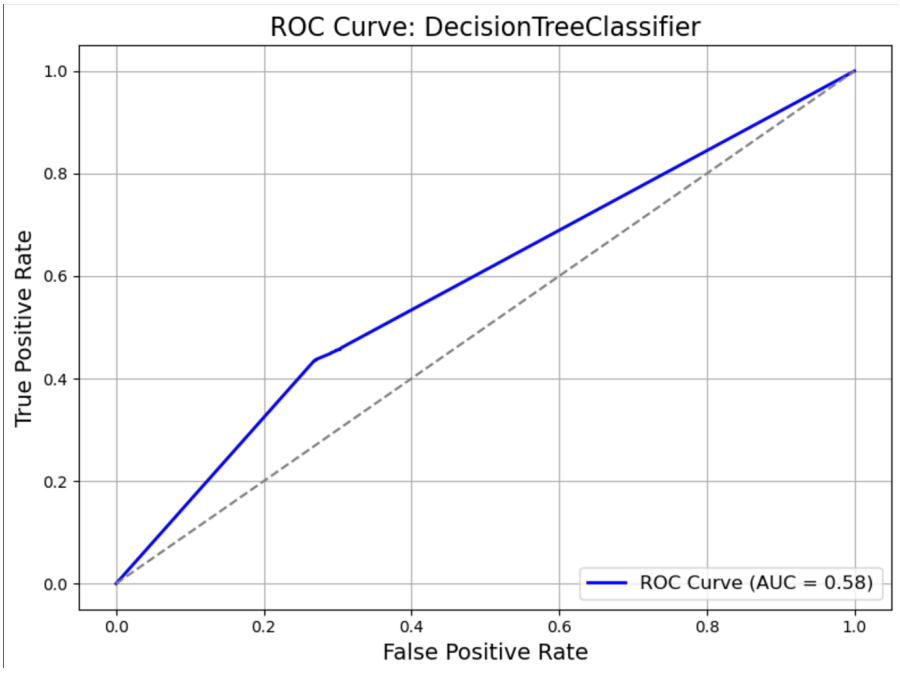

Fig 25: ROC Curve for Decision Tree Classifier

2. Random Forest Classifier:
    - Accuracy: 65.74%
    - Precision: 46.54%
    - Recall: 35.33%
    - F1-Score: 40.17%
    - ROC-AUC: 0.5950

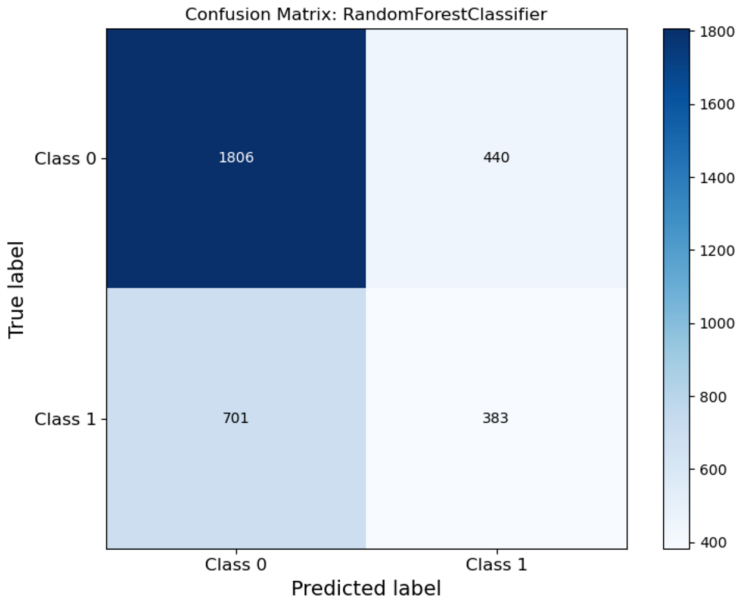

Fig 26: Confusion Matrix for Random Forest Classifier

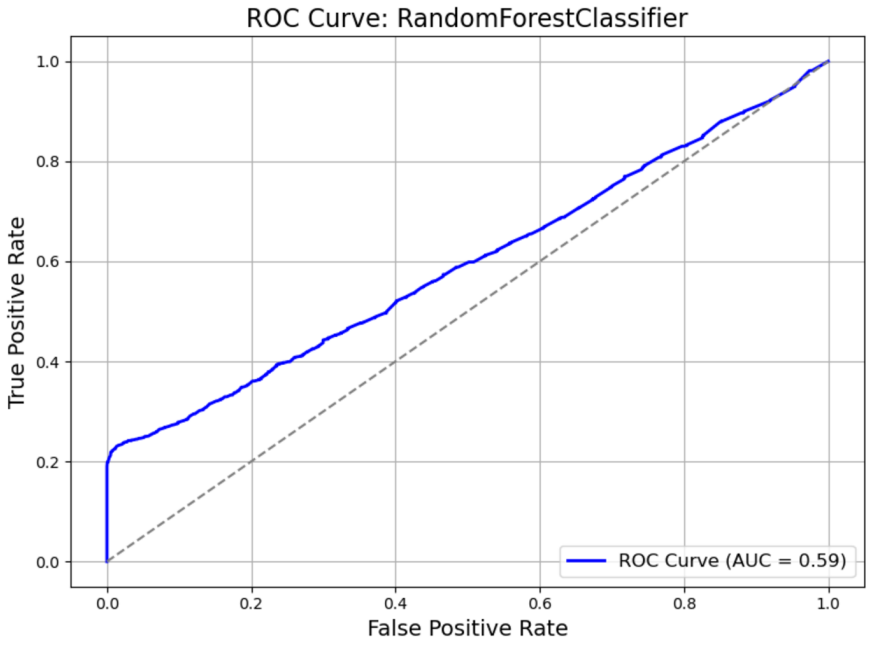

Fig 27: ROC Curve for Random Forest Classifier

3. XGBoost Classifier:
    - Accuracy: 72.82%
    - Precision: 72.89%
    - Recall: 26.29%
    - F1-Score: 38.64%
    - ROC-AUC: 0.5978

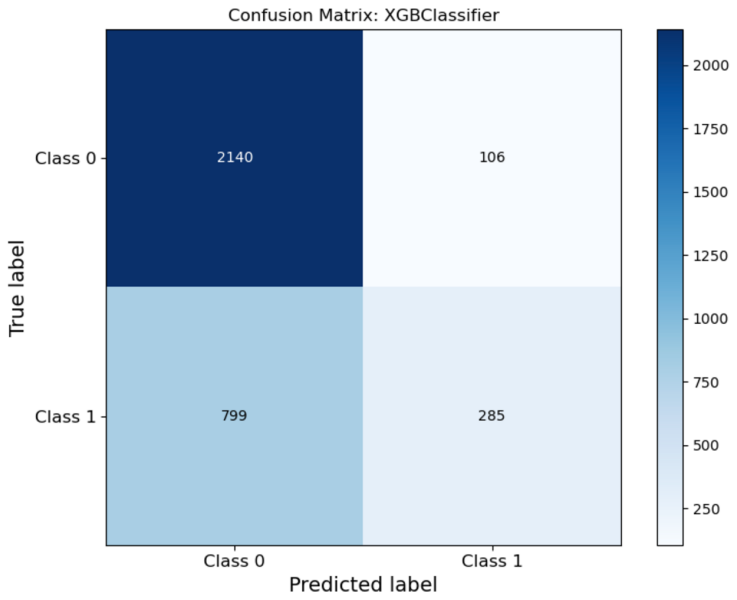

Fig 28: Confusion Matrix for XGBClassifier

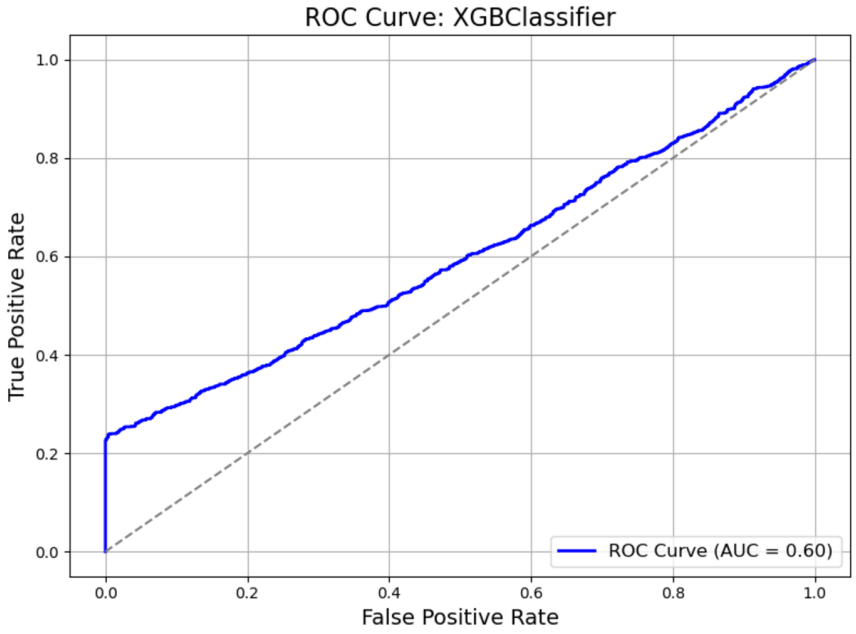

Fig 29: ROC Curve for XGBClassifier

4. AdaBoost Classifier:
    - Accuracy: 73.21%
    - Precision: 79.27%
    - Recall: 23.99%
    - F1-Score: 36.83%
    - ROC-AUC: 0.5956

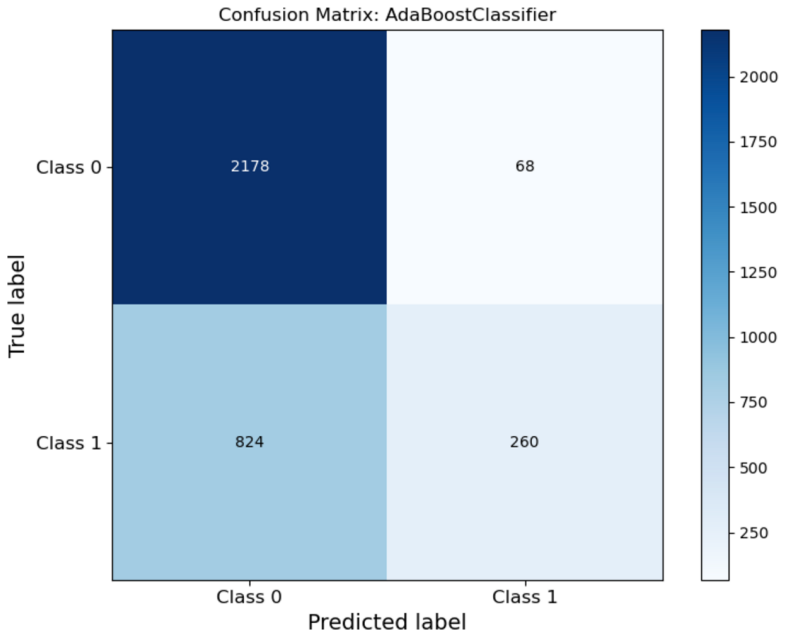

Fig 30: Confusion Matrix for AdaBoostClassifier

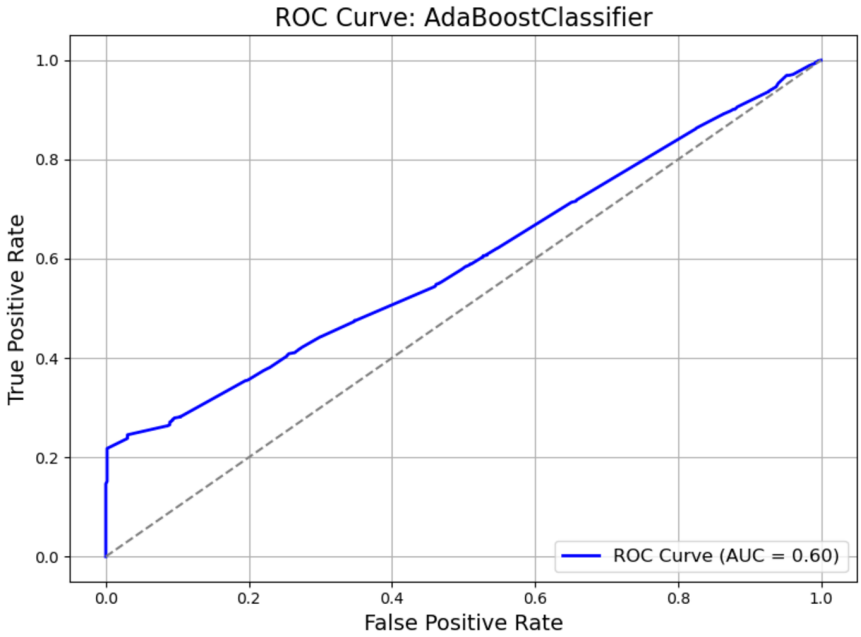

Fig 31: ROC Curve for AdaBoostClassifier

Key Observations:

1. Regression Models: Gradient Boosting Regressor performed much better with higher values of R^2 and lower errors than Linear Regression, and hence was considered the model of choice in predicting property prices.
2. Classification Models:
    - Overall Accuracy: The models which yielded the highest accuracy were AdaBoost and XGBoost (~73%).
    - Precision-Recall Tradeoff: In terms of accuracy AdaBoost proved slightly higher (79.27%) but on recall AdaBoost was notably lower (23.99%) which shows the specifics of AdaBoost in minimizing false positives.
    - ROC-AUC: Accuracy measurements of all the models were relatively okay, implying that there is potential to improve the features used or tuning the algorithm parameters.
3. General Trends: Generally, their performance favoured tree-based ensemble methods; Gradient Boosting, Random Forest and AdaBoost among others, proving how the methods perform better when they encounter complex patterns in the data.

Hyperparameter Tuning with GridSearchCV:

At this stage, further scroll through parameter arrays to the AdaBoostClassifier and GradientBoostingRegressor models for better test accuracy and prediction accuracy.

1. **AdaBoostClassifier** tuning results:

- Best Parameters: {'estimator__max_depth': 2, 'learning_rate': 0.01, 'n_estimators': 50}
- Best Cross-Validation Accuracy: 0.7572
- Test Accuracy: 0.7483

2. **GradientBoostingRegressor** tuning results:

- Best Parameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.8}
- Best RMSE (Cross-Validation): 0.2164
- Test RMSE: 0.2147
- Test R^2: 0.7259

Cross Validation:

1. **AdaBoostClassifier**:

- Cross-Validation Accuracy Scores: [0.7019, 0.7008, 0.7046, 0.7020, 0.7035]
- Mean CV Accuracy: 0.7026

2. **GradientBoostingRegressor**:

- Cross-Validation RMSE Scores: [0.2158, 0.2229, 0.2144, 0.2216, 0.2092]
- Mean CV RMSE: 0.2168

These results demonstrate the models' stability and performance across multiple data splits.

Feature Selection:

Feature selection is the process of constructing a new, but smaller, dataset to improve the outcomes of the model and its interpretability. For this purpose, Recursive Feature Elimination (RFE) was used.

1. **AdaBoostClassifier**:
   Selected features:
   - Renovation_needed_No
   - Availability_season_Spring
   - Availability_season_Summer
   - Availability_season_Winter
   - Affordability_ratio_scaled

As for these selected features, it appears that they target categorical characteristics and price. Notably, some of the features are not selected, including location and BER (Building Energy Rating); this shows that the model is focusing on more promptly influential features connected with the distribution and renovation requirements of properties.

2. **GradientBoostingRegressor**:
   Selected features:
   - Size
   - Bath
   - Property_Scope_Extended Coverage
   - Property_Scope_Land Parcel
   - Affordability_ratio_scaled

Once again, we see that Affordability_ratio_scaled is highly important, presumably because it could feed directly into Affordability. The model also incorporates size and bath count, which are conventional factors of price.

These selected features focus on the aspects affecting classification and regression models thus easing complication while preserving precision.

Using Feature Importances:

Feature importance helps one understand which elements greatly affect the predictions in the model and thereby provides a selective understanding of the developed dataset. Below are the results from the AdaBoostClassifier and GradientBoostingRegressor models:

1. **AdaBoostClassifier:**
   Top influential features and their importance scores:
   - Location_Dun Laoghaire: 0.3000
   - Renovation_needed_No: 0.2000
   - BER_B: 0.1800
   - BER_A: 0.1600
   - Location_South Dublin: 0.1200

Insights:

a. Location (especially Dun Laoghaire): This is the most decisive feature in the case of classification in the given occurrence. This gives an indication that to an extent, location is an important factor in determining if a given property gets sold or not.

b. Renovation Needed: Renovation_needed_No and BER (Building Energy Rating) also seems to be significant too.

2. **GradientBoostingRegressor:**
   Top influential features and their importance scores:
   - Affordability_ratio_scaled: 0.7502
   - Size: 0.1437
   - Bath: 0.0644
   - Property_Scope_Land Parcel: 0.0188
   - Property_Scope_Extended Coverage: 0.0090

Insights:

a. Affordability Ratio: Of all the constructed features, Affordability_ratio_scaled turns out to be the most crucial, which supports the hypothesis about the importance of affordable prices for the prediction of goods' prices.

b. Size and Bath: Another factor is the number of rooms, and the total area of the object is also huge.

c. Property Scope: Residing with positive arguments for Land Parcel and Property_Scope_Extended Coverage, stating that the type of property really does influence it, using Equation 3 which determines price per square foot.

Conclusion:

This analysis of Dublin's housing market looked at all its complicated data with a special eye. We used predictive analytics on lots of different local, national, and even international data. We did this because we really wanted to understand what drives people to buy properties here and what stops them from buying.

Throughout this report, we tried to tell a clear story. First, we talked about the data issues we faced, and then we showed the main results of our predictive analysis. In the end, we put together the two big talking points that keep coming up in discussions among housing researchers and policymakers: Affordability and Location.

Models that use ensembles, like the Gradient Boosting Regressor and the AdaBoost Classifier, have proven they can really understand how the market works. They show this by finding the right kind of factors. These factors that predict market value aren't what you would expect, but they are good predictors. They include things like how affordable something is and various categorical (or group) characteristics.

Predictive analytics is really full of potential when it comes to figuring out housing markets that are getting more and more complicated. In our report, we show that using these data skills can help plan neighborhood strategies in Dublin and other places around the world. This will give local leaders a much clearer view of the housing market's future.

References:

- **Research Paper**: Scikit-learn, 2024. GradientBoostingRegressor. Available at: https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (Accessed: 28 November 2024)
- **Research Paper**: Wang, R., 2012. AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. Procedia - Social and Behavioral Sciences, 40, pp. 493-497. Available at: https://www.sciencedirect.com/science/article/pii/S1875389212005767 (Accessed 28 November 2024)
- **Research Paper**: Aggarwal, C.C., 2016. Outlier Analysis. 2nd ed. Springer. Available at: https://doi.org/10.1007/978-3-319-47578-3 (Accessed: 26 November 2024)
- **Article**: Tiwari, R., 2024. Advanced Machine Learning with Scikit-Learn: A Deep Dive. *Medium*. Available at: https://medium.com/@rahultiwari065/advanced-machine-learning-with-scikit-learn-a-deep-dive-81f5e89158e5 (Accessed: 26 November 2024)