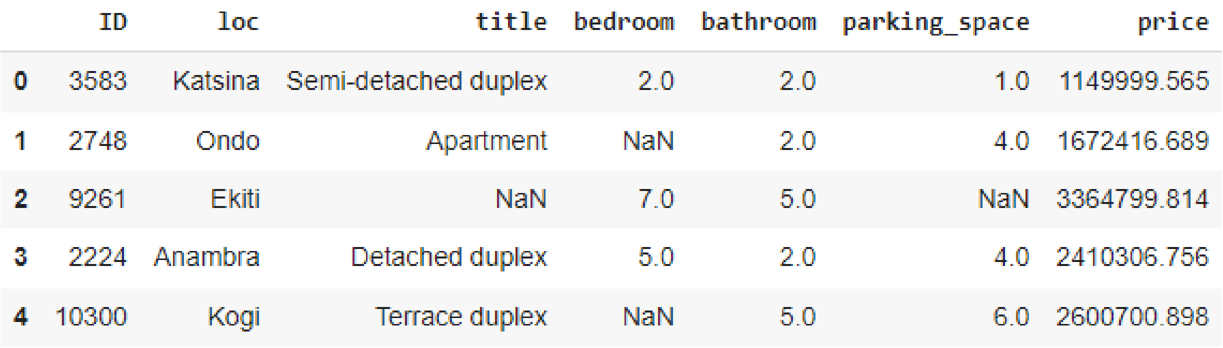
House Price Prediction Model Report

Introduction

This report presents the analysis and development of the house price prediction model for the 2023 DSN Bootcamp qualification project. The objective of this project is to build a model that accurately predicts house prices based on various features such as title, number of bedrooms, location, and other relevant factors. The dataset used for training and testing the model contains information on various houses and their corresponding prices.

Dataset Overview

The dataset contains 7 features, 4 numerical (one of which is the target variable) and 3 categorical features. One of which was dropped, given it will not add any value to the model. Some of the features with missing values were filled with descriptive values.

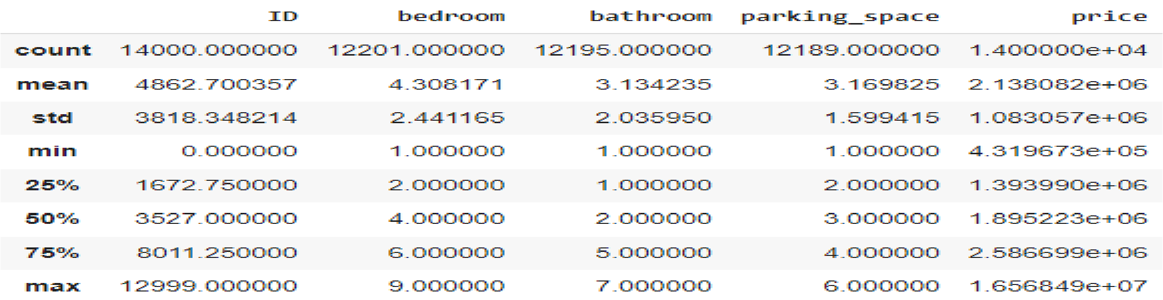


	ID	loc	title	bedroom	bathroom	parking_space	price
0	3583	Katsina	Semi-detached duplex	2.0	2.0	1.0	1149999.565
1	2748	Ondo	Apartment	NaN	2.0	4.0	1672416.689
2	9261	Ekiti	NaN	7.0	5.0	NaN	3364799.814
3	2224	Anambra	Detached duplex	5.0	2.0	4.0	2410306.756
4	10300	Kogi	Terrace duplex	NaN	5.0	6.0	2600700.898

Features Description

- **ID**: This is the unique identity representing each house details
- **Loc:** The location of the house in Nigeria
- **Title**: The type of house
- **Bedroom**: Number of bedrooms in the house
- **Bathroom**: Number of bathroom in the house
- **Parking Space**: Number of parking space in the house
- **Price**: This is the price the houses were sold at. (Target variable)

Data Summary



	ID	bedroom	bathroom	parking_space	price
count	14000.000000	12201.000000	12195.000000	12189.000000	1.400000e+04
mean	4862.700357	4.308171	3.134235	3.169825	2.138082e+06
std	3818.348214	2.441165	2.035950	1.599415	1.083057e+06
min	0.000000	1.000000	1.000000	1.000000	4.319673e+05
25%	1672.750000	2.000000	1.000000	2.000000	1.393990e+06
50%	3527.000000	4.000000	2.000000	3.000000	1.895223e+06
75%	8011.250000	6.000000	5.000000	4.000000	2.586699e+06
max	12999.000000	9.000000	7.000000	6.000000	1.656849e+07

Data Exploration

The first step in building the model is to explore and preprocess the data. This involves understanding the structure of the dataset, checking for missing values, handling outliers, and transforming categorical variables if present. We also perform data visualization to gain insights into the relationships between different features and the target variable (house prices).

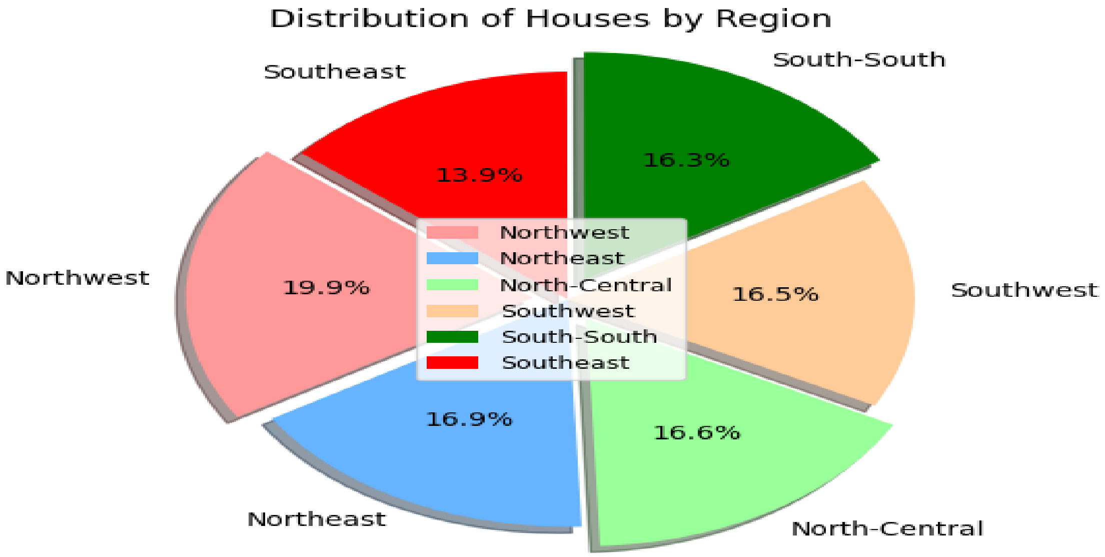


***Fig. 1**: Distribution of Houses by Region*

The chart in ***Fig. 1*** above shows that the six regions exhibit approximately equal proportions of houses. This suggests that all regions hold a significant share of the housing market. Such balanced distribution indicates that housing opportunities are fairly distributed across the area, presenting a positive outlook for potential homebuyers.

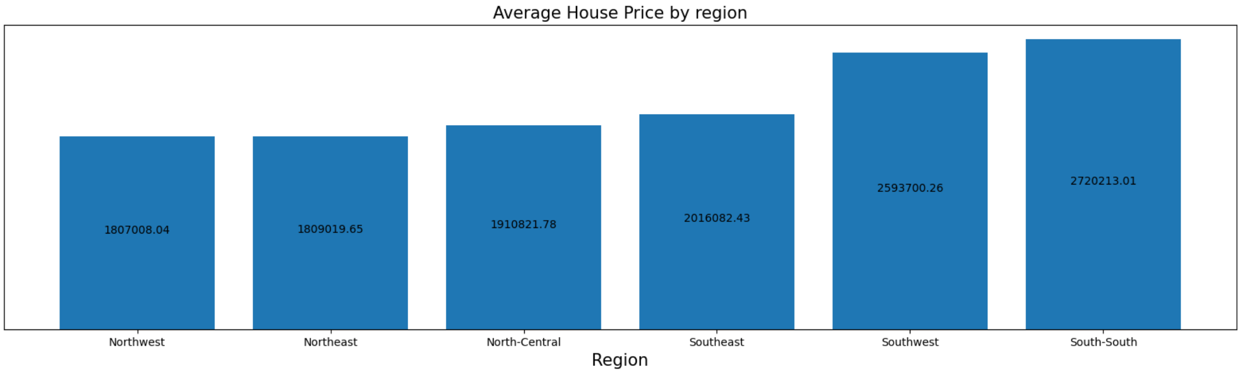


Fig. 2: *Average House price by Region*

From Fig.2 above, the two regions with the highest average house prices are Southwest and South-South. Houses in these regions are generally more expensive compared to other areas. The premium pricing in these regions may be attributed to factors such as proximity to amenities, renowned schools, or desirable neighborhood characteristics. Homebuyers in search of upscale properties and prime locations may find these regions appealing, albeit at a higher cost.

On the other hand, the regions of Southeast, North-Central, Northeast, and Northwest demonstrate the lowest average house prices among all regions. These areas present more affordable housing options, making them attractive to budget-conscious home buyers seeking cost-effective choices.

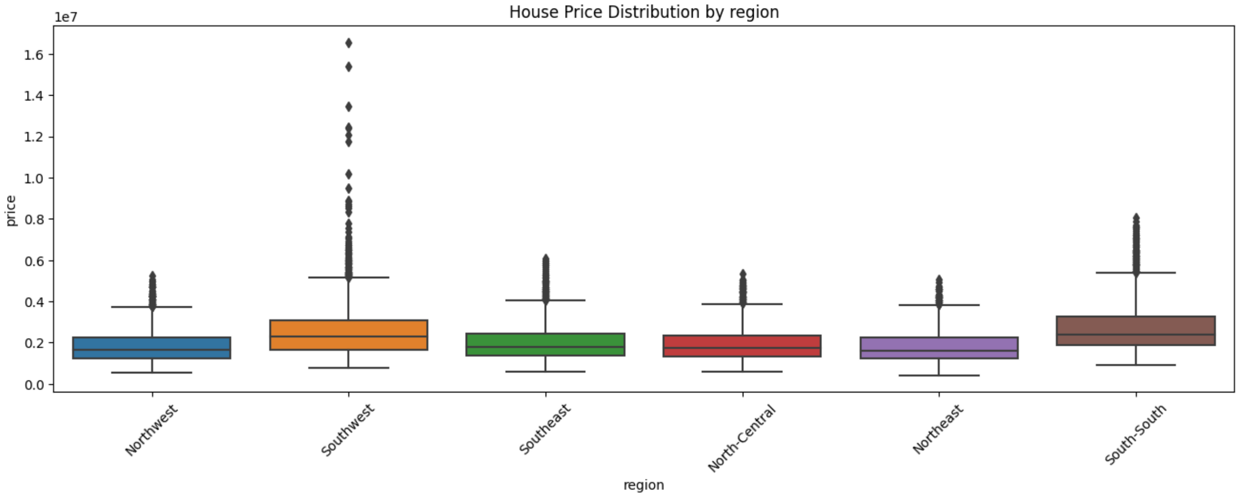


Fig. 3: *House price distribution by Region*

Fig.3 above further explains Fig.2. The box plot presents a clear picture of the distribution of house prices by region. Its symmetrical shape and presence of outliers across all regions indicate a relatively stable market with high price variability.

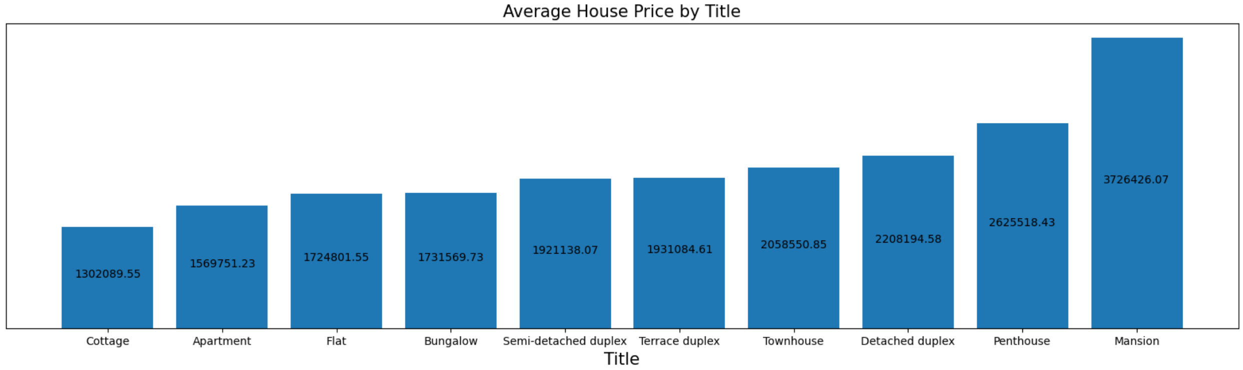


***Fig. 4**: Average House price by Title*

Fig.4 displays the average house prices categorized by different titles. The bar corresponding to the **Mansion** title shows the highest average house price among all the categories. This indicates that mansions generally command the highest prices in the market, making them a premium option for homebuyers. Following the mansion, the **Penthouse** title exhibits the second-highest average house price. Penthouse units are known for their luxury features and exclusive amenities, which contribute to their relatively higher prices. The **Detached Duplex** title secures the third position in terms of average house price. These properties typically offer more space and privacy compared to other house types, which can contribute to their higher price point.

The remaining house types show comparatively lower average house prices compared to mansions, penthouses, and detached duplexes.

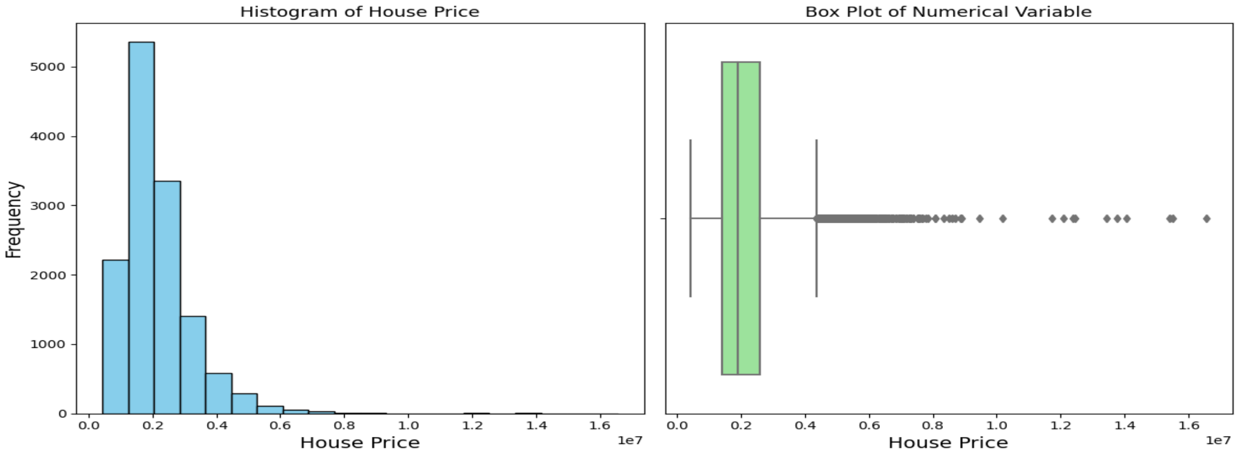


***Fig. 5**: Distribution of House Price*

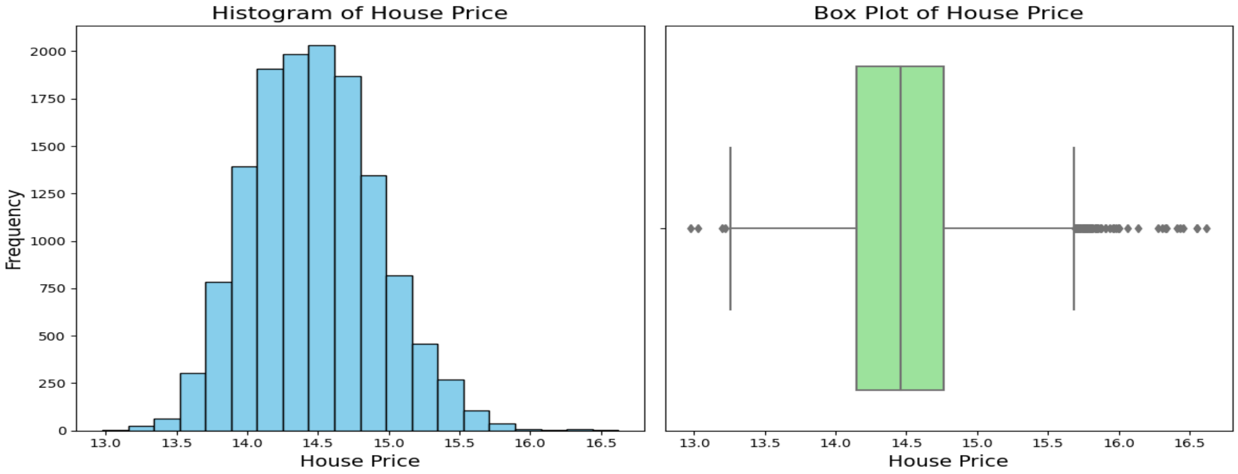


Fig. 6*: Distribution of Log Transformed House Price*

Figures 5 and 6 display the distribution of the house price and the corresponding logarithmic transformation of the house price. The original distribution of the house price exhibited skewness, which could potentially present challenges for the machine learning model. To address this issue, a log transformation was applied to the price, resulting in an improved distribution. Figure 6 illustrates that the transformed price now adheres to a normal distribution, making it suitable for the machine learning model.

Model Selection

Different machine learning regression models were trained on the dataset. The individual regression models (XGBoost Regressor, CatBoost Regressor, Random Forest Regressor, and LightGBM Regressor) encountered difficulties in accurately predicting house prices due to non-representative data features. To address this challenge, a model ensemble approach using a Voting Regressor was implemented, resulting in improved prediction performance.

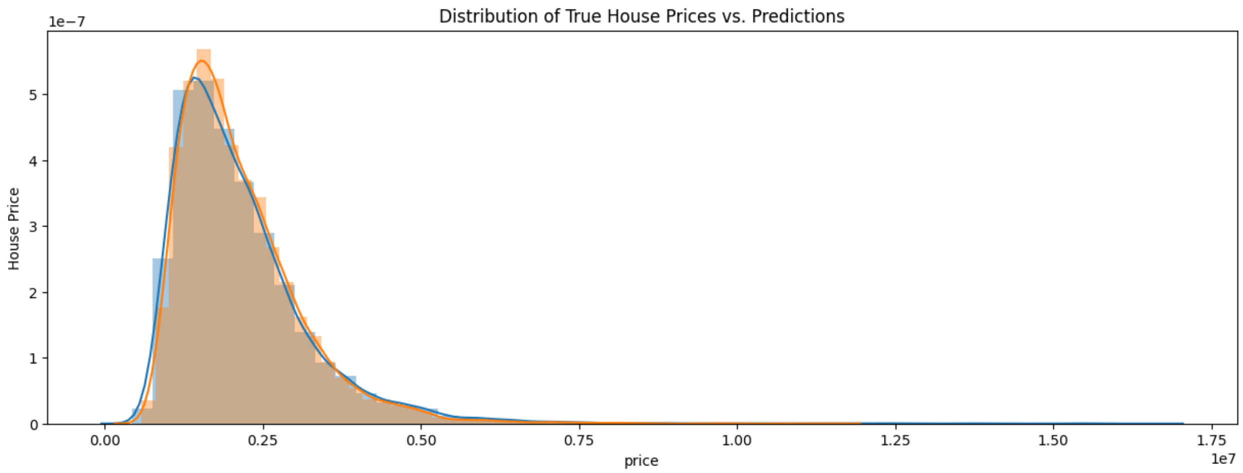
The ensemble technique allowed the models to work collaboratively, taking advantage of their respective strengths and mitigating their weaknesses. As a result, the ensemble is better equipped to capture the complex relationships between features and house prices, leading to more reliable and accurate predictions.

Steps Taken in Model Ensemble

- **Model Selection**: Three XGBoost Regressor, one CatBoost Regressor, one Random Forest Regressor, and three LightGBM Regressor were chosen as the base regression models for the ensemble.
- **Ensemble Construction:** The predictions from each individual model were combined using the Voting Regressor, which employs a weighted averaging to arrive at the final prediction.

Model Evaluation

To evaluate the model's performance, we use the mean squared error (MSE). The goal is to minimize the errors and achieve the best possible accuracy in predicting house prices. We also perform cross-validation to assess the model's generalization capability.



From the plot, we can see that the predicted prices generally follow the trend of the true prices, but there are some discrepancies. For example, there are some houses that were predicted to be sold for a higher price than their true price. Conversely, there are some cars that were predicted to be sold for a lower price than their true price.

Overall, the plot suggests that the machine learning model is performing reasonably well, but there is still room for improvement.

Challenge Faced

The existing features in the dataset used for house price prediction, such as the number of bedrooms, location, title, bathroom count, and parking space, were found to have limited representativeness. This means that these features may not fully capture the complex relationships and patterns that determine house prices, resulting in suboptimal predictive performance of the models.

To improve the accuracy of house price prediction, it is crucial to incorporate additional relevant features. Some suggested features to consider are property size and dimensions, property condition, nearby amenities, property view and orientation, neighborhood demographics, market trends, energy efficiency, upcoming developments, crime rate and safety, nearby environmental factors, historical sales data, and interest rates/mortgage data.

By integrating these additional features and employing advanced machine learning techniques, we can enhance the model's ability to capture intricate relationships and achieve more accurate house price predictions.