DSN and Microsoft Skills for Job Hackathon
House price prediction model
Executive Report
DSN and Microsoft Skills for Job Hackathon

Description

Wazobia Real Estate Limited is a prominent real estate company operating in Nigeria. With a vast portfolio of properties, they strive to provide accurate and competitive pricing for houses. The goal is to build a powerful and accurate predictive model that can estimate the prices of houses in Nigeria and provide Wazobia Real Estate Limited with an effective tool to make informed pricing decisions and enhance their competitiveness in the market.

Introduction

This executive report presents the findings and insights from the development and analysis of a CatBoost Regression Model for predicting house prices. The primary objective of this project was to build an accurate and reliable predictive model that can assist in estimating house prices based on various features such as:

- Number of bedrooms
- Number of bathrooms
- Location
- Title
- Parking spaces

Exploratory Data Analysis (EDA)

In this section, I conducted Exploratory Data Analysis to gain a comprehensive understanding of the dataset used in training the machine learning model

Data Source and Description

The dataset used in this analysis was gotten from Zindi which can be accessed here: https://zindi.africa/competitions/free-ai-classes-in-every-city-hackathon-2023/data, containing information about various houses and their corresponding selling prices. The dataset consists of 14000 rows and 7 columns. The features present in the dataset include:

- Number of bedrooms
- Number of bathrooms
- Location
- Title
- Parking spaces
- ID
- Price

Bedrooms: This simply indicates the number of bedrooms each house have.

Bathrooms: This feature indicates the number of bathrooms present in the house.

Title: This categorical feature indicates the house type. For example Bungalow, Townhouse etc.

Location: This feature is categorical and it indicates the part of Nigeria the house is situated in. That is the 36 states in Nigeria.

Parking Space: This feature indicates the number of available space for parking.

ID: The ID of each house present in the dataset

Price: This feature which is only present in the train dataset indicates the worth of each house, also it is the target i.e the feature the model aims to predict.

Below is a summary statistics of each feature:

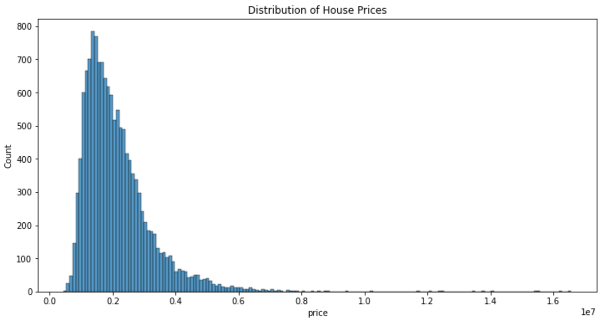

Fig 1

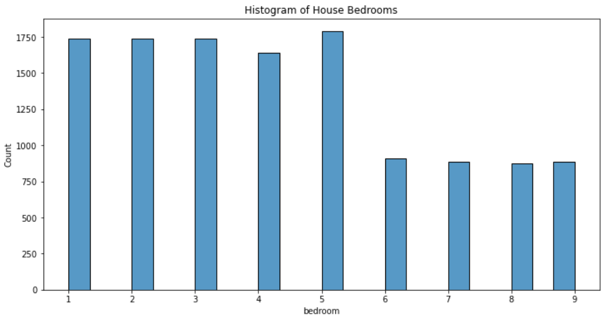

fig 2

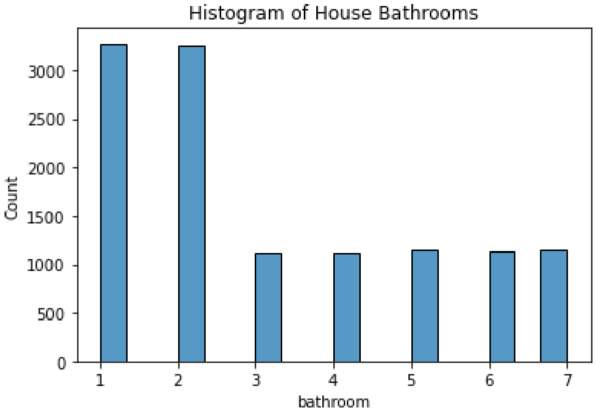

Fig 3

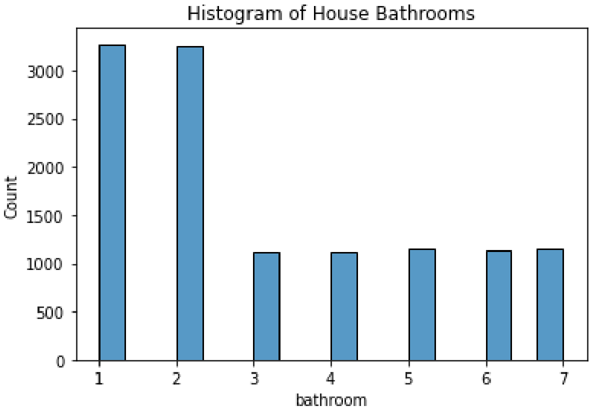

fig 4

Data Preprocessing

Prior to building the predictive model, the dataset underwent different preprocessing steps including:

- Handling of missing values and
- Feature Engineering

Handling of Missing Values: From **fig 5** below it can be seen that the features are not symmetrical and thus they are positively skewed. With this insight, filling the missing values in the numerical features with the median instead of the mode is more reasonable. This is because the mean can be significantly affected by the extreme values in the tail, pulling it away from the majority of the data points. This results in a mean that may not be representative of the typical values in the dataset. On the other hand, the median is less influenced by extreme values because it only depends on the middle value or the value at the 50th percentile.

The missing values in the numerical columns which include: bedroom, bathroom, and parking space were filled with their respective medians. Meanwhile the rows which contained the missing values in the categorical columns (location and title) were dropped.

The price column in the dataset had no missing values.

Using the median instead of the mean is advantageous when dealing with skewed data because:

1. **Robustness to Outliers:** The median is not affected by extreme values, making it more robust and providing a better estimate of the central tendency for skewed distributions.

2. **Preserving the Typical Value:** In skewed data, the majority of values are clustered in one region. The median represents the value that divides the data into two equal halves, preserving the typical value in the dataset.
3. **Interpretability:** The median is easier to interpret and understand than the mean, especially when the data distribution is skewed.

Feature Engineering: I transformed the some features present in the dataset to enhance its predictive power.

Bed to Bath Ratio: This feature was created by dividing the number of bedrooms by the number of bathrooms. This feature can potentially indicate the overall space and amenities offered by the house. A higher ratio might imply that the house has more bedrooms relative to the number of bathrooms, indicating a potentially larger living area or more private spaces. A higher ratio of bedrooms to bathrooms might indicate more living space and potentially larger bedrooms. Houses with more bedrooms relative to bathrooms could be perceived as more comfortable, especially for larger families or households with multiple members. As a result, such houses may command a higher price in the market.

House size: This feature was created by summing up the number of bedrooms, bathrooms, and parking space. This feature collectively reflects the size and amenities offered by the property. Larger houses with more bedrooms and bathrooms and ample parking spaces are generally perceived as more desirable and may command higher prices. The presence of multiple bathrooms and parking spaces adds to the convenience and comfort of the property. Houses with more bathrooms are more functional for families or households with multiple occupants, and ample parking spaces are attractive for car-owning buyers, both of which can positively influence the house price.

Convenience: This feature was created by multiplying the bed to bath ratio with the house size. This feature represents the balance between bedrooms and bathrooms in the house. A higher ratio indicates more bedrooms relative to the number of bathrooms, potentially suggesting larger bedrooms or a more private living arrangement. This can influence house prices, as some buyers may prefer properties with more bedrooms per bathroom, while others may have different preferences.

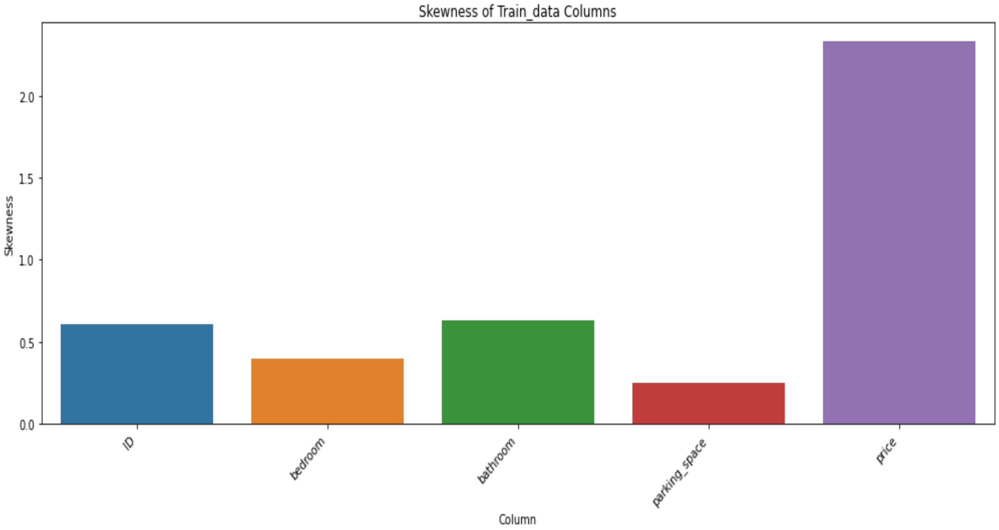

Fig 5

We use the median instead of the mean for skewed data because the median is more robust to extreme values or outliers that can heavily influence the mean and lead to a misleading representation of the central tendency. The mean can be significantly affected by the extreme values in the tail, pulling it away from the majority of the data points. This results in a mean that may not be representative of the typical values in the dataset. On the other hand, the median is less influenced by extreme values because it only depends on the middle value or the value at the 50th percentile.

Feature Importance

To better understand the impact of each feature on the house prices, I analyzed the feature importance obtained from the trained model. This analysis helps identify which variables have the most significant influence on the predicted house prices. Take a look at **fig 6** below:

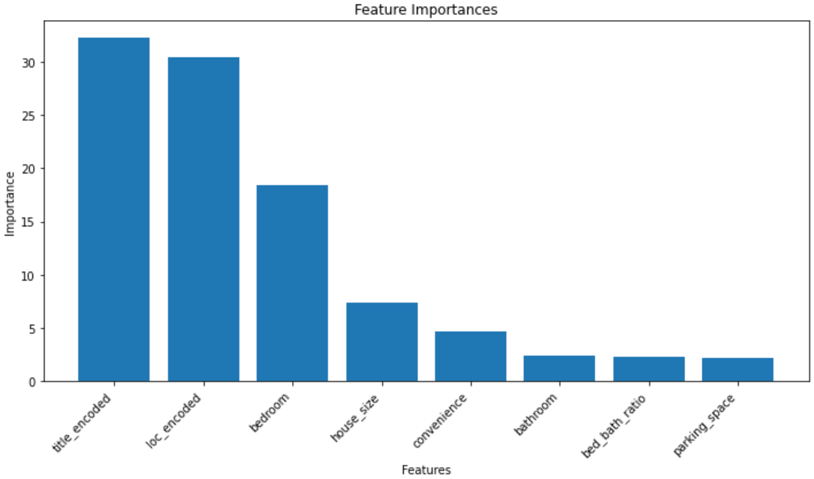

Fig 6

Distribution of the Target Variable

I examined the distribution of the target variable (house prices) to gain insights into its characteristics and assess if it follows a normal distribution. But from **fig 1** above, it appears that the price is not normally distributed.

Model Architecture

CatBoost is a powerful gradient boosting library that is well-suited for regression tasks. Its ability to handle categorical features and automatically handle missing values makes it a popular choice for house price prediction.

Model Training and Evaluation

In the model training and evaluation part, we utilized the cross-validation technique to assess the performance of the CatBoost Regression Model for predicting house prices. Cross-validation is a powerful method that allows us to make efficient use of the available data by repeatedly splitting it into multiple subsets for training and validation.

The cross-validation process involved the following steps:

1. **Data Splitting:** The dataset was randomly divided into 5 folds. Each fold acts as a holdout set for validation, while the remaining data is used for training.

2. **Model Training:** The CatBoost Regression Model was then trained on each training fold, learning the patterns and relationships between the features and the target variable.
3. **Model Validation:** The trained model was then evaluated on the corresponding validation fold to calculate its performance using the Root Mean Squared Error (RMSE). **The root mean squared error of my model was seen to be 311491.9931, which means that the CatBoost model's predictions deviate by approximately 311491.9931 units from the true house prices**
4. **Performance Aggregation:** The performance metrics from each fold were aggregated to calculate the average performance, providing a robust estimation of the model's generalization ability.
5. **Hyper-parameter Tuning:** I performed hyper-parameter tuning using random search during each fold to identify the best set of hyper-parameters for the model.
6. **Final Model:** After cross-validation, the final CatBoost Regression Model was trained on the entire dataset using the best hyper-parameters obtained from the cross-validation process.

The use of cross-validation ensures that the model is robust and less prone to overfitting or underfitting, as it tests the model's performance on multiple subsets of the data. It also helps better understand the model's stability and variance across different training-validation splits.

Model Interpretability

One of the key advantages of CatBoost is its built-in feature importance mechanism, which allows us to understand the contributions of each feature to the model's predictions. This interpretability is valuable for gaining insights into the driving factors behind house prices. This can be seen in **fig 6** above.

Biases behind the Model

It is essential to acknowledge the presence of potential biases that might influence the predictive model's outcomes. Biases can arise from various sources, including but not limited to:

- **Data Bias**: The training data might be skewed towards certain types of houses or locations, leading to biased predictions for specific segments of the housing market.
- **Feature Selection Bias**: The selection of features may inadvertently favor certain attributes over others, potentially leading to biased predictions.
- **Sampling Bias**: If the data collection process is biased or non-representative, the model may not generalize well to new, unseen data.

Conclusion

The CatBoost Regression Model presented in this report provides a reliable means for predicting house prices based on the provided dataset. The model's interpretability will allows Wazobia Real Estate Limited to gain insights into the factors driving house prices and better understand the dynamics of the housing market.