**A PROJECT REPORT ON HOUSE PRICE PREDICTION BY OSEDEME KENOSE EKAOSE**

1. Introduction; Aims and Objectives
2. Methodology
3. Findings & Results
4. **Introduction: Aims and Objectives**

In this PWC-Xerde bootcamp project, participants are tasked with carrying out Data Management tasks, Insights and visualization tasks as well as Advance Analytics (Machine Learning) tasks on a House Price Prediction dataset.

1. **Methodology**

The dataset was already provided, so it made the work a lot easier. To build advanced analytics and predictive models the following steps were taken.

* 1. Problem Statement & PowerBI Visualizations:

*Assuming Powell is a real estate Agent, and he is in charge of selling a new house. The price is unknown, but he wants to infer it by comparing it with other houses.*

* *What features of a house e.g., size, number of rooms, location, crime rate, school quality, distance to commerce, etc. could influence the price?*
* *At the end of the day, can we get a formula on all these features which gives us the price of the house, or at least an estimate for it?*



Figure 1: Visualizing the House Price Dataset

* 1. Importing Data into R.

After loading the ggplot2, readr, gplots and repr libraries the data was imported and stored in the ‘train’ variable.

* 1. Check for Missing Data

I found that every row in this dataset has missing value, and I will treat it later on.

* 1. Select Variables

The next step is to select variables that may have greater Impact on the price of the house, and then build a subset of train dataset for prediction

To get a better understanding of this dataset, I summarized all important variables in terms of maximum, first quartile, median, mean, third quartile and maximum value.

* 1. Predictive Analysis: Building the regression model

In our LR model, the relationships between the dependent and independent variables are expressed by an equation that has coefficients. The objective of this model is to minimize the sum of squared residuals. 16 variables were selected to fit into this model. After this, the next step is to choose the variables and transfer SalePrice into log term.

* 1. Divide datasets into two parts -training and validation

The validation set is the data used to provide an unbiased evaluation of a model fitted on the training dataset while tuning model hyperparameters. It also plays a role in other forms of model preparation, such as feature selection, threshold cut-off selection.

* 1. Run Regression, Forecast and check for accuracy

After creating our regression model that appears to produce unbiased predictions and can predict new observations nearly as well as it predicts the data used to fit the model, we then go ahead to use the model to make a prediction and assess the precision.

* 1. Storing the R Model in SQL and Importing it to Power BI

The regression model was saved as a csv file and imported to the HomePrediction database, after which the data was imported from the server into Power BI. Visualizations were then carried out afterward.

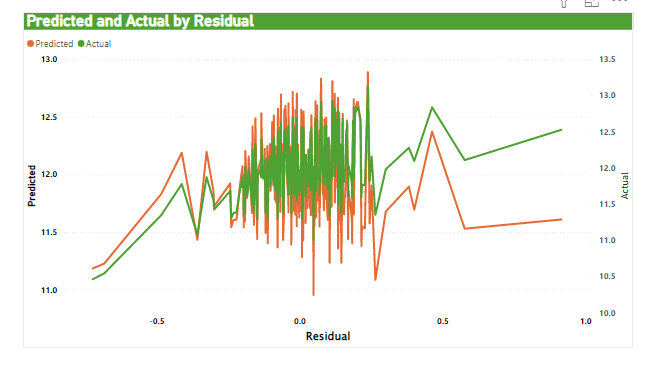


Figure 2:Visualizing the Prediction Data

1. **Findings and Results**

The objective of this project was to build models to predict housing prices of different residences. Our best model resulted in an RMSE of 0.1071, which translates to an error of about $9000 (or about 5%) for the average-priced house.

While this error is quite low, the interpretability of our model is poor. Each model found within our ensembled model varies with respect to the variables that are most important to predicting Sale Price. The best way to interpret our ensemble is to look for shared variables among its constituent models. The variables seen as most important or as strongest predictors through our models were those related to square footage, the age and condition of the home, the neighborhood where the house was located, the city zone where the house was located, and the year the house was sold.