* **PROJECT TITLE: PREDICTING HOUSE PRICE USING MACHINE LEARNING**



**Introduction:**

* Development of civilization is the foundation of the increase in demand for houses day by day. Accurate prediction of house prices has been always a fascination for buyers, sellers, and bankers also. Many researchers have already worked to unravel the mysteries of the prediction of house prices. Many theories have been given birth as a consequence of the research work contributed by various researchers all over the world. Some of these theories believe that the geographical location and culture of a particular area determine how the home prices will increase or decrease whereas other schools of thought emphasize the socio-economic conditions that largely play behind these house price rises.
* We all know that a house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task. To forecast house prices one person usually tries to locate similar properties in his or her neighborhood and based on collected data that person will try to predict the house price.
* All these indicate that house price prediction is an emerging research area of regression that requires the knowledge of machine learning. This has motivated me to work in this domain.
* Realestate appraisal is an integral part of the property buying process. Traditionally, the appraisal is performed by professional appraisers specially trained for real estate valuation. For the buyers of real estate properties, an automated price estimation system can be useful to estimate the prices of properties currently on the market. Such a system can be particularly helpful for novice buyers who are buying a property for the first time, with little to no experiences.

**Content for project phase 2:**

Consider exploring advanced regression techniques like gradients boosting or XGBoost for improved predicting accuracy.

**Data Source:**

A good data source for house price predicting using machine learning should be accurate,complete,covering the geographic area of interest,accessible.

**Data Collection and Preprocessing:**

* Importing the dataset: Obtain a comprehensive dataset containing relevant features such as square footage, number of bedrooms, location, amenities, etc.
* Data preprocessing: Clean the data by handling missing values, outliers, and categorical variables. Standardize or normalize numerical features.

**Exploratory Data Analysis (EDA):**

* Visualize and analyze the dataset to gain insights into the relationships between variables.
* Identify correlations and patterns that can inform feature selection and engineering.
* Present various data visualizations to gain insights into the dataset.
* Explore correlations between features and the target variable (house prices).
* Discuss any significant findings from the EDA phase that inform feature selection.

**Feature Engineering:**

* Create new features or transform existing ones to capture valuable information.
* Utilize domain knowledge to engineer features that may impact house prices, such as proximity to schools, transportation, or crime rates.
* Explain the process of creating new features or transforming existing ones.
* Showcase domain-specific feature engineering, such as proximity scores or composite indicators.
* Emphasize the impact of engineered features on model performance.

**Advanced Regression Techniques:**

* **Ridge Regression:** Introduce L2 regularization to mitigate multicollinearity and overfitting.
* **Lasso Regression:** Employ L1 regularization to perform feature selection and simplify the model.
* **ElasticNet** **Regression:** Combine both L1 and L2 regularization to benefit from their respective advantages.
* **Random Forest Regression:** Implement an ensemble technique to handle non-linearity and capture complex relationships in the data.
* **Gradient Boosting Regressors (e.g., XGBoost,** **LightGBM):** Utilize gradient boosting algorithms for improved accuracy.

**Model Evaluation and Selection:**

* Split the dataset into training and testing sets.
* Evaluate models using appropriate metrics (e.g., Mean Absolute Error, Mean Squared Error, R-squared) to assess their performance.
* Use cross-validation techniques to tune hyperparameters and ensure model stability.
* Compare the results with traditional linear regression models to highlight improvements.
* Select the best-performing model for further analysis.

**Model Interpretability:**

* Explain how to interpret feature importance from Gradient Boosting and XGBoost models.
* Discuss the insights gained from feature importance analysis and their relevance to house price prediction.
* Interpret feature importance from ensemble models like Random Forest and Gradient Boosting to understand the factors influencing house prices.

**Deployment and Prediction:**

* Deploy the chosen regression model to predict house prices.
* Develop a user-friendly interface for users to input property features and receive price predictions.

**PROGRAM:**

**HOUSE PRICE PREDICTION**

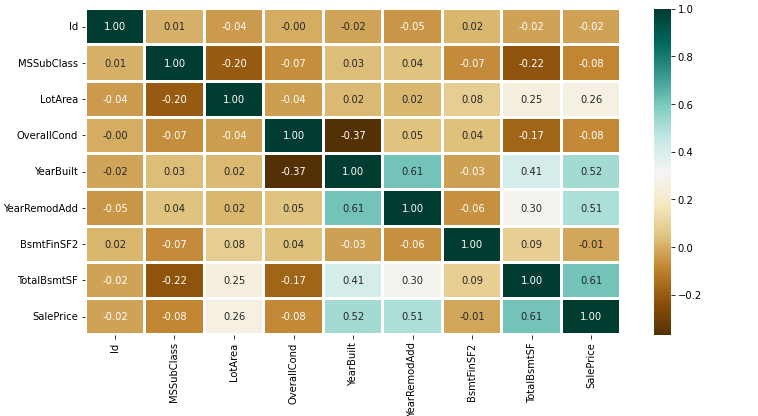
**Exploratory Data Analysis**

[EDA](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/) refers to the deep analysis of data so as to discover different patterns and spot anomalies. Before making inferences from data it is essential to examine all your variables.

So here let’s make a [heatmap](https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python/) using seaborn library.

|  |
| --- |
| **plt.figure(figsize=(12, 6))**  **sns.heatmap(dataset.corr(),**  **cmap = 'BrBG',**  **fmt = '.2f',**  **linewidths = 2,**  **annot = True)** |

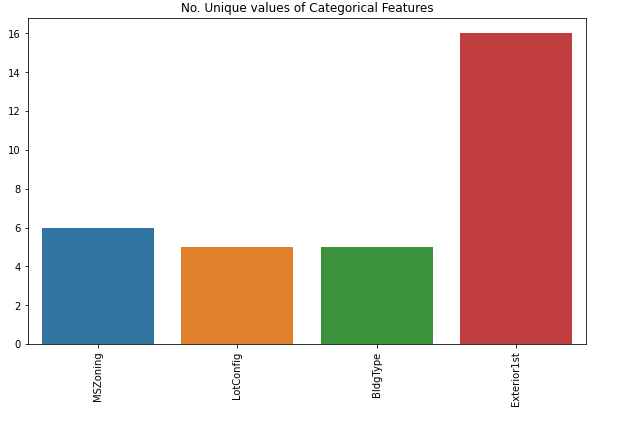
**Output:**



To analyze the different categorical features. Let’s draw the [barplot](https://www.geeksforgeeks.org/bar-plot-in-matplotlib/).

|  |
| --- |
| **unique\_values = []**  **for col in object\_cols:**  **unique\_values.append(dataset[col].unique().size)**  **plt.figure(figsize=(10,6))**  **plt.title('No. Unique values of Categorical Features')**  **plt.xticks(rotation=90)**  **sns.barplot(x=object\_cols,y=unique\_values)** |

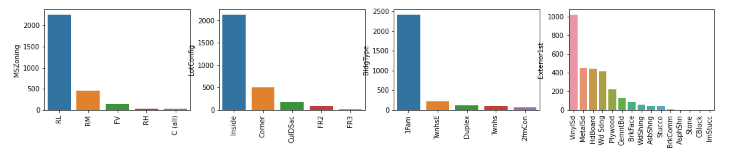
**Output:**



The plot shows that Exterior1st has around 16 unique categories and other features have around  6 unique categories. To findout the actual count of each category we can plot the bargraph of each four features separately.

|  |
| --- |
| **plt.figure(figsize=(18, 36))**  **plt.title('Categorical Features: Distribution')**  **plt.xticks(rotation=90)**  **index = 1**    **for col in object\_cols:**  **y = dataset[col].value\_counts()**  **plt.subplot(11, 4, index)**  **plt.xticks(rotation=90)**  **sns.barplot(x=list(y.index), y=y)**  **index += 1** |

**Output:**



**Data Cleaning:**

* Data Cleaning is the way to improvise the data or remove incorrect, corrupted or irrelevant data.
* As in our dataset, there are some columns that are not important and irrelevant for the model training.So,we can drop that column before training. There are 2 approaches to dealing with empty/null values.
* We can easily delete the column/row (if the feature or record is not much important).
* Filling the empty slots with mean/mode/0/NA/etc. (depending on the dataset requirement).

As Id Column will not be participating in any prediction. So we can Drop it.

|  |
| --- |
| **dataset.drop(['Id'],**  **axis=1,**  **inplace=True)** |

Replacing SalePrice empty values with their mean values to make the data distribution symmetric.

|  |
| --- |
| **dataset['SalePrice'] = dataset['SalePrice'].fillna(**  **dataset['SalePrice'].mean())** |

Drop records with null values (as the empty records are very less).

|  |
| --- |
| **new\_dataset = dataset.dropna()** |

Checking features which have null values in the new dataframe (if there are still any).

|  |
| --- |
| **new\_dataset.isnull().sum()** |

**Output:**



**Conclusion and Future Work (Phase 2):**

**Project Conclusion:**

* In the Phase 2 conclusion, we will summarize the key findings and insights from the advanced regression techniques. We will reiterate the impact of these techniques on improving the accuracy and robustness of house price predictions.
* Future Work: We will discuss potential avenues for future work, such as incorporating additional data sources (e.g., real-time economic indicators), exploring deep learning models for prediction, or expanding the project into a web application with more features and interactivity.