**House Price Prediction Model**

**About Project**

I have made this House Price Prediction Model using Linear, Lasso and Ridge Regression. The main objectives of this case study are as follows:

1. To apply data preprocessing and preparation techniques in order to obtain clean data (EDA).
2. To build machine learning models able to predict house price based on house features.
3. To analyze and compare models performance in order to choose the best model.

**Problem Statement:**

A machine learning model is to be proposed to predict a house price based on data related to the house i.e., its area type, availability, location, size, society, total square ft, bath and balcony using Regression.

**Introduction**

**Linear Regression**

It is used when we want to make predictions for continuous variables (in our case it's house price). The independent variable (or target column) is predicted based on how the value of the dependent variable is changing according to the value of the independent variable.

Lasso and Ridge are all part of the Linear Regression family where the x (predictors) and y (target) are assumed to have a linear relationship. In sklearn, LinearRegression refers to the most ordinary least square linear regression method without regularization (penalty on weights).

**Lasso Regression**

Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros. Lasso introduced a new hyperparameter, alpha, the coefficient to penalize weights.

**Ridge Regression**

Ridge takes a step further and penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed.

**Data:**

Link to the data: <https://www.kaggle.com/datasets/bhavik0901/bangalore-house-price-prediction>

This data is from Kaggle. It has 13,320 rows and 9 columns as described below:

1. area\_type -> There are 4 area types here i.e., plot area, built-up area, super built-up area and carpet area
2. availability -> This column shows whether the house is ready to move or not and if not, the date of availability is given.
3. location -> This column indicates the location of the house.
4. size -> This column tells the number of rooms in the property.
5. society -> This column gives the name of the society where the aprtment is located
6. total\_sqft -> This column conveys the area measurement of house in square feet.
7. bath -> This column shows the number of bathrooms in the house.
8. balcony -> This column shows the number of balconies in the house.
9. price -> This is our target column which indicates the price of the house

**Steps followed are:**

📌 Applied data preprocessing and preparation techniques in order to obtain clean data which includes following steps:

1. Importing and understanding of data
2. Checking for duplication
3. Missing value check: Since null values are present, simply dropped the null value rows for 'size' column and also used imputation techniques to fill null values for 'location' , 'bath' and 'balcony' columns. Also dropped the columns with excessive null values.
4. Correlation Check -> The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity. An example is linear regression, where one of the offending correlated variables should be removed in order to improve the skill of the model. Here 'bath' and 'bhk' are highly correlated, so I dropped 'bhk' column.
5. Data Reduction -> Some columns or variables are dropped if they do not add value to our analysis. Here I dropped 'balcony' and 'availability' columns.
6. Data Cleaning/Wrangling -> Some data may have data entry errors, and some variables may need data type conversion . The column 'total square feet' needs type coversion and also all other area units should be converted to square feet as well.
7. Feature Engineering -> Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model. The main goal of Feature engineering is to create meaningful data from raw data. Here I transformed location and size column.
8. Univariate Analysis -> Analyzing/visualizing the dataset by taking one variable at a time.
9. Bivariate Analysis -> Bivariate Analysis helps to understand how variables are related to each other and the relationship between dependent and independent variables present in the dataset.

📌 Built machine learning model to be able to predict house price based on house features using Linear regression, Lasso regression and Ridge regression which includes following steps:

1. Splitting data into training and testing data
2. Building and training the model
3. Making predictions from the model
4. Testing the performance of the model -> There are three main performance metrics used for regression machine learning models:

🔗 R square/Adjusted R square

🔗 Mean absolute error (MAE)

🔗 Mean squared error (MSE) / Root mean square error (RMSE)

**Model Evaluation**

Linear regression and Ridge regression both performed well giving a training and testing accuracy of almost 81% along with R square and adjusted R square values around 0.81. Also MSE,RMSE and MAE values are very low i.e. 0.045,0.212,0.163 respectively.