Bengaluru House Price prediction

-Rajdeep Shil

In this data science project, I tried to create a simple machine learning model to predict real estate prices in Bengaluru. The database used for this project was downloaded from Kaggle (https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data).

Some critical stages of the project:

Wrangling: All the columns with Null data points were dropped. Inconsistency in location, total_sqft and size columns were also handled by writing appropriate functions.

Feature Engineering: Some additional features like 'price_per_sqft' were created that helped me in filtering out outliers.

Dimensionality Reduction: In location many places had only one data point. This could lead to a lot of independent variables for the ML model. Any location with more than 10 data points were clubbed together.

Outlier detection: Outlier detection phase, some outliers were removed just by doing preliminary exploration, some were removed using respective mean and standard deviation.

One hot encoding: One hot encoding was used for managing categorical data which was location for me in this project

K fold Cross validation: Shuffle split with 10 folds with cross_val_score was used to check the performance of our regression model.

GridSearchCV: GridSearchCV was used for choosing the best algorithm amongst lasso, Decision Tress regressor and linear regressor. It is also used for Hyper parameter tunning for the best algorithm.

Major Python Libraries use: sklearn, pandas, numpy , matplotlib , cross_val_score, GridSearchCV

Outcome: Linear regression algorithm gave the best score and hence used for doing actual prediction.

No	model	Best_score	Best_params
0	linear_regression	0.821906	{'normalize': False}
1	lasso	0.703225	{'alpha': 1, 'selection': 'cyclic'}
2	decision_tree	0.755886	{'criterion': 'mse', 'splitter': 'best'}