21AIE205 Python for Machine Learning

1. Problem Definition

This project is done in an effort to predict the real estate prices of a particular region taking several attributes into account. There are three factors that influence the price of a house which includes physical conditions, concepts and location. The current framework includes estimating the real estate prices without any expectations of market prices and cost increment. By breaking down past market patterns and value ranges, and coming advancements, future costs will be anticipated.
The algorithm used to predict the real estate prices is multiple linear regression. This algorithm takes only one independent and one dependent variable into consideration.

2. Datasets

link

Attributes

The dataset has 13 fields.
date - date of publication of the announcement;
time - the time when the ad was published;
geo_lat - Latitude
geo_lon - Longitude
region - Region of Russia. There are 85 subjects in the country in total.
building_type - Facade type. 0 - Other. 1 - Panel. 2 - Monolithic. 3 - Brick. 4 - Blocky. 5 - Wooden
object_type - Apartment type. 1 - Secondary real estate market; 2 - New building;
level - Apartment floor
levels - Number of storeys
rooms - the number of living rooms. If the value is "-1", then it means "studioapartment"
area - the total area of the apartment

Prepare Data

- Preprocessing:- Data was preprocessed with the help of pandas library. The .head() function and .info() function in pandas library were used to take an intital peek into the dataset. Unwanted columns and null values were removed and the data was cleaned, normalized and standardized using MinMaxScaler() and StandardScaler().
- Summarization:- The .describe() function in pandas were used to summarize the dataset and the several types of values present in it. The unique values were found out using .unique() function.
- Visualization:- Several functions from the matplotlib and seaborn libraries were used to give a better understanding of the dataset using scatterplots, pairplots, correlation heatmaps and histogram plots.

3. Learning Algorithms

- The ML algorithms used in this solution are multi linear regression and LASSO regression. Linear regression is a ML algorithm in which a variable or a list of variables are used to predict the value of another variable. The linearity assumption in linear regression means the model is linear in parameters. LASSO regression is a type of linear regression that uses shrinkage. This model reduces unwanted values to null and predicts the value of the output variable using the remaining attributes.
- The dataset was split into training and testing using train_test_split function from sklearn library. The size of the test dataset was set as 30% while that of the train dataset was set as 70%.