National University of Computer and Emerging Sciences

Plot Price Prediction based on Zameen.pk

Group Members

Saad Waseem	19L-1003
Muhammad Hasaan	19L-1011
Laiba Gohar	19L-2367

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

Dataset Overview

The dataset consists of different CSV files based on locations in Lahore that are present in Zameen. pk. The dataset consists of the following significant fields after cleaning.

- Type (Commercial or Residential)
- Price
- Location
- Area
- Creation Date
- Amenities
- Installment Values

Data Gathering

The data was scraped using the Beautiful Soup Library of Python. The target was Lahore hence all the available areas gathered are from Lahore. Link to the plots of every location was saved and opened one by one to scrape the data. The whole process took about 4 days. Threading was used to speed up the process of data gathering.

Data Cleaning and Wrangling

This section explains the data cleaning and wrangling part.

- Unnecessary columns were dropped (Baths, Purpose and Bed)
- Amenities were sorted so that they could be categorically encoded
- The *Punjab, Lahore* part was removed from the location since this is common in all rows
- The price was converted to a standard Rs unit (Lakh and Crore were dropped)
- The area was converted to a standard marla unit
- The datatypes of the price and area were converted to numeric
- Advance payment, monthly instalments and remaining instalments were calculated and added as separate columns
- Instalment value was standardised (Lakh, Crore and Thousand were removed)
- The instalments were merged into a single column
- Amenities were categorically encoded
- Price, area and instalment value were all standardised

Conclusion

In conclusion, data was scrapped from Zameen.pk for the plots located at Lahore. The data in raw form was collected in about 4 days and then it went through a cleaning and wrangling process to make it ready for visualization and ML. The data was cleaned to the best of our knowledge and any future discovery may lead us to change the cleaning/scrapping process. The data in the raw and cleaned form is submitted to make it easy to compare.