Introduction to Data Science

Design Project

Housing Data Analysis: Prediction, Segmentation, and Trends

Submitted by:

Zaeem ul Islam	22L-7452
Ahmed Abdullah	22L-7503

Submitted to:
Mr. Saif ul Islam

26/11/2023

Department of Computer Science

National University of Computer and Emerging Sciences, Lahore

Problem Statement

The increasing trend of property prices in Pakistan, and the complex real estate market are troublesome for buyers and investors. Plus, the most common market scams of overfitting, overselling, and fraudulent marketplaces, all present formidable challenges for the clients.

In the absence of an all-in-one solution, our project aims to implement and deploy a fully web-hosted application that leverages our data knowledge, this application seeks to read property value patterns using historical data and use the data to estimate property value and mortgages as well.

Upon successful implementation and dedication to this project, the web application should be a valuable tool for individuals planning homeownership in Pakistan. This open-source and public application can help obtain accurate estimates and major support in informed decisions.

Dataset Source

The housing data was collected from "Zameen. pk" and "olx.com" using traditional Python script methods. It enabled the collector to capture the information on the residential property transactions, with details like values, size, location, etc.

This dataset is publicly available for all users, and in fact, this data is in completely raw form and no notebook now has been processed on it. The data is also up to date, with the most recent entries processed in 2018.

Dataset Description

This dataset is a comprehensive collection of information and attributes related to property and real estate in Pakistan. It deals with various facets of the residential property market, including property prices, size, locations, and additional features. Gathered from leading and reputed websites with ethical scrapping methods, this dataset provides a diverse and deep understanding of house listings around the different regions of Pakistan.

The dataset is designed to support analyses and generate insights into patterns of the historical market. It is completely in raw form, so all preliminary cleaning processes, to ensure consistency, were completed in our notebook. Additionally, geospatial information and transaction dates explore the fluctuations over time and area. The dataset's purpose is to serve as a valuable resource for real estate marketers and researchers to derive insights from this complex structure of residential property in Pakistan

Attribute Description

Here is the complete description of all attributes and their significance in our model:

- **Date**
    - The date attribute indicates the timestamp of the property transaction.
    - Beneficial for modeling trends, read the timely trends and fluctuations in the market.
- **Price**
    - The price attribute represents the listing price of the property.
    - Undoubtedly valuable for understanding and deriving correlations. It is our base variable or the variable we'll predict.
- **Bedrooms/Bathrooms**
    - The bedrooms/bathroom attribute denotes the number of bedrooms in the property.
    - Note this attribute is in float, representing 1 as a complete bedroom/bathroom and ½ representing a half bedroom/bathroom.
    - Valuable for understanding how it influences its price, contributing to more accurate predictions.
- **Sq. Footage**
    - The sqft_living and sqft_lot represent the total living area and lot size, respectively.
    - One of the essential features for deriving insights based on area and price.
- **Floors**
    - The floor attribute specifies the number of floors in the property.
    - It impacts in determining the property trends, multi-story houses, and pricing.
- **Waterfront**
    - The waterfront attribute is a binary indicator of whether the property has waterfront views.

- **View**
  - The view attribute indicates the level of view from the property.
  - It ranges from 0-4 representing the number of views available in a house.
- **Condition**
  - The condition attribute represents the overall condition of the property.
  - It is relatively discrete in nature and accountable for variations in property maintenance.
- **Grade**
  - The grade attribute signifies the overall grade assigned to the property.
  - It is relatively discrete in nature and accountable for variations in property maintenance.
- **Sq. Footage (Above/Basement)**
  - The sqft_above and sqft_basement indicate the square footage of the interior living space.
  - Valuable for understanding the distribution of living space, contributing to accurate property value predictions.
- **Year Built/Renovated**
  - The yr_built represents the year the property was built, and the yr_renovated indicates the year of the last renovation.
  - This data is left nulled in most places since most of the houses haven't had any renovations in past years.
- **Zip Code**
  - The zip code attribute provides the ZIP code of the property location.
  - This attribute is not such an important variable since we're not dealing with Geospatial Data or visual representation.
- **Longitude/Latitude**
  - These attributes provide geospatial coordinates of the property.
  - This attribute is not such an important variable since we're not dealing with Geospatial Data or visual representation.

Objectives

The project aims to develop a predictive model for property values in Pakistan. Our key objectives for this project include feature engineering, feature selection, picking influential predictors, analyzing patterns and historical trends using data, and ensuring model inter-predictability.

The project also drives a user-friendly web application, allowing users to obtain personalized property estimates with the lowest margin of error. Emphasis will be placed on scalability and efficiency Through these objectives, our project seeks to contribute to Pakistan's real estate market.

References

https://www.zameen.com/

https://www.olx.com.pk/

https://www.kaggle.com/datasets/ahmedembedded/pakistan-houses-pricing-data-web-scrapped

https://github.com/ahmedembeddedx/Mortgage-Predictor

https://www.kaggle.com/datasets/ahmedembedded/usa-mortgage-dataset