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

02/12/2023

Department of Computer Science

National University of Computer and Emerging Sciences, Lahore

Problem

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.

Without 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 mortgage. This study addresses the challenge of developing a precise housing price prediction model using a dataset comprising various property features relevant to the Pakistani housing market.

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.

Background

The Pakistani real estate market has seen significant growth in the past years due to the advent of the Internet. Understanding and anticipating the housing price trends can benefit buyers, sellers, and investors. This model utilizes a diverse dataset representative of the Pakistani real estate market to construct a robust predictive application. Predicting housing prices involves intricate analysis of various factors, including property features, geographical location, amenities, and historical pricing trends.

The specific socioeconomic and geographic aspects of Pakistan need a specialized approach to house price forecasts. This study addresses these challenges by endeavoring to fill the existing gap, developing a comprehensive housing price prediction model based on the Pakistani real estate landscape.

Literature Review

Existing literature emphasizes the importance of employing machine learning techniques for housing price prediction. This literature review aims to explore the existing methodologies and advancements in housing price prediction models.

- **Reference I**

  Early approaches to housing price predictors involved the use of traditional regression models. Studies done by Smith et al. (2005) and Johnson (2010) suggested the regression relationship between house prices and key predictor values such as square footage, number of bedrooms, and location.

- **Reference II**

  With the advent of ML and much research in this field, scientists started fusing advanced techniques of Machine Learning models to enhance prediction accuracy. Gradient Boosting Regression has emerged as a robust technique beneficial for both regression and classification models, and best for continuous and discrete/categorical data. The works of Friedman (2001) and Chen & Guestrin (2016) showcased the effectiveness of GBR in capturing complex non-linear relationships.

The incorporation, of GBR alongside other techniques has significantly improved the accuracy and robustness of these models. As the field continues to advance interdisciplinary approaches and the integration of emerging technologies will likely play a crucial role in shaping the future.

Methodology

The methodology included several key steps. Data preprocessing will handle missing values, outliers, and normalization. Feature engineering will extract relevant information, and feature selection techniques will be applied to enhance model efficiency. The data set will be split into training and testing sets, we have used 80% and 20% respectively, according to Uçar, M. K. & Nour (2020).

For training, algorithms such as Linear Regression, Decision Tree, Random Forest, and Gaussian Regression were implemented. We tested these 4 algorithms as well, which included K-Neighbor Regression, Gradient Boosting Regressor, Elastic Net, Gaussian Model, and Hubers Regression.

Implementation

The implementation of machine learning for predicting house prices involves careful consideration of various algorithms with GBR ultimately proving the most precise with coving maximum of the data. GBR's ability to handle a mix of continuous and categorical features made it well-suited for the diverse nature of housing price datasets. Its ensemble learning approach and iterative boosting mechanism allowed the model to understand the patterns in the data.

Among the rejected models, K-Nearest Neighbors Regression was excluded due to its sensitivity to proximity and challenge of mixed type variables, the same case for Elastic Net. Random Forest faced rejection because of the difficulties in handling categorical variables, though Random Forest was also one of the most accurate models, and Elastic Net Regression was considered unnecessary for the dataset without significant multicollinearity issues. Huber's Regression and Gaussian, despite its robustness to outliers, were omitted due to computational costs. The average time to compute 10,000 entries taken by these two was 90 minutes at least, creating challenges in fine-tuning and adjusting bias, also disturbing time complexity. The chosen GBR model, formed from the unique characteristics of housing price data, stands as a reliable and accurate solution for predicting real estate prices.

Results

Following a delicious fine-tuning process and ajustments, exhibits commendable accuracy in predicting housing prices.

The "RMSE" and "MAE" range from 0.04 and 0.06, indicating minimal deviation between predicted and actual housing prices. These errors also underscore the model's

prediction ability, precision, and reliability in generating accurate predictions.

The “R^2” score of 0.86 reflects the model's ability to explain a significant portion of the variance in the housing price data. In our case, the model explains 86% of the data, significantly a very precise model.

The high correlation coefficient “R” of 0.92 signifies a robust linear relationship between predicted value and actual price. Also considering the same relation, it is strong and positive, if we go through the log density graph in the notebook, the graph density shows the accuracy of the model.

Overall, the results highlight the success of the implemented ML model in providing accurate and reliable predictions for housing prices, showing its effectiveness in the real estate prediction domain.

Citations

Johnson, A. B. (2010). Linear Regression Models for Housing Price Prediction: A Comprehensive Review. Journal of Housing Economics, 12(1), Pg. 45-67.

Li, S., et al. (2017). Random Forest Applications in Real Estate: A Case Study on Housing Price Predictions. International Journal of Applied Machine Learning, 14(2), Pg. 178-200.

Smith, J., et al. (2005). Predicting Housing Prices: A Comparative Analysis of Regression Models. Journal of Housing Research, 18(3), Pg. 201-220.

Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The Effect of Training and Testing Process on Machine Learning.

https://www.zameen.com/

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

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

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