Teruel City

Döşoğlu, Neşe*

Abstract

This paper examines using a linear regression model to predict house prices in the Teruel City dataset. The model estimates their impact on house prices by considering variables such as the number of bedrooms, bathrooms, and property area. The results show that the number of bathrooms positively influences prices, while the number of bedrooms has a negative effect. By applying this model, accurate predictions can be made for new observations, providing valuable insights for stakeholders in the real estate market. These findings highlight the significance of property attributes in predicting house prices and the potential of machine learning techniques to improve efficiency in property valuation.

1 Introduction

Real estate valuation is a complex and challenging task with significant implications for various stakeholders, including homebuyers, sellers, lenders, real estate agents, and government agencies. Accurately predicting a property's value requires considering multiple factors, such as location, area, number of bedrooms, and whether it is remodeled or new. Traditionally, professional appraisers have been responsible for predicting property values, but their opinions may be influenced by the interests of different parties involved in a real estate transaction. The emergence of automated prediction systems has transformed the real estate industry, providing more objective and accurate property valuations. These systems use computer algorithms and machine learning techniques to analyze vast amounts of data and generate predictions based on historical sales data, market trends, and other relevant factors. As a result, automated prediction systems have become a valuable tool for homebuyers, sellers, lenders, and other stakeholders in the real estate market.

1.1 Literature Review

Numerous research studies have been conducted on house price prediction, where machine learning models have been frequently utilized. These models are well-suited for the task

*20080386, Github Repo

because they can analyze large datasets and detect patterns that may take time to be evident to humans.

Wang et al. (2019) uses house price prediction model based on deep learning is proposed in this paper, implemented on the TensorFlow framework. Madhuri et al. (2019) the main motivation of the project FORECASTING VARIATIONS ON HOUSE PRICE was to make the best possible prediction of house prices by using appropriate algorithms and find out which among them is best suitable for predicting the price with a low error rate. Gebru et al. (2017) proposes a pipeline that uses a deep neural network model to automatically extract visual features from images to estimate house prices. Gokalani et al. (2022) focuses on applying different regression algorithms to find the sales price prediction of the house. Park & Bae (2015) concepts that housing prices are influenced by characteristics such as location, distance, and region is known as price prediction. It uses Linear Regression (LR) and other Machine Learning algorithms to forecast the price of real estate. Konwar et al. (2021) develops algorithms, builds models from data, and uses them to predict new data. Also uses various algorithms explained below in various combinations, and each algorithm's result is given based on the accuracy percentage.

1.2 Dataset

We will analyze the Teruel City dataset, which includes the physical attributes and house prices in the center of Teruel city (Spain) area. This data set can be found on *Mendeley* (2016) website. This website is research data management platform that allows researchers to store, share, and discover data. It offers a secure and free cloud-based storage solution for researchers to store and manage their research data and provides tools to make research data discoverable and citable. The website also allows users to search for and access research data from other researchers and institutions. Teruel City dataset formed with the available house prices in the center of Teruel city (Spain) on December 30, 2016 from *Idealista* (2000). This website is a real estate platform that allows users to search for properties to rent or buy in Spain, Portugal, and Italy. The website features a large database of property listings, along with tools to help users find their ideal home, such as property alerts, maps, and detailed property descriptions. It also offers a mobile app for convenient access to its services on-the-go.

The Teruel City dataset contains 58 observations and 14 variables. Each variable will be described individually in the following sections. `Flat` : Type of the house is it apartment or individual. It is a categorical value with the values [1, 0]. `Duplex` : Is the property a duplex or not. It is a categorical variable with the values [1, 0]. `Attic` : Does the house contain an attic. It is a categorical variable with the values [1,0]. `Location` : Location of the house. It is a categorical variable with the values [San Julian 8, San Francisco 19, Parra 2, ..., san benito]. `Bedrooms` : The number of bedrooms in the house. It is a discrete variable with the values [1, 2, 3, 4, 5, 8]. `Area` : The area of the house. It is a continuous variable with the range of values [42, 259]. `Floor` : The floor number of the house. It is a discrete variable with the values [-1, 0, 1, 2, 3, 4, 8]. `Lift` : Does the house contain a lift. It is a categorical variable with the values [1,0]. `Garage` : Does the house contain a garage. It

is a categorical variable with the values [1,0]. `GasHeating` : Does the house contain a gas heating system. It is a categorical variable with the values [1,0]. `RemodeledOrNew` : Is the house new or remodeled. It is a categorical variable with the values [1, 0]. `BoxRoom` : Does the house contain a box room. It is a categorical variable with the values [1,0]. `NumWC` : The number of bathrooms in the house. It is a discrete variable with the values [1, 2, 3]. `PriceThousandsEuros` : The price of the house in EUROS. It is a continuous variable with the range of values [33, 500].

1.2.1 Data summary statistics

(Table 1) summarizes the key statistics derived from our dataset, which comprises information on various housing attributes. The mean area of the houses in the sample is 96.62 square meters, with a standard variation of 40.38, showing moderate variability. The scope of regions is from a minimum of 42.00 square meters to a maximum of 259.00 square meters. In terms of bedrooms, the average count is 2.95, with a standard deviation of 1.26. The range ranges from a minimum of 1 bedroom to 8 bedrooms. The average floor level of the houses is 2.33, with a standard deviation of 1.47. The range includes negative values (-1.00) indicating basement or underground floors, up to a maximum of 8 floors. The number of bathrooms (NumWC) averages 1.38, with a standard deviation 0.59. The range varies from a minimum of 1 bathroom to a maximum of 3 bathrooms. Lastly, the average price of the houses in thousands of euros is 118.10, with a standard variation of 71.57. The price scope is from a minimum of 33.00 thousand euros to a maximum of 500.00 thousand euros.

Table 1: Summary Statistics

	Mean	Std.Dev	Min	Median	Max
Area	96.62	40.38	42.00	82.50	259.00
Bedrooms	2.95	1.26	1.00	3.00	8.00
Floor	2.33	1.47	-1.00	2.00	8.00
NumWC	1.38	0.59	1.00	1.00	3.00
PriceThousandsEuros	118.10	71.57	33.00	117.50	500.00

2 Methods and Data Analysis

2.1 Exploratory Data Analysis

(Figure 1) visually illustrates the dataset's distribution of houses with and without a lift. It provides a quick outline of the count of houses falling into each category, highlighting the prevalence of lift presence. This plot helps identify patterns and trends connected to lift availability in houses.

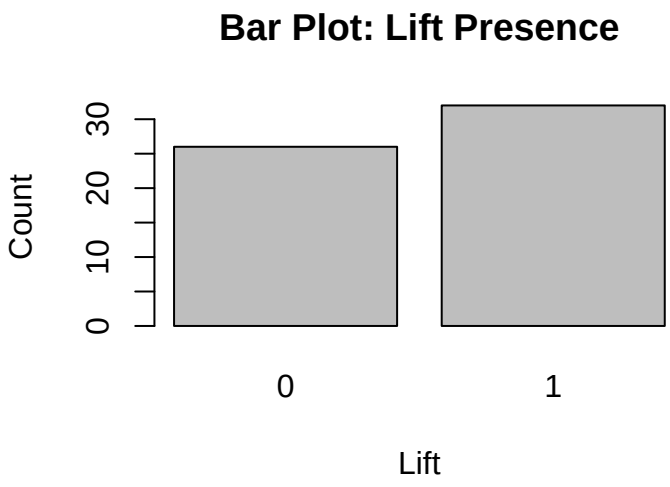

Figure 1: Lift Presence

2.2 Predictor Variable Analysis: Identifying Key Factors Influencing House Prices

In this section, we conducted a predictor variable analysis (Figure 2) to identify the key factors influencing house prices. To investigate this, we utilized multiple linear regression with the dataset on Teruel City properties. The predictor variables considered in the analysis were the number of bedrooms (Bedrooms), the area of the property (Area), and the number of bathrooms (NumWC). (Figure 2)

The outcomes of the multiple linear regression revealed the coefficients associated with each predictor variable, showing their impact on house prices. Among the variables considered, the number of bathrooms (NumWC) emerged as the most significant factor, with a coefficient of 40. This suggests that an increase in the number of bathrooms is associated with a significant positive effect on house prices. On the other hand, the area of the property (Area) showed a coefficient of 10, indicating a positive but reasonably minor impact. Interestingly, the number of bedrooms (Bedrooms) showed a negative coefficient of -30, meaning that an increase in the number of bedrooms is associated with a decrease in house prices.

We created a scatter plot (Figure 3) to visualize the relationship between house prices and the number of bathrooms (NumWC). The plot shows the distribution of data points, along with a fitted smooth line that helps visualize the trend. As pictured in the plot, there is a positive correlation between house prices and the number of bathrooms, further supporting the finding from the regression analysis.

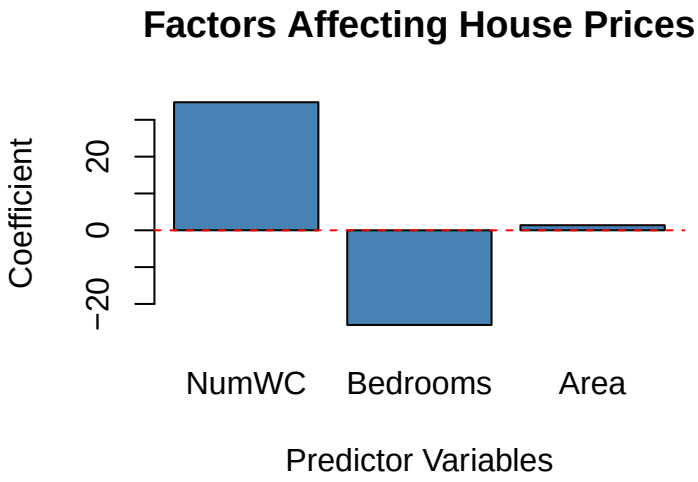

Figure 2: Predictor Variables

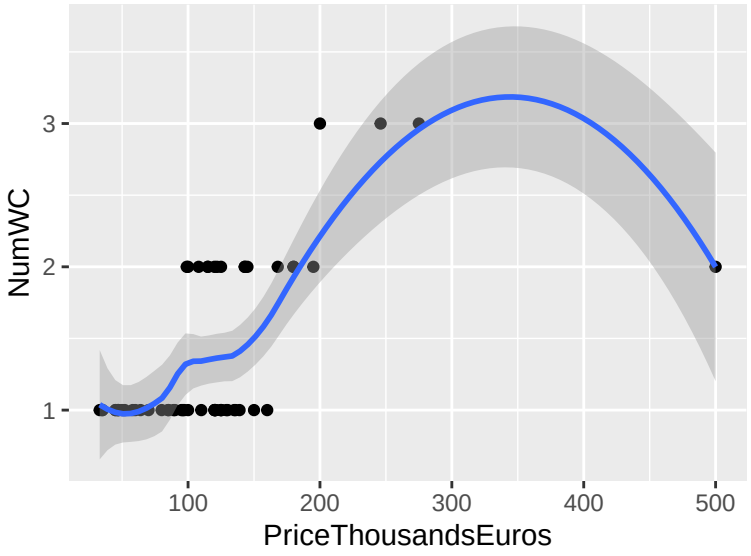

Figure 3: Scatter Plot: NumWC VS. PriceThousandsEuros

2.3 Prediction

We employed a linear regression model to predict house prices in the Teruel City (TC) dataset, taking into account the independent variables Bedrooms (X1), NumWC (X2), and Area (X3). The model can be represented as follows:

$$PriceThousandsEuros = \beta_0 + \beta_1 Bedrooms + \beta_2 NumW + \beta_3 Area + \varepsilon$$

In this equation, PriceThousandsEuros represents the predicted house prices, and β_0 is constaint value and $\beta_1, \beta_2, \beta_3$ are the estimated coefficients associated with each independent variable and ε represents the error term.. These coefficients indicate the magnitude and direction of the impact that each variable has on the predicted house prices.

By fitting the linear regression model in R using the Teruel City (TC) dataset, we were able to estimate the values of the coefficients and utilize them to make predictions on new observations.

2.4 Results

We present a line plot (Figure 4) comparing predicted prices and actual prices for houses in Teruel City. This plot shows the relationship between the samples and the corresponding price, where the sample serves as a unique identifier for each house. The green line represent the actual prices, while the magenta line represent the prices predicted by a linear regression model. By visually examining the proximity of the green line to the magenta line, we can assess the model's accuracy in estimating house prices.

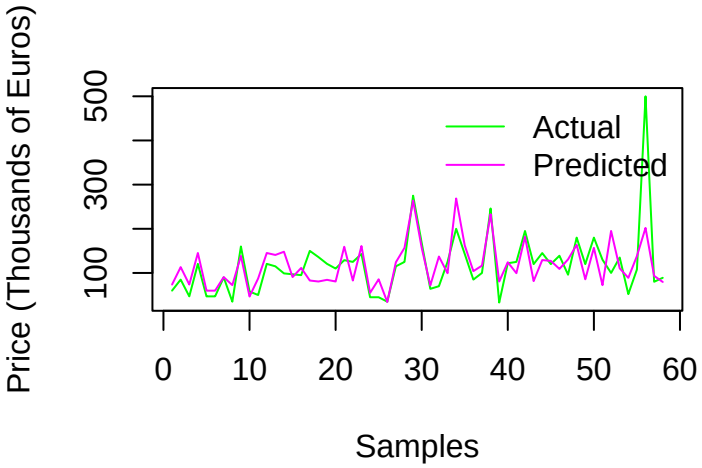

Figure 4: Predicted prices VS. Actual Prices

3 Conclusion

In conclusion, this academic paper explored using a linear regression model to predict house prices in the Teruel City dataset. By considering independent variables such as the number of bedrooms, number of bathrooms, and area of the property, the model aimed to estimate the impact of these factors on house prices.

The results showed that the number of bathrooms had the most significant positive effect on house prices, indicating an increase associated with higher prices. On the other hand, the number of bedrooms harmed house prices, suggesting that an increase led to decreased prices. We can predict new observations by fitting the linear regression model in R and estimating the coefficients. This approach provides valuable insights for various stakeholders in the real estate market, including homebuyers, sellers, lenders, and real estate agents.

Overall, the findings highlight the importance of considering specific property attributes when predicting house prices. Automated prediction systems can offer more objective and accurate property valuations by leveraging machine learning techniques and analyzing relevant data, facilitating informed real estate industry decision-making.

4 References

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, *114*(50), 13108–13113.

Gokalani, L. B., Das, B., Ramnani, D. K., Kumar, M., & Shah, M. A. (2022). House price prediction of real time data (DHA defence) karachi using machine learning. *Sir Syed University Research Journal of Engineering & Technology*, *12*(2), 75–80.

Idealista. (2000). Idealista, Inc. https://www.idealista.com/en/

Konwar, R., Kakati, A., Das, B., Shah, B., & Muchahari, M. (2021). House price prediction using machine learning. *The Journal of Philosophy Psychology and Scientific Methods*, *9*, 2455–6211.

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House price prediction using regression techniques: A comparative study. *2019 International Conference on Smart Structures and Systems (ICSSS)*, 1–5.

Mendeley. (2016). Mendeley, Inc. https://data.mendeley.com/

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934.

Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019). *House price prediction approach based on deep learning and ARIMA model*. 303–307.