Duke Forest House Price Prediction

A.A.Barakat, Hamza Aslan*

1 Introduction

When individuals are in the market for a new home, they typically seek out properties that are reasonably priced and possess the desired features they are looking for. A variety of factors contribute to a property's value, including its location, number of bedrooms, appearance, and size. While professional appraisers have traditionally been responsible for predicting property values, their opinions may be influenced by the interests of the lender, mortgage broker, buyer, or seller. As a result, an automated prediction system serves as an unbiased third-party source of information.

Access to accurate property value predictions is crucial for central banks, financial supervision authorities, investors, and homeowners. A house price prediction system can aid in their decision-making process by helping them determine whether the property is worth the asking price. For those looking to sell their homes, a price prediction system can assist them in identifying which features they should add to their property to increase its value.

1.1 Literature Review

Given the significance of the housing market in the global economy, there has been extensive research on the development of house price prediction models, with a particular focus on using machine learning and statistical techniques to accurately forecast house prices.

H. Ahmed & N. Moustafa (2016) propose using visual and textual features to predict the house price. Then, these extracted features are passed through a neural network model. Through experiments the visual features increased the performance. Different 11 machine learning regression models are used by Park & Bae (2015) and evaluated to determine the best model for a better house price prediction of the capital Islamabad. Alfiyatin et al. (2017) rely on three factors: physical conditions, concept, and location to predict the price of a house. Linear regression and particle swarm optimization methods are leveraged to predict the house price

Yang et al. (2018) build an array of machine learning models to predict the price of a product given its image. Additionally, this work visualizes the features that result in higher

*21080637, Github Repo

or lower price predictions. For price regression, they use linear regression on histogram of oriented gradients, convolutional neural network features. Kumar et al. (2021) compare different algorithms, such as linear regression, support vector regressor, XGBoost regressor, and CatBoost regressor to predict the house prices. CatBoost Regressor has the highest accuracy by 90%. Recently, Mysore et al. (2022) use the Decision Tree machine learning algorithm to predict house prices in Mumbai city, where additional features, such as air quality and crime rate are used to help in predicting the prices.

1.2 Dataset

We will utilize the Duke Forest (DK) dataset, which includes the physical characteristics and sale prices of houses in the Duke Forest area. This dataset is publicly accessible on the OpenIntro (2021) website. This website provides free and open-source statistics and data science educational materials, including textbooks, videos, and online courses. The organization aims to make high-quality educational resources accessible to all students and promote statistical transparency. Their materials are widely used in academic courses and by self-learners.

The DK dataset comprises data on houses that were sold in November 2020, within the Duke Forest community located in Durham, North Carolina (NC). It is collected from Zillow (2021) which is a popular online real estate marketplace that provides information on homes for sale, rent, and recently sold properties in the United States. It offers users various tools to search for and estimate the value of a home, as well as resources for home buyers and sellers. The website is widely used by those interested in buying, selling, or investing in real estate.

The DK dataset consists of 98 observations and 13 variables. In the following, we will describe each of the variables separately.

- `address` : Address of house. Contains the address of the house. It is a mixed variable that contains house number, street name, city, state, and postal code. House number is a continuous variable with a range of [1, 3009]. Street name is a categorical variable with the possible values of [Learned, Pinecrest, Wrightwood, …, Sevier]. City is a categorical variable with only one value [Durham]. State is a categorical variable with only one value [NC]. Postal is a discrete variable with only one value [27705].

- `price` : Sale price, in USD. Contains the sale price of the house. It is a continuous variable with a range of [95000, 1520000].

- `bed` : Number of bedrooms. Contains the number of bedrooms in the house. It is a discrete variable with the values [2, 3, 4, 5, 6].

- `bath` : Number of bathrooms. Contains the number of bathrooms in the house. Contains the number of bathrooms in the house. It is a discrete variable with the values [1, 2, 2.5, 3, 4, 5, 6].

- `area` : Area of home, in square feet. Contains the area of the house. It is a continuous variable with the range of [1094, 6178].
- `type` : Type of home (all are Single Family). It is a categorical variable with only one value.
- `year_built` : Year the home was built. Contains the year that the property was built. It is a discrete value with the range of [1923, 2020].
- `heating` : Heating system. Contains the type of the heating system that the property contains. It is a categorical with the values [forced air, electric, gas, heat pump, baseboard, other].
- `cooling` : Cooling system (other or central). Contains the type of the cooling system that the property contains. It is a categorical value with the values [central, other].
- `parking` : Type of parking available and number of parking spaces. Contains the type of parking space. It is a caterogical value with the values [0 space, garage, attached, off-street, covered, carport, garage-detached, other].
- `lot` : Area of the entire property, in acres. It is a continuous variable with the range of [0.15, 1.47].
- `hoa` : If the home belongs to an Home Owners Association, the associated fee (NA otherwise).
- `URL` : URL of the listing. It is categorical value.

2 References

Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, east java, indonesia. *International Journal of Advanced Computer Science and Applications*, *8*(10).

H. Ahmed, E., & N. Moustafa, M. (2016). House price estimation from visual and textual features. *8th International Conference on Neural Computation Theory and Applications*, 62–68.

Kumar, G. K., Rani, D. M., Koppula, N., & Ashraf, S. (2021). Prediction of house price using machine learning algorithms. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1268–1271.

Mysore, S., Muthineni, A., Nandikandi, V., & Behera, S. (2022). Prediction of house prices using machine learning. *Int. J. Res. Appl. Sci. Eng. Technol*, *10*(6), 1780–1785.

OpenIntro. (2021). *OpenIntro*. https://www.openintro.org/

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.

Yang, R. R., Chen, S., & Chou, E. (2018). AI blue book: Vehicle price prediction using visual features. *arXiv Preprint arXiv:1803.11227*.

Zillow. (2021). *Zillow*. https://www.zillow.com/