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 algorithms.

H. Ahmed & N. Moustafa (2016) proposes a novel approach for automatic house price estimation by combining visual features extracted from house photographs with textual information, which outperformed existing models that rely solely on textual data. Park & Bae (2015) develops a house price prediction model using machine learning algorithms, based on the analysis of housing data of 5359 townhouses in Fairfax County, Virginia, and compares the classification accuracy performance of C4.5, RIPPER, Naïve Bayesian, and AdaBoost algorithms, with RIPPER consistently outperforming the other models in housing price prediction. The study by Alfiyatin et al. (2017) aims to predict house prices in Malang city based on NJOP houses using regression analysis and particle swarm optimization (PSO), which resulted in a suitable combination that achieved a minimum prediction error of 14.19.

*21080637, Github Repo

Varma et al. (2018)'s study aims to accurately predict housing prices using various regression techniques and real factors, incorporating real-time neighborhood details from Google maps. The price of a house depends on both its features and the desirability of its location, which is difficult to measure directly. To tackle this, Chopra et al. (2007) used a method that combines a model that predicts the intrinsic price of a house based on its description, with a model that captures the desirability of the location. The two models are trained together resulting in more accurate predictions. More recently, Bency et al. (2017) proposed a Convolutional Neural Network (CNN) framework to model geo-spatial data and learn spatial correlations automatically for housing prices prediction, achieving a 57% improvement on top of a baseline model without satellite images.

1.2 Dataset

We will utilize the Duke Forest (DF) dataset, which includes the physical characteristics and sale prices of houses in the Duke Forest area. This dataset is publicly accessible on the *OpenIntro* (2009) 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 DF 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* (2006) which is a popular online real estate marketplace that provides information on houses 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 DF 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. 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. It is a categorical variable with only one value [Single Family].
- `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. 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. 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.
- `URL` : URL of the listing within the *Zillow* (2006)

1.2.1 Data tidying

The DF dataset is cleaned and tidied using various functions in R. The initial dataset was cleaned by removing the `type`, `hoa`, and `url` columns as they contained the same value for all the observations. The `address` column was split into separate columns (`area`, `city`, and `state`) because it contained multiple variables. The `city` and `state` columns were then removed as they contained the same value for all the observations. The `area` column was further split into separate columns (`number`, `street`, and `type`), and the `type` column was removed as it contained the same value for all the observations. The `heating` and `parking` columns were processed by creating a new column for each unique heating and parking type, respectively, with an initial value of "no". The unique list of heating and parking types was obtained, and for each observation, "yes" was put for each heating and parking type that exists in that observation. The `heating` and `parking` columns were then removed, and the column names were cleaned up by removing blanks and adding `heating_` and `parking_` prefixes to the names of the columns related to heating and parking, respectively. The rows containing missing values were dropped, and the resulting dataset was saved as a new csv file ("duke_forest_tidy.csv"). The final dataset contained information on the number of bedrooms, bathrooms, square footage, rental price, heating (gas, electric, forced air, heat pump, and other) and parking (carport, covered, garage, garage attached, garage detached, off-street, and on-street).

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).

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. (2017). Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Chopra, S., Thampy, T., Leahy, J., Caplin, A., & LeCun, Y. (2007). Discovering the hidden structure of house prices with a non-parametric latent manifold model. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 173–182.

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.

OpenIntro. (2009). OpenIntro, Inc. 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.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

Zillow. (2006). Zillow Group, Inc. https://www.zillow.com/