Duke Forest House Price Prediction

A.A.Barakat, Hamza Aslan*

Abstract

This paper explores house price prediction using the Duke Forest (DF) dataset obtained from Zillow. The dataset includes physical characteristics and sale prices of houses. After data cleaning and analysis, linear regression is applied, revealing strong correlations between the number of bathrooms, lot size, and house prices. The model achieves an average prediction error of $201,129.21 (RMSE). The study demonstrates the feasibility of using linear regression for house price prediction and suggests further research on integrating visual features for improved accuracy.

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.

*21080637, Github Repo

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. 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).

1.2.2 Data statistics

Our analysis (Table 1) of the DF dataset revealed key insights. The average number of bathrooms was 3.10, with a standard deviation of 0.93. Bedrooms averaged at 3.74, with a standard deviation of 0.75. Lot sizes had an average of 0.57, with a standard deviation of 0.22. The average house price was $560,417.01, with a standard deviation of $226,560.52. Houses in the DF dataset were built between 1923 and 2020, with an average construction year of 1966.52 and a standard deviation of 17.70. These findings provide a comprehensive understanding of the DF dataset's attributes.

Table 1: Summary Statistics

	Mean	Std.Dev	Min	Median	Max
bath	3.10	0.93	1.00	3.00	6.00
bed	3.74	0.75	2.00	4.00	6.00
lot	0.57	0.22	0.15	0.55	1.47
price	560417.01	226560.52	95000.00	540000.00	1520000.00
year_built	1966.52	17.70	1923.00	1962.00	2020.00

According to Table 2, the most expensive house in the dataset has a price of $1,520,000, with 3 bedrooms and 4.0 bathrooms. It is located on Learned Street and occupies a lot size of 0.97. On the other hand, the cheapest house in the dataset is priced at $95,000, featuring 4 bedrooms and 4.5 bathrooms. It is situated on Mcdowell Street and has a lot size of 0.63.

Table 2: Most expensive and cheapest houses

Type	price	bed	bath	street	lot
Most Expensive	$1520000	3	4.0	Learned	0.97
Cheapest	$95000	4	4.5	Mcdowell	0.63

2 Method

2.1 Feature selection

We conducted a correlation analysis (Figure 1) to identify the most predictive features for the house prices. The correlation matrix analysis revealed that several columns in the DF dataset exhibit a noteworthy relationship with the house prices. Notably, the variables "bath," "lot," and "bed," displayed the highest correlations with the price. The positive correlation coefficients indicated that an increase in the values of these features corresponded to higher house prices.

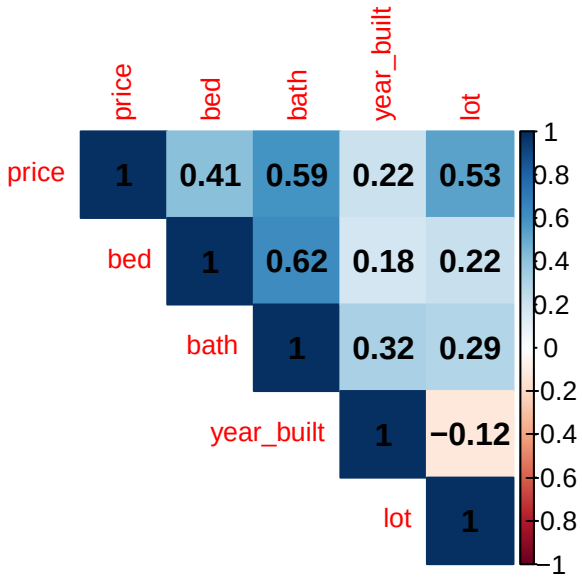

Figure 1: Correlation matrix

The "bath" variable, representing the number of bathrooms, was found to be strongly associated with price, suggesting that houses with more bathrooms tend to command higher prices (Figure 2). Similarly, the "lot" variable, representing the size of the property, exhibited a significant positive correlation with price, indicating that larger lots were often associated with higher prices (Figure 3).

2.2 Price prediction

We used linear regression to predict the house prices in the DF dataset. Linear regression is a widely-used statistical technique that aims to establish a linear relationship between the independent variables (features) and the dependent variable (house prices). Linear regression is represented by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

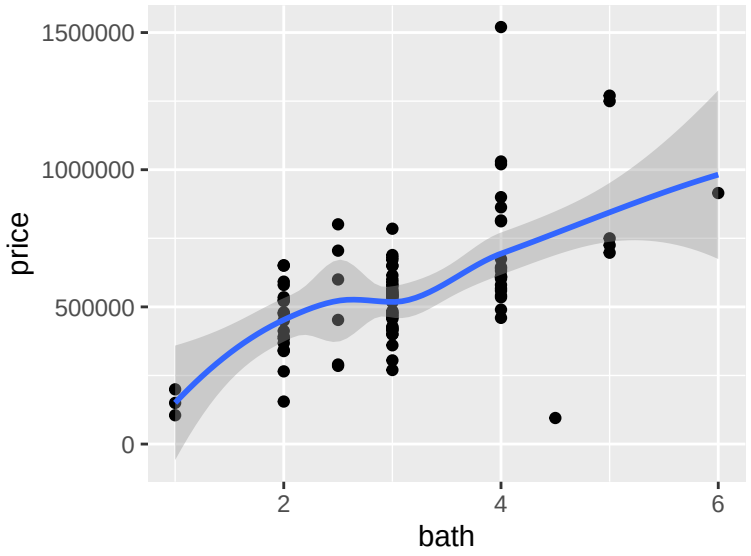

Figure 2: Scatter plot of house prices against number of bathrooms

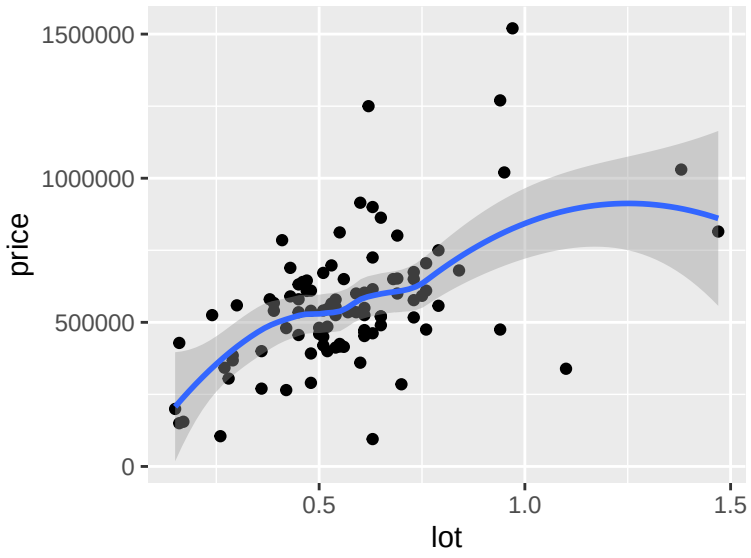

Figure 3: Scatter plot of house prices against lot area

where Y denotes the predicted house prices, β_0 represents the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients associated with the independent variables $X_1, X_2, \ldots, X_n$, respectively, and ε represents the error term. The objective of linear regression is to estimate the coefficients that best fit the given data, allowing us to make predictions on new observations.

2.3 Evaluation

The DF dataset consists of 98 samples, each representing a distinct house. To assess the performance and generalization ability of our predictive model, we split the dataset into a training set and a test set. The training set, comprising a 80% of the samples, was used to train the model on the underlying patterns present in the data. This process allowed the model to learn and optimize its parameters based on the training samples. The remaining portion of the dataset, referred to as the test set, was held back and not used during the model training phase. Instead, it served as an independent dataset for evaluating the model's performance and assessing its ability to predict house prices on unseen data. Splitting the dataset into train and test sets enables us to estimate how well our model will perform on new, unseen houses to ensure its reliability in real-world scenarios.

2.4 Results

The root mean squared error (RMSE) obtained on the test set was \$201129.21. This metric provides an assessment of the average prediction error of our model when applied to the unseen data in the test set. A lower RMSE value indicates better predictive performance, as it suggests that the model's predictions are closer to the actual house prices. In our case, the RMSE of \$201129.21 signifies that, on average, our model's predictions deviated from the true house prices by approximately \$201129.21.

We plotted a scatter plot (Figure 4) to visually compare the actual prices of houses with the predicted prices generated by our model. The scatter plot provided a clear visualization of the relationship between the two sets of prices. The actual prices were represented by blue dots, while the predicted prices were denoted by red dots. Ideally, the predicted prices would fall along a straight line, indicating a perfect match with the actual prices. Any deviations from this line indicated differences between the predicted and actual prices. This scatter plot served as a valuable tool for evaluating the performance of our model in accurately predicting house prices. The proximity of the red dots to the blue dots revealed the model's ability to capture and estimate the underlying patterns and trends in the data. Overall, the scatter plot helped to assess the effectiveness of our predictive model and provided insights into the degree of accuracy achieved in predicting house prices within our dataset.

3 Conclusion

Our results demonstrated a significant relationship between various features and house prices, allowing us to develop a predictive model. We began by performing a correlation analysis,

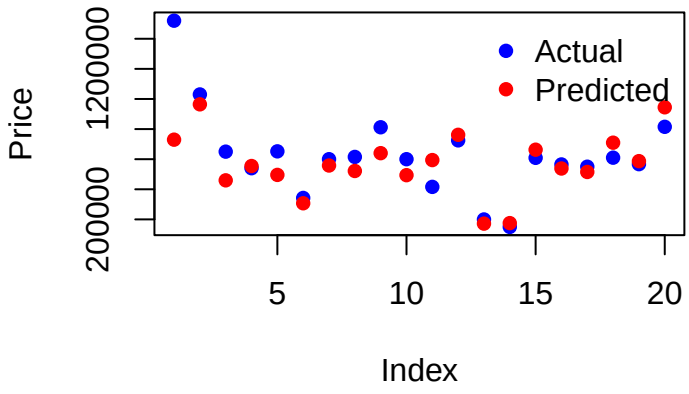

Figure 4: Scatter plot of actual and predicted prices

which revealed that variables “bath”, “lot”, and “bed” exhibited the strongest correlations with house prices. Based on these findings, we employed linear regression to predict house prices. The model achieved an RMSE of $201129.21 on the test set, indicating the average prediction error of the model. This performance suggests that our model can reasonably estimate house prices, although there is room for improvement.

To enhance future analyses, integrating photos and visual features using convolutional neural networks (CNNs) could yield more accurate predictions. Integrating visual features with CNNs could lead to more precise predictions and a deeper understanding of the visual aspects influencing house prices.

4 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/