Duke Forest House Price Prediction

A.A.Barakat, Hamza Aslan[*]

1 Summary

The academic paper investigates the development of a house price prediction system using Duke Forest (DF) dataset obtained from Zillow. The dataset comprises information on the physical characteristics and sale prices of houses in the Duke Forest community in Durham, North Carolina. The research aims to develop a model that predicts house prices based on various features.

The paper begins with a literature review, highlighting previous studies that have utilized computational algorithms for house price prediction. These studies have explored different factors, such as visual features, textual information, and neighborhood details. The review emphasizes the importance of property value predictions in decision-making processes for investors and homeowners.

The DF dataset consists of 98 observations and 13 variables, including address, price, number of bedrooms and bathrooms, area, type of home, year built, heating and cooling systems, parking availability, lot size, and URL of the listing. The dataset undergoes data tidying to remove redundant and unnecessary columns, split address information into separate columns, and process categorical variables related to heating and parking.

A correlation analysis is conducted to identify the most predictive features for house prices. The study reveals that the number of bathrooms, lot size, and number of bedrooms exhibit the highest correlations with house prices. Houses with more bathrooms and larger lots tend to command higher prices.

The paper then employs linear regression as the methodology for predicting house prices. Linear regression is a statistical technique that establishes a linear relationship between independent variables (features) and the dependent variable (house prices). The model is trained on a subset of the dataset (80% as the training set) to learn the underlying patterns and optimize its parameters. The remaining portion of the dataset (20% as the test set) is used to evaluate the model's performance and assess its ability to predict house prices on unseen data.

The study results demonstrate that the linear regression model achieves an average prediction error, measured by the root mean squared error (RMSE), of $201,129.21 on the test set. This

[*] 21080637, Github Repo

indicates that, on average, the model's predictions deviate from the actual house prices by approximately $201,129.21. A scatter plot is used to visually compare the actual house prices with the predicted fees generated by the model, providing insights into the model's accuracy.

In conclusion, the research develops a house price prediction system using the DF dataset. The study highlights the strong correlations between the number of bathrooms, lot size, and house prices. The linear regression model demonstrates the feasibility of predicting house prices based on these features. Further research suggests incorporating visual elements using convolutional neural networks (CNNs) for more accurate predictions and a deeper understanding of the visual aspects influencing house prices.