Housing Price Prediction Data Analysis using Decision Trees, Gradient Boosting and K-Nearest Neighbors

Student: Turcu Ciprian-Stelian

Master: High Performance Computing and Big Data Analytics

E-mail: ciprian.turcu@stud.ubbcluj.ro

The present research work is on predictive modelling techniques for housing price estimation using Decision Trees, Gradient Boosting, and K-Nearest Neighbors (KNN). The dataset contains 545 properties with 13 numeric and categorical features, including property area, bedrooms, and furnishing status. The analysis tries to identify the best algorithm for the prediction of housing prices while discussing strengths and limitations of each approach.

Results show that Gradient Boosting Regression is the best model, as it explains 66.5% of the variance in housing prices (R-squared = 0.665). Its ensemble-based approach captured complex patterns in the data but at higher computational costs. K-Nearest Neighbors performed moderately well, explaining 61.3% of the variance (R-squared = 0.613). However, its accuracy was sensitive to the choice of hyperparameters and scaling of input features. Decision Trees showed less predictive power, explaining 46.6% of the variance (R-squared = 0.465). The model was very easy to interpret, but it had the tendency either to overfit or to underfit the data.

The methodology comprised data preprocessing to handle the missing values properly and encoded categorical features in a compatible format with regression models. Numerical features were analysed in terms of their distribution and their correlation to the target variable—house price. Model evaluation is carried out based on the Mean Squared Error and the R-squared scores to ensure that the model is accurate and robust.

The following strategies can be followed to improve the performance of the model: addition of more interaction terms and polynomial features in order to capture non-linear relationships; state-of-the-art techniques such as XGBoost or LightGBM should be tried in order to handle complex patterns and interactions; increase the dataset to include more diverse entries for better generalization and model robustness; and finally, multicollinearity among features can be looked upon for better interpretation of the model and less redundancy.

In conclusion, Gradient Boosting Regression outperformed other models in this study and showed its appropriateness for housing price prediction. However, the analysis brought forth the importance of feature engineering, hyperparameter tuning, and careful algorithm selection to achieve optimal results. Future work could involve leveraging advanced algorithms and incorporating additional features to further enhance prediction accuracy. This study exemplifies the ability of machine learning techniques in the real estate analytics domain and is very important for practitioners and researchers aiming at improving house price prediction.