House Prices Prediction

Abstract

Background: Understanding factors influencing house prices is crucial in real estate markets. This study aims to identify key determinants affecting house selling prices.

Objectives: Investigate the significance of lot area, living space, house quality, construction year, basement exposure, heating and air conditioning types, fireplaces, garage capacity, wooden floors, and porch size on house prices.

Methods: Employed EDA and machine learning algorithms on housing dataset. Developed regression models to quantify relationships between variables and prices.

Results: Found lot area, living space, quality, construction year, basement exposure, heating, air conditioning, fireplaces, garage capacity, wooden floors, and porch size as significant factors, explaining 87% to 89.6% of price variability.

Conclusion: Insights aid real estate stakeholders in making informed decisions regarding property investments and transactions.

Data

The data used in this research paper is sourced from the "House Prices - Advanced Regression Techniques" dataset, which is available on Kaggle. The dataset can be accessed through the following link: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data . This dataset contains a comprehensive collection of housing features, including lot area, living space, quality, construction year, basement exposure, heating and air conditioning types, number of fireplaces, garage capacity, wooden floor area, porch size, and other relevant variables.

Methodology & Results

Result 1: Initially, missing values were removed from the dataset to ensure data integrity. Following this, Ordinary Least Squares (OLS) regression was applied to eliminate any variables that were not statistically significant at a 95% confidence level (p-value < 0.050). Heteroskedasticity was detected in the data, prompting the use of Weighted Least Squares (WLS) and Heteroskedasticity-Robust Standard Errors Regression to address this issue (Breusch-Pagan test statistic = 0.561276, p-value = 0.0000).

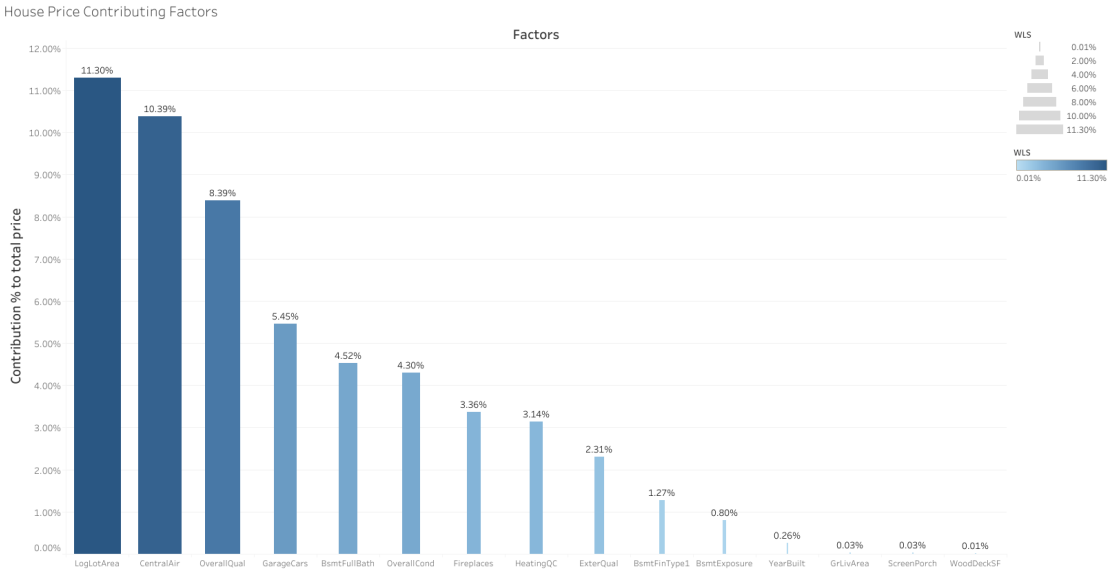
Result 2: WLS and Heteroskedasticity-Robust Standard Errors Regression techniques were implemented, revealing significance for lot area and the total living space, the quality and condition of the house, the construction year, the exposure of the basement, the heating type and air conditioning, the number of fireplaces, the number of cars fitted in the garage, the surface of wooden floors and the surface of porch. The analytical results are summarized in Table 1 and depicted graphically in Graph 1, showcasing the relationships between the significant variables and house prices.

Table 1. Regressions Results

variables	WLS	HRSE
const	4.10425528047157*** (0.000)	4.2116220985934305*** (0.000)
OverallQual	0.08391099516710529*** (0.000)	0.08600243435714322*** (0.000)
OverallCond	0.04295511588641006*** (0.000)	0.0445974691551288*** (0.000)
YearBuilt	0.0026462719031549696*** (0.000)	0.002628557335254365*** (0.000)
ExterQual	0.023088992152172638** (0.038)	0.04981288100958248*** (0.000)
BsmtExposure	0.008005121817776434* (0.065)	0.008897782468292387** (0.049)
BsmtFinType1	0.012678688052116749*** (0.000)	0.010615731152267285*** (0.000)
HeatingQC	0.03142318389041521*** (0.000)	0.021619930276594534*** (0.000)
CentralAir	0.10386158561377642*** (0.000)	0.07851657529009012*** (0.001)
GrLivArea	0.0002696866846166482*** (0.000)	0.0002139344049373754*** (0.000)
BsmtFullBath	0.045219489053571316*** (0.000)	0.04171840999971965*** (0.000)
Fireplaces	0.033613878165606625*** (0.000)	0.04004091058756176*** (0.000)
GarageCars	0.05453563090663116*** (0.000)	0.07076034070098577*** (0.000)
WoodDeckSF	9.024*10^-5*** (0.009)	9.54*10^-5*** (0.006)
ScreenPorch	0.0002682153847007611*** (0.001)	0.00030511582834294427*** (0.000)
LogLotArea	0.11302167318880194*** (0.000)	0.10611458021492158*** (0.000)
R-squared	0.896	0.871

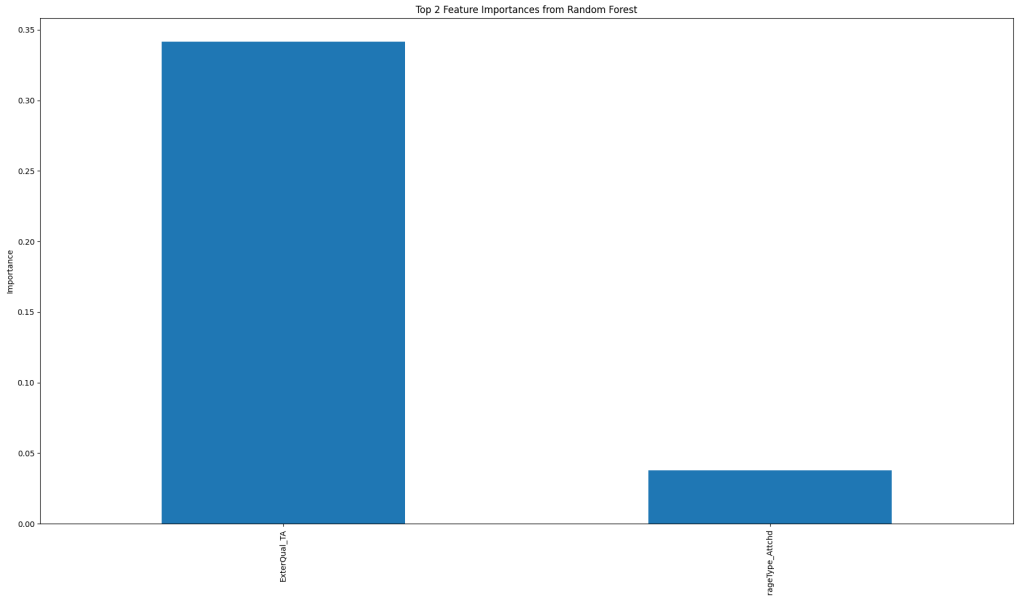
Notes: *, ** and *** indicate statistical significance of 90%, 95% and 99%, respectively.

Graph 1. Regressions Results



Result 3: Categorical variables underwent cleaning procedures. Utilizing a random forest algorithm, the top qualitative features impacting house prices were identified as external quality and garage location, with significance levels of 34% and 3.8%, respectively. The Mean Squared Error (MSE) of the Random Forest model was calculated to be 0.0388, indicating strong predictive performance. These findings are visually presented in Graph 2, highlighting the influence of qualitative features on house prices.

Graph 2. Random Forest Best Features



Conclusion

This study identifies significant factors affecting house prices through thorough data analysis and regression modeling. Utilizing techniques such as Weighted Least Squares and random forest algorithms, key determinants like lot area, living space,

house quality, and others are highlighted. These findings offer valuable insights for real estate stakeholders, aiding in informed decision-making for property transactions. This research contributes to a deeper understanding of house price dynamics, emphasizing the importance of both quantitative and qualitative features in determining property values.