ITM Web of Conferences **78**, 03005 (2025) https://doi.org/10.1051/itmconf/20257803005
CSEIT 2025

Predicting Housing Prices in Sweden: A Comparative Study of Linear Regression and Machine Learning Models

Luoping Feng

School of Data Science and Bigdata Technology, Hainan University, Haikou, China

Abstract. This study investigates the factors influencing the asking prices in the Swedish housing market, focusing on property-specific variables such as the number of rooms, land area, living area, and price per square meter. Using a linear regression model, the analysis reveals that these factors explain 18.9% of the variation in asking prices, with all coefficients showing significant positive relationships. To address the limitations of linear models, this study further employs nonlinear machine learning approaches, including decision trees and random forests, to capture complex interactions in the data. The decision tree model achieves perfect fit on training data ($R^2 = 1.000$) but shows reduced generalization on test data ($R^2 = 0.800$), suggesting potential overfitting. In contrast, the random forest model demonstrates robust performance, with high explanatory power ($R^2 = 0.892$ on test data) and minimal prediction errors, highlighting its superiority for housing price forecasting. Diagnostic tests confirm the absence of multicollinearity and autocorrelation in the linear model, while the machine learning models provide deeper insights into nonlinear relationships. These findings offer valuable guidance for investors and policymakers, emphasizing the importance of model selection in housing market analysis. Future research could integrate spatial analytics and macroeconomic shocks to further improve predictive accuracy.

1 Introduction

The housing market plays a crucial role in modern economies, acting as both an essential part of financial markets and a key driver of national economic stability [1]. Housing prices are influenced by a complex interplay of factors, including macroeconomic conditions, social trends, and local market dynamics [2,3]. These factors make accurate predictions of housing prices vital for a variety of stakeholders, including policymakers, investors, and consumers. Governments, for instance, rely on housing price forecasts to stabilize real estate markets, preventing large fluctuations in prices, which can have detrimental effects on the broader economy [4]. This stability is essential for fostering long-term economic growth, as the housing sector directly affects employment, wealth distribution, and consumer spending.

20223000406@hainanu.edu.cn

The complexity of housing price prediction lies in the multifaceted nature of the housing market [5, 6]. Housing serves as both a consumption good (homes for living) and an investment good (assets), meaning that prices are influenced by both personal and economic factors. Demand-side factors include household income, population growth, and shifts in preferences, while supply-side factors, such as construction costs, land availability, and zoning laws, also play a significant role. Furthermore, macroeconomic variables like interest rates, inflation, and income inequality can significantly affect the affordability and desirability of housing.

Traditionally, econometric models such as linear regression and variance analysis have been used to predict housing prices [7]. Linear regression is particularly useful for understanding the relationship between housing prices and a set of independent variables. Through this method, it is possible to identify how changes in certain factors, such as location or square footage, influence housing prices. Variance analysis, on the other hand, allows the extent to which variations in housing prices are explained by the included factors to be examined, helping to assess the robustness of predictions and the impact of each variable [8]. However, these linear models may struggle to capture complex, nonlinear interactions present in real-world housing markets. Where factors like speculative demand, policy shocks, and spatial dependencies can lead to abrupt price changes.

To address these limitations, recent studies have turned to nonlinear models, particularly Machine Learning (ML) approaches, which can better handle complex patterns in housing data. ML techniques such as random forests and neural networks have shown promise in improving prediction accuracy by modeling intricate relationships between economic, demographic, and spatial variables [9]. However, the trade-off between interpretability and predictive performance remains a key consideration in model selection.

Comparing how well linear and nonlinear models forecast Swedish home prices is the main goal of this study. Using a dataset that includes property characteristics (e.g., type, land area, rooms, location) and macroeconomic indicators (e.g., interest rates, inflation), this paper aims to assess how well traditional econometric methods generalize compared to more flexible ML approaches. By evaluating their predictive accuracy and interpretability, this research seeks to provide insights into the strengths and limitations of each modeling framework in capturing the dynamics of the Swedish housing market [10].

2. Methodology

2.1 Data Source and Description

The dataset used in this study is sourced from publicly available Swedish housing data on Kaggle [11]. The original dataset contains multiple features, including: Ad ID number (`ad_id`) ; Publication date (`date_published`) ; Property type (`typology`, e.g., house, apartment); Asking price in SEK (`asking_price_sek`); Land area in square meters (`land_area_sqm`); Living area in square meters (`living_area_sqm`); Price per square meter (`sqm_price_sek`); Number of rooms (`number_rooms`); Address, location, and coordinates. From these, this paper selected five key features for analysis, and the detailed information of them is shown in Table 1.

Table 1 Variable Information

Variables	Data
Property Typology	Agricultural_Estate; Apartment; Estate_Without_Cultivation; Foresting_Estate;

	Homestead; House; Linked_House; Plot; Terraced_House; Row_House; Twin_House; Vacation_Home; Vacation_House; Winterized_Vacation_Home; Other
Asking Price (SEK)	From 1000000 to 5000000
Land Area (m²)	From 50 to 2000
Living Area (m²)	From 0 to 300
Price per Square Meter (SEK)	From 10000 to 90000

These features were chosen because they directly impact on housing prices and are straightforward to analyze. In addition, macroeconomic indicators (e.g., interest rates, inflation) were incorporated from external sources to enhance the model's predictive power. These features are used in the study in this paper.

The analysis of the housing dataset shows considerable variation across the variables in Table 2. The average land area is 897.092 sqm, but with a high standard deviation of 2000.000 sqm, suggesting the presence of outliers. The living area has an average of 94.096 sqm, indicating most properties have smaller living spaces. The average asking price is approximately 2575100 SEK, with a high standard deviation, highlighting price disparities. Similarly, sqm price averages 36141.034 SEK, but the variance is large, indicating diverse pricing across properties.

Table 2 Insights from the Housing Dataset Analysis

Name	Minimum	Maximum	Average	Standard deviation
living_area_sqm	0.000	4650.000	100.224	86.391
land_area_sqm	0.000	31600000.000	12510.635	437962.384
asking_price_sek	24000.000	53750000.000	3116812.028	2642988.778
sqm_price_sek	0.270	12666666.670	48766.135	289577.738

2.2 Linear and non-linear prediction models

2.2.1 Linear Regression

Linear regression is a fundamental statistical technique that examines and quantifies the association between a response variable and its explanatory variables. This approach establishes a predictive model by fitting a linear equation to observed data, allowing for the analysis of how changes in predictors influence the outcome. The model assumes that the first three predictors exhibit a linear relationship with the asking price, as described by the equation.

$$asking\ price\ sek = \beta_0 + \beta_1 * \text{land area sqm} + \beta_2 *\ living\ area\ sqm + \beta_3 * sqm\ price\ sek + \varepsilon \quad (1)$$

Additionally, property typology is analyzed separately. Data preprocessing includes applying Min-Max normalization to standardize the input features. All analyses are conducted using SPSSAU, which provides tools to calculate regression coefficients, p-values, and R^2 statistics to assess the model's fit and performance. These statistical metrics help evaluate how well the model explains the variance in the dependent variable, thus guiding the interpretation of the results [12].

2.2.2 Decision Tree

A well-liked machine learning model for classification and regression applications is the decision tree. In order to arrive at a judgment or prediction at the leaf nodes, it first divides the data into subgroups according to the value of the features that were input. The first three variables and the asking price are presumed to be related by the model. A decision tree will use these features to iteratively split the data, identifying the most significant predictors of askingpricesek. By visualizing and interpreting the tree, it possible to gain insights into how different variables contribute to price prediction, helping inform decisions in real estate and property valuation.

2.2.3 Random Forest

Random Forest, a widely employed ensemble learning algorithm, is applicable to both classification and regression problems. During the training phase, this method generates many decision trees and aggregates their predictions to produce the final result. To enhance model generalization and mitigate overfitting, each individual tree is trained using a randomly sampled portion of the training dataset, thereby improving the overall robustness of the model. The results from all trees are aggregated, typically by averaging for regression tasks, to provide a final prediction. The model's ability to handle non-linear relationships and its high accuracy make it a powerful tool for real estate price prediction. By evaluating the importance of each feature, it is possible to gain deeper insights into which variables most significantly impact property prices.

3 Results and Discussion

3.1 Linear regression

From the Table 3, it can be seen that number_rooms, land_area_sqm, livingareasqm, and sqmpricesek are the independent variables, whereas the variable that is dependent in an analysis of linear regression is askingprice_sek. The following is the model formula:

$$askingpricesek = 0.035 + 0.506 * landareasqm + 1.011 * livingareasqm + 0.210 * sqmprice\_sek. \quad (2)$$

The model's R-squared value is 0.189, meaning that landareasqm, livingareasqm, and sqmpricesek explain 18.9% of the variation in askingpricesek. During the F-test, the model passed the test (F = 32.938, p = 0.000), landareasqm, livingareasqm, sqmpricesek have impact on askingpricesek. Additionally, the multicollinearity checks shows that the VIF values for all variables are below 5, indicating there is no multicollinearity problem. Furthermore, the Durbin-Watson value is around 2, which suggests there is no autocorrelation, and the sample data is not correlated, indicating that the model is well-fitted.

Table 3 Linear Regression Analysis Results

	Regression Coefficients	95% CI	Multicollinearity Diagnostics	
			VIF	Tolerance
Intercept	0.035** (55.048)	0.034 ~ 0.036	-	-
land_area_sqm	0.056** (17.009)	0.048 ~ 0.054	1.000	1.000
living_area_sqm	1.011 (44.889)	0.967 ~ 1.055	1.030	0.971
sqm_price_sek	0.210** (11.457)	0.174 ~ 0.246	1.030	0.971
Sample Size	11549			
R 2	0.189			
Adjust R 2	0.189			
F -value	F (3,11545)=897.447,p=0.000			

Notes: Dependent variable = asking_price_sek
D-W value = 1.840
* p<0.05 ** p<0.01 (t-values in parentheses)

3.2 Variables correlation based on Variance

The Fig. 1 of the ANOVA results indicate significant differences in the asking prices across various property types, with a p-value of 0.000 and an F-value of 64.53, the differences are statistically significant because they are below the 0.01 significant threshold. Agricultural estates have the highest average asking price at 8676 SEK, followed by estates without cultivation and foresting estates. In contrast, vacation homes and vacation houses have much lower asking prices. The data suggest that property type plays a major role in determining asking prices, with agricultural estates having the highest prices, and vacation properties generally being lower. The variation in standard deviations also highlights the differing price ranges within each property type.

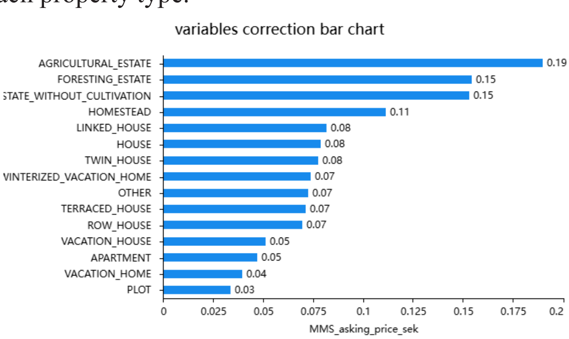

Fig. 1 Variables correlation bar chart (Picture credit : Original)

3.3 Decision Tree

Based on the provided decision tree analysis results, the model's performance varies between the training and test datasets. Despite the reduced performance on the test set, the model's Explained Variance Score (EVS) remains at 0.800, suggesting that the model can explain 80% of the variance in the test data. The Mean Absolute Percentage Error is null on the test set, which may imply that the model cannot provide an effective percentage error estimate in certain scenarios.

Overall, the simulation performs very well on the set used for training but somewhat worse on the test set, which may be a sign of overfitting. Additional regularization procedures or more model parameter adjustment may be required to improve the model's capacity for generalization. The detailed performance metrics and trends can be observed in Fig. 2.

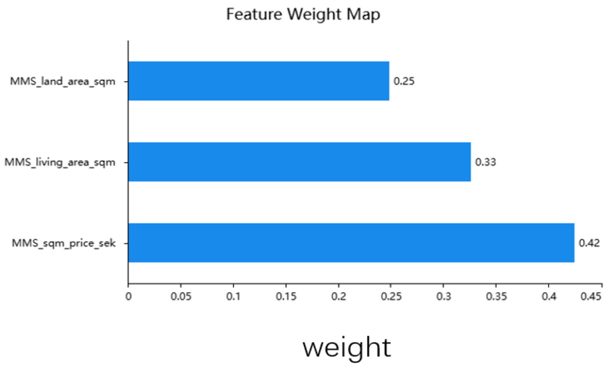

Fig. 2 Feature Weight Map-decision tree (Picture credit : Original)

3.4 Random Forest

With an R-squared value of 0.892, the Random Forest model's analysis shows excellent predictive ability on the testing dataset and accounts for a sizable amount of the volatility in house prices. With a Mean Absolute Error (MAE) of 0.003, the model demonstrates a low average divergence between the expected and actual values. The correctness of the model is further supported by the extremely low Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values.

The Median Absolute Deviation (MAD) is 0.000 for both datasets, highlighting the model's robustness against outliers. The Mean Absolute Percentage Error (MAPE) is 2.008 for the training set, which is relatively low, though it is not available for the test set. Consistent with the R-squared metric, the Explained Variance Score (EVS) further validates the model's effectiveness in capturing data variability. The Mean Squared Logarithmic Error (MSLE) is also 0.000, indicating minimal error in logarithmic terms.

Overall, the model demonstrates excellent predictive performance and robustness, making it a reliable tool for the given dataset. The detailed performance metrics and trends can be observed in Fig. 3.

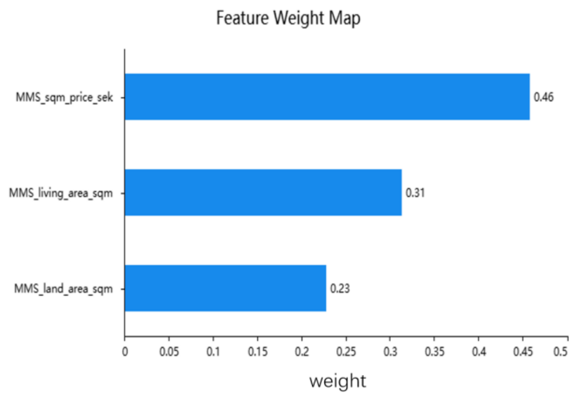

Fig. 3 Feature Weight Map-random forest (Picture credit : Original)

4 Conclusion

The study examines how property-specific factors like the number of rooms, land area, living area, and price per square meter influence asking prices in the Swedish housing market. The linear regression model explained 18.9% of the variation in asking prices, with significant positive impacts from all factors. Living area and price per square meter were particularly influential. Agricultural estates were found to command the highest average prices, highlighting the importance of property type. The model passed multicollinearity and autocorrelation tests, indicating robustness. However, the relatively low R-squared value of 0.189 suggests other factors, such as location or macroeconomic influences, contribute to the unexplained variation. The decision tree model showed overfitting, while the Random Forest model had an R-squared value of 0.892, indicating superior performance. These findings are useful for investors and policymakers in understanding housing price dynamics. Future research could integrate spatial analytics to better account for location-specific factors and use advanced machine learning methods like boosting or neural networks to improve predictive accuracy and capture more complex market patterns.

References

[1] Strouhal, J., Dvorakova, D.: 'Financial instruments' revaluation in financial statements some issues for the financial management'. Proc. IEEE Int. Conf. Inf. Manag. Eng., Chengdu, China, April 2010, pp. 94–98
[2] Nguyen, H.T., Nguyen, D.T.: 'Transfer Learning for Macroeconomic Forecasting'. Proc. NAFOSTED Conf. Inf. Comput. Sci., Ho Chi Minh City, Vietnam, November 2020, pp. 332–337
[3] Baatarjav, E.-A., Dantu, R.: 'Current and Future Trends in Social Media'. Proc. IEEE Int. Conf. Privacy Secur. Risk Trust Soc. Comput., Boston, USA, October 2011, pp. 1384–1385
[4] Ge, X.J., Mangioni, V., Shi, S., Herath, S.: 'House price forecasting using the multi-level modelling method in Sydney', Int. J. Housing Mark. Anal., 2024, 17, (2), pp. 287–306
[5] Clapp, J.M., Giaccotto, C.: 'The influence of economic variables on local house price dynamics', J. Urban Econ., 1994, 36, pp. 161–183

[6] Borwn, L., Thomas, H., James, C., et al.:'The title of the paper, IET Communications, 2012, 6, (5), pp 125-138

[7] Yan, J., Lei, F., Mo, D., Xue-qin, T.: 'Housing market cycle in China: An empirical study based on stock-flow model'. Proc. Int. Conf. Manag. Sci. Eng., Melbourne, Australia, November 2010, pp. 1671–1676

[8] D’haene, T., Pintelon, R., Schoukens, J., Van Gheem, E.: 'Variance analysis of frequency response function measurements using periodic excitations'. IEEE Trans. Instrum. Meas., 2005, 54, (4), pp. 1452–1456

[9] Abdelhalim, A., Traore, I.: 'A New Method for Learning Decision Trees from Rules'. Proc. Int. Conf. Mach. Learn. Appl., Miami, USA, December 2009

[10] Dharmarajan, K., U.N., A.: 'Diabetes Prediction using Random Forest Classifier with Different Wrapper Methods'. Proc. Int. Conf. Edge Comput. Appl., Tamilnadu, India, October 2022

[11] Borges, H.: 'Sweden Housing Prices'. Available: https://www.kaggle.com/datasets/henriquejborges/house-prices-in-sweden, accessed 13 February 2025

[12] The SPSSAU project: 'SPSSAU'. (Version 25.0). Available: https://www.spssau.com, accessed 20 February 2025