PIONEER
ACADEMIC PUBLISHING LIMITED

Journal of World Economy
ISSN 2709-3999
www.pioneerpublisher.com/jwe
Volume 2 Number 3 September 2023

Housing Price Prediction Using Machine Learning Algorithm

Ling Zhang[1]

[1] Business School, Hong Kong Baptist University, Hong Kong SAR, China
Correspondence: Ling Zhang, Business School, Hong Kong Baptist University, Hong Kong SAR, China.

doi:10.56397/JWE.2023.09.03

Abstract

This study examines the effectiveness of various machine learning models, including K-Nearest Neighbors (KNN), Ridge Regression, Random Forest, and Extreme Gradient Boosting (XGBoost), in predicting housing prices, using Multiple Linear Regression as a traditional baseline for comparison. Utilizing a dataset of residential property sales in Ames, Iowa, the XGBoost model was found to significantly outperform the baseline, highlighting the efficacy of machine learning techniques in this domain. Furthermore, the study applied the TreeSHAP package to enhance the interpretability of the XGBoost model, effectively bridging the gap between prediction accuracy and interpretability. The results underscore the potential of machine learning methods, particularly XGBoost, in housing price prediction, suggesting a need for further research to validate these findings using different datasets and exploring other machine learning algorithms.

Keywords: multiple linear regression, machine learning, SHAP values, housing price prediction

1. Introduction

The housing market, characterized by its complexity and unpredictability, is influenced by numerous factors such as economic indicators, government policies, demographic shifts, and market sentiments, often leading to heightened risks and uncertainties for stakeholders. In the ever-changing real estate landscape, precise and up-to-date housing price prediction has become increasingly critical. Real estate participants such as agents, appraisers, assessors, mortgage lenders, brokers, property developers, investors, fund managers, policy makers and both current and prospective homeowners can greatly benefit from accurate housing price predictions. These insights can aid in informed decision-making, optimizing strategies, improving risk management, and guiding sound policy-making.

This study aims to investigate the utility and efficacy of different machine learning techniques in predicting housing prices and to compare their performance with traditional multiple regression analysis, using a dataset of residential property sales in Ames, Iowa. The dataset consists of 2919 observations with 80 distinct variables categorized into physical features, locational characteristics, and neighborhood and amenity aspects. The investigation explores various models including K-Nearest Neighbors, Ridge Regression, Random Forest, and Extreme Gradient Boosting.

Notably, the application of machine learning techniques in real estate prediction extends beyond simply achieving high predictive accuracy. A critical challenge lies in the interpretability of these models, a quality highly desired in the real estate market for understanding the factors influencing property values. Consequently, this research also employs the Shapley Additive Explanations (SHAP) approach, aiming to shed light on the influence of various factors on the predictions made by the machine learning models.

The paper is organized as follows. After this introduction, a review of related literature is provided, which is succeeded by the research question, and a description of the data collection process and the chosen machine learning models. Subsequent sections discuss the data preprocessing steps, model results, and interpretations of the results, particularly through the lens of the SHAP values. The paper concludes with a discussion on the implications of the findings, the limitations of the study, and directions for future research.

In summary, this study demonstrates the feasibility and effectiveness of machine learning in predicting housing prices, proving its superiority over traditional methods. It bridges the gap between prediction accuracy and interpretability, underscoring the extensive potential of these advanced techniques.

2. Literature Review

Beginning in the 1950s, as part of an endeavour to appraise a multitude of property values, the United States, as a pioneer, widely implemented mass appraisal methods domestically. These models drew upon foundational real estate characteristics such as size, age, and geographic location for value estimation. By the 1970s and 1980s, the Hedonic model within the field of economics began to see widespread application, with the advancement of computers enabling scholars to employ more complex statistical models and appraisal methods. Historically, hedonic regression models have been utilized in efforts to interpret and replicate the pricing mechanisms within residential real estate. However, the reliability of these methods has been challenged by prior research. The primary challenge with hedonic regression or multiple regression analysis often lies in selecting the appropriate functional form for the dependent variable. Additionally, there may be instances where the assumptions concerning the error term in the regression model are not met (Chun Lin, C. & Mohan, S. B., 2011).

Geo-information system (GIS) was subsequently introduced, such as geographically weighted regression (GWR), geographically weighted principal component analysis and spatial auto-regressive models (SAR), which took into account the impact of geographic location on housing prices. In their research, McCluskey and Borst employed GWR to identify submarkets, underscoring the importance of segmenting the real estate market for the purposes of mass appraisal (McCluskey, W. J. & Borst, R. A., 2011). Leveraging the benefits of geographic weighting, Wu et al. incorporated Geographically Weighted Principal Component Analysis into a revised data-driven method with the aim of addressing issues related to housing submarkets (Wu, C., Ye, X., Ren, F. & Du, Q., 2018)0. Some scholars currently exhibit a preference for integrating these methodologies with machine AI-based models, forming hybrid models to enhance prediction accuracy.

With the explosive advancements in internet and computer technologies in the 21st century, large-scale data mining and machine learning models have begun to be employed in real estate price prediction. The real estate market has been proactively experimenting with techniques such as Support Vector Machines (SVM), Decision Trees (DT) and Artificial Neural Network (ANN). Compared to traditional linear multiple regression models, Chen et al. found that SVM algorithm was capable of predicting residential housing prices more accurately, given its ability to handle nonlinear relationships and high-dimensional data, while also better capturing spatial dynamics of housing prices. Additionally, the SVM algorithm can enhance prediction accuracy by reducing data dimensions through the identification of support vectors—all encapsulated in a single, unified framework (Chen, J. H., Ong, C. F., Zheng, L. & Hsu, S. C., 2017). A Random Forest constitutes an exemplar of ensemble learning, wherein multiple decision trees amalgamate to form a ‘forest’. This model demonstrates efficiency when handling expansive property datasets, managing input variables without the need for deletion. Antipov and Pokryshevskaya pioneered the application of the Random Forest model in mass appraisal and discovered its superior performance relative to other models (Antipov, E. A. & Pokryshevskaya,

E. B., 2012). When considering the accuracy of housing price predictions in Turkey, the ANN model outperforms the Hedonic Regression model, presenting a more precise forecast from both rural and urban viewpoints (Selim, H., 2009). The traditional multiple regression or feature models are gradually being supplanted, not only due to the advent of emerging technologies but also owing to an assortment of inherent issues that prevent it from always being the optimal modelling approach. These issues encompass model specification errors, inability to satisfy stringent fundamental assumptions, and subpar performance in the face of missing data or inconsistent data quality. Nonetheless, considering factors such as operational convenience and simple understanding, governmental fiscal systems across various nations continue to utilize the hedonic or multiple regression model.

Concurrently, advanced technology does not always hold the upper hand under all circumstances. Given that the SVM delivers optimal performance with a limited amount of data, typically in the range of thousands of sample datasets, it could require extensive operational time when processing massive data sets, those amounting to hundreds of thousands or more (Lam, K. C., Yu, C. Y. & Lam, C. K., 2009). When considering submarket variables simultaneously, Bourassa et al. found the prediction outcomes from both Ordinary Least Squares (OLS) and geographical statistical models are comparable, suggesting that the latter does not demonstrate the anticipated superiority (Bourassa, S. C., Cantoni, E. & Hoesli, M., 2007). Drawing on research by William McCluskey and colleagues, which concentrated on the Lisburn District Council area's real estate market in Northern Ireland, a non-linear regression model was found to surpass the ANN in terms of predictive accuracy. Furthermore, the opacity of the ANN's output did not provide an unequivocal model for defending predicted values against objections (McCluskey, W., Davis, P., Haran, M., McCord, M. & McIlhatton, D., 2012)0.

3. Research Questions

In the scholarly realm, the effectiveness of machine learning in forecasting prices is still a subject of debate. Some scholars assert that machine learning techniques surpass the hedonic model in terms of predictive performance, while others hold that the results are either on par or even inferior. This gives rise to key inquiries: (1) Is the predictive power of machine learning genuinely effective? It would be satisfactory to utilize a model achieving an estimate between 60-70% of the total value, acknowledging the fact that the pursuit of models predicting 90-100% of market value may not always be crucial or practical (J. McCluskey, W., Zulkarnain Daud, D. & Kamarudin, N., 2014). This implies that can machine learning satisfy or even surpass this 60-70% accuracy threshold? In line with this, we also need to consider if (2) machine learning techniques truly outperform the traditional multiple regression model? Additionally, the fiscal departments of governments, with their preference for the more conservative hedonic model due to its operational convenience and simple intelligibility, do not necessarily reflect a lack of interest in model interpretability among those in the real estate market who utilize advanced techniques. Thus, the question arises – (3) how can the impact of various factors on pricing within a complex machine learning model be interpreted? These aspects will be elaborated upon in the subsequent sections of this paper.

4. Methodology

The methodology employed in this study initially establishes a fundamental linear regression model based on least squares. Setting this baseline model is instrumental as it offers an objective measure for gauging performance enhancements in the subsequently specified models. The purpose of all models is to minimise the predictive error, which measures the proximity of the model's estimations to their actual values.

4.1 Multiple Linear Regression

The formula of the multiple linear regression is:

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

Where: Y = the dependent variables (sale price); $X_1 \dots X_n$ = independent variables; $\beta_0 \dots \beta_n$ = estimated parameters; and ε = error term.

4.2 K-Nearest Neighbors (KNN)

The regression prediction principle of KNN is to calculate the average or median of the 'k' closest points. Its algorithm can be broken down into the following steps:

(1) choosing a value of k;

(2) computing the Euclidean distance between the target data point, for which a prediction

is desired, and each data point in the training set. The equation to calculate Euclidean distance is as follows:

$$d(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

(3) finding the k nearest neighbour;

(4) calculating the prediction.

4.3 Ridge Regression

Ridge Regression, also known as Tikhonov regularization, is an improved version of the ordinary least squares regression model. It introduces an L2 penalty term (i.e., the sum of squares of the coefficients) into the least squares estimation to address the issue of multicollinearity in multiple linear regression.

In multiple linear regression, when there is a high degree of correlation among features (multicollinearity), the estimated regression coefficients become unstable, to the point that minor data changes might cause large fluctuations in the coefficients. Ridge Regression mitigates the effects of multicollinearity among features by imposing a penalty on the magnitude of the coefficients, rendering the model more stable. The regression coefficient is calculated as $\widehat{w} = (X^T X + \lambda I)^{-1} X^T y.$

4.4 Random Forest Regression

Random Forest Regression utilizes an ensemble of decision trees, where each tree is dependent on the values of a random vector sampled independently. The core idea is that the trees all share the same distribution within the forest. By generating a multitude of decision trees and averaging their predictions, the Random Forest algorithm reduces the variance, which in turn improves prediction accuracy and helps to avoid overfitting.

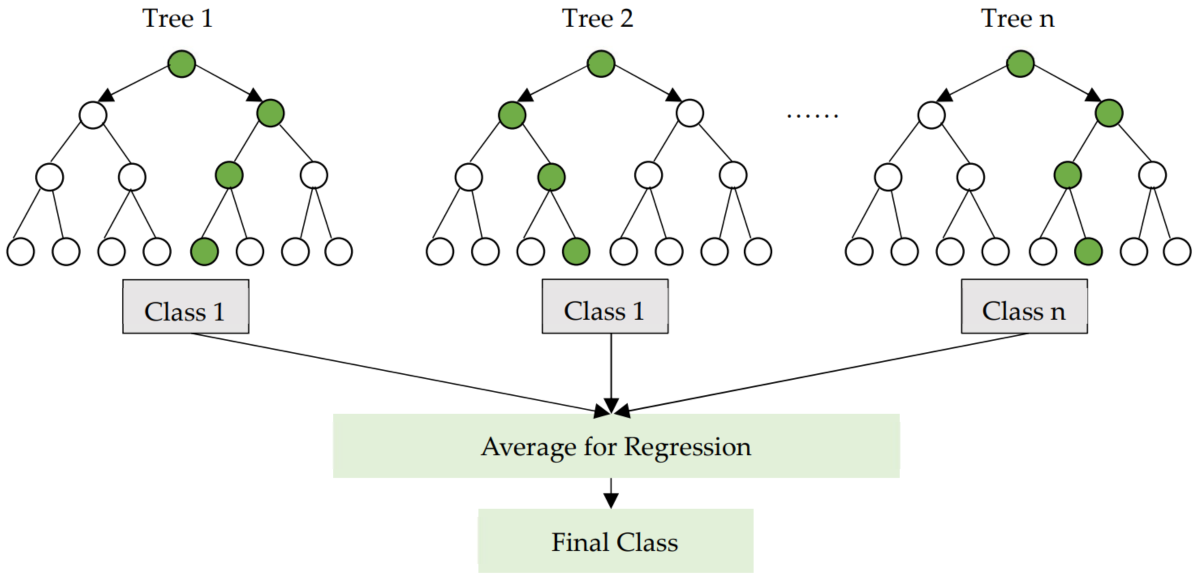

Figure 1. Random Forest Flow Chart

4.5 Extreme Gradient Boosting (XGBoost)

XGBoost is a high-performance machine learning algorithm based on gradient boosting, where multiple weak prediction models, typically decision trees, are sequentially trained to correct preceding errors and create a strong predictive model. XGBoost introduces regularization into the objective function to control complexity, reducing overfitting. It also has built-in handling of missing values, learning the best imputation during training. An advantage over other boosting methods is its tree pruning technique, which enhances efficiency by eliminating non-beneficial splits. Additionally, XGBoost leverages a Column Block structure for efficient sparse data handling, and employs parallel computing for speed.

4.6 Model Evaluation Metric

This research uses Root Mean Square Error (RMSE) as a metric for model evaluation. RMSE quantifies the difference between predicted and actual values by computing the square root of the average squared discrepancies. It's computed using the formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - actural_i)^2}{N}}.$$

Lower RMSE values indicate a model's superior predictive capacity and reduced error. RMSE, in contrast to the r-squared measure, delivers more granular insights into predictive errors, directly gauging the deviation between predicted and actual outcomes, instead of solely tracking their relative shifts. Furthermore, RMSE imparts a harsher penalty on significant errors compared to Mean Absolute Error (MAE), making it more responsive to larger discrepancies.

5. Data Collection

The dataset employed in this study encompasses 2919 observations and 80 distinct variables related to residential property sales in Ames, Iowa. The data is publicly accessible via the Kaggle competition platform, specifically under the following URL:

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.

The variables in the housing dataset are partitioned into three principal categories: (i) Physical features, emphasizing specific dwelling attributes like the age of the building, its quality, the lot area, views, among others. (ii) Locational characteristics, entailing service accessibility parameters like proximity to the central business district (CBD), access to public transport, and the distance to the nearest retail center. (iii) Aspects related to the neighborhood and available amenities.

Table 1. Selected Attribute Information Table

Attribute Name	Description
OverallQual	Assessment of material quality and finishing details of the dwelling
YearBuilt	Date of Initial construction
RoofMatl	Roof material
ExterQual	The caliber of the exterior material composition
BsmtCond	Assessment of the overall state of the basement
HeatingQC	Heating quality
Fireplace	Number of fireplaces
GarageCars	Size of garage in car capacity
YrSold	Year of Sale (in four-digit format, YYYY)
SaleCondition	Classifications of sale types

6. Result

6.1 Data Pre-Processing

Prior to the implementation of machine learning algorithms, data preprocessing constitutes a crucial step in fortifying the dependability of the data, as the data quality bears a direct influence on the precision of the learning outcomes. This study encompasses the following processes in data preprocessing and feature engineering:

I. Imputation of missing data;

II. Conversion of categorical variables into ordinal ones;

III. Transformation of categorical variables into dummy variables;

IV. Generation of new variables;

V. Removal of variables causing multicollinearity and outliers.

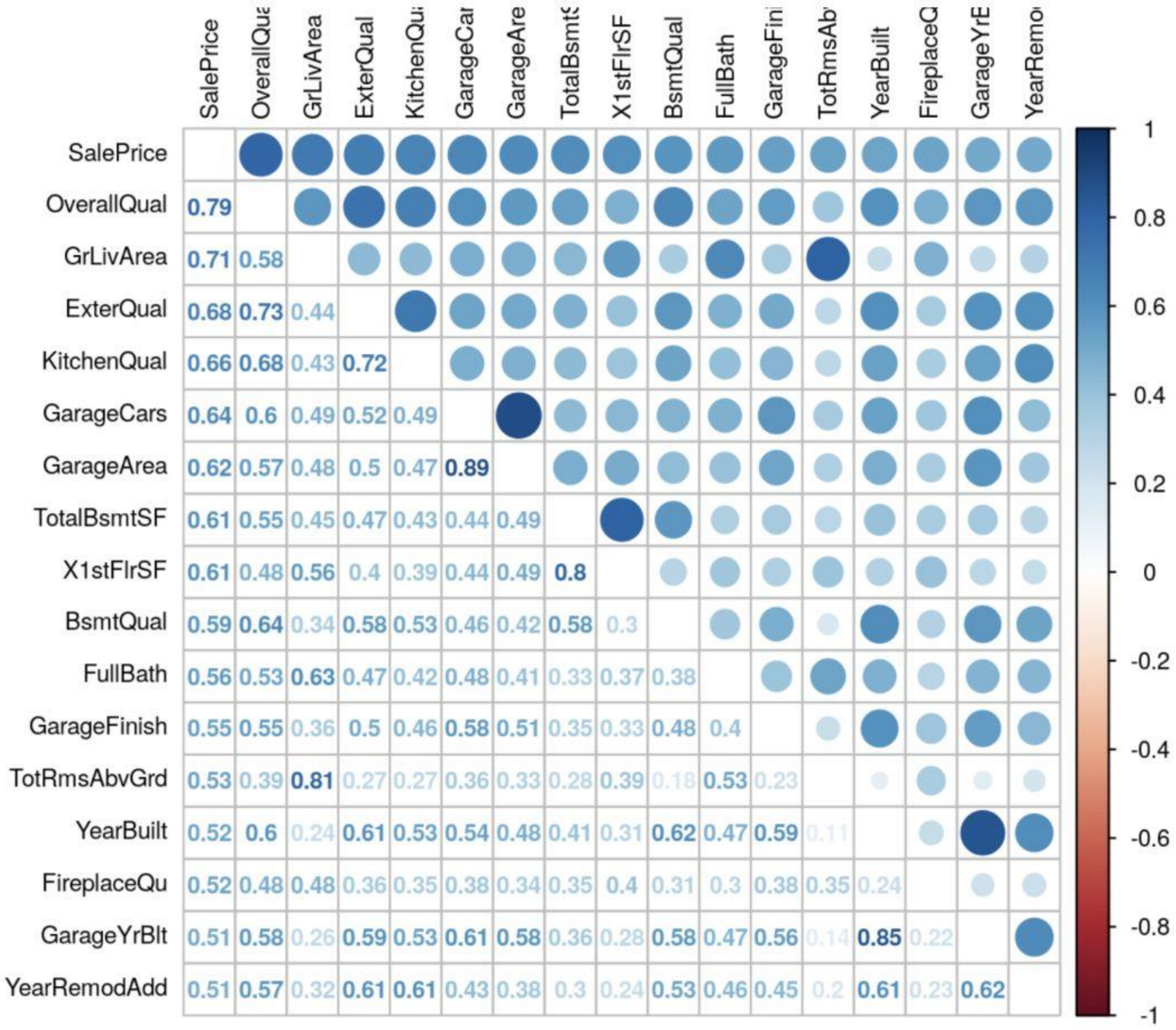

Figure 2. Correlation Between Variables After Data Processing

Following these steps, the data set was then randomly partitioned into two segments: a majority (80%) was employed for the development of models, while the remaining fraction (20%) was designated for model testing purposes, as commonly done in previous studies (Wilson, I. D., Paris, S. D., Ware, J. A. & Jenkins, D. H., 2002; Morano, P.I.E.R.L.U.I.G.I., Tajani, F. & Torre, C.M., 2015).

6.2 Model Results

In light of the earlier assertion that a model achieving a predictive estimate between 60-70% of total value could be considered satisfactory, the results showcase that machine learning models, especially XGBoost, surpass this benchmark significantly. In fact, the XGBoost model demonstrates the lowest RMSE of 0.12745, signifying the highest predictive accuracy among all models evaluated, thereby validating the potency of machine learning models in this context.

Table 2. Results of Models

Model	RMSE
Multiple Linear Regression	0.16530
KNN	0.23003
Ridge Regression	0.13370
Random Forest	0.14807
XGBoost	0.12745

6.3 Interpretation of Best Model

6.3.1 Feature Importance

Table 3 outlines the relative importance of various features as determined by the XGBoost model, the top-performing model in our study. The most influential feature in predicting the sale price is ‘GrLivArea’, with a feature importance value of 21.30, followed by ‘TotalBsmtSF’ (14.96), ‘MSSubClass’ (11.21), ‘OverallQual’ (10.75), and ‘TotRmsAbvGrd’ (10.02).

Table 3. Feature Importance of XGBoost

Model	Feature Importance
GrLivArea	21.30
TotalBsmtSF	14.96
MSSubClass	11.21
OverallQual	10.75
TotRmsAbvGrd	10.02

6.3.2 Practical Explanation

In order to provide further insight into the impact of these five important factors, as well as the remaining factors, on housing prices, this study randomly selects a subset of observations from the testing data as samples for analysis (Figure 3 and Table 4). To accomplish this task, the R package TreeSHAP was utilized within R Studio version 2022.07.2. TreeSHAP is specifically designed to compute SHAP (Shapley Additive Explanations) values for tree ensemble models, including XGBoost, LightGBM, gbm, ranger, and random forest. SHAP values are derived from game theory and serve as a method to explain model predictions by quantifying the contribution of each feature to the value of a single prediction. By utilizing SHAP, this paper aims to elucidate the relative importance of each feature in influencing the predicted housing prices.

Figure 3 illustrates the impact of variables on the sale price of the 1445th observation. Table 4 presents the price impact (in USD) of selected variables across different observations. For instance, in the 105th observation, the variable ‘GrLivArea’ positively impacts the sale price by $6,561.85, while ‘NeighborRich’ has a negative impact of $6,523.19. The variables ‘Fireplaces,’ ‘MasVnrArea,’ ‘ExterQual,’ and others also contribute to the overall price impact. The specific effect of each feature on the sale price varies across different observations, highlighting the intricate relationship between different features and their influence on housing prices.

By utilizing SHAP values, this study effectively unlocks the “black box” of the XGBoost model, shedding light on how different features contribute to the sale price of individual properties within the selected observations.

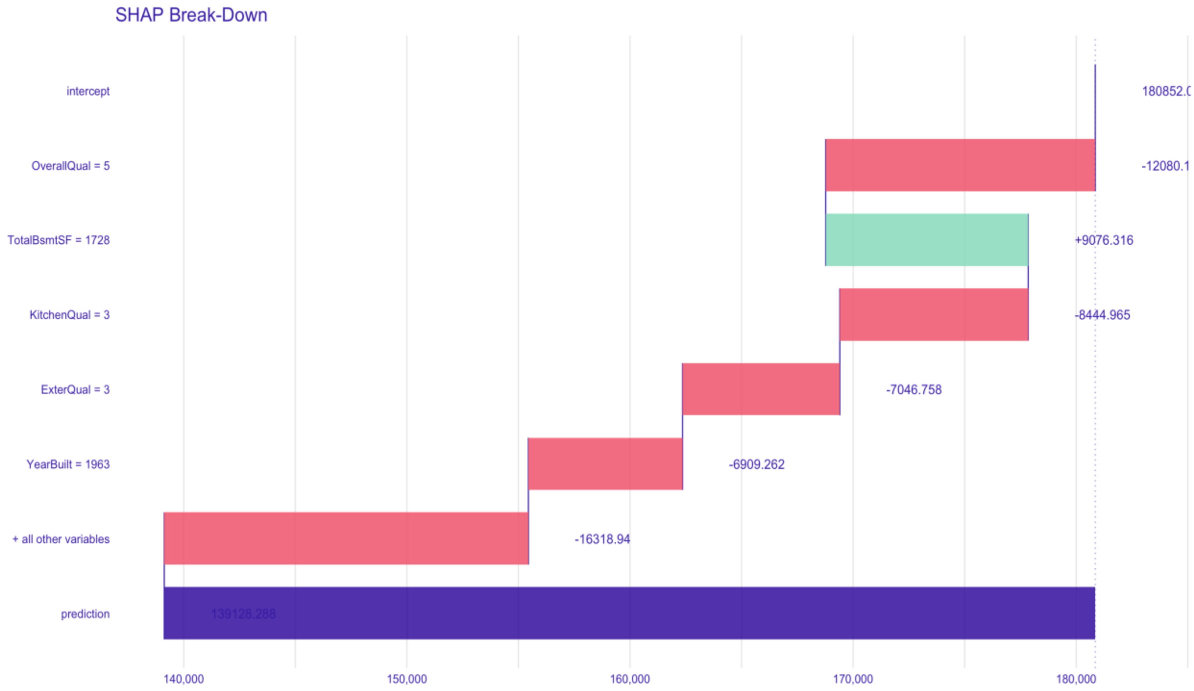

Figure 3. The Impact of Variables on Sale Price of the 1445th Observation

Table 4. Variables Price Impact (USD) on Selected Observations

Observation No. Independent Variables	105	125	145	165	225
Intercept	180856.78	180856.78	180856.78	180856.78	180856.78
GrLivArea	6561.85	N/A	N/A	4497.23	18721.14
NeighborRich	-6523.19	4347.67	-5244.64	-4713.12	N/A
Fireplaces	5393.44	N/A	N/A	N/A	N/A
MasVnrArea	5116.36	N/A	N/A	N/A	N/A
ExterQual	-4795.30	-4348.44	-5000.58	-3825.94	17827.99
LotArea	N/A	5623.24	N/A	N/A	N/A
KitchenQual	N/A	-4116.64	N/A	N/A	N/A
OverallQual	N/A	-3732.85	-6906.65	N/A	26876.71
TotalBsmtSF	N/A	N/A	8276.30	N/A	22260.80
TotRmsAbvGrd	N/A	N/A	5319.47	N/A	N/A
GarageCars	N/A	N/A	N/A	-4031.31	17724.65
Age	N/A	N/A	N/A	-3950.26	N/A
All other variables	-25245.57	-1177.37	-46421.08	-24208.51	105900.2

7. Discussion and Conclusion

The primary objective of this research was to investigate the efficacy of machine learning techniques in predicting housing prices and to compare their performance with traditional multiple linear regression model. Utilizing various machine learning models, including KNN, Ridge Regression, Random Forest, and XGBoost, the study explored their predictive capacities.

The results presented in this paper provide clear evidence that machine learning techniques, particularly the XGBoost algorithm, can effectively be employed to forecast housing prices with high accuracy. In fact, the XGBoost model outperformed all other models, yielding the lowest RMSE value of 0.12745, significantly surpassing the established benchmark of 60-70% predictive accuracy. While the results affirm the robust predictive capabilities of machine learning models, they also underline the crucial role of interpretability in these models. Using the TreeSHAP package, the research demonstrated how SHAP values could unravel the complexity of machine learning models like XGBoost. By quantifying the contribution of each feature to the predictions, the study successfully illuminated the influence of various factors on housing prices, making the seemingly "black box" model more transparent and interpretable.

Nevertheless, it is imperative to acknowledge the limitations of the study. While the dataset used offers a substantial number of observations, it is confined to residential property sales in Ames, Iowa. Therefore, the generalizability of the findings to different regions or countries may be limited. Additionally, although machine learning models exhibited superior performance in this study, it's worth noting that their utility may vary depending on the nature and quality of data available in different contexts.

In conclusion, this research underscores the potential of machine learning techniques in real estate price prediction, providing a promising avenue for future studies. However, further research is recommended to validate these findings using datasets from different regions or countries, and to explore more sophisticated machine learning algorithms or ensemble methods. Similarly, it would be intriguing to delve deeper into the interpretability of these models, ensuring that advanced predictive models can also offer insightful explanations of their predictions.

Reference

Antipov, E.A. and Pokryshevskaya, E.B. (2012). Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), pp. 1772-1778.

Bourassa, S.C., Cantoni, E. and Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*, pp. 143-160.

Chen, J.H., Ong, C.F., Zheng, L. and Hsu, S.C. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management*, *21*(3), pp. 273-283.

Chun Lin, C. and Mohan, S.B. (2011). Effectiveness comparison of the residential property mass appraisal methodologies in the USA. *International Journal of Housing Markets and Analysis*, *4*(3), pp. 224-243.

J. McCluskey, W., Zulkarnain Daud, D. and Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. *Journal of Financial Management of Property and Construction*, *19*(2), pp. 152-167.

Lam, K.C., Yu, C.Y. and Lam, C.K. (2009). Support vector machine and entropy-based decision support system for property valuation. *Journal of Property Research*, *26*(3), pp. 213-233.

McCluskey, W., Davis, P., Haran, M., McCord, M. and McIlhatton, D. (2012). The potential of artificial neural networks in mass appraisal: the case revisited. *Journal of Financial Management of Property and Construction*, *17*(3), pp. 274-292.

McCluskey, W.J. and Borst, R.A. (2011). Detecting and validating residential housing submarkets: A geostatistical approach for use in mass appraisal. *International journal of housing markets and analysis*, *4*(3), pp. 290-318.

Morano, P.I.E.R.L.U.I.G.I., Tajani, F. and Torre, C.M. (2015). Artificial intelligence in property valuations. An application of artificial neural networks to housing appraisal. In *Advances in environmental science and energy planning* (pp. 23-29). WSEAS Press.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert systems with Applications*, *36*(2), pp. 2843-2852.

Wilson, I.D., Paris, S.D., Ware, J.A. and Jenkins, D.H. (2002). Residential property price time series forecasting with neural networks. In *Applications and Innovations in Intelligent Systems IX: Proceedings of ES2001, the Twenty-first SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2001* (pp. 17-28). Springer London.

Wu, C., Ye, X., Ren, F. and Du, Q. (2018). Modified data-driven framework for housing market segmentation. *Journal of Urban Planning and Development*, *144*(4), p. 04018036.