İSTANBUL UNIVERSITY PRESS

Acta Infologica
acin 2024, 8 (2), 245–259
DOI: 10.26650/acin.1543650

Research Article

House Value Estimation using Different Regression Machine Learning Techniques

Tarek Ghamrawi[1], Müesser Nat[2]

[1]Cyprus International University, School of Applied Sciences, Haspolat, Lefkosa, Mersin 10 Turkey

Corresponding author : Müesser NAT
E-mail : mnat@ciu.edu.tr

ABSTRACT

This study investigates the effectiveness of various regression algorithms in estimating house values using a dataset sourced from Zillow.com, encompassing 15,000 residential properties from Denver, Colorado. Comparisons of different models such as linear regression, Ridge regression, Lasso regression, Elastic Net, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. The models were evaluated using R-squared (R^2) and Mean Absolute Error (MAE) as performance metrics. The results demonstrated that the Random Forest Regressor and XGB Regressor outperformed other models, achieving the highest R^2 scores and the lowest MAE values. These findings underscore the potential of these models for accurate house price estimation, which can be instrumental for the real estate market. Accurate valuations can help prevent overpricing, which causes properties to remain unsold for extended periods, and under-pricing, leading to financial losses. Implementing these regression models can enhance pricing strategies, ensuring efficient buying and selling processes and contributing to the overall financial health of the real estate market. Future research will explore the use of a broader range of regression models with fewer features to assess their performance and robustness in house price prediction.

Keywords: "House Price Estimation", "Machine Learning", "ElasticNet", "Lasso Regression", "Decision Tree Regressor", "Random Forest Regressor", "Linear Regression", "Ridge Regressor", "Gradient Boosting Regressor", "XGB Regressor"

Submitted : 05.09.2024
Revision Requested : 09.12.2024
Last Revision Received : 16.12.2024
Accepted : 17.12.2024

1. INTRODUCTION

Determining how much a house is worth, also known as home valuation, involves figuring out its market value in a fair and objective way. It's more than just checking the price tag; it uses different methods to get a good estimate. It's super important for pretty much every house deal. Sellers need accurate valuations to set reasonable prices, buyers want to make sure they're not paying too much, and agents need the right info to sell properties well, banks also need to know a house's value to decide how much of a risk it is to lend money for it. The value affects how big the loan is and what the interest rate will be. Insurance, companies, as well, use valuations to figure out how much coverage a homeowner needs. Getting it right means homeowners are protected financially if something bad happens to their house (Binu, 2020).

Accuracy matters because if a house is priced too high, it might sit on the market for ages, and if it's too low, the seller could lose out financially. Accurate valuations keep the real estate market healthy and stop people from getting ripped off, it is the basis for a well profound decision-making process (Thomas, 2023).

Machine learning (ML) offers exciting possibilities to overcome traditional house valuation methods limitations and enhance house value estimation accuracy by using the predictive performances of Linear Regression, Ridge Regression, Lasso regression, Decision Tree regressor, Random Forest regressor, ElasticNet, Gradient Boost Regression, and XG Boost Regression as regressor analytics

2. LITERATURE REVIEW

Multiple studies have used regressor methods on different data sets and a literature review was conducted on the following:

Wang (2018) proposed the utilization of Random Forests, a machine-learning approach, for developing a house price estimation model. The primary objective was to compare its performance against a benchmark linear regression model. The findings revealed that the Random Forest model excelled in capturing hidden nonlinear relationships between house prices and their features, resulting in more accurate estimations than the linear regression model. In this study, 27649 data points were used for the house assessment price of 2015 in Arlington country, Virginia USA

Similarly, Li (2023) focused on predicting house prices in King County, Washington, employing four machine learning models: linear regression, Random Forest (RF), Artificial Neural Network (ANN), and XGBoost. The aim was to identify key features influencing house prices and provide insights for future investments. Their results indicated that XGBoost achieved the highest accuracy among the models tested. Additionally, the study identified grade, square footage of living space, and latitude as the most influential features affecting house prices using 21,611 pieces of data and 21 features.

In a comparative study (Rana, 2020), various regression techniques were evaluated for their effectiveness in house price prediction. Support vector regression, XGBoost, Decision tree regression, and Random Forest regression. The results demonstrated that advanced regression techniques provided more accurate predictions compared to traditional linear models. The study underscored the importance of selecting relevant features to enhance model performance using 13,320 data points and 9 features.

Furthermore, Maida (2022) applied the Xtreme Gradient Boosting Model (XGBoost) to predict house prices in Karachi, Pakistan, using data from an open real estate portal. The XGBoost model achieved a remarkable 98% accuracy, highlighting its efficiency and flexibility in house price prediction. This study emphasized the model's superior performance in comparison to other models utilized for similar purposes using 38,961 data points and 14 features.

In another study by Truong et al. (2020) explored the performance of advanced machine learning models, including Random Forest, XGBoost, LightGBM, and ensemble techniques like Hybrid Regression and Stacked Generalization. Using a dataset of 231,962 housing records and 19 features from Beijing, they reported that Stacked Generalization achieved the best test set performance, leveraging multiple model outputs for robust predictions. The study underscores the effectiveness of combining machine learning models to enhance accuracy in house price estimation even though in this study the R2 was addressed as high or decent accuracy instead of providing real values but a root mean square logarithmic error was provided (RMSLE).

Additionally, Hernes et al. (2024) analysed the primary residential real estate market in Wroclaw, Poland, using a dataset of 15,000 records collected via a web scraping approach. The study implemented multiple machine learning models, including Gradient Boosting Regression, Random Forest Regression, LASSO, Multiple Linear Regression, and Simple Linear Regression, to predict housing prices based on attributes like area, number of rooms, year of construction, and more. Gradient Boosting Regression achieved the highest performance with an R^2 of 0.989, followed

closely by Random Forest Regression with an R^2 of 0.986. The study highlighted the utility of machine learning models for achieving high prediction accuracy and their potential application for real estate investors and institutions through a developed web-based prediction tool.

On the other hand, in this study Ali Soltani et al. (2022) analysed 428,000 residential transactions over a 32-year period (1984–2016) in Metropolitan Adelaide, Australia, using 38 explanatory variables and additional spatiotemporal lag features. The performance of four machine learning models—Linear Regression, Decision Tree, Random Forest, and Gradient-Boosted Tree—was evaluated. The Gradient-Boosted Tree model delivered the best results, achieving an R^2 of 0.896 and an RMSE of 0.086. This research underscored the significance of incorporating spatiotemporal dependency into machine learning models to enhance predictive accuracy in house price estimation.

3. AIM

The foundation of artificial intelligence is data. Models cannot be trained without sufficient data, which means pricey and sophisticated technology is left idle. The information found in data is what the models use to identify trends, glean insights, make predictions, and grow into more sophisticated models.

Although gathering reliable, and high-quality data can be an expensive and time-consuming procedure because there is a science to it. Certain types of data are extremely controlled, requiring lengthy lead times to obtain access and authorization Additionally the size of the data might be so small, that even when secured, training models might not end up finding it useful (Li, 2023). This research aims to demonstrate that, with a relatively small dataset (approximately 11,078 data points) and a limited set of carefully chosen features (7) from zillow.com for the state of Denver, machine learning models can achieve highly accurate house price estimations. This is in contrast to existing research, which often utilizes significantly larger datasets (average of 25,385.25 data points) and a broader range of features (average of 13).

In contrast to the methodologies adopted in the aforementioned studies, the current research focuses on employing a broader range of regression models with a significantly reduced feature set. This strategy aims to investigate whether a smaller set of features, which usually leads to underfitting (IBM, 2024), can still produce high-performing scores, thereby simplifying the model-building process and potentially enhancing interpretability without sacrificing accuracy.

4. METHODOLOGY

A. Data preperations

In this project, the dataset is gathered straight from Zillow.com, a trusted hub for real estate information. It was a dataset packed with details about residential properties, totalling 15,000 entries from the city of Denver in Colorado. To dig into what makes house prices tick,

One big task when preparing the data was handling missing data. missing values were not replaced with the mean, so a trick called "dropna()." Basically, it's a way to drop the blanks. It has been done to keep things legit and make sure the data stays reliable and after this dropping only 11,078 data points were left with a minimum value of $147,767 and a maximum value of $1,600,000.

B. Feature Engineering

In this study the following numerical features were selected which are bedrooms, bathrooms, rooms, square footage, lot size, year built, and last sale amount for this study. These features directly influence house prices by representing key aspects of a property's size, condition, and historical transaction values. Non-numerical features like id and address, longitude, altitude, and prior sale Date, were excluded as they do not directly impact house prices,

C. Technology and tools

For this study, JupyterNotebook was used, an open-source web application, as the primary technology platform. JupyterNotebook provides an interactive environment for conducting data analysis, exploratory research, and model development. Its flexibility and ease of use allowed the study to efficiently explore the dataset, perform feature engineering, and train predictive models. The dataset was first divided into training, validation, and testing sets to ensure the robustness of the model and to prevent overfitting. The data was split as follows:

- Training Set: 70% of the data
- Test Set: 30% of the data

This splitting was done for all learning models. Additionally, to enhance the predictive performance of the regression models, polynomial features were applied engineering and normalization. The features were expanded to polynomial degrees 1, 2, and 3, capturing non-linear relationships in the data. After polynomial feature expansion, Standard Scaler

was applied to normalize the features, ensuring that they have a mean of zero and a standard deviation of one, which helps improve the convergence and performance of the regression models.

The models were assessed using the R^2 Test score to determine the proportion of variance in the dependent variable predictable from the independent variables in order to maintain transparency. This methodology is represented in the workflow figure 1 below (Géron, 2019)

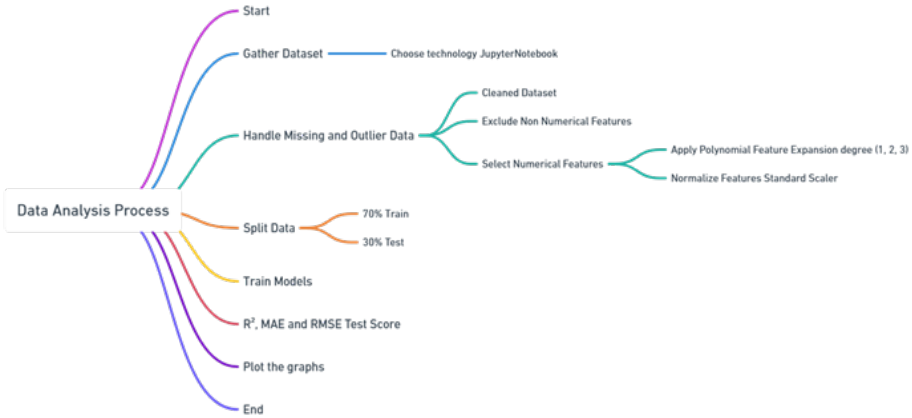

Figure 1. *Methodology Workflow (Géron, 2019)*

D. Model Selection and Training

Several regression techniques were used in this work to estimate house values; these strategies were selected based on their individual merits and contributions to the overall performance of the model. The chosen techniques consist of the following: Gradient Boosting Regressor, XGBoost Regressor, Decision Tree Regressor, Random Forest Regressor, Ridge Regression, Lasso Regression, Elastic Net Regression, and Linear Regression. The following factors led to the selection of these models:

Diverse Capabilities: The integration of linear and non-linear models guarantees the collection of straightforward and intricate interactions within the data. While more sophisticated models like Decision Trees and ensemble techniques successfully manage non-linear relationships, linear models like Linear Regression offer a baseline and findings that are easy to understand models (Montgomery, Peck, & Vining, 2021).

Regularization Techniques: The inclusion of Ridge, Lasso, and Elastic Net regressions stems from their capacity to provide regularization, which lessens overfitting and enhances the model's generalizability. These methods also help with feature selection, which improves the interpretability and performance of the model (Zou, 2005).

Ensemble Methods: Models like Random Forest and Gradient Boosting Regressor leverage the power of ensemble learning to improve prediction accuracy and robustness. These methods reduce variance and bias by combining multiple weak learners to form a strong predictive model (Breiman, 2001).

Advanced Optimization: XGBoost is incorporated for its advanced optimization techniques and efficiency in handling large datasets. Its flexibility in hyperparameter tuning allows for fine-tuning the model to achieve superior performance (Chen, 2016).

By using a diverse set of regression methods, the study ensures a comprehensive evaluation of the dataset. This approach allows for comparing the effectiveness of different algorithms in predicting house values, ultimately identifying the best-performing model.

1) Linear Regression

This study uses a linear regression model as a baseline to forecast house prices.

linear regression helps predict how a dependent variable changes as the independent variable changes. The equation for linear regression looks like this:

$$y = \alpha + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \varepsilon.$$

Here, the betas represent the estimated parameters for the independent variables, y is the dependent variable. In this case, y represents the predicted house price, and xi represents the different features chosen.

The reason behind choosing linear regression is that it provides a straightforward approach to understanding the relationship between features and target variables. Also, it is useful for comparing the performance of more complex models (Montgomery, Peck, & Vining, 2021).

2) Ridge regression

Aso known as Tikhonov regularization, is a technique used to analyze multiple regression data that suffer from multicollinearity. When it occurs, least squares estimates are unbiased, but their variances are large, which may lead to overfitting. Ridge Regression adds a penalty to the regression coefficients, effectively shrinking them and reducing the model complexity.

$$\beta = (X^T X + \lambda I)^{(-1)}\{X^T\}\_y$$

where:

- X is the matrix of input features,
- y is the vector of target values,
- λ is the regularization parameter,
- I is the identity matrix.

The term λ added to *XTX*

The reason this method has been chosen it is because it ensures that the matrix is invertible, thus stabilizing the coefficient estimates to better generalizability on unseen data (Montgomery, Peck, & Vining, 2021).

3) Lasso regression

Least Absolute Shrinkage and Selection Operator is a type of linear regression that includes L1 regularization. This technique is particularly useful for feature selection and regularization, making it effective for models with many predictors, especially when some of those predictors are irrelevant or redundant.

The main idea behind Lasso Regression is to minimize the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. The reason this method has been chosen is because it encourages sparsity in the model by shrinking some coefficients to zero, effectively selecting a simpler and more interpretable model.

$$min\beta((\| y - \{Xb \|\}\_2^2 + \lambda \| \beta \| \_1)$$

where:

- y is the target variable,
- X is the feature matrix,
- β is the coefficients,
- λ is the regularization parameter.

By adjusting the λ parameter, Lasso can control the strength of the regularization. A larger λ value increases regularization, resulting in more coefficients being shrunk to zero. This can lead to models that are simpler and potentially less prone to overfitting, especially in high-dimensional datasets.

Lasso Regression is powerful for predictive modelling because it not only prevents overfitting but also performs automatic feature selection, enhancing the interpretability and performance of the model (Hastie, Tibshirani, & Wainwright, 2015).

4) Elastic Net Regression

A regularized regression method that linearly combines the penalties of Lasso (L1) and Ridge (L2) regression techniques. It aims to address the limitations of both methods by balancing the trade-off between feature selection (Lasso) and coefficient shrinkage (Ridge). This makes Elastic Net particularly effective when dealing with datasets with highly correlated predictors or when the number of predictors exceeds the number of observations. The Elastic Net objective function is given by:

$$min\beta(\| y - X\beta \| \_2^2 + \lambda\_1 \| \beta \| \_1 + \lambda\_2 \| \beta \| \_2^2)$$

where:

- y is the target variable,
- X is the feature matrix,
- β are the coefficients,
- λ 1 is the L1 regularization parameter (Lasso)

- λ 2 is the L2 regularization parameter (Ridge).

By incorporating both L1 and L2 penalties, Elastic Net encourages a sparse model with fewer features (like Lasso) while maintaining regularized coefficients to prevent overfitting (like Ridge). This dual approach helps in improving model performance and interpretability, especially in high-dimensional datasets the reason it was chosen is because it is effective when dealing with datasets with highly correlated predictors or when the number of predictors exceeds the number of observations. (Hastie et al., 2015).

5) Decision tree regressor

Decision tree regression is a non-parametric technique that predicts continuous values by building a tree-like model. Each internal node represents a feature, and branches represent decisions based on feature values. The terminal nodes (leaves) hold predicted target values. This model is constructed through recursive partitioning:

Feature Selection: The algorithm selects the most informative feature (e.g., using information gain) to split the data at each node, aiming to maximize the difference in the target variable between resulting child nodes.

- Splitting: The chosen feature is used to create a split rule, separating data points.
- Recursion: This splitting continues recursively until a stopping criterion (e.g., minimum data points) is met

The reason it was chosen is because it provides a non-parametric approach that is easy to visualize and interpret. (James, Witten, Hastie, & Tibshirani, 2021).

6) Random Forests

Random forests as shown in the figure 2 leverage decision trees by building an ensemble of them. Each tree is trained on a random subset of data (bootstrapping) and utilizes a random subset of features at each split. This injects randomness to improve model generalization:

- Diversity: Random subsets create diverse trees capturing different data aspects.
- Aggregation: Predictions from all trees are aggregated (e.g., averaged) for the final prediction, reducing variance and improving robustness, the reason it was chosen is because it reduces variance and improves robustness by aggregating predictions from multiple trees. (Hastie, Tibshirani, & Friedman, 2021).

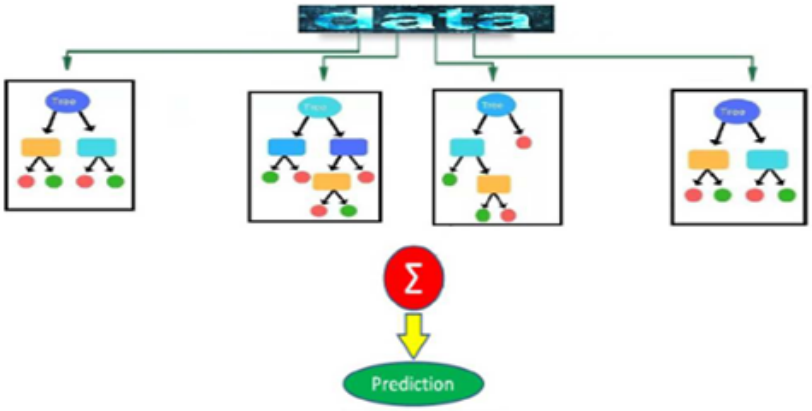

Figure 2. *Random Forest Model (Hastie, Tibshirani, & Friedman, 2021)*

7) Gradient Boost Regressor

It is an ensemble learning technique that builds a model in a stage-wise fashion from multiple weak learners, usually decision trees. It is built by iteratively fitting new models to the residual errors of the combined model built so far. It is powerful for predictive modelling because it combines the strengths of multiple models, reducing variance and bias. The main idea is to minimize the loss function by adding weak learners using a gradient descent-like procedure:

$$F\_m(x) = F\_(m - 1)(x) + h\_m(x)$$

where:

- *Fm(x)*: is the combined model after m iterations,
- *hm(x)*: is the weak learner added at the *m*-th iteration

(Friedman, 2001).

8) Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm belonging to the family of ensemble methods, specifically gradient boosting. It excels in both regression and classification tasks it builds an ensemble of decision trees, sequentially combining them to create a more robust and accurate model than any single

tree. Each tree learns from the errors (residuals) of the previous tree, focusing on improving the overall prediction accuracy. Formula:

$$f\_m(x) = f\_(m-1)(x) + \gamma * h\_m(x)$$

- *Fm(x)*: Prediction of the model at the m-th iteration (ensemble prediction) for a given input x.
- *Fm(m-1)(x)*: Prediction of the model at the previous (m-1) th iteration for input x.
- $\gamma(gamma)$: Learning rate (hyperparameter)
- *hm(x)*: Prediction of the m-th decision tree in the ensemble for input x.

XGBoost builds upon the core idea of gradient boosting by incorporating several advancements as represented in figure 3 below (Chen & Guestrin, 2016): Regularization techniques like L1 and L2 regularization that are employed to prevent overfitting, a common challenge where the model becomes too specific to the training data and performs poorly on unseen data. XGBoost leverages parallel processing and feature subsampling to handle large datasets efficiently, making it suitable for big data applications.

- Flexibility: A wide range of hyperparameters are available for tuning, enabling customization to optimize performance for specific tasks and data characteristics.
- Scalability: In essence, XGBoost combines the strengths of ensemble learning with regularization and efficient algorithms, making it a powerful tool for various machine-learning problems (Chen & Guestrin, 2016).

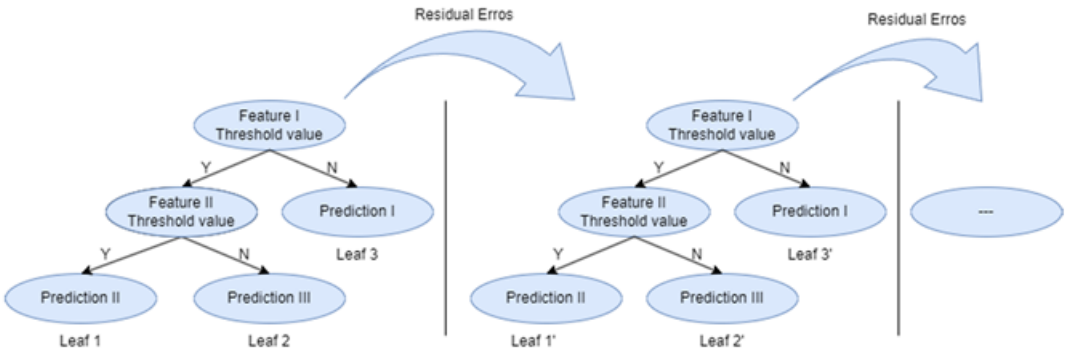

Figure 3. *XGB Model (Chen & Guestrin, 2016).*

5. RESULTS

In this study for an accurate house price prediction model, a comparison performance of various regression algorithms was conducted. a diverse range of models were applied, including:

Linear models: Elastic Net, Lasso, and Ridge regression are well-established techniques that provide interpretable results.

Decision tree and ensemble methods: Decision Tree Regressor and Random Forest Regressor offer flexibility in capturing complex relationships within the data.

Gradient boosting methods: Gradient Boosting Regressor and XGB Regressor are powerful ensemble methods known for their ability to handle non-linear relationships and complex datasets.

This study evaluated each model's performance on a separate test set using two key metrics:

R-squared (R^2): This metric measures the proportion of variance in the target variable (house prices) explained by the model's predictions. A higher R^2 signifies a better fit between the predicted and actual values.

Mean Absolute Error (MAE): This metric calculates the average absolute difference between the predicted and actual house prices. A lower MAE indicates a closer match between the predictions and the real values by analysing these metrics, the aim is to identify the model that delivers the most accurate and reliable house price predictions, ultimately assisting in informed decision-making.

Table 1. Test R^2 Scores

	Result Table			
	Model Name	***Test R²***	***Test MAE***	***Test RMSE***
1	ElasticNet	0.61	108,700.36	315735.74
2	Lasso	0.75	131,959.56	253678.56
3	DecisionTreeRegressor	0.74	114,648.88	256545.37
4	RandomForestRegressor	0.82	85,124.75	211055.52
5	LinearRegression	0.75	131,959.93	253678.55
6	Ridge	0.83	88,409.49	206209.01
7	GradientBoostingRegressor	0.82	83,888.93	218895.59
8	XGBRegressor	0.83	85,236.33	209536.04

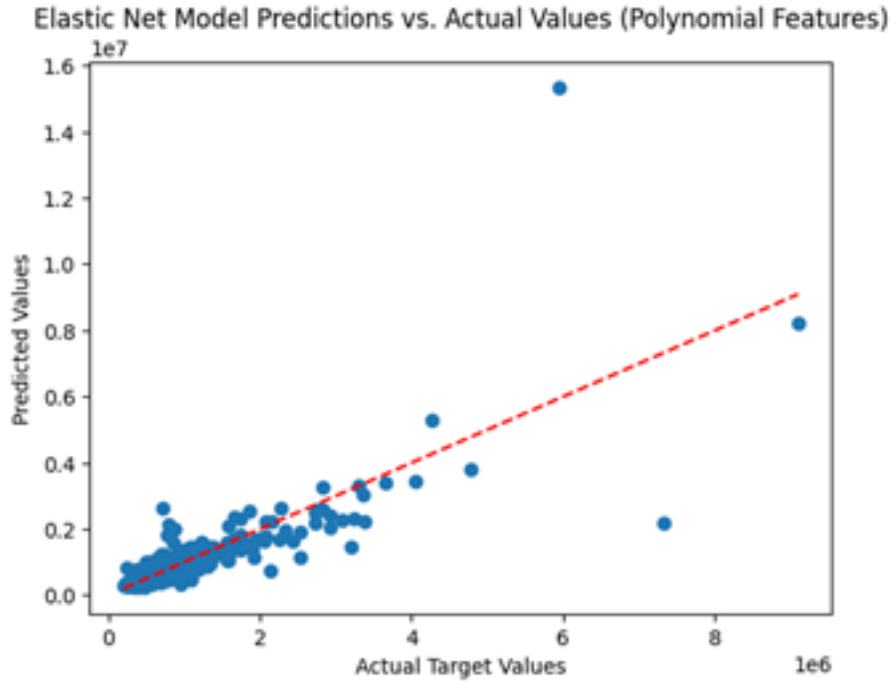

Figure 4. *Elastic Net*

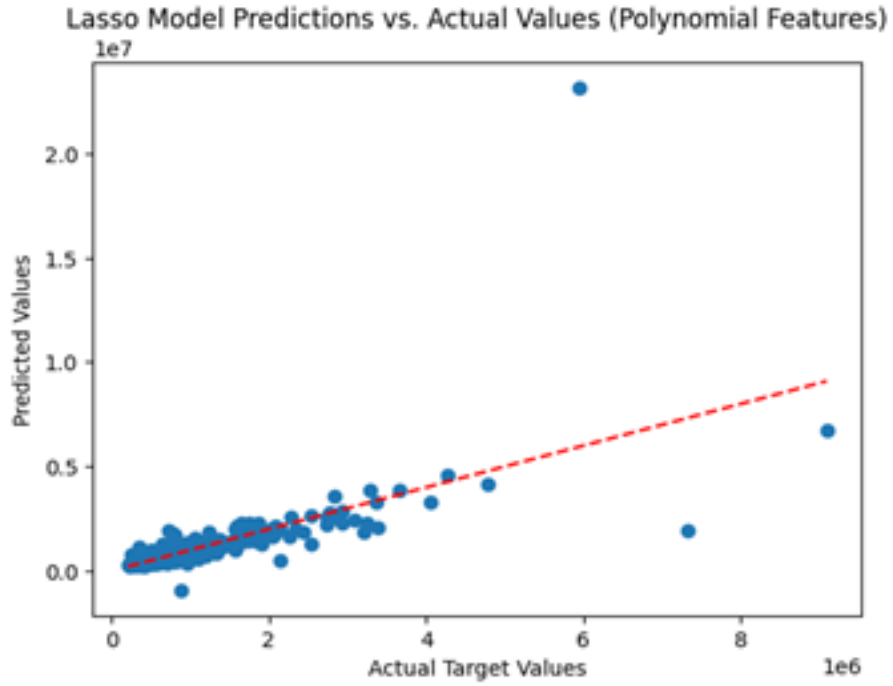

Figure 5. *Lasso Model*

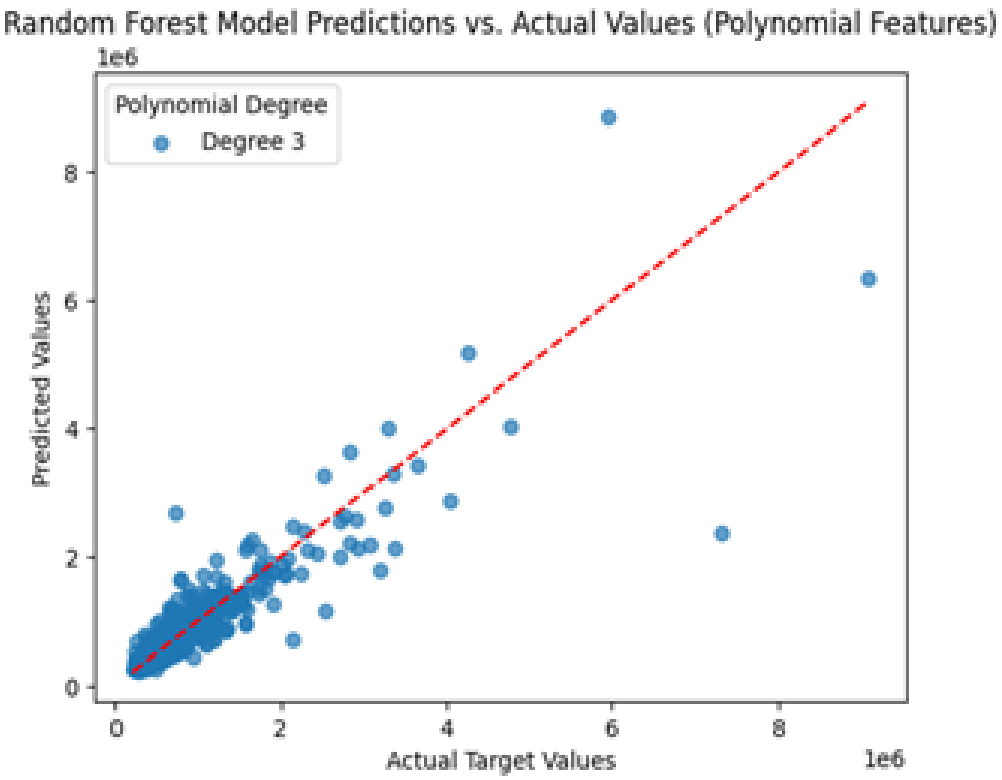

Figure 6. *Random Forest*

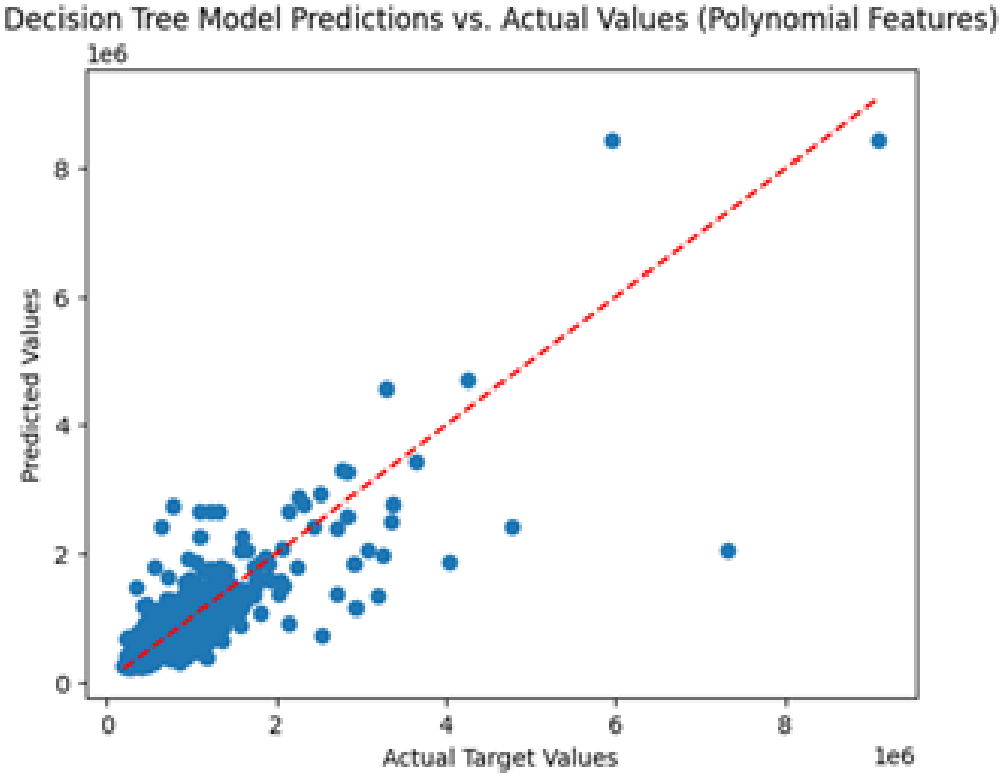

Figure 7. *Decision Tree*

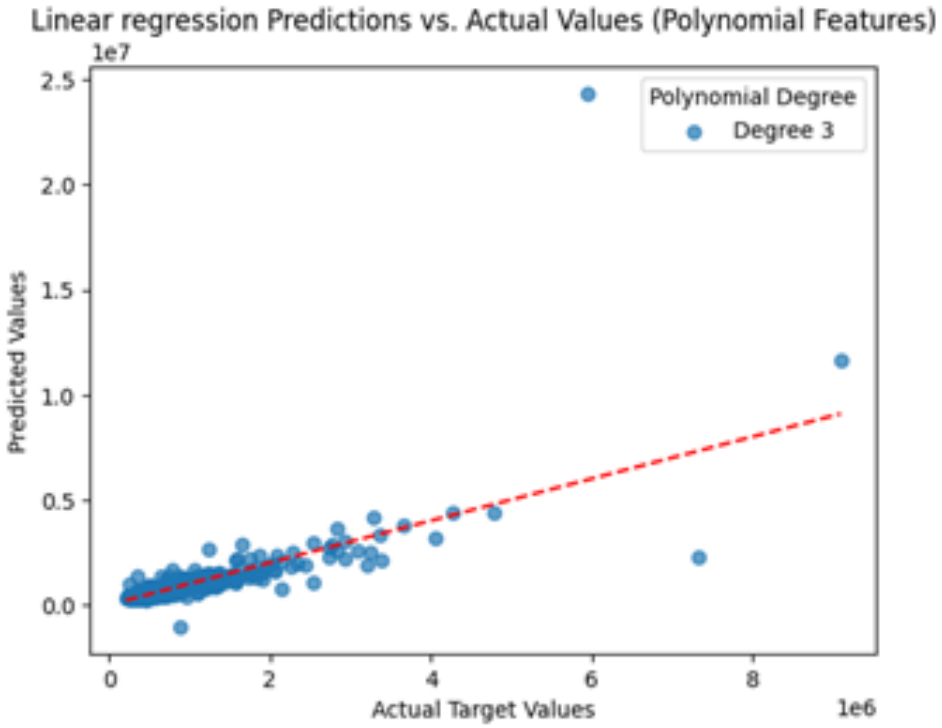

Figure 8. *Linear Regression*

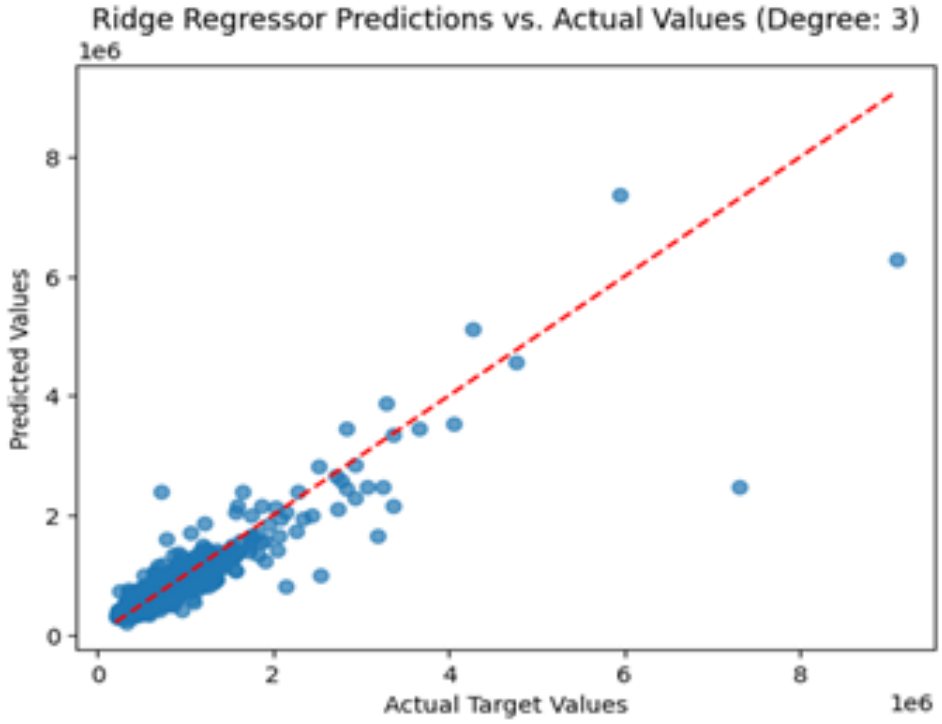

Figure 9. *Ridge Regressor*

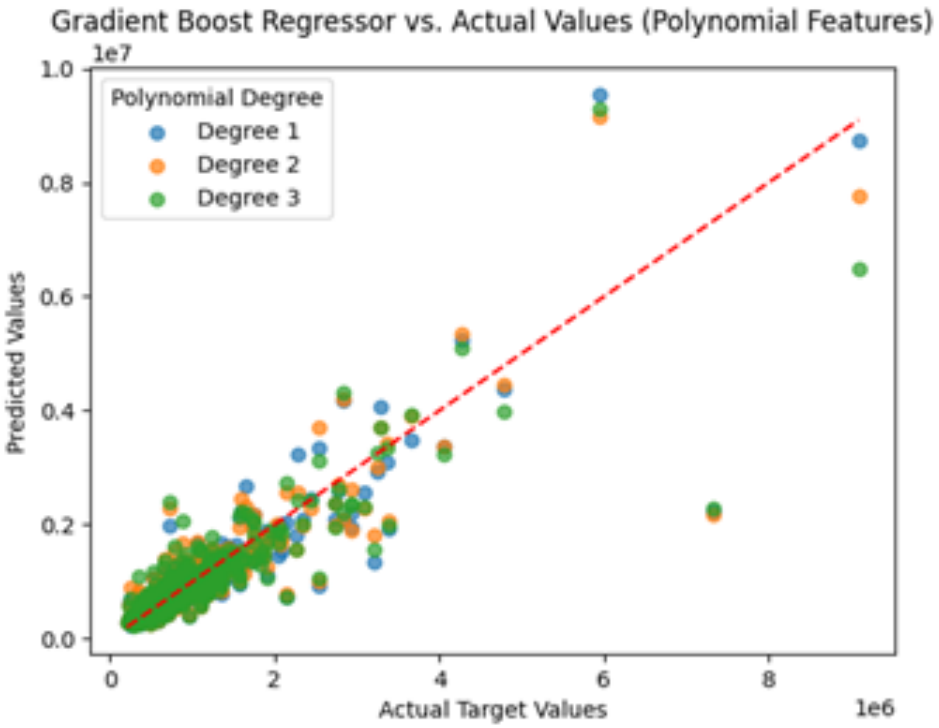

Figure 10. *Gradient Boost Regressor*

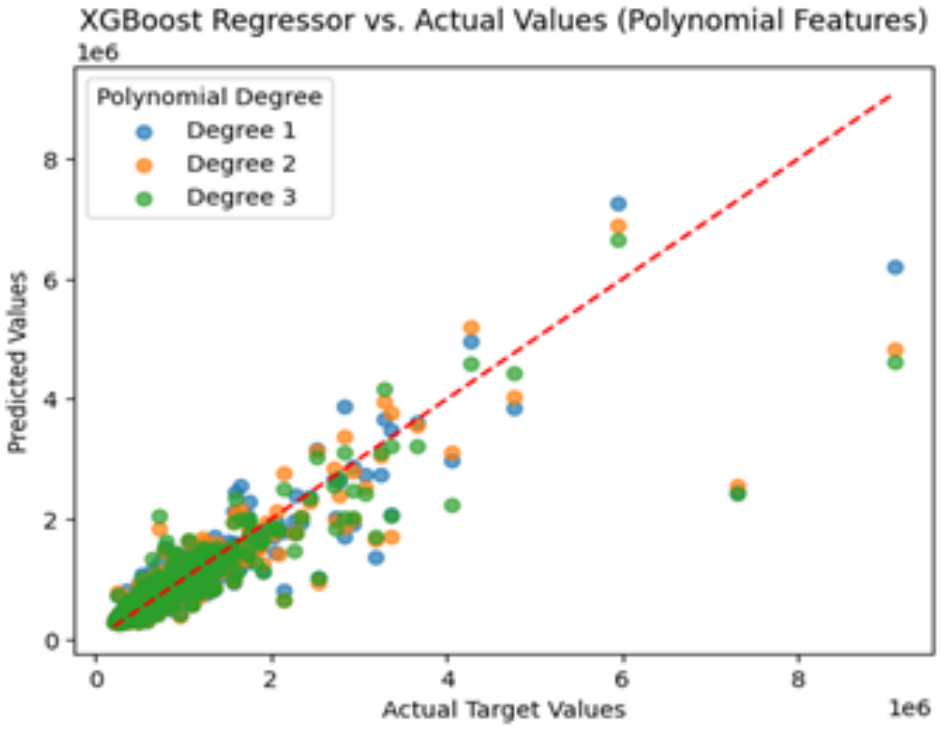

Figure 11. *XGBoost*

6. FINDINGS

In the quest to identify the most suitable model for house price prediction, an empirical evaluation was conducted comparing the performance of various regression algorithm as represented in the findings of this study represented in the result table 1 Test R^2 Scores. The candidate models included:

Linear models: ElasticNet, Lasso, and Ridge regression are established techniques known for their interpretability.

Decision tree and ensemble methods: Decision Tree Regressor and Random Forest Regressor offer flexibility in capturing complex relationships within the data.

Gradient boosting methods: Gradient Boosting Regressor and XGB Regressor are powerful ensemble methods known for their ability to handle non-linear relationships and complex datasets.

Separate test set for model evaluation, ensuring the models were not biased towards the training data. Two key metrics were used to assess performance:

R-squared (R^2): This metric measures the proportion of variance in the target variable (house prices) explained by the model's predictions. A higher R^2 signifies a better fit between the predicted and actual values.

Mean Absolute Error (MAE): This metric calculates the average absolute difference between the predicted and actual house prices. A lower MAE indicates a closer match between the predictions and the real values.

Random Forest Regressor and XGB Regressor achieved the highest R^2 scores (0.82 and 0.83, respectively), indicating a strong correlation between their predictions and the actual house prices.

Random Forest Regressor and XGB Regressor also achieved relatively low MAE values (85,124.75 and 83,888.93 respectively), suggesting a close match between the predicted and actual house prices in terms of absolute difference.

While Ridge regression achieved a comparable R^2 score to the Random Forest Regressor and XGB Regressor, its MAE value was slightly higher.

One notable aspect of this study is the relatively small number of features used compared to other literature. Despite the limited feature set, the models still produced highly accurate predictions. This demonstrates the efficiency and robustness of the Random Forest Regressor and XGB Regressor models, which can achieve strong predictive performance even with fewer input variables. This efficiency is particularly advantageous for real-world applications where gathering extensive data can be time-consuming and costly.

Figures 4 through 11 display the actual vs. predicted values for various models helping visually represent the accuracy of each model, they include, ElasticNet, Lasso Regression, Random Forest Regressor, Decision Tree Regressor, Linear Regression, Ridge Regressor, Gradient Boosting Regressor, and XGB Regressor. These figures were generated through machine learning techniques and visualized using Matplotlib in Jupyter Notebook.

A. Comparison and Analysis

Five studies were taken into consideration for comparison presented in the table 2 Comparison Table which show the Test R2 Score the Test MAE and the Test RMSE of each model used in each study the following studies taken into consideration are:

- Wang, 2018
- Chen Li, 2023
- Madhuri 2019
- Hernes et al., 2024
- Soltani et al., 2022

1) ***Strengths:***

a) XGBRegressor: In this study, the R^2 (0.83) is competitive, though slightly lower than the highest R^2 in the Chenxi Li study (0.888).

The XGBRegressor in this study has one of the highest R-squared values across all studies, indicating strong performance even though less data and features were used.

b) Ridge Regression:

The R^2 in this study (0.83) outperforms the Ridge Regression in the comparative study by CH. Raga Madhuri (0.732164).

The Ridge model in this study has a high R-squared value, showing its robustness.

c) Random Forest Regressor:

The R^2 in this study (0.82) is higher than in the Changchun Wang and Hui Wu study (0.701310346391) and close to the Chenxi Li study (0.878).

This indicates that Random Forest model is highly effective in this study as well.

d) Gradient Boosting Regressor:

The R^2 in this study (0.82) is strong and comparable to the Gradient Boosting Regression in CH. Raga Madhuri's study (0.9177022).

The Gradient Boosting model performs well, indicating its capability to handle non-linear relationships.

e) Gradient Boosting, Random Forest

Hernes et al. achieved 98% precision in predicting housing prices using Gradient Boosting, Random Forest. This precision reflects the use of ensemble methods that are well-suited to capturing non-linear relationships and ensuring robust performance.

f) Gradient Boosting

Soltani et al. incorporated a spatiotemporal lag variable to improve prediction accuracy, achieving an R^2 of 0.896 with Gradient Boosting. RMSE and MAE metrics for Soltani et al. showed improvements when the spatiotemporal lag variable was included, with MAE reduced to 0.058 and RMSE to 0.086 on the training set.

2) ***Weaknesses:***

a) ElasticNet:

The R^2 in this study (0.61) is lower compared to the Elastic Net Regression in CH. Raga Madhuri's study (0.665228).

This suggests that ElasticNet might not be the best fit for this dataset.

b) Linear Regression:

The R^2 in this study (0.75) is higher than the values in the other studies (0.539887986037 in Changchun Wang and Hui Wu, 0.706 in Chenxi Li, and 0.732072 in CH. Raga Madhuri).

However, it still doesn't match the best-performing models like XGBRegressor or Gradient Boosting but because of the feature selection and the standardization used it was able to get a better result than other studies.

c) Lasso Regression:

The R^2 in this study (0.75) is higher than in the CH. Raga Madhuri study (0.732072) but still lower compared to your top models like XGBRegressor and Ridge.

d) Decision Tree Regressor:

The R^2 in this study (0.73) is comparable but not superior to the other models. This suggests it may be less effective with less data and less feature selection compared to ensemble methods like Random Forest and Gradient Boosting.

e) Simple Linear Regression

Hernes et al. achieved 60.7% accuracy using their Simple Linear Regression technique while comparing to this study which had 75% and additionally, the study of Hernes et al. relied on web-scraped data for real-time updates, which provided dynamic insights but posed risks of instability due to changes in website structures. This study's reliance on pre-existing datasets offers stability and consistency. While Hernes et al. focused on a single market (Wroclaw primary real estate), this study's broader range of datasets and methods underscores its adaptability to diverse scenarios without losing predictive accuracy.

f) Decision Tree

Soltani et al.'s results concerning the decision tree were much less than this study with R2 of 0.579 compared to ours which was 0.73 even though it had relied on a 32-year dataset of 428,000 records that provided a historical depth that can also lead to legacy biases. This study focuses on fewer features and more recent data, achieving predictions that are both accurate and timely. Soltani et al.'s preprocessing efforts included handling extensive spatiotemporal dependencies, which increased computational demands. This study's streamlined approach reduces computational overhead while maintaining high accuracy, making it more practical for real-time applications.

This study's strengths lie in the performance of the XGBRegressor, Ridge Regression, and Gradient Boosting Regressor, which show high R-squared values, indicating strong predictive power. Random Forest Regressor also performs well but slightly below the top models in other studies. Weaknesses include ElasticNet and Linear Regression, which have lower R-squared values, indicating they might not capture the complexity of the data as effectively as other models.

The models in this study generally perform well, with certain models like XGBRegressor and Ridge standing out. Focusing on improving models like ElasticNet and exploring ensemble methods further could enhance the study's predictive accuracy.

The accuracy strength of this study lies in the polynomial features when they were created and normalized as part of feature engineering. The data was expanded to include polynomial degrees of 1, 2, and 3, capturing non-linear patterns. Following this expansion, the features were normalized using Standard Scaler to achieve a mean of zero and a standard deviation of one, enhancing the performance and convergence of the regression models.

3) Comparison Table

Table 2. Comparison Table

Study	*Algorithm*	*R^2 Score*	*MAE*	*RMSE*	*MSE*
Madhuri, 2019	Multiple Linear Regression	0.732072	-	48.88446	391,875,744
	Ridge Regression	0.732164	29.73141	51699	391,740,496
	LASSO Regression	0.732072	34.32263	46466	391,875,537
	Elastic Net Regression	0.665228	85.00798	76781	489,642,930
	AdaBoost Regression	0.7801099	79.94242	32161481	-
	Gradient Boosting	0.9177022	88.27804	109,713	179,336
Chen Li, 2023	Linear Regression	0.706	-	210,649.771	44,373,326,009.81
	Random Forest	0.878	-	136,170.257	18,542,338,996.45
	Neural Network	0.846	-	143,075.123	22,825,808,222.92
	XGBoost	0.888	-	130,281.363	16,973,233,719.78
Wang, 2018	Random Forests	0.701310	-	352.065	-
	Linear Regression	0.539887	-	381.280	-
Soltani et al., 2022	Decision Tree (ST Lag)	0.579	0.106	0.135	-
	Decision Tree (+ST Lag)	0.797	0.064	0.094	-
	Gradient-Boosted Tree (ST)	0.674	0.084	0.101	-
	Gradient-Boosted (+ST Lag)	0.896	0.058	0.086	-
	Random Forest (ST Lag)	0.662	0.086	0.109	-
	Random Forest (+ST Lag)	0.875	0.059	0.087	-
Hernes et al., 2024	Simple Linear Regression	0.607	-	-	101,789.11
	Gradient Boosting	0.989	-	-	5,173.12
	Multiple Linear Regression	0.912	-	-	41,404.69
	LASSO	0.912	-	-	41,405.14
	Random Forest	0.986	-	-	5,493.95
This Study	ElasticNet	0.61	108,700.36	315735.74	
	Lasso	0.75	131,959.56	253678.56	
	DecisionTreeRegressor	0.74	114,648.88	256545.37	
	RandomForestRegressor	0.82	85124.75	211055.52	
	Linear Regression	0.75	131,959.93	253678.55	
	Ridge	0.83	88,409.49	206209.01	
	GradientBoostingRegressor	0.82	83888.93	218895.59	
	XGBRegressor	0.83	85,236.33	209536.04	

7. CONCLUSION

This study evaluated the effectiveness of various regression algorithms for estimating house values, comparing models such as linear regression, decision trees, ensemble methods, and gradient-boosting techniques. The models were assessed on a separate test set using R-squared (R^2) and Mean Absolute Error (MAE) as performance metrics.

The results showed that the Random Forest Regressor and XGB Regressor models achieved the highest R^2 scores and relatively low MAE values, indicating a strong correlation between their predictions and actual house values with minimal absolute difference. While models like Ridge regression also performed well, Random Forest Regressor and XGB Regressor emerged as the top performers.

The choice between Random Forest Regressor and XGB Regressor should be guided by the specific needs of the application:

Interpretability: If model transparency and the ability to explain predictions to clients are important, Random Forest Regressor is preferable due to its interpretable decision tree structure.

Absolute Error Sensitivity: If minimizing the absolute difference between predicted and actual house values is crucial, the XGB Regressor is the better choice due to its slightly lower MAE.

The application of these regression models is particularly valuable for evaluating house prices in small estate markets such as North Cyprus. Accurate and reliable house value estimations can significantly enhance pricing strategies, preventing overpricing that can lead to properties remaining unsold for extended periods and avoiding under-pricing that results in financial losses. By providing precise value estimates, these models can aid real estate professionals in setting competitive listing prices, ensuring properties are bought and sold efficiently, and maintaining the financial health of the real estate market.

Real estate professionals benefit from enhanced pricing strategies and more informed client consultations, which contribute to greater market efficiency. Investors gain clearer insights for better decision-making through reliable value assessments. Meanwhile, accurate property valuations empower buyers, sellers, and homeowners to make well-informed choices regarding transactions and refinancing...

8. FUTURE WORK

These models have demonstrated the ability to anticipate house estimation values with great accuracy, thus future research in real estate markets will incorporate these models, and North Cyprus is one such market where such

research will be carried out. Future research could further enhance these models by including additional features like neighbourhood amenities, school quality, and property condition to capture a more comprehensive range of factors influencing house values and refine-tuning the hyperparameters for each algorithm to improve model performance. Applying the models to different regions to understand price variations across locations continuing to refine these models will contribute to the development of robust and precise house value estimation, ultimately leading to better decision-making in the housing market.

Peer Review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

ORCID IDs of the authors

Tarek Ghamrawi 0009-0008-9107-7375
Müesser Nat 0000-0002-1539-3586

REFERENCES

Ahtesham, M., Bawany, N., & Fatima, K. (2020). *House price prediction using machine learning algorithm - The case of Karachi City. Pakistan*, 1-5. doi:10.1109/ACIT50332.2020.9300074

Ali, S., Mohammad, H., Fatemeh, A., Christopher, J. P. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities*, 131, 103941. https://doi.org/10.1016/j.cities.2022.103941.

Binu, J. (2020). *A data analytics model for extended real estate comparative market analysis.* Pace University, New York.

Breiman, L. (2001). *Random forests.* Machine Learning, 45, 5-32. doi:10.1023/A:1010950718922

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi:10.1145/2939672.2939785

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine.* IMS 1999 Reitz Lecture. Modified March 15, 2000, April 19, 2001.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations.* CRC Press.

IBM. What is underfitting? Retrieved from https://www.ibm.com/cloud/learn/underfitting

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R (2nd ed.)*. Springer.

Li, C. (2023). *House price prediction using machine learning*. Proceedings of the 4th International Conference on Signal Processing and Machine Learning. School of International Education, GuangDong University of Technology, Guangzhou, China. doi:10.54254/2755-2721/53/20241426

Li, D. (2020, December 25). *Overcoming data scarcity and privacy challenges with synthetic data.* InfoQ. Retrieved from https://www.infoq.com/articles/overcoming-privacy-challenges-synthetic-data/#idp_register/

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). *House price prediction using regression techniques: A comparative study.* 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, pp. 1-5. doi:10.1109/ICSSS.2019.8882834

Marcin, H., Piotr, T., & Mateusz, S. (2024). Prediction of residential real estate price on primary market using machine learning. *Procedia Computer Science*, 246, 3142-3147. https://doi.org/10.1016/j.procs.2024.09.358.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis (6th ed.)*. Wiley.

Quang, T., Minh, N., Hy, D., & Bo, M. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science,* 174, 433-442. https://doi.org/10.1016/j.procs.2020.06.111.

Rana, V. S., Mondal, J., Sharma, A., & Kashyap, I. (2020). *House price prediction using optimal regression techniques*. 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, pp. 203-208. doi:10.1109/ICACCCN51052.2020.9362864

Thomas, D. (2023). *The importance of data in property valuation and the key role of comparative method.* doi:10.13140/RG.2.2.35313.86881

Wang, C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences*, 6(4).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320. doi:10.1111/j.1467-9868.2005.00503.

How cite this article

Ghamrawi, T., & Nat, M. (2024). House Value Estimation using Different Regression Machine Learning Techniques. *Acta Infologica, 8*(2), 245-259. https://doi.org/10.26650/acin.1543650