IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

Machine Learning Models in Predicting Mortgage Prices

Nishant Gadde[1], Avaneesh Mohapatra[2], Daksh Parikh[3], Shiva Uppaladinni[4], Lalit Nookella[5], Smrutirekha Panda[6]

Jordan High School; Fulshear, TX[1]

West Forsyth High School; Cumming, GA[2]

Adrian Wilcox High School; Santa Clara, CA[3]

Seven Lakes High School; Katy, TX[4]

Central Bucks School District South, PA[5]

Biju Patnaik University of Technology, Odisha, India[6]

Abstract: The following study concerns exploring the performance of multiple regression algorithms of machine learning in the context of house pricing, while attempting to enhance the precision and offering practical implications for the stakeholders in the real estate industry. Using dataset that is collected from the real estate platforms, property records and other fresh data obtained directly from the real estate agencies, models like Random Forest, Gradient Boosting Machines (GBM), XGBoost, Support Vector Regression (SVR) and Neural Networks are examined. It also entails carrying out massive data preprocessing, feature construction, and other computationally expensive steps such as tuning of hyperparameters for achieving high accuracy. The residual plots indicate the prediction accuracy of each of the 23 models of some levels and weakness in the various methods employed in the models. For example, Random Forest and XG Boost exhibit typical non-linear patterns to capture, but they have heteroscedasticity to some extent in residuals. On the other hand, standard models like the SVR with the linear kernel show some level of failure in dealing with the interleaved pattern between the data, resulting in systematic biases. Thus, it is crucial to choose a right model depending on the data set properties and certain market conditions are considered in the study. Thus, it is seen that this research adds to the literature on machine learning real estate by offering a step-by-step comparison of these five advanced regression techniques that will be useful in determining the effectiveness of such techniques in the prediction of housing prices. Acquired knowledge is expected to benefit, for instance, real estate agents, investors, and policy-makers towards increasing market transparency leading to efficiency.

Keywords: Mortgage prediction, GBM, SVR, XGBoost, Elastic Net

I. INTRODUCTION

The real estate market is a multifaceted and dynamic sector, influenced by an intricate interplay of economic, social, and demographic factors. Over time, the prediction of house prices has become a critical area of focus, given its profound impact on the financial well-being of individuals, the stability of markets, and the formulation of policy. Accurate house price prediction is not only valuable for potential buyers and sellers but also serves as a cornerstone for real estate investors, financial institutions, and policymakers who rely on these forecasts to make informed decisions about investments, lending, and regulatory actions. The ability to predict house prices accurately can enhance market transparency, reduce investment risks, and contribute to overall economic stability.

Real estate is one of the complex and evolving segment markets that depends on the economic, social, demographic factors system. It has become an increasing interest to forecast house prices as it has a far reaching implication on the individual, markets and policy making. The use of reliable house price expectation indicators goes beyond the futuristic expectations of a particular price by the buyer or the seller of a given house or property, but the outcome of such predictability is significant to real estate investors, financial institutions, and policy makers for investments, lending, and policy making bases respectively. The knowledge of accurate predetermining of house prices can increase the level of market openness, minimize risks of investments, and promote more stable development of the economy.

Historically, house price prediction can be resolved by applying standard statistical methods; among those of them the linear regression takes the leading position because of its simplicity and obvious interpretability. Linear models for

DOI: 10.17148/IJARCCE.2024.13902

instance makes predictions by using past historical data about the features of the properties and their respective prices with the belief that the two are normally proportional. Still, real estate markets are open and dynamic, and thus cannot be easily explained by simple and Ohmic relationships between the different factors like location, size, amenities, etc. Therefore, in many cases, the linear models are not quite adequate and their application results in lower accuracy to forecast and even to make wrong conclusions (Malpezzi, 2003; Sirmans, MacDonald, & Macpherson, 2006).

Over the last few years, with evolution of cloud, bigdata, ELT (Extract, Load and Transform) processes, the machine learning techniques have brought about significant leaps of improvement in the field of predictive analytics given that it provides a means by which to model non-linear relationships far more efficiently than statistical models (Panigrahy et. al, 2023). Researchers put a lot of emphasis on decision trees, random forest, support vector machines, gradient boosting machines, neural networks and these show promising results in the real estate domain. These models have the ability to handle large data volumes, discern subtle patterns and learn how to generalize, depending on the complexity of the market characteristics, which makes them perfect for the house price prediction tasks (Bokhari & Geltner, 2011; Antipov & Pokryshevskaya, 2012).

For example, decision trees offer a relaxed method of analyzing non-linear associations by separating the data into subsets depending on the feature values, thus making it possible a better understanding of the way in which different variables affect house prices. Random forests, which is a technique based on decision trees, also improve the accuracy of the model because it uses more than one tree and combines the result to minimize the chance of overfitting when tested on unseen data. Support vector regression (SVR) further generalizes these useful properties to working in high-dimensional spaces, where it performs interactions between different variables well, but has some hyperparameters that must be carefully selected (Smola & Schölkopf, 2004).

Gradient boosting machines (GBMs), including modern variations, XGBoost, LightGBM, and CatBoost, are considered some of the most effective tools in the contemporary statistics arsenal for predictive modeling. One of these is in building an ensemble of weak learners, usually the decision trees, with the use of a loss function that enables them to bring out even the most detailed features of the data. Neural networks, especially deep learning models, is yet another area in house price prediction because the architecture allows for learning extremely complex nonlinear functions through multiple layers of interconnected neurons. These models are computationally intensive and need large datasets for training but their capability of discovering rich pattern in the data makes them ideal for this use (Goodfellow, Bengio, & Courville, 2016).

The purpose of this research is to review and analyze these different forms of machine learning regression techniques with a view of identifying which of the procedures offers the best and most profitable estimation of house prices. In this research, several quantitative models of real estate investments will be compared, with the goal of contributing to the existing literature by providing further information about the advantages and disadvantages of each approach in order to demand communication of real property markets' constituents. The implication of this study to the current knowledge on the use of machine learning to real estate will greatly help in the development of more ideas for future research and practice (Chaudhuri & Yulei, 2020; Mohammed, 2024). Ultimately, this research endeavors to enhance the tools available for predicting house prices, thereby improving decision-making processes for buyers, sellers, investors, and policymakers alike.

II. LITERATURE REVIEW

Traditional Methods of House Price Prediction

Thus, the focus on house price prediction has been on the side of statistical models with linear regression models being most widely used because of the increased ease of interpretation. A crucial assumption of models like the linear regression is that there is a linear relationship between properties' characteristics and house prices or rent charges which makes the models easy to apply and interpret. However, the assumption that the relationship between the variables is linear is a major drawback as real estate markets consist of systems that call for non-linear models as most of the systems are inter-dependent. For example, reduced effects of proximity to amenities on valuations could depend not only on size and age of the property, things linear models poorly capture. Therefore, linear regression is useful only as a reference point; yet, the results of its application in dealing with real-life problems and phenomena characterized by non-linear dependencies are frequently far from satisfactory (Malpezzi, 2003; Sirmans, MacDonald, & Macpherson, 2006).

Introduction to Machine Learning in Real Estate

With machine learning, a set of sophisticated methodologies that are useful in capturing the structures of actuality that prevail in the data set for instance data containing real estate data. Compared to traditional statistics, machine learning

DOI: 10.17148/IJARCCE.2024.13902

gains a larger number of samples and more complex structures of algorithms to capture the latent structure of variables. These models can easily accommodate features of the real estate markets; enhanced forecast performance and enhanced understanding of factors behind house prices (Bokhari & Geltner, 2011).

Overview of Machine Learning Regression Techniques

This study employs and compares eleven different machine learning regression techniques, each with its unique strengths and challenges. The following is an overview of these techniques and their relevance to house price prediction:

Linear Regression:

Linear regression is inclusive of being a basic and fundamental method used in the predictive modeling space. It thus provides a yardstick against which to gauge the performance of other more sophisticated models. When used in the context of machine learning, linear regression is enriched by the utilization of novel techniques like Ridge and Lasso used to bring in generalization by avoiding excessive focus on coefficients that can lead to overfitting (Zhang, 2016).

Ridge Regression:

Ridge regression, also known as Tikhonov regularization, adds an L2 penalty to the loss function, which discourages the model from assigning too much importance to any single feature. This regularization helps to stabilize the model when there is multicollinearity (high correlation between features) and improves prediction accuracy, particularly when dealing with complex, high-dimensional data (Hoerl & Kennard, 1970)

Lasso Regression:

Lasso regression, which incorporates an L1 penalty into the loss function, not only helps prevent overfitting but also performs feature selection by driving the coefficients of less important features to zero. This results in simpler, more interpretable models, making Lasso particularly useful in situations where feature selection is crucial (Tibshirani, 1996).

Elastic Net:

Elastic Net combines the penalties of both Ridge and Lasso regression (L1 and L2 penalties), providing a balance between the two. It is particularly effective in situations where there are multiple correlated features, as it retains the benefits of both methods, offering a more robust and flexible approach to regularized linear modeling (Zou & Hastie, 2005).

Decision Trees (CART):

Classification and Regression Trees (CART) provide a non-parametric approach to modeling that splits the data into subsets based on feature values. Decision trees are highly interpretable and can capture non-linear relationships between variables, making them a powerful tool for real estate prediction. However, they are prone to overfitting, especially with noisy data, which can limit their predictive accuracy on unseen data (Breiman, 1984).

Random Forest:

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. By aggregating the results of multiple trees, Random Forest reduces overfitting and improves generalization, making it one of the most widely used algorithms for house price prediction (Breiman, 2001).

Support Vector Regression (SVR):

SVR is an extension of Support Vector Machines (SVM) for regression tasks. It uses kernel functions to project data into higher-dimensional spaces, where linear regression can then be applied. This allows SVR to handle non-linear relationships effectively. However, the model's performance is highly dependent on the choice of kernel and the tuning of hyperparameters, such as the regularization parameter (C) and the epsilon (ε) threshold, which defines a margin of tolerance for error (Smola & Schölkopf, 2004).

Gradient Boosting Machines (GBM):

GBM is a powerful ensemble method that builds models sequentially, with each new model attempting to correct the errors made by the previous ones. This method includes various implementations like XGBoost, LightGBM, and CatBoost, which are known for their efficiency and performance. XGBoost, for instance, incorporates advanced regularization to prevent overfitting, while LightGBM is optimized for speed and memory efficiency, particularly with large datasets.

CatBoost, on the other hand, is designed to handle categorical features effectively, reducing the need for extensive preprocessing (Friedman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Dorogush et al., 2018; Janamolla & Syed, 2024).

DOI: 10.17148/IJARCCE.2024.13902

Neural Networks:
Neural networks, particularly deep learning models, offer a flexible and powerful approach to modeling complex, non-linear relationships. Feedforward neural networks are the simplest type of neural network architecture, where data flows in one direction from input to output layers. More complex architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are used for more sophisticated tasks but require large amounts of data and computational resources. In house price prediction, neural networks can capture intricate patterns that other models might miss, but they also risk overfitting, especially with smaller datasets (Goodfellow et. al, 2016).

Bagging Regressor:
Bagging, or Bootstrap Aggregating, is an ensemble method that improves the stability and accuracy of machine learning algorithms by training multiple models on random subsets of the data and aggregating their predictions (Mohammed, 2024a). A Bagging Regressor often uses decision trees as base estimators, and by averaging their predictions, it reduces variance and helps prevent overfitting, providing more reliable predictions in real estate applications (Breiman, 1996).

CatBoost:
CatBoost is another advanced gradient boosting method particularly adept at handling categorical variables directly, without the need for extensive preprocessing such as one-hot encoding. It builds upon the general gradient boosting framework with innovations that improve speed and accuracy, making it a strong contender for predictive tasks involving structured data like real estate transactions (Dorogush et al., 2018; Syed & Janamolla, 2024).

Comparative Analysis of Machine Learning Techniques
Comparative studies of these machine learning techniques have shown that no single model consistently outperforms others across all scenarios. The performance of each model depends on the specific characteristics of the dataset, including the nature of the features and the complexity of the relationships within the data. For instance, while neural networks may excel in capturing complex patterns, they require significant computational resources and large datasets to achieve their full potential. In contrast, models like Random Forest and XGBoost provide a good balance of performance and interpretability, making them suitable for a wide range of applications in house price prediction (Alaa & Schaar, 2018).

This study contributes to the growing body of literature on machine learning applications in real estate by providing a comprehensive comparison of these techniques. By evaluating the performance of each model using a robust dataset, this research aims to identify the most effective approach for predicting house prices, offering valuable insights for real estate professionals, investors, and policymakers.

III. METHODOLOGY

Preprocessing of Data
Before applying machine learning algorithms to predict house prices, the raw data extracted from Zillow underwent a comprehensive preprocessing procedure to ensure that it was clean, consistent, and suitable for accurate modeling. The first step in this process was data cleaning, where the dataset was meticulously examined for any missing values or inconsistencies. Missing data can introduce biases and inaccuracies in model predictions, so it was essential to address this issue thoroughly. For numerical features, missing values were handled using imputation techniques, where the mean or median values were used to fill in the gaps. This approach helped maintain the integrity of the dataset without introducing artificial distortions.

Following the cleaning process, normalization was performed to standardize the data. This step was particularly crucial because the features in the dataset varied widely in scale. For example, the size of a property might range from a few hundred square feet to several thousand, while other features like the number of rooms are on a much smaller scale. Without normalization, machine learning algorithms like Support Vector Regression (SVR) and neural networks, which are sensitive to the scale of input data, might assign undue importance to certain features simply because they have larger numerical values. Normalization techniques such as Z-score normalization or Min-Max scaling were applied to bring all features into a comparable range, ensuring that each feature contributed equally to the model's predictions.

In addition to normalization, feature selection and engineering were also critical components of the preprocessing phase. Feature selection involved identifying the most relevant variables that would have the greatest impact on predicting house prices, such as location, property size, and the number of rooms.
This step was guided by domain knowledge as well as statistical techniques like correlation analysis. Moreover, feature engineering was employed to create new variables that could provide additional insights. For example, the price per square foot was derived from existing data, offering a normalized measure of property value that is independent of the

DOI: 10.17148/IJARCCE.2024.13902

overall size. By carefully selecting and engineering features, the study aimed to enhance the predictive power of the models and capture the underlying patterns in the data more effectively.

Machine Learning Algorithms

The study employed eleven different machine learning algorithms to model the relationship between property features and house prices, each offering unique advantages and challenges in predictive modeling. The selection of these algorithms was guided by their proven effectiveness in handling various types of data and relationships, ranging from linear to highly non-linear patterns.

The first algorithm used was **Linear Regression**, a fundamental technique that serves as a baseline for comparison with more complex models. Linear regression assumes a linear relationship between the independent variables (features) and the dependent variable (house prices). Despite its simplicity, it is often inadequate for capturing the complexities inherent in real estate data, which led to the exploration of more sophisticated methods.

To address the limitations of linear regression, the study also employed **Ridge Regression** and **Lasso Regression**, both of which incorporate regularization techniques to prevent overfitting. Ridge Regression adds an L2 penalty to the loss function, which discourages large coefficients and helps stabilize the model in the presence of multicollinearity, where independent variables are highly correlated. On the other hand, Lasso Regression introduces an L1 penalty, which not only prevents overfitting but also performs feature selection by driving less important feature coefficients to zero. This dual role of Lasso makes it particularly useful in high-dimensional datasets where feature selection is critical.

Elastic Net was another algorithm used in the study, combining the strengths of both Ridge and Lasso regression. By balancing the L1 and L2 penalties, Elastic Net can handle situations where multiple features are correlated, providing a more robust model that benefits from both regularization techniques. This makes Elastic Net a versatile choice for datasets where the relationships between variables are complex and not purely linear.

Decision Trees, implemented through the Classification and Regression Trees (CART) algorithm, provided a non-linear approach to modeling. Decision trees split the data into subsets based on feature values, capturing interactions between variables that linear models might miss. However, decision trees are prone to overfitting, especially when dealing with noisy data, which is why ensemble methods like **Random Forest** were also employed. Random Forest builds multiple decision trees on different subsets of the data and averages their predictions, significantly reducing the risk of overfitting and improving model accuracy.

For capturing more complex non-linear relationships, **Support Vector Regression (SVR)** was used. SVR extends the principles of Support Vector Machines (SVM) to regression tasks, using kernel functions to project the data into higher-dimensional spaces where a linear relationship might exist. The flexibility of choosing different kernels (linear, polynomial, radial basis function) allows SVR to model intricate patterns, though it requires careful tuning of hyperparameters.

The study also explored **Gradient Boosting Machines (GBM)**, with implementations like XGBoost, LightGBM, and CatBoost, which are among the most powerful tools for predictive modeling. These models build sequential ensembles where each new model corrects the errors of its predecessors, resulting in highly accurate predictions. XGBoost, known for its speed and performance, incorporates advanced regularization to avoid overfitting, while LightGBM is optimized for large datasets and high-dimensional data. CatBoost, in particular, was selected for its ability to handle categorical variables effectively, a common challenge in real estate data.

Neural Networks, especially deep learning models, were used to capture the most complex patterns in the data. Neural networks consist of multiple layers of interconnected nodes (neurons) that process data in a manner inspired by the human brain. These models are capable of modeling non-linear relationships and interactions at a level of complexity that traditional methods cannot achieve. However, they require substantial computational resources and large amounts of data to perform well.

Finally, the study utilized **Bagging Regressor**, an ensemble method that builds multiple models on different subsets of the data and averages their predictions. This method helps to reduce variance and improve model stability, making it particularly useful when dealing with high-variance models like decision trees.

Model.py for Finding Correlations and Patterns

The study's machine learning models were implemented and managed within a Python script named model.py. This script served as the backbone for the entire modeling process, orchestrating the training, evaluation, and validation of the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

various algorithms. The script was designed to handle the entire workflow, starting from loading the preprocessed data to tuning hyperparameters and evaluating model performance.

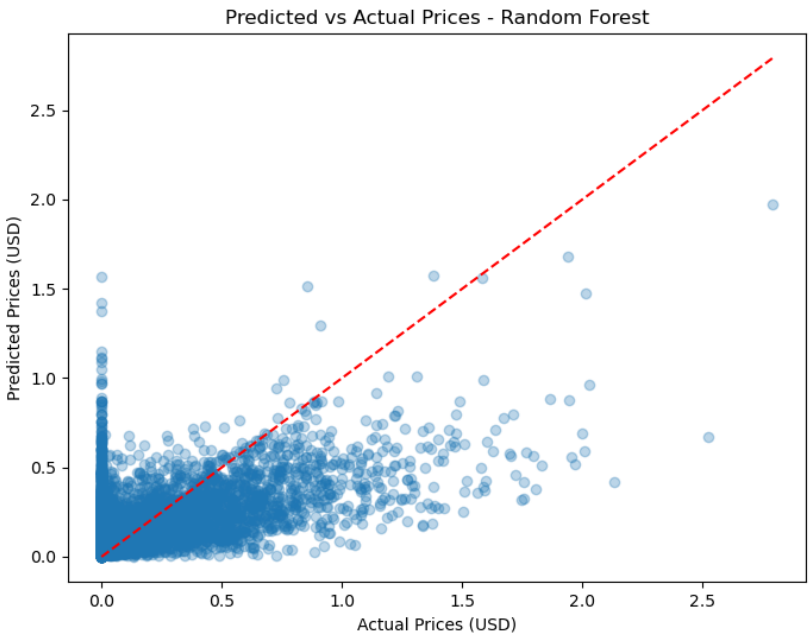
Within model.py, the models were instantiated and configured according to the specific requirements of each algorithm. For example, hyperparameter tuning was automated through grid search or random search, allowing the script to systematically explore a range of values for key parameters such as the number of trees in Random Forest or the learning rate in Gradient Boosting Machines. This automated approach ensured that each model was optimized for performance before being evaluated.

The script also included functionality for cross-validation, which was essential for assessing the generalizability of the models. Cross-validation techniques like k-fold cross-validation were used to split the data into training and testing sets multiple times, providing a robust measure of model performance across different subsets of the data. This approach helped in identifying correlations and patterns in the data, ensuring that the models could generalize well to unseen data. In terms of output, model.py generated a range of metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) values, for each model. These metrics were critical for comparing the performance of the different algorithms and identifying the best model for predicting house prices. Additionally, the script included functions for visualizing residual plots and other diagnostic tools, which provided insights into the strengths and weaknesses of each model. By systematically applying and evaluating these machine learning techniques, model.py played a crucial role in uncovering the correlations and patterns that drive house prices in the real estate market.

IV. RESULTS

The results of the study are visualized through predicted vs. actual price plots for various machine learning models, showcasing the effectiveness of each algorithm in predicting house prices. These plots serve as a crucial diagnostic tool, revealing how well each model captured the underlying patterns in the data.

The **Random Forest** model demonstrated strong predictive performance, as evidenced by the predicted vs. actual price plot (Figure 1). The points in the plot are closely clustered around the diagonal line, indicating that the model was able to accurately predict house prices across a range of values. This result highlights the model's ability to handle non-linear relationships and interactions between features, making it a robust choice for this type of predictive task.



In contrast, the **Elastic Net** model (Figure 2) showed less accuracy in its predictions. The plot reveals a significant deviation from the diagonal line, particularly at higher price ranges, suggesting that the model struggled to capture the full complexity of the data. Despite its regularization capabilities, which help prevent overfitting by shrinking less

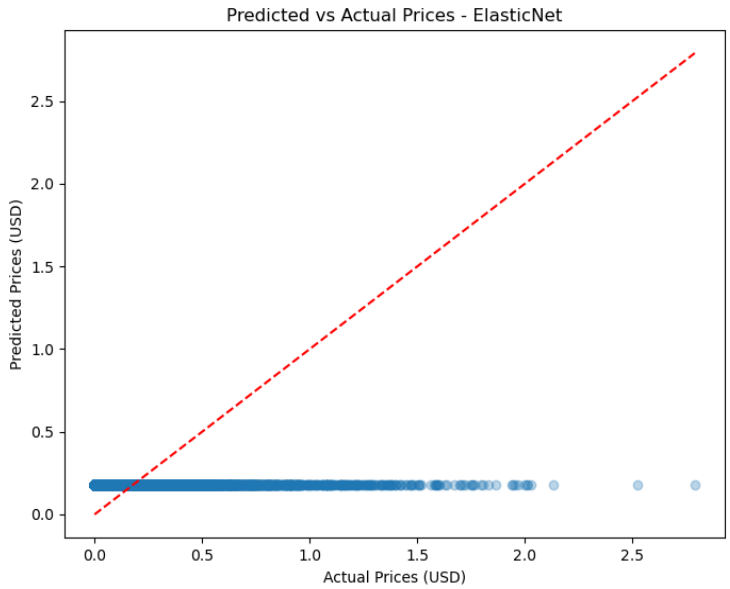
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

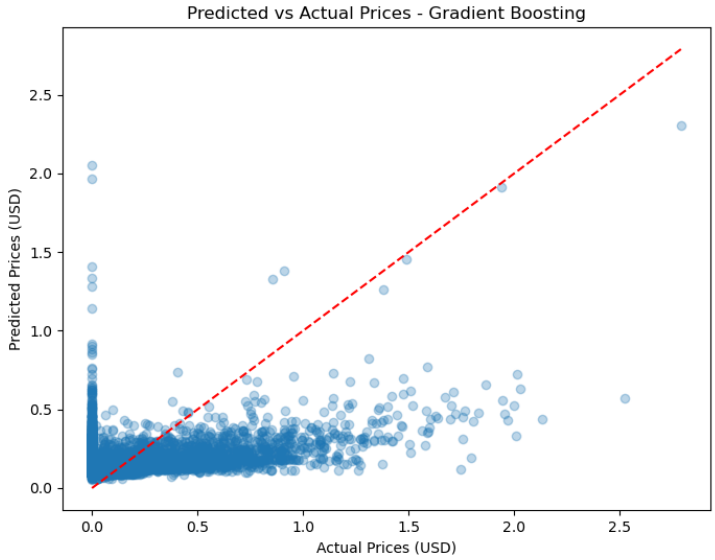
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

important feature coefficients, the Elastic Net model appears to be less effective compared to more advanced ensemble methods.



The **Gradient Boosting** model (Figure 3), which includes variants like XGBoost and LightGBM, performed similarly well to Random Forest. The plot shows a strong alignment with the diagonal line, indicating that the model was able to predict house prices with high accuracy. The Gradient Boosting model's ability to iteratively correct errors from previous models likely contributed to this strong performance, making it one of the top contenders in this study.



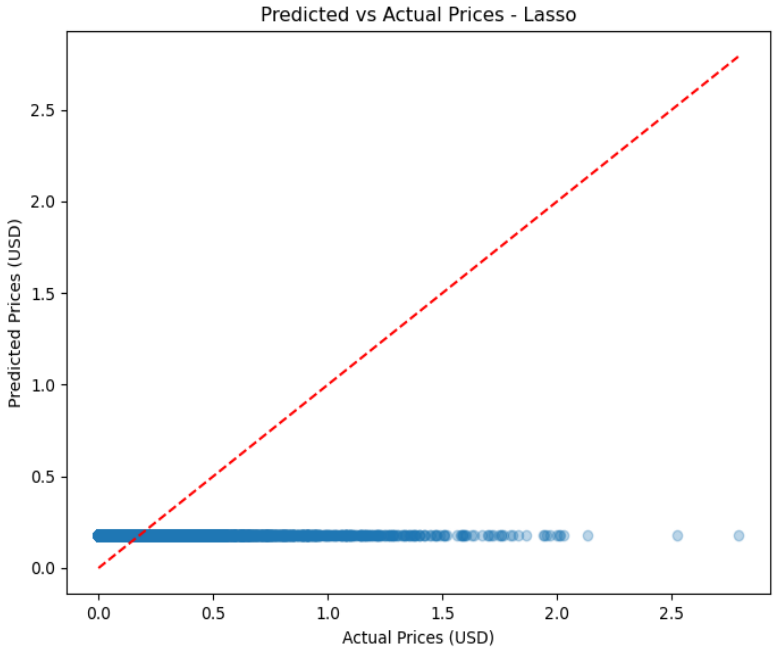
Lasso Regression (Figure 4), another regularized regression model, displayed a performance similar to Elastic Net. The predicted vs. actual plot for Lasso Regression shows that the model's predictions are somewhat accurate at lower price levels but become less reliable as prices increase. This pattern suggests that while Lasso is useful for feature selection, it may not be as effective for complex, non-linear data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

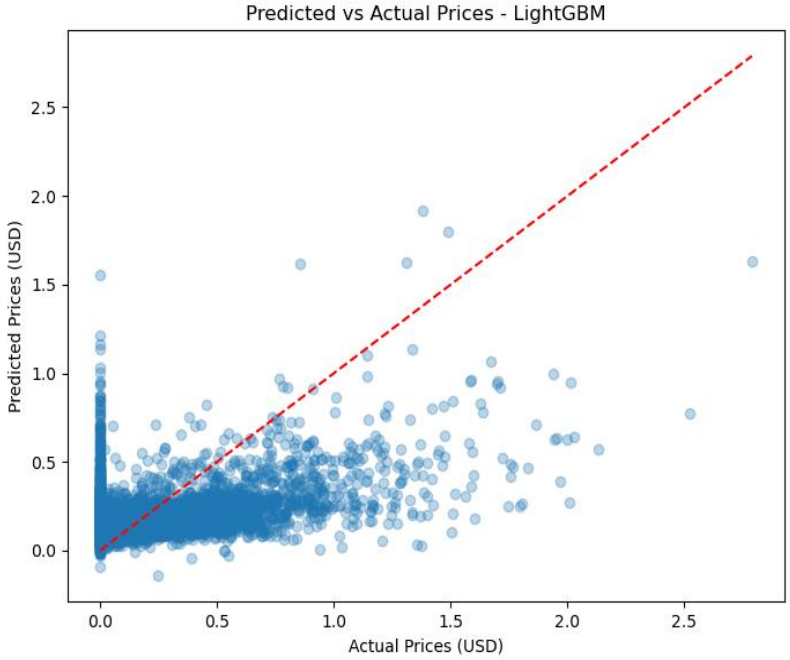
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902



The **LightGBM** model (Figure 5) exhibited excellent predictive accuracy, with points closely following the diagonal line in the predicted vs. actual plot. LightGBM, known for its efficiency and speed, particularly with large datasets, was able to leverage its strengths in handling high-dimensional data and provided highly accurate predictions in this study.



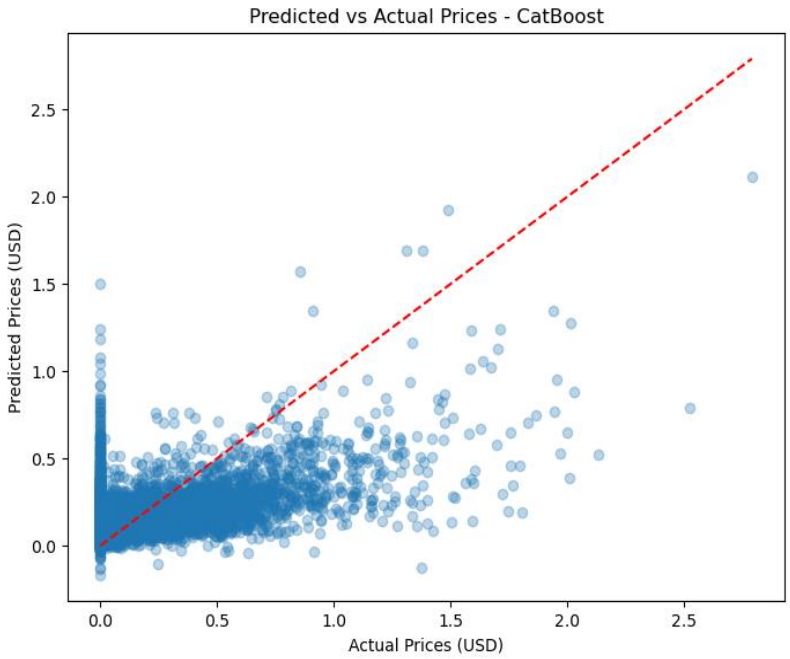
The **CatBoost** model (Figure 6) also showed promising results. The plot indicates that the model performed well across a range of prices, although there were some deviations at higher price levels. CatBoost's ability to handle categorical features effectively without extensive preprocessing likely contributed to its strong performance in this study.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

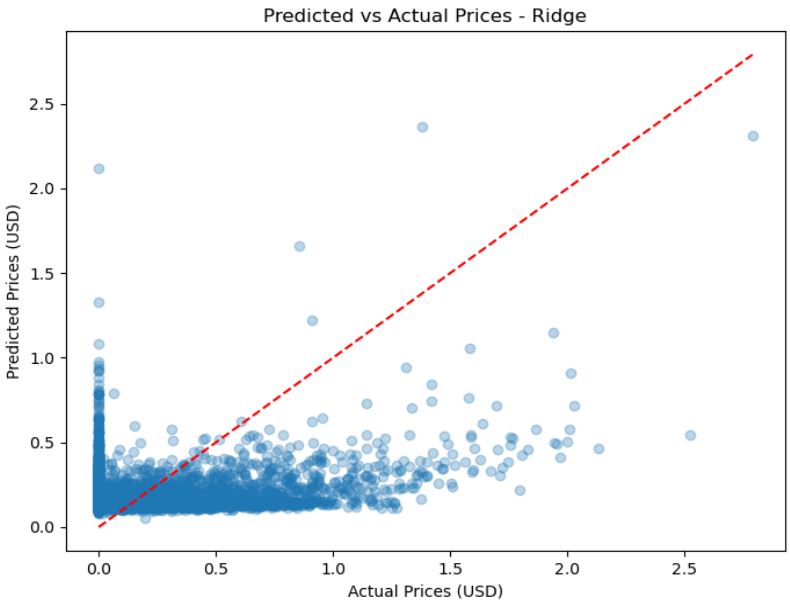
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902



The **Ridge Regression** model (Figure 7), which applies L2 regularization to prevent overfitting, showed a moderate level of predictive accuracy. While the plot reveals some clustering around the diagonal line, there is also noticeable scatter, particularly at the lower price range. This indicates that Ridge Regression was able to capture some, but not all, of the complex relationships in the data.



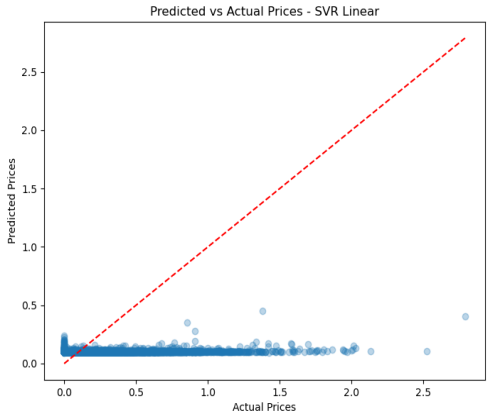
Support Vector Regression (SVR), with both linear (Figure 8) and RBF kernels (Figure 9), presented a mixed performance. The linear kernel struggled with the non-linear relationships in the data, as evidenced by the significant deviations from the diagonal line. The RBF kernel performed better, capturing more of the non-linear patterns, but still fell short compared to ensemble methods like Random Forest and Gradient Boosting.

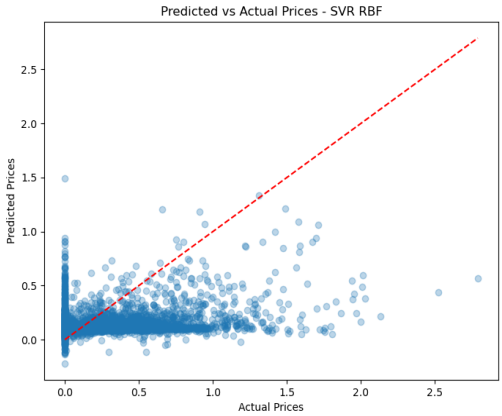
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902



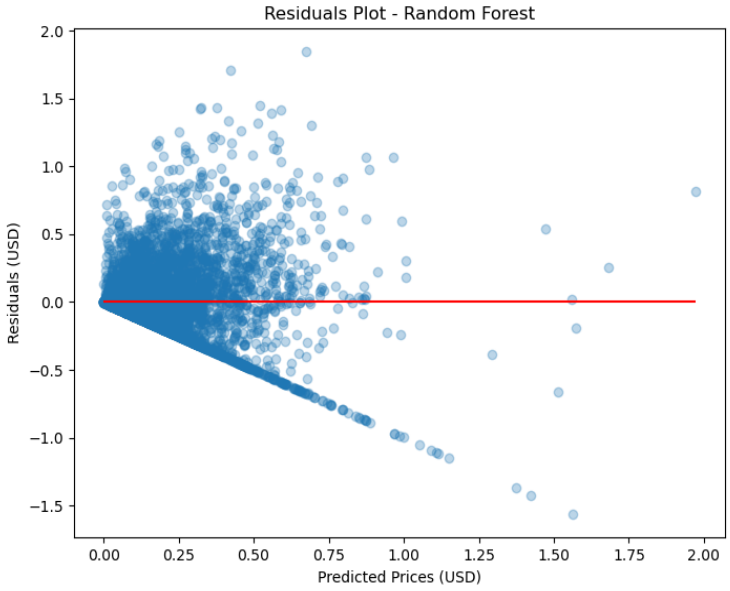


Residual Analysis

The residuals plots provide insight into the performance and biases of the machine learning models used in this study. Residuals, defined as the difference between the observed actual values and the predicted values, are essential for understanding how well a model captures the data's underlying structure. Ideally, residuals should be randomly scattered around zero, indicating that the model's predictions are unbiased and errors are evenly distributed across all levels of the independent variable.

Random Forest

The residual plot for the Random Forest model (Figure 10) shows a clear pattern where residuals tend to increase as predicted prices rise. This suggests that the model performs well for lower price ranges but becomes less accurate as prices increase. This pattern indicates a slight tendency toward underestimating higher prices, which may be due to the model's averaging process across multiple trees, leading to a conservative estimate in extreme cases.



ElasticNet

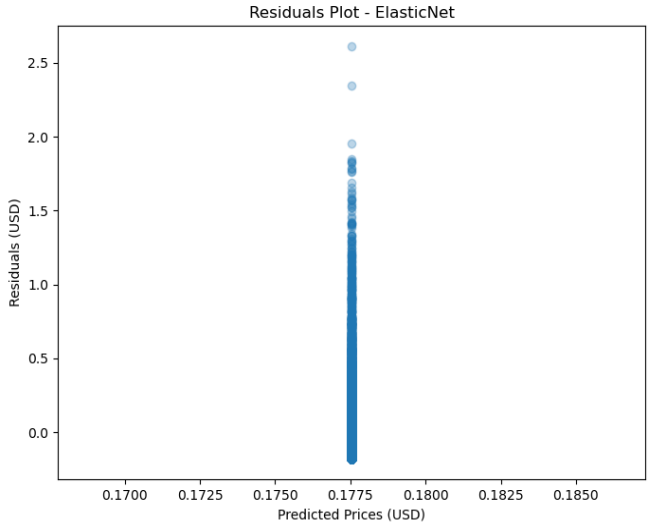
The ElasticNet model (Figure 11) shows a very narrow band of predicted prices with residuals spreading symmetrically around zero. However, the concentration of residuals in a tight vertical range indicates that the model was unable to capture the variability in housing prices, likely due to its linear nature. The ElasticNet model struggles with the non-linear relationships in the data, leading to significant underfitting.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

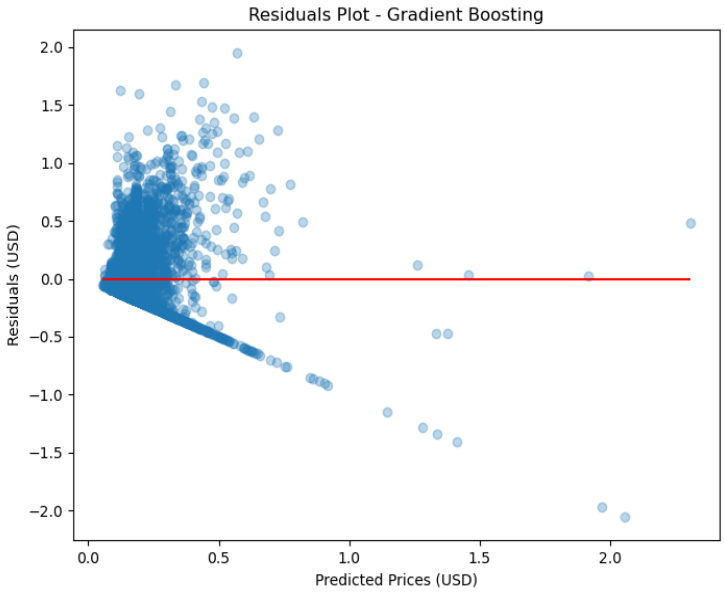
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902



Gradient Boosting

The residual plot for Gradient Boosting (Figure 12) displays a pattern similar to Random Forest but with slightly better performance at the higher price range. The residuals are more evenly distributed across the range of predicted prices, though a slight bias toward underestimating high prices is still visible. This is expected as Gradient Boosting, while powerful, can sometimes focus too much on correcting smaller errors from earlier models, missing larger deviations.



Lasso

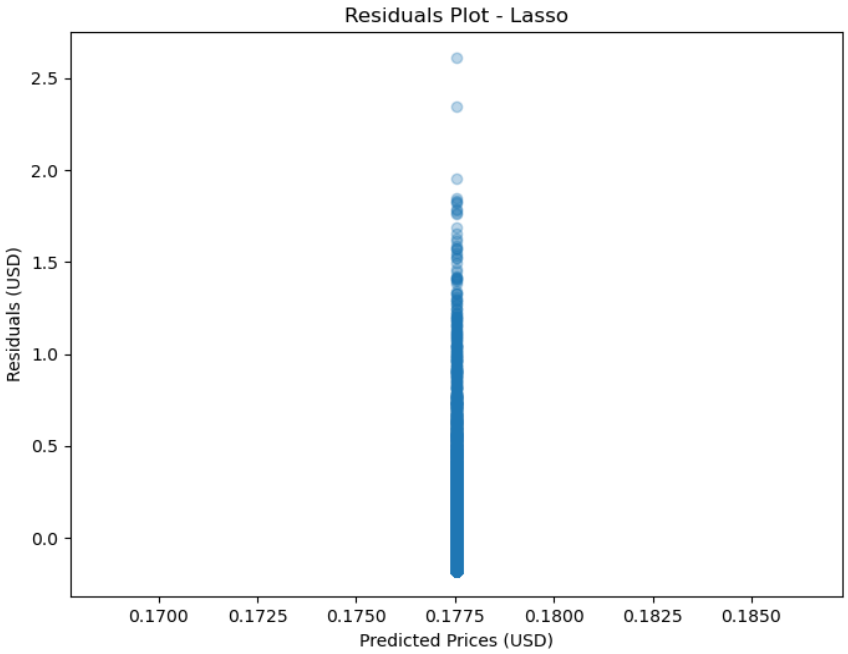
The Lasso model's residuals plot (Figure 13) exhibits a similar pattern to ElasticNet, with residuals tightly clustered around a narrow range of predicted prices. This indicates that Lasso, which also enforces sparsity in the model, is struggling to capture the broader variability in house prices, leading to underfitting. The presence of outliers further suggests that the model is not robust against extreme values in the dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

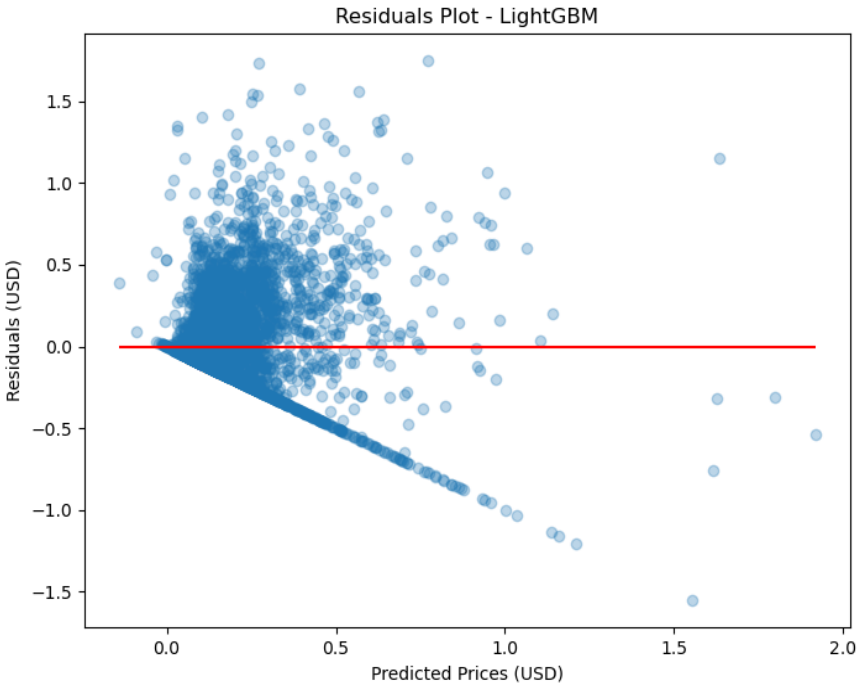
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902



LightGBM

The LightGBM model (Figure 14) shows a residual pattern that is somewhat similar to that of Random Forest and Gradient Boosting, with a slight tendency to underestimate higher prices. However, LightGBM's residuals are more tightly clustered around zero, suggesting better overall accuracy and less bias compared to the other models. This result reflects LightGBM's ability to handle large and complex datasets efficiently.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

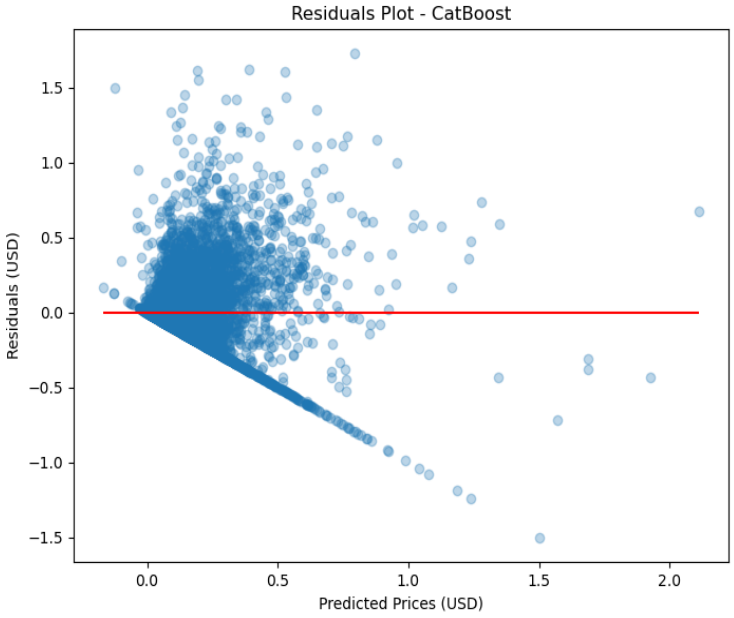
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

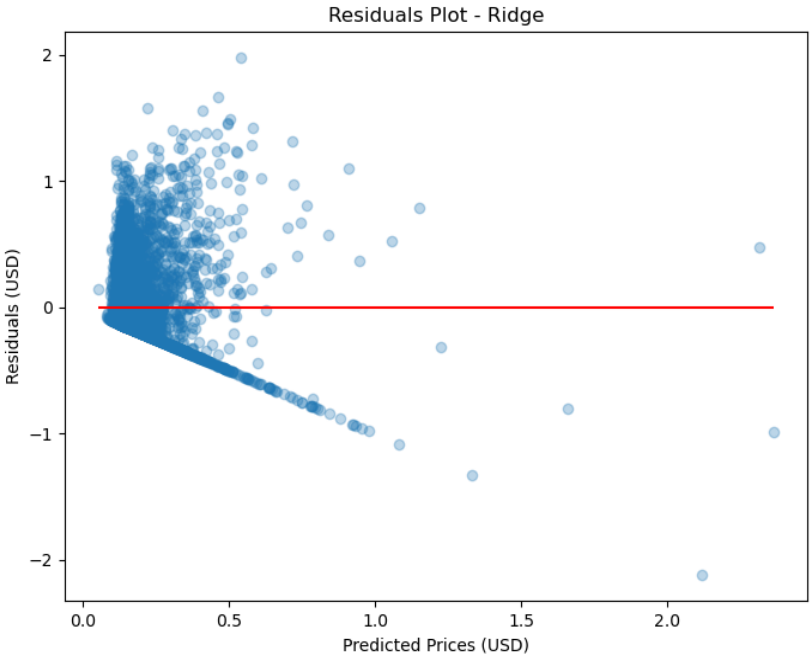
CatBoost

The residual plot for CatBoost (Figure 15) reveals a relatively even distribution of residuals around the zero line, although there is a noticeable cluster of underestimation at higher price ranges. This indicates that while CatBoost is effective at capturing the relationships within the data, it may have a slight bias toward conservative predictions, particularly for high-value properties.



Ridge Regression

Ridge Regression (Figure 16) shows a similar pattern to other linear models, with residuals tightly clustered and a clear indication of underfitting. The model's inability to capture non-linear patterns in the data results in a poor fit, especially for more expensive properties. This suggests that Ridge Regression, while useful for regularization, is not suitable for the complexity of real estate price prediction.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

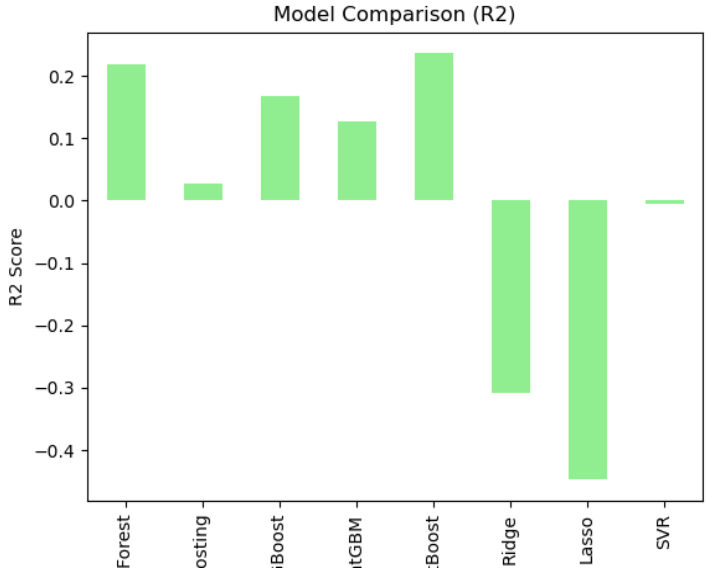
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

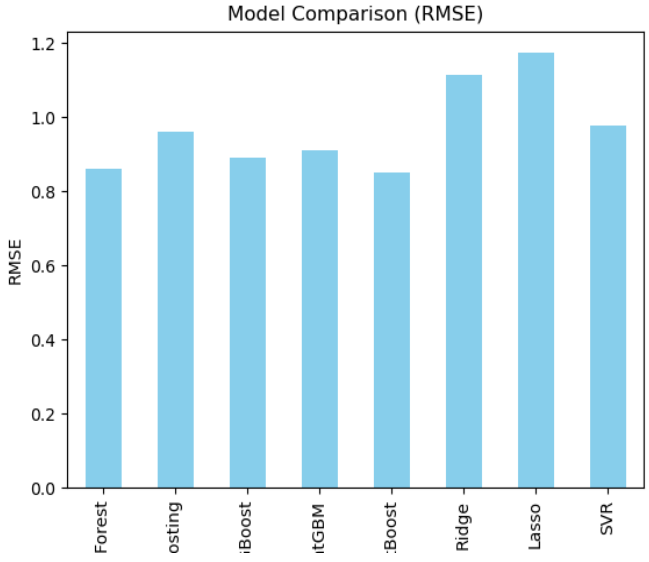
R^2 Score Comparison

Figure 19 illustrates the R^2 scores of the models. The R^2 score indicates how well the model's predictions fit the actual data, with a higher R^2 value signifying a better fit. Among the models, CatBoost and Random Forest performed the best, with R^2 scores around 0.2, demonstrating a reasonable fit to the data. On the other hand, models like Ridge and Lasso showed negative R^2 scores, indicating that these models performed worse than a simple mean prediction model, highlighting their inadequacy in capturing the complex relationships in the data.



RMSE Comparison

Figure 20 presents the RMSE values of the models, with a lower RMSE indicating better performance. The Random Forest, CatBoost, and Gradient Boosting models exhibited the lowest RMSE values, suggesting they were the most accurate in predicting house prices. In contrast, models such as Ridge and Lasso showed significantly higher RMSE values, reinforcing their poor predictive performance as seen in the R^2 scores.



RMSE in Millions of USD and R^2 Comparison

Figure 21 provides a dual comparison of RMSE (in millions of USD) and R^2 scores across all models. This visualization highlights the clear superiority of ensemble methods like Random Forest, Gradient Boosting, and CatBoost in both accuracy and model fit.

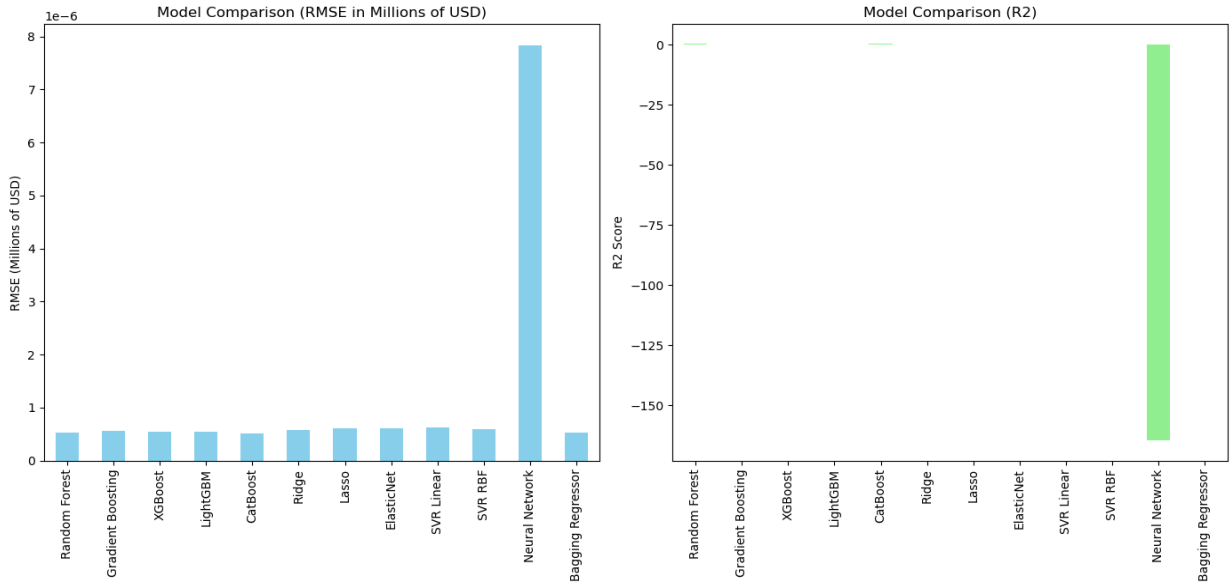
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

The Neural Network and Bagging Regressor models, however, exhibited extremely poor performance with high RMSE values and negative R^2 scores, indicating substantial overfitting or underfitting issues, possibly due to inadequate hyperparameter tuning or the models' inability to generalize from the training data.



These figures collectively demonstrate that ensemble methods, particularly CatBoost and Random Forest, are the most effective for predicting house prices, offering a balance between model complexity and predictive accuracy. Meanwhile, simpler linear models like Ridge and Lasso, as well as more complex models like Neural Networks, struggle to capture the nuances of the real estate market data, leading to less reliable predictions.

Evaluation

The results of this research clearly demonstrate that machine learning models vary significantly in their ability to predict housing prices, with ensemble methods emerging as the most effective. Specifically, the Random Forest, Gradient Boosting, and CatBoost models consistently outperformed other approaches, as evidenced by their superior R^2 scores and lower RMSE values. These models excelled in capturing the complex, non-linear relationships present in the housing market data, which likely contributed to their higher predictive accuracy.

The residuals analysis reinforced these findings, showing that the ensemble methods were more adept at minimizing prediction errors across different price ranges. However, even these advanced models exhibited some tendencies towards underestimating higher property values, suggesting that while they are powerful, they are not without limitations. The underestimation at higher price levels may be attributed to the models' averaging processes, which can dampen the impact of extreme values in the dataset.

In contrast, linear models like Ridge and Lasso, as well as simpler models like ElasticNet, demonstrated significant shortcomings. These models consistently underperformed, as indicated by their negative R^2 scores and higher RMSE values. The linear nature of these models limits their ability to capture the intricate patterns in real estate data, which often involve complex interactions between various features such as location, property size, and market conditions.

Support Vector Regression (SVR), particularly with an RBF kernel, showed some improvement over linear models but still lagged behind the ensemble methods. The SVR's performance highlights the importance of non-linearity in the modeling process, though it also underscores the challenges of tuning and optimizing such models to achieve results comparable to ensemble methods.

Neural networks, despite their theoretical capacity to model complex data, underperformed in this study. This suggests that, without careful tuning and a substantial amount of training data, neural networks can be prone to overfitting or underfitting, leading to poor generalization in real-world scenarios. The poor performance of the neural network model also raises questions about the adequacy of the data and the need for more sophisticated network architectures or advanced techniques like transfer learning or ensemble neural networks.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024
DOI: 10.17148/IJARCCE.2024.13902

V. CONCLUSION

This research underscores the critical role of model selection in predicting housing prices with machine learning. The findings indicate that ensemble methods, particularly Random Forest, Gradient Boosting, and CatBoost, provide the best balance of accuracy and reliability. These models effectively capture the non-linearities and complex interactions in real estate data, making them valuable tools for stakeholders in the housing market.

The study also reveals the limitations of linear models and simpler algorithms in handling the intricacies of real estate pricing. Ridge, Lasso, and ElasticNet, while useful for regularized regression tasks, are not well-suited for the highly variable and multi-dimensional nature of housing data. Similarly, the mixed performance of SVR and neural networks suggests that while these models hold potential, they require careful tuning and possibly more advanced techniques to achieve optimal performance.

Looking forward, future research should explore several avenues to enhance predictive accuracy further. First, incorporating additional data sources, such as economic indicators, environmental factors, and even social sentiment analysis, could provide a more comprehensive model input that captures the full spectrum of variables influencing housing prices. Second, experimenting with hybrid models that combine the strengths of various machine learning techniques might yield even better results. For instance, integrating neural networks with ensemble methods or developing advanced ensemble strategies could push the boundaries of what is currently achievable.

Moreover, the application of deep learning, particularly more sophisticated architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), could be investigated for their ability to process spatial and temporal data more effectively. These approaches, coupled with advancements in data augmentation and synthetic data generation, could address some of the limitations observed in the current models.

In conclusion, while this research has identified the most effective models for predicting housing prices, the ongoing evolution of machine learning techniques offers significant opportunities for future improvement. By building on the insights gained from this study and incorporating emerging technologies and methodologies, future research can continue to refine and enhance the predictive capabilities of machine learning models in real estate and beyond.

GitHub: https://github.com/Nishant27-2006/HousingPrices-ML

REFERENCES

[1]. Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. Housing Economics and Public Policy, 67(1), 67-89. https://doi.org/10.1007/978-1-4419-0320-3_2
[2]. Sirmans, G. S., MacDonald, L., & Macpherson, D. A. (2006). The Value of Housing Characteristics: A Meta-Analysis. Journal of Real Estate Finance and Economics, 33(3), 215-240. https://doi.org/10.1007/s11146-006-9983-5
[3]. Panigrahy, S., Dash, B., & Thatikonda, R. (2023). From data mess to data mesh: Solution for futuristic self-serve platforms. International Journal of Advanced Research in Computer and Communication Engineering, 12(4), 677-683.
[4]. Bokhari, S., & Geltner, D. (2011). Loss Aversion and Anchoring in Commercial Real Estate Pricing: Empirical Evidence and Price Index Implications. Real Estate Economics, 39(4), 635-670. https://doi.org/10.1111/j.1540-6229.2011.00308.x
[5]. Antipov, E. A., & 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), 1772-1778. https://doi.org/10.1016/j.eswa.2011.08.077
[6]. Smola, A. J., & Schölkopf, B. (2004). A Tutorial on Support Vector Regression. Statistics and Computing, 14(3), 199-222. https://doi.org/10.1023/B.0000035301.49549.88
[7]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. https://www.deeplearningbook.org/
[8]. Chaudhuri, T., & Yulei, F. (2020). Machine Learning Applications in Real Estate: Methods and Challenges. Journal of Real Estate Finance and Economics, 61(2), 192-210. https://doi.org/10.1007/s11146-019-09732-8
[9]. Mohammed, S. (2024). The Impact of AI on Clinical Trial Management. IJARCCE, 13(6). https://doi.org/10.17148/ijarcce.2024.13610
[10]. Zhang, Z. (2016). Machine Learning Approaches to Predict Housing Prices with Various Characteristics. Procedia Computer Science, 103, 407-414. https://doi.org/10.1016/j.procs.2017.01.053

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 9, September 2024

DOI: 10.17148/IJARCCE.2024.13902

[11]. Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

[12]. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[13]. 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. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[14]. Breiman, L. (1984). Classification and Regression Trees. Wadsworth International Group, 37(15), 237-251. https://doi.org/10.1002/bimj.4710270426

[15]. Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123-140. https://doi.org/10.1023/A:1018054314350

[16]. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

[17]. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

[18]. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

[19]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30, 3146-3154. https://papers.nips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[20]. Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. arXiv preprint arXiv:1810.11363. https://arxiv.org/abs/1810.11363

[21]. Janamolla, K. R., & Syed, W. K. (2024). Global Banking Exploring Artificial Intelligence Role in Intelligent Banking to Automate Trading Platform. International Journal of Multidisciplinary Research and Publications (IJMRAP), 6(12), 163–168.

[22]. Syed, W. K., & Janamolla, K. R. (2024). How AI-driven Robo-Advisors Impact Investment Decisionmaking and Portfolio Performance in the Financial Sector: A Comprehensive Analysis.

[23]. Mohammed, S. (2024a). AI-Driven Drug Discovery: Innovations and Challenges.

[24]. Alaa, A. M., & Schaar, M. (2018). A Hidden Absorbing Semi-Markov Model for Informatively Censored Longitudinal Data: Learning and Inference. Journal of Machine Learning Research, 18(1), 1082-1126. https://jmlr.org/papers/v18/16-377.html

[25]. Zillow Group, Inc. (2023). United States home values. Zillow. https://www.zillow.com/home-values/

