Data-Driven Real Estate Validation: Advanced Predictive Modeling Techniques

Gunakala Archana[1], Kancharla Stephen[2], Vuyyuru Sriram Lokesh[3] and Beaulah Jala[4*]
{archu.gunakala@gmail.com[1], kancharlastephen123@gmail.com[2], vuyyurulokesh99@gmail.com[3], beulahjala123@gmail.com[4]}

Department of Advanced Computer Science and Engineering, Vignan's Foundation for Science, Technology & Research (Deemed to be University), Vadlamudi, Guntur (Dt), Andhra Pradesh, India[1, 2, 3, 4]

Abstract. Accurate prediction of house price is an important problem in the real estate market, with significant implications to buyers, sellers and investors. This research investigates creating a reliable regression-based model using state of the art machine learning algorithms in order to improve the accuracy of estimating residential property values. The proposed approach incorporates elaborate data preprocessing, feature engineering, and model stacking stages to overcome problems of missing data, high-dimensionality, and data model-specific feature transformations. The method proposed consists of a detailed data preprocessing, feature engineering, and model stacking to overcome issues related to missing data, high dimensionality, and the necessity of performing a model-appropriate feature transformation. For better performance and predictability, a model end with an ensemble for the purpose of prediction and explaining the factors that can cause housing prices is presented in way that would help us in decision making. Neighborhood, size, and condition of the property is among the most important factors that contribute to determining the price. Due to the complexity of real-estate markets, and the also varied nature of related influencing features this project aims at linking raw housing data and accurate price estimation by employing data-driven, predictive regression models. The results of this study may help inform decision making by stakeholders in the real estate universe.

Keywords: House Price Prediction, Machine Learning, Regression Models, Ensemble Learning, Feature Engineering, Predictive Modeling, Data Preprocessing.

1 Introduction

House price prediction is a classical and popular problem in academies and industries. With the added complexity or unpredictability of housing markets, accurate and data-driven estimates of housing values are increasingly valuable. Common models like linear regression have been commonly used to address the problem. However, such models can have difficulty interpreting data from the real world that is often messy and has non-linear relationships. Chen [4] had noted these types of models are simple and interpretable, but not adept at accommodating large and complex datasets. The use of Machine Learning (ML) models represents a potential alternative. They can accommodate many types of data such as place trends, time aspects and background demographics to make it easier for them to detect patterns that classical models might have missed. For instance, Varma et al. [2] improved prediction performance by utilizing the location

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358066

features extracted from Google Maps, and Patil et al. [3] used automated software tools to more effectively gather information.

For house price prediction tasks, models such as XGBoost, Random Forest, and Gradient Boosting are particularly effective. These models handle complex, non-linear relationships and provide high accuracy in forecasting property prices [5]. Studies by Lu et al. [12] and Sharma et al. [13] have demonstrated that, properly tuned, these models are capable of obtaining state-of-the-art results.

More recently, researchers are also beginning to explore how easy these models are to interpret and use. As machine learning is used by more people whether homeowners, buyers or real estate agents the ability to explain how a prediction is made is as crucial a part of the result to be used in a product. Sanyal et al. [8] and Eze et al. [11] studied the interpretability of various models, and Madhuri et al. [14] compared various regression methods to identify the most appropriate in different situations. Other studies by Jain et al. [6], Chen [7] and Wu [9], the cleaning and choosing the right features of a column can have a significant impact on the accuracy and efficiency of a model. Overall, the research indicates that, to predict house prices, smart models are best when combined with good data, and clear explanations.

The research builds on the complex nature of housing price prediction. To address this challenge, the methodology incorporates advanced feature engineering techniques to enhance the model's predictive power. Furthermore, multiple regression models are combined to improve the accuracy and robustness of the predictions. It creates personalized composite features, to improve the model's performance and compares the performance of commonly used algorithms (Ridge, XGBoost, LightGBM) both independently and as ensemble. The results are intended to help facilitate the decision process for buyers, sellers and investors in an information-based housing market.

This project aims to create a robust regression model that would help estimating the potential price of houses based on important features of the house and property. By doing so, the model has the potential to guide buyers, sellers and investors in their decisions. It also has the power to mitigate financial risks and enable a sounder and data-driven real estate market.

This project has multiple sections to it. Section 2 discusses related work and summarizes different machine learning algorithms applied in previous studies for price prediction of houses. Section 3 covers the dataset. Section 4 is on methodology and baseline of the project, 5 is on preprocessing, and feature engineering. In Section 6, we present the machine learning models Ridge, XGBoost, LightGBM, and ensemble models along with the training and testing process followed for each. Results are shown in Section 7, including performance comparison based on RMSE and R^2 among the methods and plots. And Section 8 ends the project by providing a summary of the main process, the comparison of different model performances, and some thoughts on the advantages of ensemble learning for the forecast performance.

2 Literature Survey

Predicting house prices has become an important ap- plication of machine learning (ML) that assists in market analysis as well as in decision-making for buyers and sellers. Early research in

this domain was concerned with using classical \'statistics-based\' models, while more recent work has moved towards data-rich ML techniques able to take advantage of the complexities and multi-dimensionality of the input data. comprehensive review indicating a distinct shift towards spatial features and deep learning for predictive modelling. In Spain, Mora-Garcia et al. [1] applied XGBoost, LightGBM, and Gradient Boosting to georeferenced housing data, demonstrating that ensemble models are better at capturing nonlinear relationships than traditional regressors. Varma et al. [2] went one step further and combined neural networks and geographic APIs, and Patil et. [3] showed how RPA could be used to make data collection simpler, and enabled better predictions over CatBoost.

A variety of comparative methods have been employed by different investigators. Chen [4] used the Boston housing data and found XGBoost performed better than SVM and Random Forest on the challenge. Rana et al. Multiple models like SVR, Decision Trees etc., have been used to predict house prices in Bangalore and THEN the best has been chosen based on error metrics [5]. Jain et al. [6] constructed stacked regression models with cross-validation, and Chen [7] mapped K-Fold validation to the California data-set using tree-based regressors. Sanyal et al. [8] focused shrinkage of features and generalization, and proved that Lasso regression is good at reducing overfitting. inational B) choice are growing. Wu [9] demonstrated the potential of multivariate linear regression to enable the analysis of the socioeconomic factors, as well as Bhagat Although these claims have beentestet al. [10] increasing the accuracy by the use of heavy preprocessing and model tuning. Eze et al. [11]), who compared with the Boston dataset simple models, but found Random Forest to be best. Lu et al. [12] applied a composite model with Lasso and Gradient Boosting which accurately classified the activity type for a Kaggle competition. Sharma et al. [13] provided additional evidence by testing five algorithms on the Ames dataset and highlighting the gains from hyperparameter tuning, and once again found XGBoost to be the most accurate.

To give a wider perspective, Madhuri et al. [14] performed a comparative study comparing different regression techniques such as Multiple Linear Regression, Ridge, Lasso, Elastic Net, AdaBoost, and Gradient Boosting, was tested on the King County housing dataset. The performance was demonstrated, not only in terms of "classical" measures, MSE and RMSE, but also the real-world application, the Python and Jupyter tools used in that train.studies suggested. They concluded that Gradient Boosting Regression performed better than the other ones as for the best accuracy score and prediction error and thought it would be the most suitable model for their case study. This work adds to prior work by showing that, for regression families not only do boosting methods outperform (if properly visualized) but moreover the visualization and evaluation methods are also very effective.

A second common theme to all these studies is the growing relevance of preprocessing and feature engineering. Several papers highlighted that data cleaning, outlier processing, and feature selection are as important as the ML algorithms. For example, Bhagat et al. [10] and Lu et al. [12] which mention that if the feature spaces are well-engineered, then the type of the classifier is secondary compared to the improvement provided by the input features characteristics: house location, floor area or distance to services. Third, adding geographic APIs as demonstrated in Varma et al. [2], or social factors, examined by [9], that can assist models to model market dynamics that plain numeric datasets are not able to.

Model interpretability, too, is increasingly important, particularly in real estate markets where stake- holders prefer clear decision support tools. Although the XGBoost and neural network-based black-box models achieve high predictive performance, several studies have attempted to trade off accuracy and interpretability. Research such as the work of Sanyal et al. [8] and Madhuri et al. [14] used regression coefficients, error distribution plots, and SHAP value visualizations to gain insight into what affects housing prices. Such methods aim to close the divide between technical modeling and real-world usability, enabling ML tools to be more usable to non-expert users like realtors, buyers and policymakers.

In conclusion, the literature surveyed above reveals several dominant trends: the transition towards ensemble learning meth- ods (particularly XGBoost and Gradient Boosting), the importance of robust data preprocessing, the incorporation of location-aware and socioeconomic features, and the increasingly important concern for interpretability of model output s. All together, these studies demonstrate that pricing houses is not just a technical problem— it's a multidisciplinary problem that benefits from well-considered model design, consideration of appropriate data sources, and real world applicability. As machine learning technology develops, future work could investigate more on how explainable AI and real-time data flow can improve the accuracy, interpretability and use value of the house price prediction models.

3 Dataset

The dataset is a robust foundation for predictive modeling and house price prediction, as it provides rich information on residential houses. With over 80 features and sample size of 1460, realignment was necessary to reduce dimensionality and enhance model performance. Quantitative features such as lot area, year built, square foot, and rooms provide measurable information on property attributes, while qualitative features such as architectural style, overall quality, and neighborhood provide additional SalePrice estimation information. With ordinal features such as material and finish quality ratings, these features provide a realistic estimate of a property's market value. Richness of data and fine-grained labeling allow for greater interpretation using feature engineering, which allows for better predictive modeling. Inclusion of location information such as neighborhood and proximity to essential services captures the effect of external factors on property attractiveness. Time-varying features such as construction year and dates of renovation also provide information on lifecycle trends. Through the merging of accurate measurement with feature engineering, the dataset provides a robust foundation for the identification of meaningful patterns and the formulation of accurate predictive models.

4 Methodology

The analysis began with Ridge Regression to establish a performance baseline. Advanced machine learning algorithms such as XGBoost, CatBoost, and LightGBM were then explored due to their effectiveness in capturing nonlinear relationships. Additionally, An ensemble model was created by averaging predictions from multiple individual models. While this is sometimes referred to as stacking or voting, in this case, a simple mean of outputs was used instead of a meta-model or majority vote approach. The data set is perfectly suited to residential property price analysis and forecasting with the inclusion of detailed information regarding residential properties. The data set has over 80 features, which had to be rescaled to reduce dimensionality and improve model performance. Significant quantitative features, including square foot, lot

area, number of rooms, and year built, offer quantifiable information closely related to property attributes. Qualitative variables, including architectural style, overall property condition, and neighborhood features, also offer useful contextual information, which assists in SalePrice prediction more precisely. The proposed methodology is shown in Fig. 1.

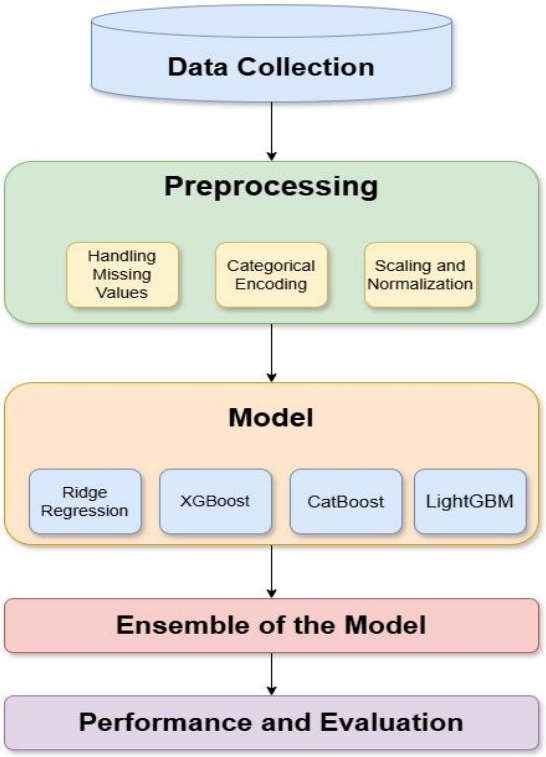

Fig. 1. Proposed Methodology.

4.1 Ensemble Learning

Ensemble Learning combines multiple models to improve predictive performance. Techniques such as Stacking, Bagging, and Boosting help to leverage the strengths of different models. The final prediction $\hat{y}$ in stacking is obtained as

$$\hat{y} = \sum_{i=1}^{n} w_i f_i(x) \quad (1)$$

Where w_i represents the weight assigned to each base model $f_i(x)$.

4.2 Preprocessing

Preprocessing is the process of transforming raw data into a clean and usable format before feeding it into machine learning models. It is important because real-world data often contains inconsistencies, missing values, and irrelevant information that can negatively affect model

performance. Proper preprocessing ensures that the data is well-structured, meaningful, and optimized for accurate and efficient learning. The data described in Section 3 are taken, and preprocessing techniques such as handling missing data, encoding categorical variables, and feature scaling are applied. The preprocessed data are then used as input to baseline models including Ridge Regression, XGBoost, CatBoost, and LightGBM, along with a proposed ensemble combination of these base models. The ensemble model is evaluated using K-fold cross-validation with K = 10.

For handling missing data, suitable imputation techniques were implemented contextually, with numerical features imputed using the median and categorical variables addressed through mode imputation or marked as 'NA' where applicable. Categorical encoding was performed using a combination of one-hot and ordinal encoding to preserve semantic meaning while optimizing model performance. Feature scaling and normalization were applied to ensure uniformity across numerical variables, with log transformations used to mitigate skewness where necessary.

4.3 Feature Engineering

Feature engineering further enhanced predictive accuracy and computational efficiency through the creation of composite features. Notably, the TotalBath feature aggregates full and half baths from both basement and above-ground levels as given in Eq 2:

$$TotalBath = BsmtFullBath + FullBath + 0.5 \times (BsmtHalfBath + HalfBath) \quad (2)$$

Similarly, the TotalPorchSF feature combines all porch and deck areas into a single metric which is given in Eq 3:

$$TotalPorchSF = OpenPorchSF + 3SsnPorch + EnclosedPorch + ScreenPorch + WoodDeckSF \quad (3)$$

The formulation of these features was guided by insights from a correlation matrix heatmap, visualized using Matplotlib, as shown in Fig. 2.

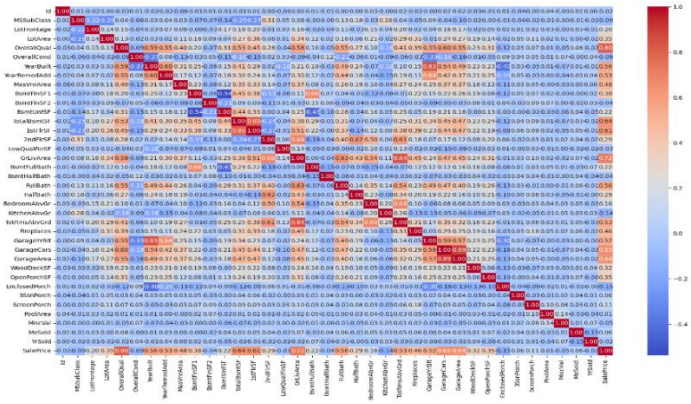

Fig. 2. Correlation matrix heatmap showing relationships between different variables.

5 Machine Learning Algorithms

5.1 Ridge Regression

Ridge Regression is a regularized linear regression model that introduces an L2 penalty to the loss function to prevent overfitting. It minimizes the sum of squared residuals along with the regularization term. The calculation of Ridge Regression is given in Eq. 4.

$$\widehat{\mathrm{B}} = \arg\min_{\mathrm{B}} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \mathrm{X_i B})^2 + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{B_j^2} \quad (4)$$

Where λ is the regularization parameter that controls the penalty on the magnitude of coefficients and B represents weights.

5.2 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses tree-based learning. It grows trees leaf-wise rather than level-wise, leading to faster training. The objective function in LightGBM is given by Eq. 5:

$$\mathcal{L} = \sum_{i=1}^{n} \mathrm{l}(y_i, \hat{y}_i) + \sum_{j=1}^{p} \Omega(f_j) \quad (5)$$

Where $\ell(\mathrm{yi}, \hat{y})$ is the loss function and $\Omega(f_j)$ is the regularization term.

5.3 XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized version of gradient boosting that includes regularization to improve generalization. The optimization objective is given by Eq. 6.

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (6)$$

Where $\Omega(f_t)$ penalizes model complexity to prevent overfitting.

5.4 CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm optimized for categorical data.To calculate CatBoost equation is given in Eq. 7. It employs ordered boosting to reduce overfitting and uses the following loss function for regression.

$$L = \sum_{j=1}^{m} \left(y_j - f\left(x_j\right)\right)^2 + \alpha \sum_{k=1}^{n} \beta_k^2 \quad (7)$$

Where α controls regularization, and $f_{(x_j)}$ represents the predicted output.

These models, particularly boosting-based algorithms, provide robust solutions for complex predictive tasks.

5.5 Components of the Machine Learning System

The machine learning pipeline implemented in this project is minimal but efficient. The data pre-processing process is devoted to clean and arrange the dataset and fill in missing values in different ways that varying depending on the feature type. Feature engineering improved the quality of the input data by createing new variables and summarizing relationships between input farther information in a better way.

We trained several regression models including tree-based ensemble approach which is particularly good at capturing nonlinear relationships in housing data. The models were then ensembled to achieve higher accuracy. Hyperparameter optimization was conducted to fine-tune the models and improve generalization. The performance of the model was tested using standard regression metrics, and cross-validation was applied to check if the models could be reliably generalised to new splits. Model validation by residual analysis confirmed the equalisation in predictions capacity of the model and the absence of heavy bias. Juxtaposed, these moves translated into a balanced and strong model for predicting house price.

5.6 Evaluation Metrics

5.6.1 Mean Squared Error (MSE):

It measures the average squared differences between the actual and predicted values.

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(a_i - \hat{a_\iota})^2 \quad (8)$$

Where a_i is actual value and $\hat{a}_i$ is predicted value.

5.6.2 Root Mean Squared Error (RMSE):

It is the square root of the MSE and provides an interpretation of the error in the same unit as the target variable.

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(a_i - \hat{a_\iota})^2} \quad (9)$$

Where a_i is actual value and $\hat{a}_i$ is predicted value.

5.6.3 R^2 Score (Coefficient of Determination):

It evaluates how well the model fits the data, with values closer to 1 indicating a better fit.

$$R^2 = 1 - \frac{\sum(a_i - \bar{a_i})^2}{\sum(a_i - \bar{a})^2} \quad (10)$$

Where a_i is actual value, $\hat{a}_i$ is predicted value and $\bar{a}$ is mean.

6 Results and Discussions

Table 1. Regression Performance Metrics.

Model	MAE	RMSE	R^2 Score
Ensemble Model	0.078834	0.115796	0.920572
CatBoost	0.080880	0.119591	0.915281
XGBoost	0.085762	0.123696	0.909366
Ridge Regression	0.091902	0.125340	0.906940
LightGBM	0.088146	0.127870	0.903145

Table 1 The regression model performance comparison shows the Ensemble Model leading with the lowest error metrics (MAE: 0.078834, RMSE: 0.115796) and highest R^2 score (0.920572), demonstrating the effectiveness of combining models for this task. CatBoost follows closely as the second-best performer (MAE: 0.080880, RMSE: 0.119591, R^2: 0.915281). XGBoost ranks third (MAE: 0.085762, RMSE: 0.123696, R^2: 0.909366), followed by Ridge Regression (MAE: 0.091902, RMSE: 0.125340, R^2: 0.906940), with LightGBM showing the weakest performance (MAE: 0.088146, RMSE: 0.127870, R^2: 0.903145). This ranking indicates that for this regression task, ensemble modeling provides superior predictive accuracy compared to individual algorithms.

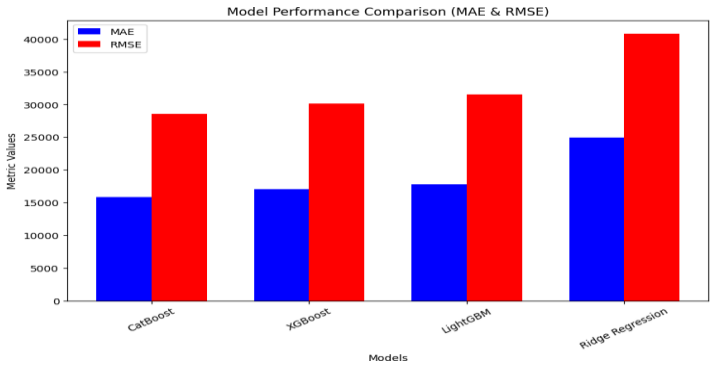

Fig. 3. Model Performances comparison (MAE & RMSE).

The scatter plot in Fig. 4 validates the predictive strength of the Stacking Regressor model, with predicted values closely aligning with actual outcomes. Which represents perfect prediction, indicating that the model is performing quite well. Most of the points cluster tightly around this line, suggesting that the model captures the underlying trends in the data with good accuracy. While there are a few deviations, especially at the lower and higher ends of the price range, they are relatively minor and expected in any real-world prediction scenario. Advanced models including CatBoost, XGBoost, LightGBM, Ridge Regression, were also evaluated and

compared with the proposed ensemble model for enhanced performance. Advanced models including CatBoost, XGBoost, LightGBM, Ridge Regression, and a custom Ensemble Model were evaluated to enhance predictive performance. The comparative results, depicted in Fig. 3, illustrate both the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each model. The Ensemble Model achieved the most favorable performance with a MAE of 0.079 and RMSE of 0.116, outperforming the individual base models. CatBoost and XGBoost followed closely, with MAE value of 0.081, RMSE value of 0.120 and MAE value of 0.086, RMSE value of 0.124, respectively. LightGBM recorded a MAE of 0.089 and RMSE of 0.128, while Ridge Regression exhibited the highest error metrics among the group, with a MAE of 0.092 and RMSE of 0.126. These results underscore the effectiveness of the ensemble strategy in reducing both average and squared prediction errors across the evaluated dataset.

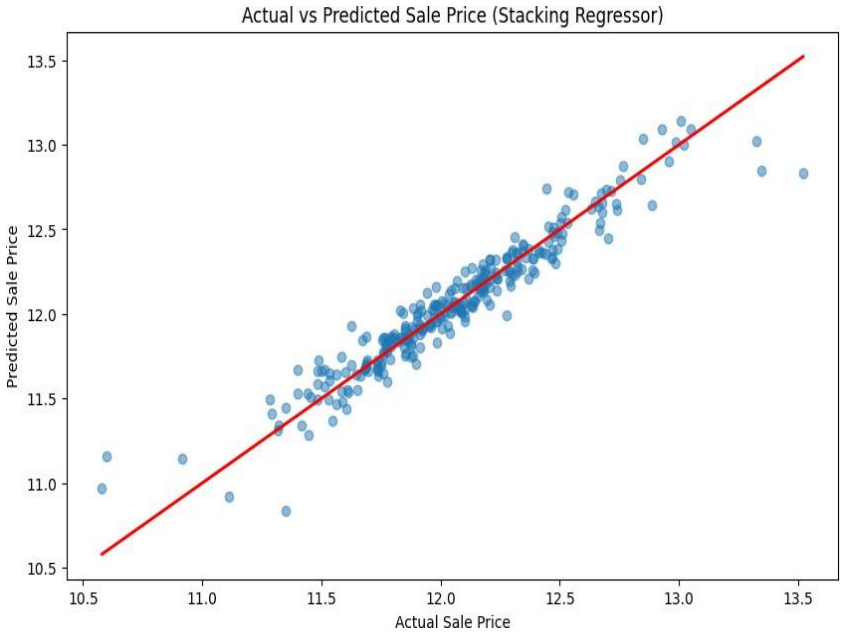

Fig. 4. Actual Vs Predicted (Stacking Regressor).

From the figures (Fig. 5, Fig. 6, Fig. 7, and Fig. 8), Ridge Regression stands out with the most stable performance, achieving the lowest average RMSE of 0.1087 and showing minimal variation across folds, making it a strong and reliable choice for consistent predictions. CatBoost also performs well, with an average RMSE of 0.1226, and demonstrates the ability to achieve low errors in several folds despite some variability, particularly around fold 6. LightGBM and XGBoost follow with average RMSEs of 0.1326 and 0.1370 respectively. While they exhibit greater fluctuations across folds, they also show potential for strong performance in specific splits, highlighting their capability to capture complex relationships in the data. Overall, Ridge offers the best consistency, while the boosting models showcase high predictive power with room for further tuning and optimization.

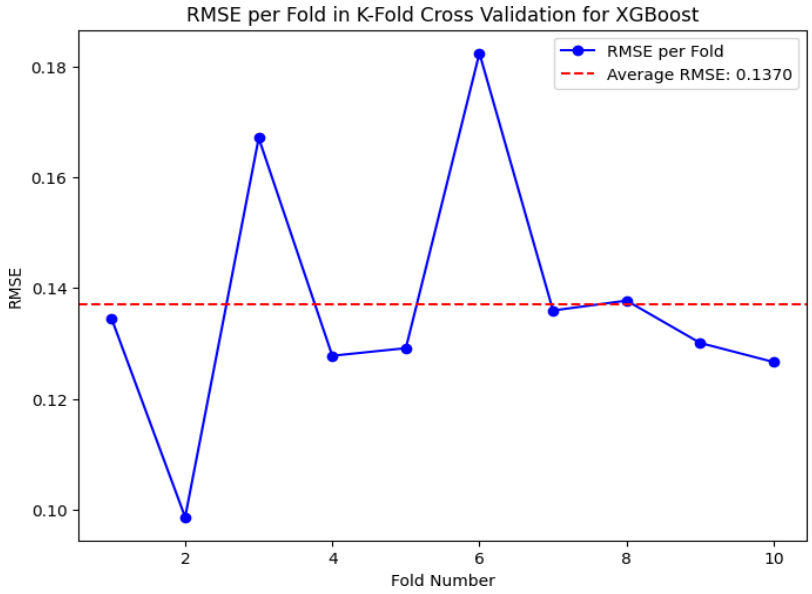

Fig. 5. RMSE per fold in K-Fold Cross Validation for XGBoost.

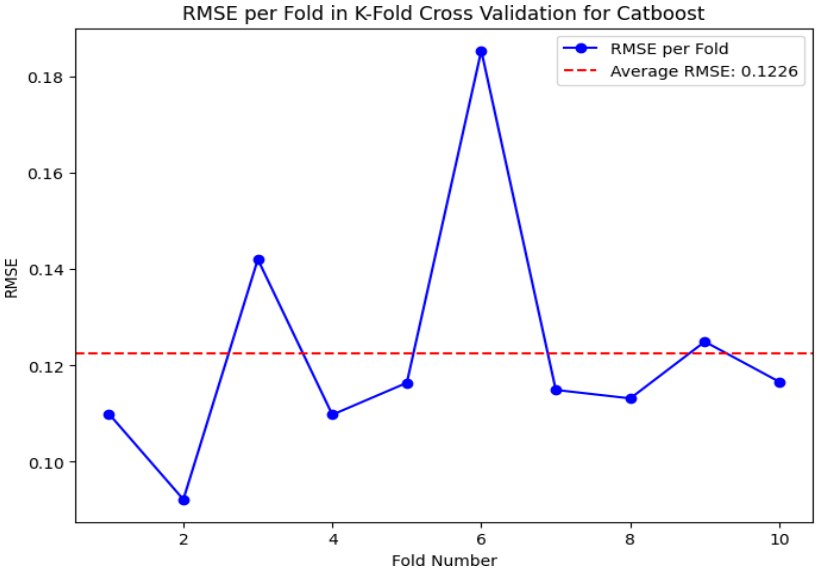

Fig. 6. RMSE per fold in K-Fold Cross Validation for CatBoost.

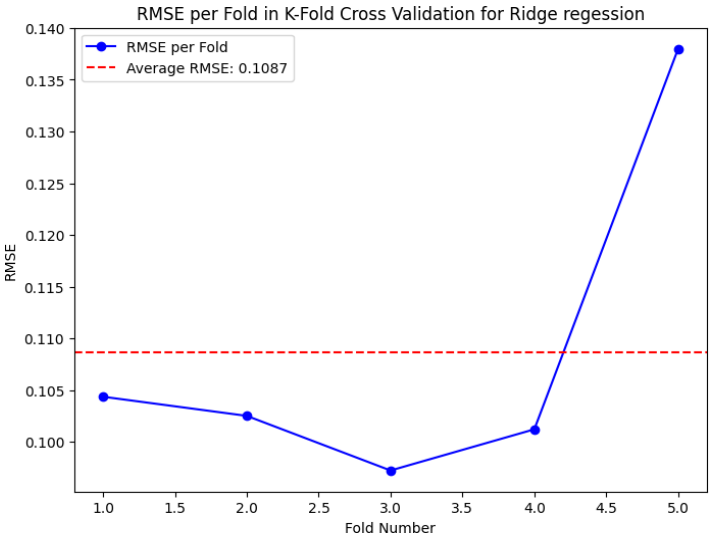

Fig. 7. RMSE per fold in K-Fold Cross Validation for Ridge Regression.

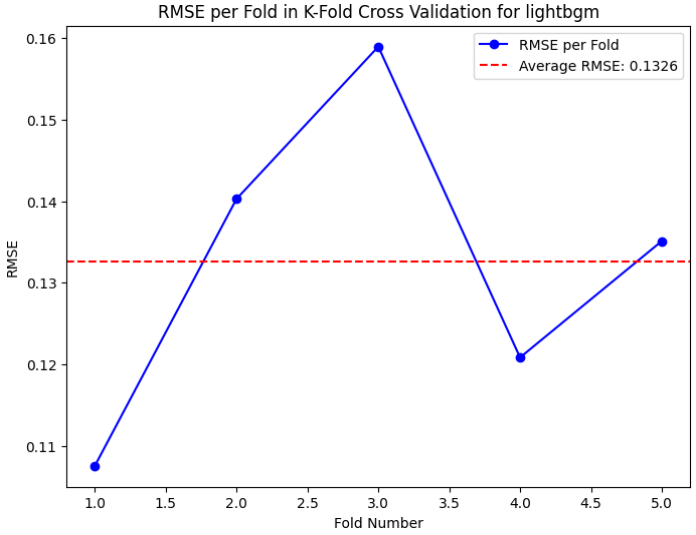

Fig. 8. RMSE per fold in K-Fold Cross Validation for LightGBM.

Fig. 8 contains the average RMSE for four machine learning models for house price prediction: CatBoost, LightGBM, Ridge Regression and XGBoost. Of these, Ridge Regression obtained the least RMSE value at around 0.109, which was the most accurate in predication. CatBoost was very close to that with an RMSE of 0.123 and LightGBM a little higher 0.132. XGBoost again had the greatest RMSE value, with an RMSE of about 0.141 and proving to be the least accurate among the models tested here. These findings reveal that Ridge Regression is the most efficient model w.r.t. error reduction, and second is CatBoost which can be given preference in

the case of categorical data in gradients boosting pros. Fig. 9 shows the Average RMSE for CatBoost, LightGBM, Ridge Regression, Xgboost.

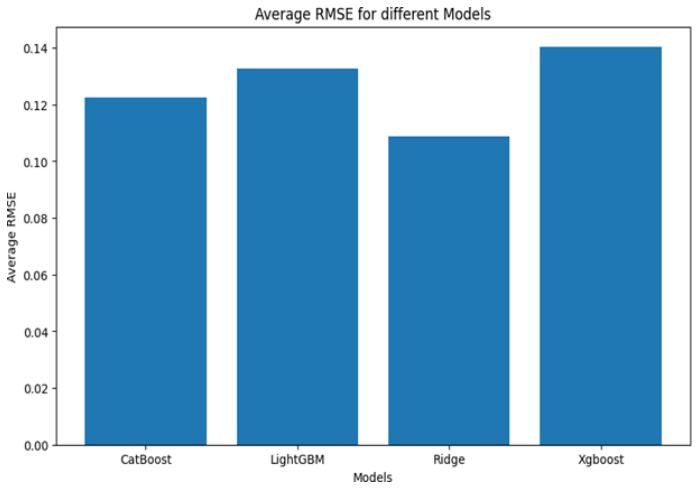

Fig. 9. Average RMSE for CatBoost, LightGBM, Ridge Regression, Xgboost.

7 Conclusions

This project objective was to develop CRISP-DM based house price prediction system with a number of machine learning regression techniques. After pre-processing and cleaning the data, several models such as Ridge Regression, LightGBM, XGBoost and CatBoost as well as ensemble of these models were implemented and tested. Each model has been evaluated by using the MAE, RMSE and the R^2 for a fair comparison. The combined classifier exhibited the highest performance, significantly superior to any single classifier for prediction accuracy. Results were strong for both CatBoost and XGBoost on standalone models, as did Ridge Regression which appeared stable with cross-validation. These results demonstrate how ensemble techniques enable optimal trade-offs between individual performance characteristics of the strategies. To summarise, the study suggests that the use of a combination of multiple regression methods can result in more robust and precise forecasting of the price. While the findings are encouraging, further research may consider more comprehensive tuning of the hyperparameters, evaluation with more various datasets, and methodologies for enhancing the model interpretability for practical applications.

8 Future Work

We are going to improve our feature selection techniques in the future so we can end up with more interpretable models, that are also simpler. With the aid of tools like SHAP (SHapley Additive exPlanations), the secrets behind predictions can be revealed, leading to more intuitive understanding of how the model works. These enhancements are designed to provide stakeholders with meaningful, data-driven insights that enable practical decision making in the real estate market.

References

[1] R. T. Mora-Garcia, M. F. Cespedes-Lopez, and V. R. Perez-Sanchez, "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times," Land, vol. 11, p. 2100, 2022.

[2] A. Varma, S. Doshi, A. Sarma, and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," Unpublished manuscript.

[3] P. Patil, D. Shah, H. Rajput, and J. Chheda, "House Price Prediction Using Machine Learning and RPA," International Research Journal of Engineering and Technology (IRJET), vol. 7, no. 3, 2020.

[4] Y. Chen, "Research on the Prediction of Boston House Price Based on Linear Regression, Random Forest, XGBoost and SVM Models," Highlights in Business, Economics and Management, 2023.

[5] V. S. Rana, A. Sharma, J. Mondal, and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," in Proc. ICACCCN, 2020.

[6] M. Jain, H. Rajput, N. Garg, and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," in Proc. ICESC, IEEE, 2020.

[7] Y. Chen, "Analysis and Forecasting of California Housing," Highlights in Business, Economics and Management, 2023.

[8] S. Sanyal, S. K. Biswas, D. Das, M. Chakraborty, and B. Purkayastha, "Boston House Price Prediction Using Regression Models," in Proc. CONIT, IEEE, 2022.

[9] Z. Wu, "Prediction of California House Price Based on Multiple Linear Regression," Academic Journal of Engineering and Technology Science, vol. 3, no. 7, 2020.

[10] A. Bhagat, M. Gosavi, A. Shahasane, N. Mishra, and A. Nerurkar, "House Price Prediction Using Machine Learning," SSRN, 2023.

[11] E. Eze, S. Sujith, W. Elmedany, and M. S. Sharif, "A Comparative Study for Predicting House Price Based on Machine Learning," in Proc. ICDABI, IEEE, 2023.

[12] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, "A Hybrid Regression Technique for House Prices Prediction," in IEEE IEEM, 2017.

[13] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," Analytics, vol. 3, no. 1, pp. 30–45, 2024.

[14] C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," in IEEE Int. Conf. on Soft-Computing and Network Security (ICSSS), 2019.