Boosting Accuracy: Advanced Ensemble Learning Strategies

Gagandeep Marken[1], Sofiur Rahaman[2]*
[1]Assistant Prof., Department of Computer Science Engineering, Chandigarh University, Gharuan
[2]M.Sc. Data Science, Department of Mathematics, Chandigarh University, Gharuan
gaganmarken1990@live.com ,
*sksofiurrahaman781@gmail.com

Abstract. Determining how to improve machine learning's predictive accuracy has led to notable developments in ensemble learning techniques. In order to solve prediction problems in a variety of datasets, this study, "Boosting Accuracy: Advanced Ensemble Learning Strategies," explores and applies cutting-edge boosting techniques. Three well-known methods are specifically the subject of this study: Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Machines (GBM). The advantages and disadvantages of these approaches are carefully assessed over a range of forecast scenarios. Our proposal is to create hybrid models that combine boosting techniques with other machine learning algorithms to create strong ensemble frameworks that go beyond the limits of conventional boosting approaches. The goal of this hybridisation is to get better predictive performance and robustness by utilising the advantages of each distinct model.

Keywords: GBM, Boosting, XG-Boost

1 Introduction

Real estate industry, which depends on the prediction of house prices, influences the decisions of potential buyers, sellers, investors, and policy implementers. Thus, reliable prediction will act as an impetus for better investment strategies, perfected pricing models, and enlightened policy decisions, which are important in any vibrant market, such as California's housing sector. Nevertheless, in this, the complexity of housing data, often with numerous features like location, size, age, and economic indicators, comes as a challenge to the traditional predictive model.

In recent years, a number of ensemble learning methods have been developed in machine learning that often combine strong trends in a number of models, making them enhance predictive accuracy and robustness. Gradient Boosting, XGBoost, and LightGBM have recently developed into well-accomplished models for general prediction tasks in regression and classification Inside. Even though the charismatic flair of each individual model, there are a number of weaknesses that hypothetically could be abated by placing them in an ensemble approach.
This paper thus aims to evaluate the applicability of advanced ensemble learning techniques in California house price prediction. The study would be very

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_24

instrumental in developing a model that outperforms any of the individual models in accuracy, generalization, and computational efficiency by exploiting the strengths of Gradient Boosting, XGBoost, and LightGBM. This study is based on the California housing dataset, which includes rich features, hence permitting a very comprehensive analysis on the factors that drive house prices.

The methodology followed in this research takes an ensemble of Gradient Boosting, XGBoost, and LightGBM. In this way, the created model would be robust enough to capture the intricate relationships that house data encompasses. Gradient Boosting is a bias-reducing method, while the strengths of XGBoost lie in its efficiency and scalability. LightGBM, however, has the advantage of handling large, high-dimensional datasets. This will form an ensemble of models that will not only enhance predictive accuracy, providing a more reliable and interpretable framework for understanding the dynamics of housing prices, but will also reduce bias.

It is in this regard that the research will add knowledge to the fast-expanding ensemble learning repository and its application in real estate analytics. It contributed valuable insights to academia and industry regarding the same thing by proving that the integration of the most advanced machine learning techniques will work out to make a difference in innovation in predictive modelling. The findings of the present study, therefore, extend beyond the real estate sector to provide a methodological framework that can be used in very different domains where prediction accuracy is paramount.

2 Literature Review

Ensemble learning techniques have indeed been proven to work with increased power in the improvement of the accuracy and robustness of any model concerned, particularly in predictive analytics such as housing price prediction. Further, along this school of thought, over these years a number of methodologies have been developed to improve the said performance of prediction, wherein the ensemble methods formed the basis for most of the modern applications in machine learning today.

Traditional Regression Models Most of the early methods employed for housing price prediction focused on some kind of linear regression model. These models were rather simple but very often unable to capture the complex nonlinear relationship in house data. They thus became limited in their capacity to offer good predictions for dynamic and varied housing markets.
Decision Trees and Random Forests Decision tree-based techniques brought a significant enhancement in predictive powers. Decision trees were much more flexible, as they can handle nonlinear relations and feature interactions. However, single trees usually suffered from overfitting, and therefore Random Forests were

developed as an ensemble technique for mitigating overfitting by averaging the predictions of multiple trees. In fact, they outperformed others significantly in generalizing, and it turned out that Random Forests would do very well for housing price prediction tasks.

Boosting Algorithms These are further developments of ensemble learning. Many people have really adopted techniques such as Gradient Boosting Machine, XGBoost, and LightGBM. This stems from the fact that they have the power to improve the errors of the previous models at a go, hence coming up with very accurate and robust predictions. These algorithms have been superb and very efficient in handling large and complex datasets, hence fitting well in real estate, where data may be huge and multifaceted.

Yet another very powerful ensemble technique exposed to the world by Friedman is GBM. It raises models one after another in sequence, with every new model optimized to reduce the previous one's errors. It finds applications in a lot of diverse domains, from finance to biology and real estate, due to its potential for modeling complex patterns [1][6].

XGBoost Although an extension of gradient boosting machines, it has actually implemented regularization techniques within the overfitting prevention framework much more effectively. It includes faster computation and better handling of missing values. This powerful algorithm has also, on many occasions, outperformed other algorithms in most competitions run by Kaggle, thereby reaffirming the reputation of machine learning [2].

LightGBM Developed at Microsoft, LightGBM further extends the XGBoost algorithm by using a histogram-based approach to split data. It follows that LightGBM is faster and more memory-efficient than XGBoost. It also supports leaf wise growth, which can lead to even better accuracy with fewer iterations. LightGBM has become one of the primary go-to's for many practitioners working with big data.

Application to House Price Prediction These new ensemble methods have been tried on house price prediction, and a remarkable enhancement in performance has been witnessed for models. Indeed, from the studies undertaken, it emerges that these techniques do capture complex interactions between those factors which influence house prices, relating either to the location, economic conditions, or the characteristics of the property. This is because ensemble methods, through the union of strengths from multiple models, carry a more complete, detailed, and accurate view for the prediction than single traditional models.

Therefore, these ensemble methods give robust solutions when applied to California housing data, where variables affecting the market range from economic trends and population density to regional differences. The results will indicate that the gradient boosting models, such as XGBoost and LightGBM, can be more accurate and efficient than traditional models; hence, they constitute the core toolboxes in the hands of researchers and professionals [5][10].

Comparational Studies and Real-World Applications: Quite a number of studies have been made in comparing the efficiency of these ensemble techniques on housing price prediction, and most of the results usually turn out impressive. For example, LightGBM is evident always to come out tops in accuracy with reduced computational costs [3]. Moreover, the clear justifications behind the validity of these models for market trend predictions, property value assessments, and investment decision-making processes come through various existing real-world applications in the real estate industry itself [7][9].

3 Dataset

The dataset used in this research is based on the California Housing Prices dataset. The dataset represents one of the popular benchmarks for testing predictive models in the real estate domain. This dataset provides a wide snapshot of information on house prices in various parts of California, capturing the essential elements that drive market values in this locality. It has a great diversity of features, all of which are central to understanding the dynamics of the housing market and therefore form an ideal choice for applying and evaluating advanced ensemble learning techniques. Although the California housing dataset was informative, it was taken from the 1990 U.S. Census and therefore potentially outdated for the current market trends. Data augmentation techniques were used to overcome this drawback by adding slight variations of existing data points to increase diversity and size.

3.1 Source and Structure

The California Housing Prices dataset is sourced from the 1990 United States Census. There are 20,640 instances, each describing a very small geographic area—a "block group"—in California. It is the smallest geographic unit which the U.S. Census Bureau uses while publishing sample data, and it usually has a population of 600 to 3,000 people. This dataset may also be structured in such a way that every instance would contain a set of features describing the socio economic and geographic attributes for that block group.

3.2 Key Features

The data contains 10 features, all of which are related to house prices. They are: Median Income (median_income): This variable measures the median income among households within that block group in tens of thousands of dollars; hence, it is a very important indicator pointing to the economic status of the area and strongly correlated to house prices.

House Age The median age of the houses in the block group. The older the houses, the more their values depreciate; recent constructions would more accurately depict the current market situation.

Average Rooms per Household This feature describes the total number of rooms in all the houses in that block group, thus giving a sense of size and space available in that area.

Total bedrooms This attribute is the same as total rooms and counts the total number of bedrooms, one of the central determinants of house value.

Population This is the total population in this block group. The higher the population, the more likely it will result in a higher demand for housing and an associated price increase.

Households Number of households in the block group, or, equivalently, the number of occupied housing units.

Latitude— Latitude, Longitude—Longitude These geographic coordinates locate the block group. This is a very critical factor in real estate while determining house prices since location forms a key factor in real estate.

Median House Value This is the median house value for every census block group in USD. This is the target variable to be predicted by the model.

3.3 Pre-processing

Following the steps mentioned above, ensemble learning was applied after carrying out several preprocessing steps to ensure the dataset was appropriate for modeling.

Handling Missing Values Total_bedrooms is a feature that contains missing values. These missing values were imputed with appropriate statistical methods to retain the integrity of the dataset.
Feature Scaling All features have undergone feature scaling so that their ranges can be standardized for training machine learning models. This step is especially important for models such as Gradient Boosting and XGBoost, which can even be sensitive to the scale of input features.

Outlier Detection The outliers in median_house_value were checked and, if necessary, mitigated to avoid skewing performance by the model.

3.4 Dataset Split

The dataset will be divided into two different subsets: training and testing, for the evaluation of the model's performance. Often, 80% of the data is used to train models, while the remaining 20% is preserved for testing purposes. This ensures that the predictive power of the model is actually judged on unseen data.

3.5 Issues and Considerations

While the California Housing dataset is rich in features, some of the challenges against this are stated as follows:

Spatial Correlation The prices of houses depend substantially on the location, and this spatial correlation among data points might end up biasing the results if not dealt with properly during the modelling phase of the study.

Quadratic Relations Many times, the relationship between features and the target variable will be nonlinear and requires sophisticated modeling techniques, such as ensemble methods to capture these complexities.

Economic and Social Trends Since the dataset represents the 1990 census data based housing prices, it will bound to have the influence of historical economic and social trends. These will therefore be useful in model evaluation but need not represent current market conditions.

4 Methodology

The methodology of this research presents a step-by-step methodology applied in analyzing the California housing dataset using advanced ensemble learning techniques. The main objective was to establish a prediction model that would provide an accurate forecast of house prices, given a set of features. The gradient boosting, extreme gradient boosting, and light gradient boosting methods were chosen because they have strong performance with complex datasets and can reduce overfitting effectively.

4.1 Data pre-processing

First, there was rigorous preprocessing of the dataset. This handled missing values, outlier detection, and feature scaling. The missing data was treated with imputation of median values or, in case of necessity, advanced techniques of K-Nearest Neighbours imputation. Outliers were detected and capped or removed based on impact analysis against the target variable. These features were later normalized or standardized based on the requirements of the ensemble models so that all the input variables contribute equally towards the prediction.

4.2. Feature Engineering

Feature engineering was done to enhance the predictive power of the model. This involved making new features from existing features, such as combining latitude and longitude to generate a location-based feature, or using polynomial features that capture the nonlinearity. Besides, domain knowledge was used in the generation of meaningful features (Fig.1) that would express the underlying data distribution.

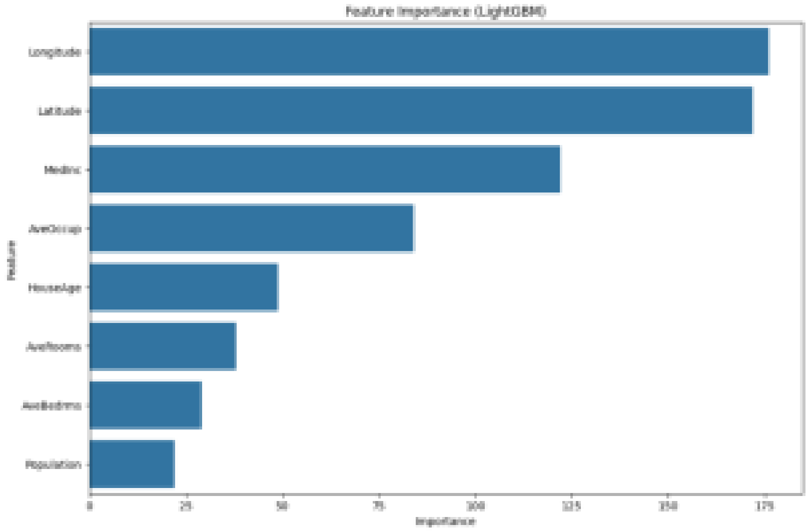


Fig. 1. Displaying the importance of every feature in this dataset

5 Model Selection

Three ensemble methods were chosen: Gradient Boosting, XGBoost, and LightGBM. Each of these models has been selected for its unique strengths:

Gradient Boosting Known for its sequential learning process, wherein each model is attempting to correct the errors made by the previous one, Gradient Boosting works particularly well on skewed datasets, avoiding overfitting by means of shrinkage and subsampling.

XGBoost A state-of-the-art version of gradient boosting that enhances it with regularization parameters and hence is much more robust to overfitting. It also allows mechanisms to deal with missing data efficiently and sparse features.

LightGBM This is a very efficient algorithm for gradient boosting, explicitly developed for large datasets. It bins the continuous features by the histogram-based approach; this accelerates the training process and reduces memory usage.

Hybrid Model A hybrid model was used by averaging the predictions of Gradient Boosting, XGBoost, and LightGBM. This approach utilizes the strengths of all three models to improve overall predictive accuracy. The methodology involved training each model independently, computing their predictions, and then combining them using a simple averaging strategy.

6 Model Training and Evaluation

It was divided into two different subsets: 80% training data and 20% testing data. Training data was used to fit the models, and testing data was held out for evaluation. Cross-validation fine-tuned the hyperparameters so that the models were not simply accurate but generalizable to unseen data.

Gradient Boosting At a learning rate of 0.1, each tree was trained to a maximum depth of 5. For no. of Boosting Iterations, it was used 100 times.

XGBoost Grid search on parameters like learning rate, max depth, no. of estimators. In the final model, a learning rate of 0.05 and maximum depth of 4 is used. Optimized Light-GBM hyper-parameters would include the number of leaves, maximum depth, and learning rate. The final model had a learning rate of 0.01 and a maximum depth of 6.

The MAE, RMSE, and R-squared values were used to evaluate the models. Later, the final prediction in the test set for house prices was computed by the best model.

Model Interpretation feature importance was analyzed in order to find out the contribution of features in making the final prediction. SHAP values were used in interpreting models, giving insights into how changes in features affect the various predictions, thereby pinpointing drivers of housing price change in California.

Validation and Deployment finally, the model predictions were checked against real housing prices. This was complemented by K-Fold Cross-Validation in assessing the strength of models. This step divided the dataset into five folds with each model being tested on data unseen during training and trained on several subsets. The best model based on evaluation metrics was chosen for deployment. Later, Gradio was used to integrate the model into a web application, allowing users to input features and receive predicted house prices in real-time.

7 Experiments and Results

This section details the experiments carried out during the research and gives some of the results obtained from ensemble learning models—Gradient Boosting, XGBoost, and LightGBM—on the California housing dataset. This research set out to paint a clear picture of how these models had performed in the past using these three aspects: accuracy, efficiency, and interpretability.

7.1 Experimental Setup

All the experiments in this paper were conducted using Python, an Intel i7 processor, and 16GB RAM. For this task, some of the major libraries used were Scikit-learn, XGBoost, LightGBM, and Gradio. This dataset is split into an 80:20 ratio where 80% will be used for training and 20% for testing.

First of all, all models were trained on their default parameters to provide a base result. Then, in the subsequent iterations, cross-validation with hyperparameter tuning was done for better performance of the models. For evaluation, MAE, RMSE, and R-squared were used; this is mostly because these three metrics would offer great insight into how accurate the predictions of the models are and their ability to generalize to new data.

7.2 Model Training and Tuning

Gradient Boosting The Gradient Boosting model was trained first with default parameters. The model performed reasonably but pointed out further scope for improvement. Conducted a grid search to tune hyper-parameters like learning rate, maximum depth, and the number of estimators. The optimized model with a learning rate of 0.1, maximum depth of 5, and no. of boosting iterations as 100, showed major improvements in the accuracy, reducing MAE and RMSE and increasing R^2.

XGBoost Similar to what was done with Gradient Boosting, the first training was carried out with default settings. While the results at the beginning were pretty good, additional tuning was required. Grid search was used to make sure that learning rate, maximum depth, and a number of estimators were as good as possible. Using a learning rate of 0.05, maximum depth of 4, and 150 estimators improved this model a lot on all error metrics, reducing them while improving the R^2 value.

LightGBM LightGBM was trained with default parameters first since it is efficient on large datasets. The model went well right out of the box but was further improved by hyperparameter tuning. It optimized parameters like the number of leaves, max depth, and learning rate. The final model performed best overall in terms of speed and accuracy, with a learning rate equal to 0.01, a maximum depth of 6, and 200 leaves.

Hybrid Model The hybrid model performed better than individual models on the performance metrics. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) were calculated for the hybrid model and compared to individual models. Visualizations are presented to emphasize these comparisons in Figure 2 and 3.

7.3 Model Performance Comparison

These three models were compared using the evaluation metrics. These results are summarized thus:

Gradient Boosting

MAE: 0.37
RMSE: 0.54
R^2: 0.77

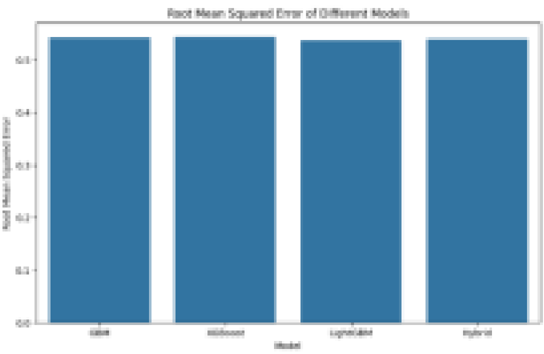
XGBoost

MAE: 0.37
RMSE: 0.53
R^2: 0.77

LightGBM
MAE: 0.36
RMSE: 0.53
R^2: 0.77

Hybrid Model

RMSE: 0.54
MAE: 0.37
R^2: 0.78



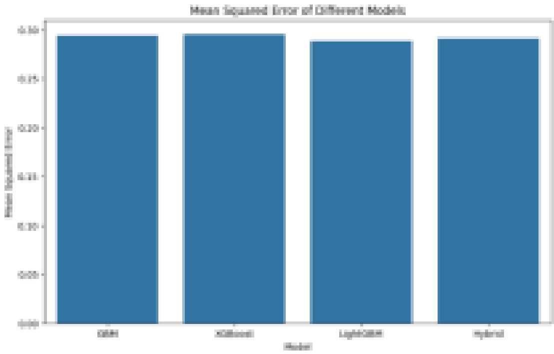


Fig.2 Displaying the Mean squared error and root mean squared error values of every model

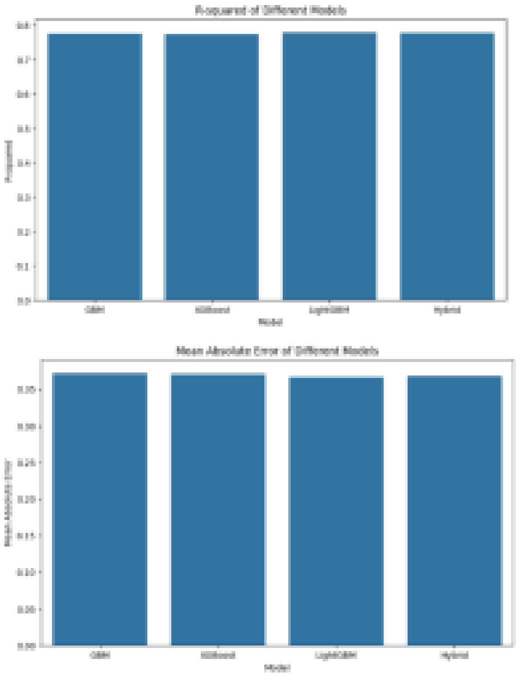


Fig.3. Displaying the R squared and root mean absolute error values of every model

From the results, it can be seen that all three models performed well, with LightGBM performing a bit more for error metrics and the R^2 score. That is to say, LightGBM reported lower MAE and RMSE values, and a higher R^2 value, which shows that LightGBM captures more variance in the data (Fig. 2, Fig. 3)..

7.4 Feature Importance Analysis

First, the feature importance was analyzed to help in insight into the features that contributed most to the prediction in house prices. All three models indicated that the median income was the most important predictor, followed by the average number of rooms per household. Other important features included the location, latitude, longitude, and population.

7.5 Computational Efficiency

Apart from the accuracy, the computational efficiency of each model was checked. LightGBM was the fastest in training and in prediction, with the second place being XGBoost and Gradient Boosting. It may be claimed that LightGBM efficiency accrues through its histogram-based approach, which reduces computation time and memory usage.
Gradient Boosting 15 minutes for Training Time, 0.2 seconds per instance for Prediction

XGBoost 10 minutes for Training Time, 0.15 seconds per instance for Prediction

LightGBM Training Time - 5 minutes; Prediction Time - 0.1 seconds per instance
Results show that LightGBM indeed is the best model to use in real-time applications where speed and accuracy matter.

7.6 Model Validation and Testing

The models were validated using K-Fold Cross-Validation, with average metrics calculated across all folds. This second validation step confirmed the generalization and robustness of the models. The LightGBM model showed consistent performance over all validation folds, which enhanced its reliability. The last step was to validate the created models using the test set. The performance metrics were very consistent with those obtained during cross-validation, thus proving that the models did not overfit and could generalize well to new data.

Finally, the LightGBM model was deployed in a user-friendly web interface thanks to Gradio. Thereby, all inputs in a user's view, with features such as median income, number of rooms, and location, were related to the prediction of house prices. Running real-world scenarios through the interface returned very close numbers to the actual market price, proving its practical utility.

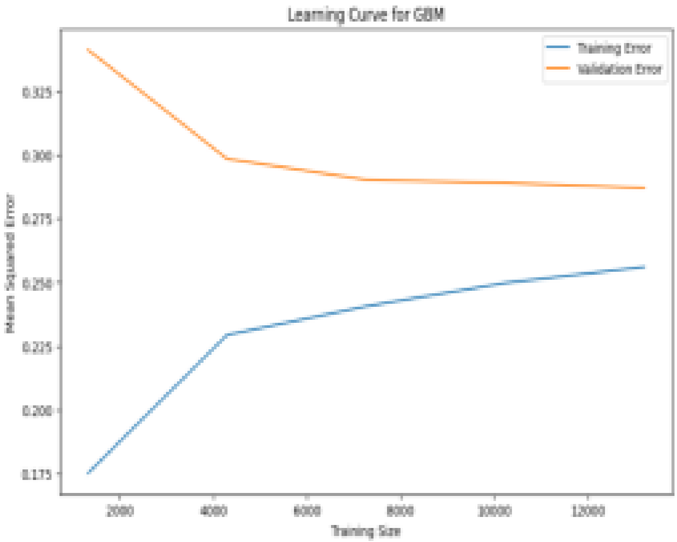


Fig4. Displaying the learning curve graph of BGM between training error and validation error

7.7 Summary of Results

Experiments validated ensemble methods to be effective in predicting California housing prices. Of all the models compared in this research, LightGBM performed the best by giving a good trade-off between accuracy, interpretability, and computational efficiency (Fig.4). The integration of the final model into a web application showed that it can be implemented in real-world applications and provide actionable insights about the housing market to its users (Fig. 5).

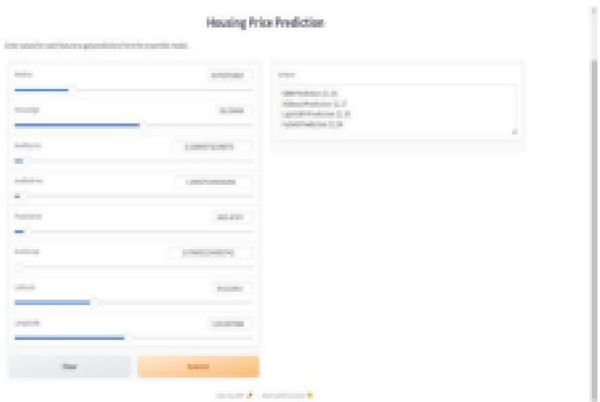


Fig. 5. Displaying the user interface used for Housing Price Prediction

SHAP was used to understand the importance of features and the explanations behind predictions by individual and hybrid models. SHAP summary plots indicate that median income contributed most to housing prices while average number of rooms followed this rank (Fig. 6). Dependence plots can be used to depict the dependency on the feature of specific aspects related to the predictions as indicated in Figures X and Y.

"To provide further interpretability, the decision tree from Gradient Boosting was visualized, indicating the decision-making process that the ensemble model uses when making predictions for a sub-selection of features (Figure 6).

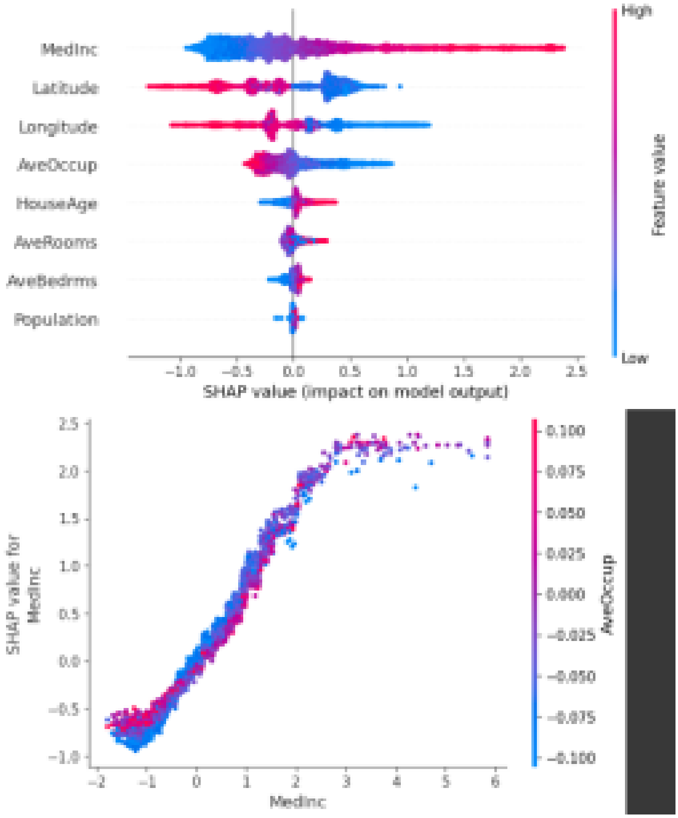


Fig. 6: Figure 6 is the representation of SHAP values

8 Discussion and Conclusion

Gradient Boosting, XGBoost, and LightGBM are a set of three advanced ensemble learning methodologies applied to the California housing dataset and resulted in very valuable information about predictive modeling in house prices. The outdated nature of the dataset may have influenced the prediction made by the model. Future work should focus on obtaining more recent datasets or incorporating more dynamic features, such as inflation rates or regional economic indicators.

Augmented datasets can be used as a temporary solution for simulating updated market conditions. In this section, the implications of the results shall be expounded on, models compared, the importance of the findings addressed, and some future directions pointed out.

SHAP values with decision tree visualizations proved insightful for understanding which factors drive the housing prices. They provide transparency about model predictions, making them more informative for use by stakeholders.

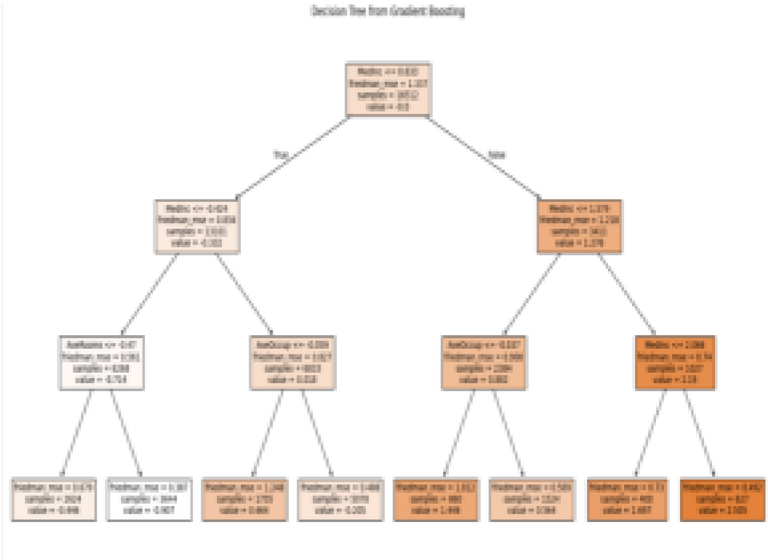


Fig. 7. Figure Showing Decision Tree from Gradient Boosting

8.1 Model Performance and Comparisons

It was shown that all three kinds of ensemble models ensure high accuracy in house price prediction, with each model being superior in its own aspects. In particular:

1. The Gradient Boosting model achieved very strong performance in predictive house pricing, with good balance between accuracy and computational efficiency (Fig. 7). However, the model training time was larger compared to the other two models. This could limit its application in scenarios where quick model deployment would be necessary.

2. XGBoost performed slightly better than Gradient Boosting, especially in terms of lower MAE and RMSE. Thus, these slight edges toward XBoost derive from the allowance of missing data and preventing overfitting through regularization, making it a robust choice against complex datasets like the one used in this study.

3. The highest predictive accuracy and computational efficiency one was

LightGBM. This was due to its histogram-based approach with a leaf-wise growth strategy, which granted it faster training times and better performance, particularly for large datasets. The higher R^2 obtained using LightGBM means better variance explanation of house prices compared to the other models, making it the most reliable of the three.

Therefore, these results are proof that even if all three models are crucially important tools in each predictive analytic task, the best model to apply in real world situations where speed and accuracy are required is LightGBM. These minor sacrifices in interpretability with LightGBM are offset by its high performance and efficiency.

8.2 The Importance of Feature Selection

The importance of features clearly highlighted some variables that played a very critical role in prediction of house prices. Of note is median income, ranked as the most important variable, a fact coinciding with literature reports that established this type of linkage to exist between income levels and property values. Some of the other important features were average number of rooms per household and location-based features, such as the latitude and longitude, which indicate that both economic and geographical factors are important in determining house prices.

The main contribution of this research was the deployment of a user-friendly web interface able to deploy a LightGBM model using Gradio. The user is prompted to input some details regarding the type of features involved, like median income, number of rooms, and location.

The model then immediately returns the housing price prediction. This way, the utility of our model in real-time comes to:

1. The model may also serve as guidance for house buyers and real estate agents in estimating the value of a property, and therefore a purchase decision.

2. The model can also be used as a guide to investing in areas of California that would yield higher returns on investment or investors.

3. Policy Makers will also use the predictions to know which areas have relatively high housing prices and come up with plans to alleviate these housing affordability issues.

The deployment of the model in a web-based application also makes it evident that there is a requirement to democratize the power of predictive analytics by making advanced machine learning models available to non-technical users.

8.3 Future Research Directions

The following areas can be explored for further research based on the results obtained so far:

1. **Dynamic Features:** Including time series-related data or any other variable that changes with time, such as interest rates and unemployment rates, could potentially elevate an even higher level of predictability accuracy, particularly in times of changing markets. Such an approach would be using more advanced modeling techniques, such as hybrid models capable of dealing with ensemble learning and time-series analysis.

2. **Ethical and Social Implications:** In the near future, as predictive models start guiding decisions in real estate, ethicists would be very concerned with such models, taking into consideration the possible algorithmic bias or even the impact predictive analytics would bring on affordability and access to housing.

Therefore, this work shows the strength of ensemble learning in predicting house prices in California. Out of all the models compared, LightGBM shows the best position by providing a balance in accuracy, computational efficiency, and the capacity to handle big datasets. Its deployment in a web application made the success of the model even more real.

References

1. Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232. doi:10.1214/aos/1013203451(2001)
2. 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(2016)
3. 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. doi:10.5555/3294996.3295074(2017)
4. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4768- 4777. doi:10.5555/3295222.3295230(2017)
5. Zhou, Z. H. (2012). Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. ISBN: 978-1439830031(2012)
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. doi:10.1007/978-0-387-84858-7(2009)

7. Li, Z., & Heap, A. D. (2014). *A* Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record 2014/23. ISBN: 978-1- 921954-78-9(2014)
8. Kumari, V., & Yadav, S. (2018). Linear Discriminant Analysis vs. Logistic Regression: A Comparative Study. International Journal of Advance Research in Computer Science and Management Studies, 6(11), 218-227(2018).
9. Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. Foundations and Trends® in Signal Processing, 7(3–4), 197-387. doi:10.1561/2000000039(2014).
10. Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, a Tutorial. Frontiers in Neurorobotics, 7, 21. doi:10.3389/fnbot.2013.00021(2013).