**Journal of Business and Management Studies**
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# Explainable AI and Machine Learning Model for California House Price Predictions: Intelligent Model for Homebuyers and Policymakers

**MD Azam Khan[1] ✉ Pravakar Debnath[2], Abdullah Al Sayeed[3], Md Fakhrul Islam Sumon[4], Arifur Rahman[5], MD Tushar Khan[6] and Laxmi Pant[7]**

*[145]School of Business, International American University, Los Angeles, California, USA*
*[2]School of Business, Westcliff University Irvine, California, USA*
*[3]Masters of Business Administration in Project Management, Central Michigan University*
*[6]Masters of Science in Business Analytics, Trine University*
*[7]MBA Business Analytics, Gannon University, Erie, PA, USA*

**Corresponding Author:** MD Azam Khan, **E-mail**: khanazamgro@gmail.com

**| ABSTRACT**

California's housing and real estate market is one of the most valuable markets in the USA. Many shareholders such as individual homebuyers, home sellers, real estate agents, lenders, and policymakers depend on high-volume information regarding the dynamics at work and their correct estimation. The research project aimed at developing an Explainable AI machine-learning model for California house price predictions. Data on house prices were collected from reliable sources such as California home estate websites, land sites, and public datasets. Features of the data included location, size, number of rooms, area type, availability, sale prices, and oceanic proximity. In this research project, credible, proven, and renowned machine learning algorithms were used most notably, Linear Regression analysis, XG-Boost, and Random Forest. The Random Forest came up quite impressively with a superior accuracy score and low MAE and MSE; thus, it was good for learning the underlying best patterns and relationships that may exist within the data for house price predictions. XG-Boost also did relatively well, showcasing moderately high accuracy and relatively low MSE and MAE, compared to the Linear Regression.

**| KEYWORDS**

Explainable AI; California house pricing; Home buyers; Home Sellers; Random Forest; XG-Boost; Linear Regression

**| ARTICLE INFORMATION**

**ACCEPTED:** 01 September 2024 **PUBLISHED:** 14 September 2024 **DOI:** 10.32996/jbms.2024.6.5.9

## 1. Introduction

The California housing and real estate market is one of the biggest and most indispensable markets in the USA. Stakeholders, including homebuyers, sellers, real estate agents, lenders, and policymakers, rely on proper information about the dynamics at work and their correct estimation. Although traditional hedonic regression models are illuminating in this respect, they suffer from limitations due to the potential unavailability and complexity of data. Recent advances in machine learning and large amounts of available real estate transaction data power more powerful predictive modeling (Chen et al. 2021). However, one of the major challenges with most ML models pertains to the issue of explainability, reducing transparency and trust in the insights derived from those models. This paper aims to fill this gap by discussing a model for explainable machine learning that is balanced between the accuracy of house price prediction and interpretability for California.

This research project proposes the application of explainable AI and Machine Learning techniques, which shall build into an intelligent, and transparent predictive model for house price estimation in California. The proposed model shall bring predictive visibility concerning the dynamics in house pricing concerning a variety of factors rightly to the homebuyers and policy framers in

terms simple to comprehend. AI models using ML are built on the California house sales transaction data and relevant economic, demographic, and other home-specific datasets. Popular explainable Machine Learning techniques are used to explore and explain the model predictions(Chen et al. 2021). The explained model will provide transparency to the pricing predictions to homebuyers for informed decisions and provide insights to policymakers on housing market trends to craft policies. In general, the development of a model shall balance prediction accuracy with explainability to enhance trust and usability by stakeholders.

### *1.1 Problem Statement*

The California housing market presents a myriad of challenges attributed to its volatility and complexity, driven by various factors such as demographics, economic conditions, and government policies. Traditional predictive models are ineffective in terms of interpretability, which constrains homebuyers from making proper decisions as well as policymakers to design efficient housing strategies. The incomprehensibility of the machine learning models further exacerbates this issue, as stakeholders struggle to make sense of the influencing factors that necessitate house prices (Chordia, 2022). Therefore, this calls for an urgent need for an Explainable AI framework that shall employ advanced machine learning mechanisms to predict the prices of houses while explaining the decisions made by these models. This study, therefore, seeks to fill this gap by constructing a robust and interpretable machine learning model specifically for house price prediction in California, which will enable homebuyers to make better, well-informed decisions and support policymakers as they develop acts regarding housing policies.

This research project explores the designing of an Explainable AI machine-learning model for house price prediction in California. The proposed model is targeted to be much in line with advanced algorithms and interpretable elements to provide comprehensive insights on decision-making for homebuyers and policymakers. It focused on the aspect of explainability within AI models, showed a broad methodology concerning price prediction, and discussed the implications for stakeholders in the housing market.

## 2. Literature Review of Related Works

### *2.1 Explainable AI*

As per IRJET (2023), explainable AI is a fast-emerging research area concerned with the development of techniques and methods that make the decision-making process of AI transparent, interpretable, and understandable by human beings. While AI technologies have become integral to most critical facets of human life, including real estate, healthcare, finance, and criminal justice, demands for interpretable models have grown high of late. Traditional AI models, mainly in algorithms relating to deep learning, often work out results as "black boxes" without giving any insight into how such an outcome was achieved. This opacity raises several difficulties, especially when AI systems make decisions affecting human life. XAI seeks to bridge this gap by providing techniques and tools that explain the internal mechanism of AI models, and thus consolidate them with much more transparency and trust.

One of the prime motivations behind Explainable AI [XAI] is developing trust in AI systems. Once users can make sense of how decisions are made by a model, they are more likely to use them and have more faith in these technologies. In domains like healthcare, trust will be very crucial when AI might be used for diagnosis or treatments (IRJET, 2023). In these cases, the practitioners have to know why the AI system recommended a particular diagnosis to make informed decisions. Explainable AI includes techniques such as SHAP and LIME, which explain feature importance and individual predictions, respectively. Such tools tend to make complex models more interpretable, providing a clearer understanding of how input features relate to the outcomes.

### *2.2 Machine Learning in Real Estate*

In their empirical study, Jain et al., (2020), by adopting the Colorado housing dataset using four different regression algorithms, namely Lasso Regression, Logistic Regression, Decision Tree, and Support Vector Regression. Considering all these algorithms with different error metrics like R-squared value, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error, amongst all the estimated algorithms, the Decision Tree algorithm showed the best results when all these algorithms were used, having the highest accuracy score of 86.4 and lowest error values. On the other hand, the worst performance was achieved by Lasso Regression, which only elicited an accuracy score of 60.32.

Kumar et al. (2021), presented the concept of classification algorithms in predicting the resale value of houses. This study predicted property-selling prices by considering multiple classification models: Linear Regression, Decision Tree, K-means, and Random Forest. A house price was determined by the subject's physical characteristics, location, and general economic scenario. Testing these methods was done using RMSE as the performance metric across different available data sets concerning finding the most accurate model to improve the predictions.

Madhuri et al. [2019] presented a comparison of six machine learning housing price prediction models, most notably, Linear Regression, Decision Tree, Random Forest, Gradient boosting, Support Vector Machine, and Neural Network. The author applied the respective models to Melbourne real estate data. The best predictions in housing prices emerged to be neural network models.

An empirical study by Phudinawala (2024), illustrated the superiority of ML algorithms such as Random Forests and Gradient Boosting Machines (GBM) over conventional methods. It was through the introduction of a wide array of variables, including economic indicators, property characteristics, and location data in these models, that superior results in real price predictions were achieved. Besides, the author demonstrated that neural networks were able to represent complex nonlinear relationships to further increase predictive performance.

A study by Rana et al. (2020), in which the authors used k-means clustering to segment neighborhoods. according to their investigations socio-economic and housing-demand factors were beneficial for real estate practitioners in making market strategy and investment decisions, and have been proven to identify emerging trends and market opportunity.

Studies by Sinha. (2020) employed SVM and logistic regression for the probability of default in mortgages. These models gave a better risk assessment by considering data on borrowers, economic conditions, and property features, and fared much better than the traditional system of credit scoring. This application is very important to lenders in their effort to reduce financial risk and make rational decisions on lending.

## 3. Methodology

### *3.1 Data Collection*

In this stage, relevant data about house Prices were collected from reliable sources such as California real estate websites, land sites, and public datasets. The data features included location, size, number of rooms, area type, availability, sale prices, and oceanic proximity. Besides, it was also imperative that the dataset had characteristics, such as Location, cover area, built-up area, age of the property, postal state, and so forth (Pro-AI-Rokibul, 2024). The researcher ensured that the data was representative of the target and diversified. Dataset legitimacy was of paramount prerequisite in our research project.

### *3.2 Data Pre-processing*

This preliminary data processing mainly involved cleaning the collected data and making it ready for modeling. Some key activities included handling missing values, removing outliers in the dataset, normalizing numerical features, and encoding categorical variables. Feature selection techniques were also employed to identify the most relevant attributes that bear close relation to the prediction of house prices (Pro-AI-Rokibul, 2024). Stratified sampling data splitting techniques were subsequently employed to break down the cleaned information into training and testing datasets that were later used in the fitting models and objectively assessing their predictive accuracy.

### *3.3 Feature Engineering*

Scaling the features contained in the dataset is recommended as a best practice in most experiments so that they are all on a similar scale. This can be done using methods such as standardization or normalization. The California house price dataset had a host of features, and it was suitable for feature engineers by choosing only the relevant features for training the model (Pro-AI-Rokibul, 2024). This was done using techniques such as forward selection or backward elimination. Subsequently collected dataset is encoded so that it may be used in any model training. One such way could have been the use of one-hot encoding. The method appends a new binary column to the list of unique categories of each variable.

| Features | Description |
|---|---|
| Buying Price | Transaction price/sqft [RM] |
| Selling Price | Disposing price/sqft [RM] |
| GC | Green Certificate |
| Floor | Floor |
| MFA | Main floor area |
| Distance | Distance from CBD |
| Bed | Number of bedrooms |
| BC | Building Category |
| CA | Category Area |
| Age | Age of the building |
| AC | Area Classification |
| Sell | Seller |
| Buy | Buyers |
| Ownership | Own |
| Ocean Prox | Oceanic proximity |

***Table 1: Displays Feature Engineering Scaling***

| Avg. Price/ Sqft | lotsize | Bedrooms | bathroom |
|---|---|---|---|
| 2,371 | 850 | 3 | 1 |
| 3,761 | 4000 | 4 | 2 |
| 3,724 | 3060 | 3 | 1 |
| 3,724 | 6650 | 3 | 1 |
| 3,097 | 6360 | 3 | 2 |
| 4,093 | 7383 | 6 | 3 |
| 3,396 | 6734 | 5 | 2 |
| 3,396 | 9866 | 4 | 1 |
| 3,396 | 7888 | 4 | 1 |

***Table 2: Showcases Price Factors***

### *3.4 Model Training*

The prime focus of this stage was to implement various machine learning algorithms to construct a predictive model capable of estimating house prices. The models were trained on preprocessed data by optimizing their parameters and fitting them to a subset of training data. Furthermore, tuning hyperparameters was also performed with the view of improving predictive accuracy. Performance was measured by appropriate metrics-mean squared error or R-squared-which determined how well each model learned from the information provided to generalize to unseen data during the actual model-building process.

In the modeling phase of the project execution, most of the preprocessed data comprised the training which was utilized to train and fit the machine learning algorithm as it learned the pattern and relationships between predictive input variables such as the number of rooms, location, square footage among others and the goal or target variable of house prices. By processing the information and correlations in the training subset, the model learned how variations in those explanatory attributes were associated with changes in housing values. Subsequently, it enabled the model to make an intelligent prediction regarding a particular property's estimated market price based on logic and inferences of decisions developed during training from the explanatory patterns within the training dataset.

### *3.5 Algorithms Implemented*

In our research project, proven and popular machine learning algorithms were applied, most notably, Linear regression analysis, Lasso regression, XGBoost, random forest, and support vector machines. Each algorithm was trained separately using part of the preprocessed data. Then, the trained models were tested on an independent dataset for their predictive accuracy. This therefore gave us the avenue to assess various modeling approaches and select the best-fitting solution for our house price forecasting problem. Among the pre-processing and testing activities was the tuning of hyperparameters for the algorithms using many methods like cross-validation and grid search that optimize performance. It gives an objective comparison and evaluation of the predictive capability and accuracy of various trained models by computing different evaluation metrics, which include mean squared error, root mean squared error, and R-squared.

**I. XG-Boost**

XG-Boost is among the most efficient and popular algorithms of machine learning for solving regression and classification problems. It is an efficient library interfacing with graduate-boosted decision trees for production goals, aiming at speed and accuracy. XG-Boost adopts this sequential ensemble approach whereby each added tree in the sequence reduces the progressive errors that were made by the predecessors (Sinha, 2020). It further introduces the regularization methods to ensure that there is no overfitting during training, especially in highly dimensional and very big data. Some of the reasons it finds such extensive use in competitive modeling situations are parallel processing, tree pruning, and handling missing values.

**II. Random Forest**

Random Forest is an ensemble learning method that could be used for classification and regression. It constructs many decision trees during the training process and combines their results anticipating better accuracy with a reduced risk of overfitting. At each step, a random subset of data and features is used to form trees, which increases the diversity among trees (Rana et al., 2020). The final prediction is conducted by averaging the results (regression) or taking a majority vote (classification). It reduces variance and enhances the generalizing capability of the model on complicated datasets and therefore improves its robustness and performance when there are noise and missing values.

**III. Linear Regression**

Linear regression is a form of supervised machine learning in which an ongoing output is normally forecasted for a given variable input or sometimes multiple variable inputs. It models the relationship between the dependent variable, otherwise known as the target, and the independent variables, also known as the features, by fitting a linear equation to the observed data. That is, one variable will provide a line, and several will provide a hyperplane that minimizes the sum of the squared differences between the predicted and actual values (Chordia, 2022). This model trains a coefficient estimate for every input feature, with the goal that it captures the linear relationship that underlines it. Linear regression has become a staple in many developers' toolkits because it's very simple to conceptualize and interpret, and, where appropriate, can be used as the basis for predictive analytics.

## 4. Experimentation Results

### *4.1 Importing Libraries*

```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
import warnings

# Ignore all warnings
warnings.filterwarnings('ignore')
```

**Output:**

| | longitude | latitude | housing_median_age | total_rooms | total_bedrooms | population | households | median_income |
|---|---|---|---|---|---|---|---|---|
| **0** | -122.23 | 37.88 | 41.0 | 880.0 | 129.0 | 322.0 | 126.0 | 8.3252 |
| **1** | -122.22 | 37.86 | 21.0 | 7099.0 | 1106.0 | 2401.0 | 1138.0 | 8.3014 |
| **2** | -122.24 | 37.85 | 52.0 | 1467.0 | 190.0 | 496.0 | 177.0 | 7.2574 |
| **3** | -122.25 | 37.85 | 52.0 | 1274.0 | 235.0 | 558.0 | 219.0 | 5.6431 |
| **4** | -122.25 | 37.85 | 52.0 | 1627.0 | 280.0 | 565.0 | 259.0 | 3.8462 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **20635** | -121.09 | 39.48 | 25.0 | 1665.0 | 374.0 | 845.0 | 330.0 | 1.5603 |
| **20636** | -121.21 | 39.49 | 18.0 | 697.0 | 150.0 | 356.0 | 114.0 | 2.5568 |
| **20637** | -121.22 | 39.43 | 17.0 | 2254.0 | 485.0 | 1007.0 | 433.0 | 1.7000 |
| **20638** | -121.32 | 39.43 | 18.0 | 1860.0 | 409.0 | 741.0 | 349.0 | 1.8672 |
| **20639** | -121.24 | 39.37 | 16.0 | 2785.0 | 616.0 | 1387.0 | 530.0 | 2.3886 |

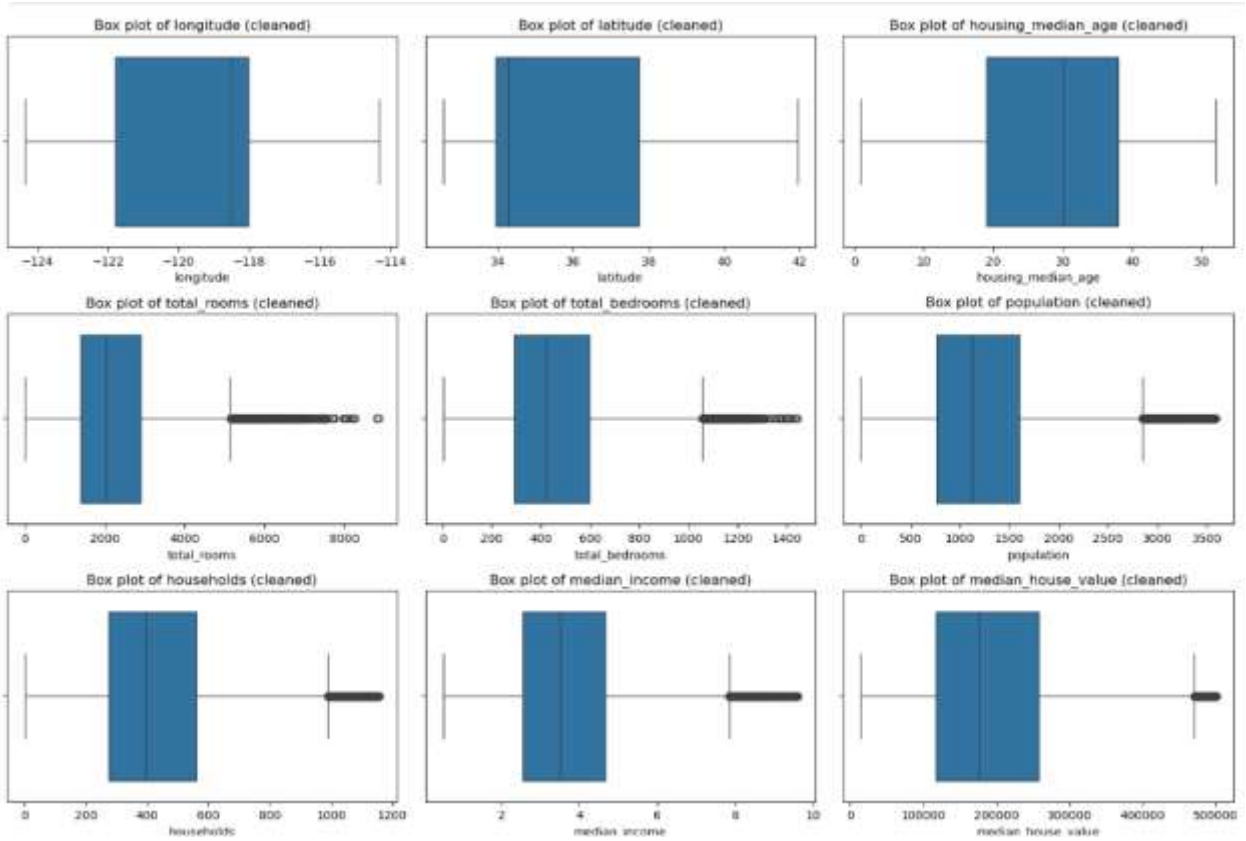
As showcased above, structural changes were performed during the loading process of data to be organized in a way that it was correctly stocked for input specifications for each algorithm. The dataset initially consisted of rows of attributes including longitude, median housing age, latitude, total rooms, total bedrooms, population, households, and median values. Some of the structural changes were done on the attributes in a way to prepare them for use as inputs to the classification algorithm by considering the nature of predictor variables or features of the dataset at hand.

To visualize the cleaned data respective code snippets were applied by the analyst to visualize the cleaned data using box plots as showcased below:

```
# Visualize the cleaned data using box plots
plt.figure(figsize=(15, 10))
for i, column in enumerate(numeric_columns, 1):
    plt.subplot(3, 3, i)
    sns.boxplot(x=df_cleaned[column])
    plt.title(f'Box plot of {column} (cleaned)')

plt.tight_layout()
plt.show()
```

**Output:**



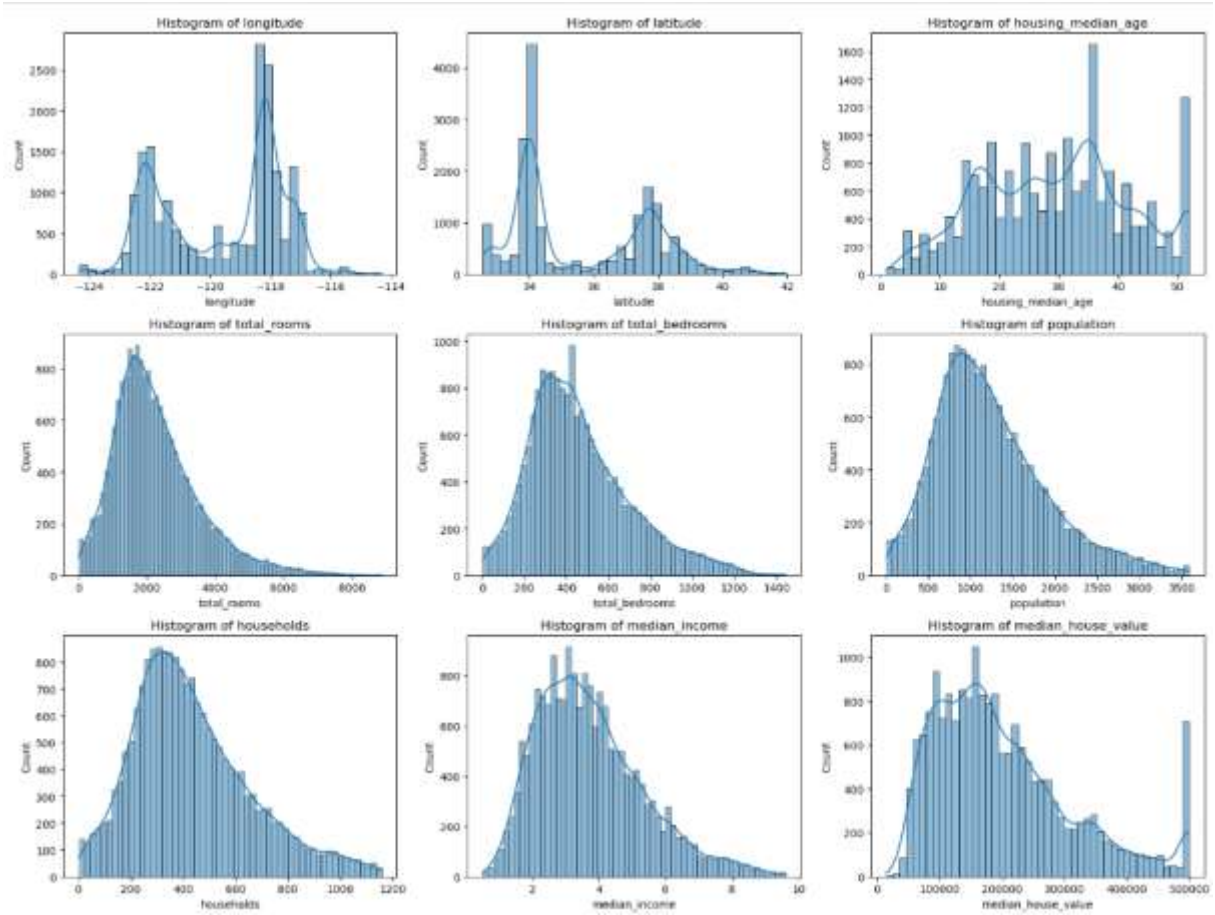

To understand the distribution of the dataset, a code snippet was imposed to generate plot histograms as showcased below:

```
# Plot histograms for numerical columns to understand the distribution
numeric_columns = ['longitude', 'latitude', 'housing_median_age', 'total_rooms',
                   'total_bedrooms', 'population', 'households', 'median_income',
'median_house_value']

plt.figure(figsize=(16, 12))
for i, column in enumerate(numeric_columns, 1):
    plt.subplot(3, 3, i)
    sns.histplot(df_cleaned[column], kde=True)
    plt.title(f'Histogram of {column}')

plt.tight_layout()
plt.show()
```

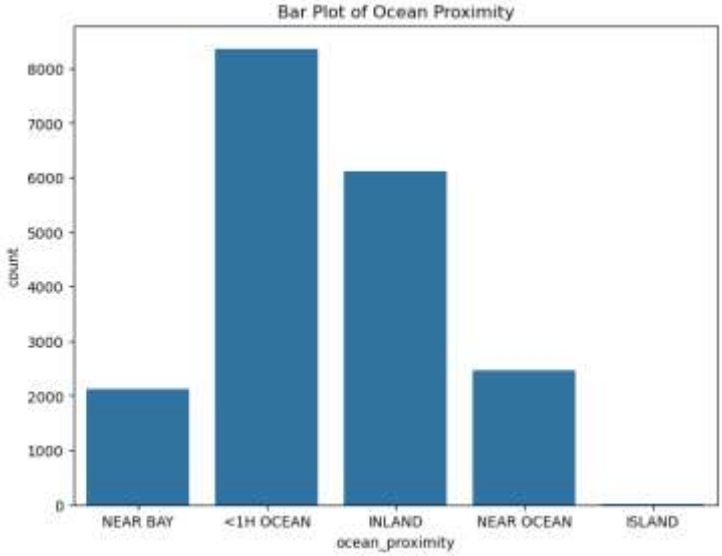
**Output:**




To ascertain the correlation between house pricing and ocean proximity, a suitable code snippet was imposed to generate a bar plot for the categorical column as displayed below:

```
# Bar plot for categorical column 'ocean_proximity'
plt.figure(figsize=(8, 6))
sns.countplot(x='ocean_proximity', data=df_cleaned)
plt.title('Bar Plot of Ocean Proximity')
plt.show()
```

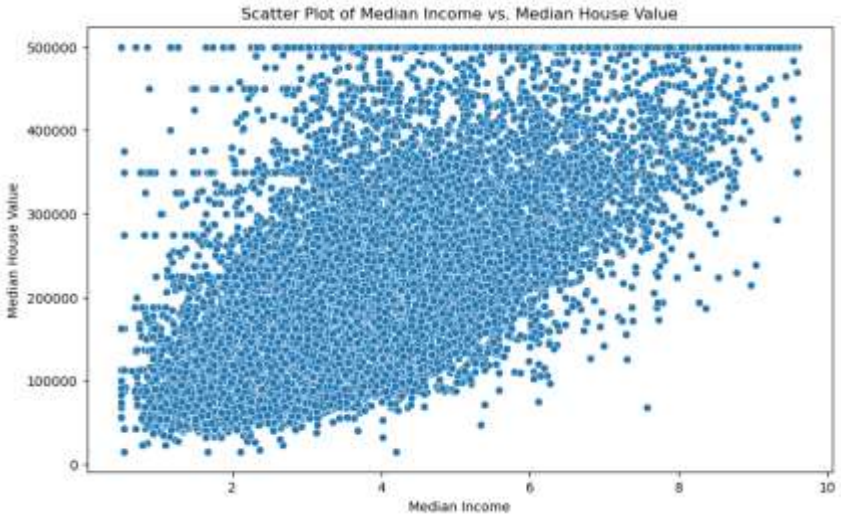
**Output:**



To determine the relationship between median income and median house value, appropriate scatter plots were generated to visualize the association between key numerical variables as displayed below:

```
# Scatter plots to visualize relationships between key numerical variables
plt.figure(figsize=(10, 6))
sns.scatterplot(x='median_income', y='median_house_value', data=df_cleaned)
plt.title('Scatter Plot of Median Income vs. Median House Value')
plt.xlabel('Median Income')
plt.ylabel('Median House Value')
plt.show()
```

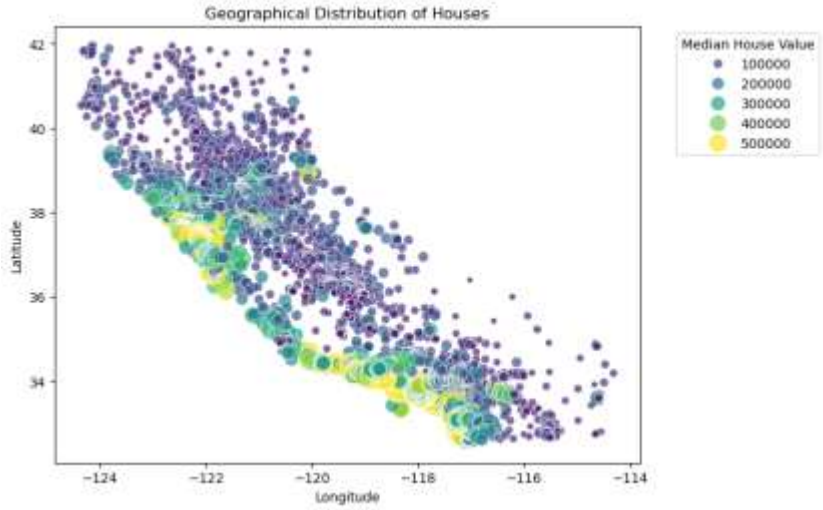
**Output:**




Apart from that, the analyst was also keen to generate a geographical visualization of the targeted houses, comparing longitudes and latitudes, as exhibited below:

```
# Geographical Visualization: Longitude vs Latitude
plt.figure(figsize=(8, 6))
sns.scatterplot(x='longitude', y='latitude', hue='median_house_value',
palette='viridis', size='median_house_value', sizes=(20, 200), data=df, alpha=0.7)
plt.title('Geographical Distribution of Houses')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.legend(title='Median House Value', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

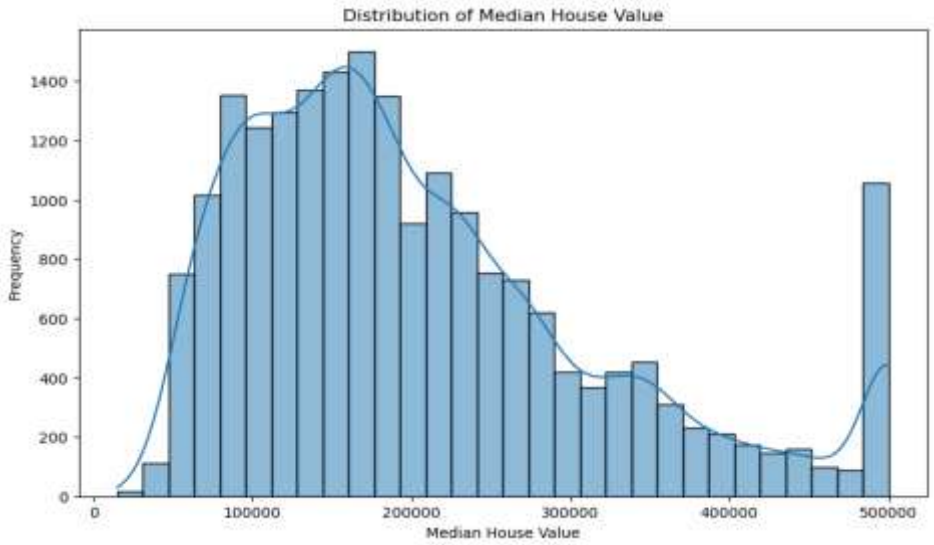
**Output:**




Moreover, the analyst was also keen to pinpoint the distribution of median house value where suitable code snippets were applied to produce histograms of median house value:

```
# Univariate Analysis
## Histogram of Median House Value
plt.figure(figsize=(10, 6))
sns.histplot(df['median_house_value'], kde=True, bins=30)
plt.title('Distribution of Median House Value')
plt.xlabel('Median House Value')
plt.ylabel('Frequency')
plt.show()
```

**Output:**




### *4.2 Performance Evaluation Metrics*

The trained models were evaluated using a separate testing dataset that had been utilized in the training phase. Calculations of performance metrics included mean squared error or MSE, mean absolute error or MAE, and R-squared-all to establish the predictive power and accuracy of each model. These metrics correlated to providing important insights into the models' efficacy in generalizing to new, unseen data.

### *4.2.1 Mean Absolute Error (MAE)*

The MAE is a measure of the average magnitude of the errors within a set of predictions without regard to the direction of the errors. That is, it is the average absolute difference between the predicted values and the actuals. It reflects better model performance in case the MAE is lower.

### *4.2.2 Mean Squared Error (MSE)*

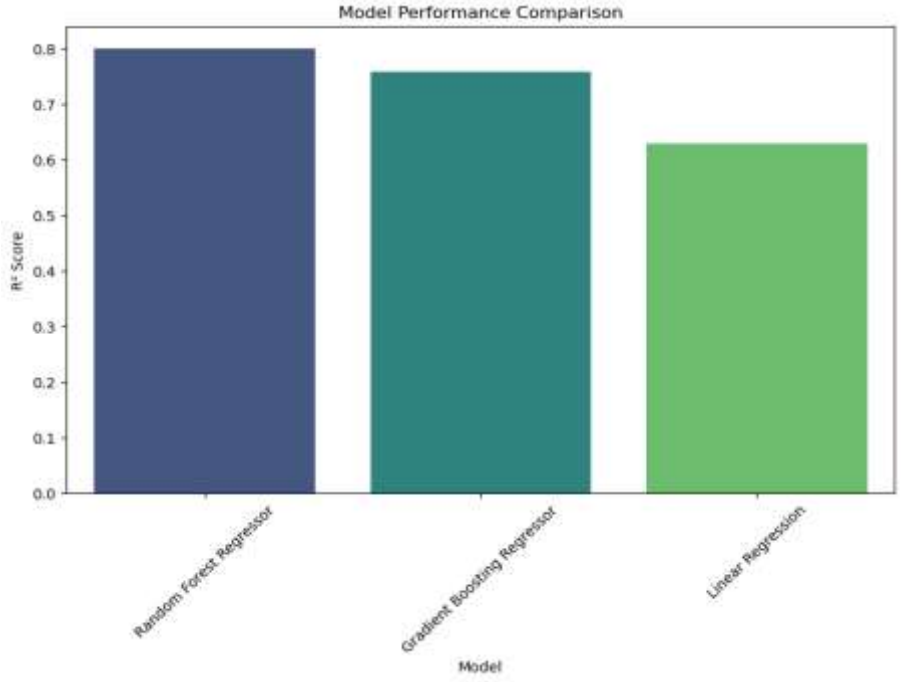
Mean Squared Error is one of the most used measures in regression analysis, representing the mean of the squared difference between the forecasted and actual values. By underscoring errors, MSE provides a comprehensive evaluation of prediction accuracy. The formula of MSE includes the sum of squared errors divided by the number of records, and can be expressed as follows:

### *4.2.3 R-Squared ($R^2$)*

R-squared ($R^2$), or the coefficient of determination, describes what percent of the variation in the dependent variable is described by independent variables in a regression model. It is considered a measure of model fit because it records how well the model curves fit the data set. The $R^2$ is the ratio of explained variance against the total variance, expressed as:

**Model Performance Summary**

| Model | Mean Absolute Error (MAE) | Mean Squared Error (MSE) | $R^2$ Score |
|---|---|---|---|
| **Linear Regression** | 50,291.37 | 4,542,940,578.84 | 0.63 |
| **Random Forest** | 32,589.59 | 2,450,928,245.67 | 0.80 |
| **Gradient Boosting Regressor** | 37,894.89 | 2,972,837,585.10 | 0.76 |




From the table and chart above, the MAE for the Random Forest Regressor is lower than that of the Linear Regression model, meaning that, on average, predictions from the Random Forest model are closer to the actual values. The MSE is also comparatively lower in the case of a Random Forest model than in the case of the Linear Regression model, again suggesting that the Random Forest model has fewer large errors. Hence, for the Random Forest model, the $R^2$ score amounts to 0.80, specifying that 80% of the variance of the target variable is accounted for by this model, whereas its value is higher on average in comparison with the Linear Regression model and corresponds to better performance.

Concerning, Gradient Boosting Regressor, the mean absolute error was higher than in the Random Forest model but lower than in Linear Regression. This implies that this model has a pretty reasonable performance in terms of average prediction error. MSE on the Gradient Boosting model is higher compared to the Random Forest model, indicating that it does have slightly higher errors

in predictions. A higher $R^2$ score of 0.76 implies that 76% of the variance in the target variable is explained by the Gradient Boosting model. It, therefore, falls between the score of the Linear Regression and Random Forest models, hence it would indicate a fairly good fit.

The impressive performance demonstrated by both Random Forest and XG-Boost can be explained by their capability of handling high-dimensional datasets, capturing complex relationships, and managing efficiently the interaction between features. These algorithms represent the class of tools very well known for robustness, scalability, and adaptability across many different machine learning applications.

### *4.3 How to Implement the Models*

**Step 1: Understand the Business Objectives-** Define Goals: Determine what specific outcomes the business wants from the model (e.g., House price predictions, market trends).
**Step 2: Data Collection**- Access data and gather Data on California house prices, with specific features for house location, square footage, number of bedrooms, oceanic proximity, bathrooms, age of property, and local amenities.
**Step 3: Data Preparatio**n Data, replace the missing values, remove cases of duplication, and deal with outliers.
**Step 4: Feature Engineering:** Include new features that may improve the model's performance. For instance, neighborhood crime rate and rating of the school's Categorical data may be encoded using one-hot encoding or any other technique.
**Step 5: Exploratory Data Analysis (EDA):** Visualize Data, plots will help in understanding the distribution of and the relationship between different features and target variables. Examine the correlations and trends that could potentially affect the price of a house.
**Step 6: Model Selection:** Choose Algorithms: From the literature, select Random Forest and XG-Boost as the foundation for this model based on their strengths. Random Forest is suitable for handling big datasets with a large number of features; reduces overfitting. XG-Boost has demonstrated high speed and performance, especially while dealing with non-linear relationships.
**Step 7: Model Training-**Split Data, divide the data into training and test sets. Perform hyperparameter tuning using Grid Search or Random Search to choose the most optimal parameters for both models. Perform, k-fold cross-validation to ensure the robustness of the model.
**Step 8: Model Evaluation:** Model evaluation should be done using metrics like RMSE, MAE, and $R^2$ score. Compare the performance of the Random Forest with XG-Boost and choose the best model.
**Step 9: Deployment**: Deploy via the cloud such as AWS, Azure, or on-premises servers. Create APIs that will enable users to make predictions in real time. Design an intuitive dashboard that the stakeholder will use to input data and predictions.
**Step 10: Monitoring and Maintenance:** Enable the tracking of model performance continuously over time to ensure that it works with accuracy. Schedule its regular update with new data that will keep it relevant. Incorporates user feedback for model improvement and usability.

### *4.4 Benefits for U.S. Businesses*

**High Accuracy:** The strong predictive performance of both algorithms leads to more accurate property prices.
**Handling Non-linearity**: XG-Boost models ascertain complex relationships in the data, which makes it suitable for many of the variables affecting real estate prices.
**Noise Resistance:** Random Forest is resistant to overfitting by averaging multiple decision trees, the model generalizes better.
**Speed and Efficiency:** XG-Boost provides fast computation and execution of the algorithm, which enables it to handle big data in less time and thus help in real-time feature applications.
**Flexibility:** Both algorithms can be used in very different areas other than regression, such as classification and ranking, which enables application in many different tasks in the sphere of real estate.
**Scalability:** These models scale with greater volumes of data, making their performance remain consistent with more data coming in.

### *4.5 Benefits to the U.S Economy*

**Elevated Market Analysis:** These models can effectively analyze the trend-setting in the housing market, which can help the organization identify whether or not there is a potential boom or bust that is being witnessed to affect overall economic growth.
**Informed Policy Decisions:** The insights drawn from model predictions may provide the government with input on housing policies in terms of zoning laws and tax incentives that spur economic activities.
**Risk Reduction:** Recognition of housing-market risk factors by these models would further close the avenues to certain avenues of economic decline associated with housing bubbles or crashes.
**Investment optimization:** Investors supported by efficient forecasts, can optimally adjust their portfolios for better capital allocation in the economy.
**Construction and Development:** With correct demand forecasts, construction and development projects will emerge, boosting job opportunities and economic development.

**Supporting financial institutions:** By hiring banks and lenders, the risk of mortgage can be judged based on these models, thereby infusing good, prudent lending that will be beneficial to economic health.

## 5. Conclusion

The California housing market presents a myriad of challenges attributed to its volatility and complexity, driven by various factors such as demographics, economic conditions, and government policies. This research project explored the designing of an Explainable AI machine-learning model for house price prediction in California. Relevant data about house Prices were collected from reliable sources such as California real estate websites, land sites, and public datasets. The data features included location, size, number of rooms, area type, availability, sale prices, and oceanic proximity. In this research project, proven and popular machine learning algorithms were applied, most notably, Linear regression analysis, XG-Boost, and Random Forest. The Random Forest yielded very impressive results, with a high accuracy score and a low root MAE and MSE; therefore, it was good at learning underlying best patterns and relationships that may exist within the data for house price predictions. XG-Boost also performed fairly slight errors and lower $R^2$ score compared to the Random Forest. Conversely, Linear Regression had the highest errors and the lowest $R^2$ score, demonstrating that it was less effective compared to the other algorithms.

**Funding:** This research received no external funding.
**Conflicts of Interest:** The authors declare no conflict of interest.
**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

## References

[1] Ahmad, M., Ali, M. A., Hasan, M. R., Mobo, F. D., & Rai, S. I. (2024). Geospatial Machine Learning and the Power of Python Programming: Libraries, Tools, Applications, and Plugins. In *Ethics, Machine Learning, and Python in Geospatial Analysis* (pp. 223-253). IGI Global.

[2] Chen, Y., Xue, R., & Zhang, Y. (2021, September). House price prediction based on machine learning and deep learning methods. In *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)* (pp. 699-702). IEEE.

[3] Chordia, P. (2022). Prediction of house price using machine learning. *www.academia.edu*. https://www.academia.edu/84479157/Prediction_of_House_Price_Using_Machine_Learning

[4] Gazi, M. S., Nasiruddin, M., Dutta, S., Sikder, R., Huda, C. B., & Islam, M. Z. (2024). Employee Attrition Prediction in the USA: A Machine Learning Approach for HR Analytics and Talent Retention Strategies. *Journal of Business and Management Studies*, *6*(3), 47-59.

[5] Hasan, R., Islam, Z., & Alam, M. (2024). Predictive Analytics and Machine Learning Applications in the USA for Sustainable Supply Chain Operations and Carbon Footprint Reduction. Journal of Electrical Systems, 20(10s), 463-471.

[6] IRET Journal, I. (2023). House price prediction using machine learning. *Irjet*. https://www.academia.edu/103759158/House_Price_Prediction_Using_Machine_Learning

[7] Jain, M., Rajput, H., Garg, N., & Chawla, P. (2020, July). Prediction of house pricing using machine learning with Python. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 570-574). IEEE.

[8] Kumar, G. K., Rani, D. M., Koppula, N., & Ashraf, S. (2021, June). Prediction of house price using machine learning algorithms. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1268-1271). IEEE.

[9] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: A comparative study. In *2019 International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-5). IEEE.

[10] Phudinawala, H. (2024). Predicting House Price with Deep Learning: A Comparative Study of Machine Learning Models. *www.academia.edu*. https://www.academia.edu/100621001/Predicting_House_Price_with_Deep_Learning_A_Comparative_Study_of_Machine_Learning_Models

[11] ProAIrokibul. (2024). *California-Housing-Trends-And-Analysis/Model/California Houses Trends And Analysis. ipynb at main · proAIrokibul/California-Housing-Trends-And-Analysis*. GitHub. https://github.com/proAIrokibul/California-Housing-Trends-And-Analysis/blob/main/Model/California%20Houses%20Trends%20And%20Analysis.ipynb

[12] Rana, V. S., Mondal, J., Sharma, A., & Kashyap, I. (2020, December). House price prediction using optimal regression techniques. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 203-208). IEEE.

[13] Sinha, A. (2020). Utilization of machine learning models in real estate house price prediction. *www.academia.edu*. https://www.academia.edu/43737997/Utilization_Of_Machine_Learning_Models_In_Real_Estate_House_Price_Prediction

[14] Vidya, S. (2024). Machine learning approach for house price prediction. *www.academia.edu*. https://www.academia.edu/122914693/Machine_Learning_Approach_for_House_Price_Prediction