Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

# Predicting the Rise in California Home Prices and Factors Affecting

Yufan Bai *

School of Data Science Capital University of Economics and Business, Beijing, 1000000, China

* Corresponding Author Email: yufanbai@ldy.edu.rs

**Abstract.** California's real estate market has been in the spotlight for its unique geography and economic advantages, and home prices have been volatile. Therefore, accurate prediction of house price increases is of great importance to home buyers, investors and policy makers. This paper utilizes the California house price dataset, combines macroeconomic indicators and micro property attributes, and analyzes house price prediction through random forest, decision tree and neural network models. The study results show that the Random Forest model performs the best in predicting house prices with an $R^2$value of 0.817210, which is significantly higher than the other models at 0.817210. the influence of $R^2$leads to a slightly poorer neural network performance (CNN). Further through visualization studies and feature importance analysis median income, longitude, latitude, and distance from the sea are the key factors affecting the house price in California. House prices are strongly positively correlated with median income, while distance from the sea is negatively correlated with house prices. These key factors can also directly affect future home buying trends and house price direction. The research in this paper provides a strong reference for homebuyers, investors, and policymakers to help them make more informed decisions in the real estate market and promote a smooth and healthy market development.

**Keywords:** *California home prices; home price prediction; influencing factors; random forest; machine learning.*

## 1. Introduction

The California real estate market has long been a topic of interest and speculation. The state's unique landscape, thriving economy, and multicultural offerings make it an attractive place to live, resulting in a competitive and vibrant real estate market.

As the overall U.S. economic environment changes, the trend of California home prices is also affected by many factors, including home prices. Data from the U.S. Department of Commerce indicates that despite the spring 2024 home buying season approach, the U.S. real estate market has yet to show significant signs of rebound, with new home sales falling 0.3% in February [1]. This trend reflects the dual suppression of demand and supply in the real estate market by the Federal Reserve's interest rate hiking cycle [1].

Predicting the increase in California house prices is of great practical significance and theoretical value. During the epidemic, the Federal Reserve's loose monetary policy pushed U.S. home prices up sharply; for example, the median U.S. home sales price reached $346, 900 in 2021, a year-on-year increase of 16.9% [2]. However, since 2022, when the Federal Reserve tightened monetary policy, the home price appreciation rate has slowed [2]. In this context, accurately predicting the rate of increase in California home prices can provide a decision-making basis for home buyers and investors and a reference for policy makers to achieve a stable and healthy development of the real estate market [3].

Currently, most studies on house price forecasting focus on the analysis of macroeconomic factors and market supply and demand relationships. Li Neng [3] pointed out that the healthy development of the real estate market is of great significance to the economic and financial stability, and emphasized the importance of the policy to guide the expectations of the real estate market. In addition, Jiaqi Wang [4] deeply analyzed the multiple factors affecting the price of houses in the United States through the method of machine learning, and found that micro factors such as house area, decoration situation and supporting facilities have a significant impact on the price of houses.

Current research often ignores the combination of macro and micro factors when forecasting house prices. This study will fill this gap and construct a more comprehensive and accurate house price prediction model by combining macroeconomic indicators and micro property attributes. An in-depth investigation of the factors that influence of housing prices is of great significance to formulating effective housing regulation policies and promoting the healthy and stable development of the real estate market [5].

Understanding and predicting California home prices is critical for homeowners, potential buyers, real estate investors, and policy makers. In this paper, we compare and analyze the performance of models such as random forests, decision trees, and neural networks for home price prediction using the California home price dataset.

This paper hopes to reveal the key factors affecting California home prices, and through in-depth analysis of housing characteristics and market data, build a forecasting model to anticipate future trends in California home prices accurately. This study will refine existing house price forecasting methods, explore the micro and macro influences on house price increases, and propose corresponding policy recommendations to provide valuable references for home buyers, investors, and policy makers, and help stakeholders make informed decisions when buying, selling, and investing in real estate.

## 2. Data and Methods

### 2.1. Data source

The data for this paper is characterization data on California housing prices and the factors corresponding to each house, which is a dataset from the kaggle website (https://www.kaggle.com/datasets/hosammhmdali/house-price-dataset/data). The California housing dataset contains information about various socioeconomic characteristics of neighborhood groups in California. The purpose of the dataset is to provide detailed information about the housing market in different regions for further analysis and research. The fields in the data cover a wide range of geographic locations, demographics, housing characteristics, and economic conditions, which makes it valuable for a wide range of applications in areas such as real estate market research, urban planning, and economic analysis. Moreover, the dataset has detailed and specific continuous and discrete data, which provides sufficient data support for subsequent modeling and analysis, and also makes the data results more diversified. Each row in the dataset represents a block group with 20,640 observations and each observation has 10 attributes.

**Table 1.** Attributes of the dataset

| Listings | Implied meaning | Data Characteristics |
|---|---|---|
| longitude | Longitude | Floating Point |
| latitude | Latitude | Floating Point Numbers |
| housing_median_age | Median age of houses | Floating Point |
| total_rooms | Total number of rooms | Floating Point |
| total_bedrooms | Total number of bedrooms | Floating point number (with missing values) |
| population | Population | Float |
| households | Total number of households | Float |
| median_income | Median Income | Float |
| median_house_value | Median Home Value | Float |
| ocean_proximity | Distance to Sea | Category |

Table 1 provides a generalized summary and an explanation of each column of data for this dataset, and the last column of Table 1 also describes the characteristic features of each column of data.

## 2.2. Data preprocessing

This study first needs to deal with missing values in the data. Dealing with missing values is very important in data analysis and machine learning because missing values can have multiple negative impacts on the analysis results and model performance, and correctly dealing with missing values can improve the accuracy of data analysis and the predictive ability of the model to ensure the reliability and interpretability of the results.

In this study, Simple Imputer class is used, for other features with small number of missing values, the data are generally supplemented by the way of missing value completion, the way of data completion is selected based on the specific type of missing values, commonly used to take the mean, median, or take the mode and so on [6]. First, the categorical variable ocean_proximity is deleted and only the numerical variables are retained. Then, Simple Imputer is applied to calculate and fill in the missing values in these numerical variables.

For the categorical variable ocean_proximity, this paper uses OneHotEncoder for solo heat coding. OneHotEncoder is the technical process of converting a categorical variable into a numerical variable, and this treatment will make our subsequent analytical study easier. The categorical variable is first extracted and then converted into the form of solo hot encoding using OneHotEncoder.

After filling the missing values, the numerical variables are merged with the coded categorical variables. Since OneHotEncoder outputs a sparse matrix, it is first converted to array form. Then the numeric and coded categorical variables are horizontally spliced together using np. hstack.

Convert the merged data into a new DataFrame. column names include the original numeric variable column names and the uniquely coded categorical variable column names. The final output is the first few rows of the processed data frame.

The dataset was successfully preprocessed through these steps, missing values were dealt with, and categorical variables were converted into a numerical form that could be used in machine learning models.

## 2.3. Methods and models

For this dataset, this study attempted to select three models for model building and training: random forest, decision tree and neural network. These three models can be compared in order to select a model that is most suitable for this dataset for subsequent prediction and analysis.

Random Forest is an integrated learning method that improves the accuracy and robustness of a model by constructing multiple decision trees and combining their predictions. It reduces the risk of overfitting and is more stable, not noticeable for small data changes. It can also handle large amounts of features and missing data, but at the same time it is slower to train and has a large memory footprint.

A decision tree is a tree-structured supervised learning model for classification and regression tasks. It consists of nodes and edges, where each node represents a test of a feature, each edge represents a branch of the test result, and leaf nodes represent the final output.

The decision tree model is characterized by its simplicity, ease of visualization, and the ability to handle both numerical and categorical data. However, compared to random forest, it is sensitive to small changes in data, and sensitive to small changes in data, and it is prone to overfitting.

A neural network is a computational model inspired by the biological nervous system and is particularly suitable for complex pattern recognition and prediction tasks. It consists of neurons (nodes) layers, including input, hidden and output layers.

The special feature of neural network is that it can train high-dimensional complex data and fit complex nonlinear relationships. However, it also suffers from parameter selection and tuning difficulties, and is prone to overfitting, which needs to be mitigated using methods such as Regularization and Dropout.

### 2.4. Assessment of indicators

In order to evaluate the model performance, this paper chooses the commonly used Mean Squared Error (MSE) and R squared score (R^2) metrics to analyze it. These metrics can analyze the accuracy of the algorithmic model.

MSE measures the average squared error between a model's predicted and actual values. The larger the error, the larger the MSE, the worse the model's predictive ability; the smaller the error, the smaller the MSE, the better the model's predictive ability. The formula for MSE is shown below:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{1}$$

R^2 is used to assess the goodness of fit of the regression model. It represents the proportion of the total variation in the dependent variable explained by the independent variable and ranges from 0 to 1. An indicator of the degree of fit of the trend line, the magnitude of the value reflects the degree of fit between the estimated value of the trend line and the corresponding actual data [7]. The closer the value of $R^2$ is to 1, the better the model is at explaining the data. Higher R^2 values indicate a better fit of the model to the data and higher prediction accuracy. Conversely, lower R^2 values indicate that the model fits the data poorly and that more features or a more complex model may be needed. Select a model with a higher R^2 to improve prediction accuracy. The formula for R^2 is shown below:

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{2}$$

## 3. Analysis of results

### 3.1. Visualization

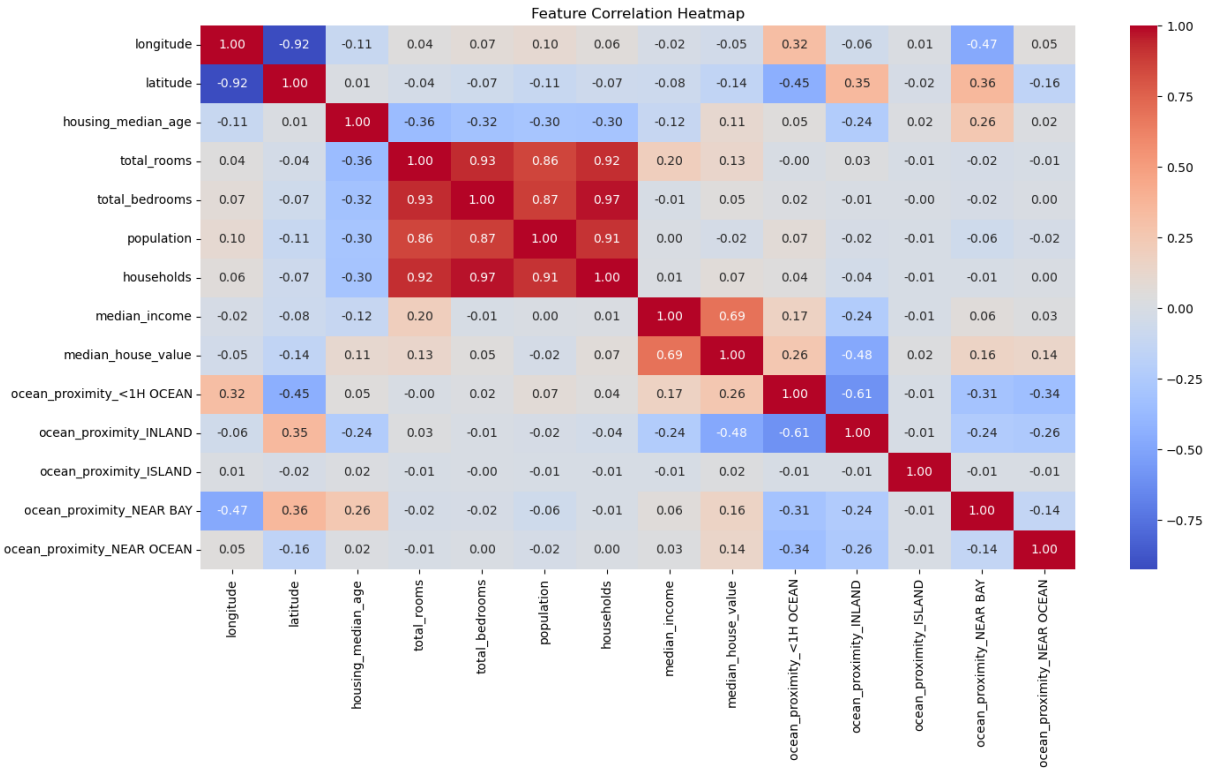


**Figure 1.** Feature correlation heatmap

Figure 1 is heating map of feature correlation. This figure shows the correlation between the variables. Correlation analysis belongs to the exploratory analysis of the data analysis process and refers to

uncertain dependencies between variables [8]. In Figure 1, blue color represents negative correlation and red color represents positive correlation, they take values between 1 and -1, the darker the color, the stronger the correlation. From Figure 1, it can be observed that median_house_value and median_income have a strong positive correlation (0.69), indicating that the higher people's income is, the higher the house price they are likely to live in. median_house_value and ocean_proximity_INLAND have a strong negative correlation (-). 0.48), indicating that the closer inland you are, the lower the house price.

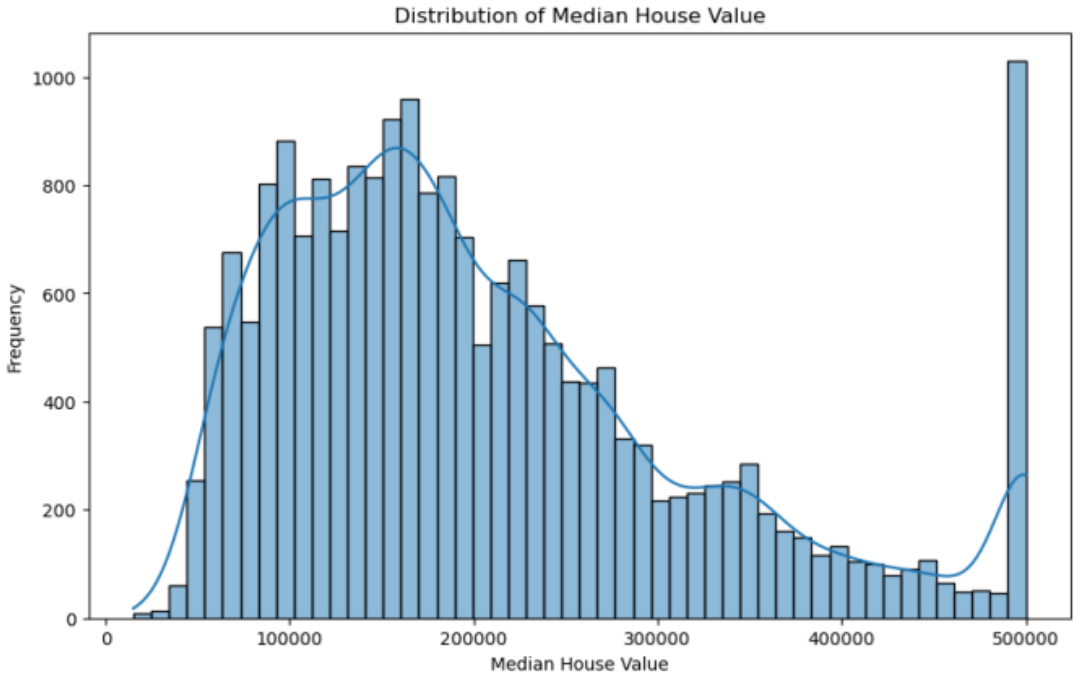


**Figure 2.** Distribution of median house value

Figure 2 shows the distribution of median house prices. The graph shows the distribution of individual house prices. The above graph shows a right-skewed distribution, implying that the median price of most homes is concentrated in the low-to-moderate price range, with a smaller number of homes in the higher price ranges. There is a noticeable peak near 500,000, indicating that a significant number of homes have median prices at or near this value. Of course, this could also be because the upper limit for home prices in the dataset is set at 500,000. The majority of home prices are clustered between 50,000 and 300,000, showing the main distribution of home prices. As can be seen from the graph, the largest number of homes are found between 100,000 and 200,000, creating a peak.

The height of the frequencies shows that the number of houses at different price points is unevenly distributed. The number of houses at the lowest and highest price levels is low, while the number of houses at the middle price level is high.

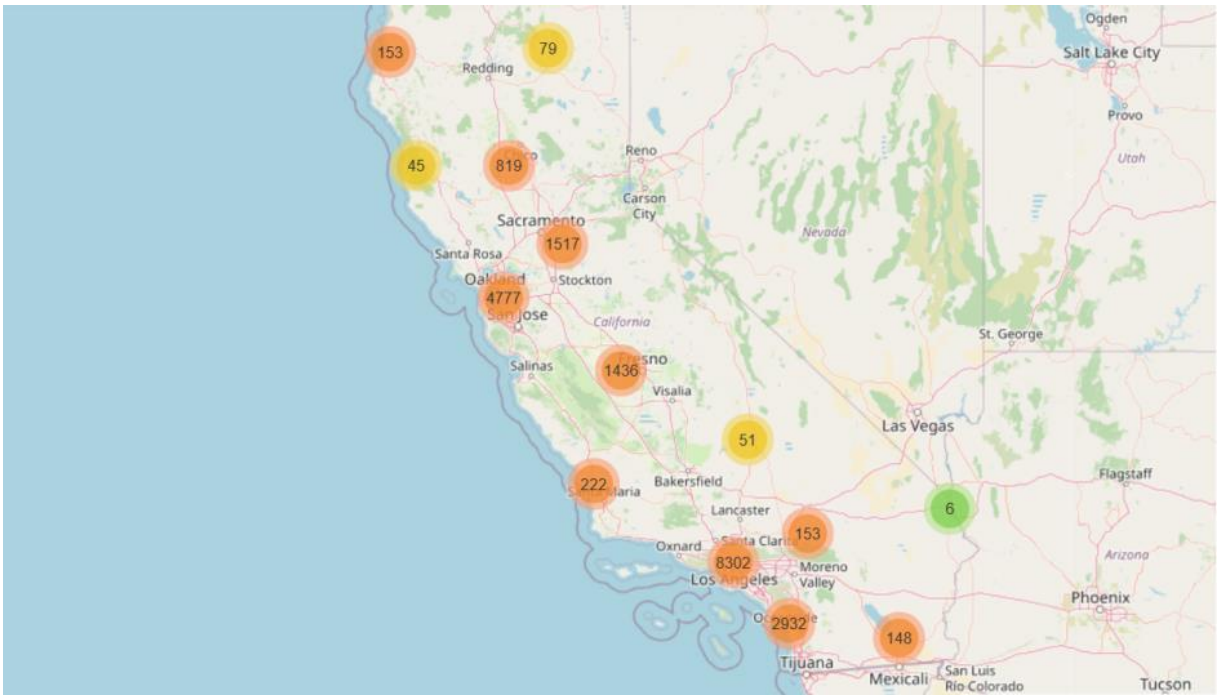


**Figure 3.** Map

Figure 3 is a map drawn from the latitude and longitude based on the dataset for this study and labeled with data on the distribution of the number of houses in each region of the United States. Figure 3 is a dynamic map that can be zoomed in and out, allowing the researcher to conduct a more specific, more in-depth regionalized study. As can be seen from the map, over 15,000 houses in the United States are located along the coastline, with sparser distribution of houses inland. Drawing the above map because we can see the distribution of houses in California, Figure 1 also shows the distribution of houses and housing prices of the relationship between the higher coefficient, we can use this map to more clearly see the distribution of houses in the coastal areas of the denser, from the side can also be analyzed in the coastal areas of the housing prices may be higher.

### 3.2. Performance analysis of the algorithm

**Table 2.** Model performance indices

| model | Mean Squared Error | R^2 score |
|---|---|---|
| Random Forest | 2.395290e+09 | 0.817210 |
| Decision Tree | 4.785287e+09 | 0.634825 |
| Neural Network | 4.618103e+09 | 0.647583 |

This study selects three models for comparison, and an optimal one is chosen for subsequent analysis. Table 2 shows the values of MSE and R^2 for the training output of these three models for the dataset. From Table 2, it can be seen that Random Forest is more advantageous and its R^2 score is significantly higher than the other two models, indicating that Random Forest has the best fit to the data and the best prediction accuracy. Although the MSE is not as high as the other two models, the MSE for all three models is very small, and they are all in a better state regarding model prediction, so this paper considers that their differences are not very large overall. In summary, choosing Random Forest is the optimal choice.

### 3.3. Characteristic Importance Analysis

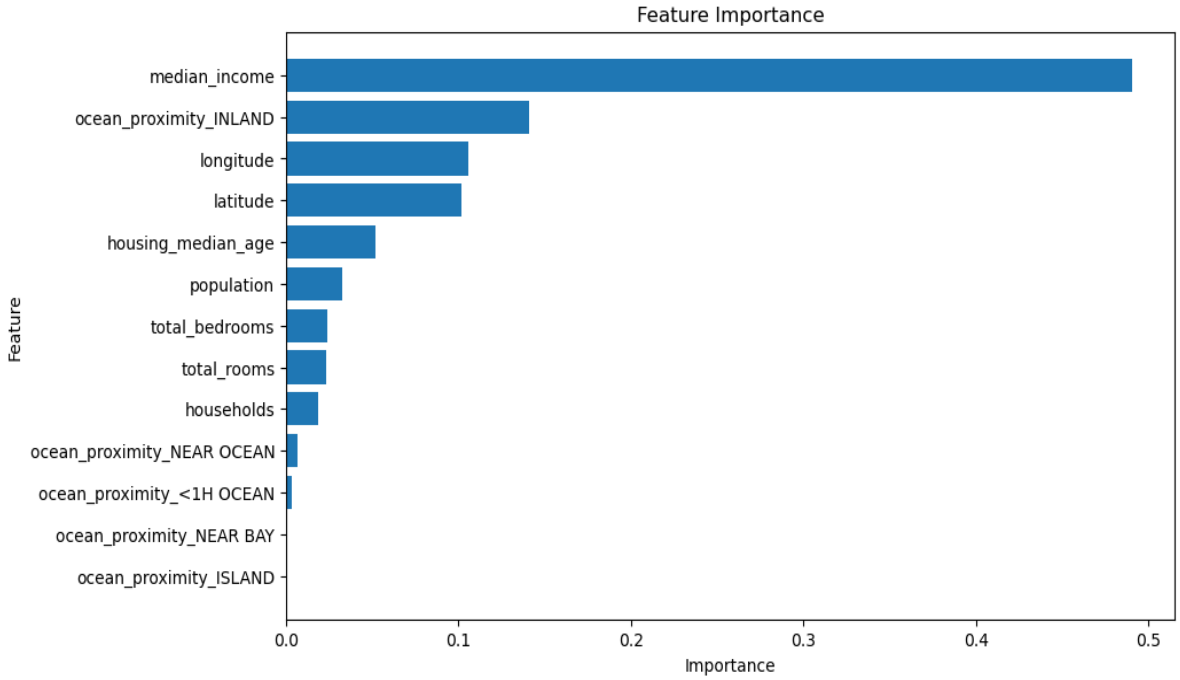


**Figure 4.** Feature importance analysis

In this study, the Random Forest model is used for feature importance analysis to assess and quantify the degree of influence of each feature (variable) in the dataset on the predicted target variable (house price), so that the explanatory variables that have a significant impact on the explanatory variables can be effectively selected [9]. As shown in Figure 4, there is no single influence on real estate prices [10]. Median_income is the most prominent and obvious characteristic variable for U.S. house prices, with a value more than three times ahead of the other characteristic variables, with a value as high as 0. 490642.This is followed by ocean_proximity_INLAND (0.140925), LONGITUDE (0.105889), and LATITUDE (0.101597).

Next, this paper focuses on analyzing only the top four characteristic variables in importance.

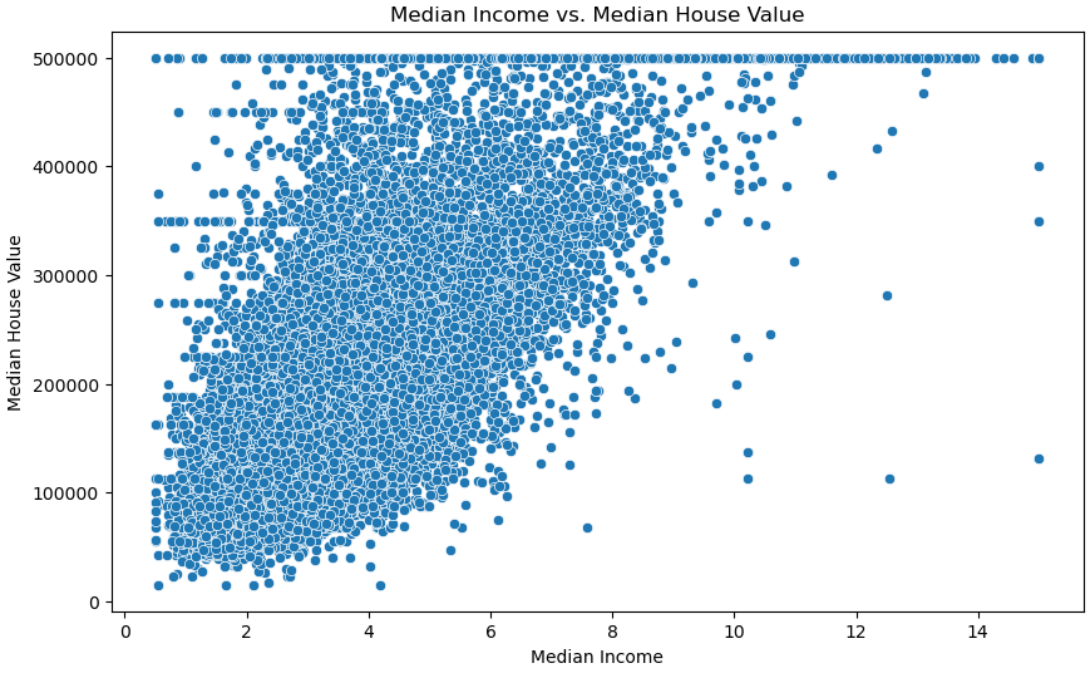


**Figure 5.** Scatterplot between house prices and median income

Figure 5 presents a scatterplot of the relationship between house prices and median resident income in California. As can be seen in Figure 5, residential income shows a positive correlation, mainly with house prices. This means that the higher people's income is, the higher their standard of living is, the better house they live in, and the higher their house price is.

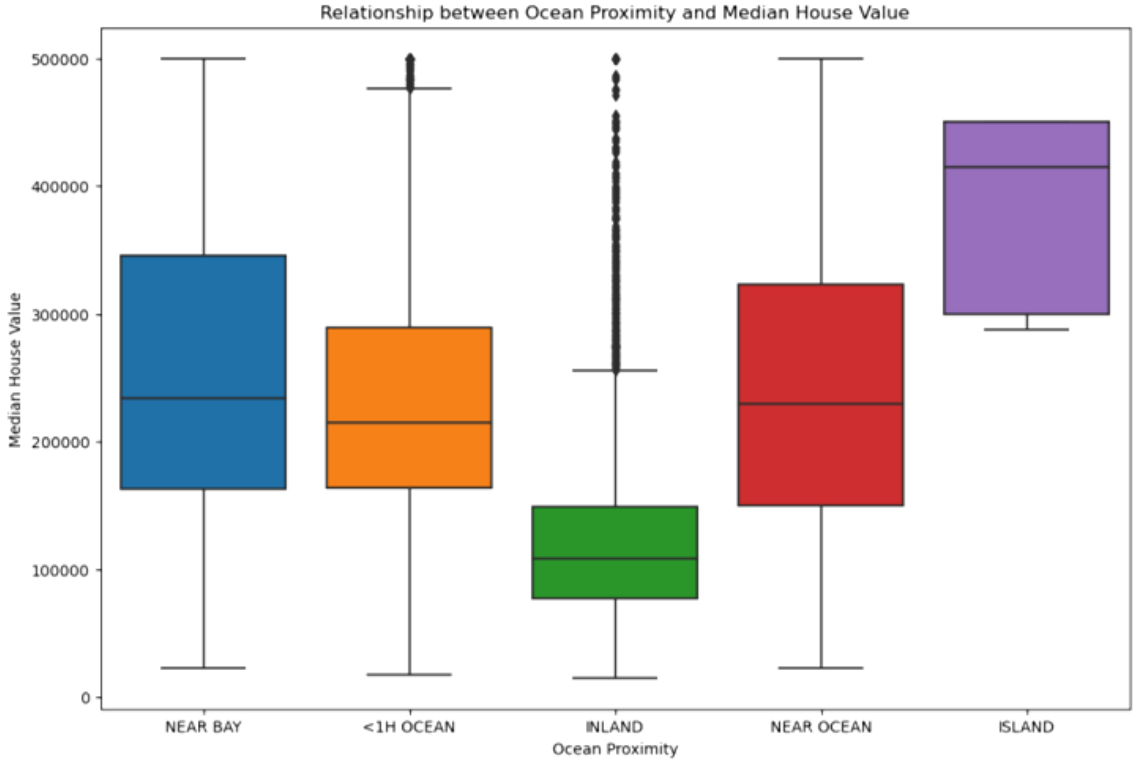


**Figure 6.** Box-line diagram of position distribution

For the indicator of whether or not it is by the sea, Figure 6 divides it into five categories so that it is clear what characteristics each type of house has. As can be seen in Figure 6, house prices are generally low for houses located in the interior of the U.S. House prices are highest on the U.S. islands, possibly because the islands, as tourist destinations, are small and have beautiful views, so house prices there are above average. The price of houses located near the coastline is higher than inland house prices, the intrinsic reason for this could be because of the price advantage of ocean view houses in the overall real estate market.

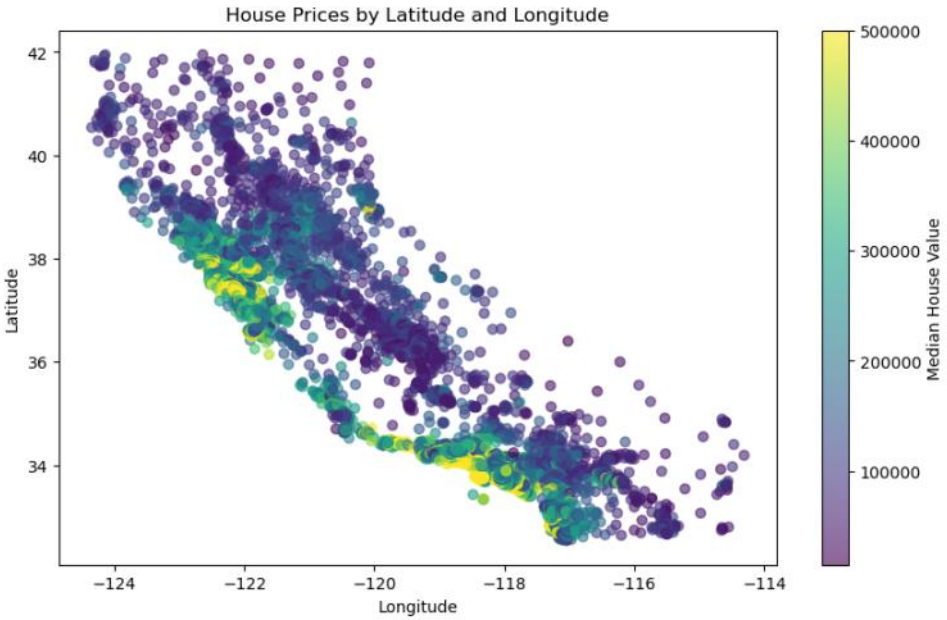


**Figure 7.** Scatterplot

Figure 7 illustrates a scatter plot between California house prices and latitude and longitude. As can be seen in Figure 7, each point represents a geographic location, and the longitude and latitude of the point indicate its geographic coordinates, respectively. The point's color indicates the median house price, and the color bar ranges from purple to yellow to indicate a change in house price from low to high. Lighter colors (closer to yellow) indicate higher house prices and darker colors (closer to purple) indicate lower house prices.

In Figure 7, areas of higher house prices are concentrated in specific latitude and longitude ranges, especially in the more western (longitudes between -122 and -118) and more southern (latitudes between 34 and 37) locations of the figure. These locations are likely to be cities or regions with high house prices. Figure 7 also shows that house prices are concentrated in a lower range in certain areas, with these points mainly in the more northern and eastern locations. Geographic location significantly impacts house prices, with generally higher house prices close to a particular geographic area (which may be a coastal city or metropolitan area). A clear geographic gradient can also be seen in Figure 7, with house prices gradually decreasing from the southwest to the northeast. This may be related to local economic development, natural environment, infrastructure and other factors.

## 4. Conclusion

This paper examines the major influences on house prices by analyzing California house price data and develops several prediction models to assess their accuracy. The results of the study show that the Random Forest model performs best in house price prediction, with significantly higher $R^2$ values than the Decision Tree and Neural Network models. Generally speaking, it should be the neural network model that is better, but because the dataset in this study is not high-dimensional and has a large number of linear relationships, this leads to its poor performance. In particular, median income shows a strong positive correlation with house prices, indicating that the higher the income of residents, the higher the house prices. This finding is in line with expectations, as income levels are usually directly related to housing demand and affordability. On the other hand, the effects of longitude and latitude reveal the importance of geographic location on house prices, with proximity to major cities and coastal areas significantly higher than inland areas. Distance from the sea is negatively correlated with house prices, suggesting that houses close to the coastline are usually more expensive due to the advantages of the landscape and living environment.

Through visualization and analysis, this paper further validates the role of these influencing factors. For example, the scatterplot of house prices versus median income shows a clear positive correlation trend, while the map distribution of house prices versus latitude and longitude reveals that high-price areas are mainly concentrated in coastal cities. In addition, a box-and-line plot of house prices versus distance from the sea supports this finding.

The innovation of this study is the integrated consideration of macroeconomic indicators and micro property attributes, and the use of multiple machine learning models for comparative analysis. Compared with previous studies, this paper not only focuses on traditional economic and market factors, but also introduces specific property characteristics such as geographic location, which improves the accuracy and reliability of the house price prediction model.

Future research can further extend the scope and dimension of the dataset to include more time series data to analyze house price trends dynamically. In addition, more advanced deep learning models such as CNNs and LSTMs can be incorporated to capture more complex non-linear relationships and spatio-temporal features. The study can also further refine the regional analysis to explore the drivers of house prices in different cities and neighborhoods to provide more targeted policy recommendations and market strategies for local governments and real estate companies.

## References

[1] M. Molly. "U.S. real estate market 'spring' has not arrived," Financial Times, March 27, 2024, p. 008.

[2] M. Molly. “Where the US housing market is headed in 2022,” Financial Times, January 26, 2022, p. 008.

[3] L. Neng. “Building a new model for stable and healthy development of real estate market,” China Development Observatory, vol. 2023, no. Z2, pp. 84 – 90.

[4] W. Jiaqi. “Analysis of influencing factors of U.S. housing prices based on machine learning method,” M.S. thesis, Yunnan University, 2020.

[5] T. Runze. “Boston house price prediction based on multiple machine learning algorithms,” China New Communication, vol. 21, no. 11, pp. 228 – 230, 2019.

[6] U. Xiaomin, J. Zhou, D. Zhou. “Empirical analysis of factors influencing housing prices in China,” Real Estate World, vol. 2024, no. 03, pp. 5 – 8.

[7] J. Shaotaki, M. Wang, Z. Wang, et al. “Econometric analysis of factors influencing housing prices - based on Stata panel data,” Investment and Entrepreneurship, vol. 34, no. 12, pp. 49 – 51, 2023.

[8] M. Huan. “Research on the impact of real estate bubble on financial risk in the context of new crown epidemic,” M.S. thesis, Shandong University of Finance and Economics, 2023. DOI: 10.27274/d.cnki.gsdjc.2023.001396.

[9] X. Ouyang, Z. Lv. “Linear regression model analysis of factors influencing real estate prices in Nanning,” Residential and Real Estate, vol. 2023, no. 27, pp. 110 – 112.

[10] K. Meng. “Analysis of Influencing Factors on Housing Prices and Policy Recommendations—Taking Beijing as an Example,” Journal of Economic Research, vol. 2023, no. 22, pp. 32 – 34.