HOUSE PRICE PREDICTION

ALT SCHOOL AFRICA

Capstone Project

2024

MOMOH VICTOR SOKOLAYAM

momohvictor2018@gmail.com

Table of Content

Table of Content

1.0 Introduction

1.1. Project Overview

The real estate market is a dynamic and multifaceted domain where home prices fluctuate due to a multitude of factors. Accurate prediction of housing prices is crucial for various stakeholders, including buyers, sellers, real estate agents, and investors. An effective prediction model can significantly enhance decision-making processes by providing reliable price estimates based on a thorough analysis of relevant features.

This project focuses on developing a robust House Price Prediction Model using the California Housing Pricing dataset. The model aims to offer dependable estimates of median house values by leveraging a comprehensive set of features, including geographical, demographic, and housing characteristics. By integrating advanced machine learning techniques, this model seeks to assist real estate professionals and investors in accurately valuing properties, thereby improving market efficiency and investment strategies.

1.2 Problem Statement

Accurately forecasting house prices presents a major challenge in the real estate market, influenced by a range of factors such as property types, zoning regulations, lot attributes, property conditions, and sale circumstances. Conventional valuation approaches often rely on limited datasets and subjective assessments, which can result in inconsistencies and inaccuracies. This uncertainty affects various stakeholders, including home buyers, sellers, real estate agents, and investors, potentially leading to financial losses and inefficiencies for real estate companies.

1.3 Aim

To address the complexities of predicting housing prices, our goal is to develop a robust House Price Prediction Model utilizing advanced machine learning techniques. This model will analyze a diverse set of property features and market data to achieve a high level of accuracy. Specifically, we aim to attain an overall prediction accuracy of 85%, with a maximum deviation of $25,000

between actual and predicted prices. By providing precise, data-driven price estimates, our model is designed to equip stakeholders—including buyers, sellers, real estate agents, and investors—with reliable information for making informed decisions in the real estate market.

1.4 Objective

The objective of this project is to systematically develop and refine a House Price Prediction Model through the following processes:

- Data Preprocessing: Clean and prepare the California Housing Pricing dataset by handling missing values, performing feature engineering, and normalizing data to ensure consistency and reliability.
- Feature Engineering: Apply techniques such as logarithmic transformation, creation of new features, and one-hot encoding of categorical variables to enhance the model's ability to interpret and predict housing prices.
- Model Selection and Training: Evaluate and train various machine learning models, including Linear Regressor, Random Forest Regressor and XGBoost Regressor models to identify the most effective approach for price prediction.
- Model Tuning and Evaluation: Utilize hyper parameter tuning and cross-validation to optimize model performance, ensuring that the model achieves high accuracy and robustness across different data subsets.
- Performance Assessment: Assess model performance using metrics such as Mean Squared Error (MSE) and R-squared, and compare results to select the best-performing model.
- Deployment Preparation: Prepare the final model for deployment by saving it and providing example usage to demonstrate its practical application for real estate stakeholders.

Through these processes, the project aims to deliver a precise and reliable House Price Prediction Model that supports informed decision-making in the real estate market.

1.5 Data Sources

The California Housing Pricing dataset is a well-known dataset used for regression tasks, particularly for predicting housing prices. It contains various features related to housing and geographical information, such as median income, housing age, and proximity to the ocean. The dataset is sourced from the 1990 U.S. Census, specifically the California regions, and is commonly used for educational and benchmarking purposes in data science and machine learning.

- Kaggle

1.6 Summary

Linear Regression, Random Forest Regressor, and XGBoost models were developed for the dataset. Among them, XGBoost outperformed the others, achieving our XGBoost model shows a Mean Squared Error (MSE) of about $23,000 and an R-squared value of 0.802 which ended up being the lowest MSE compared to other models.

2.0 Libraries and Dataset

2.1 Libraries

The following libraries were utilized in this project for Exploratory Data Analysis (EDA) and Model Development:

pandas: For data manipulation and handling.

numpy: For numerical computation and array operations.

matplotlib: For creating 2D visualizations and plotting.

seaborn: For enhanced 2D data visualizations and statistical plots.

scipy: For statistical analysis and computations.

StandardScaler: For standardizing features before model training.

train_test_split: For splitting the dataset into training and testing subsets.

LinearRegression: For developing a baseline linear regression model.

RandomForestRegressor: For implementing an ensemble regression model.

XGBoost: For developing an ensemble machine learning model.

mean_squared_error, mean_absolute_error, r2_score: For evaluating model performance using key metrics.

GridSearchCV: For selecting optimal hyper parameters through grid search.

2.2. Dataset Description

The California Housing Pricing dataset contains data on housing prices and various features that affect them. The dataset consists of:

Number of rows: 20,640

Number of columns: 9

2.3. Features

longitude: Longitude coordinate of the housing unit.

latitude: Latitude coordinate of the housing unit.

housing_median_age: Median age of the housing units.

total_rooms: Total number of rooms in the housing unit.

total_bedrooms: Total number of bedrooms in the housing unit.

population: Population of the area.

households: Number of households in the area.

median_income: Median income of the area.

ocean_proximity: Categorical feature representing proximity to the ocean

3.0 Methodology

3.1 Data Processing

- Handling Missing Values:

Missing values in the total_bedrooms column were handled by imputing with the median value.

Fig. 1.0 Identification of missing values

3.2 Exploratory Data Analysis

Exploratory data analysis was performed to uncover certain insights about the dataset.

Univariant and Bivariant analysis were performed.

3.2.1 Univariate Analysis:

A univariate analysis was carried out on all the columns

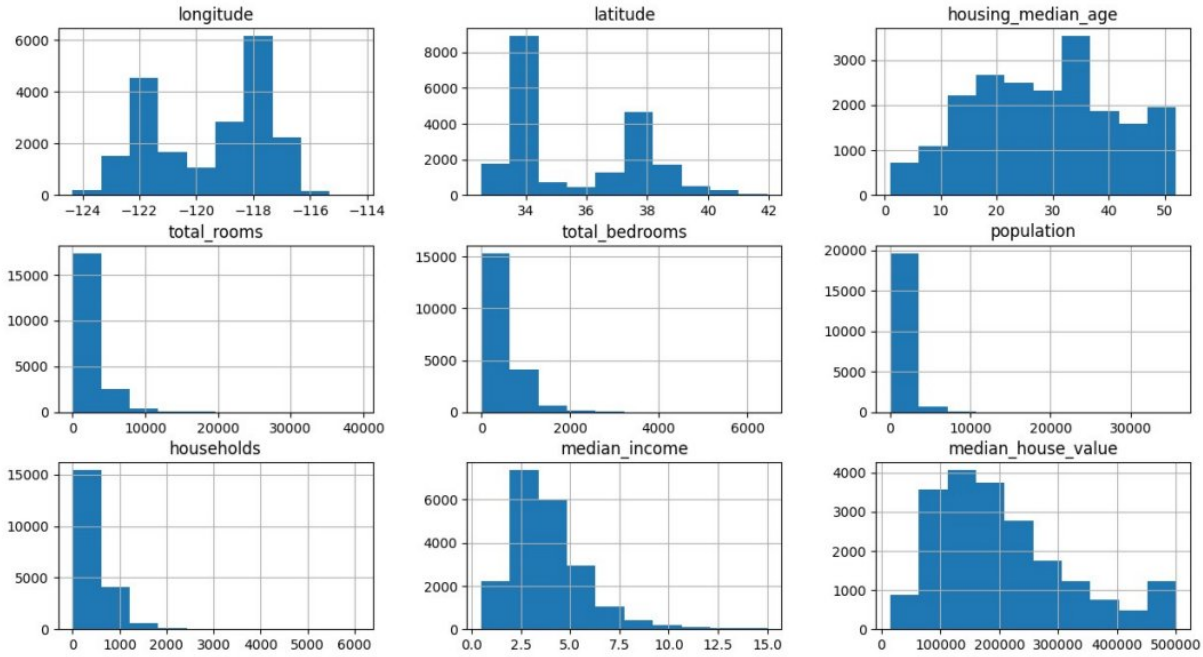

Fig. 1.1 Distribution of Features

- Features distribution visualization

Longitude: The histogram shows the distribution of houses across different longitudes in California. The data is clustered between -124 and -114, with a higher concentration around -122 and -118.

Latitude: Similar to the longitude, this histogram represents the distribution of houses across latitudes. Most houses are concentrated between 34 and 38 degrees latitude.

Housing Median Age: This histogram shows the age of houses. There's a significant peak around 30 to 35 years, indicating that many houses are relatively older.

Total Rooms: This histogram displays the total number of rooms in houses. The distribution is right-skewed, meaning most houses have a moderate number of rooms, with fewer houses having a very high number of rooms.

Total Bedrooms: The total number of bedrooms also follows a right-skewed distribution, similar to the total rooms.

Population: This histogram illustrates the population in the areas where the houses are located. The distribution shows that most areas have a population between 500 and 2500, with fewer areas having extreme populations.

Households: This represents the number of households in different areas. The distribution is similar to that of the population, indicating that most areas have a moderate number of households.

Median Income: The distribution of median income is slightly right-skewed, with most households earning between $2,000 and $6,000.

Median House Value: The histogram of median house values shows a right-skewed distribution, where most houses fall between $100,000 and $300,000, with a fewer number of houses at the higher end.

These histograms help understand the spread and central tendency of different features in the dataset, which is crucial for any data analysis or machine learning model building.

- Numerical columns

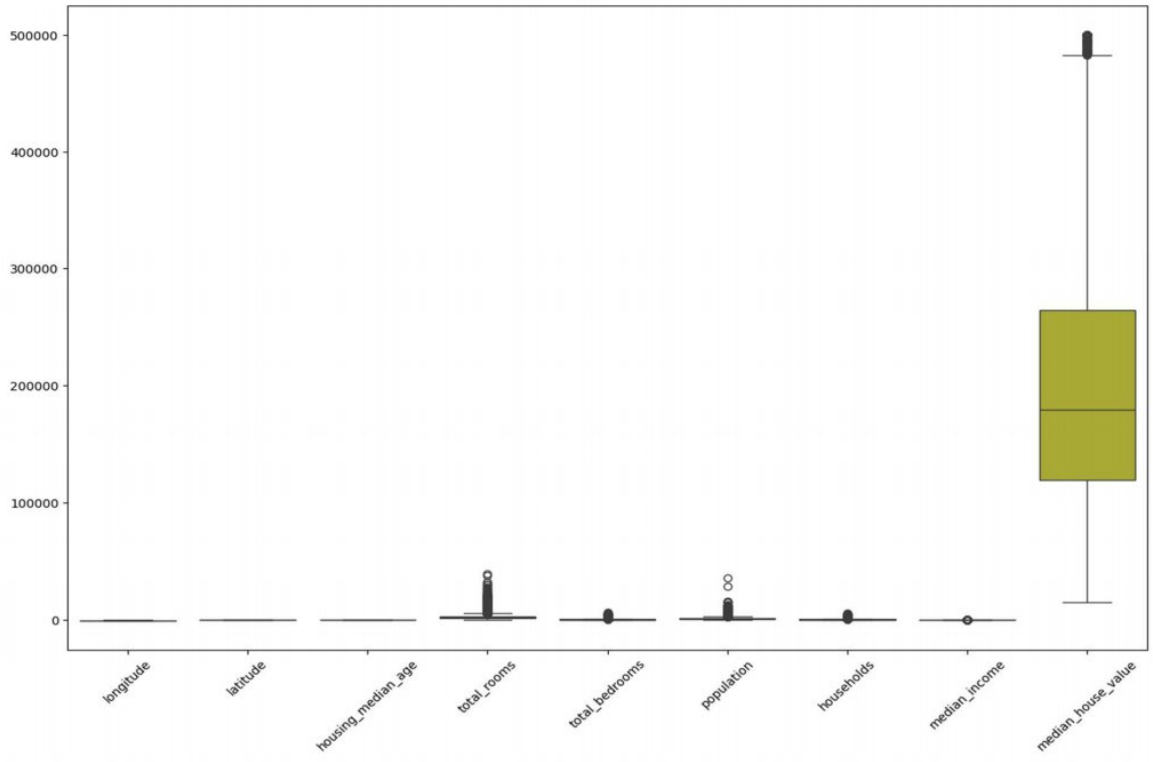

Fig. 1.3 Box Plot of Numerical Columns

The median_house_value feature has a significantly larger range compared to other features, which causes the boxplot for this feature to dominate the chart. This disparity suggests that the median_house_value needs to be visualized separately or scaled down for clearer comparison with other features.

Several features, including total_rooms, total_bedrooms, population, and households, exhibit outliers. This is indicated by the dots outside the whiskers. These outliers may impact the performance of predictive models and may require further investigation or preprocessing, such as capping, transformation, or removal.

Some features like longitude, latitude, and housing_median_age show minimal spread, indicating lower variability within the data. These features may have less impact on predicting the target variable (median_house_value) compared to others with more variability.

- Target Variable

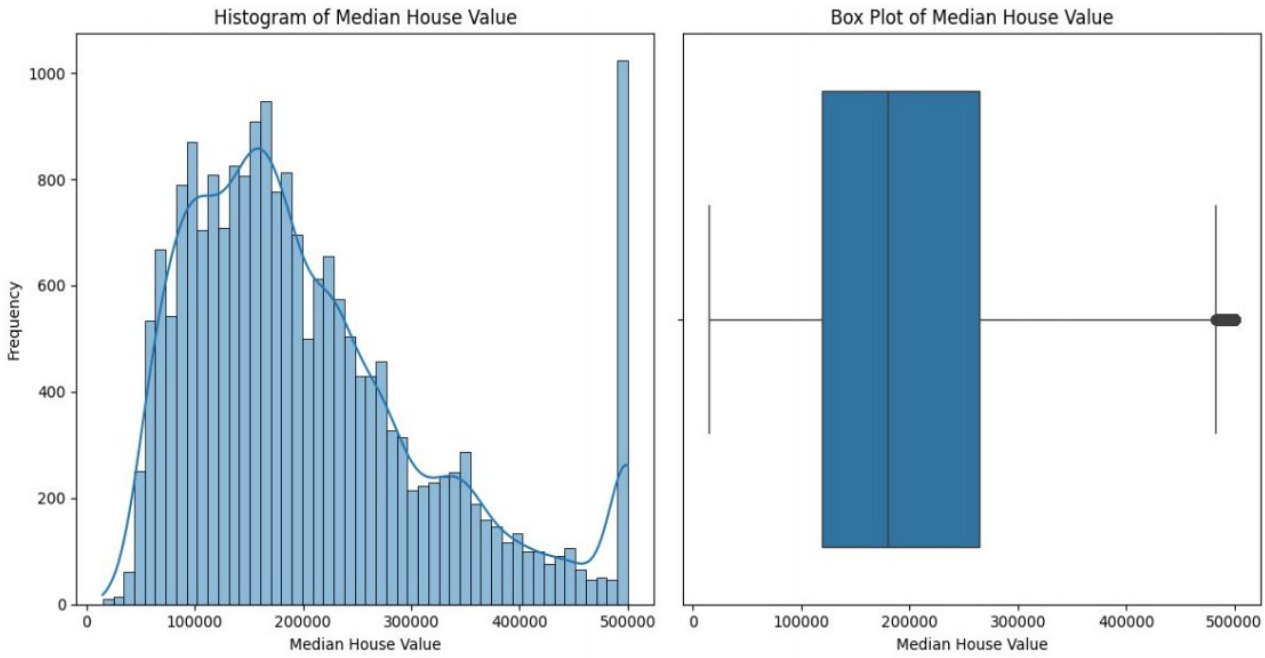

Fig. 1.4 Distribution of Median House Price Showing Skewness to the right and Box Plot showing outliers.

The distribution of the housing values is positively skewed with a peak around $150,000 - $200,000. There is a noticeable spike at $500,000, likely due to a cap in the data, indicating that many homes are valued at or above this limit.

The interquartile range (IQR) is wide, and there is a visible upper whisker extending to the maximum value of $500,000, confirming the presence of high-value outliers.

3.2.2 Bivariate Analysis

-Correlation Analysis

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
longitude	1.000000	-0.924616	-0.109357	0.045480	0.069608	0.100270	0.056513	-0.015550	-0.045398
latitude	-0.924616	1.000000	0.011899	-0.036667	-0.066983	-0.108997	-0.071774	-0.079626	-0.144638
housing_median_age	-0.109357	0.011899	1.000000	-0.360628	-0.320451	-0.295787	-0.302768	-0.118278	0.106432
total_rooms	0.045480	-0.036667	-0.360628	1.000000	0.930380	0.857281	0.918992	0.197882	0.133294
total_bedrooms	0.069608	-0.066983	-0.320451	0.930380	1.000000	0.877747	0.979728	-0.007723	0.049686
population	0.100270	-0.108997	-0.295787	0.857281	0.877747	1.000000	0.907186	0.005087	-0.025300
households	0.056513	-0.071774	-0.302768	0.918992	0.979728	0.907186	1.000000	0.013434	0.064894
median_income	-0.015550	-0.079626	-0.118278	0.197882	-0.007723	0.005087	0.013434	1.000000	0.688355
median_house_value	-0.045398	-0.144638	0.106432	0.133294	0.049686	-0.025300	0.064894	0.688355	1.000000

Fig. 1.5 showing the correlation between several features

There is a very high correlation between total_rooms and total_bedrooms (0.930). This suggests multi-collinearity, which might lead to redundancy in the model.

The housing_median_age has a weak positive correlation with median_house_value (0.106). Older houses may have slightly higher values, but this correlation is not very strong.

There is a very high correlation between total_rooms and total_bedrooms (0.930). This suggests multi-collinearity, which might lead to redundancy in the model.

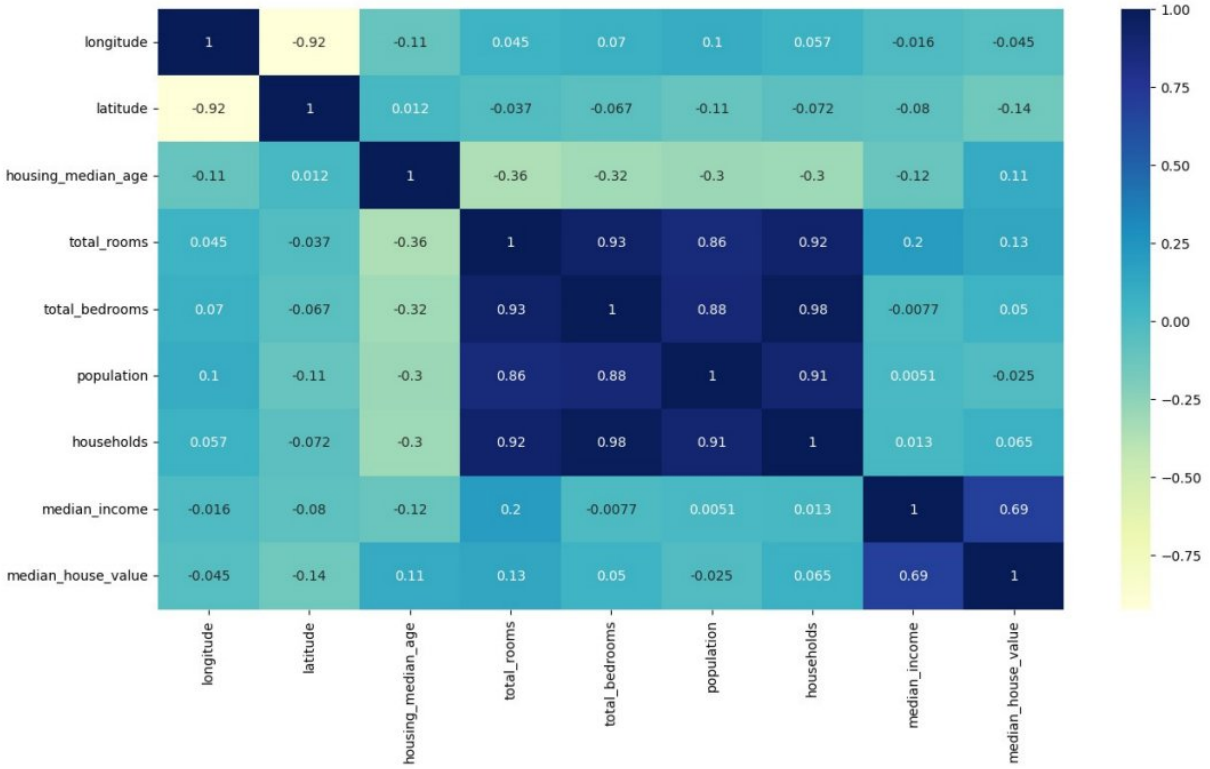

Fig. 1.6 Heatmap showing the correlation amongst features

The median_income has a strong positive correlation with median_house_value (0.688), indicating that as income increases, house values tend to increase as well. This makes median_income a crucial predictor in modeling house prices.

Longitude and latitude are negatively correlated with each other (-0.925), which is expected since they represent geographic coordinates.

Both have weak correlations with median_house_value, suggesting that location alone (in terms of coordinates) may not be a strong predictor of house value without features.

- **Categorical Analysis**

The bar chart shows the distribution of houses based on their proximity to the ocean, as indicated by the ocean_proximity categorical feature.

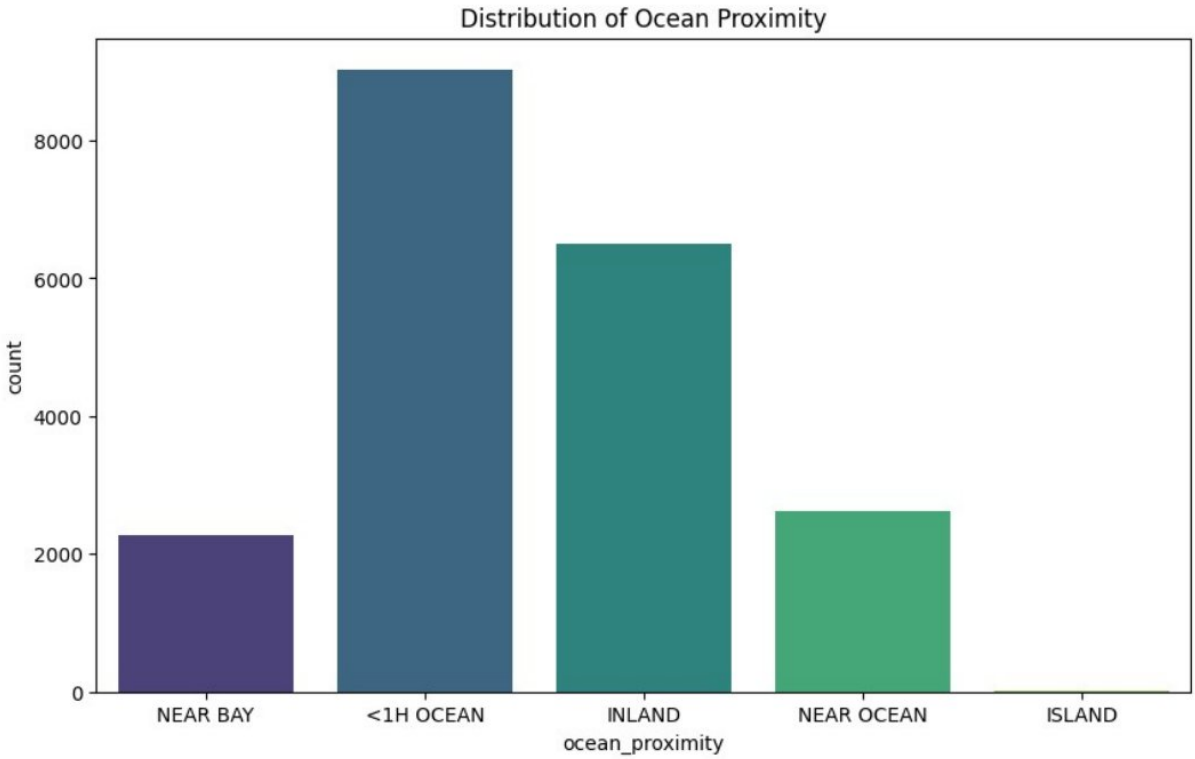

Fig. 1.7 Distribution of the Ocean Proximity features

The majority of the houses are within a one-hour drive to the ocean, as indicated by the "<1H OCEAN" category. This category has the highest count, suggesting that a significant portion of the dataset is composed of houses near the coast.

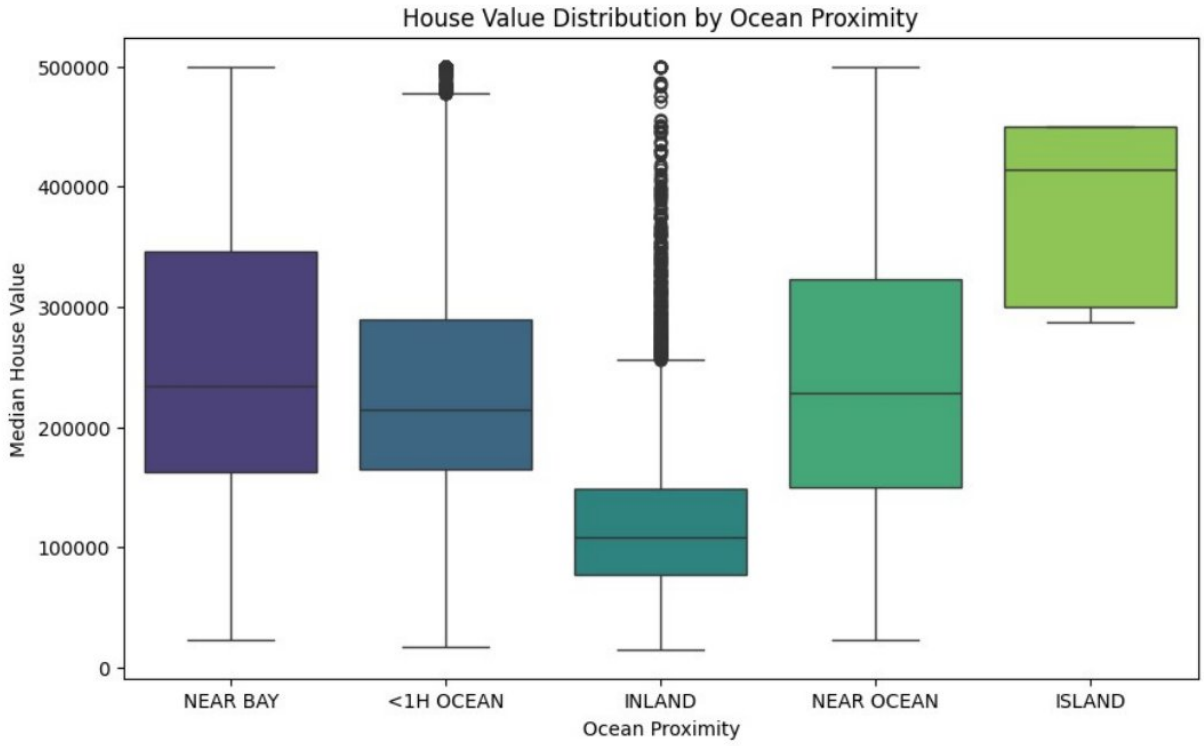

Fig. 1.8 Box plot showing the House value Distribution

The next largest category is "INLAND," followed by "NEAR OCEAN" and "NEAR BAY." This shows a broad range of housing locations, from coastal to inland areas.

Since proximity to the ocean could significantly impact house prices, this distribution indicates that the dataset is skewed toward houses near the coast, which might affect model predictions. In this case, we will be using the RandomOverSampler to duplicate samples in the minority class without creating synthetic data.

The plot highlights the significant impact of ocean proximity on housing prices. Properties closer to water, especially on islands and near bays, tend to have higher and more stable values compared to inland properties. This trend underscores the importance of location in determining property values in the California housing market.

- **Geographical Visualization**

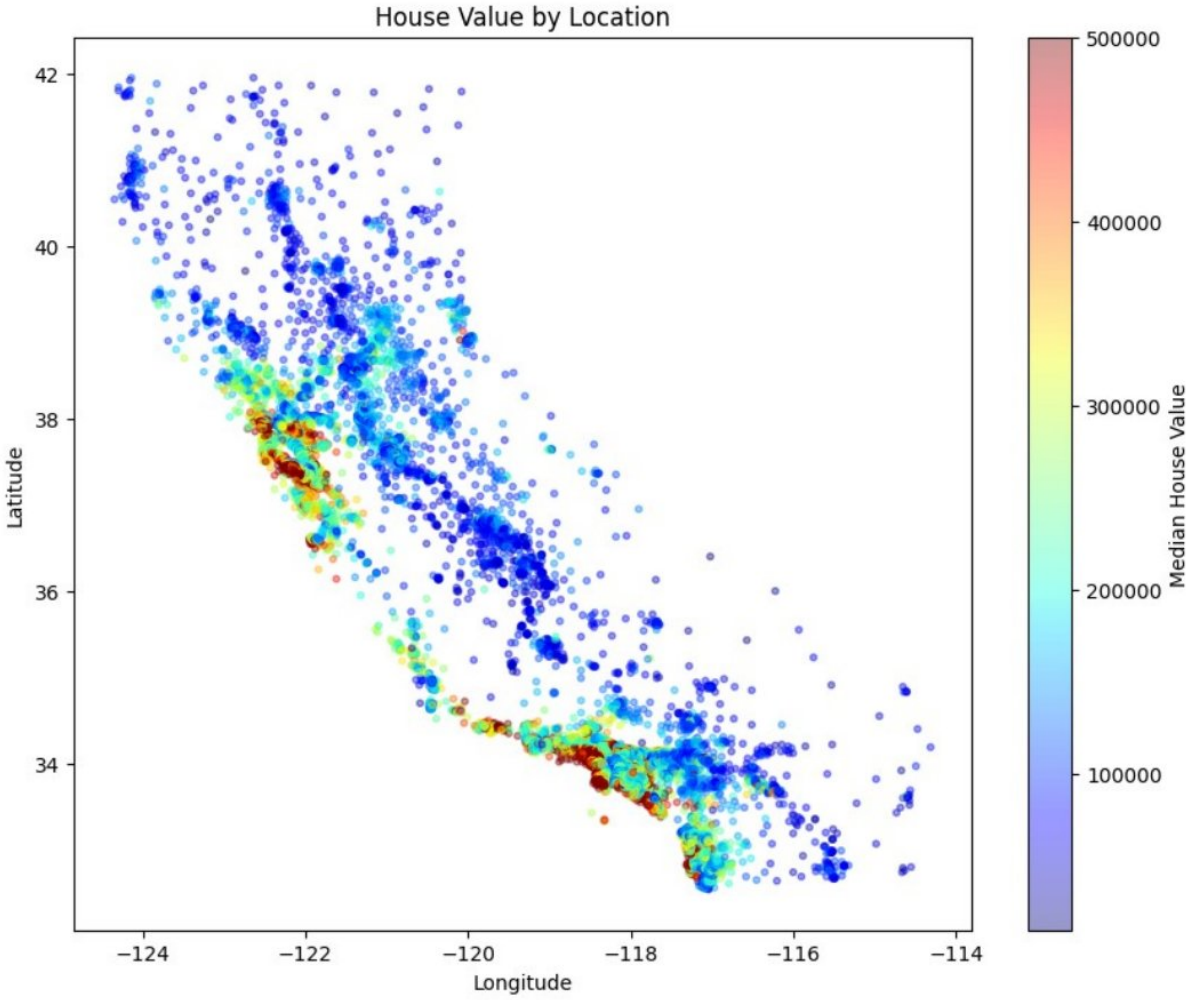

Fig. 1.9 scatter plot maps longitude vs. latitude

This scatter plot maps longitude vs. latitude with points colored based on median_house_value.

Higher house values (darker red points) are concentrated in specific geographic areas, particularly around coastal regions.

There is a gradient effect visible, with house values generally decreasing as one moves inland from the coast.

The plot visually confirms the geographic distribution of house prices in California, with higher prices near the coast and lower prices inland.

4.0 Feature engineering

4.1 Log Transformation

Applied logarithmic transformation to skewed features like total_rooms, total_bedrooms, population, and households.

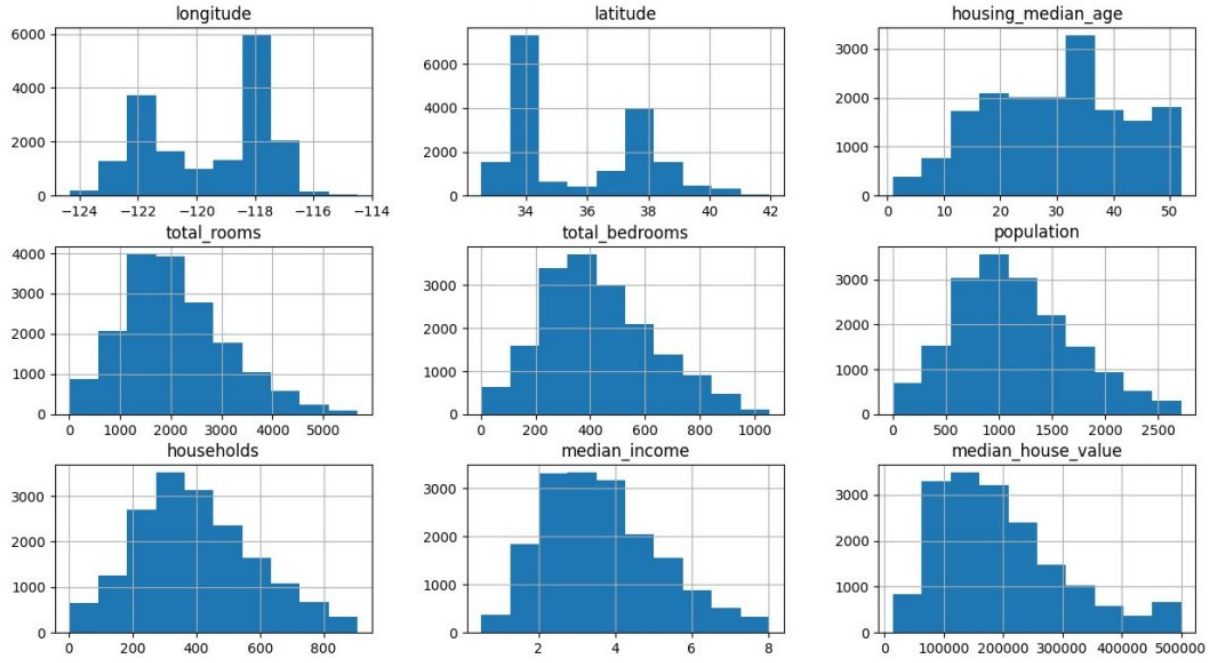

Fig. 2.0 Distribution of numerical features after skewness is adjusted

These histograms help understand the spread and central tendency of different features in the dataset, which is crucial for any data analysis or machine learning model building.

4.2 Converting Ocean_proximity columns to binary features

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity_<1H OCEAN	ocean_proximity_INLAND	ocean_proximity_ISLAND
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	0	0	0
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	0	0	0
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	0	0	0
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	0	0	0
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	3.6591	299200.0	0	0	0

Fig. 2.1 showing the Ocean Proximity features

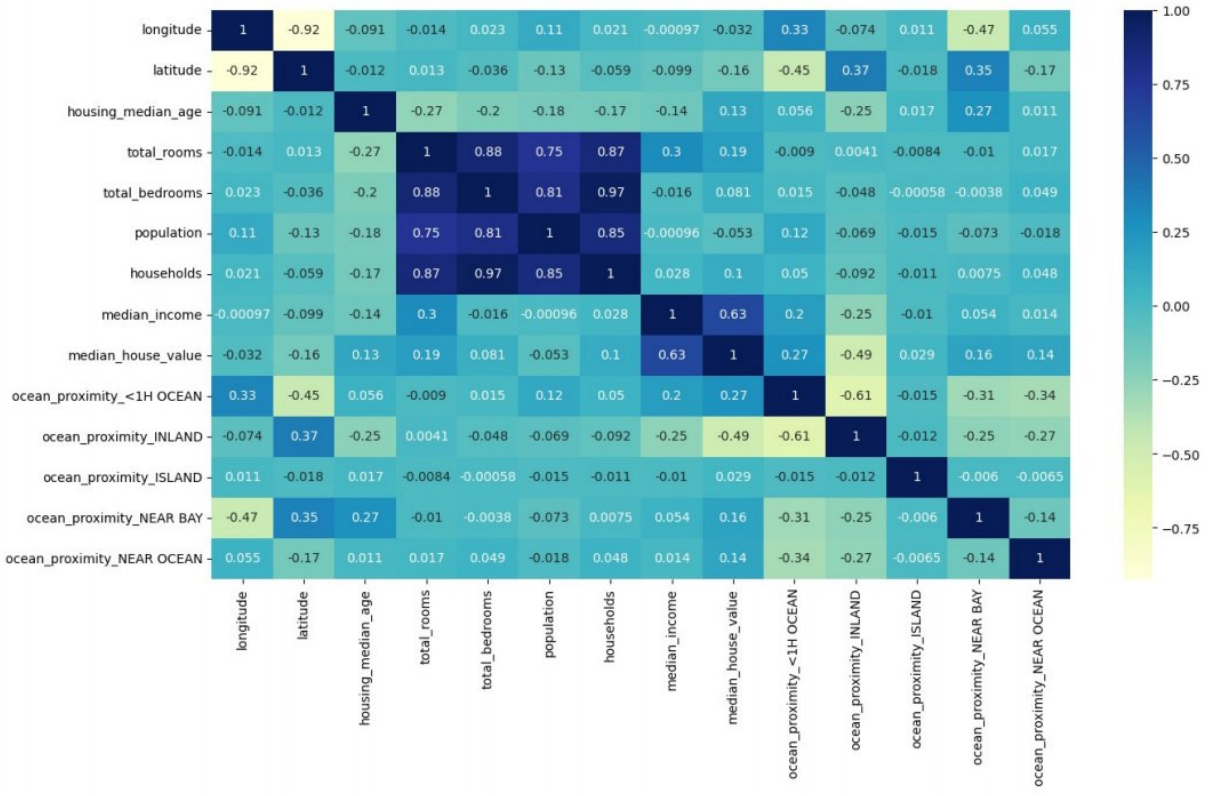

Fig. 2.2 Heat map showing the new columns from ocean proximity

Longitude and latitude have a strong negative correlation (-0.92).

total_rooms, total_bedrooms, and households are highly correlated with each other (positive correlation).

median_income has a moderate positive correlation (0.63) with

median_house_value.

Categories of ocean_proximity show various correlations with other features, like ocean_proximity_INLAND being negatively correlated with median_house_value (-0.49).

4.3 Features creation

Created new features such as bedroom_ratio (bedrooms/households), household_rooms (rooms/households), and rooms_population_interaction (interaction term between rooms and population).

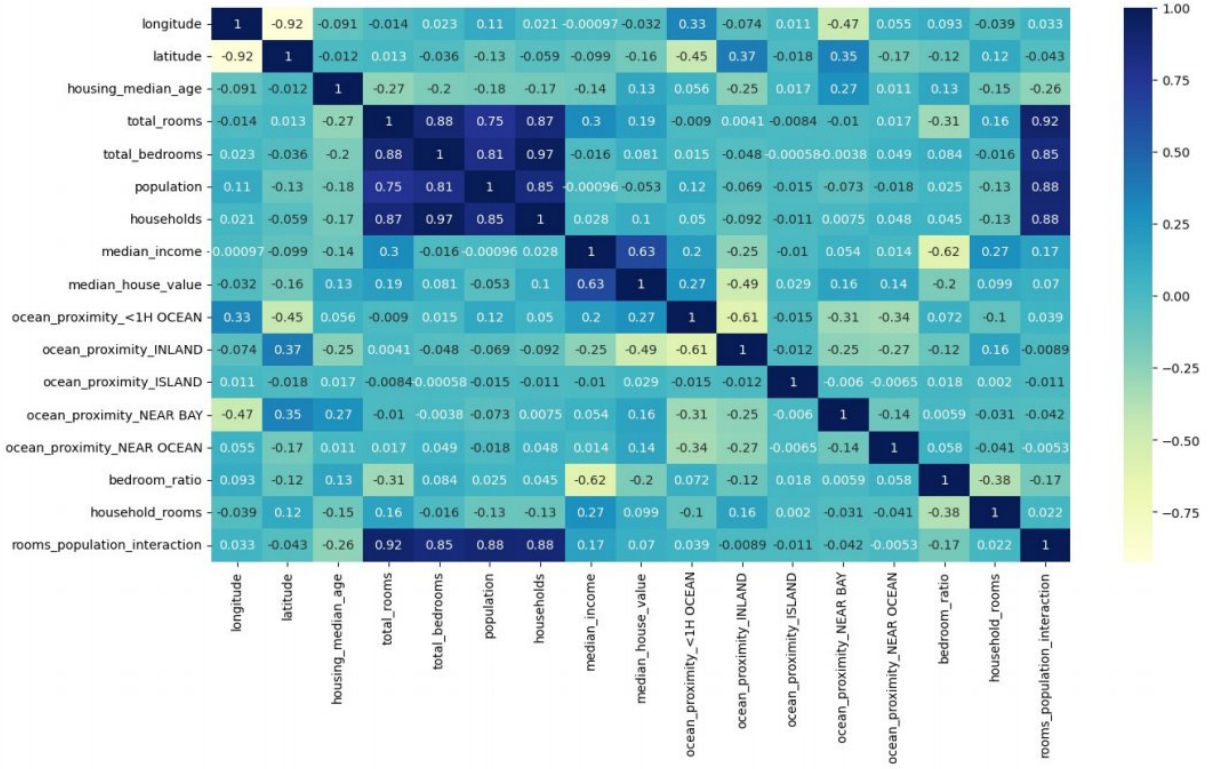

Fig. 2.3 Heatmap showing new columns (bedroom ratio, household room, rooms population interaction)

A high ratio might indicate smaller or more compartmentalized homes, while a lower ratio suggests larger or more open spaces. This feature could help the model better understand how the layout of a home influences its value, potentially capturing nuances in housing quality or appeal not directly reflected in raw counts of rooms or bedrooms.

The household_rooms can give insight into the average living space available per household, which could be a strong indicator of comfort or desirability, thus influencing house value predictions.

The rooms_population_interaction feature could highlight the relationship between housing density and value, potentially capturing the desirability of more or less crowded areas.

4.4 Normalization or standardization of numerical features

Normalized numerical features using StandardScaler to bring all features to a similar scale.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	bedroom_ratio	household_rooms	rooms_population_interaction
0	0.213996	0.564293	1.0	0.258241	0.178877	0.181651	0.193584	0.898126	0.032796	0.052756	0.048327
1	0.212982	0.564293	1.0	0.224220	0.221694	0.204495	0.240044	0.683573	0.093843	0.035241	0.047215
2	0.212982	0.564293	1.0	0.286445	0.264510	0.207074	0.284292	0.444750	0.080107	0.038534	0.061054
3	0.212982	0.564293	1.0	0.161643	0.200761	0.151069	0.211283	0.470082	0.146415	0.027757	0.025208
4	0.212982	0.563231	1.0	0.446501	0.463368	0.401990	0.566372	0.419883	0.103222	0.028964	0.184196

Fig. 2.4 showing Normalized data

5.0 Model Training and Evaluation

5.1 Model Selection

Models Evaluated:

I used the Random Forest Regressor, XGBoost and the Neural Networks which excel at handling outliers and capturing complex, non-linear patterns.

Linear regressor: A powerful machine learning model that fits a linear relationship between input features and a target variable.

Random Forest Regressor: A robust model for regression tasks, which performs well with tabular data.

XGBoost Regressor: An advanced boosting model known for its high performance in regression tasks.

Model Evaluation Metrics: They are used to assess the performance of machine learning models.

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.

R-squared: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

5.2 Train-Test Split

The data was split into Train and test set to allow for evaluation on the test set (unseen data).

The data was split 80/20 in favor of the train set

5.3 Linear Regression Model

The Linear Regression model was chosen as the base model to establish a foundation for other

ensemble models to build upon and enhance.

Evaluation

- **Mean Squared Error (MSE): 4,399,305,515.55**

R-squared (R^2): 0.6213

- Mean Squared Error (MSE): This value represents the average squared difference between the actual and predicted median_house_value. The large value here suggests that, on average, the predictions are somewhat off, possibly due to outliers or because the linear model may not fully capture the complexity of the data.

- R-squared (R^2): This value indicates that about 62.13% of the variance in median_house_value is explained by the model. While this is a reasonable start, it suggests that there is still room for improvement, as around 37.87% of the variance remains unexplained.

5.4 Random Forest

Also an ensemble learning method that combines multiple decision tress to improve the predictive accuracy and control over fitting.

Evaluation

- Mean Squared Error (MSE): 2,559,818,185.20

- R-squared (R^2): 0.7797

- Mean Squared Error (MSE): The MSE has decreased from the previous model, indicating that the Random Forest model is making more accurate predictions than the Linear Regression model.

- R-squared (R^2): The R^2 value has increased to approximately 0.7797, meaning that about 77.97% of the variance in median_house_value is now explained by the model. This is a notable improvement over the 62.13% explained by the Linear Regression model.

- Feature Importance

Checking the features that contiributed the most to the model performance. Checking the features that contiributed the most to the model performance

```
Feature Importances:
 median_income                0.423226
ocean_proximity_INLAND        0.143230
longitude                     0.103111
latitude                      0.097673
housing_median_age            0.059986
bedroom_ratio                 0.036942
population                    0.029206
household_rooms               0.028858
total_bedrooms                0.021303
total_rooms                   0.018251
households                    0.014474
rooms_population_interaction  0.010964
ocean_proximity_NEAR OCEAN    0.008842
ocean_proximity_<1H OCEAN     0.002874
ocean_proximity_NEAR BAY      0.000772
ocean_proximity_ISLAND        0.000288
```

Fig 2.5 showing the levels of features importance

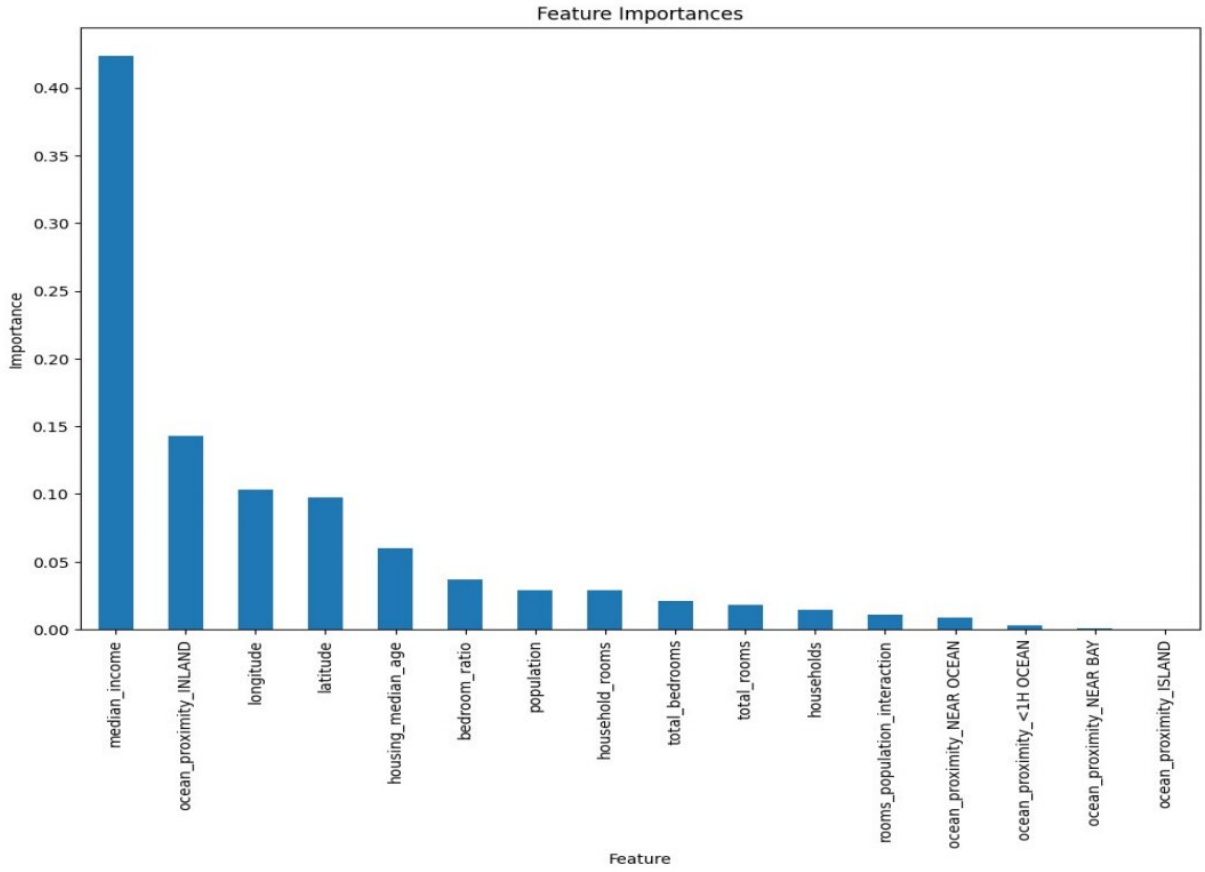

Fig. 2.6 Feature importance Distribution

5.5 XGBoost model

Powerful and highly efficient gradient boosting algorithm widely used for supervised learning tasks.

XGBoost MSE: 2298286800.948889

XGBoost R-squared: 0.802179395154528

Those are excellent results! Your XGBoost model shows a Mean Squared Error (MSE) of about

2.3×10^9 and an R-squared value of 0.802, indicating that the model is performing well in predicting housing prices and explaining a substantial amount of the variance in the target variable.

- Hyper parameter Tuning

The XGBoost hyper parameters were tuned using GridsearchCv module with an aim of finding the best hyper paramters and improving models performance

The below hyper parameters were obtained:

Fitting 3 folds for each of 10 candidates, totalling 30 fits

Best parameters: {'subsample': 1.0, 'n_estimators': 100, 'max_depth': 5, 'learning_rate': 0.1, 'gamma': 0.1, 'colsample_bytree': 1.0}

Best score: -2442903570.403489

It looks like your hyperparameter tuning helped improve the XGBoost model's performance slightly.

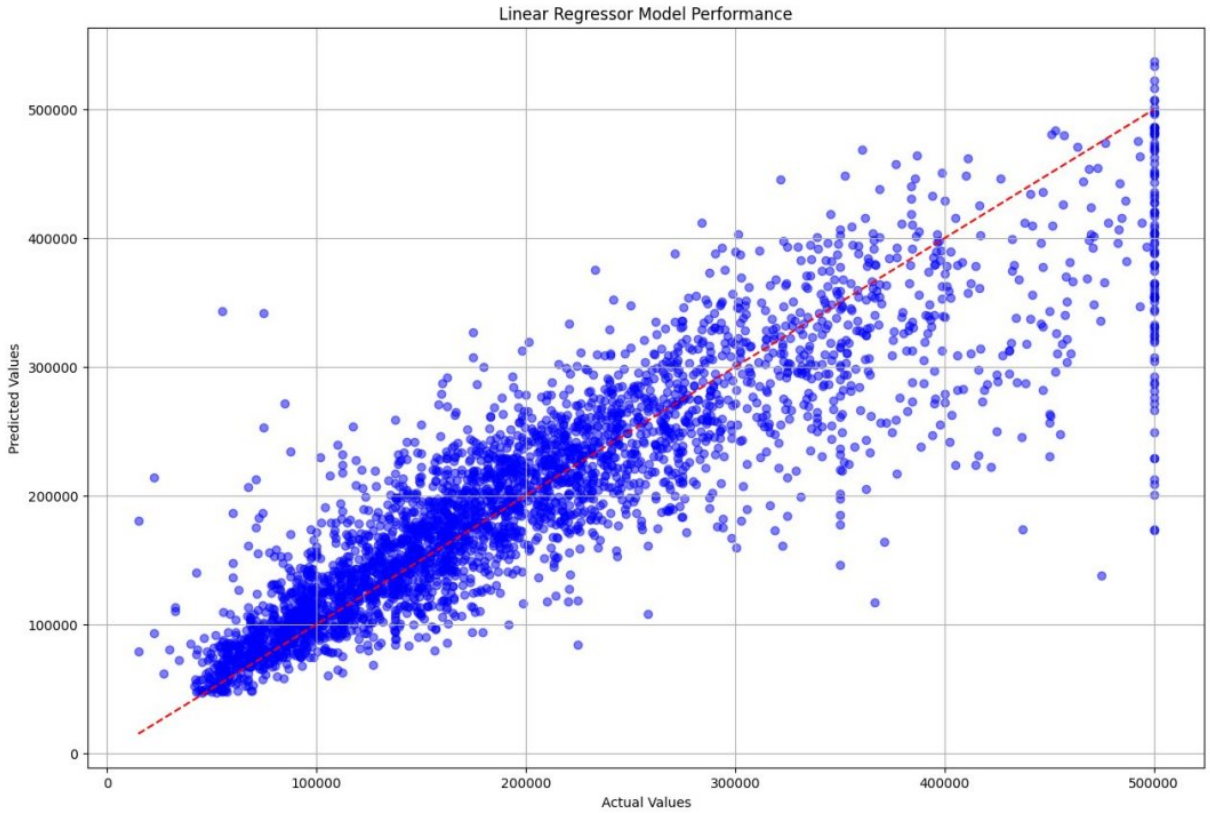

Fig. 2.7 Visualizing the model performance

The model seem to be performing well. The RMSE values suggest that the model is capable of generalizing to new data reasonably well.

- Feature Importance

Checking the features that contiributed the most to the model performanceChecking the features that contiributed the most to the model performance

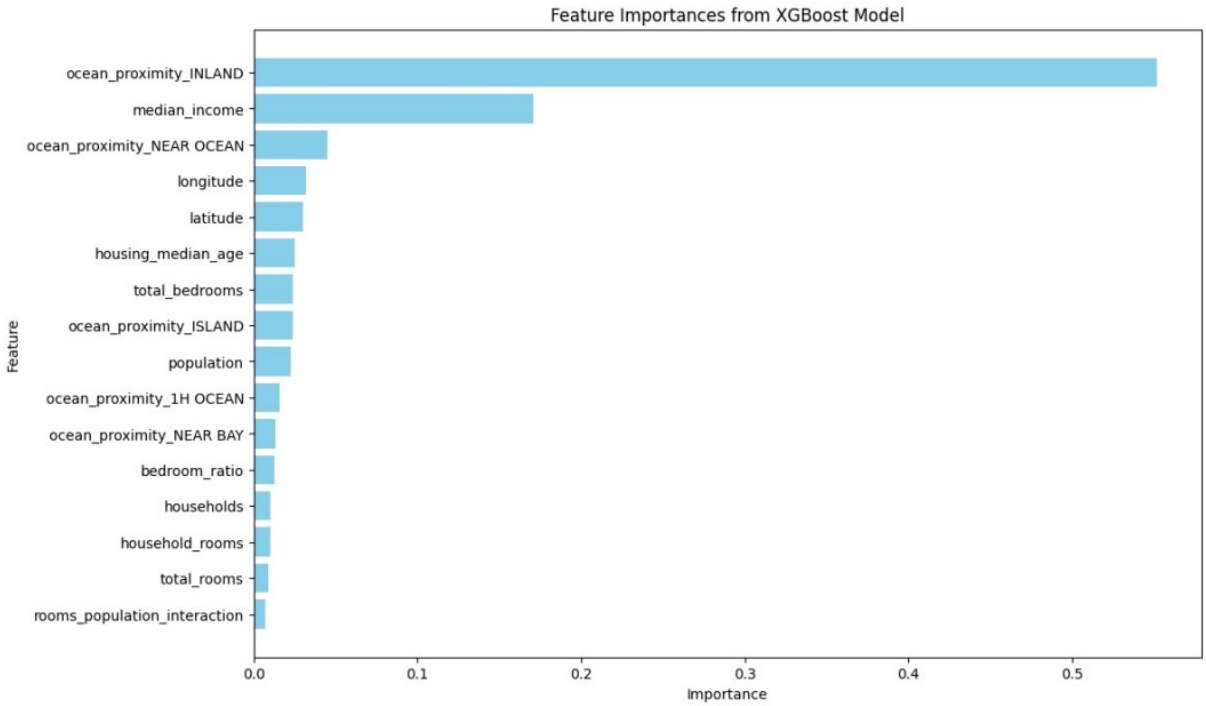

Fig. 2.8 Feature importance for XGBoost model

6.0 Conclusion

6.1. Summary

The XGBoost model demonstrated the highest performance with an R-squared value of 0.8022 and an MSE of 2,298,286,800.95, indicating it is the best-performing model among those evaluated.

6.2 Recommendations

Deployment: The XGBoost model is recommended for deployment due to its superior performance.

Future Work: Explore additional feature engineering and model tuning to further improve performance.

6.3 Limitations

Data Quality: The performance is subject to the quality and completeness of the dataset.

Model Assumptions: Assumptions made during preprocessing and model training may impact results.

Appendix

A.1 Dataset Description

The dataset used in this project is the California Housing Prices dataset, sourced from the California Housing Prices dataset provided by the California Department of Public Health. This dataset contains various features related to housing in California, including:

longitude: Longitude coordinate of the house location.

latitude: Latitude coordinate of the house location.

housing_median_age: Median age of the houses in the area.

total_rooms: Total number of rooms in the house.

total_bedrooms: Total number of bedrooms in the house.

population: Population of the area.

households: Number of households in the area.

median_income: Median income of the area.

bedroom_ratio: Ratio of bedrooms to total rooms.

household_rooms: Ratio of household rooms to total rooms.

rooms_population_interaction: Interaction term between rooms and population.

ocean_proximity: Categorical feature indicating proximity to the ocean.

A.2 Data Preprocessing

The following preprocessing steps were performed on the dataset:

Handling Missing Values: Missing values in total_bedrooms were imputed using the median of the column.

Feature Engineering:

Created new features such as bedroom_ratio, household_rooms, and rooms_population_interaction.

Applied one-hot encoding to the ocean_proximity categorical feature.

Normalization: Scaled features to a standard range using StandardScaler.

Outlier Handling: Outliers in the median_house_value were managed using statistical methods to ensure the accuracy of predictions.

A.3 Model Selection and Evaluation

Linear Regression

Mean Squared Error (MSE): 4,399,305,515.55

R-squared: 0.621

Random Forest Regressor

Mean Squared Error (MSE): 2,559,818,185.20

R-squared: 0.779

XGBoost Regressor

Mean Squared Error (MSE): 2,296,286,800.95

R-squared: 0.802

Cross-Validation Results:

Mean MSE: 2,282,792,364.03

Standard Deviation MSE: 64,912,368.19

Mean R-squared: 0.799

Standard Deviation R-squared: 0.015

Neural Network

Mean Squared Error (MSE): 3,546,161,510.05

R-squared: 0.695

A.4 Hyperparameter Tuning

For the XGBoost model, hyperparameters were tuned using Grid Search:

Parameters Tuned:

n_estimators: Number of boosting rounds

max_depth: Maximum depth of the trees

learning_rate: Step size shrinkage

subsample: Proportion of training data used for fitting

Best Parameters:

n_estimators: 100

max_depth: 10

learning_rate: 0.1

subsample: 0.9

A.5 Learning Curve

XGBoost Model Learning Curve:

The learning curve for the XGBoost model was generated to assess performance across different training set sizes.

Visualization of learning curves provided insights into model performance and overfitting.

A.6 Model Deployment

The final XGBoost model was saved to a file using joblib for deployment:

A.7 References

California Housing Prices dataset: California Housing Prices dataset

Scikit-learn Documentation: Scikit-learn

XGBoost Documentation: XGBoost