Predicting Housing Price in California

(Special Focus on San Diego)

Data Mining Project

ABSTRACT

This project aims to predict housing prices in California, with a focus on estimating the housing price in San Diego, California, and figuring out what factors impact on the price. By utilizing the California housing price dataset and employing various techniques, I am hoping to develop an accurate and reliable predictive model. This project involves data preprocessing, model training and evaluation to achieve the goals.

INTRODUCTION

Housing prices are influenced by multiple factors like location, size of a property and housing market economy. This project intends to build a machine learning model to predict housing prices in California, with a specific case study on the price prediction of housing price in San Diego and figure out which factors impact the price.

Understanding the factors that affect housing prices is important to various stakeholders like potential homebuyers and real estate agents. Accurate price predictions can also help in assessing market trend and making strategic investment decisions.

PREVIOUS WORK

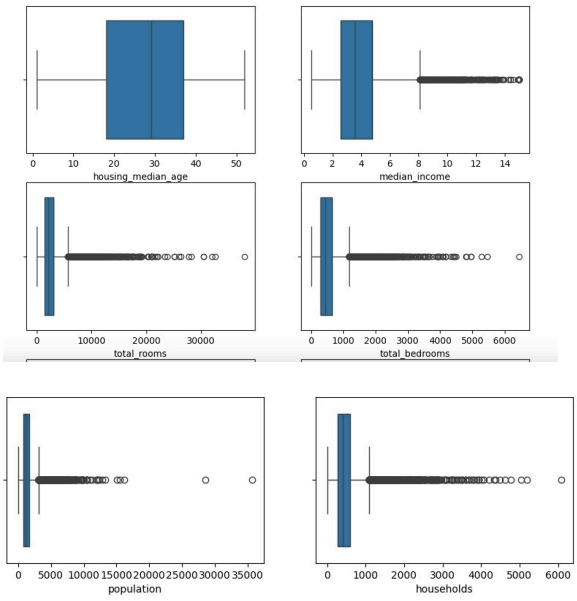
Previous studies have explored housing price prediction using various techniques and datasets, often focusing on broader geographical areas or different property types. However, this project has targeted analysis specific on the housing price in San Diego. Existing models may not fully capture the unique local factors influencing housing prices in the region. This project aims to fill this gap by providing a focused analysis and developing a tailored predictive model.

Many studies have employed regression models and machine learning algorithms to predict housing prices. Researchers have been conducted on the impact of location, property size and other features on property values. Linear regression, decision trees and ensemble methods have been used to develop predictive models.

Hedonic Pricing Models have been extensively used to evaluate the impact of various factors on housing prices. Also, spatial analysis techniques like Geographical Information System (GIS) and spatial econometrics have been used to predict housing data.

PROPOSED WORK

1. Data Collection: California_housing data from Kaggle
2. Data Preprocessing: Missing values, categorical variables and outlier were checked. The original dataset was clean. There were no missing values and also no categorical variables. The presence of outliers was checked with box plots.

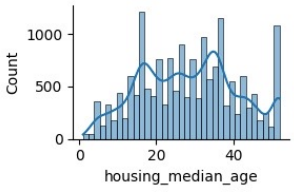


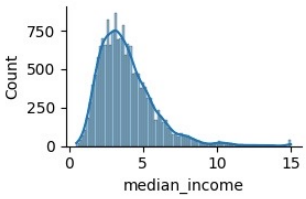
First 2 plots are housing_median_age and median_income respectively from left to right. Next are total_income and total_bedrooms. The last 2 plots are population and households.

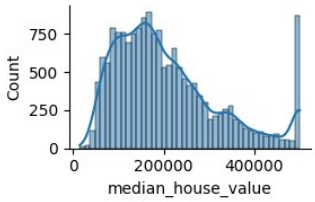
Except housing_median_age, other plots showed narrow box plots and a lot of outliers. It was impossible to remove all the outliers so I decided not to remove any of them.

Before exploring the data, I stored the rows, which belongs to San Diego based on the longitude and latitude values, separately as San_Diego. This set was about 5% of the whole original dataset. I used the longitude(-117.16 ± 0.2) and latitude(32.71 ± 0.2). The 0.2 range was determined by the location of Oceanside, which is the area of North San Diego, and Cross Border Express, which is the border crossing between San Diego and Tijuana, Mexico. This San_Diego set was stored separately to use at the end of the project for the house value prediction.

3. Exploratory Data Analysis (EDA): I selected some features to visualize their distributions: housing_median_age, median_income and median_house_values.

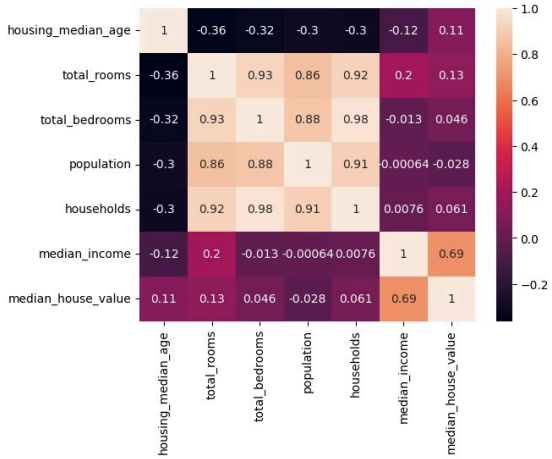






Median_income distribution was not terrible but skewed to the left side and has a long tailing. Those tailing must be the outlier data points on the boxplot.

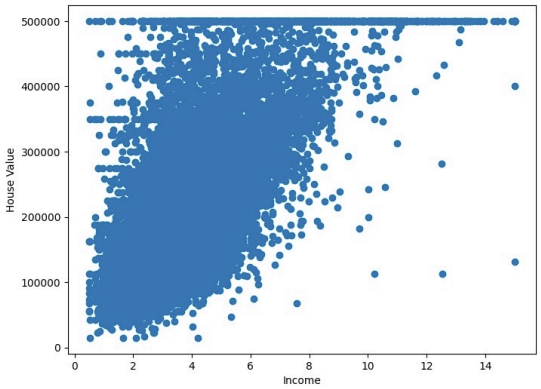
Median_house_value was the target variable. It looked like it was not normally distributed. The houses which had values over 500,000 were all counted as 500,000 so there was a big count at 500,000. Due to the big count at 500,000, the target variable needed to be normalized after EDA.



A heat map was created to check the correlation of other features with the target variable, media_house_value. The target variable was most correlated with median_income, 0.69.

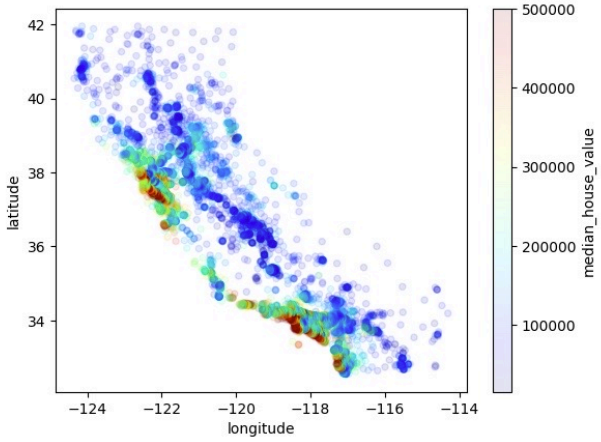
Other than that, total_rooms, total_bedrooms, population and households were all closely correlated to each other, and this make sense. If there are more people, there will be more households and bigger households need more rooms.

Since the most correlated feature with the target variable was median_income, I wanted to visualize their relationship.



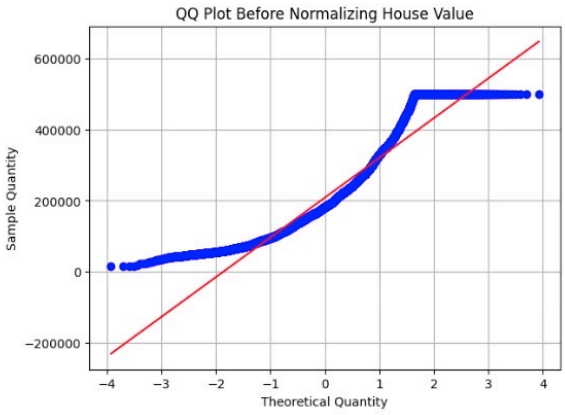
The plot(Income vs House Value) above showed that as the income increases, the house value increases. This made sense.

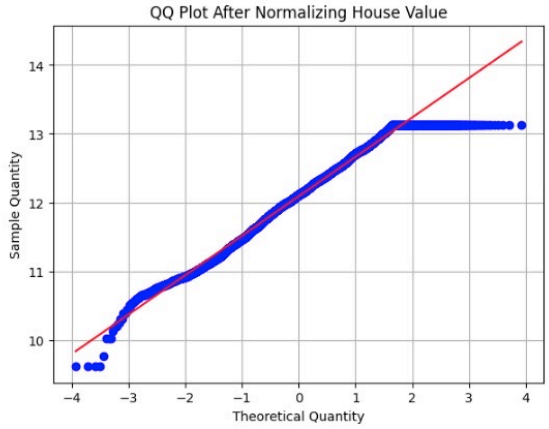
Last EDA was plotting the longitude and latitude data points.



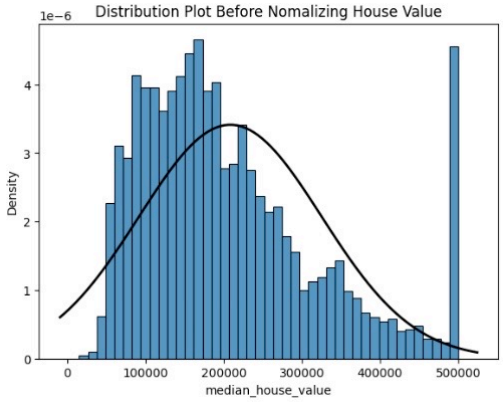
The longitude and latitude plot was the shape of California. The color indicated the median house value. More red is higher house value and more blue is lower house value. There were 3 area with high house value; the top one was San Francisco and there were Los Angeles, Orange County and San Diego at the bottom. Generally the coast line showed higher house value compared to the inner land house value.

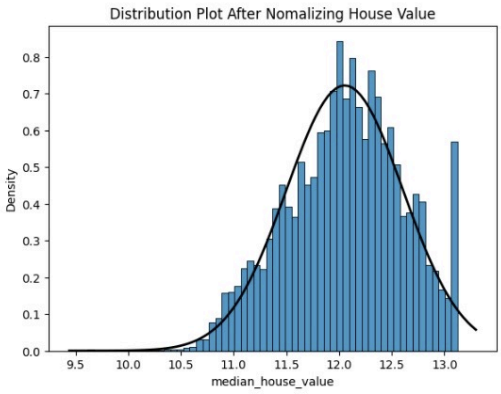
Before going into the model development step, I normalized the target variable. In order to visualize the difference between before normalization and after the normalization, I created QQ plots and histograms.





First QQ plot was before the normalization and the data points were definitely not the reference line. This indicated that the data was not normally distributed. However after the normalization, the data became more normally distributed. The target variable was normalized as log(target variable +1).





The first histogram was before the normalization and the bottom histogram was after the normalization. These histograms proved that the normalization worked well on the target variable.

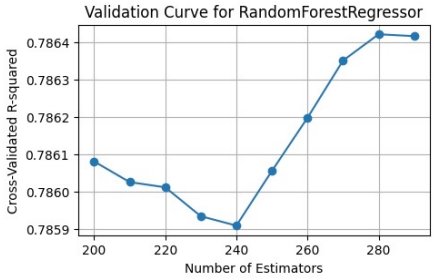
Finally, everything became ready for the model development after the normalization. The data was split into training set(80%) and test set(20%). Then, the training and test sets were scaled.

4. Model Development: Total 6 models were developed: linear regression, ridge regression, support vector regression (SVR), decision tree regressor, random forest regressor, and gradient boosting regressor(GBR).

The linear regression, SVR and decision tree regressor did not have complicated parameter selection. However, ridge regression, random forest regressor and GBR required some parameter inputs so I performed some tests to select the best performing parameters.

For the ridge regression, there was a parameter input called alpha. 3 values of alpha were tested: 1, 10 and 100. Out of these 3 values, alpha =10 performed the best by resulting the highest R-squared score and lowest Mean Squared Error (MSE).

The random forest regressor required n_estimator, max_depth and random_state as its parameter inputs. 3 different ranges of n_estimator were tested. 10 to 100, 100 to 200 and 200 to 300. From the each range, the best n_estimators were 90, 190 and 290. These 3 values were applied to the regressor and n_estimator = 280 performed the best.



Lastly, the GBR required random_state, n_estimator, learning_rate and max_depth but only n_estimator, learning_rate and max_depth were tested to find the best parameter inputs. Here are the values tested for each parameter. n_estimator: 100, 150, 190, 300. learning_rate: 0.01, 0.1, 0.25, 0.5. max_depth: 1, 5, 10, 15. The best combination of the parameter inputs was n_estimator= 300, learning_rate= 0.1 and max_depth=5.

5. Model Evaluation: Mean Squared Error (MSE) using 3 fold cross-validation and R-squared using 3 fold cross-validation were used as metrics. MSE measured how close predicted values are to actual value.

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

R-Squared measured goodness of fit.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Here are the result of MSE (first value) and R-squared scores (second value). Lower value is better for MSE and higher value is better for R-squared score.

Linear Regression: 0.116 & 0.654

Ridge Regression: 0.116 & 0.655

Support Vector Regression: 0.072 & 0.787

Decision Tree Regressor: 0.110 & 0.674

Random Forest Regressor: 0.073 & 0.783

Gradient Boosting Regressor: 0.053 & 0.842

GBR got the lowest MSE and the highest R-squared score. Therefore, GBR was chosen as the best performing model for the dataset.

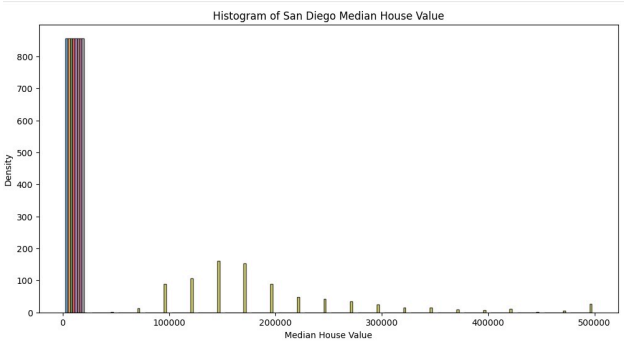
To confirm the performance of the model, the model was applied to the test set and the results were: 0.066 & 0.794.

There result using the test set was still better than other models on training set.

6. Data Analysis: The best model was chosen which is GBR. Finally, the model could be applied to the San_Diego set to predict the house value in San Diego.

The predicted median house value in San Diego was 196,739 and the actual median house value was 158,900. The difference between these values was 23.8% which is pretty high.

The predicted median_house_value seemed low but the actual value was even lower so a histogram was created to check what happened to the value.

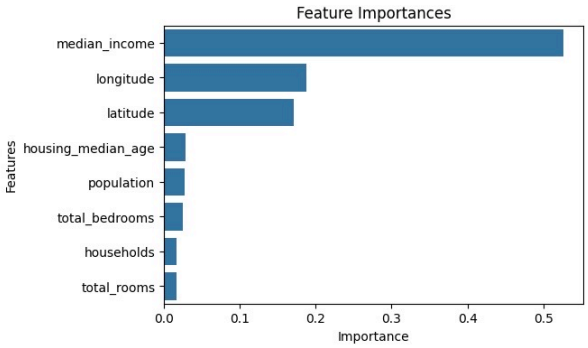


The histogram was heavily skewed to the lower value side. This is why the actual median house value was so low because there are too many low value in the dataset.

Realistically, it did not make any sense that there are that many of low valued houses in San Diego. There must be some mistakes in the original dataset or during collecting the data.

Another objective of this project was to find the influential features to the target variable.

GBR was used to create the plot.



The most influential feature to the target variable was median_income. This result matched the result from the heat map. Other than median_income feature, the location features were next influential features.

DISCUSSION

The project had two primary objectives: identifying the feature most influential to housing values and predicting house values in the San Diego area. The dataset contained eight features —median_income, longitude, latitude, housing_median_age, population, total_bedrooms, households, and total_rooms—with median_housing_value as the target variable.

To predict San Diego housing values, six models were developed: linear regression, ridge regression, support vector regression, decision tree regressor, random forest regressor, and gradient boosting regressor. Model performance was evaluated using the R-squared score and Mean Squared Error (MSE), with higher R-squared values indicating better goodness of fit and lower MSE values reflecting closer predictions to actual values.

Among the models, the gradient boosting regressor performed the best, achieving an MSE of 0.053 and an R-squared score of 0.842, while the R-squared scores of the other models did not exceed 0.8. The gradient boosting regressor was then trained on the training set and tested on the test set, yielding an MSE of 0.066 and an R-squared score of 0.794. Although there was a slight drop in performance on the test set, these results remained superior to those of the other models.

The gradient boosting regressor predicted a house value of 196,739 for San Diego. However, the median house value in the San Diego dataset was 158,900, resulting in a discrepancy of 23.8%, which is substantial. This raised concerns about the dataset's accuracy. The histogram of median_house_value for San Diego revealed a heavy skew towards lower values, with many values below 50,000. This skew suggests that the dataset might be problematic, as such low values are unrealistic for San Diego's housing market.

Additionally, the gradient boosting regressor was used to visualize feature importance. The analysis confirmed that median_income was the most influential feature on the target variable, aligning with the heat map analysis that indicated median_income had the strongest correlation with median_housing_value.

CONCLUSION

The most influential feature affecting house value was household income, with an estimated house value for San Diego of $196,739. However, this estimate is unrealistic, as San Diego is one of the most expensive cities in the U.S. To assess the model's performance, the actual median house value in San Diego was retrieved from the dataset, revealing an even lower value than the estimate. A visualization of the house value distribution in San Diego showed a significant skew toward lower values, which is inconsistent with the current market conditions. This suggests two possible root causes: the data might be outdated, or there could have been errors in data collection.

While there have been many predictions made for California housing prices, there has been limited focus on smaller regions like San Diego. This project addresses that gap, highlighting the need for localized housing price analysis.

Timeline

Week 3 - 4: Data Collection, preprocessing and EDA

Week 4 - 5: Model Development and Training

Week 5 - 6: Model Evaluation and Tuning

Week 6 - 7: Analysis and completing the report

REFERENCE AND CITATION

1. Zhang, Z., & Wang, S. (2021). "Predicting Housing Prices with Machine Learning: A Review." *Journal of Real Estate Research*, 43(2), 255-278.
2. Smith, J., & Brown, A. (2020). "Real Estate Price Prediction using Regression Models." *International Journal of Data Science and Analytics*, 12(4), 345-362.
3. California Department of Housing and Community Development. (2023). "California Housing Data."
4. Ebhuoma, E. (2021). Housing California [Data set]. Kaggle. https://www.kaggle.com/datasets/ebhu9103/housing-california
5. LatLong.net. (n.d.). San Diego, CA, USA latitude and longitude. Retrieved August 15, 2024, from https://www.latlong.net/place/san-diego-ca-usa-7073.html
6. LatLong.net. (n.d.). Oceanside, CA, USA latitude and longitude. Retrieved August 15, 2024, from https://www.latlong.net/place/oceanside-ca-the-us-27920.html
7. Wikipedia contributors. (2024, August 15). Cross Border Xpress. In *Wikipedia, The Free Encyclopedia*. Retrieved August 15, 2024, from https://en.wikipedia.org/wiki/Cross_Border_Xpress