California Housing Price Prediction

Introduction/Problem statement:

Today, investments are a business activity that most people are interested in. There are several objects that are often used for investment, such as stocks, ETFs, cryptocurrencies, and property. Specifically, property investment has increased significantly, both in demand and in sales. A house's price depends on many factors, such as its size, its area, how many bedrooms, its location, and the price of other houses. Real estate investors want to know the actual cost of a house before buying and selling. Buying a house at a higher price and selling it at a lower price will cause them to lose money. Banks often want to know the current market value of a property when they use someone's home as collateral for a loan. Loan applicants sometimes overvalue their houses in order to borrow the maximum amount from the bank. Home buyers can also obtain mortgage loans from banks and financial institutions. Home buyers can also predict the house price to determine if a seller is asking too much. Additionally, local sellers can predict the price of their houses and determine what a fair market value is. This project is to look at California housing pricing and use machine model to determine what factors impact the housing price.

Dataset source and information:

The dataset came from a csv file from Kaggle. The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. The columns are as follows; their names are described below:

1) Median House Value: Median house value for households within a block (measured in US Dollars) [$]  
2) Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]  
3) Median Age: Median age of a house within a block; a lower number is a newer building [years]  
4) Total Rooms: Total number of rooms within a block  
5) Total Bedrooms: Total number of bedrooms within a block  
6) Population: Total number of people residing within a block  
7) Households: Total number of households, a group of people residing within a home unit, for a block  
8) Latitude: A measure of how far north a house is; a higher value is farther north [°]  
9) Longitude: A measure of how far west a house is; a higher value is farther west [°]  
10) Distance to coast: Distance to the nearest coast point [m]  
11) Distance to Los Angeles: Distance to the center of Los Angeles [m]  
12) Distance to San Diego: Distance to the center of San Diego [m]  
13) Distance to San Jose: Distance to the center of San Jose [m]  
14) Distance to San Francisco: Distance to the center of San Francisco [m]

Data Wrangling and EDA

I did inspection on the dataset and check if there are any missing values. The dataset is pretty clean so I didn’t need to do much of the cleaning process. Later, I explored some variables and see if there are any interesting trend or insight. I found the median income and house value has a positive relationship (pic 1). I also used histogram to see the distribution of each variable—Tot\_Room, Tot\_Bedrroms, Populartion, Households are skewed (pic 2).

Chart, scatter chart

Description automatically generated

Pic 1

A picture containing crossword puzzle

Description automatically generated

Pic 2

Chart, scatter chart

Description automatically generated

Based on this ggplot, I can see the house price incases as it more close to the coast especially in LA (Southern California) and Bay area (Northern California)

Chart, scatter chart

Description automatically generated

Population and Total room has a positive relationship.

I also drew a heat map to see the correlated between the variables which confirmed the Tot\_room, Tot\_Bedroom, Population, householad are highly correlated. Median\_income and house value are medium correlated.

Preprocessing/Machine Learning

The target variable is the median\_housing\_value and I use other variables as X’s in the model. For the modeling purposes, the data was split into training and test sets with 80% on the training data and 20% on the test data. I noticed the scale of the dataset isn’t consistent so I used feature scaling called StandardScaler method to standardize the data.

The models I have used are Random forest, Linear Regression, Lasso, and XGBRegressor. R2 is used here to determine which model give us the best result. I have the R2 values for each model below

|  |  |
| --- | --- |
| Model | R2 |
| Random Forest | 0.98 |
| Linear Regression | 0.65 |
| Lasso | 0.65 |
| XGBRegressor | 0.80 |

As the result, the highest score is the Random Forest model so I will pick that.

I also plotted the graph to display the futured importance from the model, I conclude the location is a key as I can see Longitude, Distance to San Jose, and Distance to coast are the highest ones under the location variables. We all know that the Silcon Valley is located in San Jose and also close to the coast. They tend to have higher housing price due to prosperous job market and economy. Next feature that impacts the price is the household and follow by the income.

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

This graph also shows the Random Forest model predicts and fits the data well

Limitation/Recommendations:

* The dataset was from 1990 census data, so the model was built here may not be able to predict the trend of the current housing price. However, we can apply the same concept here to predict the housing price if there is the data is available in the near future
* There are also more variables impacting the value of a home such as the home’s condition, age and property size. It would be nice to include these variables in the dataset and they play huge roles when it comes to housing price in the real life.