Housing price prediction

Sirra Alfa Wesley

Abstract

This project aims towards predicting the price of houses in future based on several factors like average income of a city's residents, Age of the building, number of rooms, population of the city etc. I used a public housing dataset for this analysis.

Method : Multiple linear Regression analysis using machine learning has been the main technique used for this prediction and I managed to calculate the coefficient on which the real estate price is dependant upon.

Motivation

Covid-19 has made a huge impact on the world economy, along with people's purchasing power, employment , bank interest rates etc. The same can be seen in the real estate market making it almost impossible for most people to purchase their dream house. A recent article on forbes titled 'housing market predictions 2022', made me curious to understand the relation between the population of a city, income of the residents of a locality, age of a building and the price of a house.

These coefficients can be used by any one who is planning to buy a house to estimate what kind of cities they can afford a house in, how old the structure would be etc based on your income level. Or vice versa you may use this to plan on how much money you need to make based on the kind of house you plan to purchase in the future.

Dataset(s)

A Housing prices related dataset has been used for this analysis. It contains Average Income of residents of the city house is located in, Average age of Houses in same city, Average number of Rooms for Houses in same city, Average number of Bedrooms for Houses in same city, Population of the city the house is located in, Price that the house was sold at and Address for the house. It contains about 5000 entries.

Data Preparation and Cleaning

To start with, I checked if there were any empty cells in the dataset using a dropna function on the dataframe. It returned the same number of entries which confirmed that the dataset was complete. Also, the min, max and mean of all the attributes were checked to confirm that there were no outliers, as I wanted to check why there were any extremities. At this point, the data looked clean and complete, hence I proceeded to the analysis stage of my project.

Research Question(s)

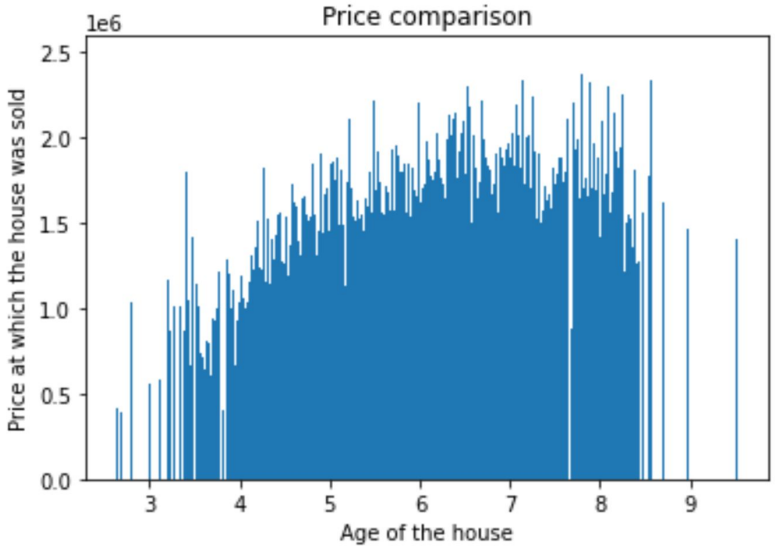
This project aims towards predicting the price of houses in future based on several factors like average income of a city’s residents, Age of the building, number of rooms, population of the city etc.

Methods

Describing the data on my jupyter notebook let me take a quick look at the different dimensions of the data I am dealing with. ‘Hvplot’ has been of good help during the analysis to visualise the whole data set and also individual attributes.

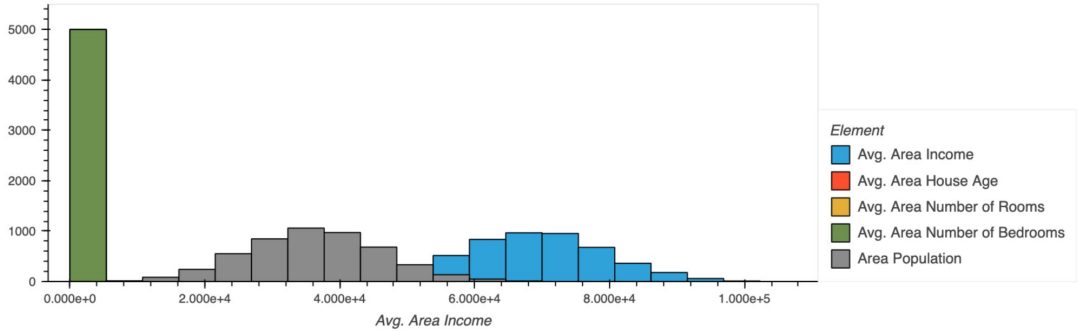
Prices of houses being numeric, linear regression seemed to be the first choice for my machine learning model. Sci kit learn was the key player during this part of the project. The whole of the dataset was divided into 2 parts, namely X and Y where they corresponded to X and Y in ‘y = mx + b’ in which ‘m’ denotes the slope of the line. Moving forward, the X and Y parts of the data set was further segregated into train and test data at 70-30 ratio to first train the model and test it respectively.

Also, functions like print_evaluate and evaluate have been defined to ease my process of evaluating the trained model

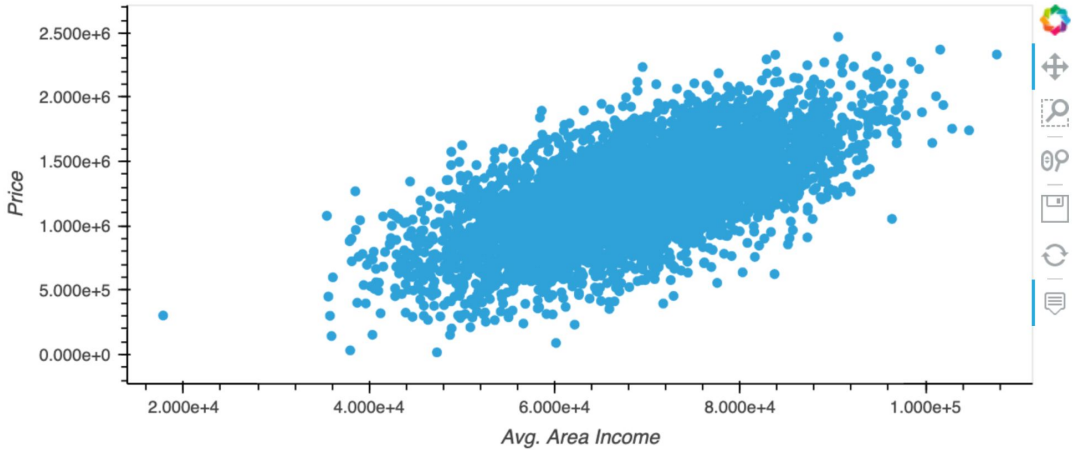


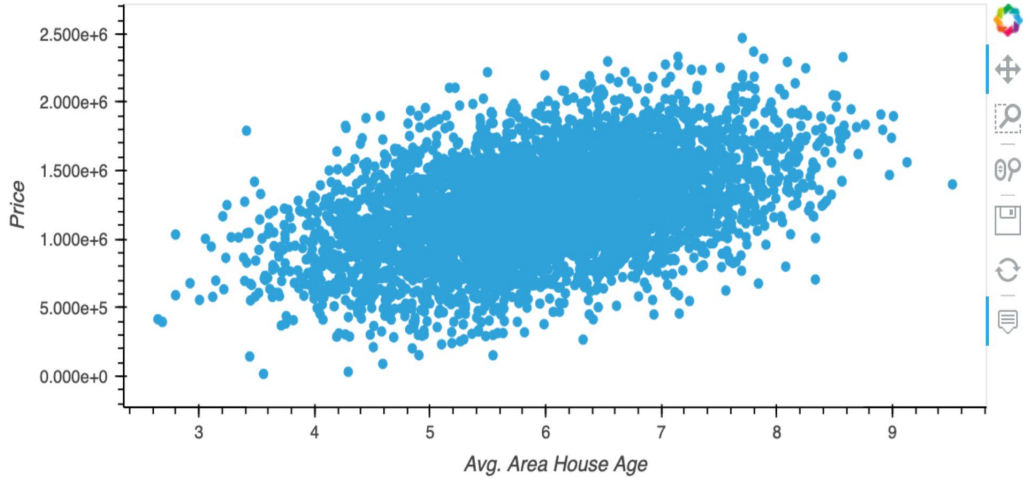
This bar graph makes a simple comparison of age of a house vs its price. As we can see, the price of the house is partially directly proportional to the age, until about 5-6 years. Once the age of the house reached 5 years, the price stayed relatively constant and din't improve much.

Contrary to my personal belief that the newer the house, the more expensive it is, this chart proves me wrong.



While the number of bedrooms seems to have a steady impact on the price of a house, the population of the city and the average income of people in that locality seem to be comparatively lower.





These 2 scatter plots are representing average income of the residents of the city, the house is located in and the average age of the house against the price of the house. They both seem to display a particularly similar trend.

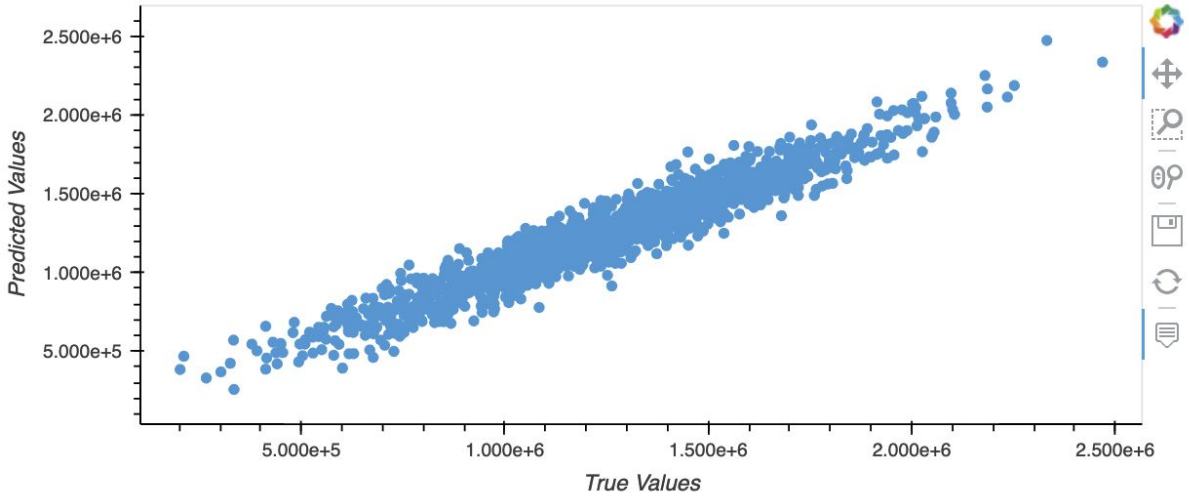
Higher the household income and the age of the house, the real estate seemed more valuable.

	Coefficient
Avg. Area Income	232679.724643
Avg. Area House Age	163841.046593
Avg. Area Number of Rooms	121110.555478
Avg. Area Number of Bedrooms	2892.815119
Area Population	151252.342377

Here are the coefficients of all the available dimensions of the dataset which we were able to calculate using the Linear regression model.

What does it mean by these coefficients?

- Holding all other features fixed, a 1 unit increase in **Avg. Area Income** is associated with an **increase of $23.26**.
- Holding all other features fixed, a 1 unit increase in **Avg. Area House Age** is associated with an **increase of $163841.04**.
- Holding all other features fixed, a 1 unit increase in **Avg. Area Number of Rooms** is associated with an **increase of $121110.55**.
- Holding all other features fixed, a 1 unit increase in **Avg. Area Number of Bedrooms** is associated with an **increase of $2892.81**.
- Holding all other features fixed, a 1 unit increase in **Area Population** is associated with an **increase of $15.12**.



Here is the scatter plot representing the values of real values of Y vs predicted values of Y using the LinearRegression machine learning model.

They seem to be pretty much on the same line, which is a good news, this means our model was more or less able to predict the Y values very close to the real ones.

Limitations

This data set is limited to just 6 attributes and I personally do not think real estate market is that simple. Price of housing mainly depends on additional factors like inflation, socio-economic stability in the state, global events like the current on-going ukraine war etc.

But this kind of analysis would become extremely extensive and could become too broad. Hence assuming my current factors are the only ones that mainly influence the prices in real estate, this analysis has be done.

Conclusions

Based on the given dataset, the coefficients for several factors like average family income, age of the house etc we found in relation to the price of the house. The same has been presented earlier.

Acknowledgements

The dataset was acquired from a databank called Kaggle. I did not get any feedback on my work from my friends or colleagues.

References

These article from forbes & bloomberg gave me a basic understanding on what major changes happened to the real estate market in 2022 and led me to choose this topic for the final project.

https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/

https://www.bloomberg.com/news/features/2022-06-21/cooling-real-estate-markets-in-us-uk-risk-deeper-global-economic-slump