CISD43 Final Project: USA Housing Price Prediction

-- with Data Analysis Method: Linear Regression & KNN

(By: Taichun Chao 06/06/2024)

This is a CISD43 Big Data and Modeling Analysis class final project. We used the provided data to work on this project. First, we need to use Jupiter Notebook to write the main codes. Second, using the Rapidminer application to analyze the data with two different Big Data methods. Lastly, we need to work the dataset in the NoSQL program, such as MongoDB or Neoj4, and find the questions that we are interested in.

The Dataset used in this project is USA_Housing.csv. It includes 7 columns and 5000 rows. The columns' names are **Avg. Area Income, Avg. Area House Age, Avg. Area Number of Rooms, Avg. Area Number of Bedrooms, Area Population, Price, and Address**.

In the EDA section, I replace the space and "." with "_", so the field name will integrate. To simplify the question I drop the Address column. As for the visualization, I pick up the histogram, to see all the items' quantity distribution, select the boxplot to find the relationship between the number of bedrooms and price, and use a heatmap to demonstrate the coefficient of each item. From the map, we can easily see that the house price has a very high coefficient with 'ave. area income'.

As for the data analysis methodology, I used Linear Regression and KNN to present my data. The reason is all the histogram present a close normal distribution curve. Linear Regression is a method to find the small error of the square root between the actual (training) value and predicted (test) value, to get a best-fit straight line, and from the best-fit straight line formula we can get the best-predicted price. The methodology of KNN (k-nearest neighborhood), is to find the new data and see which group it belongs to.

In summary, we can see that House prices are related to the average income and the number of bedrooms, if there are no address and location issues. I predicted the house prices with my Linear Regression formula, the price from the original $1,505891 jumps to $1,550,214. The conclusion is that house prices will keep growing if there is no big disaster.