Real Estate Valuation

Kun Huang, a second year graduate student in statistics
Yizhou Jin, a second year graduate student in statistics

group 6

Background

We aim to predict the house price \( \text{house\_preic} \) in New Taipei City, Taiwan using a data set containing the following variables,

- \text{trans\_date}: the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- \text{house\_age}: the house age (unit: year)
- \text{distance\_mrt}: the distance to the nearest MRT station (unit: meter)
- \text{stores}: the number of convenience stores in the living circle on foot (integer)
- \text{latitude}: the geographic coordinate, latitude. (unit: degree)
- \text{longitude}: the geographic coordinate, longitude. (unit: degree)

The response variable is \( \text{house\_price} \), the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared).

Methodology

We mainly use BigQuery ML to deal with this problem. Specifically, we use K-means clustering(\( K = 4 \)) to transform the geometric coordinates \text{latitude} and \text{longitude} into one variable \text{CENTROID\_ID} and linear regression to fit the data. Both two methods can be used directly in BigQuery ML. In linear regression, we use \text{distance\_mrt}, \text{stores}, \text{CENTROID\_ID}, and \text{house\_age} as predictor variables. We do not use \text{trans\_date} since we see from figure 1 that the transaction date has a relatively small influence on the house price.
Preprocessing and Exploratory data analysis

Since directly using the geographic coordinates as predictor variables will lead some problems. For example, the difference among geometric coordinates are small numerically but in practice they can be very different. Hence we transform the longitude and latitude into a categorical variable, CENTROID_ID by K-means clustering to indicate the location of each observation.

After looking at the map, see figure 2, we roughly observed that there are 4 clusters in the map, hence we decided to cluster longitude and latitude based on data features into 4 groups by using K-means clustering. To choose a more accurate $K$, we can divide the training data into training set and validation set, a preferred $K$ is that has less MSE on the validation set. We believe $K = 4$ is good enough, so we just let $K = 4$. This step is imposed on the whole data set, i.e., the training set and test set.

The Scatterplot of House Location

![Map with clusters](image.jpg)

By using the K-means model, we successfully cluster the latitudes and longitudes into four groups as follows.
Then we subset the result table by train group and test group and we used `CENTROID_ID`, `house_age`, `distance_mrt`, and `stores` as the predictors and `house_price` as the response variable to train the model and do the prediction.

Before training the model, we check the correlations among predictor variables, see figure 4.
We can also see from figure 4 that directly using geometric coordinates can include relatively high correlated variables, like `distance_mrt` and `Longitude`.

Model and Results

We build a linear regression model using the BigQuery ML. Then we run the regression model and we use trained model to predict house price based on test data. The result is showed as follows.
By submitting the data to Scriptedin, the score (MSE) is **13.1672** which is relatively good for a small dataset with size of 414. Hence we believe our model is good and we can use our model to predict the house price in New Taipei City, Taiwan.

Additionally, since our model is based on the observation in this area, our model is useful only in this area. For the observations in other areas, our model may fail.

**Future Work**

Our given data set only contains 290 observations. Our above work can be reproductive when the data set is large. Moreover, when we have more than 1000 observations, we can use the AutoML on google cloud platform to help us to fit the data using more flexible and tuned model, which we believe can improve the prediction results.