# Case Overview

The objective of this case is to determine what is a fair market value of properties for Cook County Assessor’s Office, home to the City of Chicago and over 130 other municipalities. We evaluate the historical residential property value using machine learning techniques and provide prediction of the value for each home as closely as possible to the actual values.

# Methodology

Our Methodology includes three parts, which are data pre-processing, model training, and finding best model by comparing mean squared errors (MSE).

* **Data Pre-Processing**

Firstly, we pre-process the data on 50,000 properties that have recently sold. We drop a total of 13 variables. To elaborate, 5 of them (char\_tp\_dsgn, char\_cnst\_qlty, char\_site, char\_repair\_cnd and meta\_cdu) have notes that describe as not useful for modeling based on the codebook. Also, we exclude 4 variables (meta\_nbhd, geo\_property\_zip, geo\_fips and geo\_municipality) that indicate location as property city. Furthermore, we decide to eliminate 4 variables (char\_apts, char\_attic\_fnsh, char\_renovationn and char\_porch) that contain more than 50% NAs, for we think it will cause bias if we replace those NAs. Then, we replace the remaining NAs with mode and mean respectively for character and numeric variable. After checking variable types, we discover there are some variables that exceed 50 levels, which can lead to mismatch levels between training and test data. Thus, we redefine the levels and transform them into around 50 levels and convert each variable to categorical or logical variables. Lastly, we conduct data partition. 60% of historic data is used to train the model and 40% of the data is used to make prediction and calculate MSE.

* **Model Training**

Since this is a regression problem, we select five different models, which are simple linear regression, stepwise regression, Lasso regression with the best lambda value, random forest with 10 variables randomly sampled at each split and regression tree. Then, we train our models with the training dataset and apply to the test dataset. In the end, we calculate MSEs for each model and choose the one with the lowest MSE for further prediction.

* **Finding the Best Model**

The following table is the summary of models we train and calculated MSE:

|  |  |
| --- | --- |
| Model | MSE |
| Simple linear regression | 15260334101 |
| Step wise regression | 15274230425 |
| Lasso regression | 15306767565 |
| Random forest | 13267269659 |
| Regression tree | 21482553307 |

Based on the table, we can tell that random forest has the lowest MSE. As a result, we predict the value of each property using our random forest model.

# Conclusion

The following table is the summary statistics for our predcitions:

|  |  |
| --- | --- |
| pid | assessed\_value |
| Length:10000 | Min. : 8559 |
| Class :character | 1st Qu. : 162023 |
| Mode :character | Median : 251484 |
|  | Mean : 326475 |
|  | 3rd Qu. : 380384 |
|  | Max. : 3658459 |

To interpret the summary statistics, we can start from looking into the home price in the United States. According to the Federal Reserve Bank of St. Louis, the median home price in the United States is $374,900 as of the second quarter of 2021. That's an increase of over $50,000 from just a year ago. Moreover, home prices increased by 16.2% from 2020 to 2021 and the median home price increased by 416% from 1980 to 2020.

We can tell that the median of our prediction is 33% lower than the median home price in the United States. The difference might be resulted from the influence of pandemic and inflation. However, based on Zillow Home Value Index (ZHVI), typical home price in the United States is $293,349, which is similar to the mean for our predictions. Thus, it is fair to conclude that our prediction is reasonable and can be improved if we consider economic factors.