**Final Project Report**

**The forecast of the house price**

Team member: Shangxing Jiang (40%, part modeling)

Xiaoning Zhou (30% part ETF)

Wenxi Zheng (30% part data processing)

目录

[Ⅰ. Introduction to the project 3](#_Toc532332777)

[Ⅱ. Challenges and Solutions 3](#_Toc532332778)

[Ⅲ. Data processing 4](#_Toc532332779)

[Ⅳ. Algorithm, results and Data Visualization 5](#_Toc532332780)

[Ⅴ. ETF 10](#_Toc532332781)

# Ⅰ. Introduction to the project

Our project is about the forecast of the house price by using specific function with python. We mainly focus on the real estate price of New York, as it is important for the financial market in US and it also has an impact on the economical condition of US as well.

New York is not only the capital of US but also the location for Wall Street, which is an important worldwide financial market. So, understand and predict the real estate price of it is very useful and important.

The process of our project mainly includes 4 parts, finding the dataset, processing the data, training model (machine learning) and the ETF in the real estate.

As for the result, we use a scatter showing the relationship between predicted sale price and real sale price.

# Ⅱ. Challenges and Solutions

We have faced several challenges during the completion of the project and we gave solutions.

1. Optimizing the data we want.

There is too much data, some are invalid, some are missing, and some are not suitable for the model.

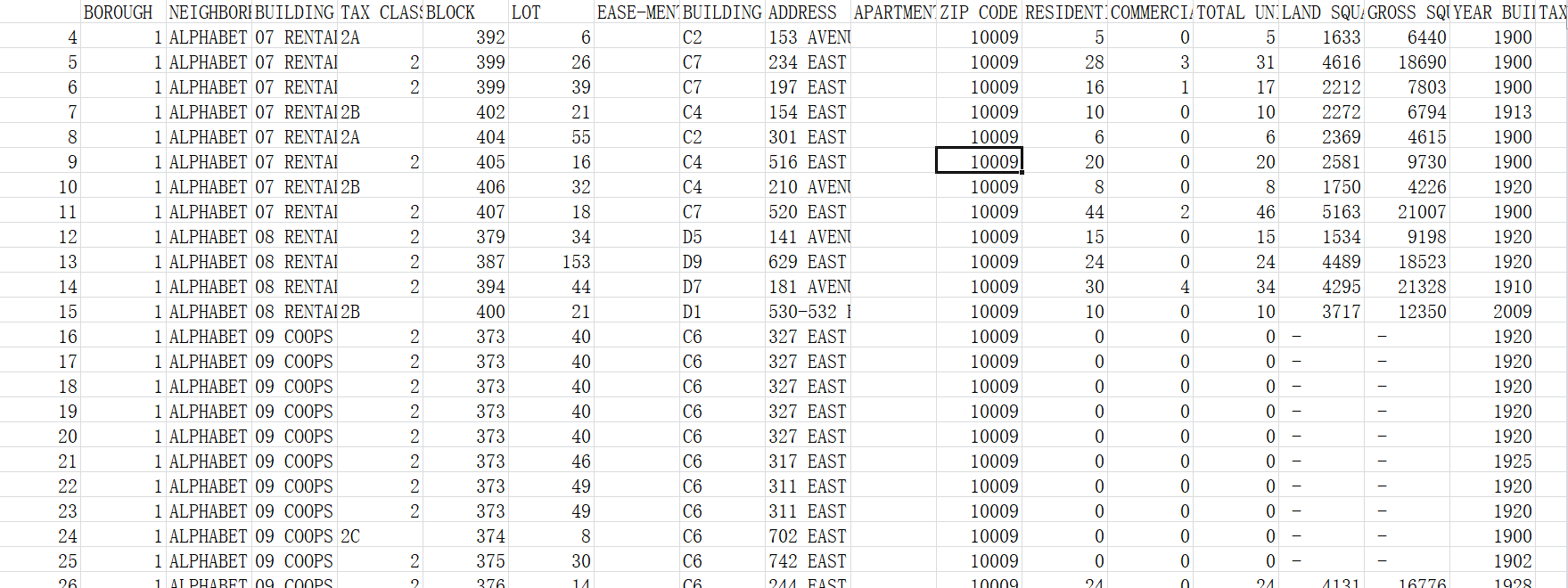
For optimizing data, we classified the data with common sense for the first step, and then we eliminate missing values and outliers. Finally, we classify the sale price into 5 categories from low to high and every categories contains 2000 samples to avoid bias. But it failed eventually. (high price is too high)

2. Building Model.

There are many kinds of models, it’s not easy to find the correct model. So we made hypothesis and tried three kinds of models.

# Ⅲ. Data processing

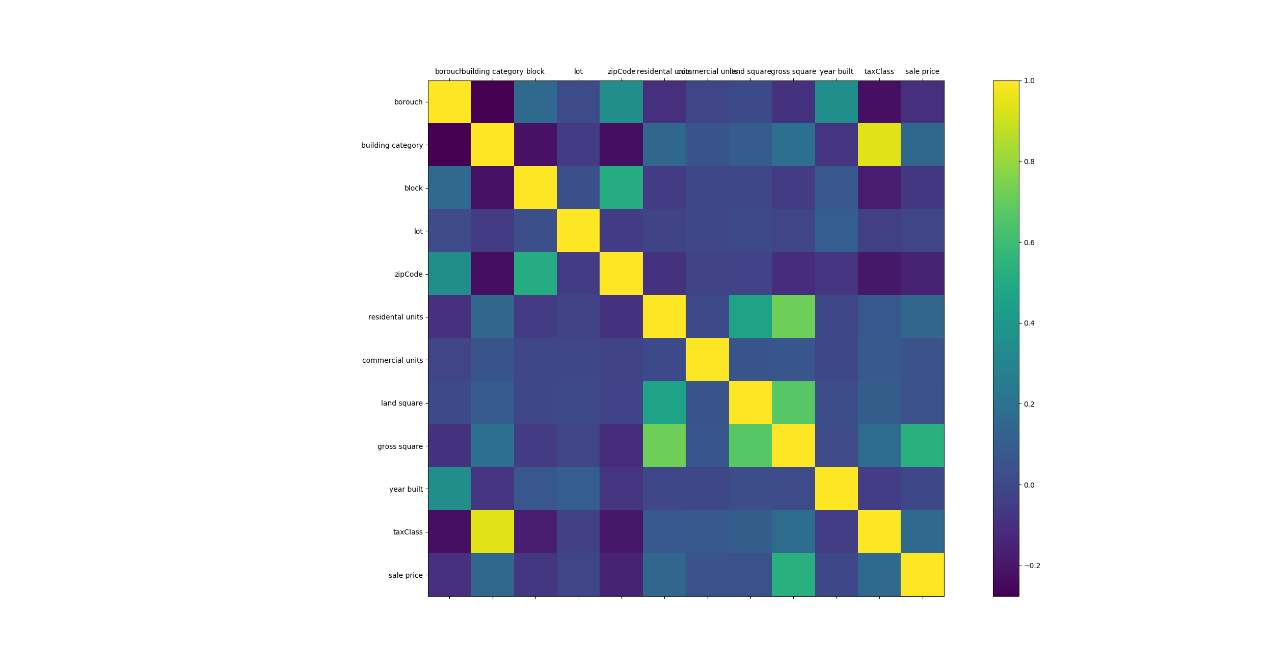
1. Downloading data from the Internet —— NYC department of finance.



The picture above is the data we downloaded from the Internet. We can see from the picture that the data cannot be used directly, since it contains some irrelevant factors, invalid values and missing values.

Firstly, we filtered some data based on common sense manually. We draw a tree map based on the data we collected form the Internet in order to find the main factors and the branch factors that affect the result. Since we are considering the general impact on real estate, we decided to eliminate nearly all the branch factors that are useless or too complex to handle such as the number of the residence. And we also deleted some data that are too relevant to each other that they have the very similar impact on the result such as the zip code and the address. Also, we divided New York into 7 zones regardless of its own division.

Secondly, use python to handle outliers. We put the data into python and use its specific function to handle the data. We build dictionaries and lists to prepare for the data processing, and we use this optimized data to do the advanced analysis such as the correlation matrix. We eliminate some data that have very low correlation to sale price as it is not relevant to the result and may add errors to the result. And also we deleted some data that are too relevant to each other since the features selected should be independent with each other, otherwise it may cause coherent problems.

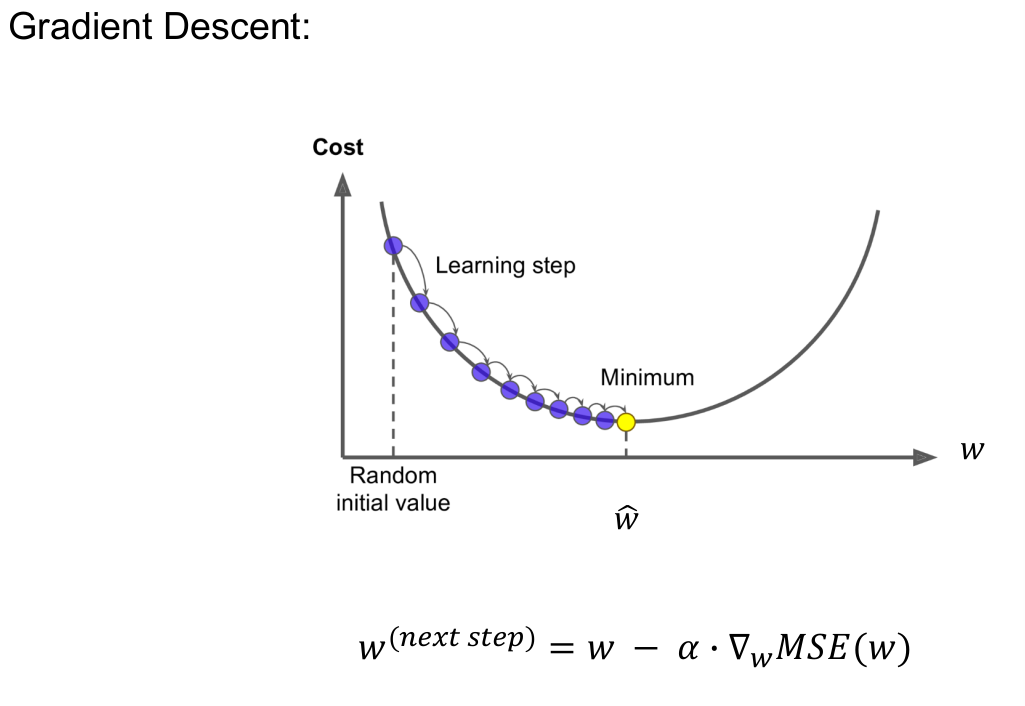


The picture above is the procedure of filtering the data. The matrix displays the correlation between factors ranged from -0.2 to 1.0. The yellow plot represents the correlation of 1.0, the darker the colors are, the less related between the two factors. Using this matrix, we find that the ‘tax class’ and the ‘building category’ are not independent.

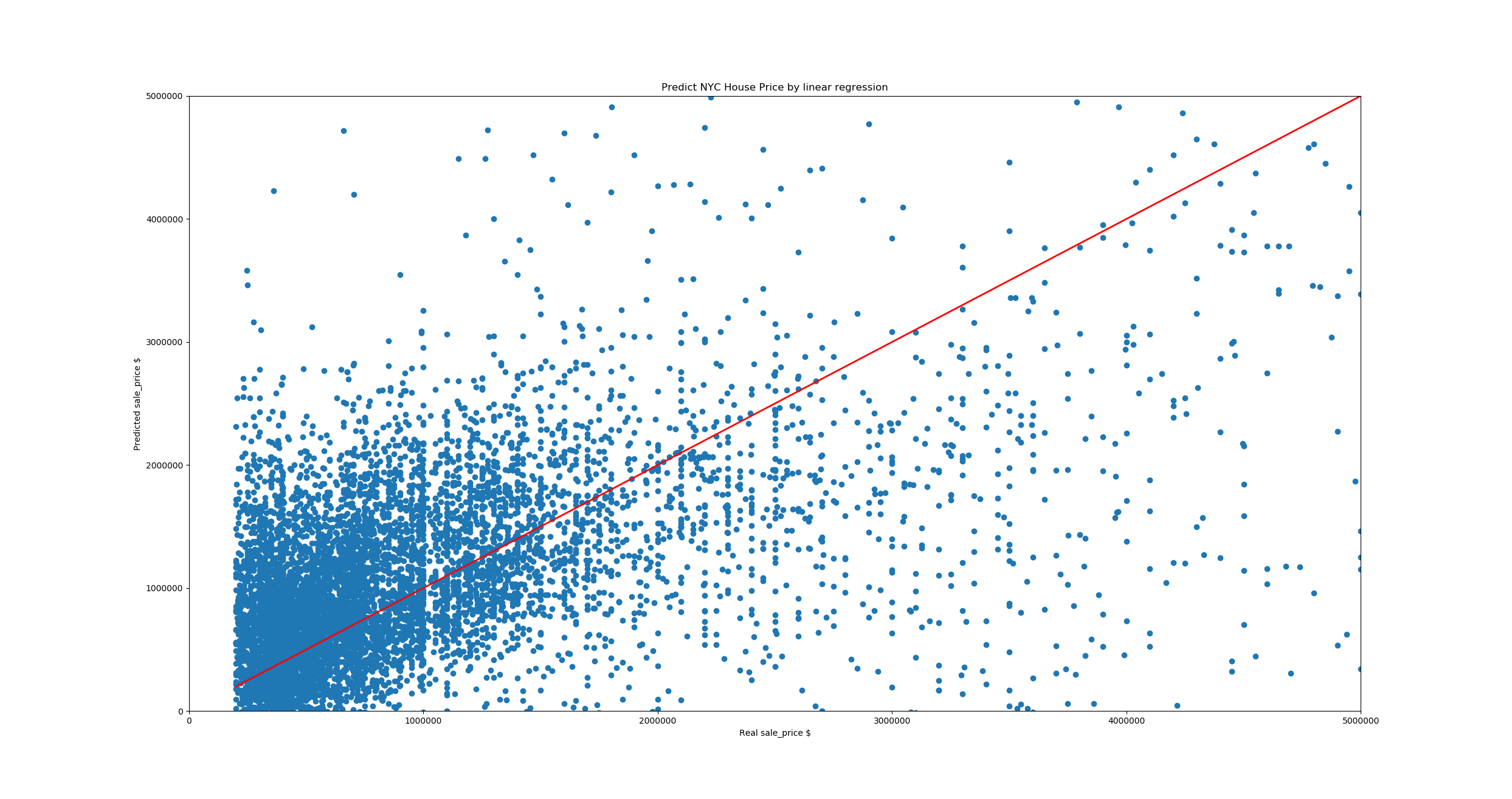
As the result, we dropped 'NEIGHBORHOOD', 'TAX CLASS AT PRESENT', 'BUILDING CLASS AT PRESENT', 'EASE-MENT', 'ADDRESS', 'APARTMENT NUMBER', 'TOTAL UNITS', 'SALE DATE', 'BUILDING CLASS AT TIME OF SALE'.

# Ⅳ. Algorithm, results and Data Visualization

1. Linear Regression



The picture below is the result of the linear regression. It shows relationship between predicted sale price and real sale price

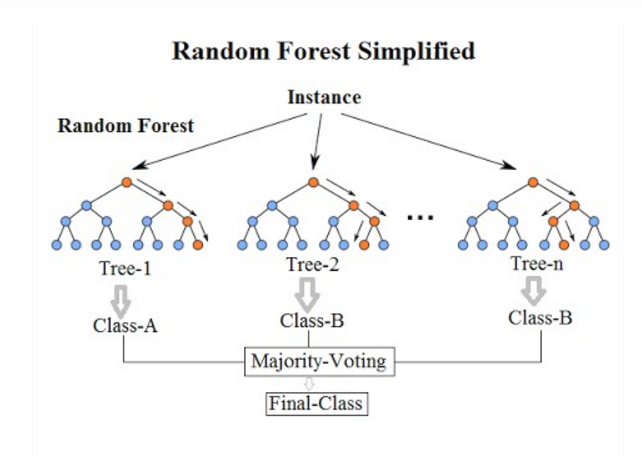


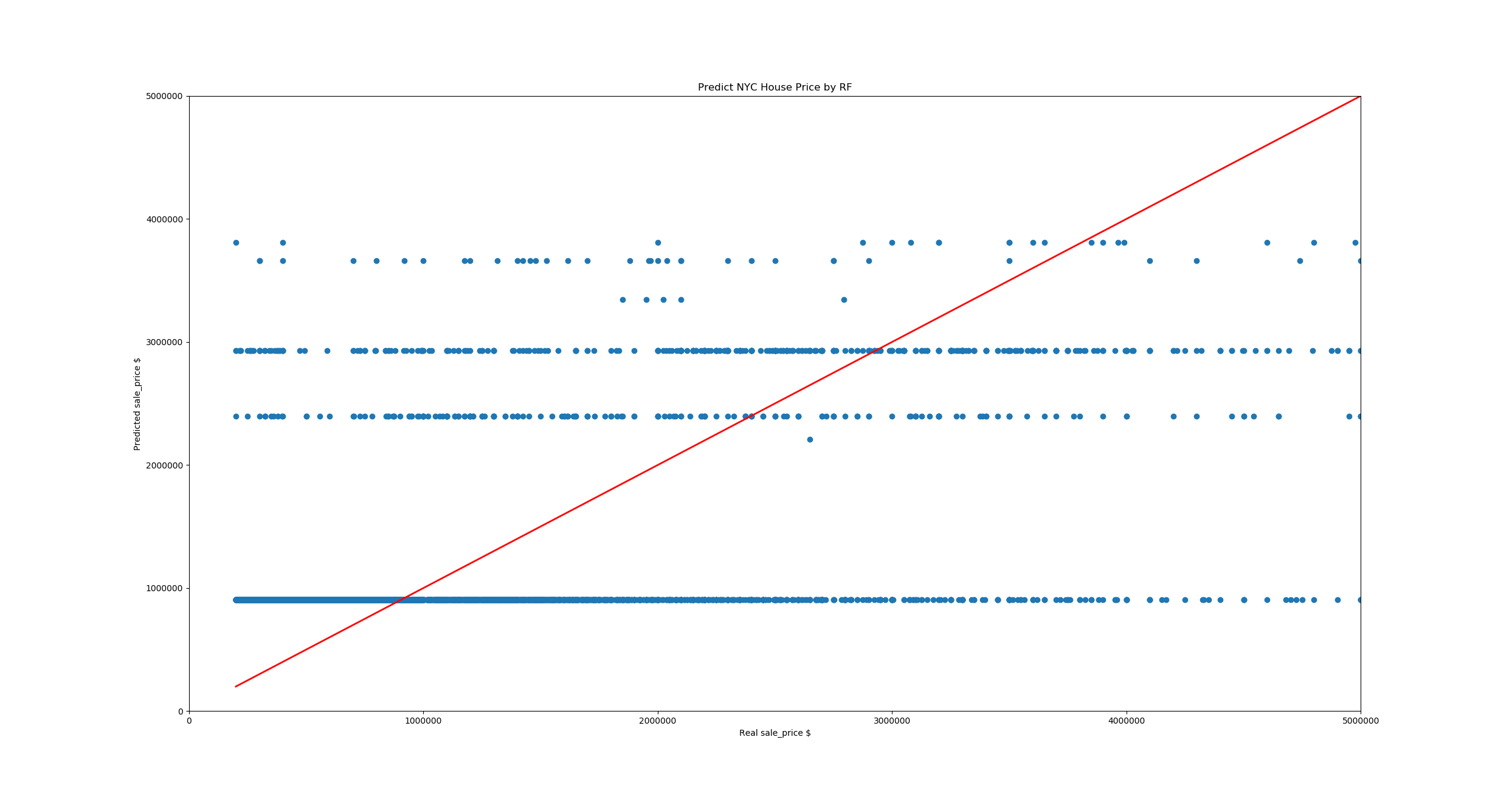
The score 0.6714874629993289 (R^2)

1. Random Forest Regression

Random Forest Regression is an ensemble learning that averages the outputs of many decision trees as output.

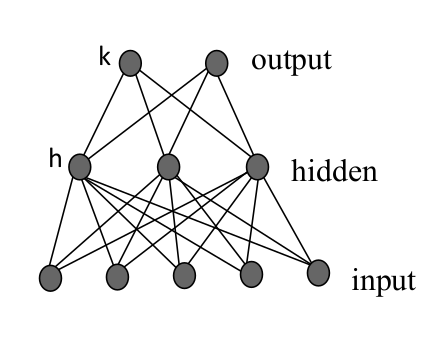
For every decision tree, classify every attribute into several categories. Then for the same category, average the target value as output.

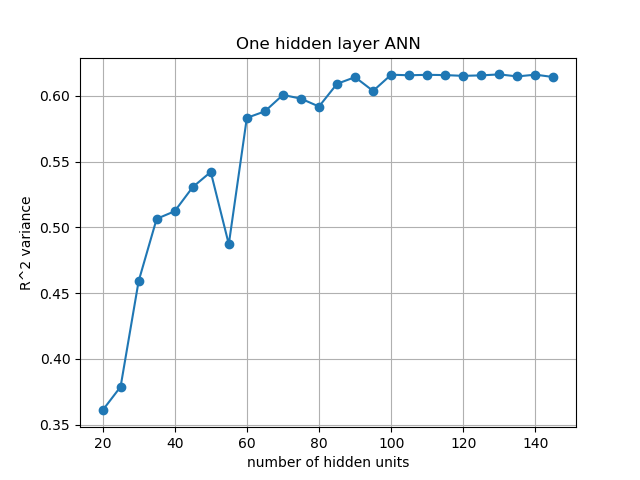




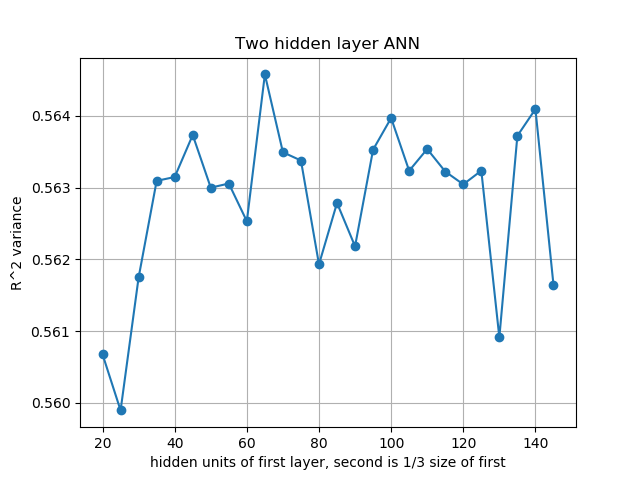
The score is 0.868 (R^2)

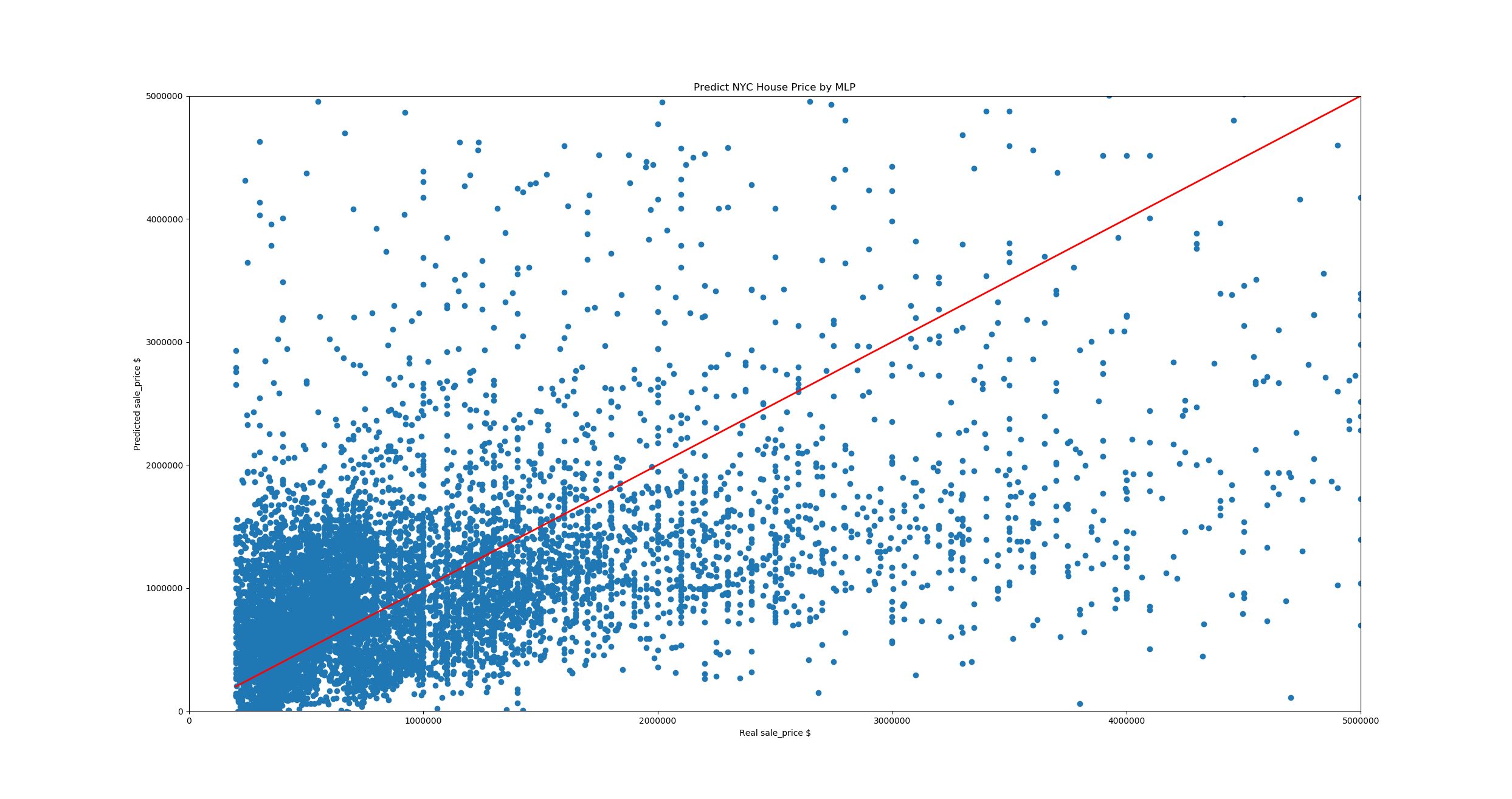
1. Multi-layer Perceptron





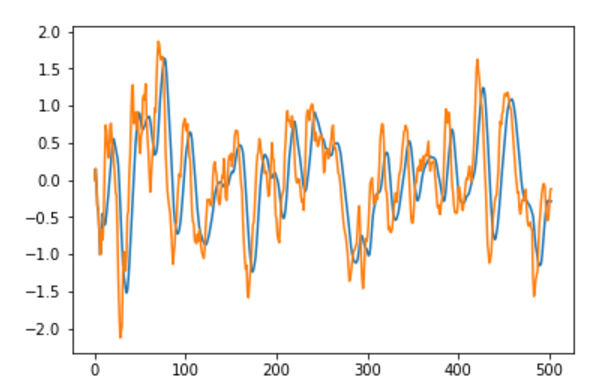
For one hidden layer MLP, we choose “hidden units = 100 “ as the best parameter.

Two hidden layers MPL has worse performance than one hidden layer. So we quit it.

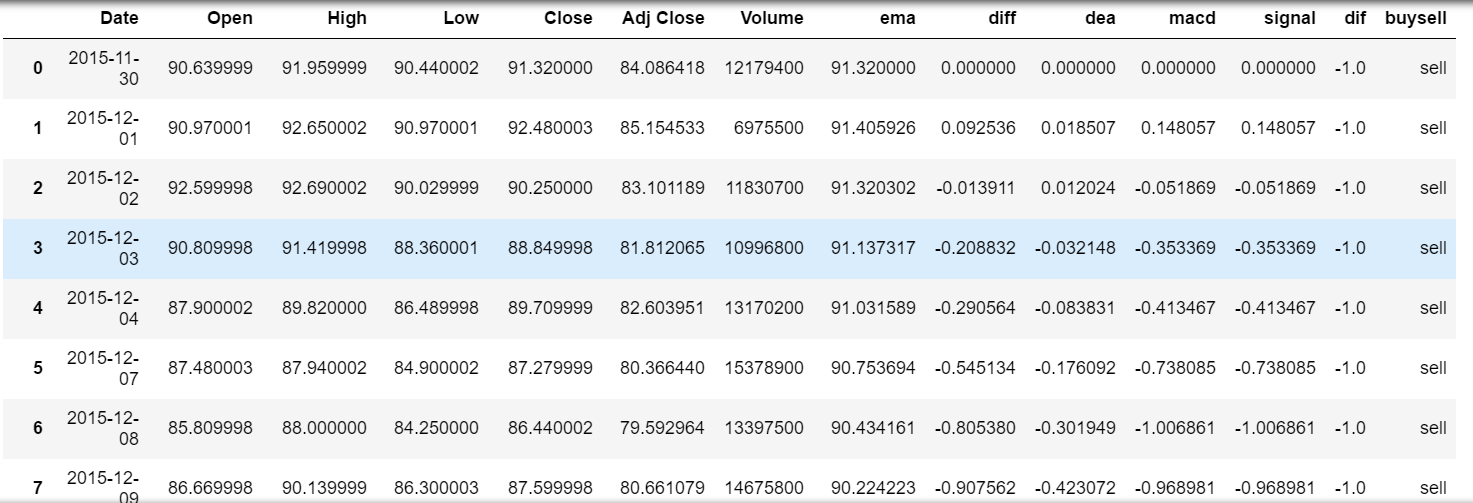


The score: is 0.625 (R^2)

# Ⅴ. ETF



The orange line in the picture is signal line, and the blue line is MACD. In this plot, when this two-line cross each other, it will give a buy or sell signal.



The real estate ETF is based on the real estate stock, which is some kind correlation with the house price. The buy and sell signal model are based on the moving average theory. As you can see when this two-line cross each other, here is buy or sell signal.so after we predict the house price, we can get a buy and sell signal in the market.

But the buy and sell signal is just a model at present. If we want to use that, we need to combine the etf stock price with the forecast price.