通常課題１

Dingming Xue, The University of Tokyo, Graduate School of Engineering, Department of System Innovation

氏名：薛丁銘

所属：東京大学　工学系研究科　システム創成学専攻

学籍番号：37-237256

1. Introduction

The common topic 1 tackles the problem of price predicting of Tokyo real estate.

Firstly, I am required to imagine that I am intending to purchase some real estate in the first quarter of 2023. From the official website of Land General Information System the detailed history data of land transaction happened in Tokyo can be acquired. These data could be used for model training and by introducing the corresponding features in test dataset, the prices in the first quarter could be predicted. Except for simply predicting the prices, another task is to predict with two different training datasets, which are the most recent 50000 pieces of transaction data (referred as recent data) as well as randomly selected 50000 pieces of transaction data (referred as random data) and compare their training accuracy.

In this topic, I introduced several models for price prediction including 3 different linear regression models, random forest and gradient boosting. For the random forest and gradient boosting, I also tried hyperparameter learning for better hyperparameter optimization. As for the conclusion, Extreme Gradient Boosting model shows the highest quality of price prediction and linear regression model preformed worst. Besides, compared to the result of price prediction in the whole Japan, the importance of features differs significantly and the performance decreased in every models. More importantly, prediction accuracy of recent data is better than random data for the case of most models, indicating that for the sequence data like price is quite time sensitive and it shares high possibility of inheriting features from nearby time periods. In addition, I streamlined the code by extracting common parts and abstracting them into functions, making the code more concise and efficient. Through the coding I gained improvements in my coding ability. Large langue model(LLM) such as ChatGPT was also used in this work mostly for model selection, reference code searching and slightly for report proofreading.

1. Problem Analysis

The common topic 1 appears to be a simple prediction problem with a sub mission of using two different training datasets for training and comparing the performance on the same test dataset. I divide the whole work into two parts: data preprocessing and model training.

In data preprocessing period, the first step is to download the history transaction datasets and divide it into the training and test datasets. There should be data cleaning for handling missing/incorrect data, data transformation for converting data into suitable format such as data of time period and room size as well as converting the input dataframe into suitable format such as two-dimensional matrix as machine learning models input, data reduction for reducing the unnecessary features.

In the model training period, I applied different linear regression models, Random Forest and Extreme Gradient Boosting for training. For Random Forest and XGBoosting model, hyperparameter learning was used for better parameter optimizations. R-squared score and MSE are mainly used for evaluating the accuracy. Considering that the features are mainly sparse matrix, lasso linear regression model was introduced expecting for better results.

1. Methodologies and Experiments
   1. Data Preprocessing

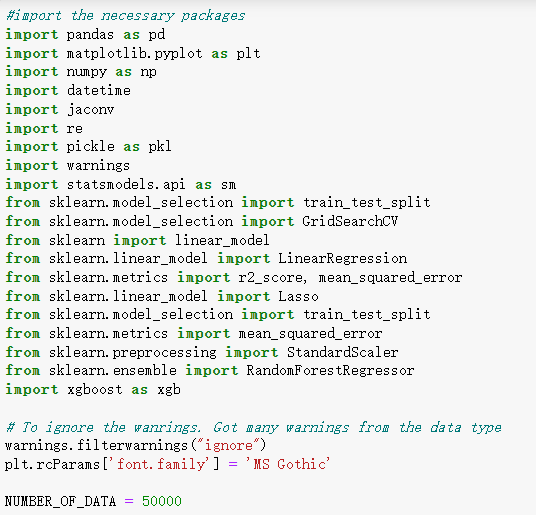
 This paragraph discusses the data preprocessing period of the work. The data preprocessing period includes data download, data cleaning, data reduction and data transformation. The necessary packages were imported and the number of data for training is set as a global variable as shown in Figure 3.1.1.

Figure 3.1.1 Package Import and Global Variables

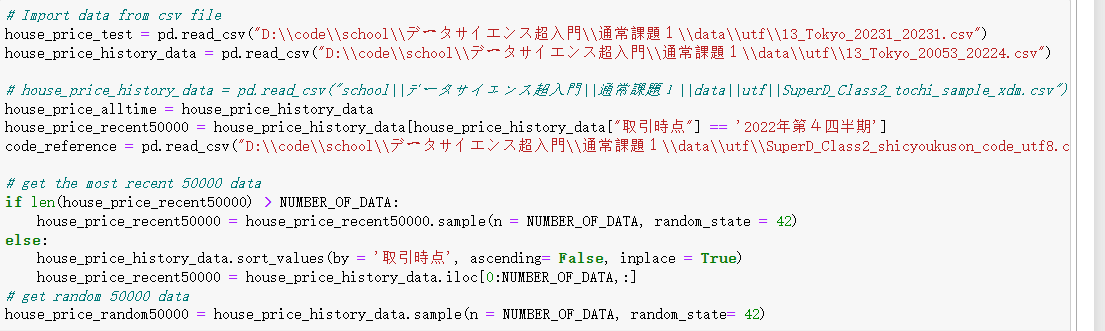
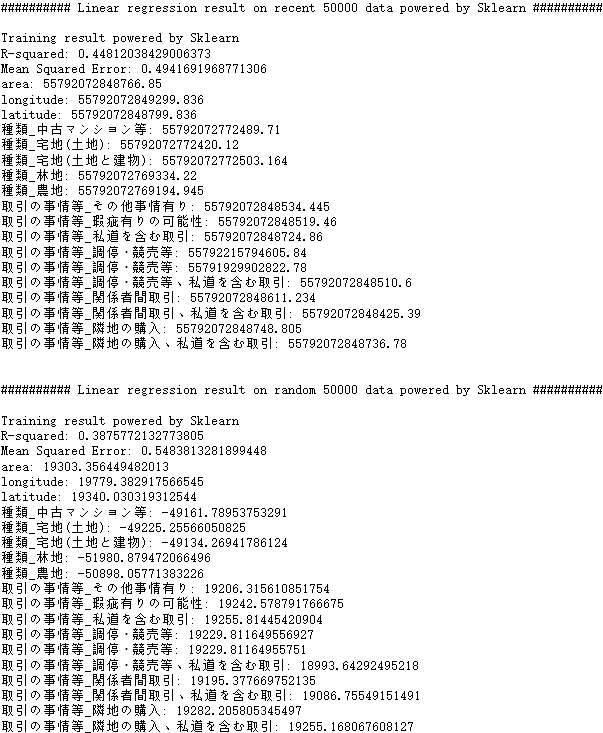
All the existing transaction data before the first quarter of 2023 was downloaded to the local but I encountered the problem of encoding error. The encoding of the csv file was JIS and I used Notepad++ to transfer the csv files into UTF-8 format. I used sample method of Dataframe to randomly select 50000 pieces of data from the history transaction data. As for the most recent data, I firstly checked whether the length of the data in the last quarter of 2022 is more than 50000 as shown in Figure 3.1.2. If so, I would randomly select 50000 pieces of data from it. If not, I would sort the history transaction data by transaction time and pick the most recent 50000 pieces.

Figure 3.1.2 Data Sampling

I created several methods for data transformation and got the input matrix I wanted. However, after the dummy convert, the dataframe was filled was Boolean data of true and false. The machine learning model could not recognize the Boolean data. In that case it was necessary to transform the Boolean data into 0 and 1 as shown in Figure 3.1.3(a) and Figure 3.1.3(b).

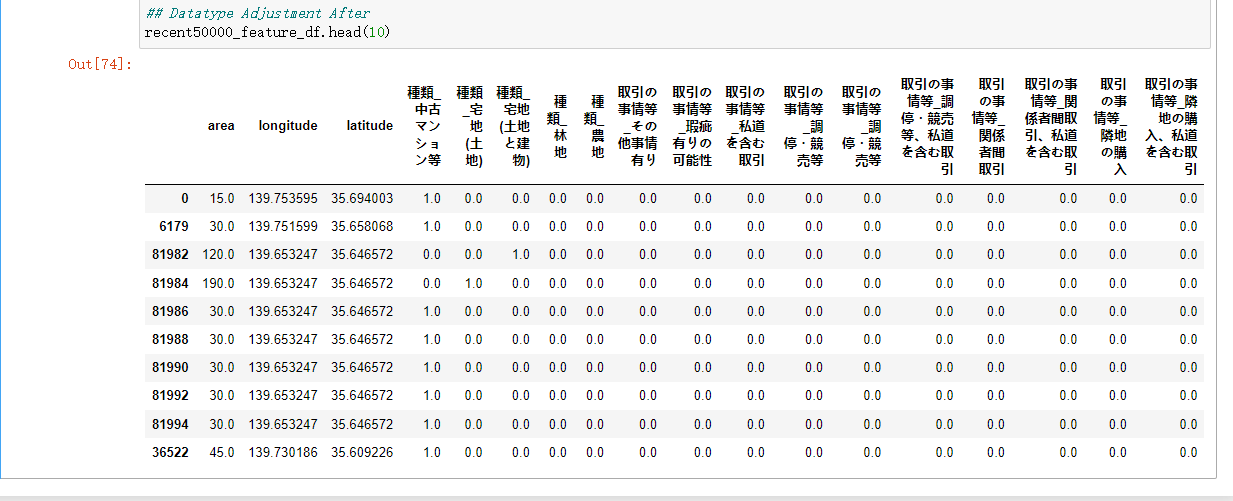
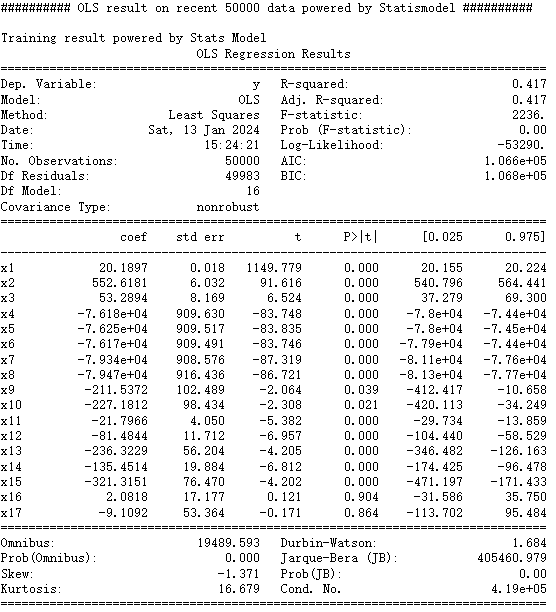
Figure 3.1.3(a) Dataframe Before Adjustment

Figure 3.1.3(b) Dataframe After Adjustment

Up till now, the feature and label dataframe of both training datasets and test datasets are prepared.

* 1. Model Selection

In this work, several models were selected and deployed for acquiring prediction results. The OLS model of Statsmodels package and LinearRegression model of Sklearn package were used firstly for prediction. However, the outcome showed very low performance on the test dataset. The average R-squared score was around 0.4. Considering that the features were sparse because of creating the dummy features, I introduced Lasso Linear Regression model hoping to solve the problem. However, the result still did not work out well. Therefore, random forest and XGBoosting was introduced and acquired relatively good results.

1. Training Results
   1. Linear Regression Model

For the prediction result of Linear Regression models, I meanly focus on the R-squared score and the MSE score. The prediction result of Linear Regression model of Sklearn is as shown in Figure 4.1.1. The result shows that using recent data as training data has better performance on the predicting result since the R-square score of recent data is larger an the MSE score is smaller. However, the coefficients indicates that the regression result is not similar between using recent data and random data.

Figure 4.1.1 Training Results of Sklearn Linear Regression Model

Then the Linear Regression model of Statsmodels Package was introduced to compared the results as shown in Figure 4.1.2(a) and Figure 4.1.2(b). In this result we can see that the random data has out-performed the recent data regarding to it’s higher R-squared score and lower MSE score. This no doubts leads to the uncertainty of the model.

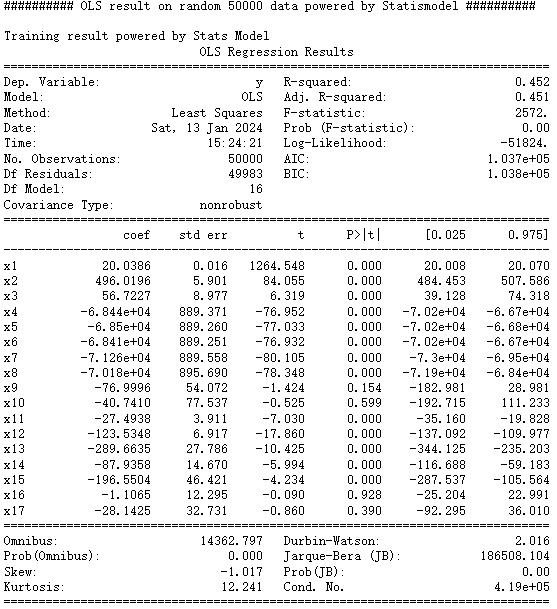
Figure 4.1.2(a) Result of Recent Data by OLS Model

Figure 4.1.2(b) Result of Random Data by OLS Model

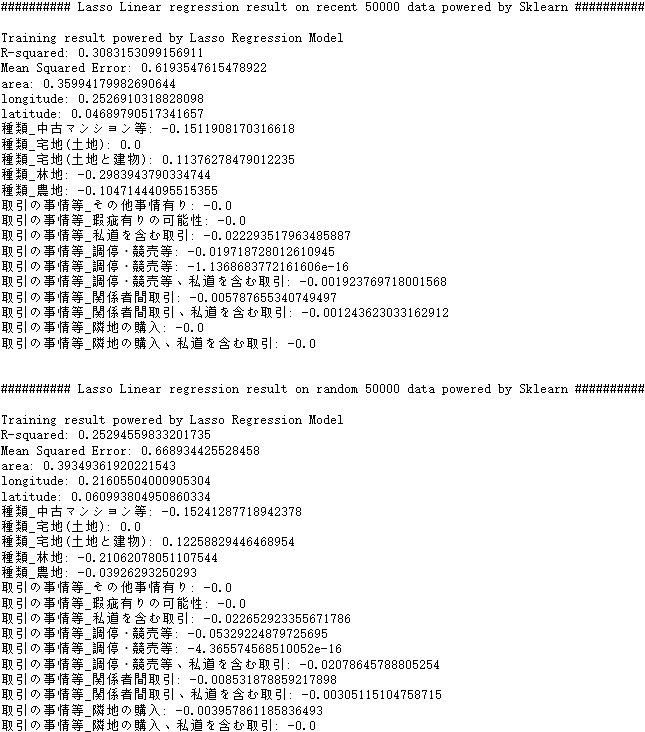
Then the Lasso linear regression model was introduced to seek for a balance. However, the result was not good as shown in Figure 4.1.3. The low quality of regression makes the result lack of reliability.

Figure 4.1.3 Training Results of Lasso Linear Regression Model

In this case, there’s not necessary to continue to experiment on the linear regression models. Then Random Forest Model and XGBoosting were used as shown in the next paragraph.

* 1. Random Forest Model

Random Forest Model is expected to acquire better results and indeed the results were better than Linear regression. The result is as shown in Figure 4.2.

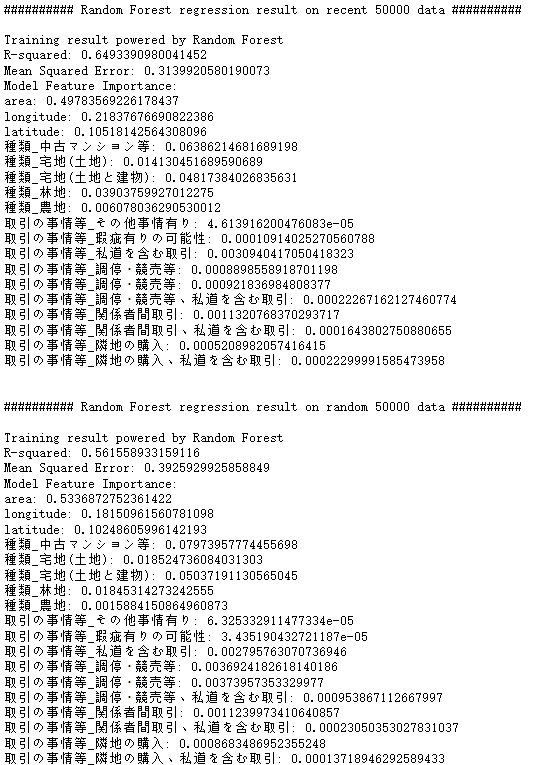
The parameters were selected by the hyperparameter learning for acquiring better results. From the importance map we can see that area, longitude, latitude and categories are the key features which effects the price. The R-square score and MSE score are improved

Figure 4.2 Training Results of Random Forest Model

by 20%. Considering that house price prediction is a very complicated task, the R-square score of 0.649 is good as the result. Meanwhile, the comparison shows that with a good prediction model recent data is better for house price prediction.

* 1. XGBoosting Model

XGboosting can be considered as an update of Random Forest Model. The training result of XGboosting model is as shown in Figure 4.3.1.

XGboosting model has a slight improvement upon Random Forest Model. The result also indicates that using recent data for prediction is better. However, as for the importance map, the result was completely different from the Random Forest Model. In this case, the category has become the most import feature as shown in Figure 4.3.2.

1. Conclusion

In this topic several models are used for real estate prediction such as linear regression model, random forest and xgboost. For most of the models, using recent data as training dataset has better prediction accuracy. For the comparison between models, xgboost is the best and linear regression models are the worst. Data normalization is required to improve the prediction result but it depends on the model. The prediction is not precise because for the same model (such as linear regression), using different can lead to different coefficients of the features. There’s another important finding that by increasing the size of training data, the result could become worse. This topic enhances my ability of manipulating the models and improved my ability of coding. I also put some efforts in making the code easy to read. ChatGPT played an important role in this work because I used it for supporting my coding. It greatly improved my speed of coding.

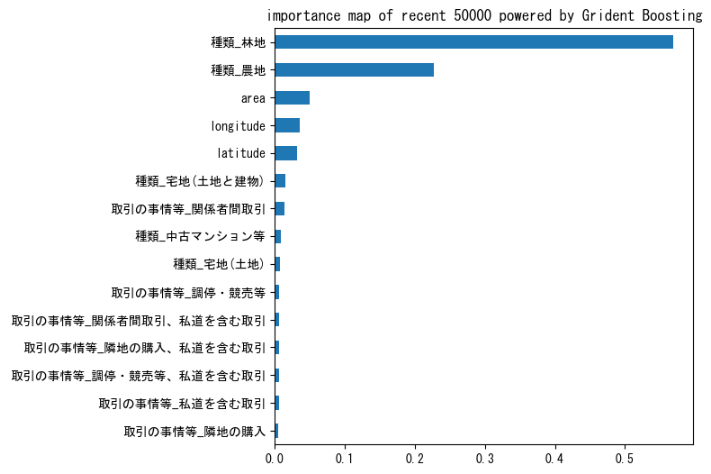
Figure 4.3.1 Training Results of XGBoosting Model

Figure 4.3.2 Importance map of XGBoosting Model