EIE3124 Report

Member：

JI ZIAO 21106839D

Li Hongjin 21097395D

MA YIming 21096467D

ZHU Houpu 20076843D

Wang Xinyang 20076348D

## Introduction

In this project, we are going to solve the problem of forecasting the price of Boston house. In order to achieve this purpose. We set the factors that may affect the housing price as independent variables, hoping to establish a housing price prediction model based on independent variables by analyzing the relationship between the changes of independent variables and corresponding housing prices. We need to do data processing and model design, in this process.

For data processing: There are four parts to data processing: Data import, data visualization, data normalization, data set partitioning. The purpose is to enable data to be invoked by the model, to work with special data in the data set that might negatively impact the performance of the model, and to help us understand the data in an intuitive way.

For model design: The model design consists of three parts. Algorithm selection, model optimization and evaluation. We assume that an algorithm can be used to express a certain functional relationship between the independent variable and housing price, and then optimize each algorithm through training, and compare the predicted results of the model with the actual results.

This is what we called Machine Learning doing and we will use parametric methods.

## Define the issue

### A clear problem formulation

Data set introduction: The Boston house price forecast data set contains 14 variables, including 13 characteristic variables and 1 target variable. Here is a breakdown of these variables:

'CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX','PIRATIO', 'B', 'LSTAT','PRICE'

These variables include a variety of information about houses in Boston, such as the location of the house, the physical characteristics of the house itself, and the characteristics of the neighborhood around the house.

Core Problem: Price is the target variable we want to predict. It can be used as an indicator of house price, which can help us understand the changing trend of house price and the expected price of a house under given characteristic conditions.

Sub-problems: We can also derive almost every characteristics a consumer or real estate agent or government would like to know about the Boston's houses with different estimation parameters to be evaluated in different dimensions.

## Methodology

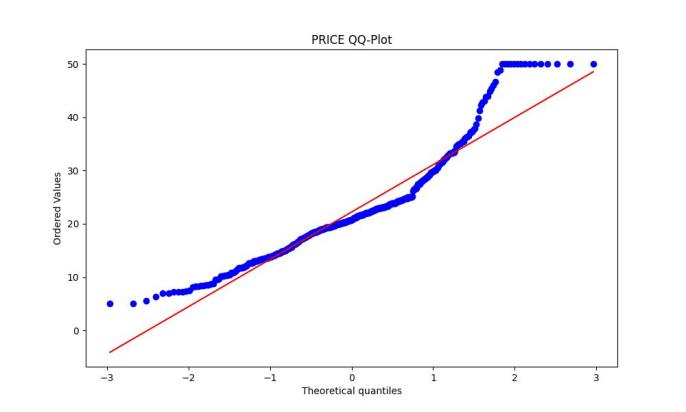
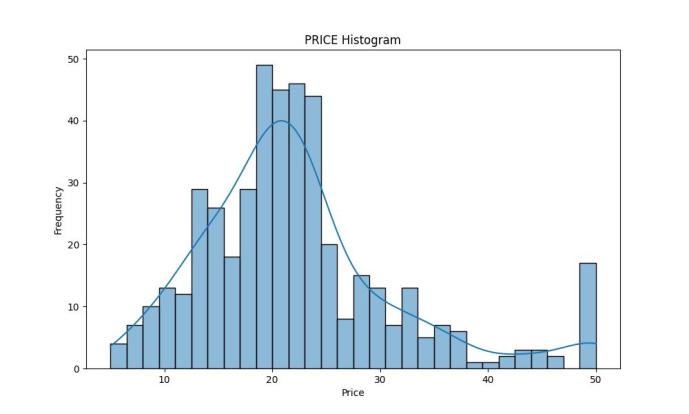
A detailed discussion of how the parametric method is used to solve the problem.

### 1. Visualization: including histograms, line plots and box plots of the distribution of all variables for outlier distribution observation

#### Data Understanding:

Visualization allows you to better visualize the data distribution, patterns, trends and relationships in the data set. This helps to have a deeper understanding of the data and thus better determine the appropriate modeling approach to the problem. Data distribution analysis picture is show in table1:

Table 1: Distribution histogram and QQ chart of the target variable PRICE

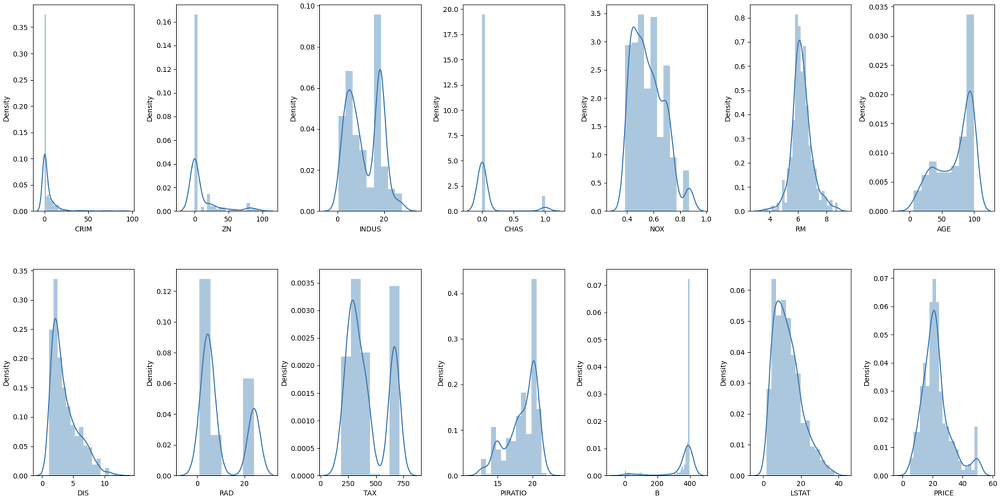


The histogram shows a slightly normal distribution of prices, but some deviation from the normal distribution at higher prices.

The QQ chart shows that the data points are mostly distributed along a diagonal, but there is some deviation in the high price area. This indicates that the data does not conform perfectly to a normal distribution in these regions.

To sum up, the dependent variable (PRICE) of the Boston housing price data is roughly close to the normal distribution, but there are some deviations in the high price region. So before we start modeling, we can consider appropriate transformations to the data to get it closer to a normal distribution.

Table 2: Feature Variable Visualization



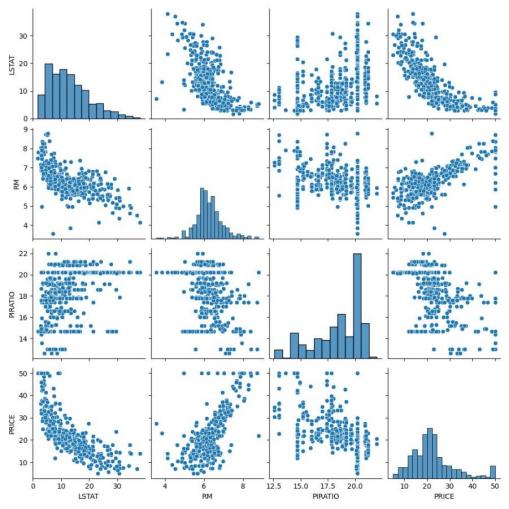
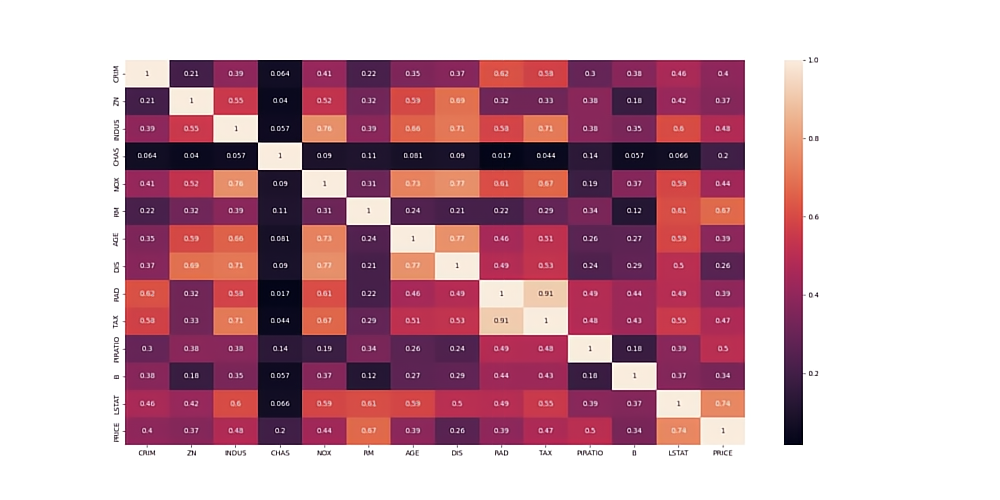
By analyzing the distribution of variables in the data set, we notice that they follow certain patterns. Using heat maps, we can gain a clearer insight into the relationship between these variables.

A significant bimodal distribution is observed in variables such as INDUS, RAD, TAX, CHAS, and PRICE. Although the bimodal heights of CHAS and PRICE were significantly different, the bimodal heights of other variables were relatively consistent.

At the same time, we also find that variables such as CRIM, ZN, NOX, DIS and LSTAT show left-skewed distribution, while variables such as AGE, PTRATIO and B show right-skewed distribution.

#### Correlation analysis:

Table3: Feature Correlation Heat map Table3: Dual variable correlation analysis



Thermal maps: Observe correlations between variables, useful in data mining and building feature engineering.

Correlation analysis: Visualization helps you discover relationships between independent and dependent variables. In this way, you can identify the important characteristics associated with the target variables, allowing you to make more informed decisions in the feature selection and feature engineering phases.

Proof of effectiveness of feature engineering. From the correlation matrix, we can see that TAX and RAD are highly correlated features. The [LSTAT, RM, PTRAIO] column has a correlation score above 0.5 with PRICE, which is a good indication for use as a predictor.

After conducting correlation analysis between the independent variable and PRICE through thermal map, we find that LSTAT, RM and PTRAIO are highly correlated with PRICE. Therefore, these three variables and PRICE are selected for analysis.

Finally, I came to a certain conclusion:

1. The lower the CRIM, the higher the housing price.

2. There is a low correlation between the proportion of residential lots (ZN) with more than 25,000 sq ft of floor space and house prices

3. The proportion of non-retail business (INDUS) in each town is higher and the housing price is lower. However, there are also lower prices in INDUS's lower locations.

4. There are high and low housing prices near the Charles River dummy variable (CHAS) basin.

5. There is no obvious correlation between nitric oxide concentration (NOX) and housing price.

6. The average number of rooms (RM) per house is positively correlated with the housing price. The more rooms, the higher the housing price.

7.The proportion of owner-occupied units (AGE) built before 1940 has little correlation with house prices

8. The higher the weighted distance between the five job centers (DIS) in Boston, the higher the house price, but the DIS of most houses is below 4.

9. Radial highway accessibility index (RAD) has no correlation with housing price

10. The full property TAX rate per $10,000 is under 400, and home prices can be high or low.

11. The ratio of students to teachers in cities and towns is higher, and there are more low-price housing.

12.1000 (Bk -0.63) ^2 Where Bk is the proportion of black people in urban areas (B), most of them are around 400, and housing prices are concentrated in the range of 20,000 to 40,000 dollars.

13. The ratio of people who are considered to be low-income (LSTAT) in a region is negatively correlated with housing prices. The higher the LSTAT, the lower the housing prices.

### 2. Feature engineering

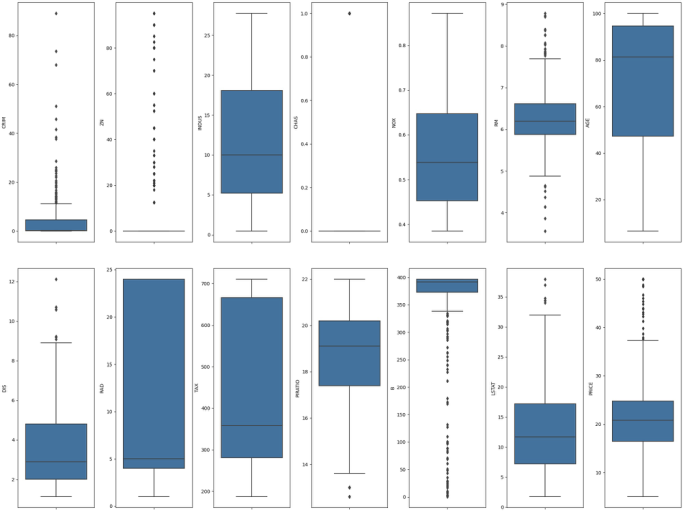
#### 1.Outlier detection:

Outlier detection in regression analysis: It can be used to judge the maximum value, minimum value, upper quartile, median and lower quartile in a set of data, by identifying outliers or outliers in the data, and removing these outliers.

Columns like CRIM, ZN, RM, B seems to have outliers. Let's see the outliers percentage in every column.

Outliers have both positive and negative effects (usually side effects) on regression analysis:

In regression analysis, outliers are generally considered data points that have a negative impact on the results. They can lead to overfitting of models, which can lead to a decline in the predictive power of models.



After experimental elimination, our data set changed from (455, 14) to (437, 14).

Outliers are usually extreme values that are far away from other data points. They can be caused by measurement errors, data entry errors, or other anomalies. An outlier is also a value that differs significantly from other data points, but is usually caused by a system error or data corruption.

In most cases, outliers and outliers have a negative effect on regression analysis. They can lead to overfitting of the model, which reduces the predictive power of the model. However, in some specific problems, outliers and outliers may have a positive effect and can be treated as special cases.

#### 2.Feature scaling

Several methods of feature scaling:

(1) min-max normalization: It uses normalization of values to the range [0, 1]

4LPD2BAAXY

(2) mean normalization: It scales the value range to the interval [-1, 1], and the average value of data becomes 0

DTPT2BAASM

(3) (standardization/z-score normalization) : Scale the value to near 0, and the distribution of the data becomes a standard normal distribution with the mean of 0 and the standard deviation of 1 (subtract the mean to centralize the features, then divide by the standard deviation for scaling).

ITPT2BAA3M

1. max abs normalization: It is changed to the scaling to unit length, and the range of values is scaled to the range [-1, 1]

#### Feature Scaling

Normalizing or standardizing data can improve model accuracy by treating each feature as equally important, especially when a feature has a larger range of values than others. This enables comparison of different feature characteristics.

Normalization or standardization can also increase the convergence rate of the gradient descent method when solving optimization problems.

Models solved by gradient descent, such as Linear Regression, Logistic Regression, Perceptron, support vector machines (SVM), and Neural Networks, require feature scaling.

#### 3.Construction feature

Construction methods include: interactive feature, scaling feature, polynomial feature

To build a new feature, we need to do this after scaling the feature. Because feature scaling can make the values between different features comparable, so as to better feature engineering.

Since the data dimension of the housing price forecast data set is low, and the potential collinearity of the dimensional information of each variable is small, we mainly choose interaction features and scaling features to construct.

Three main interaction features of construction:

X\_scaled['LSTAT\_RM'] = X\_scaled['LSTAT'] \* X\_scaled['RM']

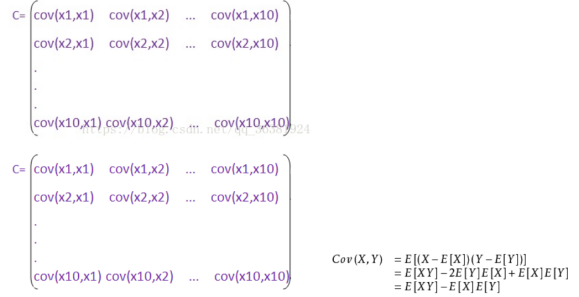
X\_scaled['TAX\_PIRATIO'] = X\_scaled['TAX'] \* X\_scaled['PIRATIO']

X\_scaled['TAXRM'] = X\_scaled['TAX'] / X\_scaled['RM']

And scale the remaining variables using log(), sqrt() methods.

#### Dimension reduction

Dimensionality reduction techniques such as PCA and correlation coefficient can help improve model training by reducing overfitting, improving efficiency, and eliminating multicollinearity. However, this can result in some information loss, which can impact model performance. Feature selection methods like variance threshold and mutual information can also be used. For Boston Housing Price prediction with only 13 features, dimensionality reduction may not be necessary, but it can be explored using PCA or t-SNE to evaluate impact on performance. Note that some models, like decision trees and random forests, can handle high-dimensional data without reduction, while others like linear regression and support vector machines may benefit from it, especially with large feature sets.



### 3.Construct regression model

In this project, we have used seven different models to predict house prices: Linear Regression, Ridge Regression, Poly Ridge Regression, SVR, Decision Tree Regression, KNN Regression, and GBR. Each of these models has been evaluated using the K-Fold Cross Validation method and the Box Plot Evaluation technique, with the evaluation metric being MSE. This part will provide a detailed introduction to each of these seven models.

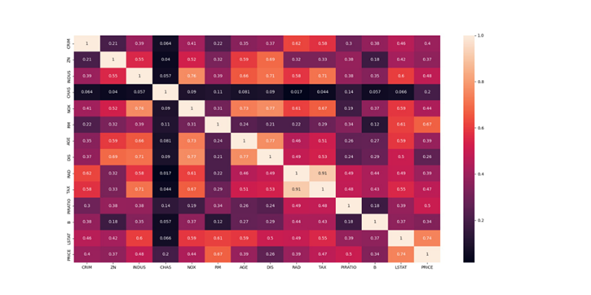
#### 1.Model introduction

**Linear regression**

Firstly, by visualizing the data, we can conclude that the data is continuous. And corresponding data has been provided for the average values and other characteristics of these related factors.

3YMT6BAAK4

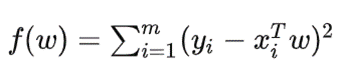
And we conducted pairwise analysis of fourteen parameters through heat maps, and found that there was a strong correlation between these parameters.



After data processing, factors with low data correlation (<0.3) were removed to make the dataset more consistent with our assumed model.

**Ridge regression**

Ridge regression mainly applies a method similar to least squares to optimize data and find correlation coefficients.



Some correlation coefficients are small, and we are worried that these factors will lead to collinearity. Therefore, we added ridge regression to the review hypothesis and finally compared the fitting degree between the two models and the data.

At the same time, we standardized the data (data scaling) to control them at [0,1], which can reduce the corresponding data processing time.

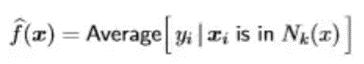
**Polynomial regression**

Similar to linear regression, turning linear regression into a curve this time may allow for more data fitting.

**SVR Support Vector Regression& K-neighbors regression**

SVR is a two-class classification model. The basic model is defined as the Linear classifier with the largest interval in the feature space. Its learning strategy is to maximize the interval. Perform data processing by finding corresponding interval lines.

K-neighbors regression is to calculate the K (usually 5 or 10) data points closest to a certain data point, and then predict this data point.



**Decision Tree**

When making predictions, a certain attribute value is used at the internal nodes of the tree to determine which branch node to enter based on the judgment results, until reaching the leaf node to obtain the classification result.

**Gradient Boosting Regression**

This estimator builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

#### 2.Univariate regression effect diagram

Model hypothesis testing: Before building a model, many models have certain assumptions. For example, linear regression models assume that there is a linear relationship between independent variables and dependent variables. It can be seen from the figure that the regression prediction effect of single variable is poor. These hypotheses are verified and valid, so other more complex models are adopted.

### 4. Model evaluation, comparison and optimization

Model evaluation and diagnosis: Evaluate the performance of a model by visually comparing the predicted results of the model with the actual results. In addition, you can use diagnostic graphs such as residual graphs to check that the model meets certain assumptions, such as independence of error terms, homogeneity of variances, and so on.

These evaluation and training methods directly determine the performance of our final model.

#### Model performance evaluation index

##### 1.MSE

Mean Squared Error (MSE) is a measure of regression model performance that calculates the average of the squared difference between predicted and actual values. It is calculated by dividing the sum of squared prediction errors by the number of samples. A lower MSE indicates higher prediction accuracy. MSE is sensitive to large errors, which are amplified by squaring, so it pays more attention to samples with large prediction errors during model optimization.

The formula for MSE is: MSE = 1/n times sigma (Pi-Yi)^2, where n is the number of samples, Pi is the predicted value of the ith sample, and Yi is the actual value for the ith sample.

##### 2.K-fold cross check

In general, K-fold cross-validation is used for model optimization to find the super parameter value that makes the model generalization performance optimal. After finding, the model is retrained on all training sets, and the independent test set is used to make the final evaluation of the model performance. K-fold cross-validation has the advantage of using the no-repeat sampling technique: each sample point has only one chance to be included in the training set or the test set during each iteration.

##### 3.Hyperparameter adjustment

GridSearchCV combines grid search and cross-validation to find the best model parameters. It works well for small datasets, but may be slower for larger datasets with multiple parameters. Grid search can be computationally expensive and may not find the global optimal value if the objective function is non-convex. A larger search range and step size can help locate the general vicinity of the global optimal value before narrowing the range for more accurate search, which can improve efficiency.

## Experiment Analysis:

### 1.The first experiment:

In experiment 1, we compare the effects of hyperparametric adjustment, feature engineering, and data outlier processing on a parametric regression model (linear regression) using the Boston house price dataset. The comparison will be made through the following steps:

1. Benchmark model: linear regression model (no adjustment of hyperparameters, no feature engineering and data outlier processing)

2. Model A: linear regression model + hyperparameter adjustment

3. Model B: Linear regression model + feature engineering

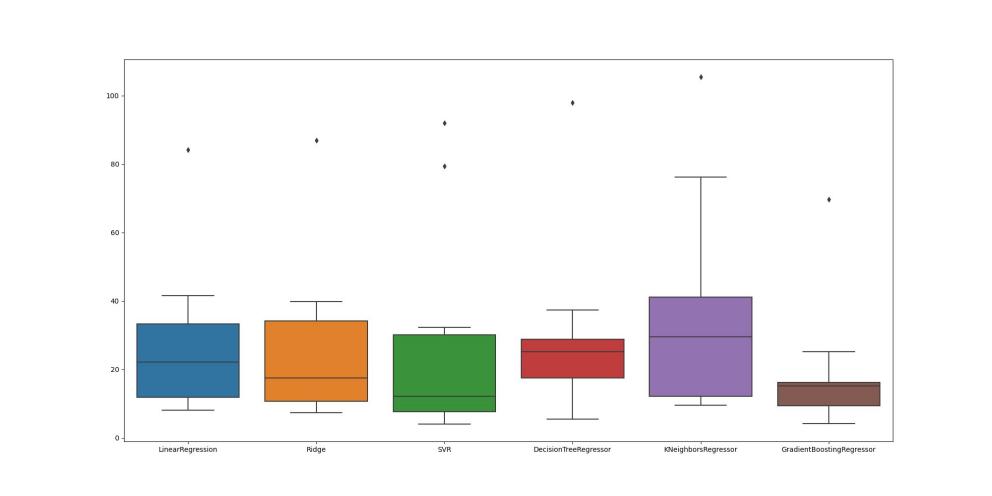
4. Model C: linear regression model + data outlier processing

5. Model D: linear regression model + hyperparameter adjustment + feature engineering + data outlier processing

|  |  |  |
| --- | --- | --- |
| ModelTrain | MSETest | MSE |
| Baseline | 21.6414 | 24.2911 |
| Model A(Hyperparameters) | 21.6428 | 24.3129 |
| Model B (Feature Eng.) | 5.1322 | 14.2936 |
| Model C (Outlier Removal) | 16.2341 | 30.2843 |
| Model D (Combined) | 16.2355 | 27.8873 |

### 2.The second experiment:

Try to use more models to solve this problem.



|  |  |  |
| --- | --- | --- |
| Model Name | MSE | Deviation |
| Linear Regression | 27.48 | (+/- 21.61) |
| Ridge | 26.35 | (+/- 22.87) |
| SVR | 27.63 | (+/- 30.34) |
| Decision Tree Regressor | 29.46 | (+/- 24.58) |
| KNN | 36.58 | (+/- 30.24) |
| Gradient Boosting Regressor | 19.06 | (+/- 17.78) |

Conclusion 1:

Through multiple attempts at data processing and optimization adjustments, we found that data processing has a significant impact on model performance. Gradient Boosted Regression (GBDR) is the most stable and efficient model. Parametric regression is not optimal due to limited learning ability and consistently high MSE. Support Vector Regression (SVR) with RBF kernel and K-Neighbors Regressor is sensitive to data changes. Decision tree regression has good stability but performs slightly below GBDR. The appropriate model should be chosen based on data characteristics and problem requirements. GBDR may be better for complex data, while other models such as parametric regression, SVR, KNN regression, and decision tree regression may have their own advantages. Flexibility in selecting and adjusting models is crucial for achieving the best performance in practical applications.

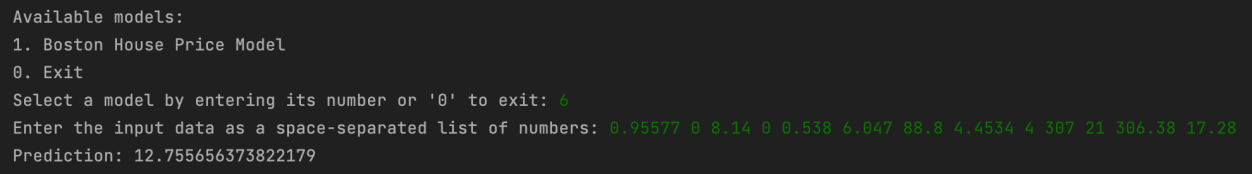
Conclusion 2:

Dimensionality reduction can benefit parametric regression models by reducing overfitting, improving efficiency, and eliminating multicollinearity. However, for integrated learning methods like gradient boosting, it may degrade model performance due to information loss during dimensionality reduction. For the Boston housing price prediction dataset with a small amount of data, dimensionality reduction is not applicable. If the dataset grows and the dimensions expand, the integrated model may perform better. We need to consider the advantages and disadvantages of dimensionality reduction based on dataset characteristics and model type to achieve the best performance in real-world problems. This can be an improvement point for future research.

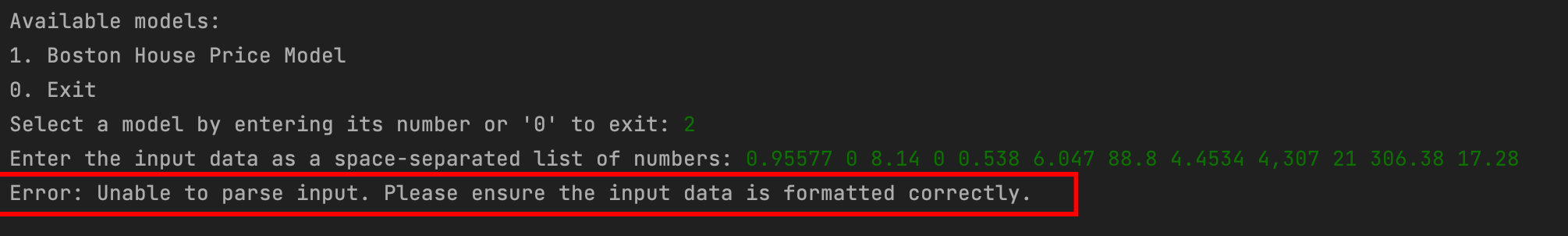
### 3.The third experiment：

We saved the model trained after data processing, feature engineering and hyperparameter adjustment, and wrote a simple terminal system. Use the terminal to call the trained model so that it can predict the input data.

The system mainly includes two parts. The function of the first part is to be able to select models 1-6, corresponding to linear regression, ridge regression, SVR support vector regression, decision tree, k-neighbors regressor, and GDBT algorithm models. If the input value is not the serial number of models 1-6, the system will identify an error and require re-input. After the model is selected, the system will prompt you to enter 13-dimensional data information separated by spaces. Similarly, when the information is entered correctly, the predicted house price result will be returned, and when the information is entered incorrectly, it will be required to re-enter. The operation process demonstration effect is shown in the figure below:：



At the same time, if an illegal input occurs, the error reason will be returned and the user will be required to re-enter the data.



## Conclusion

ML has a good performance on the regression problem of housing price prediction, and it is relatively easy to achieve some useful things.

Our contribution: combine the content learned in class with actual code analysis, and learn from practice.