Research on Housing Price Prediction Model Based on Machine Learning

Xing Qian
University of Oregon
xingq@uoregon.edu

Kai Xiong
University of Oregon

1. Abstract

With the rapid development of artificial intelligence technology, machine learning is widely used in all walks of life and has achieved good results in many fields. Housing prices are a hot issue with complex influencing factors, and it is difficult to make a comprehensive and accurate forecast. Therefore, this article attempts to apply relevant machine learning algorithms to housing price predictions, under a relatively stable condition, establish a model that meets the complex characteristics of housing prices, and analyze and predict housing prices. The main work of this paper is as follows:

1. Preprocessing the data: including missing value filling, visual exploration, and outlier processing
2. Feature engineering: including new feature screening, feature dummy code, and feature standardization
3. Model selection: training and verification of 5 models including the least square, and a reasonable evaluation of each model
4. Parameter tuning: Use grid search method to optimize the parameter combination of the model to find the best parameter combination model

Finally, through experiments, in the five regression models tried in this article, the performance results of the Random Forest Regression model finally, the R2 index can reach 0.8556, and its RMSE evaluation index value is 31574.2035, which has a higher RMSE value than other works of Kaggle. The expressive power of higher models.

2. Introduction

Machine learning is a multi-field interdisciplinary subject belonging to the science of artificial intelligence, which involves relevant knowledge of disciplines such as probability theory, optimization theory, statistics, and approximation theory. With the development of big data and computer hardware technology, machine learning has unprecedented opportunities, and the application of machine learning has become more and more extensive. For example, weather forecasting uses machine learning technology to analyze the collected weather data to predict the weather at a later date; companies use machine learning to analyze customer behavior and habits to obtain customer portraits and develop targeted marketing based on customer portraits Strategy: The web search engine uses machine learning technology to retrieve relevant content and rank the importance, giving priority to the content with high importance, so that users can quickly and accurately find the content they need. In addition, there are many aspects of machine learning technology in face recognition, environmental monitoring, recommendation systems, disease diagnosis, fault diagnosis, and automatic driving. Many facts show that machine learning is widely used in many fields such as medicine, finance, industry, and agriculture, and certain achievements have been made. Therefore, machine learning technology has important application value.

No matter in any period, housing is the most basic demand of people's life, and it is closely related to people's daily life. The rapidly developing real estate market has become a basic pillar industry that promotes economic growth and stimulates domestic demand and plays an important role in the development of the national economy. Therefore, the housing problem is not only an issue of people's livelihood, but also an economic issue, which is related to the stability of the country and society. Housing prices are the market value of real estate, which has a great impact on people's living standards and the development of the national economy. The research on housing prices has received key attention from many fields such as statistics, management, and computer science, and the forecast of housing prices has also become a research and discussion problem for many scholars. Complying with the development trend of big data and machine learning, combining network data, using machine learning algorithms to analyze and predict housing prices is more scientific.

From an overall point of view, the research on housing price prediction can be summarized into two categories. One type is qualitative valuation and forecasting of housing prices, and more inclined to economic analysis, mainly focusing on market information, and rarely using mathematical models. The other type focuses on quantitative analysis, using mathematical models to quantitatively predict housing prices. However, qualitative analysis is easily affected by subjective factors. Therefore, when analyzing and predicting housing prices, it is more scientific and reasonable to use quantitative analysis than qualitative methods. For the quantitative analysis of housing price forecasts, scholars at home and abroad generally have two ideas: First, the changes in housing prices are regarded as a time series to predict housing prices. The second is to analyze the influencing factors of housing prices and use the influencing factors to establish an index system to predict housing prices. There are two angles for predicting housing prices by constructing a predictive model based on influencing factors: one is to predict the average housing price from a macroeconomic perspective, using macro indicators such as GDP and loan interest rates. The other is to predict the housing price of a specific house from the perspective of the house itself, based on the characteristics of the house itself. This article is sampling the last method to predict house prices from the perspective of the house itself.

Establish a specific house price based on the characteristics of the house, such as the type of house, the year of construction, the area of the house, the decoration situation and other influencing factors. Rosen et al. introduced the Hedonic theory into housing price prediction for the first time, and proposed a residential characteristic model, which first studied the relationship between residential prices and the living environment. Hasan Selim analyzed the influencing factors of Turkish housing prices and used artificial neural network models and Hedonic models to predict housing prices. The comparison found that artificial neural networks are better than Hedonic models. Chen Shipeng built a random forest model based on Xiangyang mortgage data and compared it with the prediction results of the ARIMA model and the multiple linear regression model. Experiments proved that the random forest model has a better prediction effect. Chia-Chen Fan et al. established a network service system for housing price prediction and housing information sales. The system combines analysis methods and prediction models to predict housing prices. Stephen Law et al. used a combination of deep neural network models and housing features to estimate housing prices in London, England. Naalla Vineeth et al. established a housing price prediction model by using simple linear regression, multiple linear regression and neural networks in machine learning algorithms to help buyers and sellers find the best price for a house.

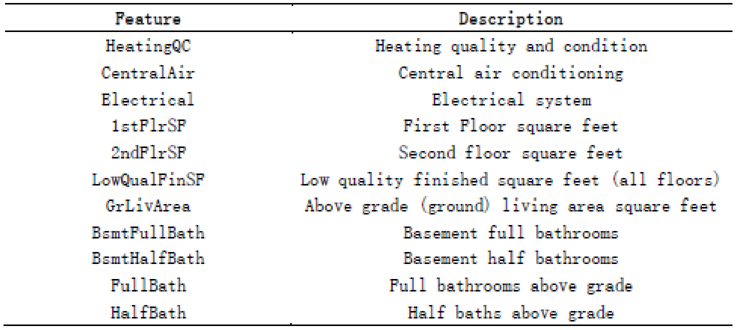
This article takes Kaggle's housing price data as a starting point, through the construction and screening of data characteristics, to find out the characteristics that are more relevant to the sale price, and then respectively perform least square regression, lasso regression, ridge regression, Progression and Random Forest Regression of integrated learning The model is trained and predicted, and it is concluded that the Random Forest Regression model has better performance than other models. Then, the model is tuned through the parameter grid search method to obtain the best model parameter combination and model result. Through the experimental link, the best R2 value of this regression model is 84.83%.

The rest of the paper is organized as follows: Section II is an overview of the data; Section III details our methodology for this work; Section IV covers our experiments and analyzes their results; and in Section V we draw our conclusions.

3. Data description

The data in this paper comes from the house price prediction competition in Kaggle. The data contains a total of 79 features. Table 1 shows the names of some features and their descriptions. Its characteristics mainly include several aspects. One is the characteristics of the house itself, such as the type of house, the year of construction, the area of the house, the type of decoration, etc., which involve all aspects related to the house, which are relatively comprehensive; the other is the spatial location of the house, such as The street where the house is located, whether the house is close to the main road, the type of street, the distance from the property, etc.

Table 1 part feature name and it's description



Feature	Description
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade

There are qualitative data such as features MSZoning, Street, Utilities, etc., which are all strings, and numerical data such as features MSSubclass, LotFrontage, and OverallQual. At the same time, there are also many differences in the scope of numerical data in the data. For example, the maximum and minimum Lotfrontage values of the feature are 313 and 21 respectively, while the maximum and minimum value of the feature Grlivarea are 5642 and 334 respectively. In addition to the above situation, there are many missing values in this data.

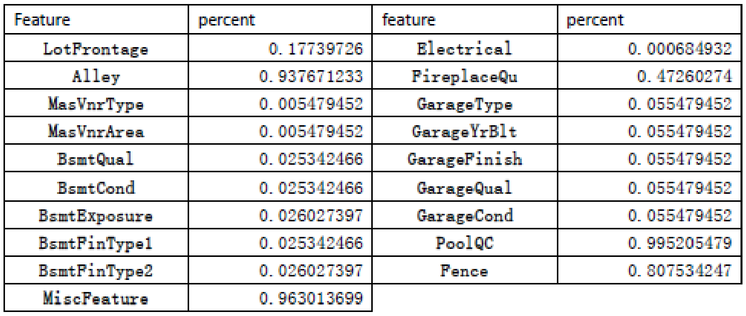
Through the basic understanding of the data, the original data cannot be directly used in the model. Therefore, this paper will transform the current data into data suitable for the model through a series of processing, such as data exploration, data cleaning, data visualization, feature engineering and so on.

4. Methods

In this section, we will conduct targeted processing according to the data characteristics and problems mentioned in the second part, explain the reasons for the processing and the reasons for such processing operation, and finally complete the visual display of the final cleaning data. It is important to note that the training set and testing set have the same problem, so this article before making a data operation will have test and training sets merged, so convenient we two data synchronization process, so below in all the description and data processing, are for the test set and the training after the collection and the overall data.

4.1 Missing value processing

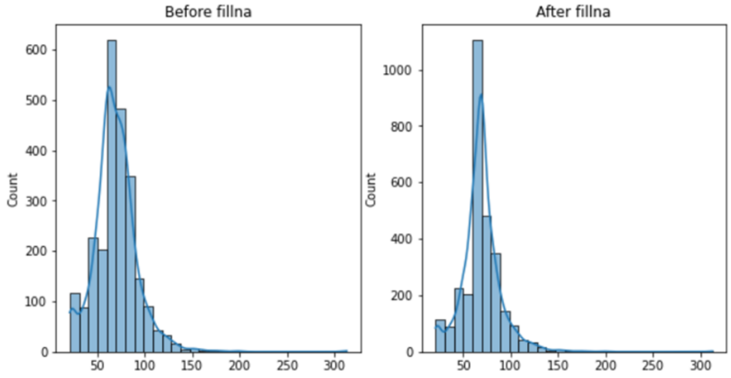
Table 2 feature with null value



Feature	percent	feature	percent
LotFrontage	0.17739726	Electrical	0.000684932
Alley	0.937671233	FireplaceQu	0.47260274
MasVnrType	0.005479452	GarageType	0.055479452
MasVnrArea	0.005479452	GarageYrBlt	0.055479452
BsmtQual	0.025342466	GarageFinish	0.055479452
BsmtCond	0.025342466	GarageQual	0.055479452
BsmtExposure	0.026027397	GarageCond	0.055479452
BsmtFinType1	0.025342466	PoolQC	0.995205479
BsmtFinType2	0.026027397	Fence	0.807534247
MiscFeature	0.963013699		

Many features in the data have missing values, of which MiscFeature, Fence, PoolQC and other missing values exceed 80%. These features can be directly determined as invalid features. Although there are many techniques for filling missing values, such as those based on existing data Simple filling methods, such as mean filling and median filling; there are also filling methods based on model fitting, etc. However, for features with a large proportion of missing values, we can reasonably believe that the feature is not important or meaningless. In this paper, the threshold of the missing proportion of feature values is set to 0.2, that is, when a feature has a value of more than 20%, we will remove the feature. Through this operation, a total of 5 features are eliminated.

For features with a small proportion of missing values, the easiest way is to delete all samples with missing values. However, since the data sample size is not large this time, direct deletion will lose a lot of information, so this article fills in. Considering that the features with missing values include numerical data and character data, two different strategies are adopted to fill them. For numerical data, we use the average value of the existing data to fill, and for character data, we use the mode value to fill, which can maintain the distribution of features to a large extent and fit the model later. Good effect. Figure 1 shows the change in the overall distribution of the feature LotFrontage before and after the missing value is filled.



Picture 1 feature distribution before and after fillna

4.2 Feature selection

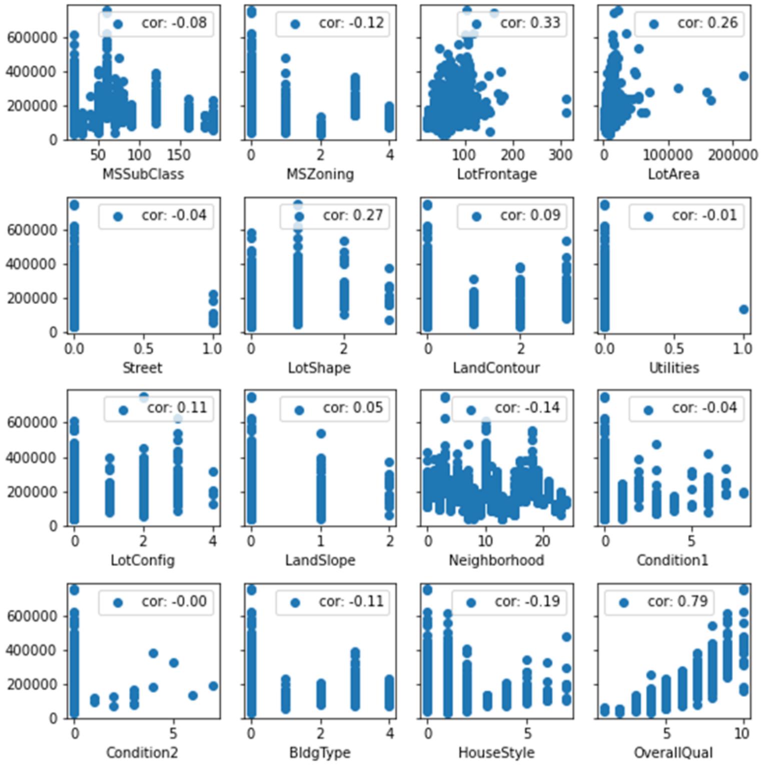
Feature selection is also called feature subset selection (FSS), or attribute selection. It refers to the process of selecting N features from the existing M features to optimize the specific index of the system, and the process of selecting some of the most effective features from the original features to reduce the dimension of the data set. It is an important means to improve the performance of the learning algorithm and also a key data preprocessing step in pattern recognition. For a learning algorithm, a good learning sample is the key to training the model.

Considering that the model in this paper is a regression model, it is considered to use the correlation coefficient to judge the strength of the correlation between feature and label. Since some of the current features are character types, the correlation coefficient cannot be calculated directly, so it is considered to convert these character features into numerical types. The operations to convert character type to numeric type mainly include one-hot encoding, dummy code encoding or directly replacing different classification values with numeric values. This paper uses the last method. For example, there are two values in feature Street, Pave and Grval, which will be represented by 0,1 in this paper. However, if there is an obvious size relationship among the values of a certain feature, for example, when the features with good or bad such as suburban apartment, ordinary apartment and high-class apartment are included, we should set higher values to represent the higher values.

(1) Scatter Diagram - Feature Removal

In this paper, all filtered features and Y (i.e., Sale price) values are made into a scatter plot to see the relationship between features and Y. Figure 2 shows the scatter diagram of some features and Y values, and the correlation coefficient between each feature and Y is described in the legend of each graph, so that it is convenient to view the correlation between different features and Y. For example, the scatter diagram in the upper left corner of

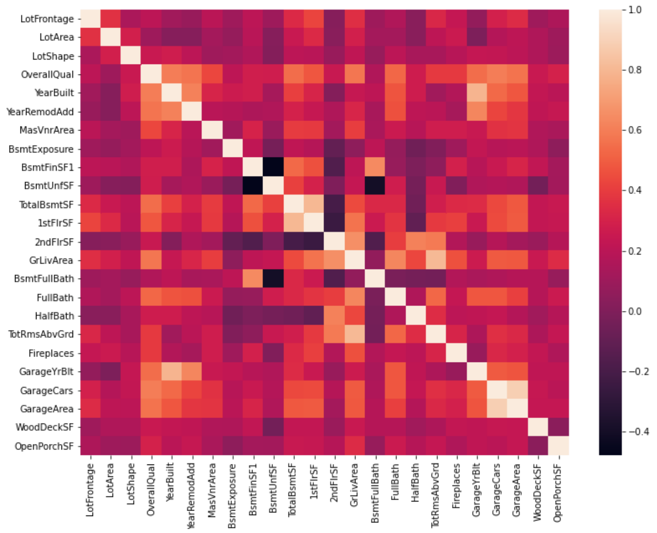
Figure 2 shows that the correlation coefficient between feature MSSubclass and Y is -0.08. It shows that there is basically no correlation between the two; At the same time, if we look at the scatter graph of the last graph in the lower right corner of the graph, it can be seen from the trend of the scatter graph that there is a strong correlation between the two. At the same time, if we look at the legend, we can find that the correlation coefficient of the two is 0.79, which belongs to a strong correlation.



Picture 2 feature and y scatter plots

We generally believe that when the absolute value of the correlation coefficient is less than 0.2, there is no correlation between the two variables. When the absolute value of the correlation coefficient is between 0.2 and 0.6, it is a weak correlation, while when the value is greater than 0.6, it can be a strong correlation. Therefore, in this paper, all features with a correlation coefficient less than 0.2 between features and y were removed by filtering through the correlation coefficient at first, because features with a relatively low correlation would not make significant contributions to the regression model. After that, there are only 24 features left in the data.

Then we drew a heat map for the correlation of all the remaining features, as shown in Figure 3. It can be seen from the figure that there are many features with strong correlation among them. Such as Gliraria (Above grade (ground) living area square feet) and TOTRMSABVGRD (Total rooms Above grade (does not include) Bathrooms). These two variables are, by definition, related. Therefore, there is the problem of collinearity between features. Faced with this problem, we can consider using dimension-reducing technologies, such as PCA and Lasso, to eliminate the existence of collinearity, so that the model can better fit the data, or use models that are not sensitive to collinearity, such as decision tree, random forest ridge and other technologies.

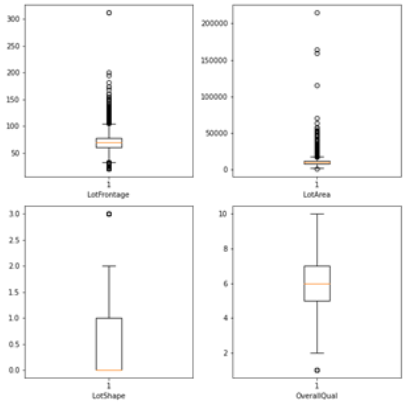


Picture 3 feature correlation hot map

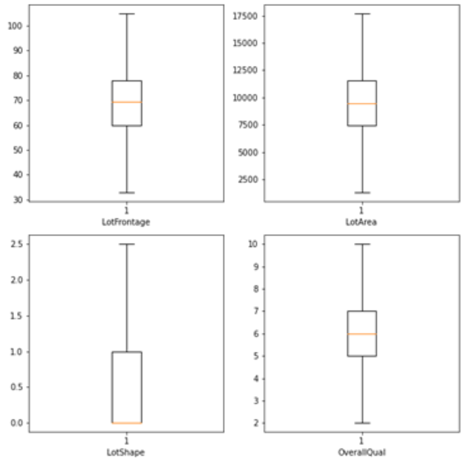
（2）Box plot-eliminate outliers

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement, or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Box plot is a very effective way to detect outliers. Therefore, this article draws box plots for all features, as shown in Figure 4. It is obvious from Figure 4 that many features have outliers., Such as feature LotFrontage, LotArea, etc., and the number of outliers of some feature’s accounts for a large proportion of the total. Considering that the overall sample size of the data is small, the direct deletion of outliers will seriously affect the number of samples. Therefore, this article truncates all outliers. When a certain value is greater than the top value of the feature box plot to which it belongs at this time, it will be cropped to the top value, and the same will be done for the value less than the bottom value, to ensure that all the values are between the top and bottom values of the box chart, while retaining all Sample. Figure 5 shows the box plot after performing the above operations on the data. All values are within a reasonable range.



Picture 4 some feature boxplot



Picture 5 some feature boxplot after delete outlier

(3) Feature reduction

In the second chapter, in the description of the data, we can know that the features of the current data have a large difference in scale, which is not conducive to the fitting of some regression models, because the model will be biased towards the influence of larger-scale features. Therefore, we need to normalize all features to the same value range. In this paper, the min-max scale method is used to make the value range of all features be between 0-1, as shown in formula (1):

$$X_{new} = \frac{X - X_{min}}{X_{min}} \tag{1}$$

5. Experiment

After the data prediction in Chapter 4, we have obtained relatively complete and reliable clean data. Next, we need to select a suitable model, fit the data, and make predictions.

5.1 Model selection

This article will use LinearRegression respectively, LassoCV RidgeCV, MLPRegressor, RandomForestRegressor these five models to fit the data, choose the reason of these models are:

1.LinearRegression, as a classical regression model, has a better interpretability and is easy to understand.

2. Both Lasso regression and Ridge regression can solve the problem of feature collinearity well

3. MLPregSOR can be used like neural network to extract deep characteristic information.

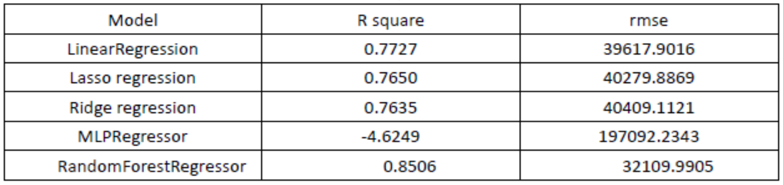
4.RandomForestRegressor Considering that feature species have many classification variables, and decision tree is good at fitting classification variables, we try to use the inheritance learning model with decision tree as weak classifier

The above models all have their own advantages and disadvantages, so in this chapter, we will first fit each model (under the default parameters of Sklearn), then select the models with better performance, and then conduct parameter tuning for them, to obtain better fitting effect.

5.2 Preliminary experimental results

Set the overall data set as 80: 20 was divided into training set and validation set, and then the basic model (that is, all parameters of model species used the default parameters of Sklearn species) was used to fit and evaluate the data (using validation set), and R square and root mean square error were used to evaluate the model.

Table 3 model performance



Model	R square	rmse
LinearRegression	0.7727	39617.9016
Lasso regression	0.7650	40279.8869
Ridge regression	0.7635	40409.1121
MLPRegressor	-4.6249	197092.2343
RandomForestRegressor	0.8506	32109.9905

As can be seen from Table 3, Linear Regression, Lasso Regression and Ridge Regression all have similar solutions in both R Square and RMSE, and they are all poor. The R square value of MLPregressor is less than 0, which indicates that the fitting effect of this model is very poor, and there is a very large gap between the predicted value and the real value. Among the five models, RandomForestRegressor was the best, and R Square could reach 0.8574, indicating that the model could explain the data to a large extent.

5.3 Parameter tuning

Through the experimental data in 5.2, we have concluded that RandomForestRegressor is the best model among the five models. Therefore, in this section, we will adjust the parameters of the model to make the model get a better effect. Grid search method is used here, that is, all the pre-set parameters are fully arranged, and then the model of each parameter combination is trained and evaluated. Finally, the parameter combination with the highest evaluation index is selected.

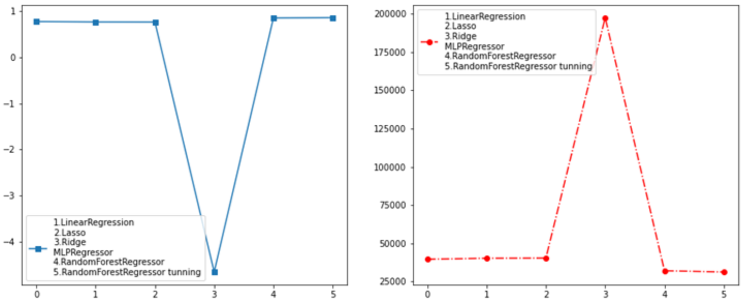
Table 4 parameter

Parameter name	Parameter definition	Parameter list
n_estimators	The number of trees in the forest.	100,300, 500
max_depth	The maximum depth of the tree	None, 1, 2, 5, 8
bootstrap	Whether bootstrap samples are used when building trees	True, False
criterion	The function to measure the quality of a split	mae, mse

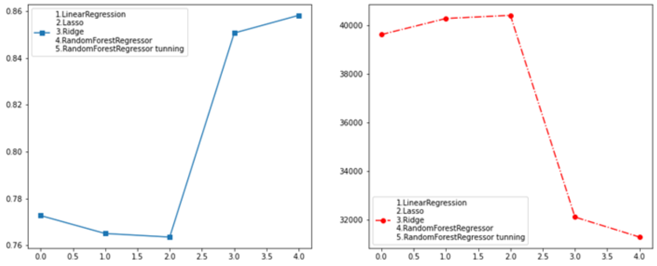
Through the combination and training evaluation of the above parameters and their values, the best parameter combination is finally obtained as follows: n_estimators = 300, max_depth=None, bootstrap = True, criterion=mae. The best model evaluation results were R square = 0.8582 and RMSE =31287.7824.

5.4 Visualization of results

Figure 6 and Figure 7 below can easily compare the performance differences between different models. MLPregression was the worst for performance resolution, while the random forest model with parameter adjustment achieved the best result.

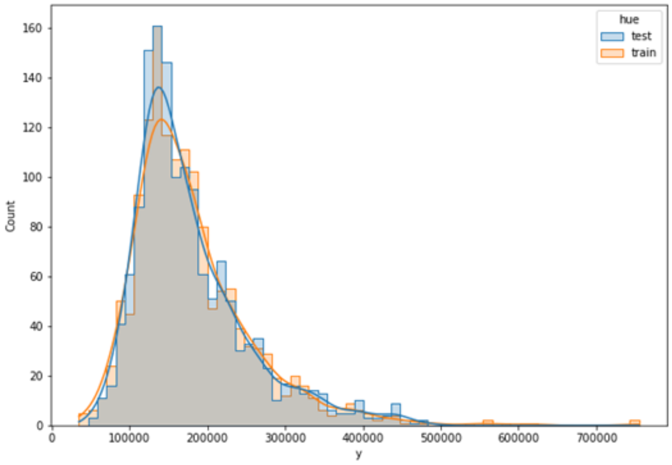


Picture 6 model compare with mlpregression



Picture 7 model compare without mlpregression

Then we use the selected best model to predict the results of the test and make a histogram of the final predicted results and the Y value in the training set. The distribution of the two is basically consistent, which indicates that the prediction result of this time is highly reliable. If the results of the two distributions are completely different, then the prediction result of the model can be poor with high probability.



Picture 8 true and prediction y distribution

6. Conclusion

Through this modeling and prediction, Table 5 can be obtained, which shows the importance of different features. It can be found that the pricing of houses is only affected by some of their own features, such as feature OverallQual and GrLivArea, which have a decisive influence on the price.

Table 5 features importance

feature	importance
OverallQual	0.3594
GrLivArea	0.1144
TotalBsmtSF	0.0660

This paper uses a variety of data science and technology, from data cleaning to model training and prediction, and obtains a relatively satisfactory model, and its prediction results are also convincing, but the overall model RMSE value is relatively large, indicating that there is still a lot of room for improvement. In the later stage, we will have a deeper understanding of feature construction and neural network and other related technologies, and hope to build a more powerful prediction model based on this current basis.

References

[1]Rosen S . Hedonic Prices and Implicit Markets: Product Differentiation in Pure

Competition[J]. Journal of Political Economy, 1974, 82(1):34-55.

[2]Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial

neural network[J]. Expert Systems with Applications, 2009, 36(2):2843-2852.

[3] Chen Shipeng. Housing Price Forecasting Based on Random Forest Model [J]. Science and Technology Innovation and Application, 2016(4):52-52.

[4]Fan C C , Yuan S M , Zhang X , et al. A House Price Prediction for Integrated Web Service

System of Taiwan Districts[C]// International Conference on Genetic & Evolutionary

Computing. Springer, Singapore, 2017.

[5]Stephen Law, Brooks Paige, Chris Russell. Take a Look Around: Using Street View and

Satellite Images to Estimate House Prices[J]. Papers, 2018.

[6]Vineeth N , Ayyappa M , Bharathi B . House Price Prediction Using Machine Learning

Algorithms: Second International Conference, ICSCS 2018, Kollam, India, April 19–20,

2018, Revised Selected Papers[M]// Soft Computing Systems. 2018,425-433.