Predicting Housing Prices Using Machine Learning Algorithm

Zeyi Sun 605645417, Xiuqi Li 605638474, Ziyi Gu 705631863, Difei Chen 305640614, Qinwen Yan 505646525

I. Introduction

In this project, our main goal is to use different machine learning models to predict housing prices and select the model with the highest accuracy. The models we used include Linear Regression, Lasso, Ridge, Elastic Net, SVM, and Random Forests. According to the accuracy of prediction, we found the Lasso model had the highest scores. Therefore, we selected the Lasso model and further analyzed the issue of undervalued house prices.

II. Data Exploration

1. Selecting Variables

The original data set has 55 independent variables. According to the completeness and relevance of variables, we finally selected 35 independent variables, which included TaxRateArea, PropertyType, GeneralUseType, SpecificUseType, YearBuilt, EffectiveYearBuilt.etc

2. Data Processing

In this part, we replaced the missing values with the mode. Then we did the data transformation. We also converted the category variable into dummy variables. Finally, we finished the data standardization to ensure the scale of each variable is on the same level.

3. Data Visualization

We used histogram for each independent variable to analyze their distribution characteristics.

日历

描述已自动生成

日历

描述已自动生成

日历

描述已自动生成

III. Models

1. Model Description

In this section we will briefly introduce the regression models we use.

The first regression model is Linear Regression. Linear regression is a linear approach for modelling the relationship between a dependent variable and one or more explanatory variables. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

Linear regression is a basic regression model which can be easily extended to other statistical models, including Lasso, Ridge and Elastic Net. Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. The difference between Lasso and Linear Regression is that there is a l1 regularization part in the loss function of Lasso. The l1 regularization function can be written as

Ridge is also a regression model generated from Linear Regression and designed to prevent overfitting issues. Ridge has a l2 regularization part in its loss function that does not exist in the loss function of Linear Regression. The l2 regularization function can be written as

Elastic Net is the combination of Lasso and Ridge, which has both l1 and l2 regularization functions in its loss function. The regularization function in Elastic Net can be written as

Boosting and Bagging are resampling methods, which are combined with other weak algorithms to create a strong prediction algorithm. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. In this case, we combine the bagging and boosting with Decision Tree Regressor (default model type in python functions).

2. Model Summary

In this section, we use the training samples to train each algorithm and select the best regression model and hyperparameters by comparing their prediction results in testing samples. The R-squared values of each model are shown in the table below.

|  |  |
| --- | --- |
| Models | R-squared |
| Linear Regression | 0.9677 |
| Lasso (alpha=0.01) | 0.9683 |
| Ridge (alpha=200) | 0.9678 |
| Elastic Net (alpha=0.001) | 0.9678 |
| Bagging Regressor | 0.9668 |
| Boosting Regressor | 0.0842 |

IV. Further Analysis: Undervalued Issue

1. Undervalued Issue

From the accuracy above, we selected the Lasso Model and used it to make the prediction.

We defined that the error is equal to the true price minus the predicted price.

If the error is larger than zero, which means the predicted value is lower than the true value and the undervalued issue exists.

2. Analysis of Undervalued Price

First, we ranked the value of error and calculated the mean, standard deviation, minimum, first quartile, median, third quartile, and maximum.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Y\_true | Y\_predict | Error |
| Mean | 28.40 | 20.53 | 7.87 |
| Std | 33.06 | 27.05 | 13.73 |
| Min | 1.92 | 1.07 | 0.68 |
| 25% | 5.87 | 2.92 | 1.00 |
| 50% | 15.50 | 9.94 | 2.07 |
| 75% | 32.93 | 22.54 | 9.05 |
| Max | 181.87 | 151.83 | 105.83 |

Then, we selected data with an error greater than 10, which is shown in the coding file.

3. Reason for the Undervalued price

First of all, the score of our Lasso model is around 0.9683, which shows that the model has a high level of accuracy. Through calculating, we can also find that the mean of the top 100 error is around 7.87, which is also at a low level. But from the error ranking table, the largest error value is 105, and the second-largest error value is only 45. Therefore, we can conclude that outliers exist, which caused a high error and severe underestimation.

Besides, through analyzing the distribution of the y variable, we find that the underestimated data are concentrated on both sides of the distribution and data with accurate predictions are concentrated in the middle of the distribution. The sample size distributed on both sides is much smaller than the sample size concentrated in the middle. Therefore, we think increasing the sample size may help address the undervalued issue.

V. Conclusions and Future Work

In this project, we used different machine learning models to predict housing prices. Through splitting data into train set and test set and selecting the best parameter for each machine learning model, we find the Lasso Model has the highest scores with 0.9683. Then we also further analyze the undervalued problem. We conclude that it is the outliers and small sample size that cause the underestimation.

In the future, we will improve the following aspects. First, increasing the sample size can help enhance the model performance. Second, this data set has many missing values, and we replaced the missing values with the mode. If we could have more complete data, maybe the accuracy could be improved. Third, we didn’t remove the outliers. In the future, we will have a more careful data exploration and remove the outliers to improve the accuracy. Last but not least, we can also try other advanced machine learning models to make a better prediction, such as the artificial neural networks.

Link of our code: <https://github.com/xiuqili29/445-project/blob/main/HousingPrice.ipynb>