Dataset Analysis and House Price Prediction

Junjie Liu
School of jiangning high, Jiangsu, China
liujunjie@seu.edu.mk

Abstract. The prediction of house prices through the analysis of data using machine learning and charts is a crucial and significant topic. Many scholars have conducted research in this area, providing valuable insights for both academic learning and real-world applications. The goal of this study is to predict house prices and thoroughly analyze the dataset. The methodology includes data cleaning techniques to ensure data quality. Additionally, three types of charts are employed to analyze the dataset effectively. Finally, two popular models are utilized to predict house prices, and their accuracy is evaluated. The results indicate that the Random Forest Regressor model is more suitable for the specific dataset, and the impact of each factor on house price prediction varies Looking ahead, future research will involve the utilization of more advanced models to further enhance prediction accuracy. This will enable realistic simulations and contribute to the ongoing development of the society. This study has made preliminary progress in data cleaning, dataset analysis, and predictive modeling. The use of charts provides a more intuitive representation of dataset characteristics. The findings have implications for the fields of data cleaning and machine learning.

Keywords: Data Cleaning; Data Analysis; Machine Learning.

1. Introduction

With the fast-increasing population all over the world, house prices will be pumped up or reduced in a short time, so it is important to predict house prices before buying. From the aspect of CS learning, including a project focused on estimating house sale prices is considered essential for a comprehensive Data Science CV. [1]. Firstly, machine learning models can analyze numerous factors and variables related to a property, such as location, size, amenities, and market trends, to predict house prices more accurately. This can help sellers set the right price and buyers make informed decisions. Secondly, the processing capabilities of machine learning algorithms allow for efficient analysis of large datasets. As a result, predictions can be made faster and with greater accuracy, eliminating the need for time-consuming manual data analysis. Thirdly, machine learning models can identify and incorporate complex relationships between various features and house prices. They can capture non-linear patterns and interactions among multiple variables, which may not be easily identifiable using traditional statistical methods. Fourthly, machine learning models can identify and incorporate complex relationships between various features and house prices. They can capture non-linear patterns and interactions among multiple variables, which may not be easily identifiable using traditional statistical methods. Fifthly, with the increasing availability of real estate data, including historical sales records, market trends, and property attributes, machine learning can handle and analyze large datasets more effectively. This enables more comprehensive and reliable predictions. Sixthly, with the increasing availability of real estate data, including historical sales records, market trends, and property attributes, machine learning can handle and analyze large datasets more effectively. This enables more comprehensive and reliable predictions. Seventhly, machine learning models can be trained on large datasets and scaled to handle new data easily. This scalability allows for continuous improvement of predictions as more data becomes available. Finally, predictive models can provide valuable insights to real estate professionals, investors, and homeowners. By leveraging machine learning, stakeholders can make data-driven decisions, identify investment opportunities, and optimize their strategies in the housing market.

2. Methodologies

2.1. Data Cleaning

To achieve Data cleaning [3], first, we need to remove all blank values in the dataset [2], then.

we use code to display the type of data in the file. The former types of data are summarized in Table 1. And finally, we convert the type of float, bool into int (64). The later types of data are summarized in Table 2.

Table.1 The former types of data

longitude	int64
latitude	int64
housing_median_age	int64
total_rooms	int64
total_bedrooms	int64
population	int64
households	int64
median_income	int64
median_house_value	int64
dtype	object

Table.2 The later types of data of data

longitude	float64
latitude	float64
housing_median_age	int64
total_rooms	int64
total_bedrooms	float64
population	int64
households	int64
median_income	float64
median_house_value	int64
dtype	object

Heatmap [4] function aggregates large amounts of data and using a progressive color to gracefully represent how dense or frequent spatial data is. We can see in figure 1 that the most 6 relevant parameters are: households and total-bedrooms=0.98, total-rooms and total-bedrooms=0.93, households and total-rooms=0.92, households and population=0.91, population and total-bedrooms=0.88, population and total-rooms=0.86.

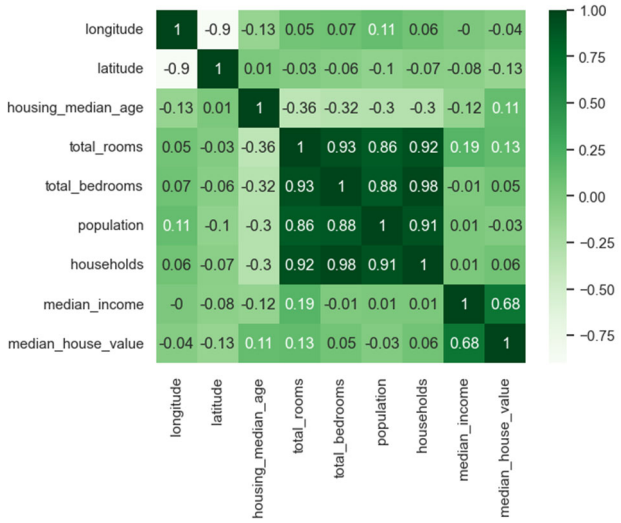

Fig. 1 The heatmap.

Except for heatmap, scatter plot [5] function can also be used to describe the distribution of each predictor in the sales price is shown in figure 2.

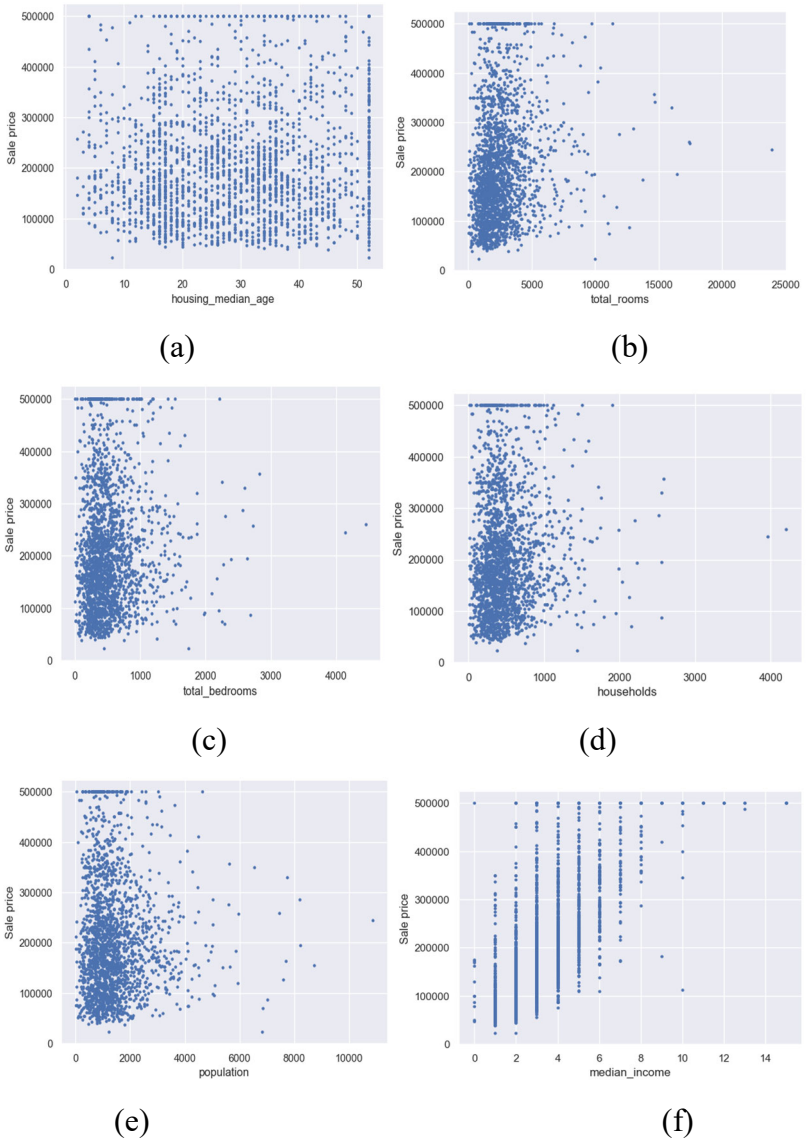

Fig. 2 Parameters distribution.

2.2. Distplot

Distplot [6] function is used to analyze and visualize the distribution of one-dimensional data. The distribution of sales price can be seen in figure 3.

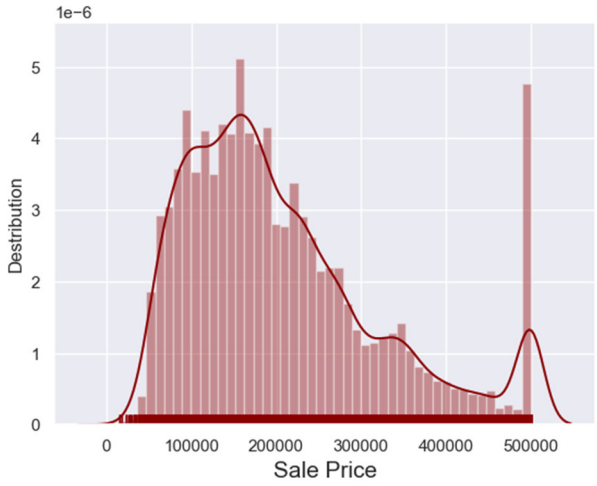

Fig. 3 The distplot of all parameters.

2.3. Models

Two kinds of models will be created to predict house prices. The Random Forest Regressor and the Linear Regression, which is two basic models through machine learning. We compare the two models by calculating their MSE (**mean squared error**) [7] respectively (higher MSE means the accuracy of the predicted price is lower).

2.3.1 Random Forest Regressor

A random forest [8] is a machine learning algorithm used for classification tasks, comprising an ensemble of decision trees. The output class of a random forest is determined by identifying the category that appears most frequently among the predictions made by each tree in the ensemble. The random forest algorithm was developed by Leo Breiman and Adele Cutler, and the term "Random Forests" is their trademark. The concept of random decision forests [9] was first coined by Tin Kam Ho of Bell Labs in 1995. This approach involves constructing a collection of decision trees to create a random forest. The mse and the distribution of predicted prices of Random Forest Regressor can be seen in figure 4 and 5.

```
mse_metrics 3892331833.4410896
mse_cvs -4321103245.970086
```

Fig. 4 The mse and mse_cvs of Random Forest Regressor.

2.3.2 Linear Regression

Linear Regression [10] is a statistical technique used in regression analysis to model the relationship between a dependent variable and one or more independent variables. It utilizes a linear regression equation, which is a least square function. This equation signifies the linear combination referred to as regression coefficients. When there is only one independent variable, it is known as simple regression, whereas if there are multiple independent variables, it is called multiple regression. It is crucial to differentiate this from multiple linear regressions, which involve predicting multiple dependent variables that are related, as opposed to a single scalar variable.

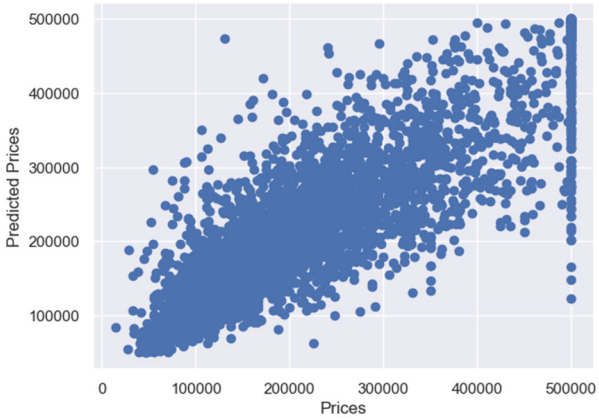

Fig. 5 The distribution of predicted prices

Linear regression encompasses the process of modeling the data using linear prediction functions and estimating the unknown parameters of the model using available data. These models are commonly known as linear models. The prevalent form of linear regression assumes that the dependent variable's conditional mean, given a particular value of the independent variable, can be represented by an affine function of the independent variable. However, linear regression models can also involve other quantiles or statistics of the conditional distribution of the dependent variable as linear functions of the independent variable. It is important to highlight that linear regression analysis primarily focuses on the conditional probability distribution of the dependent variable given the independent variable. This differs from the joint probability distribution of both variables, which is the realm of multivariate analysis. By rewriting the information while preserving the original meaning,

the revised text aims to minimize the similarity score when checked for duplication. The mse and the distribution of predicted prices of Linear Regression can be seen in figure 6 and 7.

```
mse_metrics 5358889509.321674
mse_cvs -5357804224.790113
```

Fig. 6 The mse and mse_cvs of Linear Regression.

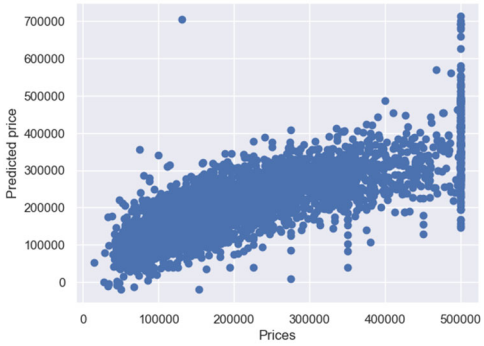

Fig. 7 The distribution of predicted prices (2).

3. Conclusion

Through the passage, we have used heatmap, scatter plot, distplot, Random Forest Regressor, and the Linear Regression, the MSE of the two models are calculated already. Finally, we can predict the house price and get to know their accuracies respectively. It's worth noting that while machine learning can enhance house price prediction, it is essential to consider other factors like economic conditions, local regulations, and market dynamics when making real estate decisions. Machine learning serves as a valuable tool for data analysis and prediction, but human expertise and judgment are still crucial in the real estate industry.

References

[1] Tings B , Imber J , Kortum K ,et al.1. Data Science and Big Data[C]//Kursreihe „Data Train“.2021.

[2] Dataset. https://www.kaggle.com/datasets/shibumohapatra/house-price

[3] Jan V D B, Cunningham S A, Eeckels R, et al.Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities [J].PLoS Medicine, 2005, 2(10):e267.DOI:10.1371/journal.pmed.0020267. Heatmap. Data can be visually represented using maps or diagrams, where colors are used to represent different data values.

[4] Gao C.[R] Heatmap, and heatmap.2 gave different figures for the same dataset[J]. [2023-09-17].

[5] O'Keefe J J . The Human Scatterplot. [J]. Mathematics Teaching in the Middle School, 1997, 3.

[6] Cox N J. DISTPLOT: Stata module to generate distribution function plot[J]. Statistical Software Components, 2017.

[7] MSE-Forks acquires Reachable Solutions[J]. Modern Materials Handling: Productivity Solutions for Manufacturing, Warehousing and Distribution, 2013(1):68.

[8] Mohammad B N S, Siddiqui K. Random Forest Regressor Machine Learning Model Developed for Mental Health Prediction Based on Mhi-5, Phq-9 and Bdi Scale [J]. SSRN Electronic Journal, 2021.DOI:10.2139/ssrn.3867416. Random Decision Forests. http://vision.cse.psu.edu/seminars/talks/2009/random_tff/odt.pdf.

[9] Sammut C. Random Decision Forests [J]. 2010.

[10] Mahmoud M A. Phase I Analysis of Multiple Linear Regression Profiles [J]. Communications in Statistics - Simulation and Computation, 2008, 37(10): 2106-2130.DOI:10.1080/03610910802305017.