Application of Multiple Linear Regression with Regularization on Boston Housing Datasets

Yuanwei Ding[1,a*], Hexing Zhou[2,b], Chak Hoi Huang[3,c], Haoxiang Zhang[4,d]

[1]Qingdao No.58 High School, Qingdao, Shandong, China
[2]Independent Schools Foundation Academy, Hong Kong, China
[3]British international Shanghai School, Shanghai, China
[4]Eastern Christian High School, New Jersy, US

[a*]3141205326@qq.com, [b]charlie_zhou1117@163.com,
[c]Huang_chak-hoi@bisspuxi.com,
[d]haozha24@student.easternchristian.org

Abstract. This paper first introduces the principle of multi-objective linear regression, and studies the Boston housing price data set with regularized multiple linear regression. Then this paper combines the knowledge of machine learning to build a prediction model. In the final forecast of the Boston house price, it was about 78 percent accurate compared to the real house price.

Keywords: Linear Regression, Multiple Linear Regression, Lasso regression, Ridge regression, Machine learning.

1 Introduction

After fully learning the relevant basic knowledge, the theory of linear regression is planned to be put into practice. In this investigation, our goal is to investigate the application of approximation on dataset. The main fields of approximation investigated is Linear regression with regularization. As linear regression is widely used among different datasets, this work aims to apply Linear Regression to a specific dataset and look at different linear regression methods. To generate a linear regression model, knowledge of programming and machine learning is applied. Our investigation is important as Linear Regression is a useful tool in data science which different data collection methods all need Linear Regression to help them approximate data and generate a general form that represents the data collected. It is not only useful, but also beneficial for our further research as data must be analyzed in order to form conclusion. Because of this, this work aims to use this investigation to analyze the data of Boston Housing and use this real-life problem to help us investigate deeper into the term of Linear Regression.

J. Liao et al. (eds.), *Proceedings of the 2024 2nd International Conference on Digital Economy and Management Science (CDEMS 2024)*, Advances in Economics, Business and Management Research 292,
https://doi.org/10.2991/978-94-6463-488-4_3

2 Methodology

2.1 Linear Regression

In this investigation, the key area is linear regression. Linear regression is a model which shows how the dependent variable changes based on the independent variables. The variables involved in this model are independent input variables and dependent output variables. In simple linear regression, variables will form linear equations in the form of:

$$Y = Ax + B \tag{1}$$

In this equation, Y value is the dependent variable which is in the form of a vector. B value is the y intercept and it is also in vector form. On the other hand, x value is the independent variable and it is in matrix form. It is similar for A value which is also a matrix.

Linear regression is able to measure the relationship between two sets of data, but the two sets of data do not need to have a clear relationship. Because of this Linear regression is useful in different areas of knowledge. Its importance can be illustrated in different directions. The most common use of linear regression is to estimate the future value or the missing value of a data set. It is also useful when measuring the strength of the relationship between two variables by using the R-square value. Besides its importance, linear regression is choosed as the tool to organize data and investigate also because it is accurate and easy to understand [1, 2].

The R-square value determines the linear regression line. The R-squared value is the value that shows the proportion of the variance in the dependent variable that can be explained by the independent variable. It is also used to determine how well the data point fits the regression line [3].

Like all mathematical algorithms, linear regression is formed based on assumptions. The first assumption is Linearity. Linearity assumes that the data set is having a linear relationship, which means the dependent variable changes constantly with the independent variable. The second assumption is normality and homoscedasticity, which means the errors are evenly distributed in the data set and have equal variance. The third assumption is no endogeneity, which means the independent variables chosen are not related to the errors presented. The fourth assumption is no autocorrelation which means the errors are not dependent on one another. The last assumption is no multicollinearity which means that the change in one variable should not affect the other variable. In this case, linear regression is an ideal model which can be used to investigate the relationship between two variables [1].

2.2 Multiple Linear Regression

Multiple Linear Regression is a statistical methodology that uses multiple variables to predict the outcome of a response variable. Multiple linear regression is also known as multiple regression and it extends from simple linear regression as it involves multiple

variables. For multiple regression, it has a more complex formula than simple linear regression.

The equation is:

$$Y_i = \beta_0 + \beta_{i1}x_{i1} + \beta_{i2}x_{i2} + \beta_{i3}x_{i3} + \ldots\ldots + \beta_{ip}x_{ip} + \epsilon \quad (2)$$

In this equation, $i = n$ observations. Y_i is the dependent variable. x_i represents the explanatory variables. β_0 represents the y-intercept and β_i represents the slope coefficient of different explanatory variables. ϵ is the error term of this equation.

To optimize multiple Linear regression, Python is used as a tool to build the model. In this work, Python is used to program our data set as code will be more reliable than human calculation. Code is planned to be learned from the experts in coding and try to change it into a form that is suitable for our data set.

2.3 Linear Regression with Regularization Term

Previously, linear regression is briefly introduced. In order to make the analysis of linear regression more scientific and accurate, lasso regression and ridge regression are adopted. They all belong to regularization techniques. The so-called regularization technique is to find the balance between model complexity and goodness of fit by adding a penalty term to the Loss function. In short, the purpose of applying lasso regression and ridge regression tools is to prevent the overfitting of data and ensure the accuracy of experimental conclusions. Below, the principles of these two technologies are explained separately and apply them in the final experiment through Python.

2.3.1 Lasso Regression

Least absolute shrinkage and selection operator regression, also known as Lasso regularization and L1 regularization, minimizes the number of coefficients that are not zero, that is, reduces the model data to make it closer to zero. The penalty term is proportional to the absolute value of the coefficient. In the end, only a portion of the input information is used as valid data for linear regression and subsequent data prediction. In other words, this approach helps us leave behind more beneficial and critical parts for data prediction [4].

$$J = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda|w| \quad (3)$$

Where: λ is the regularization parameter controlling the strength of the regularization, N is the number of samples, $\sum$denotes the sum over all observations, w represents the weight of the model, $\widehat{y_i}$ is the target value for the i-th sample, $\hat{y}_i$ is predicted value for the i-th sample, and J is an objective function that is tried to minimize [5].

Next, the derivation of the cost function for parameter w are provided:

$$J = (Y - Xw)^T(Y - Xw) + \lambda|w| \quad (4)$$

$$J = Y^TY - 2Y^TX + w^TX^TXw + \lambda|w| \quad (5)$$

$$\frac{\partial J}{\partial w} = -2x^T y + 2x^T xw + \lambda sign(w) \tag{6}$$

$$w = X^T(\hat{y} - y) + \lambda sign(w) \tag{7}$$

Where: sign(w) = {1, w > 0; 0, w = 0; -1, w < 0}.

Lasso method overcomes the shortcomings of traditional methods in selecting models. Therefore, this method has received great attention in the field of statistics [6].

2.3.2. Ridge Regression

Ridge regression, also known as L2 regularization. Ridge regression analysis is a biased estimation method specially used for collinear data analysis. It seeks a regression process with less effect but more in line with reality at the cost of giving up the unbias of least squares and some accuracy [7]. Compared to the former, ridge regression reduces the linear regression model slightly differently. It also incorporates the penalty term into the Loss function to distribute the coefficient more evenly and reduce the amplitude of the coefficient. The difference is that it can reduce the coefficient to 0, which is quite effective in preventing overfitting [5].

$$J = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda |w|^2 \tag{8}$$

After defining the objective function, this work can continue to discuss the Partial derivative of the objective function with respect to the parameter w.

$$J = (Y - Xw)^T (Y - Xw) + \lambda w^T w \tag{9}$$

$$J = Y^T Y - 2Y^T X + w^T X^T Xw + \lambda w^T w \tag{10}$$

$$\frac{\partial J}{\partial w} = -2x^T y + 2x^T xw + 2\lambda w \tag{11}$$

$$(\lambda I + w^T X)w = X^T Y \tag{12}$$

$$w = X^T(\lambda I + X^T X)^{-1} X^T Y \tag{13}$$

Where: λ is the regularization parameter controlling the strength of the regularization, (expression)^Tis transpose matrix, X is input data in the vector form, Y is in the target data in the vector form, I is the identity matrix of size D (dimensionality of data). The identity matrix is a square matrix with ones along the main diagonal and zeros everywhere else. (AI=IA) Utilized for matrix operations, and J is the objective function that is tried to minimize [5].

3 Application

3.1 Datasets

Two datasets are introduced in this study, namely the statistics of Boston housing prices with their related social factors. As the housing price in Boston are investigated, this

work handles a large range of variables. While simple linear regression is only focusing on one dependent variable and one independent variable, it is not suitable for our investigation. Multiple regression is needed to process our data.

3.2 Description of Datasets

Boston House dataset: The dataset specifically focuses on housing prices in Boston, providing valuable information for analysis and modeling in the real estate sector.

With Boston being a prominent city in the United States, understanding the factors that influence housing prices is crucial for homeowners, real estate agents, and analysts. This dataset serves as a valuable resource for exploring and investigating these factors. It is likely to contain a comprehensive set of features related to the housing market, such as the number of rooms, crime rates, accessibility to amenities, and property age, among others.

By utilizing this dataset, researchers, data scientists, and analysts can gain insights into the dynamics of Boston's housing market. They can perform various analyses, including exploratory data analysis, regression modeling, and predictive analytics, to uncover patterns, trends, and correlations between housing features and prices. The dataset's size and quality provide a robust foundation for developing machine learning models that can predict or estimate house prices based on the given features.

Researchers and enthusiasts in the field of real estate, urban planning, and data science can leverage this dataset to understand the drivers of housing prices in Boston and potentially apply the insights gained to other cities with similar dynamics. Overall, this dataset offers an excellent opportunity for data-driven exploration and modeling in the domain of Boston's housing market.

3.3 Implementation

When making a forecast of the Boston housing price problem based on the data set, a supervised linear regression model is built because the target variables are continuous. When the final regression is performed, the histogram should be a bell curve, with slight slants acceptable.

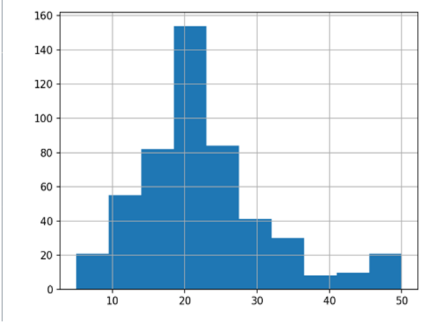

Fig. 1. Data distribution.

Figure 1 shows that the running result of the target variable meets the requirements. In this data set, each column has the following meanings:

CRIM - per capita crime rate by town
ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS - proportion of non-retail business acres per town.
CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX - nitric oxides concentration (parts per 10 million)
RM - average number of rooms per dwelling
AGE - proportion of owner-occupied units built prior to 1940
DIS - weighted distances to five Boston employment centres
RAD - index of accessibility to radial highways
TAX - full-value property-tax rate per 10,000 dollars
PTRATIO - pupil/teacher ratio by town
B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT - % lower status of the population
MEDV - Median value of owner-occupied homes in $1000's

After deleting variables with large missing values and exploring the data, the following two classification factors 'CHAS' and 'RAD' are found and classified the data according to this criterion (see Figure 2):

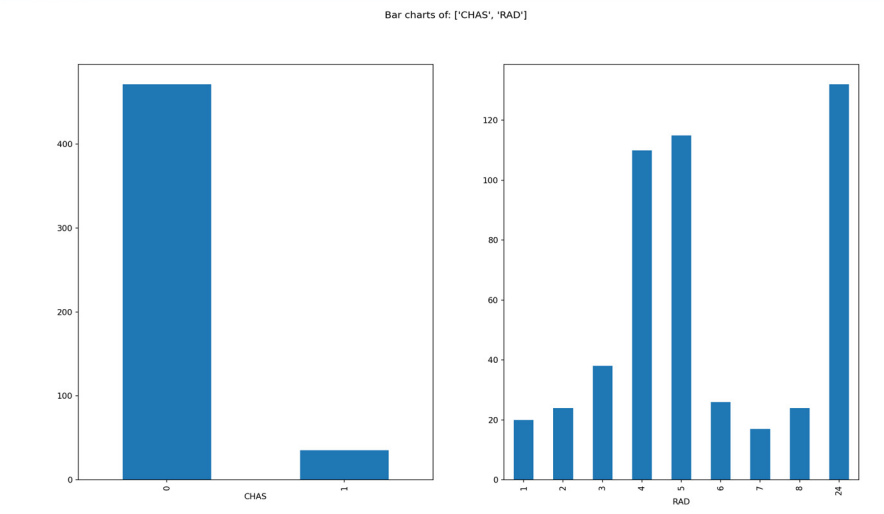

Fig. 2. Data classification.

The histogram below (Figure 3) shows us the data distribution for a single continuous variable, with the X-axis representing the range of values and the Y-axis representing the number of values within that range.

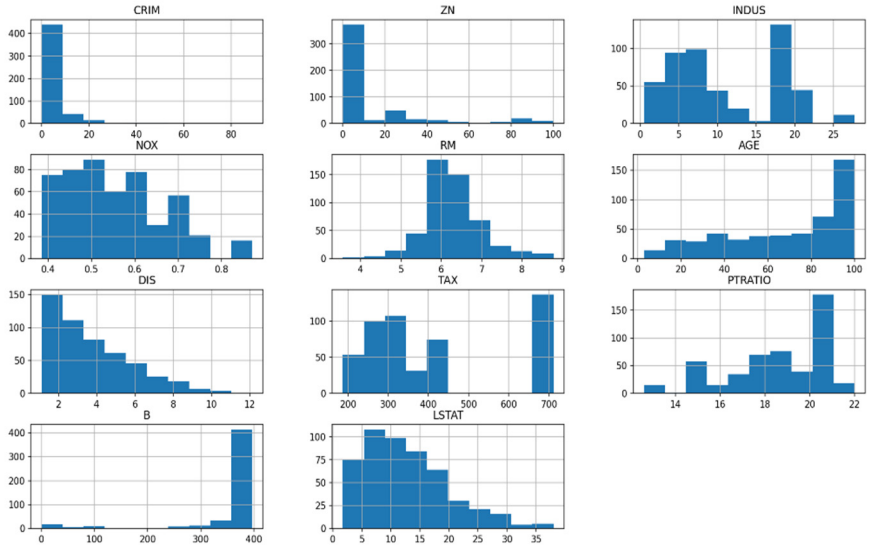

Fig. 3. The distribution of a single continuous variable.

Next, outliers that are far from most of the data and will replace the outliers by looking at the histogram to find the most logical value. The improved distribution is shown in the figure 4.

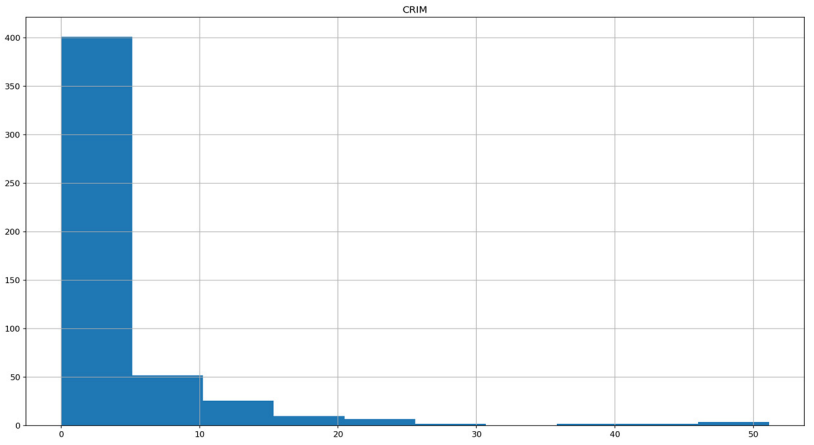

Fig. 4. Data distribution after outliers are replaced.

When the target variable is continuous and the predictor variable is also continuous, a scatter plot can be used to visualize the relationship between the two variables and use pearson's correlation values to measure the strength of the relationship, as Figures 5, 6, 7, 8, 9 and 10 shows:

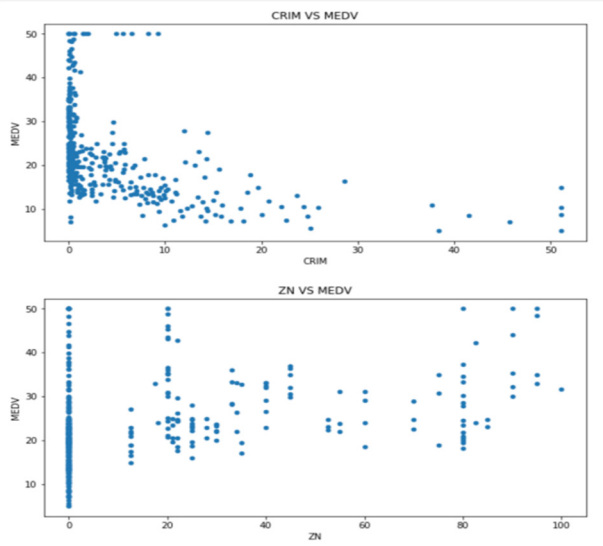

Fig. 5. Scatter plot of variable relationship.1

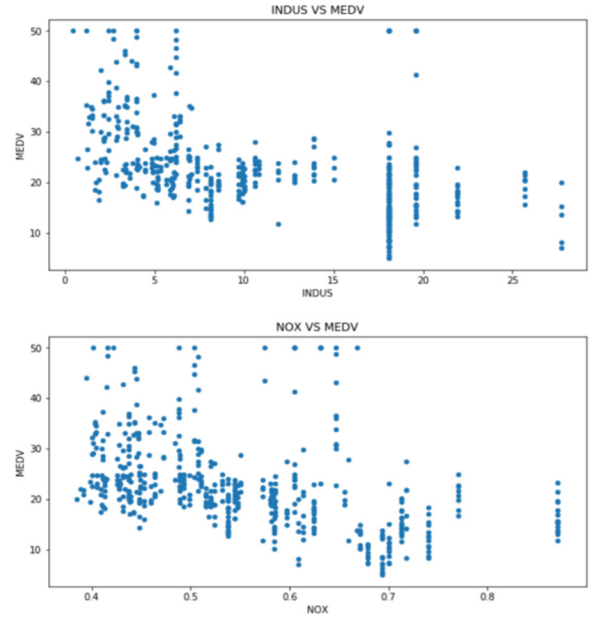

Fig. 6. Scatter plot of variable relationship.2

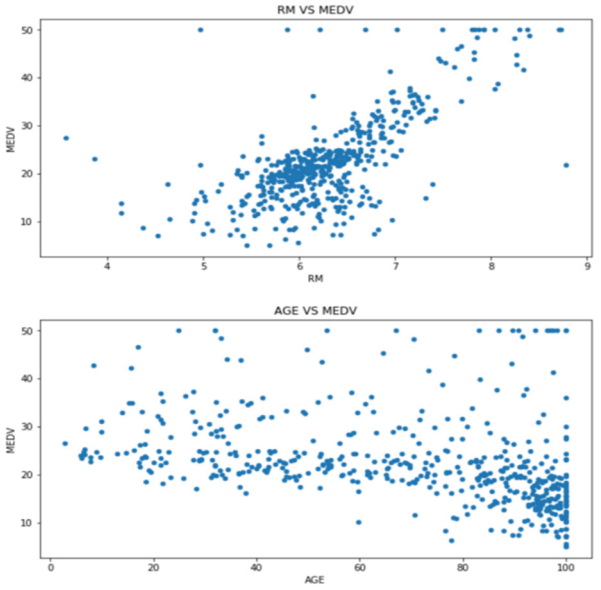

Fig. 7. Scatter plot of variable relationship.3

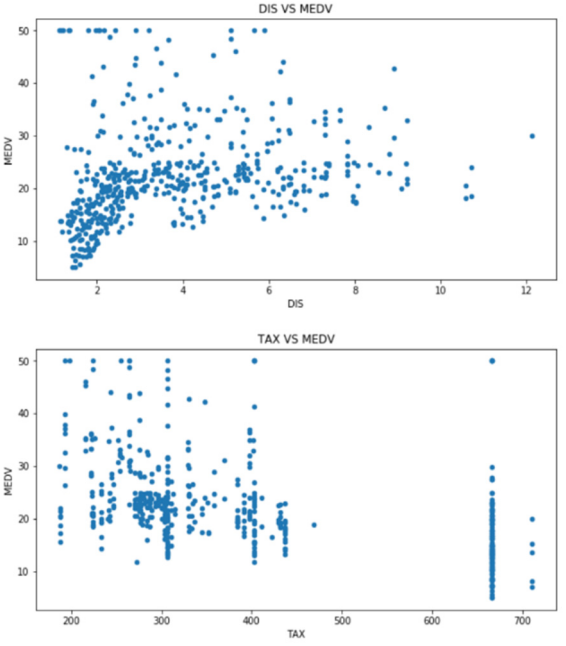

Fig. 8. Scatter plot of variable relationship.4

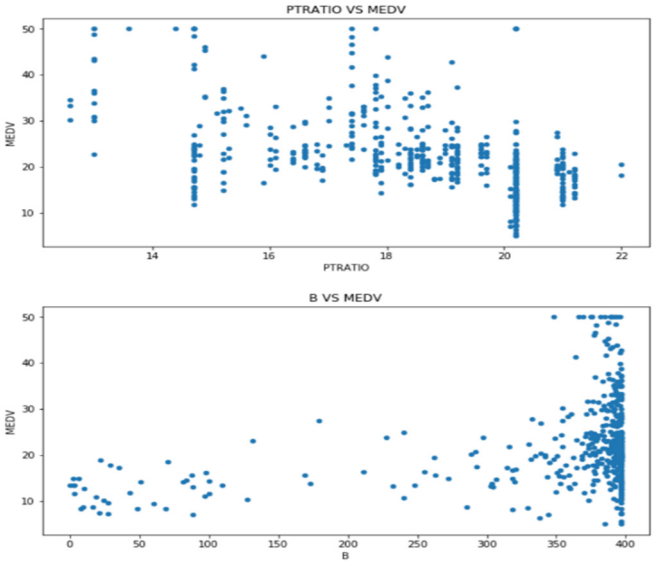

Fig. 9. Scatter plot of variable relationship.5

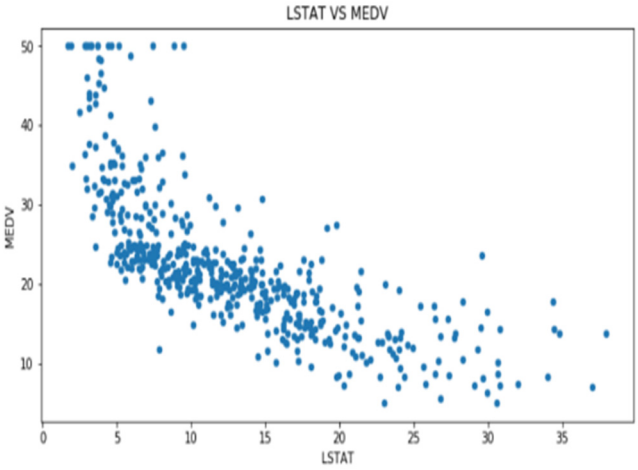

Fig. 10. Scatter plot of variable relationship.6

The scatter plot is used to roughly determine whether two variables are positively correlated, negatively correlated, or not directly correlated. Box plots are used to show the distribution of data for each category on the x-axis for each successive predictor on the y-axis. The continuous variable has no impact on the target variable if the distributions for each category are similar. As a result, there is no correlation between the variables. On the other hand, if each category's distribution is unique. This implies that these factors could be connected to MEDV.

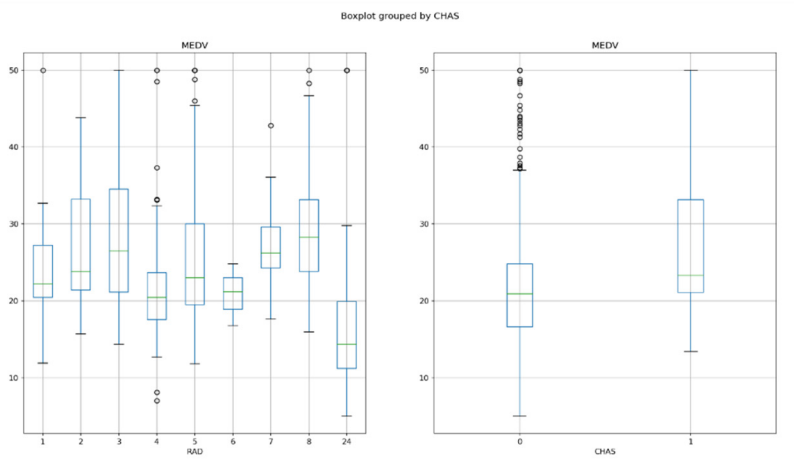

Fig. 11. Box plan.

The ANOVA results support our visual analysis presented above using box plots. The Target variable and all category variables are associated. This assumption is based on the box plots. final choice 'RAD' and 'CHAS' are categorical columns. (see Figure 11)

Choosing the final columns for machine learning based on the tests mentioned above. The data is divided into training samples and test samples; the model using the entire set of data is not built. To evaluate the model's accuracy, some randomly selected data is set aside. The remaining data is referred to as the training data on which the model is developed, and this is referred to as the test data. Usually, 70% of the data is utilized for training, while 30% is used for testing.

Then multi-objective linear regression is used to analyze and predict this set of data, and obtained the following results (see Table 1):

Linear Regression ()

R2 Value: 0.6980461431155771

Table 1. Model Validation and Accuracy Calculations.

	RM	PTRATIO	LSTAT	RAD	CHAS	MEDV	Predicted MEDV
0	0.547040	0.425532	0.201711	0.173913	0.0	23.6	27.0
1	0.612569	0.531915	0.049669	0.130435	1.0	32.4	35.0
2	0.464074	0.797872	0.450883	0.130435	0.0	13.6	16.0
3	0.479785	0.702128	0.104581	0.130435	0.0	22.8	25.0
4	0.524238	0.808511	0.428808	1.000000	0.0	16.1	18.0

Mean Accuracy on test data: 81.80618534324392

Median Accuracy on test data: 88.88888888888889

Accuracy values for 10-fold Cross Validation:

[88.78472351 90.01462311 82.93367348 82.14573469 86.66783537 88.63411206 84.88621233 48.83600797 51.6418211 85.29449921]

Final Average Accuracy of the model: 78.98

4 Conclusion

Multiple linear regression is an important part of multivariate statistical analysis. The method has been widely used in the research of society, economy, technology and many fields of natural science [8]. Our purpose of this investigation is to discover the application of linear regression and apply it to specific data sets. This work has discovered the difference between linear regression and multiple linear regression, and choose multiple linear regression to analyze our data set. In this investigation, the different methodologies of linear regression including Lasso and Ridge linear regression are also investigated. By investigating these two methods, the approach to understand these methods is their application in dataset analysis. This work also brings a deeper understanding of linear regression and the different kinds of linear regressions and their approaches. During the investigation, code is chosen to use with linear programming to analyze our data set and applicate linear regression. Some machine learning knowledge is also learned [9]. In this investigation, our team worked together to form the report of this investigation. It is expected to discover deeper into linear regression and linear programming and find out its application in a wider range of real-life situations [10].

Reference

1. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. 2021. Introduction to linear regression analysis. John Wiley & Sons
2. Weisberg, & Sanford. 2013. Applied linear regression /-4th ed. Wiley-Interscience.
3. Olive, D. J. 2003. Linear Regression Analysis (2nd ed.). World Scientific.
4. Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. . 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 6.
5. Karas, Peter. 2023. “L1 (LASSO) and L2 (Ridge) Regularizations in Linear Regression.” https://medium.com/ai-in-plain-english/l1-and-l2-regularization-lasso-and-ridge-in-linear-regression-a83b6fe07bf8
6. Gong J C. 2008. Application of Lasso and its related methods in model selection of generalized linear models.
7. Yang N. 2004. The unique role of Ridge regression analysis in solving multicollinearity problems. Statistics and Decision, 000(003), 14-15.
8. Zhang R T. 2019. Improvement of housing price forecasting model based on multiple linear regression. Electronic production (4), 3.
9. Li X L. 2020. Housing price prediction based on multiple linear regression model.] Marketing Week: Business Marketing, 000(066), P.1-2.
10. Lu Zixiang, Huang Jiashan, Tu Liyang, XU Xijia, & Zhang Daoqiang. 2018. Regularized multilinear regression algorithm based on tensor and its application. Computer science and exploration.