Boston Housing Data Statistical Analysis

Surbhi Rathore

1 Introduction

This project aims to analyze factors affecting the domestic housing property value in Boston dataset. The project evaluates and analyzes the performance and predictive power of multiple linear regression model and linear regression model trained and tested on data collected from houses in Boston using variable selection methods. The analyses are followed by correlation between the variables and target variable, shrinkage approach for getting the best fit model possible.

2 Data Description

The Boston dataset used in this project comes from the MASS package. This data was collected in 1978 and each of the 506 entries represents information about 14 features of homes from various suburbs located in Boston. The features can be summarized as follows:

CRIM: per capita crime rate by town, ZN: proportion of residential land zoned for lots larger than 25,000 sq.ft., INDUS: proportion of non-retail business acres per town., CHAS: Charles River dummy variable, NOX: nitric oxides concentration, RM: average number of rooms per dwelling, AGE: proportion of owner-occupied units built prior to 1940, DIS: weighted distances to five Boston employment centers, RAD: index of accessibility to radial highways, TAX: property-tax rate per 10,000 dollar, PTRATIO: pupil-teacher ratio by town, Black: proportion of people of African American descent by town, LSTAT: lower status of the population, MEDV: median value of owner-occupied homes in 1000 dollar

3 Questions of Interest

The main objective of the project is to address the following questions:

- What are the major factors associated with the housing value?
- Similarly, how much effect do predictors like nox, lstat etc. have on the target variable?
- What observations can be dropped, for getting a better fit model?

4 Statistical Data Analysis

4.1 Data Summary

We can observer with the summary that variables ‘crim’ and ‘black’ take wide range of values. Variables ‘crim’, ‘zn’, ‘rm’ and ‘black’ have a large difference between their median and mean which indicates lot of outliers in respective variables.

4.2 Outlier visualisation

Figure 1 shows that variables ‘crim’, ‘zn’, ‘rm’ and ‘black’ do have outliers.

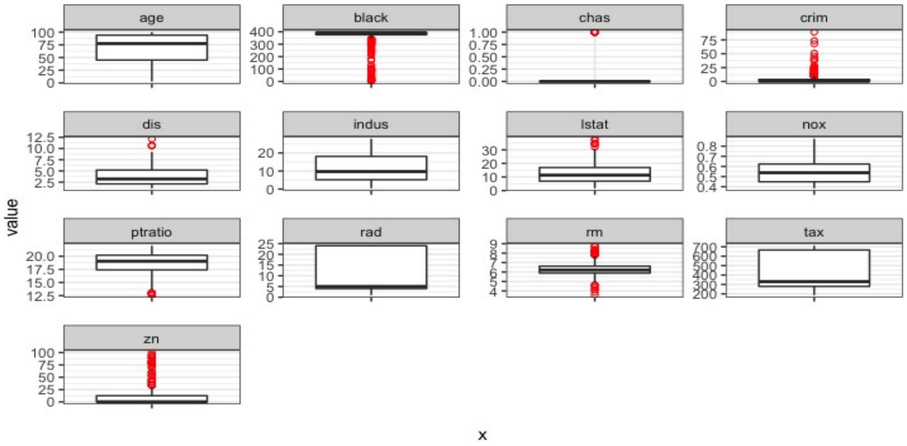

Figure 1: Boxplot plot for outlier detection

4.3 Correlation

Correlation is a statistical term which in common usage refers to how close two variables are to having a linear relationship with each other. Figure 2 shows the level of interaction between the variables. Figure 2 observations: medv

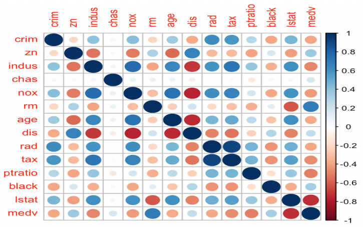

Figure 2: Correlation between variables

increases as rm increased. nox increases as indus increases, both rad and tax has strong correlation. crim is strongly associated with variables rad and tax which implies as accessibility to radial highways increases, per capita crime rate increases. indus has strong positive correlation with nox, which supports the notion that nitrogen oxides concentration is high in industrial areas.

4.4 Pearson Correlation

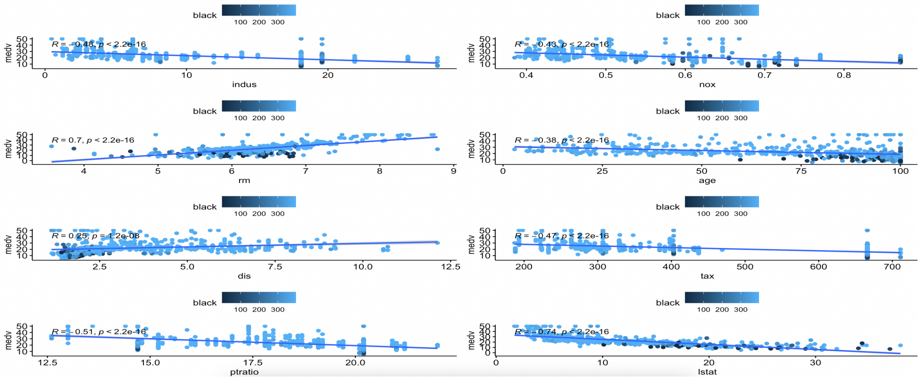

Figure 3: Pearson Correlation

In order to go ahead with liner model implementation, I checked for pearson correlation between the variables, and as shown in the Figure 3, relationship is linear but we are dealing with nonlinear association between the two variables too.

4.5 Splitting the Data for Linear Regression

Splitting the Data into train and test data set in 80:20 ratio. Summary of linear model shows that 'age' and 'indus' have very high p-value and low significance. F-statistic is 96.82, which is not very significant number, variable selection can further be implemented for better model output.

4.6 Subset Selection

Subset selection techniques for variable selection is followed by three methods listed below:

- Forward Variable selection : with forward selection method, we keep on adding influential variables to the model, lstat, rm and ptration are the most significant variables as per the results. Model with 11 variables gives the highest Adjusted R-squared value and the lowest AIC, BIC and GCV value as shown in Figure 4. We find the best model is the model with all variables except age and indus.

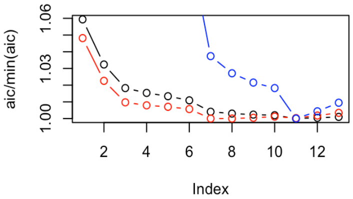

Figure 4: AIC, BIC, GCV plot

- Backward Variable Selection : with backward selection, we keep on removing non-influential variables from the model. lstat, rm and ptration are the most significant variables. Checking the 13 models with varying variable size, we plot the model metrics to find out the best model. R-squared keeps on increasing with added variables and hence will always favor model with highest number of variables. Model with 11 variables gives the highest Adjusted R-squared value and the lowest AIC and BIC values and GCV values. This result is consistent with the forward selection result.

- Exhaustive Variable Selection : exhaustive search cannot be applied to models with a large number of explanatory variables as it works on the approach of brute force. While applying exhaustive search on boston train data set, model with one best explanatory variable contained lstat variable and model with two best explanatory variable consisted of lstat and rm. Finally we got the model with 11 explanatory variables with exhaustive search as best model fit on the basis on minimun values of AIC, BIC and GCV.

Using the subset selection techniques, we select the model without indus and age as our best model. P value is statistically significant for all the variables and model has a significant F statistic of 112.6.

Mean squared prediction error, MSPE can be used on test data for model validation. It is the process of looking at how well a model has performed. By checking the MSPE on boston test data set, we get the MSPE as 25.79744.

Residual analysis Figure 5 of the selected subset model observations:

The variance is not completely constant and hence the assumption of constant variance is not totally satisfied. From the q-q plot we see that it is not completely normal and a little skewed towards the right. There is no autocorrelation observed in the model. There are no observed outliers.

4.7 Shrinkage

- Lasso Regression : Lasso regression does both shrinkage and feature selection. If there exist a group of highly correlated variables which are causing multicollinearity, lasso selects one variable from the group and

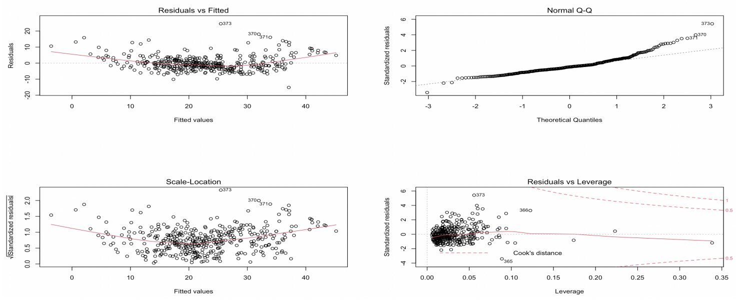

Figure 5: Residual analysis

ignores others. Lasso regression shrinks regression coefficients with some shrunk to zero.

with the implementation of lasso, we try to shrink the coefficient estimates of non-significant variables to zero. Lambda is the penalty factor which helps in variable selection and so higher the lambda, lesser will be the significant variables included in the model.

We fit the LASSO model to our data. From Figure 6, we can observe that as the value of lambda keeps on increasing, the coefficients for the variables tend to 0.

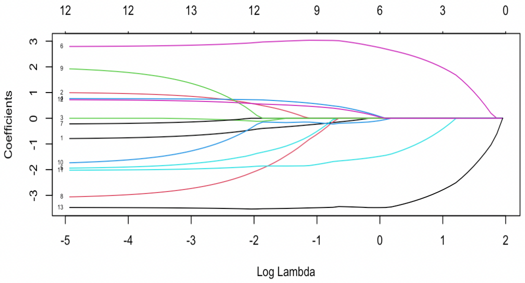

Figure 6: Lasso Plot

Using cross-validation we can find the appropriate lambda value using error versus lambda plot. value with the least error as well as the error value which is one standard deviation away from the lowest error value is considered, a models is built on the basis of both of these. For the higher error value, the number of variables selected decreases.

For model with lambda=min, coefficients of age and indus get reduced to zero. For model with lambda=1se, coefficients of indus, age, rad and tax get reduced to zero, Figure 7

Since lasso also help in variable selection, it will be interesting to observe

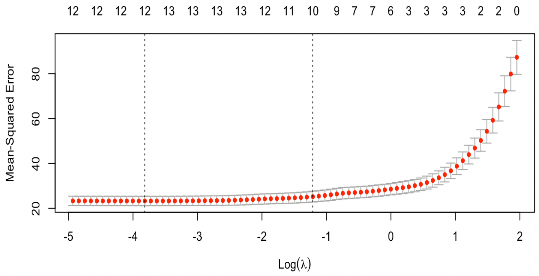

Figure 7: Lasso Model Plot

the comparison between all the three subset selection and lasso with full model.

MSE: MSE of all models stays around 23, except the LASSO model which gives a MSE of 26.39. R-Squared: Full model performs best in this category as expected, and the LASSO model performs the worst. Adjusted R-squared: A better metric for comparing models of different variable sizes, Subset selection model performs the best. Test MSPE: LASSO model performs the best here with a low MSPE of 18.88, while other models also do a pretty good job with scores around the 19.

- Ridge Regression : Ridge regression shrinks the coefficient to non zero values to prevent over fitting but keeps all variables. Lambda is a hyper parameter in ridge which is estimated using cross validation. It is the strength of penalty on the coefficients, which means if we increase the lambda we increase the penalty and if we decrease the lambda we tend to decrease the penalty. In Figure 8, on the y axis we have root mean squared error. With the plot we can see that for higher values of lambda the RMSE increases, hence the best value that we get for lambda is close to 0 between 0.4 to 0.6.

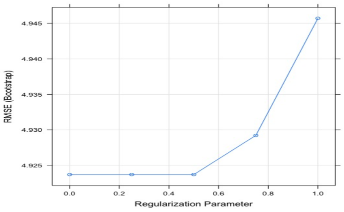

Figure 8: Regularization paramete

In Figure 9, when log lambda is around 9 or 10 all the coefficient is close to0, as the lambda is relaxed the coefficient starts growing, at every point all the 13 variables are considered in ridge regression and the coefficient keeps growing as lambda decreases, thus, increasing lambda reduces the size of coefficient but doesn't make coefficient 0 for variables which doesn't contribute in a major way.

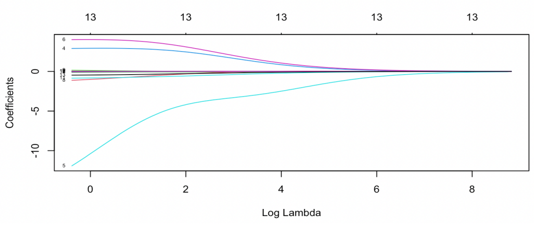

Figure 9: Ridge plot, Lambda vs Coefficient

In Figure 10, nox is the most important variable followed by rm, variable in bottom are the least significant variables, which includes - age, tax and so on.

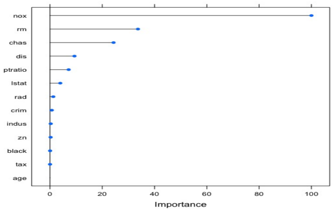

Figure 10: Variable importance

5 Limitation and Potential Extension

I believe exploring the data the other way may help in strengthening the current results or may help in exploring a different narrative associated to it. The Spearman rank-order correlation coefficient (Spearman's correlation, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. Hence, an extension to the project would be to explore the spearman correlation and find related results to strengthen the experiment.

References

⟨https://cran.r-project.org/web/packages/MASS/MASS.pdf⟩ MASS Package.

⟨https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155⟩ Towardsdatascience Article.

⟨http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r⟩ Correlation.

⟨https://bookdown.org/egarpor/PM-UC3M/lm-ii-modsel.html⟩ Informative ebook.