Gurjus Singh

MSDS 422 – Practical Machine Learning September 27th, 2020

Evaluating Regression Models Assignment # 2

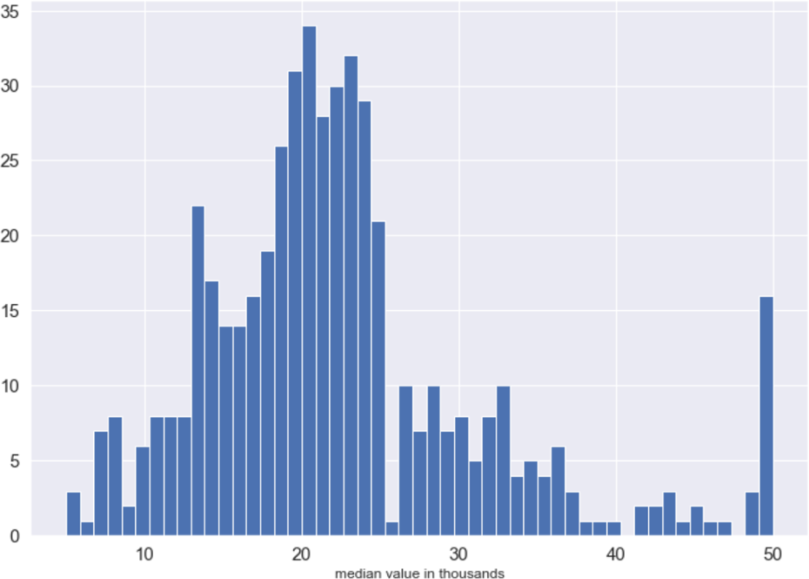
# Data preparation, exploration, visualization

I first started out by loading the dataset “Boston.csv”. I first wanted to get an initial look and feel for the data set. So, I looked at the size of the data set which was (512, 14). This means there are 512 rows and 14 columns mainly features and one variable we want to predict. After getting a feel for the size of the dataset I wanted to check out the head of data frame and types for every column. I saw non-numeric row, which we do not want to use in linear regression, so I dropped this feature column. Most of the columns were mainly floats, except for two columns which contained integer values. I used .describe() value and .isnull.sum() to find out whether the initial data had any typos, or missing value. From looking at min-max category in describe table I did not see anything obviously wrong, so I decided to continue. I also did not see any columns with NA values which is what isnull.sum() does. This counts up all the values which evaluate to True in a column.



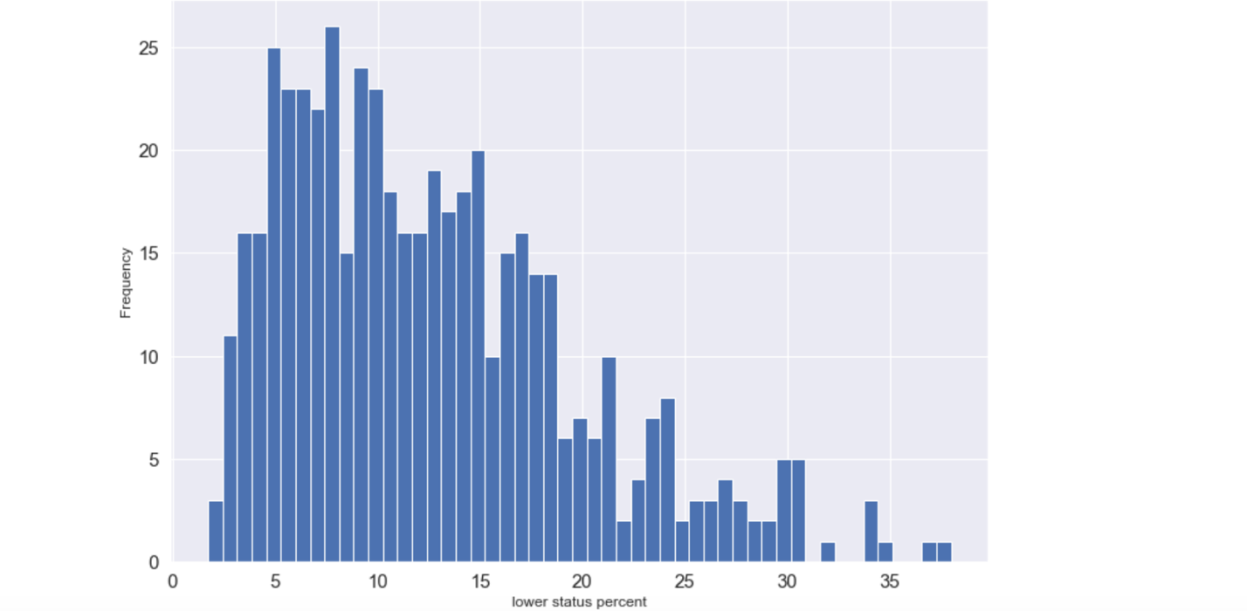
*Boxplot of Columns (Plot 1-1)*

I then decided to go to exploring through data visualizations of the data set. I first tried to boxplot as depicted in Plot 1-1. In the plot I could see many of our columns had an abundance of outliers such as our target variable MV, lstat, dis, rooms, chas, zn and crim. I wanted to play around more and look closely at some of the variables, so I decided to see distributions through histograms.

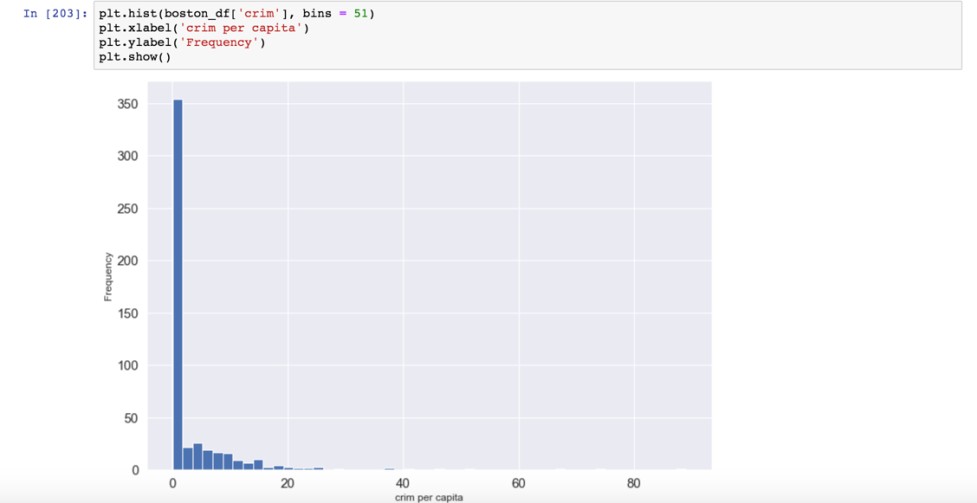


*Histogram Plot 1-2*

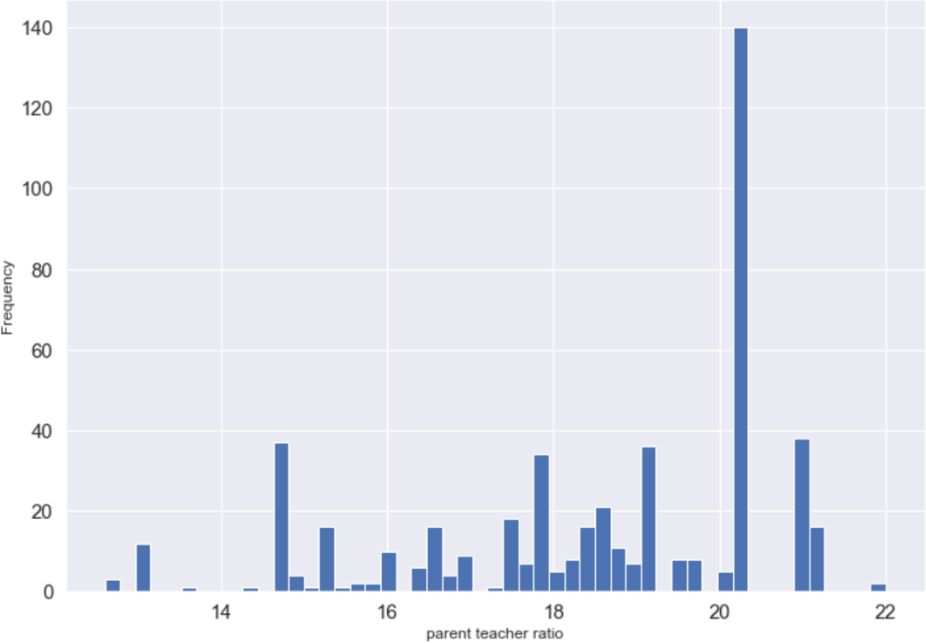
In plot 1-2, it is the median value of the houses, which we will be predicting on this using the independent variables. This looks like a less skewed distribution in comparison to doing the boxplot. I also wanted to look at features I thought were more important to predicting this dataset such as crim, which is the per capita crime rate per town, ptratio which is pupil to teacher ratio by town and lstat which is percent of lower status in an area.



*Histogram Plot 1-3*

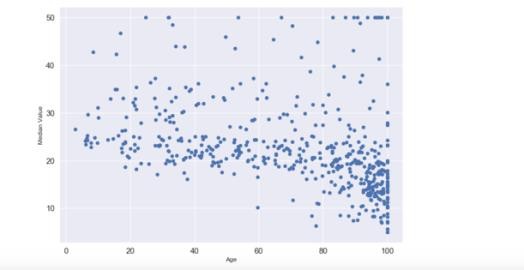


*Histogram 1-4*

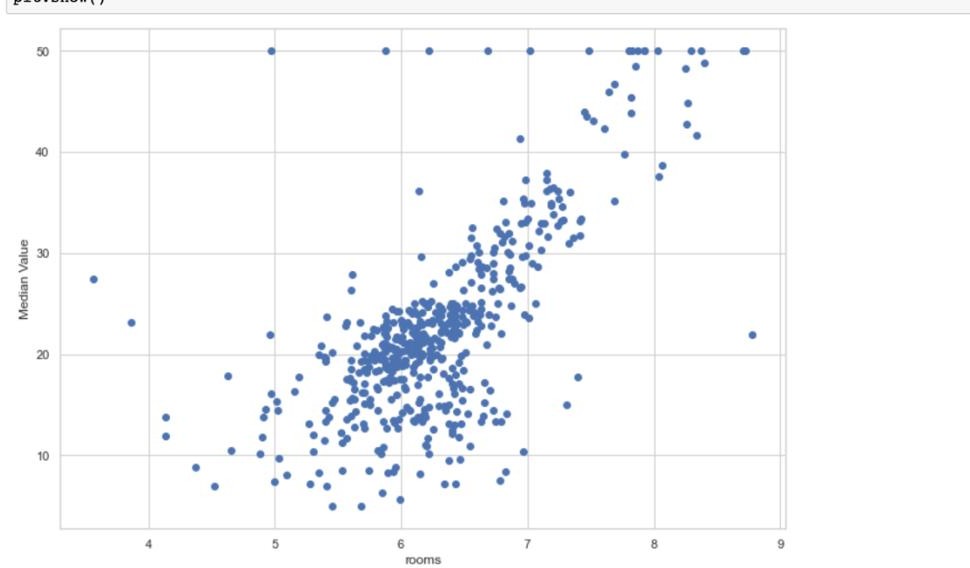


*Histogram 1-5*

The lstat variable in 1-3 seems to be skewed to the right where 10% in a given neighborhood or region are mostly of lower status. In 1-4 the histogram is right skewed as well with most of the crime happening near 0 in the areas in Boston where the data is collected. In Histogram 1-5, we see for ptratio, is near 20 which mean the difference in teacher to student ratio is high, this data is more skewed to the left. Lastly, I wanted to make some plots to see any initial correlation between MV, and these variables. I also wanted to see if age was in any way correlated.



*Dot plot of Age variable compared to MV Plot 1-6*



*Figure 1-7 of highly linear ‘rooms’ variable*

In plot 1-6, I do not see any linear shape compared to age, which suggests there is no relationship in MV, and the age variable. I also looked at crim variable which did not seem to have a linear shape as well. During my initial findings I also saw that rooms and lstat variables seemed to have a linear shape which means they were highly positive or negatively correlated as seen in figure 1-7.

After examining linear shaped data and nonlinear data, I then wanted to scale the features between 0 and 1 as it helps speed up the learning process. In order to make the data more linear I had to use the lambda function which helps to map the function to all the data in one line of code. I did not want any zeros in my data, so I added .01 to every data point. I then used “boxcox” which is great for making sure my data is more normalized and more linear to use linear models on it.

Before modeling using the linear regression methods, I first wanted to create a correlation heatmap to get an overall picture of which features in the data set might be more useful to predict median value price of a neighborhood of houses. This heatmap showed exactly what might initial

findings confirmed in that lstat was negatively correlated, and rooms was highly positively correlated as shown in figure 1-8 below.



*Figure 1-8 Heatmap*

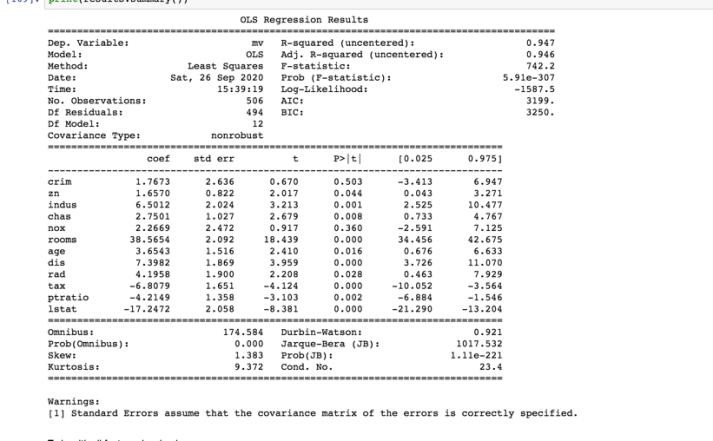
# Review research design and modeling methods

After we choose, the features I am going to use 4 modeling methods which are main linear regression learning algorithms but have some variations. The four types are Regular Linear Regression, Ridge Regression, Lasso Regression and Elastic Net Regression. The differences between the four is that Linear Regression does not add regularization term which is important in most cases to prevent overfitting [1]. Ridge Regression does incorporate regularization but is different than Lasso Regression in terms of the cost function. ElasticNet Regression is different in a way because it provides a median between both Ridge Regression and Lasso Regression [1]. I think it is important to try a range of algorithms because one might work better than the other in terms of how much it overfits which is what we do not want when using training set. We want the learning algorithm to generalize rather than to remember the data.

Before we start training the algorithm, I had to save the target variable in its own data frame to separate it from the features. In order to use the regression methods, it is important to split up the data into train data, and test data. I did the 80-20 split on the data. Our goal was to make sure that the data trained and predicted well on the test set.

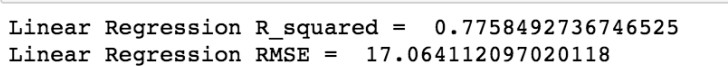
# Review Results, and Evaluate Model

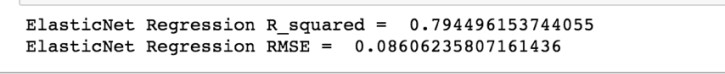
I was expecting to find Elastic Net would be the best at prediction as that is what I found out in Geron’s book [1]. I first used all the features and did not split the data initially. I used Ordinary Least Squares Regression which is a type of linear regression which involves finding the minimization of the sum of squared differences between predicted and actual value [2]. I found the model was overfitting with and R2 of around 0.947 so it would not generalize to new data as depicted in Table 1-9.



*Table 1-9*

I then split the data as I mentioned in the previous section and did not remove features. I wanted to see how the models performed on the test set. What I found was quite astonishing. I found that Ridge Regression performed slightly better than ElasticNet which was surprising. I thought Lasso Regression would also perform better but that was not the case as Ridge Regression was also better.





*Results 1-10*

I then tried taking out p-values greater than 0.05 as suggested by professor and shown in Table 1-9. I only found crim variable was above 0.05, and the assignment goal was to predict using nox variable. I saw that the models worsened. This time the best model was Regular Linear Regression. I then wanted to take a peak back at the table in 1-9 to find out if I could find anything other findings. I tried to take out values as shown in Code 1-12. The results got progressively worse. With the best model being yet again Linear Regression. So, in a way, I think keeping all the features was the best way to go about this predicting median values in housing prices.





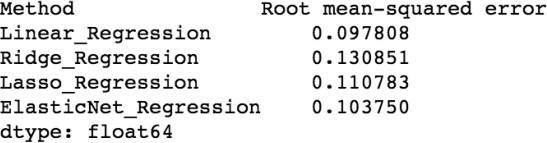


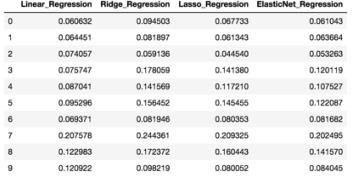


*Results 1-11 taking out p-values greater than 0.05 except for nox*



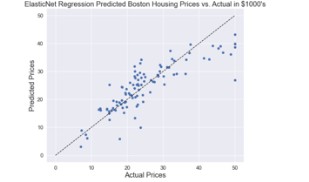
*Code 1-12*

After seeing different data models and how they ran, I wanted to see cross validation and how it ran. Out of this I think the best models were Lasso Regression and Elastic Net which is what Geron’s confirmed would be overall the best models to use. I think Linear Regression was overfitting on the validation set which had the lowest error, while Ridge Regression seemed to be too high.



*Results 1-13*



*Regressions plot 1-14*

# Implementation and Programming

Different packages I used for this data analysis insight was sklearn package which is important for doing predicting and creating the different linear regression models. In terms of visualizing the data, this is where seaborn and matplotlib came into play. This a great for

providing different kinds of barplots on categorical variables, scatter plots on more numerical type variables, and histograms to see how skewed the data is. In order to start implementing the models it is important we transform the data using boxcox which helps normalize or linearlize the data, and make sure we do not have any zeroes in our data [4]. So, in our code we have two lambda functions for doing this [3]. We also want to scale using minmax from the sklearn package. This is important for speeding up the learning algorithms. We then run our models, and it is important the results such R2 are not high for the training set and are higher for the test set. We can also use cross fold validation to see if our models accurately predict.

# Exposition, problem description and management recommendations

From examining the results, on the test set, I do believe it is better to drop features as it performs higher around 0.79 for all the models. From looking at the results in 1-10, the ridge regression results are the best, so I would use this model for the data set. This means the effect of pollution is not a good indicator on the price of houses compared to lstat for example. I do believe if we take out nox, which shows the pollution or nitrogen oxide levels, I think the regression models will perform better, but with nox included it does not perform as best as I wanted. For the future, I would like to see if the models perform higher with nox out of the features. I would also like to automize the task of dropping the features such as using stepwise or a different way to drop the features. I hope to learn other EDA mechanisms as well, as I did see a way, I could use any other EDAs in this assignment to explore the data.

References

1. Géron, Chapter 4, & Chapter 5
2. <https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols>
3. <https://www.w3schools.com/python/python_lambda.asp>
4. [https://www.geeksforgeeks.org/box-cox-transformation-using-](https://www.geeksforgeeks.org/box-cox-transformation-using-python/#%3A~%3Atext%3DIn%20short%2C%20trying%20to%20move%2Ctests%20than%20we%20could%20have) [python/#:~:text=In%20short%2C%20trying%20to%20move,tests%20than%20we%20could%20](https://www.geeksforgeeks.org/box-cox-transformation-using-python/#%3A~%3Atext%3DIn%20short%2C%20trying%20to%20move%2Ctests%20than%20we%20could%20have) [have.](https://www.geeksforgeeks.org/box-cox-transformation-using-python/#%3A~%3Atext%3DIn%20short%2C%20trying%20to%20move%2Ctests%20than%20we%20could%20have)