Poornima Andukuri

Introduction:

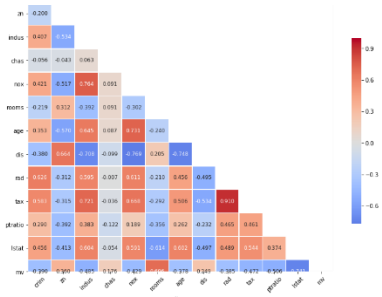
Machine learning can be used to evaluate real estate market to increase sales and profits. It is less time and resource consuming compared to other evaluation techniques. Machine learning regression for residential real estate evaluation will help improve performance.

Exploratory Data Analysis:

The data provided has 506 rows with 13 variables that describe the neighborhood, number of rooms, zone, distance, crime rate and more. Variables like number of rooms, distance, crime rate affect the market. Neighborhood is a categorical attributed and is not required. There were missing values and no outliers.

The correlation matrix shows that the response variable is most influenced by the variable rooms. This indicates that the highest value homes are bigger homes in specific neighborhoods.

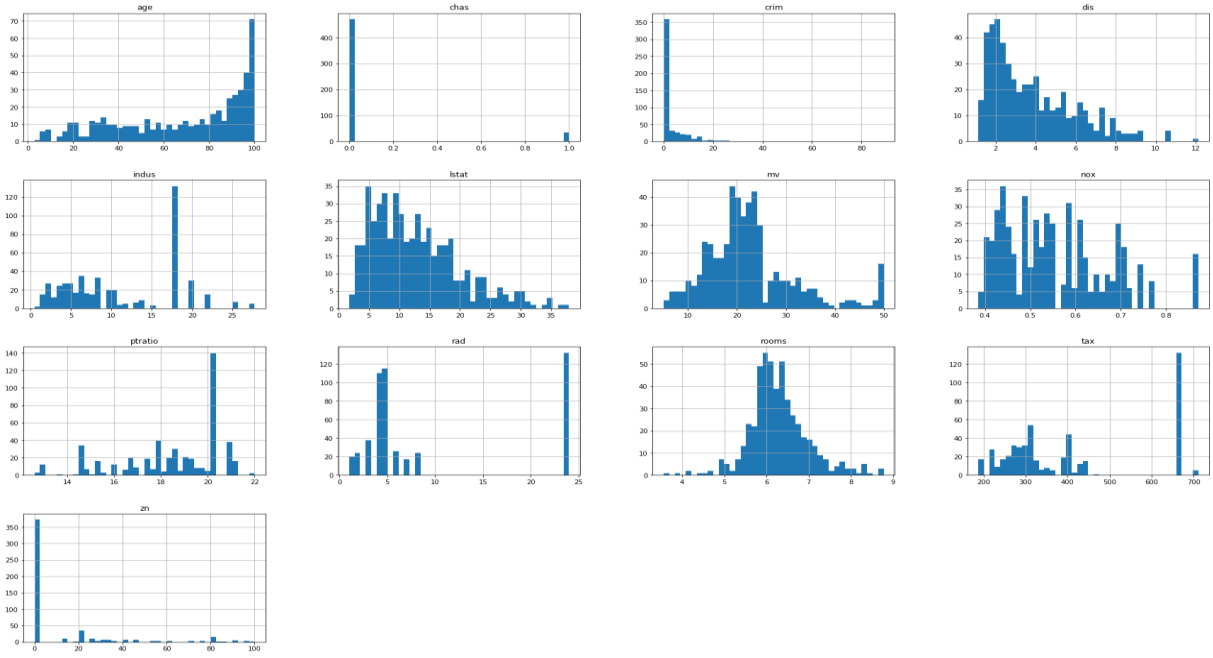
Figure 1: Correlation matrix to response variable 'mv'



The histograms below are the distribution plots for all variables.

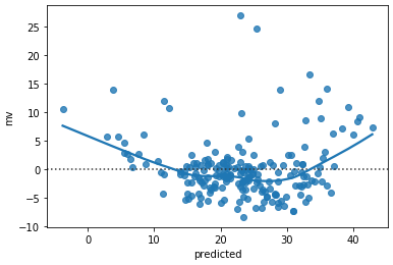
All variables may require scaling as they are not normally distributed

Figure 2: Linear Plot for variables to response variable

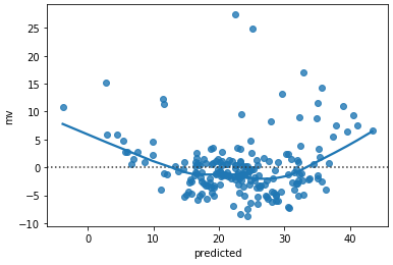


Programming:

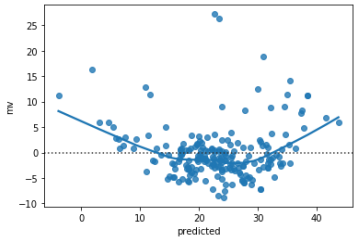
The four regression models used for this analysis are Linear, Ridge, Lasso and ElasticNet. The python packages used for data handling are Pandas, Number, SciKitLearn. The variable ‘neighborhood’ was removed to create a subset of variables and response variable. The model data was creating by using StandardScaler(). The SciKit Learn function train_test_split was used to split the data by the test size. squared error metric



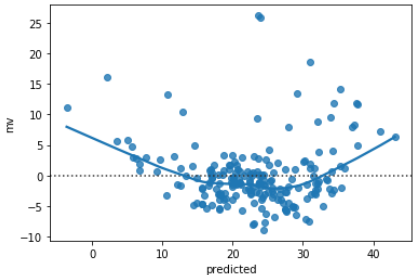
Linear Regression



Rigde



Lasso



ElasticNet

For every regression model the model score and root mean squared error metric for performance evaluation was calculated.

	Model	RMSE	Score
0	Linear	5.107567	71.196569
1	Ridge	5.151525	70.698646
2	Lasso	5.263183	69.414679
3	ElasticNet	5.271727	69.315297

Root mean squared Error Metric and Model Score

Insights & Conclusion:

The Linear model performed the best with respect to the root mean squared error metric, then Ridge regression, ElasticNet and Lasso in the order mentioned. I recommend that Linear regression model be used for real estate market evaluation.

Appendix:

The ipynb notebook and an html version of the notebook along with the output and graphs are included in the submission.