NORTHWESTERN
UNIVERSITY

SCHOOL OF
PROFESSIONAL
STUDIES

Boston Housing Study

Ranaa Ansari, Daniel Arenson, Yining Feng, Han Nguyen

Northwestern University

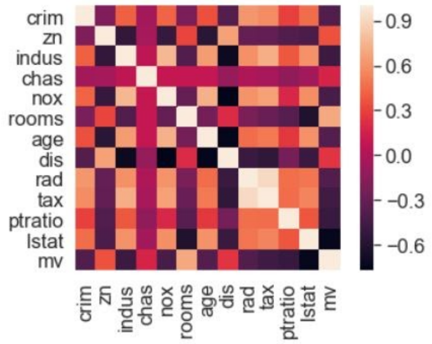
MSDS-422 Assignment 2: Evaluating Regression Models

Summary and problem definition for management

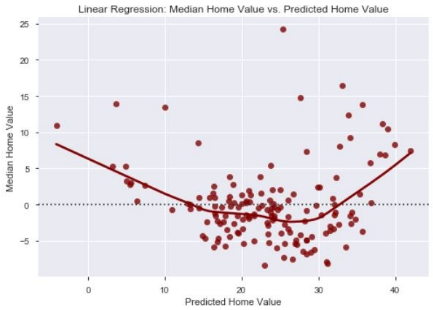
We will be exploring the data and implications of the Boston Housing Study. This data set was originally established to examine the correlation between air pollution and the value of housing prices in Boston. It includes data for 14 different variables that were used to control for and isolate the impact of air pollution, which was measured in nitrogen oxide concentration. However, our group’s focus will not focus only on the impact of pollution. Instead, we will be examining all of the variables to summarize the most important drivers of housing price. For this project, our goal is to develop a model that predicts home value in the Boston metropolitan area, and provide a recommendation on which machine learning modeling methods to use.

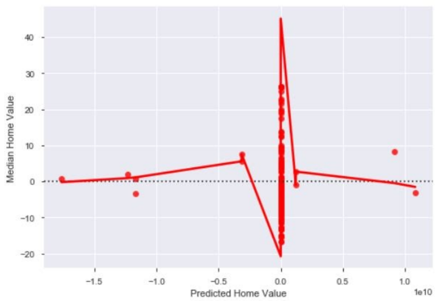
Research Design

The Boston Housing dataset consisted of 506 census tracts in the Boston metropolitan area, with each row describing a Boston town or suburb. The objective of the research was to examine 12 given attributes (features) that describe a house in Boston, and construct regression models which have the capability of predicting the median price of houses.



To fit a linear regression model, we selected those attributes which have a high correlation with our target variable `mv`. By looking at the correlation matrix on the right hand side, we could see that the feature `rooms` has a strong positive correlation with `mv`(0.70) whereas `lsata` has a high negative correlation with `mv`(-0.74). In addition, the feature `crim` has a moderate negative correlation with `mv`(-0.39). Another important point in selecting features for a linear regression model is to check for multicollinearity. For instance, features `rad` and `tax` have a correlation of 0.91. This pair of features is therefore strongly correlated to each other, and they are excluded from the training dataset of the model. Based on the linear regression model using only numeric features, the median price of houses increases as the value of `rooms` increases linearly, whereas the median price of houses inclines to decrease with an increase in the variable `lstat`. However, the linear regression model excluding the `neighborhood` categorical attribute has low accuracy and is a poor fit for predicting the median price of houses. On the other hand, we detected overfitting in the same linear regression model with the `neighborhood` variable converted into a “dummy” quantitative variable.





In order to avoid the overfitting issue, Lasso regression model was applied to prevent overemphasis of coefficients and eliminate any feature with insignificant coefficients. Ridge regression model was utilized to minimize the impact of nonessential features in predicting the target variable, whereas the Elastic Net model was used to combine feature elimination from Lasso and feature coefficient reduction from the Ridge model to improve the accuracy of model predictions. (Swamy, 2018)

Next, we calculated the coefficient of determination R^2 to quantify our model's performance since R^2 whose values range from 0 to 1, is a useful statistic that describes how "good" a model is at making predictions. The values for R^2 capture the percentage of squared correlation between the predicted and actual values of the target variable. A model with an R^2 of 0 fails to predict the target variable, whereas a model with an R^2 of 1 perfectly predicts the target variable. Any value between 0 and 1 indicates what percentage of the target variable using our regression models can be explained by selected features. (Grace-Martin, 2020)

On the other hand, the Root Mean Squared Error (RMSE) was used to illustrate how close the observed data points are to the model's predicted values. RMSE represents the standard deviation of the unexplained variance. Lower values of RMSE indicate better fit of the model. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction. (Grace-Martin, 2020)

Programming Work

Initial Exploratory Data Analysis was performed by taking the Boston Housing dataset and analyzing the data based on the distribution of 13 given attributes (features) that describe a house in Boston. Then a correlation matrix was used to view which attributes demonstrated a strong correlation with the median price of houses. To validate this, a linear regression was conducted on each selected attribute by splitting the data into training and testing datasets so that the model can be trained and tested on different data. The model was trained with 67% (337 records) of the sample data and tested with the remaining 33% (169 records) by splitting the data with the `train_test_split` function and training the model with the function `LinearRegression` from the *scikit-learn* library. Nevertheless, we discovered overfitting in the linear regression model and Ridge, Lasso, and Elastic Net regression models were employed to help reduce the model complexity and multicollinearity with `Ridge`, `Lasso`, and `ElasticNet` functions from the *scikit-learn* library. Last but not least, we evaluated the model performance based on the Root Mean Squared Error (RMSE) and R^2 with `metrics.mean_squared_error()`and `score()` functions from the *scikit-learn* library and the `sqrt()` function from the *math* library.

Results and Recommendations

Four regression models were conducted to regularize the Boston Housing dataset to predict the median home value in the Boston metropolitan area. Of the four regression models conducted, the linear regression method resulted in an overfitted model; the remaining three methods (Lasso, Ridge, and Elastic Net regression) required a reduction of features to account for overfitting in the data and yielded more accurate predictions.

The root mean squared error (RMSE) value for training and testing data are comparable for the Lasso, Ridge, and Elastic Net methods which suggests no overfitting occured in these three regression models. Based on the analysis, the Ridge regression model yielded the least RMSE value for testing data (refer to the table below), indicating that the model would most accurately predict Boston home value of the four regression models.

Regression Models	# Features Used	**R^2**		**RMSE**	
		Training	*Testing*	*Training*	*Testing*
Linear Regression	105	0.89	-6.5×10^{16}	5.403	5.502
Lasso Regression	89	0.89	0.85	2.951	3.771
Ridge Regression	98	0.89	0.85	2.947	3.771
Elastic Net Regression	94	0.89	0.85	2.985	3.842

For that reason, the recommendation is to employ machine learning methods to better predict housing value, more specifically the Ridge Regression method.

References

Grace-Martin, K. (2020, January 16). Assessing the Fit of Regression Models. Retrieved April 18, 2020, from https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/

Swamy, V. (2018, October 16). Lasso Versus Ridge Versus Elastic Net. Retrieved April 18, 2020, From https://medium.com/@vijay.swamy1/lasso-versus-ridge-versus-elastic-net-1d57cfc64b58