NORTHWESTERN
UNIVERSITY

SCHOOL OF
PROFESSIONAL
STUDIES

Boston Housing Study

Ranaa Ansari, Daniel Arenson, Yining Feng, Han Nguyen

Northwestern University

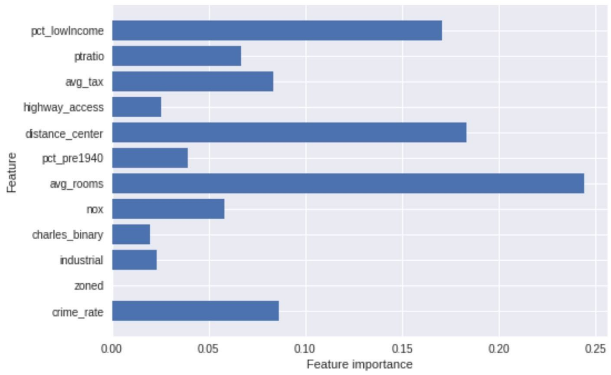
MSDS-422 Assignment 4: Random Forests and Gradient Boosting

Summary and problem definition for management

Assignment 4 builds on the linear regression models we developed for assignment 2. We explored the data and implications of the Boston Housing Study. This dataset consisted of 506 census tracts with 14 given attributes (features) that describe a house in the Boston metropolitan area. In the prior analysis, we investigated three modeling techniques that applied a standard linear regression, a Ridge Regression model and a Lasso Regression model. In the end, the recommendation was to apply a Ridge Regression method for better prediction of housing value. Although the training and test scores fall within an acceptable range, the curved arc of the residual plot suggests that we may need to work on the fit a bit more. In this week's assignment, we will experiment with decision tree, random forest, and gradient boosting methods, and provide a recommendation on which machine learning modeling methods to use.

Research Design

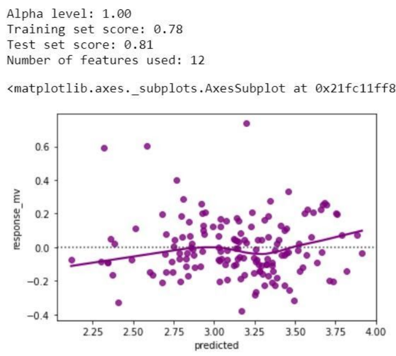
As the dataset is the same as in assignment 2, there were no additional discoveries in the exploratory data analysis. The most important points to recall were that the highest correlation with the response variable `response_mv` observed in the fields: `avg_rooms`, `pct_pre1940` and `pct_lowIncome`. These were the most 'normal' data points with the clearest linear relationship to the response variable. We can observe their relative influence remains high in this week's assignment as well from the plot below, which is the 'feature importance' mapping in the gradient boosting model.



The recommended model applies gradient boosting to determine the best relationship of the variables in a decision tree model. The point to note is that other features like distance to center, crime rate, tax and parent teacher ratio take on more importance relative to other models.

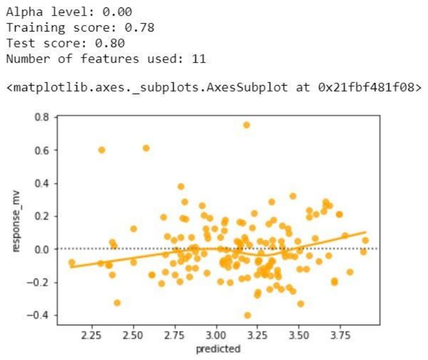
In contrast, the more simplistic linear models applied previously in assignment 2 were not able to get to the training/test scores that we see using the gradient boosting technique. This may be due to the fact that the linear models could not weigh these additional variables without making the model overly sensitive to changes, resulting in overfitting.

In this assignment, we start with the assumption that we will not use the neighborhood classifier and we will employ five methods to the regression problem based on a log transformation of the original data. We will investigate five modeling techniques: a regularized Lasso model, a regularized Ridge model, a decision tree model, a random forest model, and finally a model that uses gradient boosting. The residual plot in yellow is an example of where we start our model exploration—— a scaled Lasso model with light constraints. On the other hand, the residual plot in purple is another example of where we start our model exploration—— a scaled Ridge model with light constraints. The general performance of both models is poor with a training/ test score in the range of .80. As we will see, this falls short of what we are able to achieve using more complex methods, and is only a point of reference to improve upon.



```
Alpha level: 1.00
Training set score: 0.78
Test set score: 0.81
Number of features used: 12

<matplotlib.axes._subplots.AxesSubplot at 0x21fc11ff848>
```



```
Alpha level: 0.00
Training score: 0.78
Test score: 0.80
Number of features used: 11

<matplotlib.axes._subplots.AxesSubplot at 0x21fbf481f08>
```

Programming Work

Initial exploratory data analysis was performed on the Boston Housing dataset by analyzing the distribution of the fourteen given attributes that describe a home in Boston, and correlation between the explanatory variables and mean home value; these were conducted using primarily the *seaborn* and *pandas* library. Distribution plots were used to analyze the distribution of each attribute, and a correlation matrix was used to demonstrate which attributes had a strong correlation with the median home value. To validate the observed correlations, several regression methods from the *scikit-learn* library were conducted on the dataset, including (1) Scaled Lasso Regression, (2) Decision Tree Regression, (3) Random Forest Regression, (4) Gradient Boosting, and (5) Scaled Ridge Regression.

For each regression method, the Boston Housing data was split into training and testing data so the models would be trained and tested on different data. Each regression model was trained with 70% of the Boston Housing dataset and tested with the remaining 30% by splitting the data with the `train_test_split` function and training the model with the following functions from the *scikit-learn* library: `Lasso`, `DecisionTreeRegressor`, `RandomForestReressor`, `GradientBoostingRegressor`, and `Ridge`. The performance of each model was evaluated based on the Root Mean Squared Error (RMSE), R-squared values, and cross-validation accuracy metrics by using the `metrics.mean_squared_error()`, `r2_score()`, and `cross_val_score()` functions from the *scikit-learn* library and the `sqrt()` function from the *math* library.

Results and Recommendation

Scaled Lasso Regression, Scaled Ridge Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting techniques were evaluated on the Boston housing dataset to predict the median home value in the Boston metropolitan area. Based on the results, Gradient Boosting method was most optimal when applied to Lasso Regression and Ridge Ridge Regression models. By limiting the depth to max depth to two, the models did not overfit, and there was a satisfactory balance of emphasis across all the features in the model.

The table below reinforces the recommendation that Gradient Boosting worked best when applied to Lasso Regression and Ridge Regression. Gradient Boosting Ridge Regression had the highest cross validation scores across three, five, and ten fold that averaged out to .995. Gradient Boosting Lasso Regression followed by the second highest average score of .78. In addition, **R**^{2} yielded the highest score for both test and train datasets and also had the lowest RMSE scores, indicating that the models, specifically Gradient Boosting Ridge Recession followed by Gradient Boosting Lasso Regression would be the best recommendations to accurately predict Boston home.

Regression Models	**Cross Validation Accuracy**			**R**^{2}		**RMSE**	
	3 Fold	*5 Fold*	*10 Fold*	*Training*	*Testing*	*Training*	*Testing*
Scaled Lasso Regression	0.767	0.728	0.687	0.78	0.80	0.179	0.175
Decision Tree Lasso Regression	0.723	0.619	0.593	0.92	0.68	0.109	0.223
Random Forest Lasso Regression	0.779	0.776	0.704	0.91	0.85	0.117	0.151
Gradient Boosting Lasso Regression	0.808	0.794	0.737	0.95	0.86	0.088	0.145
Scaled Ridge Regression	0.770	0.741	0.709	0.78	0.81	0.178	0.173
Decision Tree Ridge Regression	0.715	0.641	0.624	0.87	0.66	0.135	0.228
Random Forest Ridge Regression	0.779	0.776	0.704	0.91	0.85	0.117	0.151
Gradient Boosting Ridge Regression	0.991	0.998	0.997	0.94	0.86	0.092	0.145

References

David A. Belsley, Edwin Kuh, and Roy E.Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

Géron, A. 2017. *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol, Calif.: O'Reilly. Source code available at https://github.com/ageron/handson-ml