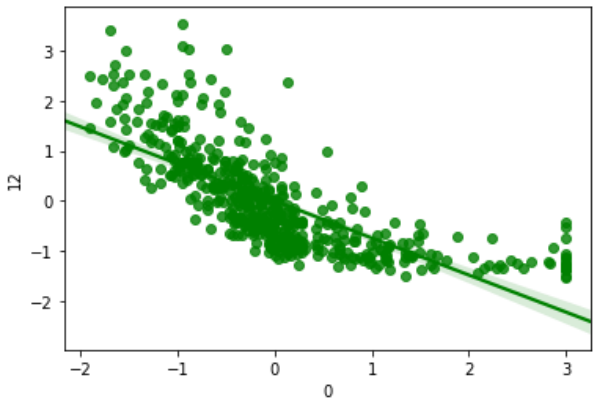
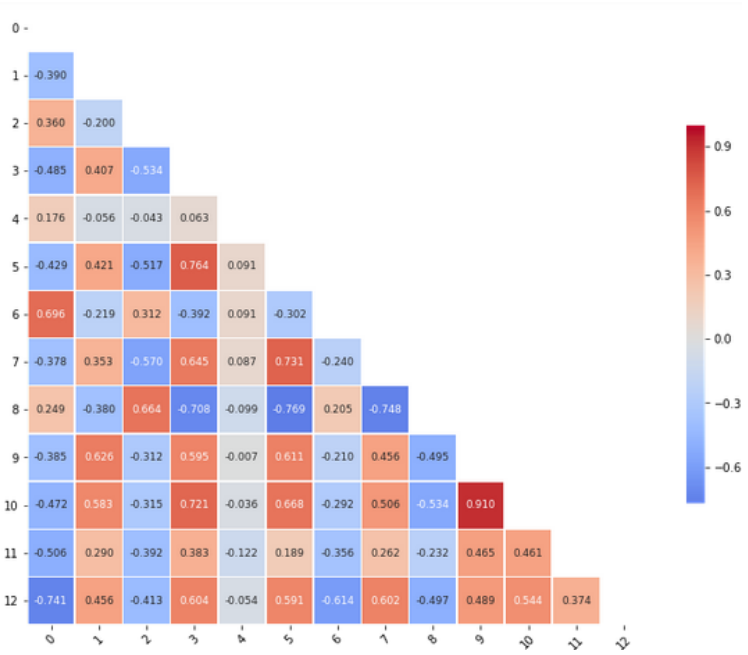
MSDS 422 Assignment 4

In a previous project, regression models were employed to predict the market value of homes in Boston for a real estate firm. In addition to this, random forest regression and gradient boosting regression models were used to further assess the value of the homes. In order to determine which regression model works best for this case, it is important to compare each method by comparing the predictions made on a sample data set. This data contained a list of 506 houses with the following attributes:

Variable Name	*Description*
neighborhood	Name of the Boston neighborhood (location of the census tract)
mv	Median value of homes in thousands of 1970 dollars
nox	Air pollution (nitrogen oxide concentration)
crim	Crime rate
zn	Percent of land zoned for lots
indus	Percent of business that is industrial or nonretail
chas	On the Charles River (1) or not (0)
rooms	Average number of rooms per home
age	Percentage of homes built before 1940
dis	Weighted distance to employment centers
rad	Accessibility to radial highways
tax	Tax rate
ptratio	Pupil/teacher ratio in public schools
lstat	Percentage of population of lower socio-economic status

One can notice that these attributes are on different scales, so it will be important use scaling techniques on the data in order to get the data on a common scale without distorting it for the Linear, Lasso, Ridge, and ElasticNet regression models Each variable was then plotted against the attribute of interest, median value of homes in order to view correlations. As seen by this heat graph and scatter plot, median value(attribute #12) is most correlated with crime rate (attribute #0).

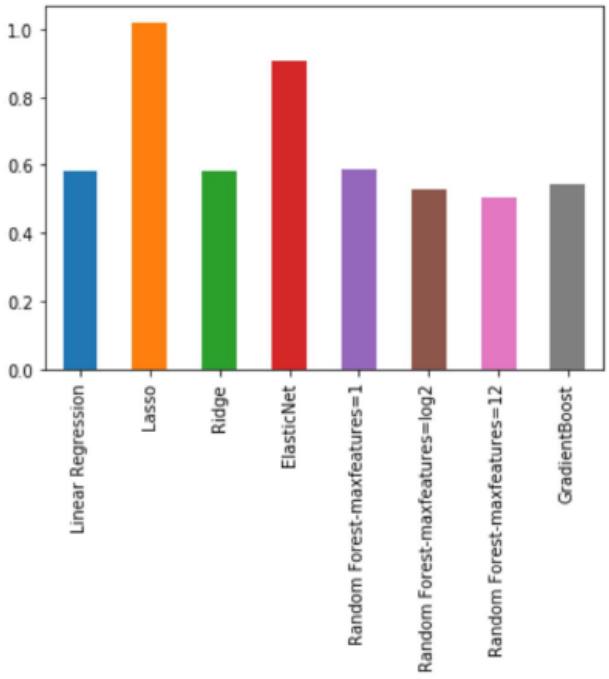




Next, multiple regression methods were compared using the scaled data and having median value as the response variable. Linear regression, Lasso regression, Ridge regression, ElasticNet Regression, Random Forest Regression, and Gradient Boosting Regression models were all evaluated using the Root Mean Square Error (RMSE). Also known as prediction error, this value is an indicator of how far off the regression line is from the test data set. The lower the RMSE value, the closer the regression fits the test data. 10 trials were taken for each of the methods and an average was taken for each model. The results of each regression are the following:

```
*******************************
Average from 10 folds
Method                 Area under ROC Curve
Linear Regression                  0.582659
Lasso                              1.019173
Ridge                              0.582095
ElasticNet                         0.907610
Random Forest-maxfeatures=1        0.589242
Random Forest-maxfeatures=log2     0.529962
Random Forest-maxfeatures=12       0.505435
GradientBoost                      0.542423
dtype: float64

Standard Deviation
Linear Regression                  0.152272
Lasso                              0.274785
Ridge                              0.152927
ElasticNet                         0.267576
Random Forest-maxfeatures=1        0.210765
Random Forest-maxfeatures=log2     0.176464
Random Forest-maxfeatures=12       0.138868
GradientBoost                      0.172814
dtype: float64
```



When recommending to the real estate firm model to choose, I would advise them to choose the Random Forest Regression model with a max feature of log2 or 12. This is due to the fact that they have similar RMSE values are the lowest of any other models. If they were to decide between the two, I would recommend to probably choose the log2 in future analysis since a max feature of 12 can lead to overfitting. Furthermore, when looking into the most important explanatory variables, crime, being on the river, and weighted distance to employment centers had the highest correlation as seen on the heat map. More analysis was completed with looking into performing regression analysis with one explanatory variable at a time. Although high RMSE values were found with each of these, it may also be worthwhile looking into air pollution, number of rooms, and lstat. Ultimately, I believe it is necessary to obtain more data if possible and further compare the different models. With only 506 entries present it is difficult to receive a sufficient analysis of the housing market. The more data present, the more accurate the model will be.