

# INTRODUCTION

The objective of the Boston Housing Study was to examine the effect of air pollution on housing prices, controlling for the effects of other explanatory variables. The response variable is the median price of homes in the census tract.

## SUMMARY AND PROBLEM DEFINITION FOR MANAGEMENT

This project *evaluates the performance and predictive power of a model that has been trained and tested* on data collected from homes in suburbs of Boston, Massachusetts.

## MEASUREMENT AND STATISTICAL METHODS

The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts. To appropriately value Boston Housing *Market Linear, Ridge, Lasso, ElasticNet Regression* modeling techniques were used and *missing value preprocessing and StandardScaler scaling* steps have been made to the dataset.

## EXPLORATORY DATA ANALYSIS METHODS

### Dimensions

Due to the missing values the neighborhood column has been dropped and modified dataset has following dimensions

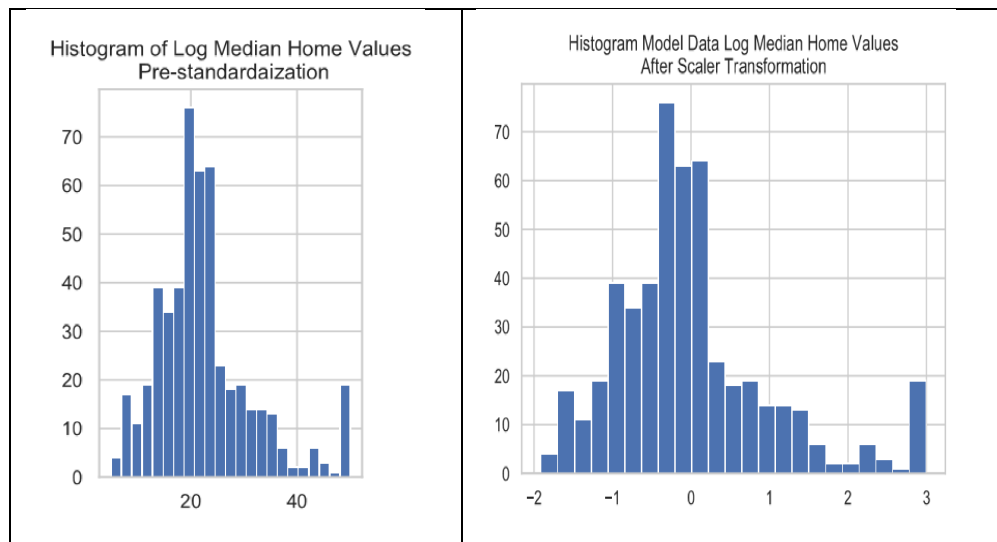
```
dataset dimensions (506, 13)
```

*StandardScaler* scaler has been used to scale entire dataset and after the transformation following basic statics has been captured.

### Dataset Basic Descriptive Statistics

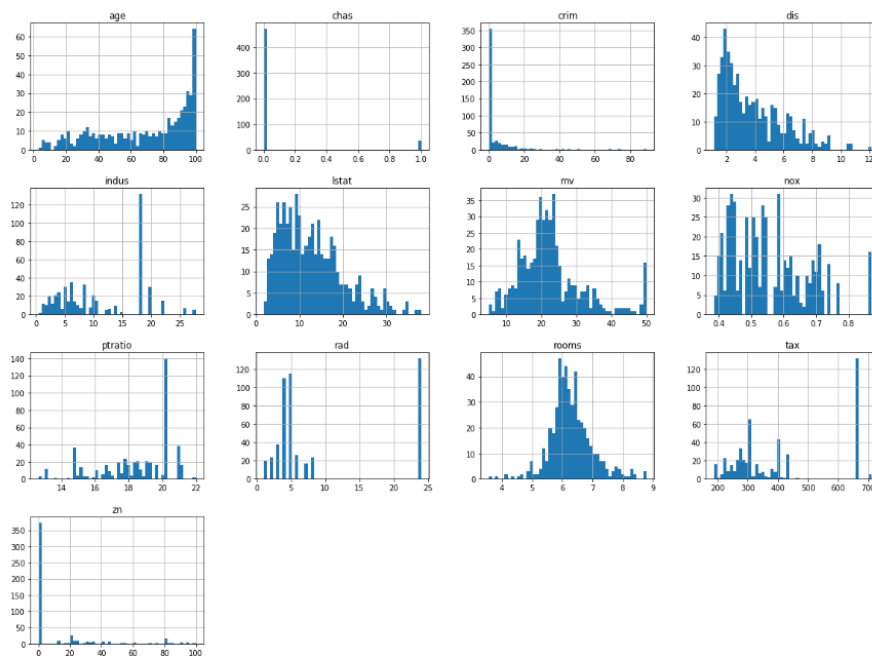
	crim	zn	indus	chas	nox	rooms	age	dis	rad	tax	ptratio	lstat	mv
count	506	506	506	506	506	506	506	506	506	506	506	506	506
mean	3.613524	11.363636	11.136779	0.06917	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	12.653063	22.528854
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.10571	8.707259	168.537116	2.164946	7.141062	9.182176
min	0.00632	0	0.46	0	0.385	3.561	2.9	1.1296	1	187	12.6	1.73	5
25%	0.082045	0	5.19	0	0.449	5.8855	45.025	2.100175	4	279	17.4	6.95	17.025
50%	0.25651	0	9.69	0	0.538	6.2085	77.5	3.20745	5	330	19.05	11.36	21.2
75%	3.677082	12.5	18.1	0	0.624	6.6235	94.075	5.188425	24	666	20.2	16.955	25
max	88.9762	100	27.74	1	0.871	8.78	100	12.1265	24	711	22	37.97	50

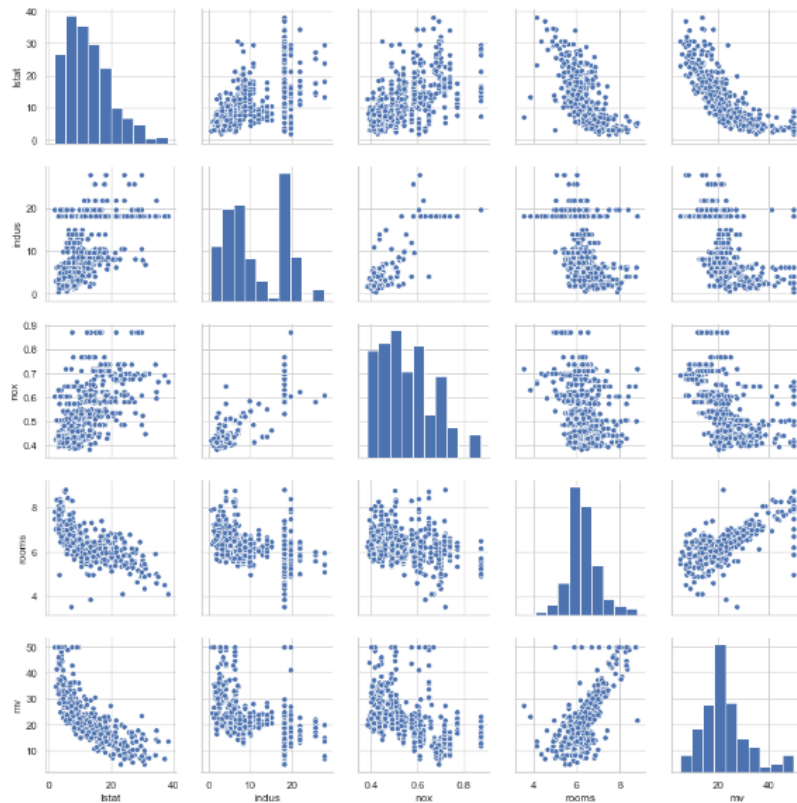
The response mv variable distribution shape has not changed after the scaling and fit to the Gaussian distribution



## Data Visualization

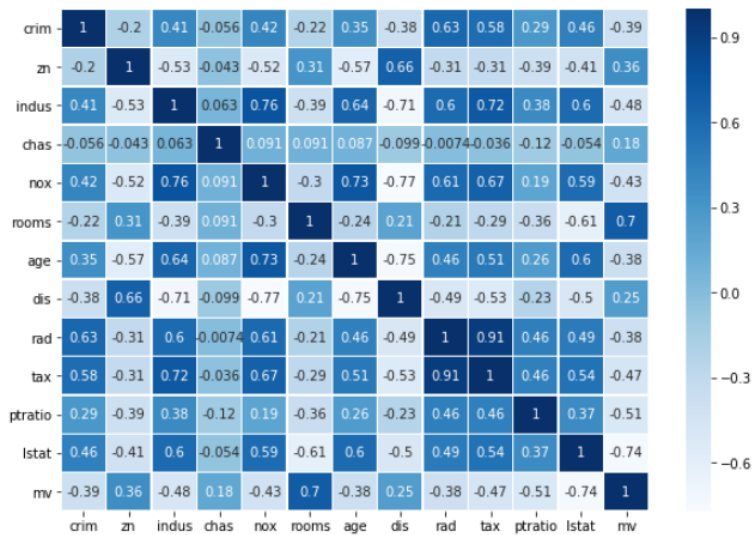
Below are basic dataset visualization histograms and pair plots showing relation between features and response variables





## Correlation

Correlation matrix shows strong relations between response variable and number of rooms, ratio Pupil/teacher ratio in public schools, Percentage of population of lower socio-economic status



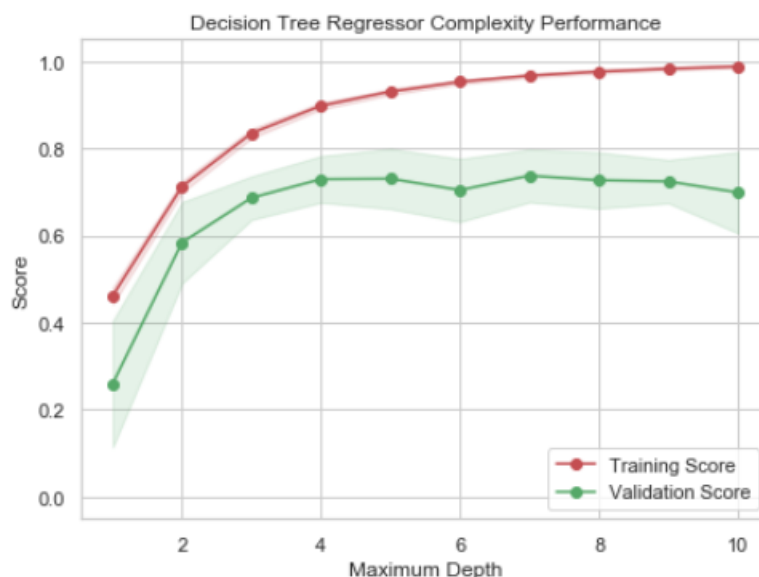
## OVERVIEW OF PROGRAMMING WORK

The Python's *Pandas* and *Numpy* for data handling, and *Scikit Learn* for machine learning and model evaluation metrics. The housing data was presented to us as a CSV file and loaded into the program using *Pandas*. The resulting data frame was put into numpy arrays *prelim\_model\_data*, so that it could be used within the *Scikit Learn* environment. Model data was obtained standardizing *model\_data* using *SciKit Learn StandardScaler()*. All four regression models: **Linear, Elastic Net, Lasso and Ridge**, along with a ten-fold K-fold cross validation design using root mean squared error metric for performance evaluation were implemented within *SciKit Learn* environment. For the re-usability, regression models were put in the array and fitting, model attributes were generalized and were include into a loop. For the model selection Model Complexity, performance parameters were analyzed, coefficient, MSR (mean squared error), AIC and BIC parameters were calculated.

## RESULTS AND RECOMMENDATIONS

### Complexity Curves

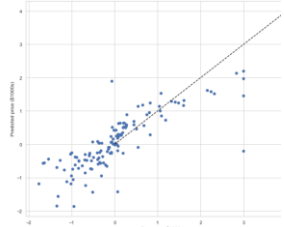
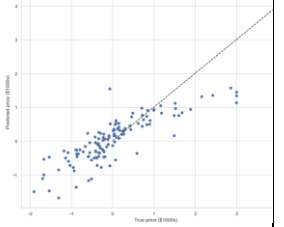
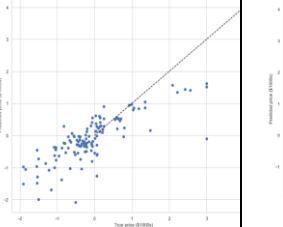
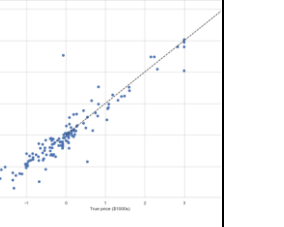
The following code graph for the model that has been trained and validated on the training data using different maximum depths. The graph produces two complexity curves — one for training and one for validation. Similar to the learning curves, the shaded regions of both the complexity curves denote the uncertainty in those curves, and the model is scored on both the training and validation sets using the **performance\_metric** function.



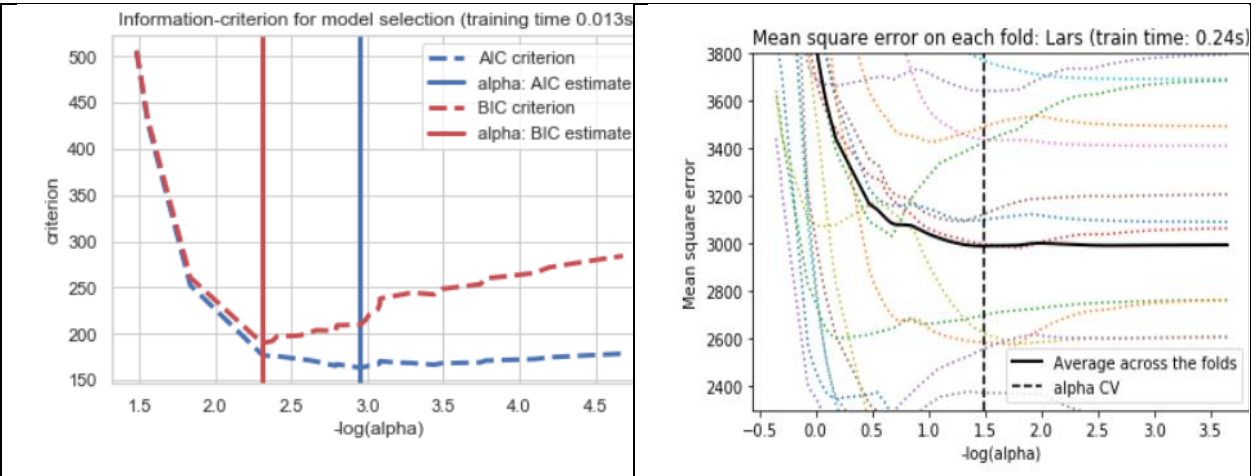
It seems that at maximum depth of 4 the training score seems to plateau here, indicating the highest possible score for the model's ability to generalize to unseen data. Gap between the training score and testing score does not seem to be substantial too, indicating that the model may not be suffering from a high variance scenario.

Results

The results from the 10-fold cross-validation in standardized units for selected models depicted in the following grid

	Linear	Ridge	Lasso	ElasticNet
Root mean-squared error	0.561940	0.560511	0.587381	0.568084
Predicted vs True				

Akaike information criterion (AIC), the Bayes Information criterion (BIC) and cross-validation to select an optimal value of the regularization parameter alpha of the Lasso estimator. Results obtained with LassoLarsIC are based on AIC/BIC criteria. Information-criterion based model selection is very fast, but it relies on a proper estimation of degrees of freedom, are derived for large samples (asymptotic results) and assume the model is correct, i.e. that the data are actually generated by this model.



## Recommendations

Based on the MSR analysis and preliminary AIC and BIC research, the ***Ridge Regression*** is recommended for the *Boston Housing Market* analysis max depth =4 and threshold=20% hyper parameters.