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 track.

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 valuate 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**

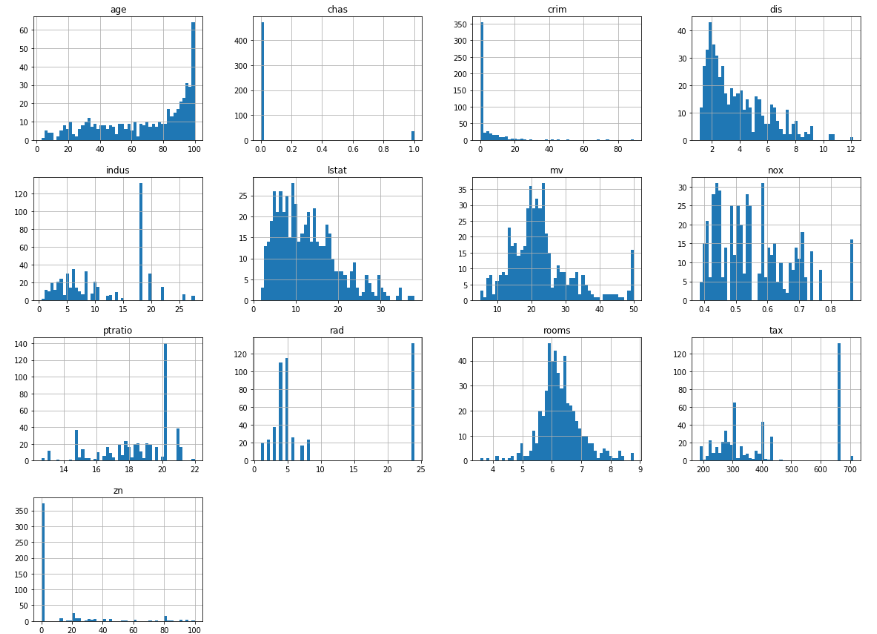


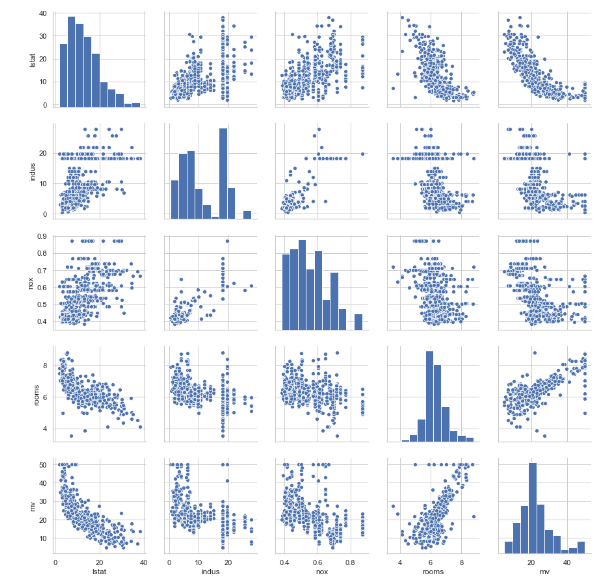
The response mv variable distribution shape has not changed after the scaling and fit to the Gaussian distribution

|  |  |
| --- | --- |
|  |  |

**Data Visualization**

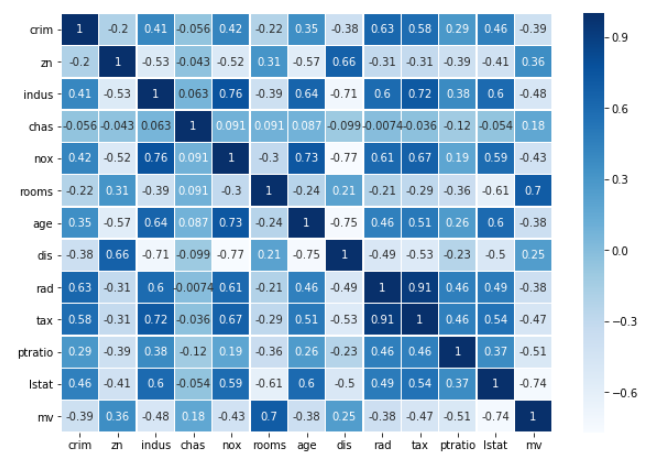
Below are basic dataset visualization histograms and pair plots showing relation between features and response variables





**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



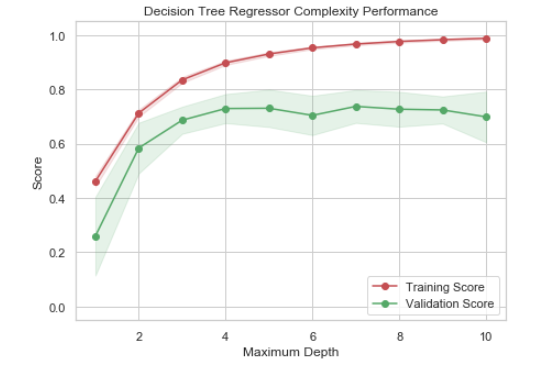
Overview of Programming Work

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 where 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 Hosing Market* analysis max depth =4 and threshold=20% hyper parameters.