1) Statistical Analysis and Data Exploration

1.1) Number of data points?

506

1.2) Number of features?

13

1.3) Minimum and maximum housing prices?

Minimum = 5.00

Maximum = 50.00

1.4) Mean and median Boston housing prices?

Mean = 22.532806

Median = 21.200000

1.5) Standard deviation?

9.188012

2) Evaluating Model Performance

2.1)

Which measure of model performance is best to use for predicting Boston housing data?

Regression metrics/Mean squared error.

Why do you think this measurement most appropriate?

I choose Regression metrics because Boston housing data is numeric continuous data. And I among the regression metrics mean squared error is the most appropriate because it is simple and it give more weight on point that is further away from the mean.

Why might the other measurements not be appropriate here?

Because Classification metrics, Multilabel ranking metrics and Clustering metrics are not for regression data.

In regression metrics, explained variance score and R^2 score are too complicate. Median absolute score is not sensitive to all point. And mean absolute error does not give weight on further away point as mean squared error does.

2.2)

Why is it important to split the data into training and testing data?

Because I want to know the performance of the model which trained on unknown data (or testing data).

What happens if you do not do this?

Cannot validate the performance of model. And we cannot see error due to variance which leads to over fitting model.

2.3) Which cross validation technique do you think is most appropriate and why?

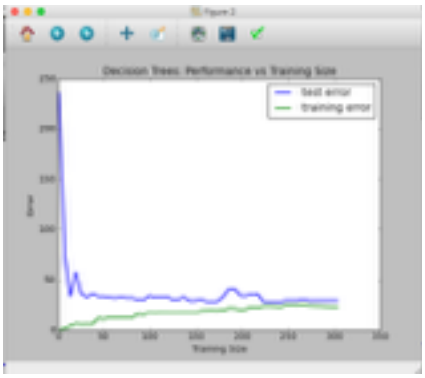
K-fold cross validation. Because Boston housing data does not have label for stratified k-fold,label k-fold technique and etc. And leave-one-out technique is too expensive

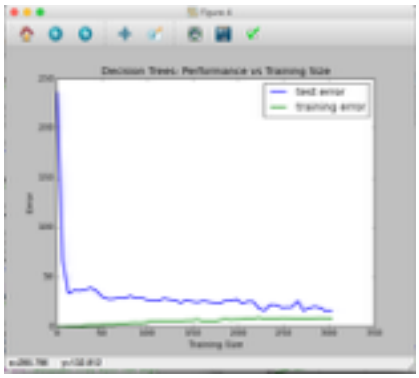
2.4) What does grid search do and why might you want to use it?

It do an exhaustive search over specified parameter values for tuning the Decision Tree Regressor and find the best evaluating estimator performance score from mean of validation score in cross-validation.Which in this case I modify to 5-fold cross-validation. I want to use grid search because I don't need to write a iterative code for finding the best parameter because grid search get parameter and run all, compare and select best one.

3) Analyzing Model Performance

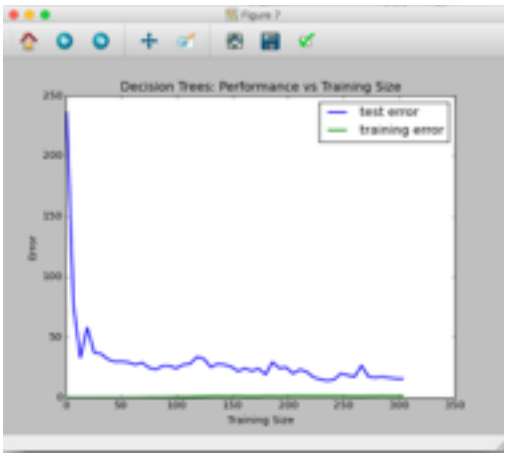
3.1) Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?





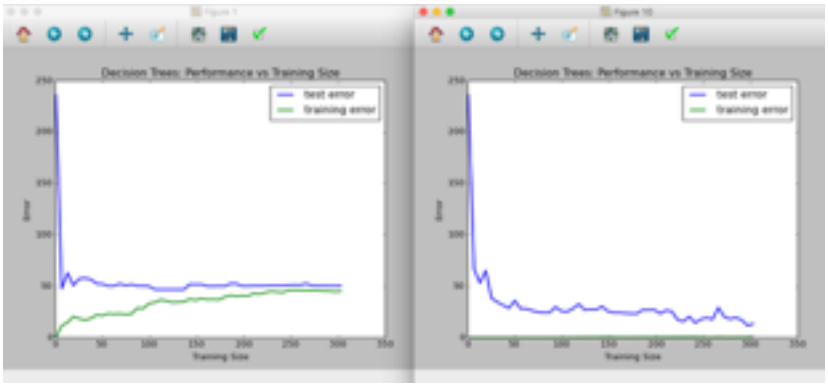
Left image is learning curve graph for max depth = 2, and max depth = 4 for right image.

The error of test error decrease as training size increase, while training error increase as training size increase.



When max depth is 7 or greater than 7. trying error is getting close to 0.

3.2) Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/underfitting or high variance/overfitting?

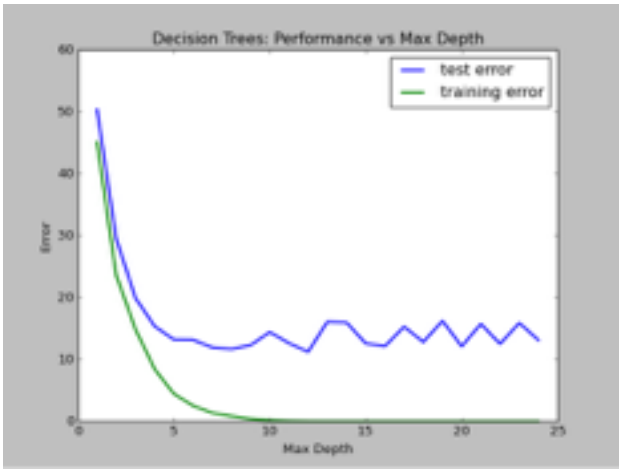


max depth 1 max depth 10

At max depth = 1, the model is suffer from high bias/underfitting because both training error and testing error are high and the trend of graph when increasing training size does not decrease the error as it go almost parallel to X-axis.

At max depth = 10, the model is suffer from high variance/overfitting because the training set is 0 error but there is still large gap between training error and test error.

3.3) Look at the model complexity graph. How do the training and test error relate to increasing model complexity?



At first training and test error get decrease rapidly until Max depth is around 3 where both test error and training error decrease slowly. Test error from max depth = 4 and greater are getting stable (the trend is parallel to X-Axis). And training error get near 0 when max depth is approaching 10.

3.4)Based on this relationship, which model (max depth) best generalizes the dataset and why?

From previous image,max depth = 6 best generalises the data set because test error trend started to get stable from model = 5 and there is not much gap between test error and train error. And both test error and training error at depth = 6 are low.

4) Model Prediction

4.1) Model makes predicted housing price with detailed model parameters (max depth) reported using grid search. Note due to the small randomization of the code it is recommended to run the program several times to identify the most common/reasonable price/model complexity.

By running 20 times, the median of best model's max depth is 6. which predict house price at 20.76598639.

4.2)Compare prediction to earlier statistics and make a case if you think it is a valid model.

I think it is a valid model because result of prediction is 20.76598639, which fall near mean and median of Boston housing dataset.