Questions and Report Structure

1) Statistical Analysis and Data Exploration

- Number of data points?
- There is 6578 Data points. (Maybe you want the Number of data points for each feature : 506 but i'm not sure)
- Number of features?
- There is 13 features.
- Minimum and maximum housing prices?
- The minimum is zero and the maximum is 711.
- Mean and median prices of Boston housing prices?
- The mean is about 70 and the median is 6.5.
- Standard deviation?
- The standard deviation is equal to 145.

2) Evaluating Model Performance

- Which measure of model performance is best to use for regression and predicting Boston housing data? Why is this measurement most appropriate? Why might the other measurements not be appropriate here?
- The use of RMSE (Root Mean Squred Error) is a good general purpose error metric for numerical predictions and this the one we have to use here because as the mean and the median can show us, there is some bad oultiers out there. So it depends on our loss function. In many circumstances it makes sense to give more weight to points further away from the mean--that is, being off by 10 is more than twice as bad as being off by 5. In such cases RMSE is a more appropriate measure of error. If being off by ten is just twice as bad as being off by 5, then MAE is more appropriate.
-
- Why is it important to split the data into training and testing data? What happens if you do not do this?

- It’s important to prevent overfitting. If we train the model on the whole dataset without splitting, the model will perform well on the data we have, but will perform really bad to predict.

- Which cross validation technique do you think is most appropriate and why?
- I usually prefer K-Fold CV because a classic train/test provides a high variance estimate of out-of-sample accuracy and K -Fold overcomes this limitation. But as you want us to just split, i will stick with the easy one.

- What does grid search do and why might you want to use it?
- We want to choose a set of hyperparameters for our DecisionTreeRegressor (the max_depth) with the goal of optimizing this one. We can use gridsearch to do that. Grid Search is simply an exhaustive searching through a manually specified subset of the hyperparameter guided by some performance metric and measured by cross validation.

3) Analyzing Model Performance

- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?
- As the training sizes increases we can clearly see that the testing errors shrink while training errors grow.

- 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?
- For maxdepth=1 we can see that the training error and testing error are very similar, this indicates that the model is underfitting the data because it doesnt have enough complexity to represent the data (high bias model).
- For maxdepth=10 there is a large separation between the training and test error which indicates overfitting / high variance.

- Look at the model complexity graph. How do the training and test error relate to increasing model complexity? Based on this relationship, which model (max depth) best generalizes the dataset and why?
- As the model get more complex, the training score is going down (less error) but the test score stay the same as the model complexity increases. At the end both diverges. This is

because a more complicated model can fit the noise better so the training error gets better but not the test error. We can say that the “sweet spot” for this model complexity is max_depth=4.

4) Model Prediction

- Model makes predicted housing price with detailed model parameters

- best_params : {'max_depth': 4}
  model : DecisionTreeRegressor(criterion='mse', max_depth=4, max_features=None,
  max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2,
  min_weight_fraction_leaf=0.0, random_state=None,
  splitter='best')
  score : -5.71587476412
  House: [11.95, 0.0, 18.1, 0, 0.659, 5.609, 90.0, 1.385, 24, 680.0, 20.2, 332.09, 12.13]
  Prediction: [21.62974359]
- Compare prediction to earlier statistics

- The prediction is a little above the median and inside the first standard deviation (68%). The mean is kind of useless due to outliers. So we can say it’s a good one ?