P1: Predicting Boston Housing Prices Report

Statistical Analysis and Data Exploration

- Total number of houses: 506
- Total number of features: 13
- Minimum house price: 5.0
- Maximum house price: 50.0
- Mean house price: 22.533
- Median house price: 21.2
- Standard deviation of house price: 9.188

1) Of the available features for a given home, choose three you feel are significant and give a brief description for each of what they measure.

**Answer: **

CRIM, RM, RAD

I took into consideration three factors: crime rate, size of a house , location of a house

CRIM: indicates the crime rate in town, higher crime rate will decrease the price of the house

RM: no. of rooms, doesn't indicate well the size of the house , but i assume that houses with more rooms will cost more than houses with less rooms

RAD: refers to how easy it is to access radial highways which indicates how good or bad the location of the house, houses with good locations will cost more than houses with bad location

2) Using your client's feature set CLIENT_FEATURES in the template code, which values correspond to the chosen features?

**Answer: **

CRIM: 11.95
RM: 5.609
RAD: 24

3) Why do we split the data into training and testing subsets?

**Answer: **
to make the training and testing data independent of each other, When i test my model on different data from my training one this gives me better results answering if my model generalizes well or not , but if my test data was already part of the training one, my model will definitely produce good results as it was previously trained on these data. Also it's a check on overfitting as my model may have very low error on the training set , but after using the test set if my model is overfitted then we will find much higher error than the training set

4) Which performance metric below is most appropriate for predicting housing prices and analyzing error? Why?

**Answer: **
Since the data values are continuous, so i went to regression metrics and chose Mean Squared Error for analyzing total error (i don't know what would be the difference if i choose mean absolute error but i guess i can choose either one of them)

5) What is the grid search algorithm and when is it applicable?

**Answer: **
Gridsearch algorithm is used to tune a function that chooses between many parameter values instead of trial and error gridsearch does it for us , we choose first the classifier that we will use, then the set of parameters that we need to go through to choose the best ones for our model , a scoring function which determines how we choose our model , a k-fold value which is 3 by default .
in each fold a new instance of the classifier is instaniated and we test the different values of the parameters given the current training and testing data and come up with the evaluation according to our scoring function.
then choose the best model with the best parameters from the k-folds done.

applicable whenever i need to choose from a set of paramaters the best ones for my model without going through trial and error

6) What is cross-validation and how is it performed on a model? Why would cross-validation be helpful when using grid search?

**Answer: **
Cross validation is the way of splitting the data into training and testing sets, in the simplest way we can split the data to 70% training and 30% testing (for example), there

is another way in cross validation for splitting data called K-fold cross validation which distributes the data to k bins and run the training & testing processes K times considering testing data a different bin each time (and training data will be the remaining bins) so when it finishes it would have used all the data for training and testing, it is generally used for increasing the prediction accuracy and decreasing overfitting.

In gridseach k-fold cross validation is used so data will be splitted into training and test set. gridsearch will fit the data then test/evaluate on the remaining data (according to the score function) , this is done for each parameters set for K times . when K-fold finishes it will average the values of each parameter in each k-fold and choose the best parameters for my model so cross-validation increases the accuracy of the gridsearch.

7) Choose one of the learning curve graphs your code creates. What is the max depth for the model? As the size of the training set increases, what happens to the training error? Describe what happens to the testing error.

**Answer: **
Max_depth = 3
As the size of the training set increases the training error increases gradually till it reaches a steady state, as the number of data points in training set increases the training error remains the same

As the size of the training set increases the testing error decreases till it reaches a steady state value, as the number of data points in training set increases the testing error is about the same.

8) Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?

**Answer: **
At max_depth = 1 -> it suffers from high bias, because the training and testing errors converge and they are both very high values, which means that the model fails to represent the complexity of the training data and it won't change anything if we increased the data which means underfitting.

at max_depth = 10 -> it suffers from high variance, because the training error is too small actually approximately equals zero but the test error is large compared to it (large gap between the two curves),which means that the model doesn't generalize well which means overfitting.

8) From the model complexity graph, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?

**Answer: **
As the maximum depth increases the training error decreases exponentially and tends to zero at high complexity
but as the depth increases the testing error decreases lineary till a certain value then it oscillates (increase & decrease) around a certain value till it reaches a steady state

Max_depth that best generalizes the dataset = 5 , to reduce overfitting at high complexity values we have to either increase the data or reduce the complexity,so when we reduce the complexity the training error will decrease and the testing error will oscillate , until we reach a point(arount max_depth = 5) where the value of the testing error is minimum and the gap between the training and testing error slightly reduced.

10) Using grid search, what is the optimal max depth for your model? How does this result compare to your initial intuition?

**Answer: **
gridsearch max_depth = 4, it's close to my initial intuition, not sure which is better 4 or 5, after looking at gridsearch max_depth i revised the maximum depth graph but i think according to the graph 5 is better.

11) With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the statistics you calculated on the dataset?

**Answer: **
Best selling price = 21.630 in $1000's
by comparing it to the basic statistics calculated on the dataset
the selling price is above the median(21.2) with a slightly small value and below the mean(22.533) with a slightly small value, it definitely didn't go above the max value or below the min value

value from mean = 0.903
value from median = 0.43

12) In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients’ homes in the Boston area.

**Answer: **
I would use this model to predict selling prices for future clients' homes as it produced a decent value for the client's home which was close to the mean/median of the the original data.