PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW | CODE REVIEW | NOTES

Requires Changes

SHARE YOUR ACCOMPLISHMENT

2 SPECIFICATIONS REQUIRE CHANGES

Data Exploration

✓ **All requested statistics for the Boston Housing dataset are accurately calculated. Student correctly leverages NumPy functionality to obtain these results.**

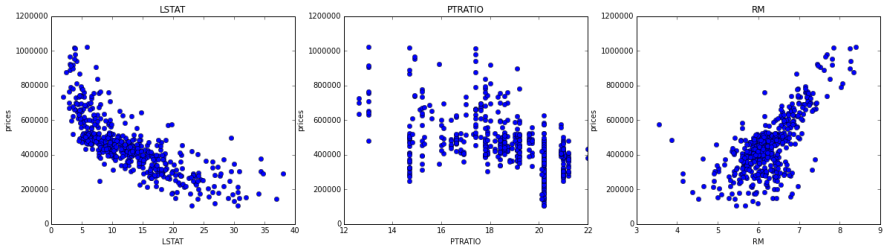
Well done! You have shown a good understanding of making use of `numpy` to get the descriptive statistics of the dataset.

✓ **Student correctly justifies how each feature correlates with an increase or decrease in the target variable.**

- Well done for identifying the correlation for the features vs the house pricing.
- We could confirm this from the scatter plot as follows:

```
import matplotlib.pyplot as plt
plt.figure(figsize=(20, 5))
for i, col in enumerate(features.columns):
  plt.subplot(1, 3, i)
  plt.plot(data[col], prices, 'o')
  plt.title(col)
  plt.xlabel(col)
```

```
    plt.ylabel('prices')
```



Developing a Model

Student correctly identifies whether the hypothetical model successfully captures the variation of the target variable based on the model's R^2 score.
The performance metric is correctly implemented in code.

- Kindly note that "This could indicate a high error variance i.e. that our model is overfit to the training data." is not accurate.
- The r2 score measures the goodness of how fit the model is, the greater the r2 score is, the better the model is fitting. However, the r2 score is not an error performance metric which does not indicate the high error variance scenario.
- Please note that for the performance metrics, we have several different ways to measure the model performance, Explained Variance Score and R2 Score are metrics to measure how well the model fits the data, explains the variability in predictions. The other possible options are mean squared error and mean absolute error etc.
- Please look at here for the list of the performance metrics.

Student provides a valid reason for why a dataset is split into training and testing subsets for a model. Training and testing split is correctly implemented in code.

The benefit of splitting Boston Housing Data into training and testing data would provide two independent dataset to

- Give estimate of performance on an independent dataset;
- Serve as check on overfitting.
- Without the testing data, you will not gain insights about how your

model is performing and will not allow you to work out issues like bias vs variance trade-offs.
- As you mentioned in the report, splitting the dataset into training and testing would be a great way to make sure the model has been generalized well for unseen data.

SEEDING YOUR ALGORITHMS:

- In order to remove randomness of your algorithms, and to make sure your results don't differ at each run, please consider to always use a random seed to seed your algorithms.
- A standard practice I've come across is to define a random seed as a global variable in your work, and to use it throughout all the algorithms/methods which require random number generation (splitting data, decision tree initialisation, neural network weight initialisation etc).
- In sklearn, as far as I know, random seeds are provided to methods and functions using the parameter `random_state`. Please seed all of your algorithms in the future if you haven't been doing so yet

Analyzing Model Performance

✓ **Student correctly identifies the trend of both the training and testing curves from the graph as more training points are added. Discussion is made as to whether additional training points would benefit the model.**

- It would be better that you could indicate that the validation score reaches to a plateau but does not become better.
- To give more context, when the training set is small, the trained model can essentially "memorize" all of the training data. As the training set gets larger, the model won't be able to fit all of the training data exactly.
- The opposite is happening with the test set. When the training set is small, then it's more likely the model hasn't seen similar data before. As the training set gets larger, it becomes more likely that the model has seen similar data before.
- It would be worth noting that the model learning rate is related to the increasing testing score, initially, the learning rate of the model is high, however, in the later stage, adding more data point

does not change the testing score significantly.

✓ **Student correctly identifies whether the model at a max depth of 1 and a max depth of 10 suffer from either high bias or high variance, with justification using the complexity curves graph.**

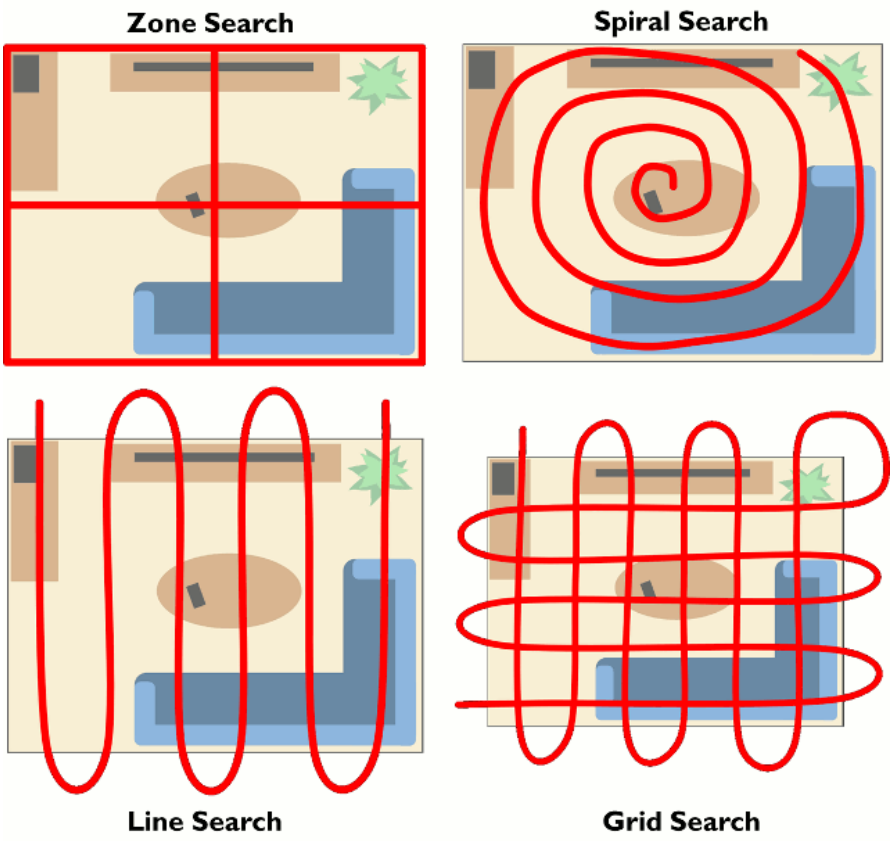
✓ **Student picks a best-guess optimal model with reasonable justification using the model complexity graph.**

- Please note that with the increasing model complexity, the model goes through two stages from underfitting to overfitting - Please consider to include this in your report
- The first phase is where the model is underfitted and the training error is extremely low.
- The second phase is where the model is overfitted and the difference between testing and training score is high.
- The optimal model is where the turning point at, which the training score is high and testing score is at global maximum.

Evaluating Model Performance

✓ **Student correctly describes the grid search technique and how it can be applied to a learning algorithm.**

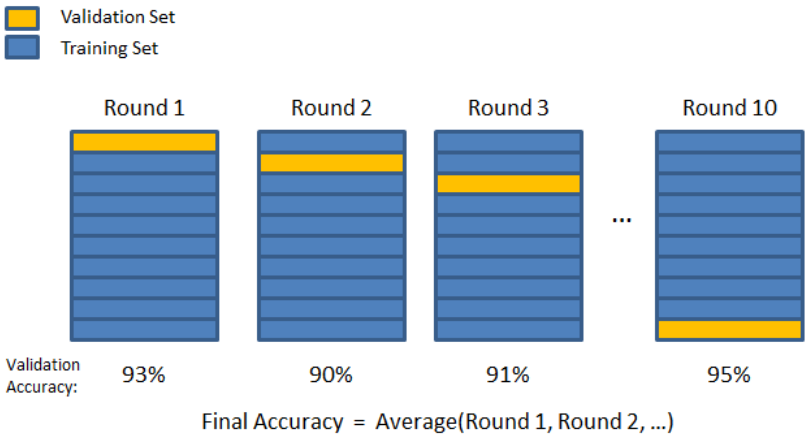
- Similar to what you have included regarding to grid search algorithm which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set
- It would be worth mentioning about fine tuning a learning algorithm for a more successful learning/testing performance in terms of the application for grid search.
  Please look at the following comparison for different space search:



Student correctly describes the k-fold cross-validation technique and discusses the benefits of its application when used with grid search when optimizing a model.

Rate this review

- Please note that there is an additional steps to average out error across all k trials is computed.
- Kindly note that for Cross validation, especially on K-fold CV - the data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times.
  Please look at the following diagram which clearly shows how the k-fold CV works:



✓ **Student correctly implements the `fit_model` function in code.**

✓ **Student reports the optimal model and compares this model to the one they chose earlier.**

✓ **Student reports the predicted selling price for the three clients listed in the provided table. Discussion is made as to whether these prices are reasonable given the data and the earlier calculated descriptive statistics.**

✓ **Student thoroughly discusses whether the model should or should not be used in a real-world setting.**

Would completely agree with your comment of

- Data is very likely not representative of Boston anymore. LSTAT, PTRATIO, RM and MDEV have changed over the last 38 years. Predictions made on homes in the Boston area is very likely not valid anymore
- Additional features such as crime rate, distance to high-tech areas, distance to universities, remodelling activities, age of the homeetc. should be added into the feature list as they influence the price of the home.
- Model trained with Boston data cannot be applied to other regions such as a rural city as the characteristics are completely different.

RESUBMIT PROJECT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Student FAQ