



[Return to "Machine Learning Engineer Nanodegree" in the classroom](#)

[DISCUSS ON STUDENT HUB](#)

Predicting Boston Housing Prices

REVIEW

HISTORY

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

Udacity student,

your submission shows how you committed you are to the course. I'm proud of being part of your journey. 😊

I tried to make the best feedback possible to reach your expectation. I hope I did well! I encourage you to keep the great work and read all additional materials and left to you. 😊

You have to do some adjustments but it's normal in your first project. Just keep in mind that you are in right path to succeed in this course! It's very important for us to know what is your opinion about the aspects of this project. In a daily machine learning engineer professional it is not all about coding, but also telling people what the data wants to show to improve the businesses decisions.

I'll share with you some extra links from [Medium](#):

[A guide to start your path in Data Science and Machine Learning:](#)

[Fundamental Python Data Science Libraries](#)

This last one is not for only a job interview, but it contains a lot of useful information about some great topics for a machine learning professional:

[Data Science and Machine Learning Interview Questions](#)



Data Exploration

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

Correct! 😊

You have answered all statistics questions using the Numpy library. Nice job! 👍

It is important here to know why Udacity ask students to use the NumPy library:

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

[Numpy Documetation](#)

It's always important to be aware of tools you use. For example, the Pandas' Series.std() will by default give you different result than numpy.std(). It's because Numpy takes in count the whole population while pandas assumes that you are evaluating the standard deviation for a sample of your dataset.

This [article](#) has a very good explanation about it.

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

Correct! 😊

You have correctly justified how each feature correlates with an increase or decrease in the target variable.

In addition, you can also plot your data to confirm your intuition and practice some coding skills using a library called matplotlib

```
import matplotlib.pyplot as plt
plt.figure(figsize=(15, 5))
for i, col in enumerate(features.columns):
    plt.subplot(1, 3, i+1)
    plt.plot(data[col], prices, 'x')
```

```
plt.title('%s x MEDV' % col)
plt.xlabel(col)
plt.ylabel('MEDV')
```

[Matplotlib Documentation](#)

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.

Great job! 😊

Since [R2 Score](#) is close to one we may say it is a good model.

The [R-Squared](#) is very common score used in Data Science / Machine Learning. But we can have another type of R-squared called [Adjusted R-Squared](#).

Also, in this [article](#) from Duke University you can find a very nice opinion about how good R^2 Score can be in a ML mode.

I'd like to share one more excellent article: [Mean Squared Error, \$R^2\$, and Variance in Regression Analysis](#)

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.

You have implemented a `random_state` for the `train_test_split` function properly.

Nevertheless, the test and train sets must be split in a certain way. Most of times, the ML engineers use 80% for Training and 20% for testing.

May you please fix your code to reflect those values?

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.

Good job! 😊

As more data we add the training score decreases and the testing score increases. Nice job explaining it.

Moreover, adding even more points will not benefit the model and only "make the computer work harder" in terms of processing the data. I recommend this reading about this topic: ["How much data is enough?"](#).

Getting more training points may be hard and require a lot of additional work. More data also requires more computing resources or making performance improvements (this is not a problem in this case since the training set is small). Getting a dataset with more features, choosing a more complex model or increasing the `max_depth` hyperparameter is likely to produce better improvements than getting more training data.

In a short medium [article](#) you may go deeper in your studies about learning curves.

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.

Correct! 😊

Good job identifying high bias and high variance for different `max_depth` parameters.

The model which is trained with a maximum depth of 1 suffers from high bias because it performs well neither on the training nor on the validation set. The graph indicates this by showing that training and validation scores are low (close to 0.4) and both training score and validation score curves are close to one another.

The model which is trained with a maximum depth of 10 suffers from high variance because it performs really well on the training set but its performance on the validation set is not as high. This can be seen in the complexity graph by looking at the big gap between the training score and validation score curves. The training score is close to 1.0 while the validation score is less than 0.7.

In this [link](#) you can read more about high bias and high variance of data and boost your understanding about this topic!

Also in Wikipedia you may find a nice [article](#) about the tradeoff regarding to a model with high bias and high variance.

You may also check this [website](#). It brought me a lot of light about this issue.

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

Correct! 😊

Evaluating Model Performance

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

You are almost there. 😊

You started explaining well some aspects of Grid Search. Now you need to complete your answer to fully meet the requirement for this question. You are on the right path, just need to improve your answer, right?

Grid Search allows you to provide various models to apply your data to in order to see which one provides the most useful insight. Similarly the individual hyperparameter of each model can be adjusted and tested to see which hyperparameter settings provide the optimal results. This optimal result can be measured as a score such as R-Squared, F1 etc. depending what you want to use as the score measure, note here that not only F1 score is a metric for Grid Search. We can have some.

Showing a different example, with a [decision tree classifier](#), there are hyperparameters such as `max_depth`, `min_samples_leaf` and `min_samples_split` available to optimize the model.

Without Grid-Search, we have to set the hyperparameter to produce the best outcome running every combination time at time. Nevertheless, with a grid search we run the one function without the "manual effort".

Below shows an example of the values in each hyperparameter which would be tested over the various sequences.

```
{ 'min_samples_leaf': [2,4,8,16,20], 'min_samples_split': [2,4,6,8,10], 'max_depth': [2,4,8,10] }
```

Once the grid search is completed we can have the best estimator for the grid search result, and check the hyperparameter settings the grid search used to achieve the best score. However, the more hyperparameters we have the more computing power will be needed to complete the grid search task.

I'd like to share some extra links to boost your learning and max 7-8 minutes reading:

1. [Official sklearn page on gridSearch](#)
2. [How gridSearch works](#)
3. [Specifying multiple metrics for evaluation](#)
4. [The scoring parameter: defining model evaluation rules](#)
5. [Defining your scoring strategy from metric functions](#)

You need to improve your question explaining the main reason to use the Grid Search and how it works to find those ideal hyperparameters.

1. Does it uses a score to check to best combination? Is this score only one kind score or depends on the Machine Learning Algorithm?
2. Does it do it randomly or test every combination?
3. May you give some examples of hyperparameters of a learning algorithm?

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.

You are almost there! 😊

You explained the basic concepts of k-fold cross validation. You need to think about some things about your answer and explain some aspects:

- 1- What is the difference between testing and validation sets? It is important to know the difference.
- 2- Do we break the whole dataset or only the training subset? How do we break the dataset?
- 3- How do we evaluate the performance of the k-fold cross validation? Is there a single score in there or the algorithm take the average of the results ?

I'd like to share some extra readings about K-fold:

- [What is Cross Validation?](#)
- [K-fold and Cross Validation](#)

K-fold cross-validation training allows you to extend the use of your data so the Train and Validation data can be treated as one. Thus only needing to remove a portion of data for final testing. This is achieved by creating multiple data sets or 'folds' of the same data. A model is trained using k-1 of the folds specified as training data and the remainder for the validation. The testing/validation is run for each of the folds. It is then possible to take an average score from across the folds to help determine the usefulness of the model.



For Example, using a data set of 12. If 4 folds are specified, the data will be split into 4 combinations. With k-1 of the folds being used for training, that provides 9 blocks of data for Training and 3 blocks for Validation. Since all data is used in each set.

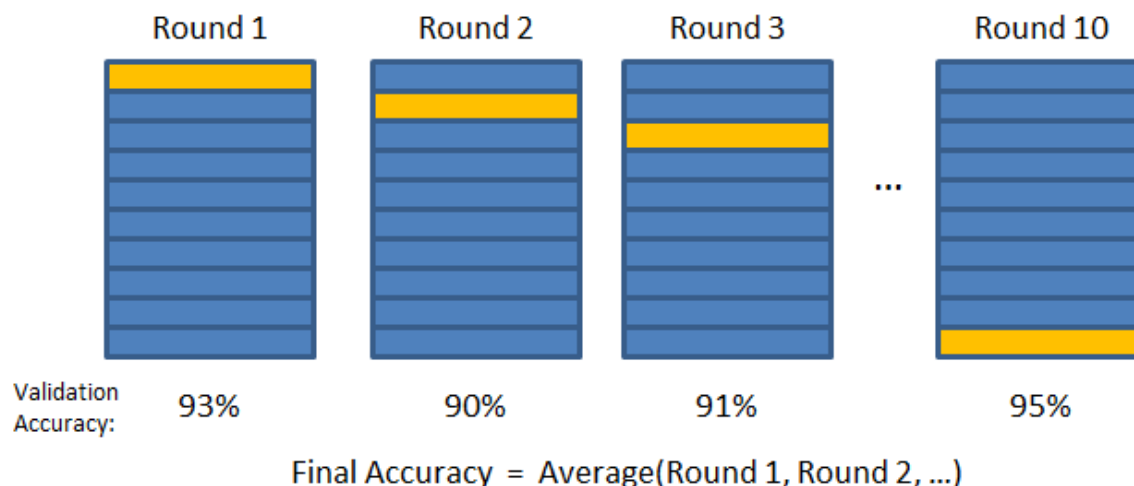
'TRAIN:', array([3, 4, 5, 6, 7, 8, 9, 10, 11]), 'VALIDATE:', array([0, 1, 2])

'TRAIN:', array([0, 1, 2, 6, 7, 8, 9, 10, 11]), 'VALIDATE:', array([3, 4, 5])

'TRAIN:', array([0, 1, 2, 3, 4, 5, 9, 10, 11]), 'VALIDATE:', array([6, 7, 8])

'TRAIN:', array([0, 1, 2, 3, 4, 5, 6, 7, 8]), 'VALIDATE:', array([9, 10, 11])

 Validation Set
 Training Set



Applying cross-validation alongside grid search, it is possible to ensure the results of a grid search are not just fitting to one set of training data, but that the results are consistent across multiple train/validates sets created by k-fold cross validation. If grid search is only run against one set of training data it is more difficult to ensure the model is working consistently.

One drawback to consider with k-fold is that you are now running train/validation multiple times. Additionally when combined with grid search running multiple parameters, this could affect processing time when determining the optimum model.

Student correctly implements the `fit_model` function in code.

Correct! 😊

Very nice implementation. 👍

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

Correct! 😊

You have explained how your answer compares with question 6.

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

Correct! 😊

The selling prices you evaluated are correct and your discussion shows your very good understanding about this project. Also, it shows that you can extract insights from data using your intuition and abilities.

You can also plot the data and the predictions to have an intuition about the model:

```
prediction_data = np.transpose(client_data)
pred = reg.predict(client_data)
for i, f in enumerate(house_features):
    plt.scatter(features[f], prices, alpha=0.25, c='green')
    plt.scatter(prediction_data[i], pred, color='red', marker='D')
plt.xlabel(f)
plt.ylabel('MEDV')
plt.show()
```

Good job!

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

Correct! 😊

Nice answers regarding to all questions in this last section.

🔄 RESUBMIT

📄 DOWNLOAD PROJECT



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

📺 [Watch Video](#) (3:01)

RETURN TO PATH

Rate this review
