UDACITY

PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Meets Specifications

Avid Learner,

Great job getting over the minor humps from the past submission! You show great potential and a firm grasp of the knowledge on our subject. Your answers are extremely satisfying and well-thought of! Keep up the great work and I wish you all the best of luck throughout the whole course! 😃👍

Data Exploration

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

Great Implementation on using NumPy function to calculate statistical data. Good Job! Here are some tips for a pro like you:

Pro Tips:

- Checking your dataset statistics is an very useful routine in applying a predictive model. This is because:
    - It helps us to check if the key assumptions of our algorithms hold (thereby helping us choose which model to apply).

    - These statistics tend to be very handy when you obtain a prediction, to check whether the predictions are reasonable, and not off-chart, compared to central values of the dataset.
- NumPy as a library might have been new for you, and not that easy to learn. In this tips section I'll give you some tips you can use when learning and picking up a new library:
    - Two functions are very useful when investigating a module (library) or a simple Python Object: the **doc** functionality and the dir() functionality.
    - If you wish to rapidly explore documentation of a library/module/function/object, you can just type print obj.**doc**, and the documentation of the function will be printed.
    - If you wish to rapidly explore what attributes and functions are available for an object, you can just type dir(obj), and you'll get a Python `list` of the object's attributes and functions.
    - Remember to always read documentation and try examples in your interpreter if you feel confused about a new library.
    - Hopes these help in your future Machine Learning Endeavors!

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

Excellent job explaining on why the `RM` value is directly proportional to the `MEDV` while `LSTAT` is inversely proportional. `PTRATIO`'s explanation is feasible as well and you got it perfectly that it may imply a lesser value of education. Great work here.

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.

Good job implementing the coefficient of determination `R^2`. Your explanation is logical since we have a score of 92.3% which is closer to 1 and considered a good fit.

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.

Awesome answer! Of course we need to test our model on an independent testing set in order for us to be able to assess how well it might perform when faced with unseen data.

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.

Amazing job describing how the training and testing score change as the training set size increases. You're right!

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.

Great job identifying that the model suffers from high bias when `max_dept = 1` and that the model suffers from overfitting (high variance) when the `max_dept = 10`.

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

Interesting guess for the optimal model! The reasons are sufficient and let us hope it would be spot on!

Evaluating Model Performance

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

Excellent work describing the Grid Search technique! Here are some pro tips for you. 😃

Pro Tips:

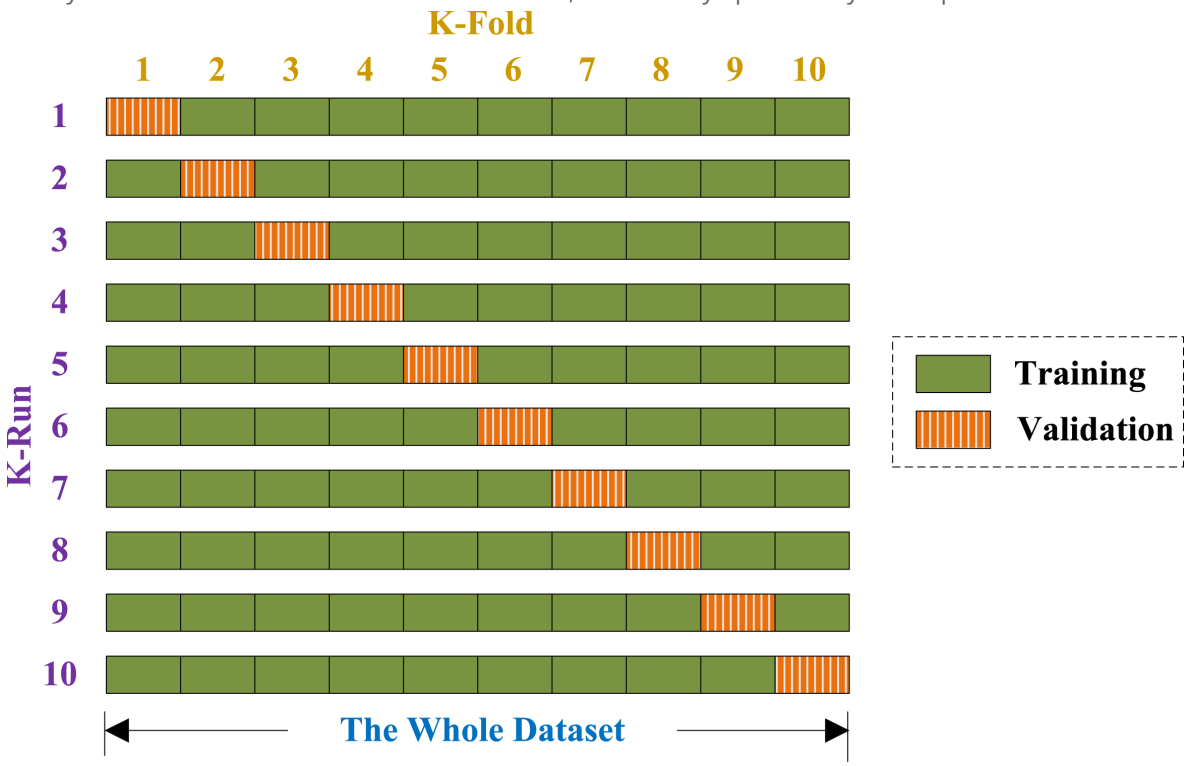
- Another very powerful parameter tuning algorithm is RandomizedSearchCV. In contrast with GridSearchCV, not all parameters are tried out, but rather a fixed number of parameter settings is sampled from the specified distributions.
- One particular advantage of RandomizedSearchCV is that it is much faster than GridSearchCV, and it is theoretically proven to find models that are as good; or even better than grid 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.

Amazing description of k-fold cross validation and how it is performed on a model! Here are some more suggestions and comments for further understanding:

Suggestions and Comments:

- Please check out the scikit-learn page about cross-validation, where a precise and concise definition of how cross-validation works is given. You may also check out cross-validation on Wikipedia) for more background information about cross-validation.
- Briefly checkout this illustration of cross-validation, which in my opinion very well captures how it works.



- Below is an explanation of how cross-validation works, with guidance from the diagram shown to the above:
  - As you can see from the diagram, the training set is divided into K-folds, or k subsets. For this particular diagram, it is 10 subsets.
  - The model is then trained and validated k times, or 10 times for this particular example.
  - At each run, one subset or fold is held out at validation set, and the other k-1 folds are used for training
  - At the end, the validation scores are collected and averaged out, to get the final score of the model being tested.
- One particular advantage of validating a model in this way is that it makes particular good use of the data available, especially if the dataset is small. So it can help mitigate overfitting.
- Also, as stated on the sklearn page for cross-validation, if a single set is used for testing, and parameter tuning, then information can leak away from the only test set into the model being tuned. So, using multiple test/validation sets can help mitigate this.
- Hope this helps you to understand what is cross-validation and how it works, as this is a quite important concept in Machine Learning.

Student correctly implements the `fit_model` function in code.

Excellent job in implementing the `fit_model` and utilizing `make_scorer` and `GridSearchCV` 👍

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

Nice work finding a valid optimal model max depth, which is very close to the one you guessed from your intuition! Below is a pro tip for you:

Pro Tip:

In order to have a more robust estimate of the best `max_depth` parameter, you might want to run the grid search algorithm multiple times.
Below I provided the code to help you do so:

```
max_depths = []
for i in range(500):
    reg = fit_model(X_train, y_train)
    max_depths.append(reg.get_params()['max_depth'])
best_max_depth = np.mean(max_depths)
print "The Best model, on average, has a max depth of:", best_max_depth
```

In general, if you had good intuition in picking your max_depth parameter from the complexity curves, your result from running the code above should be very close to your best-guess estimate of max_depth.

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.

Well done analyzing the prices from the prediction and I have to agree that they may be sold at the predicted selling prices. Great job utilizing the Data Exploration part in helping you formulate your comparison with the dataset and the statistics.

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

I agree with you wholeheartedly. Not only that the data is old and inflation happened, but also there are just not enough aspects and features that would be a factor in determining the value of a house.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

Student FAQ