This notebook is an exercise in the Intermediate Machine Learning course. You can reference the tutorial at this link.

In this exercise, you will leverage what you've learned to tune a machine learning model with **cross-validation**.

Setup

The questions below will give you feedback on your work. Run the following cell to set up the feedback system.

In [ ]:
```
# Set up code checking
import os
if not os.path.exists("../input/train.csv"):
    os.symlink("../input/home-data-for-ml-course/train.csv", "../input/
train.csv")
    os.symlink("../input/home-data-for-ml-course/test.csv", "../input/t
est.csv")
from learntools.core import binder
binder.bind(globals())
from learntools.ml_intermediate.ex5 import *
print("Setup Complete")
```

You will work with the Housing Prices Competition for Kaggle Learn Users from the previous exercise.

Run the next code cell without changes to load the training and validation sets in `X_train`, `X_valid`, `y_train`, and `y_valid`. The test set is loaded in `X_test`.

For simplicity, we drop categorical variables.

In [ ]:
```
import pandas as pd
from sklearn.model_selection import train_test_split

# Read the data
train_data = pd.read_csv('../input/train.csv', index_col='Id')
test_data = pd.read_csv('../input/test.csv', index_col='Id')

# Remove rows with missing target, separate target from predictors
train_data.dropna(axis=0, subset=['SalePrice'], inplace=True)
y = train_data.SalePrice
train_data.drop(['SalePrice'], axis=1, inplace=True)

# Select numeric columns only
numeric_cols = [cname for cname in train_data.columns if train_data[cna
me].dtype in ['int64', 'float64']]
X = train_data[numeric_cols].copy()
X_test = test_data[numeric_cols].copy()
```

Use the next code cell to print the first several rows of the data.

In [ ]:
```
X.head()
```

So far, you've learned how to build pipelines with scikit-learn. For instance, the pipeline below will use `SimpleImputer()` to replace missing values in the data, before using

`RandomForestRegressor()` to train a random forest model to make predictions. We set the number of trees in the random forest model with the `n_estimators` parameter, and setting `random_state` ensures reproducibility.

In [ ]:
```
from sklearn.ensemble import RandomForestRegressor
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer

my_pipeline = Pipeline(steps=[
    ('preprocessor', SimpleImputer()),
    ('model', RandomForestRegressor(n_estimators=50, random_state=0))
])
```

You have also learned how to use pipelines in cross-validation. The code below uses the `cross_val_score()` function to obtain the mean absolute error (MAE), averaged across five different folds. Recall we set the number of folds with the `cv` parameter.

In [ ]:
```
from sklearn.model_selection import cross_val_score

# Multiply by -1 since sklearn calculates *negative* MAE
scores = -1 * cross_val_score(my_pipeline, X, y,
                              cv=5,
                              scoring='neg_mean_absolute_error')

print("Average MAE score:", scores.mean())
```

Step 1: Write a useful function

In this exercise, you'll use cross-validation to select parameters for a machine learning model.

Begin by writing a function `get_score()` that reports the average (over three cross-validation folds) MAE of a machine learning pipeline that uses:

- the data in `X` and `y` to create folds,
- `SimpleImputer()` (with all parameters left as default) to replace missing values, and

- `RandomForestRegressor()` (with `random_state=0`) to fit a random forest model.

The `n_estimators` parameter supplied to `get_score()` is used when setting the number of trees in the random forest model.

In [ ]:
```
def get_score(n_estimators):
    """Return the average MAE over 3 CV folds of random forest model.

    Keyword argument:
    n_estimators -- the number of trees in the forest
    """
    my_pipeline = Pipeline(steps=[
        ('preprocessor', SimpleImputer()),
        ('model', RandomForestRegressor(n_estimators, random_state=0))
    ])

    scores = -1 * cross_val_score(my_pipeline, X, y,
                                  cv=3,
                                  scoring='neg_mean_absolute_error')
    return scores.mean()

# Check your answer
step_1.check()
```

In [ ]:
```
# Lines below will give you a hint or solution code
#step_1.hint()
#step_1.solution()
```

Step 2: Test different parameter values

Now, you will use the function that you defined in Step 1 to evaluate the model performance corresponding to eight different values for the number of trees in the random forest: 50, 100, 150, ..., 300, 350, 400.

Store your results in a Python dictionary `results`, where `results[i]` is the average MAE returned by `get_score(i)`.

In [ ]:
```
results = {}
for i in range(1,9):
    results[50*i] = get_score(50*i)



step_2.check()
```

In [ ]:
```
# Lines below will give you a hint or solution code
#step_2.hint()
#step_2.solution()
```

Use the next cell to visualize your results from Step 2. Run the code without changes.

In [ ]:
```
import matplotlib.pyplot as plt
%matplotlib inline

plt.plot(list(results.keys()), list(results.values()))
plt.show()
```

Step 3: Find the best parameter value

Given the results, which value for `n_estimators` seems best for the random forest model? Use your answer to set the value of `n_estimators_best`.

In [ ]:
```
n_estimators_best = min(results, key=results.get)

# Check your answer
step_3.check()
```

In [ ]:
```
# Lines below will give you a hint or solution code
#step_3.hint()
#step_3.solution()
```

In this exercise, you have explored one method for choosing appropriate parameters in a machine learning model.

If you'd like to learn more about hyperparameter optimization, you're encouraged to start with **grid search**, which is a straightforward method for determining the best *combination* of parameters for a machine learning model. Thankfully, scikit-learn also contains a built-in function `GridSearchCV()` that can make your grid search code very efficient!

Keep going

Continue to learn about **gradient boosting**, a powerful technique that achieves state-of-the-art results on a variety of datasets.

Have questions or comments? Visit the Learn Discussion forum to chat with other Learners.