Getting Started with Kaggle: Takeaways

Syntax

- Cutting bin values into discrete intervals:

```
pd.cut(np.array([1, 7, 5, 4, 6, 3]), 3)
```

- Splitting our data using scikit-learn:

```
from sklearn.model_selection import train_test_split
train_X, test_X, train_y, test_y = train_test_split(
all_X, all_y, test_size=0.2,random_state=0)
```

- Calculating accuracy using scikit-learn:

```
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(test_y, predictions)
```

- Calculating cross validation accuracy score using scikit-learn:

```
from sklearn.model_selection import cross_val_score
cross_val_score(estimator, X, y, cv=None)
```

Concepts

- Kaggle is a site where people create algorithms to compete against machine learning practitioners around the world.
- Each Kaggle competition has two key data files to work with — a training set and a testing set. The training set contains data we can use to train our model whereas the testing set contains all of the same feature columns, but is missing the target value column.
- Along with the data files, Kaggle competitions include a data dictionary, which explains the various columns that make up the data set.
- Acquiring domain knowledge is thinking about the topic you are predicting, and it's one of the most important determinants for success in machine learning.
- Logistic regression is often the first model you want to train when performing classification.
- The scikit-learn library has many tools that make performing machine learning easier. The scikit-learn workflow consists of four main steps:
    - Instantiate (or create) the specific machine learning model you want to use.
    - Fit the model to the training data.
    - Use the model to make predictions.
    - Evaluate the accuracy of the predictions.
- Before submitting, you'll need to create a submission file. Each Kaggle competition can have slightly different requirements for the submission file.
- You can start your submission to Kaggle by clicking the blue 'Submit Predictions' button on the competitions page.

Resources

- Kaggle
- Documentation for pandas.cut