**Project Overview**

In this project, you will apply basic machine learning concepts on data collected for housing prices in the Boston, Massachusetts area to predict the selling price of a new home. You will first explore the data to obtain important features and descriptive statistics about the dataset. Next, you will properly split the data into testing and training subsets, and determine a suitable performance metric for this problem. You will then analyze performance graphs for a learning algorithm with varying parameters and training set sizes. This will enable you to pick the optimal model that best generalizes for unseen data. Finally, you will test this optimal model on a new sample and compare the predicted selling price to your statistics.

Prepare for this project with our supplementary course in [**Model Evaluation and Validation**](https://classroom.udacity.com/courses/ud725-nd).

**Project Highlights**

This project is designed to get you acquainted to working with datasets in Python and applying basic machine learning techniques using NumPy and Scikit-Learn. Before being expected to use many of the available algorithms in the sklearn library, it will be helpful to first practice analyzing and interpreting the performance of your model.

Things you will learn by completing this project:

* How to use NumPy to investigate the latent features of a dataset.
* How to analyze various learning performance plots for variance and bias.
* How to determine the best-guess model for predictions from unseen data.
* How to evaluate a model’s performance on unseen data using previous data.

**Project Description**

The Boston housing market is highly competitive, and you want to be the best real estate agent in the area. To compete with your peers, you decide to leverage a few basic machine learning concepts to assist you and a client with finding the best selling price for their home. Luckily, you’ve come across the Boston Housing dataset which contains aggregated data on various features for houses in Greater Boston communities, including the median value of homes for each of those areas. Your task is to build an optimal model based on a statistical analysis with the tools available. This model will then be used to estimate the best selling price for your clients' homes.

For this assignment, you can find the boston-housing.zip file as a downloadable in the**Resources** section. You may also visit our [**Projects GitHub**](https://github.com/udacity/machine-learning) to have access to all of the projects available for this Nanodegree. While some code has already been implemented to get you started, you will need to implement additional functionality to successfully answer all of the questions included in the notebook. You can find the included questions for reference on the following slide. Unless requested, do not modify code that has already been included.

**Software and Libraries**

For this project, you will need to have the following software and Python libraries installed:

* [**Python 2.7**](https://www.python.org/download/releases/2.7/)
* [**NumPy**](http://www.numpy.org/)
* [**scikit-learn**](http://scikit-learn.org/stable/)
* [**matplotlib**](http://matplotlib.org/)
* [**iPython Notebook**](http://ipython.org/notebook.html)

**Deliverables**

The following files should be included as your submission, and can be packaged as a single .zip file:

* The boston\_housing.ipynb file with fully implemented, functional code, with all code blocks executed and showing output.
* An HTML or PDF report of the project (you may either export the notebook as HTML with your answers included, or submit a separate PDF with only the questions and your answers).

## Project Report Questions

A description of the Boston Housing dataset can be found [**here**](https://archive.ics.uci.edu/ml/datasets/Housing). Use this slide as reference to the project questions you will encounter in the notebook. These questions (and your answers) must be present in your submitted report.

### Data Exploration

This question is integrated into the project notebook output.  
Using the NumPy library, calculate a few meaningful statistics about the dataset:

* How many data points (houses) were collected?
* How many features are present for each house?
* What is the minimum housing price? The maximum?
* What is the mean housing price? The median?
* What is the standard deviation of all housing prices?

1) Using your intuition, for each of the three features present in the dataset, do you think that an increase in the value of that feature would lead to an increase in the value of 'MDEV' or a decrease in the value of 'MDEV'? Justify your answer for each.

### Developing a Model

2) For the hypothetical model presented in the project, would you consider this model to have successfully captured the variation of the target variable? Why or why not?

3) What is the benefit to splitting a dataset into some ratio of training and testing subsets for a learning algorithm?

### Analyzing Model Performance

4) Choose one of the learning curves graphs presented in the project and state the maximum depth for the model. What happens to the score of the training curve as more training points are added? What about the testing curve? Would having more training points benefit the model?

5) When the model is trained with a maximum depth of 1, does the model suffer from high bias or from high variance? How about when the model is trained with a maximum depth of 10? What visual cues in the graph justify your conclusions?

6) Which maximum depth do you think results in a model that best generalizes to unseen data? What intuition lead you do this answer?

### Evaluating Model Performance

7) What is the grid search technique and how can it be applied to optimize a learning algorithm?

8) What is the k-fold cross-validation training technique and how is it performed on a learning algorithm?

9) What maximum depth does the optimal model that you have implemented have? How does this result compare to your guess in Question 6?

10) What price would you recommend each client sell his/her home at, given the data presented in the project? Do these prices seem reasonable given the values for the respective features?

11) In a few sentences, discuss whether the constructed model should or should not be used in a real-world setting.