## **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://www.udacity.com/course/viewer#!/c-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 (and ultimately know) 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 used to estimate the best selling price for your client’s home.

For this assignment, download the boston\_housing.ipynb file in the **Downloadables** section below. 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.

You may also visit our [Projects GitHub](https://github.com/udacity/machine-learning) to have access to all of the projects for this Nanodegree.

## **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/)
* [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.
* A PDF report of the project (you may either save the notebook as a PDF with your answers, or submit a separate document 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 included PDF report.

### **Statistical Analysis and 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) Of the available features for a given home, choose three you feel are significant and give a brief description for each of what they measure.

2) Using your client’s feature set CLIENT\_FEATURES in the template code, which values correspond to the chosen features?

### **Evaluating Model Performance**

3) Why do we split the data into training and testing subsets?

4) Which performance metric below is most appropriate for predicting housing prices and analyzing error? Why?

* Accuracy
* Precision
* Recall
* F1 score
* Mean Squared Error (MSE)
* Mean Absolute Error (MAE)

5) What is the grid search algorithm and when is it applicable?

6) What is cross-validation and how is it performed on a model? Why would cross-validation be helpful when using grid search?

### **Analyzing Model Performance**

7) Choose one of the learning curve graphs your code creates. What is the max depth for the model? As the size of the training set increases, what happens to the training error? Describe what happens to the testing error.

8) Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?

9) From the model complexity graph, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?

### **Model Prediction**

*To answer the following questions, it is recommended that you run your notebook several times and use the median or mean value as your result.*

10) Using grid search, what is the optimal max depth for your model? How does this result compare to your initial intuition?

11) With your parameter-tuned model, what is the best selling price for your client’s home? How does this selling price compare to the statistics you calculated on the dataset?

12) In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients’ homes in the Boston area.

### **Evaluation**

Your project will be reviewed by a Udacity reviewer against [**this rubric**](https://docs.google.com/document/d/1b3u3HV2xMBAH7SQxdMPzr2QUaBJU8oeiCNdNh3AWLVY/pub?embedded=true). Be sure to review it thoroughly before you submit. All criteria must "meet specifications" in order to pass.

### **Submission**

When you're ready to submit your project go back to your [Udacity Home](https://www.udacity.com/me), click on Project 1, and we'll walk you through the rest of the submission process.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at **machine-support@udacity.com** or visit us in the [discussion forums](http://discussions.udacity.com/).

### **What's Next?**

You will get an email as soon as your reviewer has feedback for you. In the meantime, review your next project and feel free to get started on it or the courses supporting it!