**Project Requirement and Specification**

**on**

**Housing Price Prediction ML Project**

**(CSE ML 3rd Semester Mini project PCS-333)**

**2020-2021**

****

**Submitted by:** Pranjal Agarwal

Roll. No.: 2015301

CSE-ML-3rd -Sem

Session: 2020-2021

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA UNVERSITY, DEHRADUN**

* 1. ***About Project***

## Project Overview

In this project, we 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. We will first explore the data to obtain important features and descriptive statistics about the dataset. Next, we will properly split the data into testing and training subsets and determine a suitable performance metric for this problem. We will then analyse performance graphs for a learning algorithm with varying parameters and training set sizes. This will enable us to pick the optimal model that best generalizes for unseen data. Finally, we will test this optimal model on a new sample and compare the predicted selling price to our statistics.

## Project Highlights

This project is designed to get us 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 our model.

Things we 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.

## 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.

The dataset for this project originates from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Housing). The Boston housing data was collected in 1978 and each of the 506 entries represent aggregated data about 14 features for homes from various suburbs in Boston, Massachusetts. For the purposes of this project, the following preprocessing steps have been made to the dataset:

* 16 data points have an 'MEDV' value of 50.0. These data points likely contain **missing or censored values** and have been removed.
* 1 data point has an 'RM' value of 8.78. This data point can be considered an **outlier** and has been removed.
* The features 'RM', 'LSTAT', 'PTRATIO', and 'MEDV' are essential. The remaining **non-relevant features** have been excluded.
* The feature 'MEDV' has been **multiplicatively scaled** to account for 35 years of market inflation.

## Software and Libraries

This project uses the following software and Python libraries:

* [**Python**](https://www.python.org/download/releases/3.0/)
* [**NumPy**](http://www.numpy.org/)
* [**Pandas**](http://pandas.pydata.org/)
* [**Scikit-learn**](http://scikit-learn.org/stable/)
* [**Matplotlib**](http://matplotlib.org/)

We will also need to have software installed to run and execute a [Jupyter Notebook](http://ipython.org/notebook.html).

If you do not have Python installed yet, it is highly recommended that you install the [Anaconda](http://continuum.io/downloads) distribution of Python, which already has the above packages and more included.

## Starting the Project

For this project you can find the boston\_housing folder containing the necessary project files. Please make sure that you use the most recent version of project files when completing a project!

This project contains three files:

* boston\_housing1.ipynb: This is the main file where you will be performing your work on the project.
* housing.csv: The project dataset. You'll load this data in the notebook.
* Housing price prediction ML.

In the Terminal or Command Prompt, navigate to the folder containing the project files, and then use the command jupyter notebook boston\_housing.ipynb to open up a browser window or tab to work with your notebook. Alternatively, you can use the command jupyter notebook or ipython notebook and navigate to the notebook file in the browser window that opens. Follow the instructions in the notebook and answer each question presented to successfully complete the project.

### ----Install

This project requires **Python** and the following Python libraries installed:

* [NumPy](http://www.numpy.org/)
* [Pandas](http://pandas.pydata.org/)
* [matplotlib](http://matplotlib.org/)

### Code

Template code is provided in the boston\_housing1.ipynb notebook file. You will also be required to use the housing.csv dataset file to complete your work. While some code has already been implemented to get you started, you will need to implement additional functionality when requested to successfully complete the project.

### Run

In a terminal or command window jupyter notebook boston\_housing1.ipynb file,

This will open the Jupyter Notebook software and project file in your browser.

**Features**

1. RM: average number of rooms per dwelling
2. LSTAT: percentage of population considered lower status
3. PTRATIO: pupil-teacher ratio by town

**Target Variable** 4. MEDV: median value of owner-occupied homes

* 1. ***Requirement of Project***
     1. **Hardware Requirement**

1. Memory and disk space required per user: 512MB RAM + 1GB of disk + .5 CPU core.
2. Server overhead: 2-4GB or 10% system overhead (whatever is larger), .5 CPU cores, 10GB disk space.
3. Port requirements: Port 8000.

**1.2.2 Software Requirement**

**--Latest version of Anaconda navigator----------------------------------**

**Modules of Project**

Data Exploration

Implementation: Calculate Statistics

### Feature Observation

**Visualization**

Developing a Model  Define a Performance Metric

Shuffle and Split Data

### Training and Testing

## Analysing Model Performance

## Evaluating Model Performance

**REFERENCE**

1. MACHINE LEARNING, Tom M. Mitchell (Indian edition)