HOUSING USE CASE STUDY PROJECT

NAME: Sucharitha A M

BATCH:16 internship batch

Problem statement:

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know: • Which variables are important to predict the price of variable? • How do these variables describe the price of the house?

Problem definition:

The housing market is highly competitive, and I want to be the best real estate agent in the area. To compete with my peers, I decide to leverage a few basic machine learning concepts to assist myself and my client with finding the best-selling price for their home. Luckily, I've come across the Housing dataset which contains aggregated data on various features for houses in Greater communities, including the median value of homes for each of those areas. My 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 my client's homes.

[](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/logo.png)

**Table Of Contents:**

* [Description](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#description)  
  + [About the project](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#about-the-project)
  + [What needs to be done](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#what-needs-to-be-done)
  + [Why this project](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#why-this-project)
* [Data](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#data)  
  + [Files](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#files)
  + [Dataset file](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#dataset-file)
* [Loading Project](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#loading-project)  
  + [Requirements](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#requirements)
  + [Execution](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#execution)
* [Conclusion](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#conclusion)  
  + [What I learned](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#what-i-learned)
  + [Evaluation](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#evaluation)
  + [Results](https://github.com/JuzerShakir/Predicting_Boston_Housing_Prices/blob/master/README.md#results)

### Description

#### About the project

The housing market is highly competitive, and I want to be the best real estate agent in the area. To compete with my peers, I decide to leverage a few basic machine learning concepts to assist myself and my client with finding the best-selling price for their home. Luckily, I've come across the Housing dataset which contains aggregated data on various features for houses in Greater communities, including the median value of homes for each of those areas. My 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 my client's homes.

#### What needs to be done

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

#### Why this project

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.

### Data

#### Files

This project contains three files:

* report. ipynb: This is the main file where I have performed my work on the project.
* data.csv: The project dataset. I'll load this data in the notebook.
* visuals.py: This Python script provides supplementary visualizations for the project.
* export/: Folder containing HTML and PDF version file of notebook.

Template code is provided in the report. ipynb notebook file. I will be required to use the included visuals.py Python file and the data.csv dataset file to complete my work. While some code has already been implemented to get me started, I'll need to implement additional functionality when requested to successfully complete the project.

#### Dataset file

The dataset for this project originates from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Housing). The housing data was collected in 1460 and each of the 81entries represent aggregated data about 80 features for homes from various suburbs in Boston, Massachusetts.

For the purposes of this project, the following pre-processing steps have been made to the dataset:

* 16 data points have an 'ID' value of 50.0. These data points likely contain **MSSubClass or censored values** and have been removed.
* 1 data point has an 'LotArea' value of 8.78. This data point can be considered an **outlier** and has been removed.
* The features 'Id', 'MSSubClass', 'LotArea', and 'MSZoning' 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.

**Features**

1. ID: average number of rooms per dwelling
2. LotArea: percentage of population considered lower status
3. MSZoning: pupil-teacher ratio by town

**Target Variable**

1. Lot Frontage: median value of owner-occupied homes

### Loading Project

#### Requirements

This project requires **Python 3.6.5** and the following Python libraries installed:

* [Python 3.6.5](https://www.python.org/downloads/release/python-365/)
* [NumPy](http://www.numpy.org/)
* [Pandas](http://pandas.pydata.org/)
* [matplotlib](http://matplotlib.org/)
* [scikit-learn](http://scikit-learn.org/stable/)

You will also need to have software installed to run and execute a [Jupyter Notebook](http://jupyter.org/install)

If you do not have Python installed yet, it is highly recommended that you install the [Anaconda](https://www.anaconda.com/download/) distribution of Python, which already has the above packages and more included.

#### Execution

In a terminal or command window, navigate to the top-level project directory Predicting\_housing\_case\_study and run one of the following commands:

ipython notebook report.ipynb

or

jupyter notebook report.ipynb

or if you have 'Jupyter Lab' installed

jupyter lab

### Conclusion

#### What I learned

Things I've learned by completing this project:

* How to use NumPy to investigate the latent features of a dataset.
* How to analyse 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.