Architectural Decisions Document (ADD)

The Lightweight IBM Cloud Garage Method for Data Science

Project:

House Price Prediction – To predict house price.

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

The heart of the this project is the small dataset “Measurements on Petroleum Rock Samples” distributed by <http://vincentarelbundock.github.io/Rdatasets/> , which “is a collection of over 1300 datasets that were originally distributed alongside the statistical software environment R and some of its add-on packages.” See documentation: <http://vincentarelbundock.github.io/Rdatasets/doc/datasets/rock.html>

The heart of this project is from the Kaggle [House Prices: Advanced Regression Techniques]

Dataset. This is the link: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview>. This is the Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Justification

The data set is compiled by Dean De Cock for use in data science education.

## Enterprise Data

### Technology Choice

We will use Jupyter Notebooks on local server to analyze the data and run models.

### Justification

Jupyter Notebooks are sufficient for this small project that does not require deployment. The Jupyter Notebooks can be shared with other Data Scientists in the department.

## Streaming analytics

### Technology Choice

The data set is a finished product that does not require real-time streaming.

### Justification

The data set is historical and was generated in a small-batch research workflow, not in an automated production workflow allowing for streaming.

## Data Integration

### Technology Choice

No data integration applied.

### Justification

The draw data set used for this project comes from a single csv file and was already integrated.

## Data Repository

### Technology Choice

The data is stored as csv file and can be loaded to the Jupyter Notebook in the local server. After completion, it can be uploaded to github repository as well.

### Justification

The data is light enough to be stored and loaded either way described above.

## Discovery and Exploration

### Technology Choice

An ETL notebook was created analyzing with Pandas data frames and MatPlotLib. A more detailed exploration was started in the main notebook with Pandas data frames and seaborn plots.

### Justification

The data are given as continuous values in a csv file perfectly suited for Pandas data frames.

## Actionable Insights

### Technology Choice

The data set required scaling due to strongly different value ranges. The SciKit-Learn functions for preprocessing were used.

### Justification

The SciKit-Learn library for Python is the gold standard for preprocessing and machine learning.

## Applications / Data Products

### Technology Choice

Pandas and Numpy

SciKit-Learn (in-memory Machine Learning)

Keras/ Tensorflow (in-memory Deep Learning)

### Justification

The above technology and library is actually the de factor standard tools that are used in machine learning model building.

## Security, Information Governance and Systems Management

### Technology Choice

The completed Notebook/final product will be uploaded to github repository.

### Justification

Github is a great sharing platform and has great technology in terns of security, Information Governance and Systems Management.