Exercises

Level 10: Data Research

All the exercises in this level should be done using Anaconda/JupyterLab. Each exercise should be in its own Notebook, with subparts as individual cells. You should submit the actual Notebooks (.ipynb) files from Jupyter.

For the below exercises, you should find your own datasets; any online source is ok, and you should include your datasets with your submission as CSV files. You should find datasets other than what is provided from the lectures for this level. The criteria for the data are:

1. Either timeseries or cross-sectional data are fine.
2. There should be at least three independent variables (X) and a single variable to predict (Y).
3. There should be at least 100 rows, but preferably 1000.
4. You can either have one dataset that would can be used for both a classification or regression use case (as demonstrated in the lecture), or two separate datasets.

10.1: Data Cleaning/Bootsrapping

1) Perform data cleaning on the downloaded data. If the data is already clean, add a bunch of dummy 'bad' rows and columns which you can then demonstrate how to properly clean the data. Should perform at least the following (and other examples you can think of):

- **a.** Drop Nulls
- **b.** Remove irrelevant columns
- **c.** Standardize a date/time column
- **d.** Standardize a string column
- **e.** Remove outliers
- **f.** Winsorize outliers using both clip and a winsorize function.

2) Write a function that will bootstrap any dataset, for a given number of bootstraps, as demonstrated in the lectures. Use the function on the downloaded data, and perform the following:

- **a.** Calculate the mean, variance, mean, 5^{th} percentile, 95^{th} percentile, standard deviation, and standard error of the bootstrapped data.
- **b.** Compare against the non-bootstrapped values.
- **c.** Plot a box and whisker chart of the bootstrapped statistics.

10.2: Intro to Machine Learning

1) Use the dataset to perform a regression prediction using scikit-learn, as demonstrated in the lectures. You should split the data into test/train sets, train the model (output/comment the scores), cross validate the model (output/comment the scores), and predict using the test set (output/comment the scores and actual accuracy).

2) Use the dataset to perform a classification prediction using scikit-learn, as demonstrated in the lectures. You should split the data into test/train sets, train the model (output/comment the scores), cross validate the model (output/comment the scores), and predict using the test set (output/comment the scores and actual accuracy).