**Mark Sausville**

Verify existing empirical predictive models with neural networks

**Instructions for script**: ***v\_script.py:***

**Please run from the command line like this: 'ipthon v\_script.py'.**

**Running the script will use the models to compute the predicted values for DO, NO3, NH4 and PO4 and then store them in the current directory. In addition, it will display graphs comparing actual values and predicted values for the validation data.**

**Description of contents:**

* All four trained models are in the directory **'./models905'**. Note that these are directories. The models can be loaded with:

**model = tf.keras.models.load\_model('dir\_path')**

* + do\_dnn\_model
  + nh4\_dnn\_model
  + no3\_dnn\_model
  + po4\_dnn\_model
* The current directory contains the notebooks that build, train and save the models. All models were trained using **'./data905/training905.xlsx'** as training data.
  + model\_builder\_DO.ipynb
  + model\_builder\_NO3.ipynb
  + model\_builder\_NH4.ipynb
  + model\_builder\_PO4.ipynb
* The directory  **./val\_predictions'** contains pre-calculated predictions for each target. The features come from **'./data905/validation905.xlsx'.** 
  + dnn\_predicted\_do.csv
  + dnn\_predicted\_no3.csv
  + dnn\_predicted\_nh4.csv
  + dnn\_predicted\_po4.csv

**Thoughts on results:**

* **Predictions on validation are not great.**

The error and r-squared are much worse than validation during training. There are several possible issues:

* + The training data is quite unlike the validation data:
    - Many time gaps between observations in the validation data
    - Outliers more prevalent in the validation data
    - The time base doesn't match between training and validation data (4 obs per 15 min in training data vs. 1 obs per 15 min in the validation data)
* **What to do about it**
  + Time-series techniques

The description in the contract led me to believe that this was a standard regression, but it is, in fact, a time-series problem. There are specific techniques used to approach problems where data depends on past values of some of the variables.

After familiarizing myself with the data, I believe these techniques are applicable and very promising. If I have the opportunity to go further with this problem in a subsequent contract, I would review my past work with time series techniques and understand how they could be applied.

* + Data quality

I believe that the outlier, missing data and time-base issues are severely degrading model performance. These would need to be addressed for any usable model.

* + Model tuning

Model tuning (hyper-parameter optimization) can have a significant impact on performance. Based on my examination of the graphical output comparing predicted and actual values, I recommend applying time-series techniques, followed by tuning.