Data challenge - Information and instructions

The goal of this application is to predict a single numeric continuous response (Y). Here is a description of the dataset provided in xlsx file:

- The training set is in sheet CAL500. It contains 500 samples (rows) for which:
  - The known response value in 1st column
  - The predictors (from 2nd column to the last). The predictors are spectral measurements (interaction of a sample and light). The light wavelengths are reported in the headings of the column: the spectrum is captured between 1100 and 2498 nm (with a step of 2 nm) i.e. 700 columns
- The validation set is in sheet VAL HSI 80 x 400. It contains 80 samples for which:
  - Y value is missing, to be predicted
  - Each sample is named as SXXX (where XXX ranges from 001 to 080)
  - Each sample has 400 measurements identified as SXXX-YYY (where YYY ranges from 001 to 400). For example, S001 has 400 spectrums identified as S001-001 to S001-400.
  - Note that the device used to measure the test samples is different from the one used for the training samples. The test samples have been measured on a wavelength range between 1118.19 to 2424.92 nm (with a step of 6.28 nm) i.e. 209 columns.

Instructions:

- In Python, using the training set, develop a predictive model and assess its performance
- Predict the response value for the 80 samples (1 response value per sample i.e. you must get 1 unique predicted value for each group of 400 spectra per sample).
- Submit your answer to samd.guizani.pro@gmail.com. We require that you submit:
  - Your predictions for the test set in xls or csv file (1 column of 80 values for samples from S001 to S080). You can use the provided xslx template (Predictions_YourName.xlsx)
  - Your Python files (Jupyter notebooks and/or scripts)
  - A short presentation (max. 10 slides) explaining your model development strategy and results, in particular:
    - What data preprocessing/cleaning/modifications/transformations… did you apply on the train and/or the test datasets?
    - Which models have you tried?
    - How did you measure their performance and select the best model?