Subject: Assessing A Machine Learning Method's Predictivity
Posted by Christophe on Tue, 01 Mar 2022 12:56:02 GMT
View Forum Message <> Reply to Message

Hello everyone,

In the User Manual at the “Assessing A Machine Learning Method's Predictivity” part, I can read:

“The dataset is divided into ten fractions along the time axis. A model is build with the first fraction and used to predict the property of the second fraction's molecules. Then a second model is built from the first two fractions, which is then applied to predict the third fraction. This continues until the nineth model is built from the first nine fractions and used to predict the tenth and last fraction.”

When I apply this to a case, by clicking “Machine Learning” and then “Assess Prediction Quality” the nine linear regressions I get, “predicted vs observed”, each only contain one Time Id set of data !!!
For example if I split my data set into ten fractions according a Time Id column, I get 9 regression models containing for example prediction fraction 3 vs 2 ; 3 vs 3 ; 3 vs 4 … 3 vs 10

From the user manual I would expect “predicted vs observed” as 2 vs (3) then 2 vs (3+4) then 2 vs (3+4+5 …). Of course the number of data from the (3), (3+4) and (3+4+5…) set should equals the number contained into 2.
But in that case the differences with "use random fractions instead of time based ones" would go thinner from (3) to (3+...+9)

Did I miss something? For me, dividing the data set into 10 fractions along the time axis serves precisely to take account of batches that are very different from one another

All the best