@maxchu the moment you use the validation set for anything (e.g. learning ensemble weights, early stopping on your training base learners, picking tree architecture, etc) it ceases to become an out-of-sample and the test results are not meaningful. If you want to utilize the validation data for live models, I would suggest:

- first cut the training set into train-train, and train-validation
- automate your training, optimization, architecture selection there
- feed the validation set “once” to this pipeline, and if you like the results
- rebuild the models with the same pipeline using the train and validation sets and deploy

I think the MLDP book had a section on quantifying how much each “look” at your out-of-sample test reduces the reliability of the test results.