11/22/2018

Baseline model (All the base features + time)

The RMSE on the validation set is 0.15811 (without One Hot Max Size) with convergence around 650

The RMSE on the validation set is 0.15831 (with One Hot Max Size) with convergence around 650

This is understandable because with OHMS > 2, we are one\_hot\_encoding many features, sometimes useless for this task, and by this way we are increasing variance (the devil is in details)!

INTERESTING FINDINGS: The validation error is always LOWER than the train error. Searching on the web, this can happen basically if the validation is “easier” compared with the train, and this may be possible

* Baseline model with rare features cleaned (All the base features + time)

The RMSE on the validation set is 0.15812 (without One Hot Max Size) with convergence around 450

This may be not influent, but a faster convergence may be an interesting discover

INTERESTING FINDINGS: Like in the previous case, the validation error is always lower than the train error (This must not astonish us, the validation set is the same)

* Doing or not the label encoding doesn’t change the feature ranking order. Now I have to check if the same is true for the score