Intermediate Machine Learning: 4th lesson – Pipelines

Pipelines are a simple way to keep the pre-processed data and organize code modeling. Specifically, a pipeline bundles pre-processing and modeling steps so the user can use the whole bundle as if it were a single step.

Many data scientists hack together models without pipelines, but pipelines have some important benefits, including:

* Cleaner code

Accounting for data at each step of preprocessing can get messy. With a pipeline, you won't need to manually keep track of your training and validation data at each step.

* Fewer bugs

There are fewer opportunities to misapply a step or forget a preprocessing step.

* Easier to productionize

It can be surprisingly hard to transition a model from a prototype to something deployable at scale. Users won't go into the many related concerns, but pipelines can help.

* More options for model validation

Includes cross-validation (e.g. resampling, sample splitting), various similar model validation techniques for assess how the results of a statistical analysis will generalize to an independent data set, which use different portions of the data to test and train a model on different iterations.