Course 4 · Week 2 — ML done honestly

Cheatsheet — biostats_courses

R. Heller

Tree-based models

Model	R	Strengths
CART	`rpart::rpart`	interpretable single tree
Random forest	`ranger::ranger`	robust, OOB error, variable importance
XGBoost	`xgboost::xgb.train`	top tabular performance, careful tuning
LightGBM	`lightgbm::lgb.train`	fast on large data

```r
library(ranger)
fit <- ranger(y ~ ., data = df, importance = "permutation",
              num.trees = 500, mtry = floor(sqrt(ncol(df) - 1)))
fit$prediction.error   # OOB error
```

Interpretability

```r
library(DALEX); library(iml)

ex  <- explain(fit, data = df |> select(-y), y = df$y)
ip  <- model_parts(ex)           # permutation importance
pd  <- model_profile(ex, variables = "x1")   # partial dependence
sh  <- Shapley$new(Predictor$new(fit, data = df, y = df$y),
                   x.interest = df[1, ])     # SHAP
```

PDP assumes feature independence; if features are correlated, use ALE plots.

Tabular neural networks with `torch`

```r
# sketch only — typically in a loop on GPU
library(torch)
nn <- nn_sequential(
  nn_linear(p, 64), nn_relu(),
  nn_linear(64, 1)
)
opt <- optim_adam(nn$parameters)
```

On small biomedical tabular data, boosted trees usually beat NN. Use NN when you need end-to-end training with images, sequences, or text.

tidymodels pipeline

```r
library(tidymodels)
rec  <- recipe(y ~ ., data = df) |>
  step_normalize(all_numeric_predictors()) |>
```

```r
  step_dummy(all_nominal_predictors())

mod  <- rand_forest(mtry = tune(), trees = 500) |>
  set_engine("ranger") |> set_mode("classification")

wf   <- workflow() |> add_recipe(rec) |> add_model(mod)
cv   <- vfold_cv(df, v = 5, strata = y)
tuned <- tune_grid(wf, cv, grid = 20)
collect_metrics(tuned)
```

Decision rule for Week 2

- Tabular, < 100k rows ⟶ gradient-boosted trees.
- Need explanations ⟶ SHAP + PDP, validate with held-out.
- Images / sequences ⟶ CNN / transformer, use `torch`.
- Reproducibility ⟶ `tidymodels` pipeline + fixed seed + locked recipe.

Common pitfalls

- Reporting feature importance from a model trained on the full data.
- Interpreting PDPs when features are strongly correlated.
- Claiming NN superiority without CV on the same split as a tree model.
- Forgetting to set a seed for any tune / split operation.

Further reading

- Biecek & Burzykowski, *Explanatory Model Analysis*.
- Molnar, *Interpretable Machine Learning*.