CSX415
Data Science Principals and Practice Models

Christopher Brown
U.C. Berkeley / Decision Patterns LLC
Spring 2018

Berkeley
UNIVERSITY OF CALIFORNIA

Caret

"Misc functions for training and plotting classification and regression models."

- Really:
  - Wraps 100's of modeling functions
  - Automates tediousness of model building
  - Manages a process

- Competitors:
  - [Rattle](#) : Graham Williams et al./Togaware.com
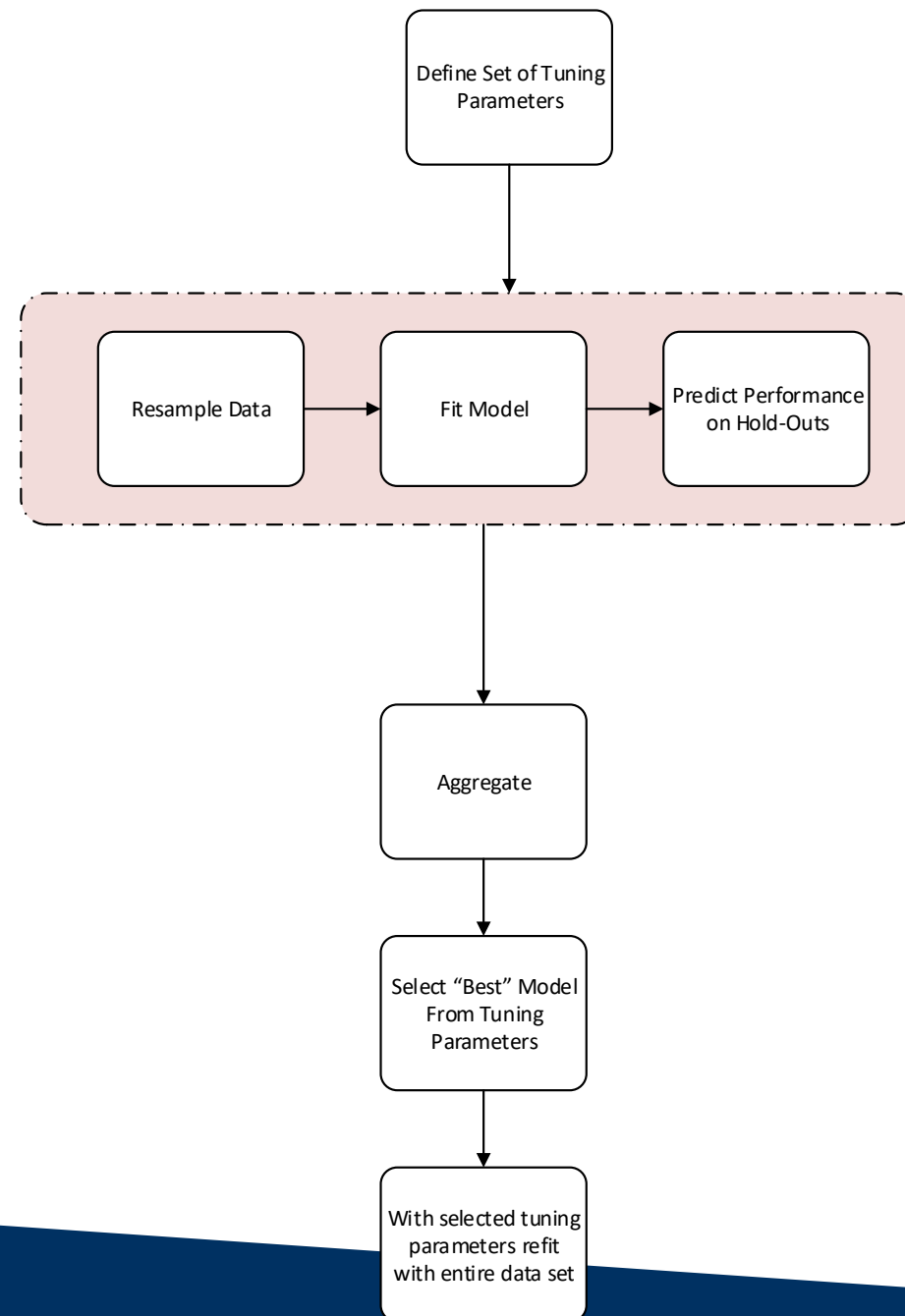 - [R Commander](#) : Statistical workbench

Caret Goals

Does a couple things:

- Preprocess data (transfroms, imputes)
- evaluate, using resampling, the effect of model tuning parameters on performance
- choose the "optimal" model across these parameters
- estimate model performance from a training set
- Variable Importance
- Aids feature selection

Berkeley
UNIVERSITY OF CALIFORNIA

Process

1 Define sets of model parameter values to evaluate
2 **for** *each parameter set* **do**
3   **for** *each resampling iteration* **do**
4   Hold–out specific samples
5   [Optional] Pre–process the data
6   Fit the model on the remainder
7   Predict the hold–out samples
8   **end**
9   Calculate the average performance across hold–out predictions
10 **end**
11 Determine the optimal parameter set
12 Fit the final model to all the training data using the optimal parameter set

Define Set of Tuning Parameters

Resample Data → Fit Model → Predict Performance on Hold-Outs

Aggregate

Select "Best" Model From Tuning Parameters

With selected tuning parameters refit with entire data set

Lots of Configurations

- Easy if you know what you are doing
- which `method`?

[Caret Model List](#)*

- Controlled mostly through
  - `train(tuneLength, tuneGrid)`
  - `trainControl - supplied to train`