Welcome to the Toolbox

MACHINE LEARNING TOOLBOX

Max Kuhn

Software Engineer at RStudio and creator of caret

Supervised Learning

- R `caret` package
- Automates *supervised learning* (a.k.a. predictive modeling)
- Target variable

Supervised Learning

- Two types of predictive models
  - Classification ⇒ Qualitative
  - Regression ⇒ Quantitative
- Use *metrics* to evaluate models
  - Quantifiable
  - Objective
- *Root Mean Squared Error* (RMSE) for regression

Evaluating Model Performance

- Common to calculate in-sample RMSE
  - Too optimistic
  - Leads to overfitting
- Better to calculate out-of-sample error (a la caret)
  - Simulates real-world usage
  - Helps avoid overfitting

In-sample error

```
# Fit a model to the mtcars data
data(mtcars)
model <- lm(mpg ~ hp, mtcars[1:20, ])
```

```
# Predict in-sample
predicted <- predict(
  model, mtcars[1:20, ], type = "response"
)
```

```
# Calculate RMSE
actual <- mtcars[1:20, "mpg"]
sqrt(mean((predicted - actual) ^ 2))
```

```
3.172132
```

Let's practice!

MACHINE LEARNING TOOLBOX

Out-of-sample error measures

MACHINE LEARNING TOOLBOX

Zach Mayer

Data Scientist at DataRobot and co-author of caret

Out-of-sample error

- Want models that don't overfit and generalize well
- Do the models perform well on new data?
- Test models on new data, or a test set
  - Key insight of machine learning
  - In-sample validation almost guarantees overfitting
- Primary goal of caret and this course: don’t overfit

Example: out-of-sample RMSE

```
# Fit a model to the mtcars data
data(mtcars)
model <- lm(mpg ~ hp, mtcars[1:20, ])
```

```
# Predict out-of-sample
predicted <- predict(
  model, mtcars[21:32, ], type = "response"
)
```

```
# Evaluate error
actual <- mtcars[21:32, "mpg"]
sqrt(mean((predicted - actual) ^ 2))
```

```
5.507236
```

Compare to in-sample RMSE

```
# Fit a model to the full dataset
model2 <- lm(mpg ~ hp, mtcars)
```

```
# Predict in-sample
predicted2 <- predict(
  model, mtcars, type = "response"
)
```

```
# Evaluate error
actual2 <- mtcars[, "mpg"]
sqrt(mean((predicted2 - actual2) ^ 2))
```

```
3.74
```

Let's practice!

MACHINE LEARNING TOOLBOX

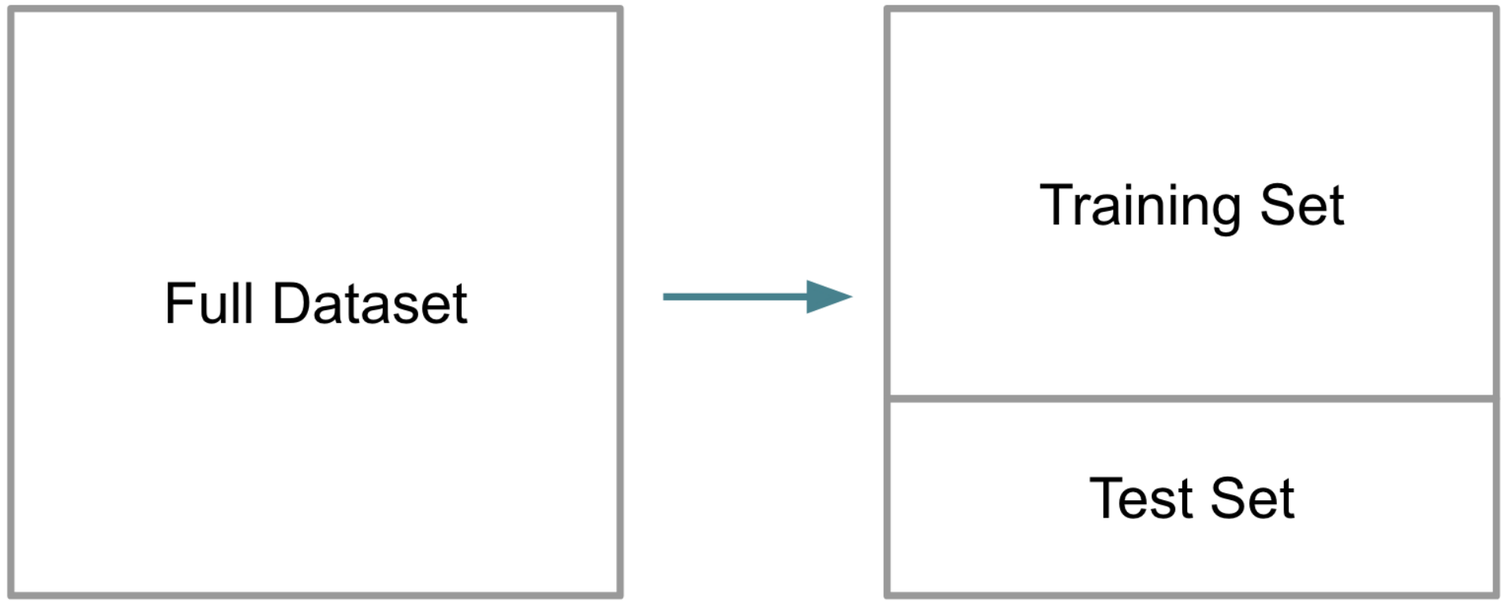
Cross-validation

MACHINE LEARNING TOOLBOX

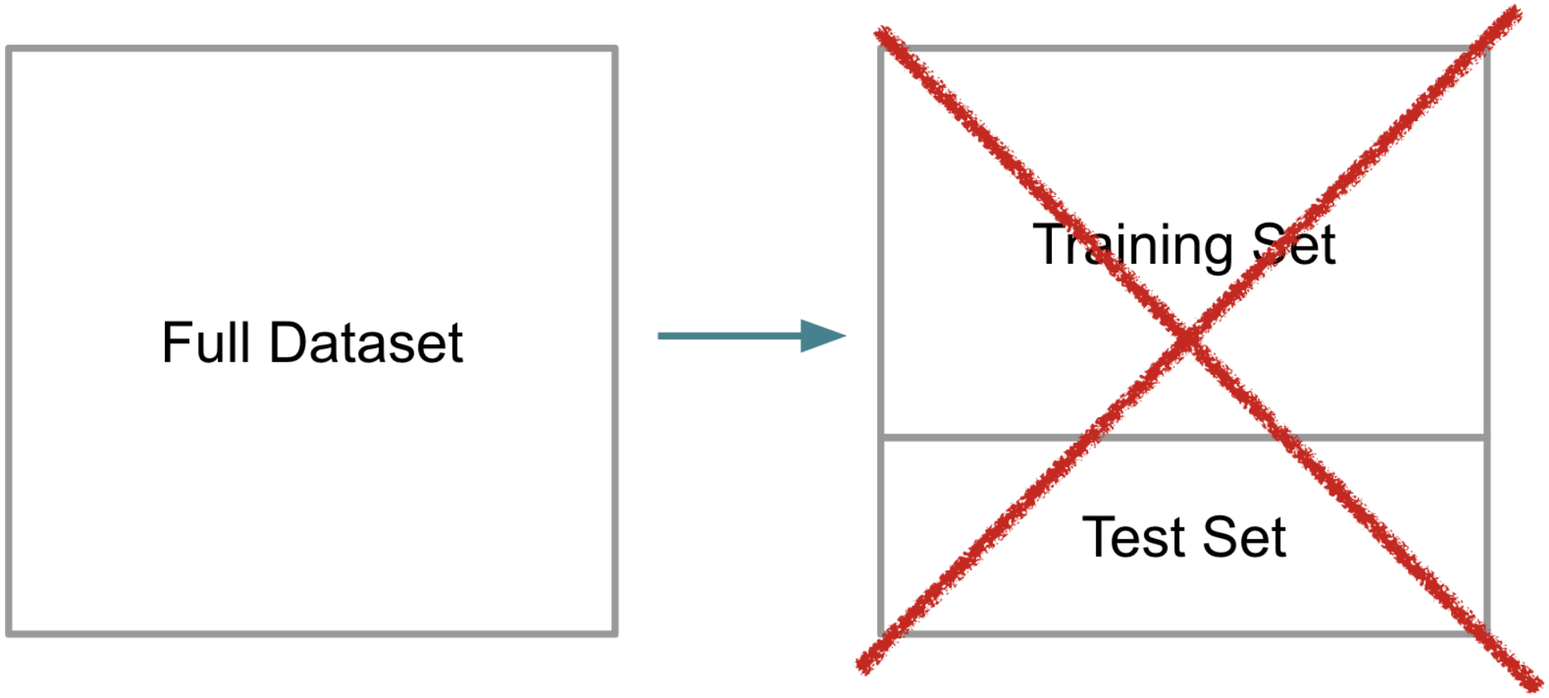
Zach Mayer

Data Scientist at DataRobot and co-author of caret

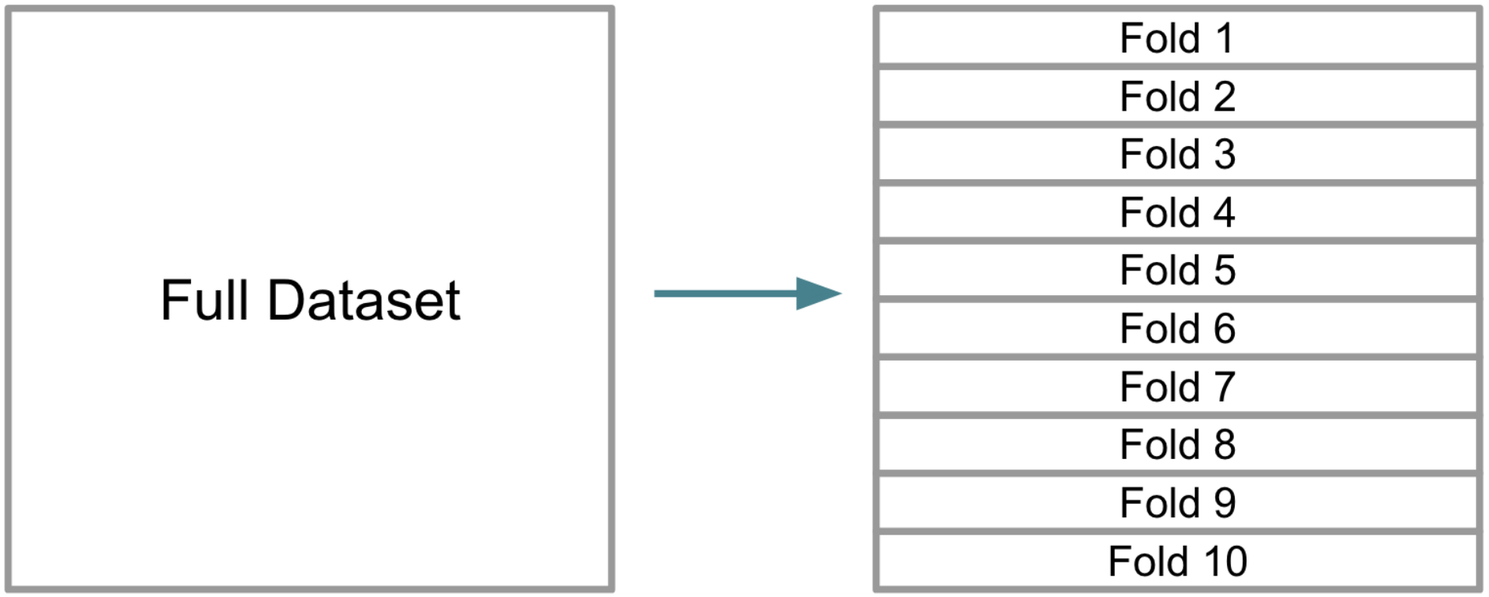
Cross-validation



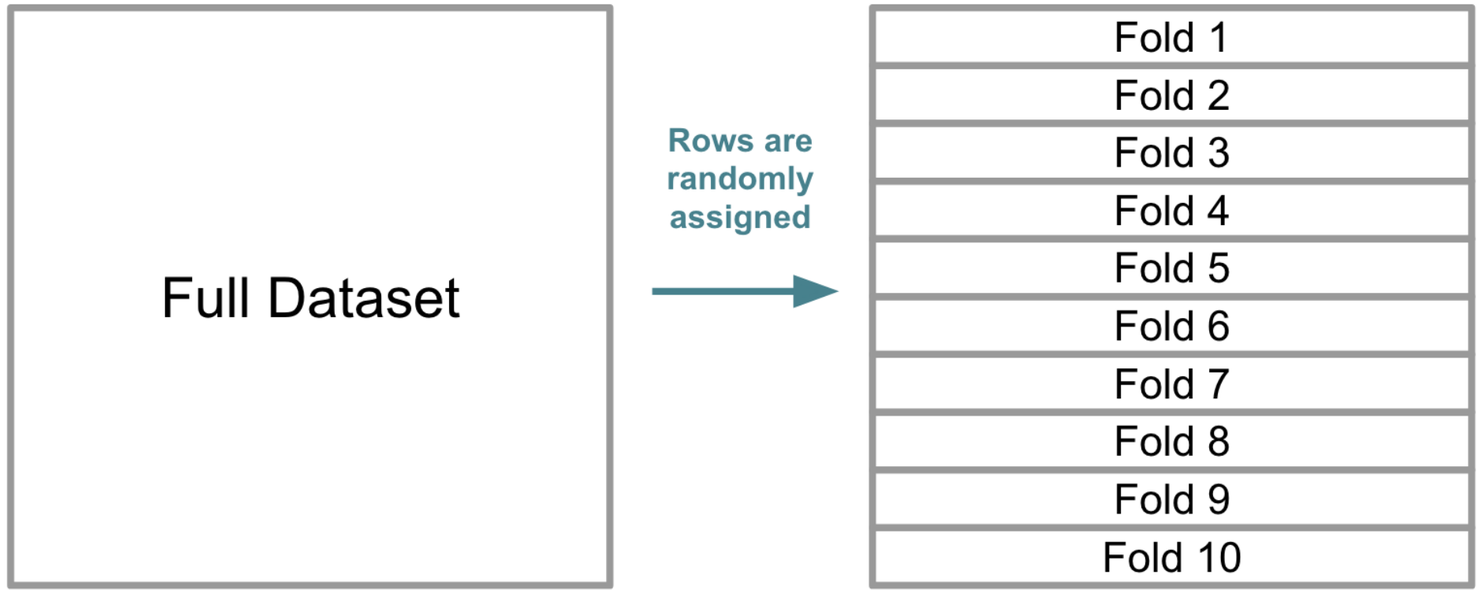
Cross-validation



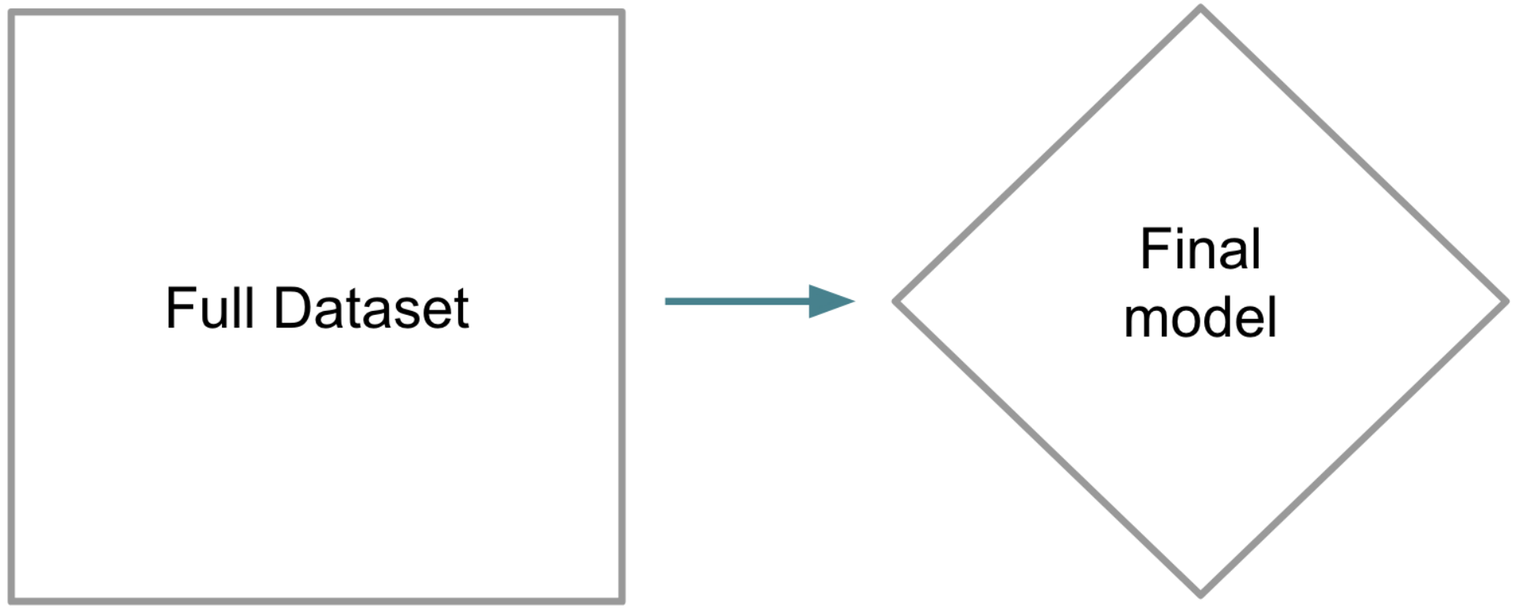
Cross-validation



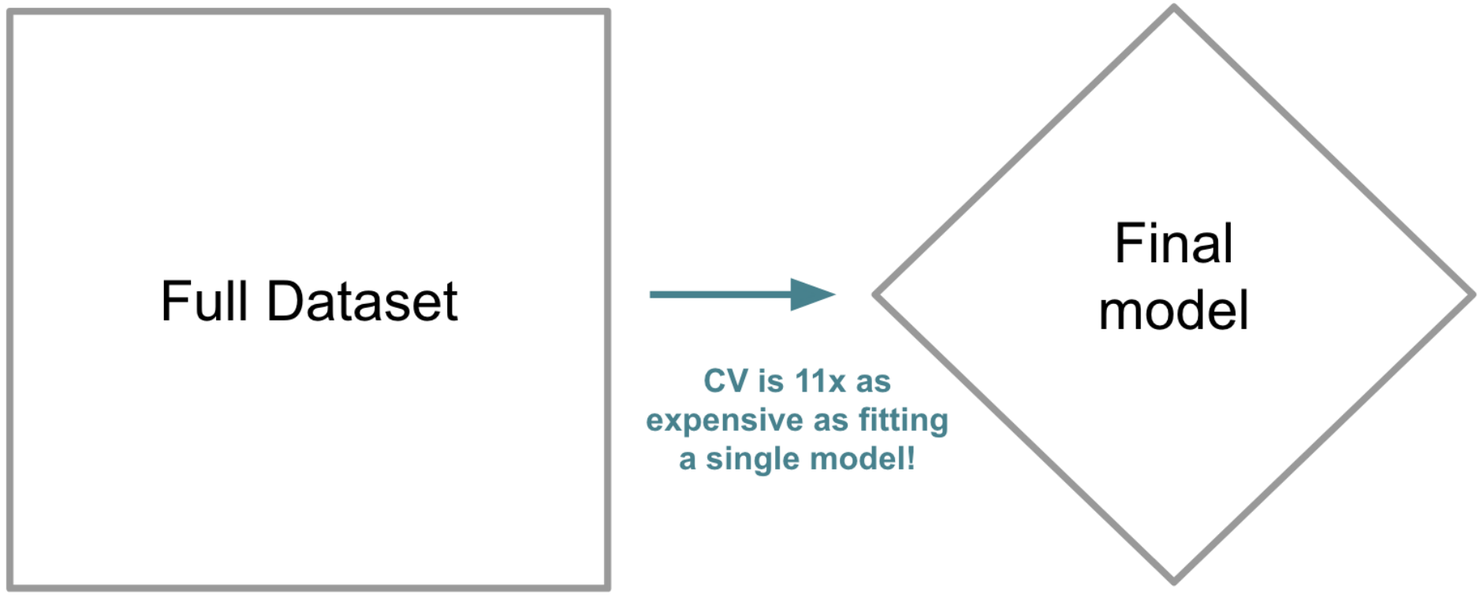
Cross-validation



Fit final model on full dataset



Fit final model on full dataset



Cross-validation

```
# Set seed for reproducibility
set.seed(42)
```

```
# Fit linear regression model
model <- train(
  mpg ~ hp, mtcars,
  method = "lm",
  trControl = trainControl(
    method = "cv",
    number = 10,
    verboseIter = TRUE
  )
)
```

```
+ Fold01: intercept=TRUE
- Fold01: intercept=TRUE
...
- Fold10: intercept=TRUE
Aggregating results
Fitting final model on full training set
```

Let's practice!

MACHINE LEARNING TOOLBOX