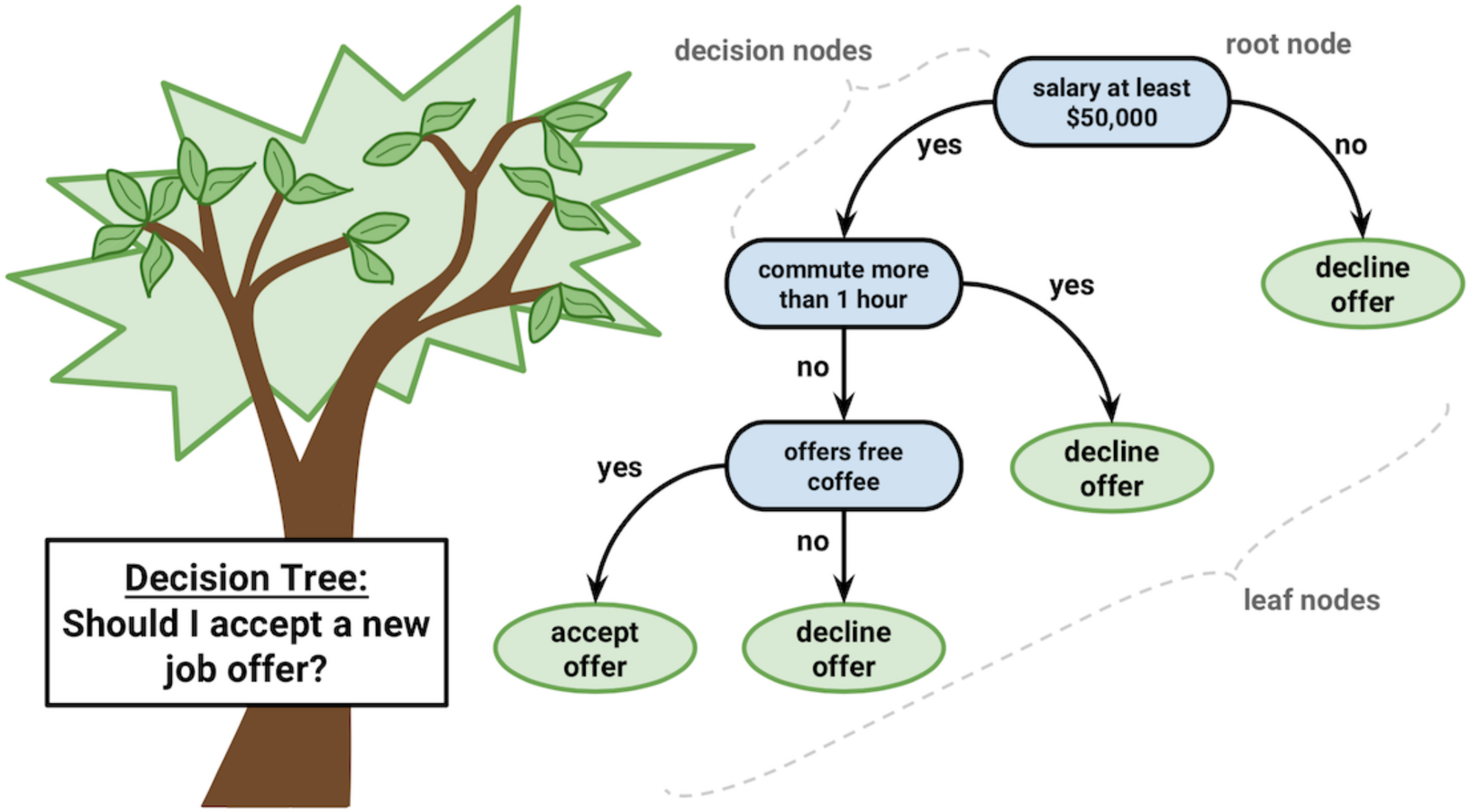
Making decisions with trees

SUPERVISED LEARNING IN R: CLASSIFICATION

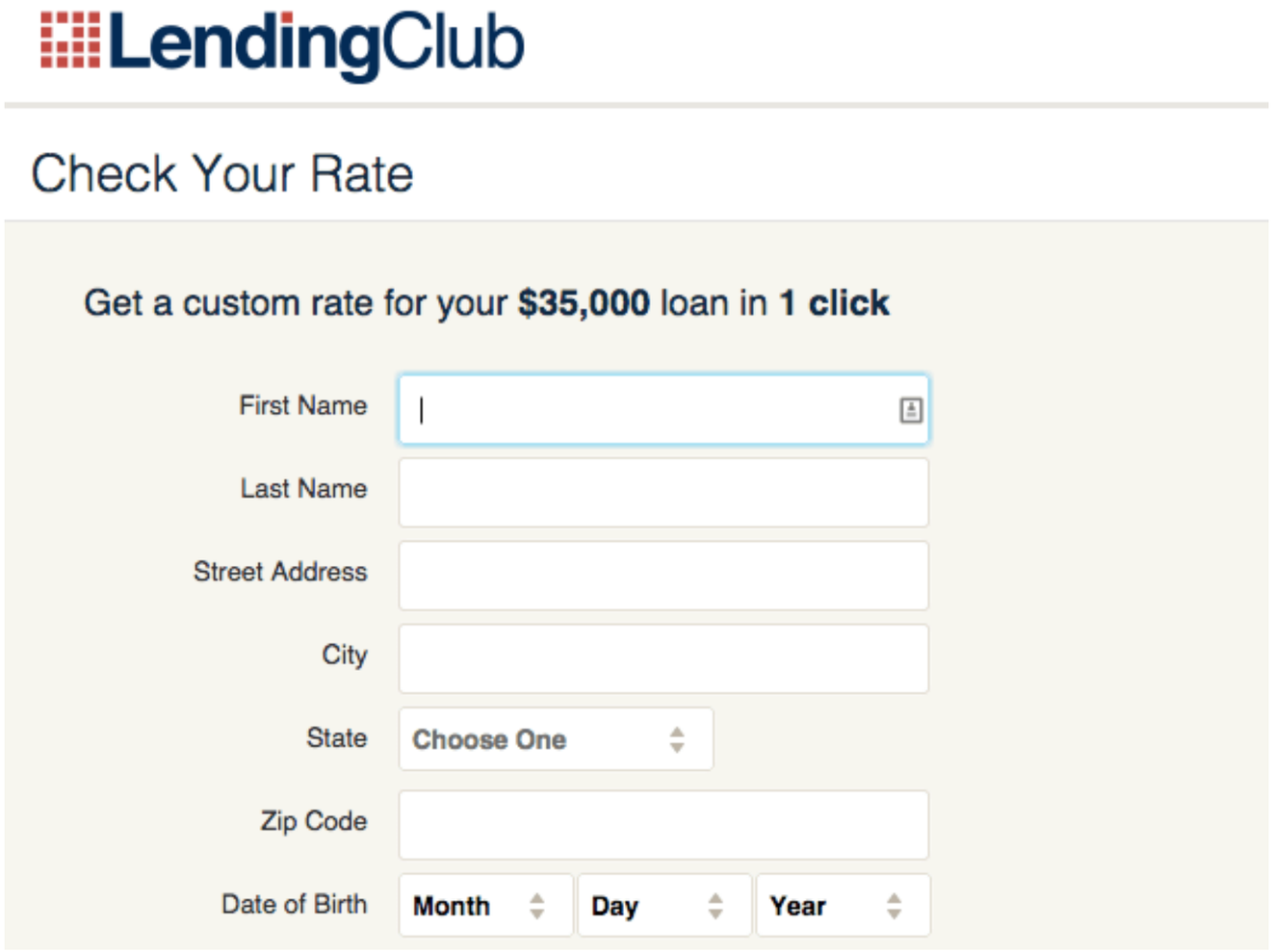
Brett Lantz
Instructor

datacamp

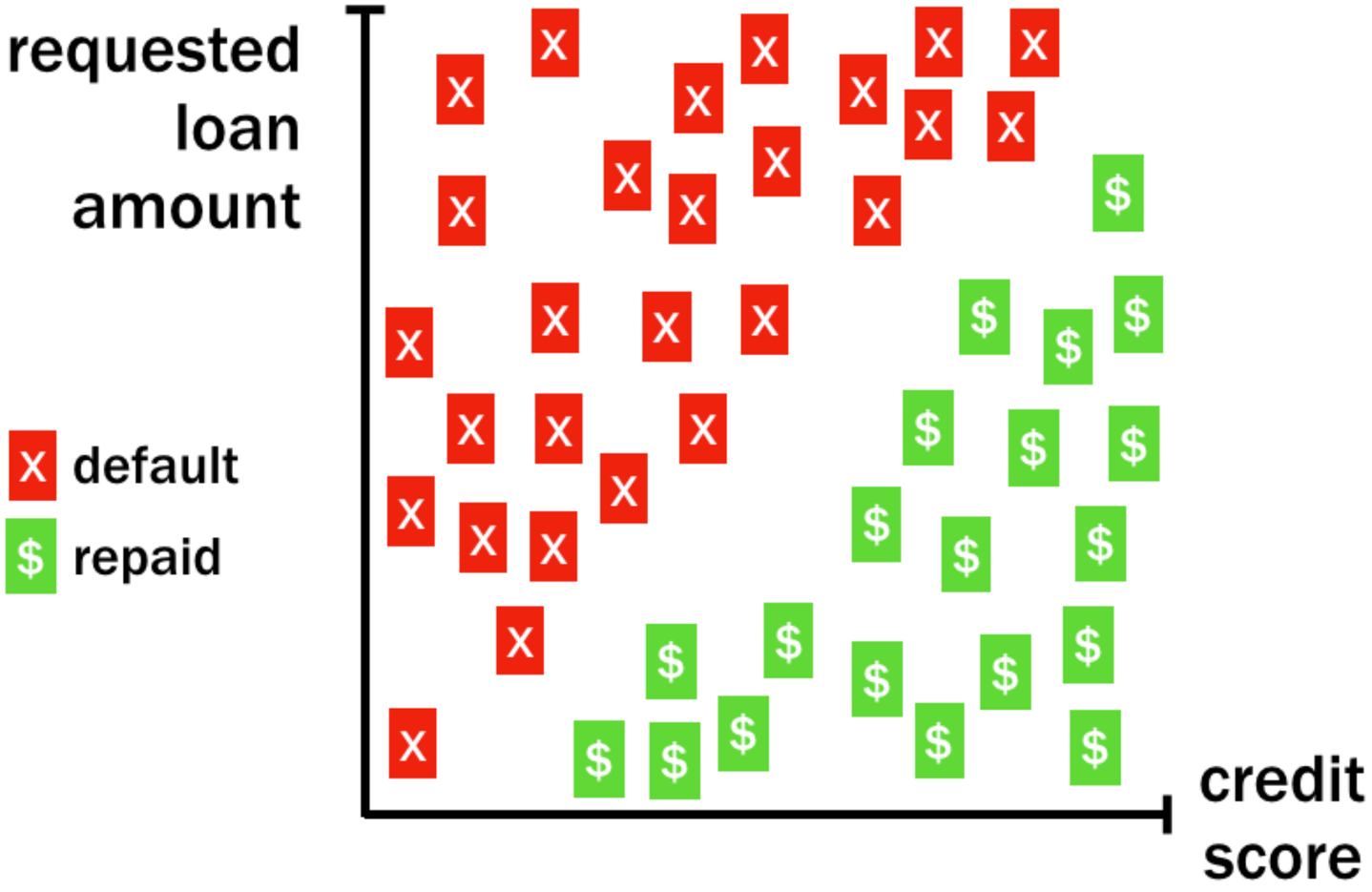
A decision tree model



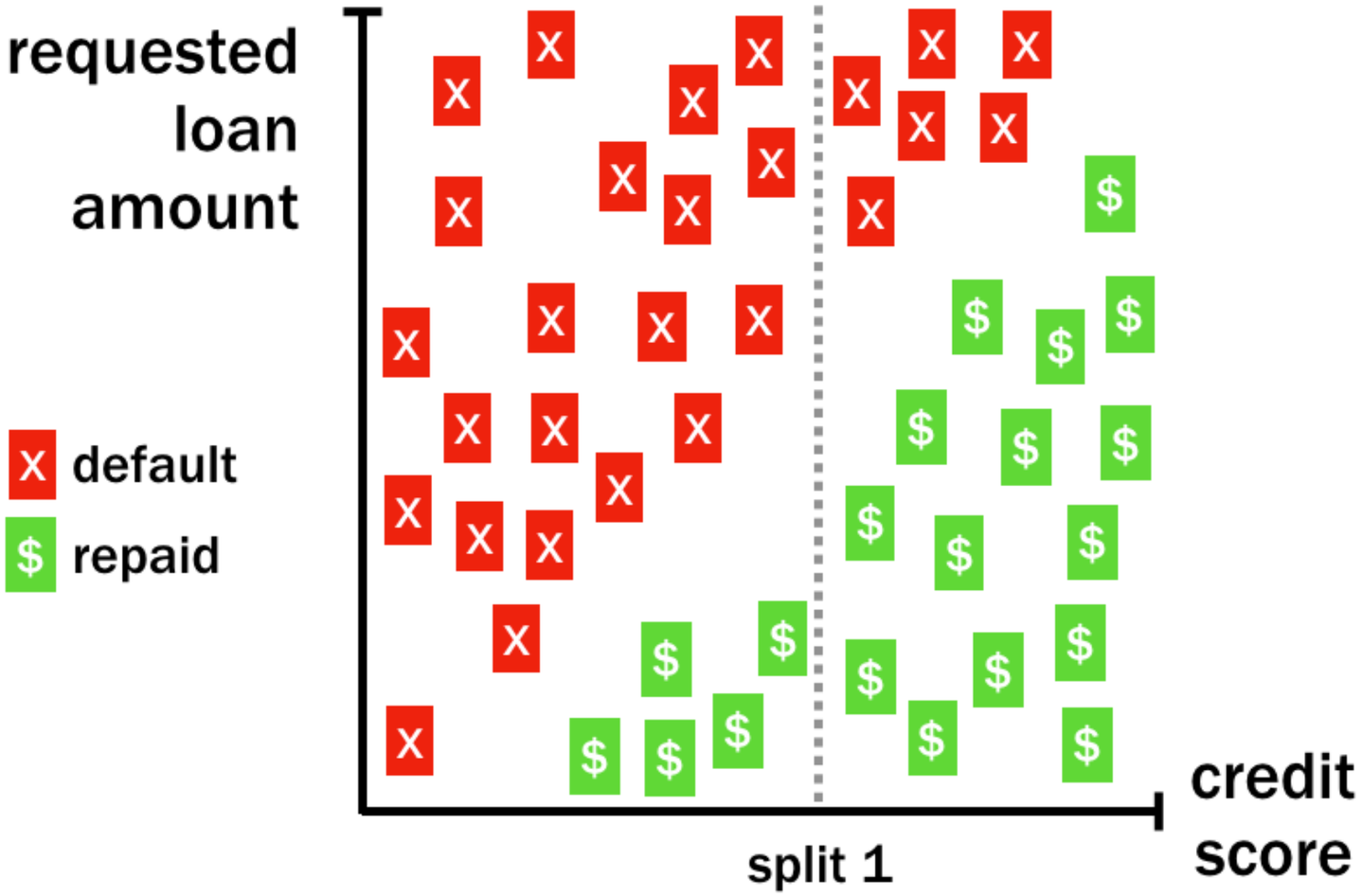
Decision trees for prediction



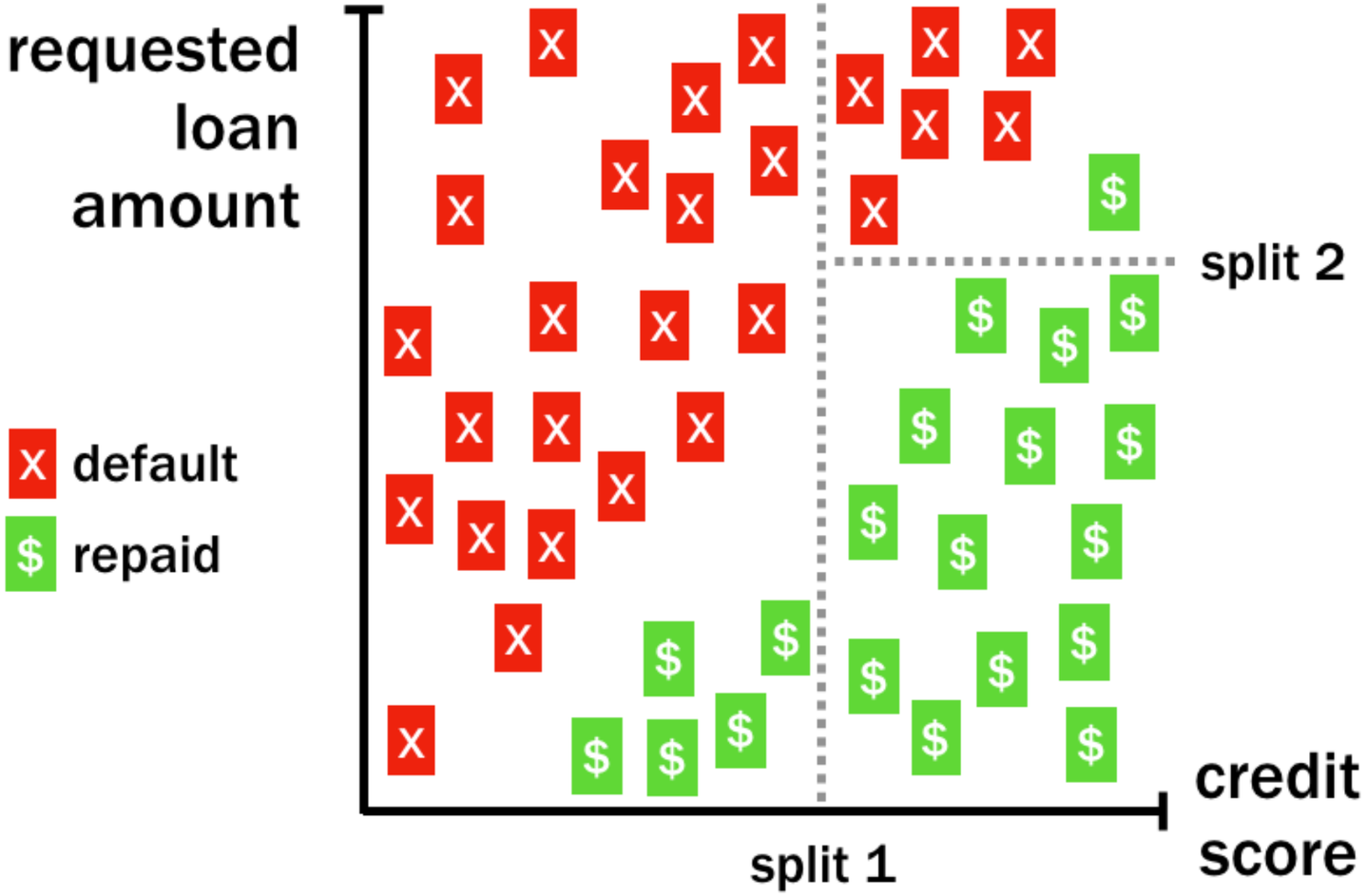
Divide-and-conquer



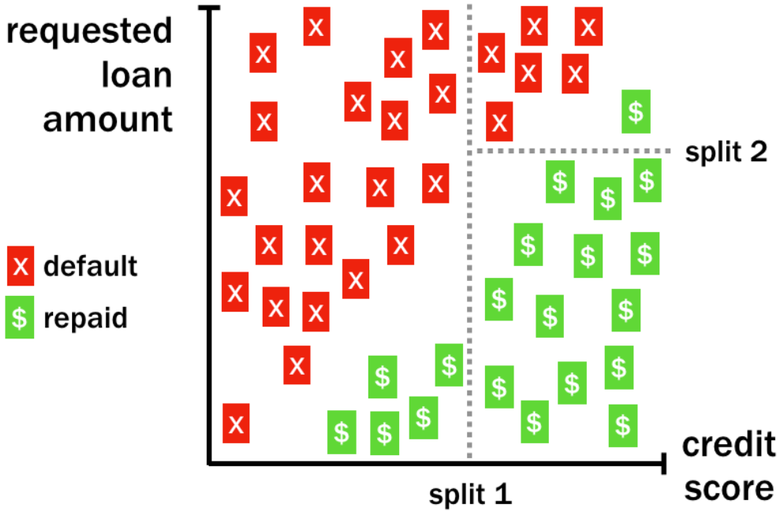
Divide-and-conquer

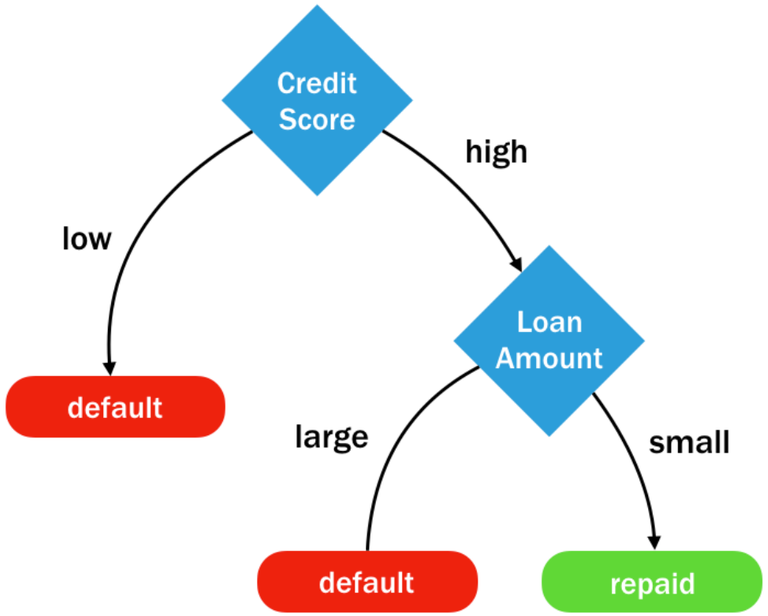


Divide-and-conquer



The resulting tree





Building trees in R

```
# building a simple rpart classification tree
library(rpart)
m <- rpart(outcome ~ loan_amount + credit_score, data = loans,
           method = "class")
```

```
# making predictions from an rpart tree
p <- predict(m, test_data, type = "class")
```

Let's practice!

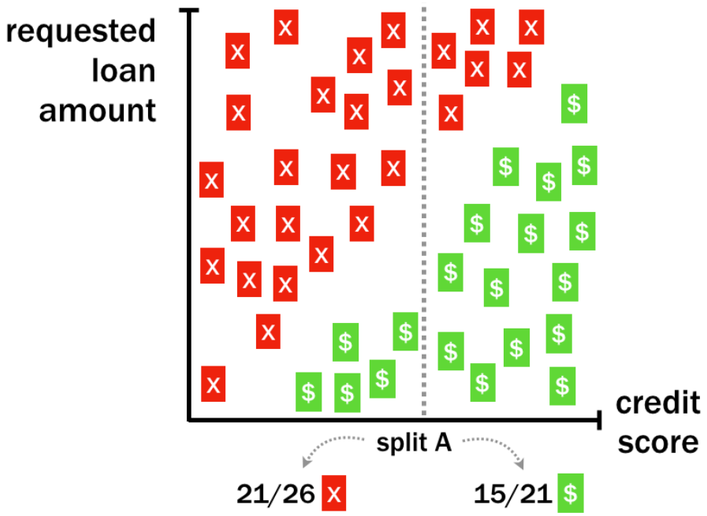
SUPERVISED LEARNING IN R: CLASSIFICATION

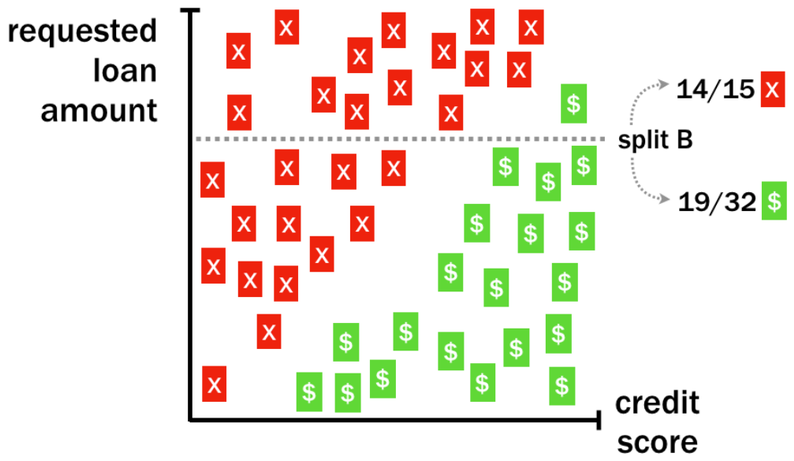
Growing larger classification trees

SUPERVISED LEARNING IN R: CLASSIFICATION

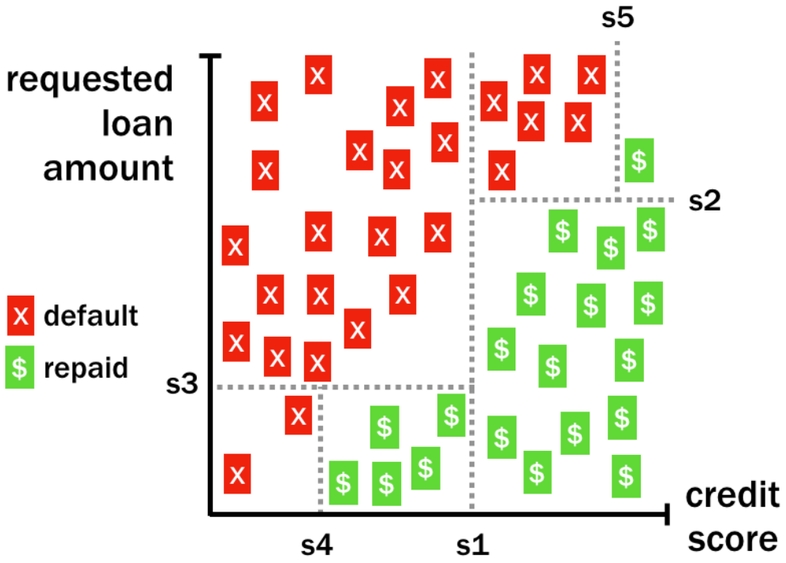
Brett Lantz
Instructor

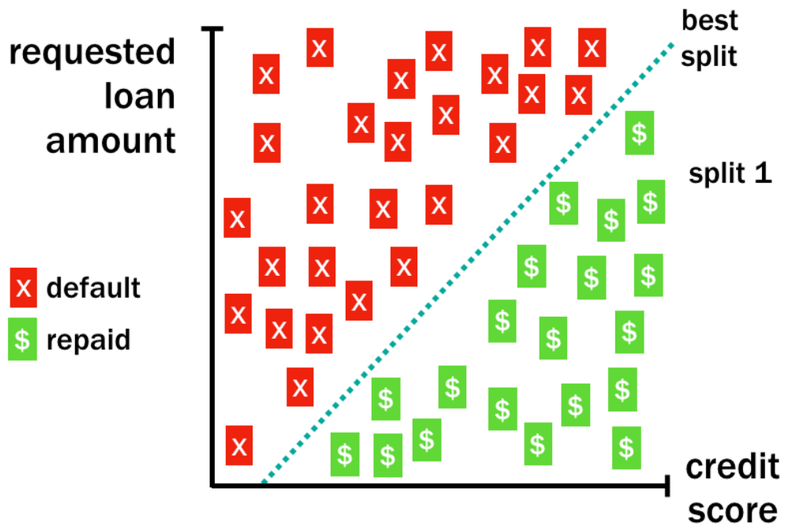
Choosing where to split



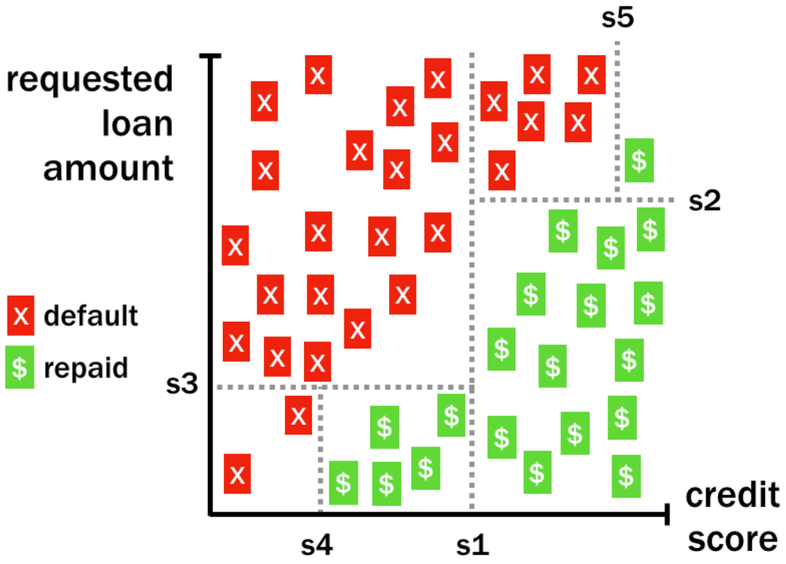


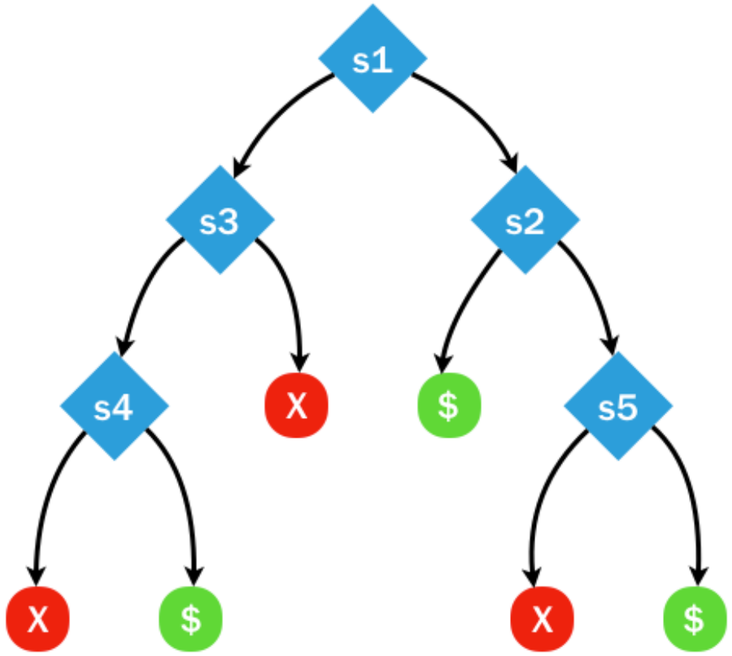
Axis-parallel splits



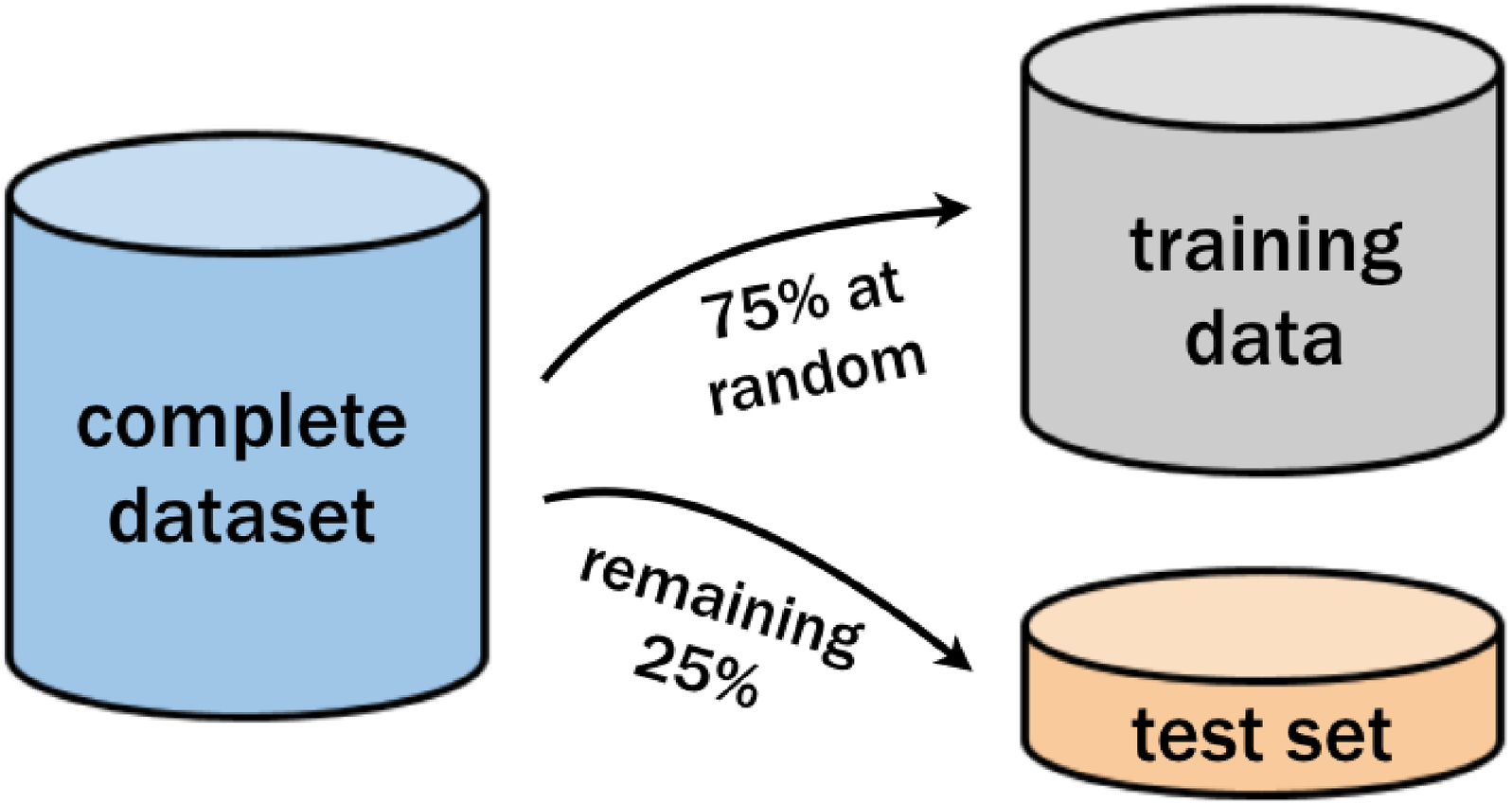


The problem of overfitting





Evaluating model performance



Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION

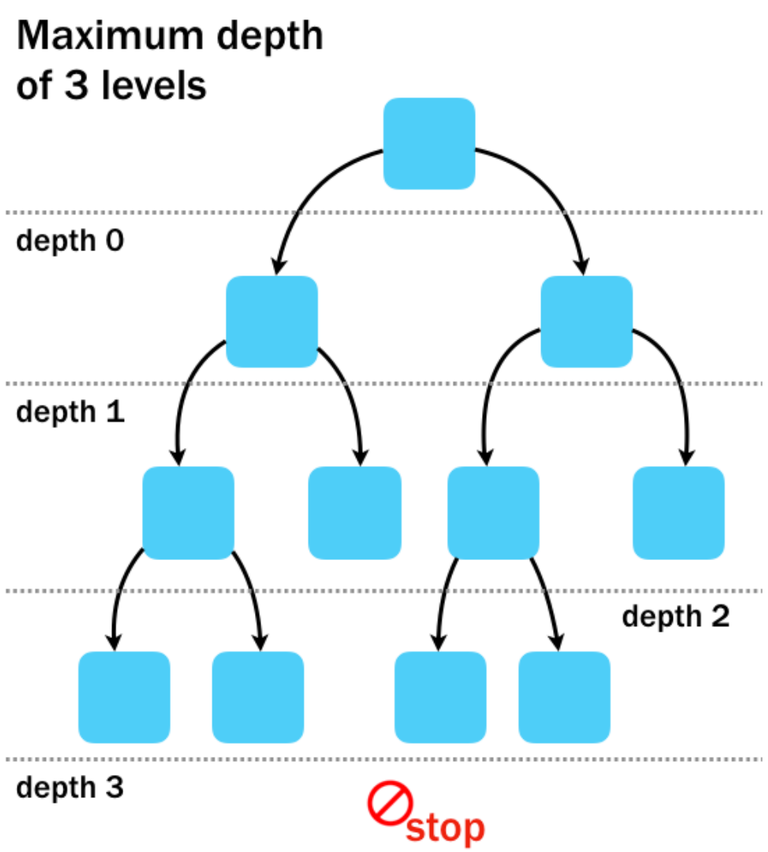
Tending to classification trees

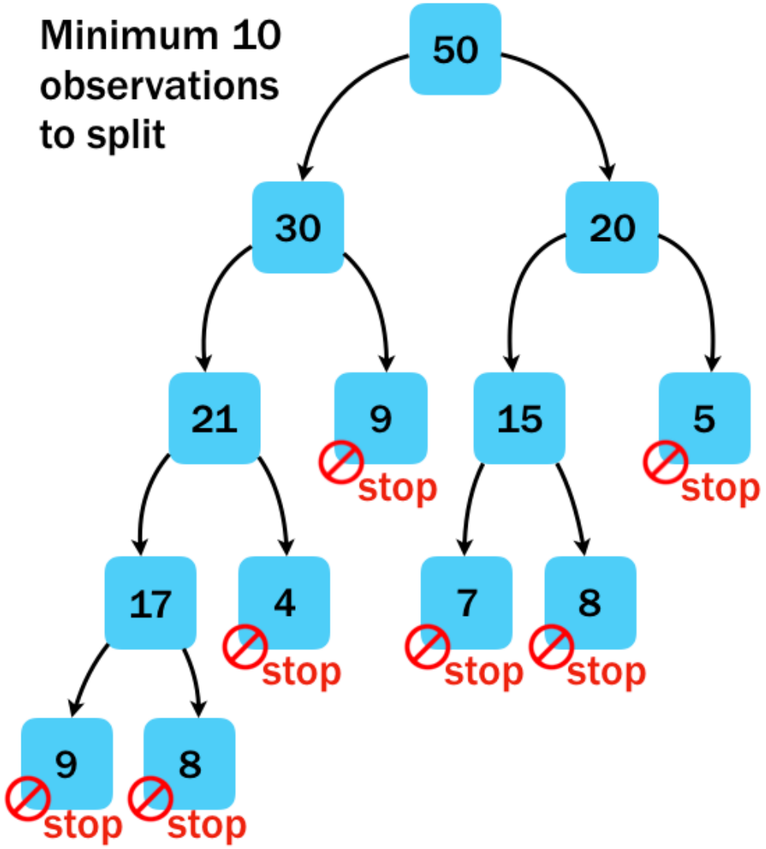
SUPERVISED LEARNING IN R: CLASSIFICATION

Brett Lantz
Instructor

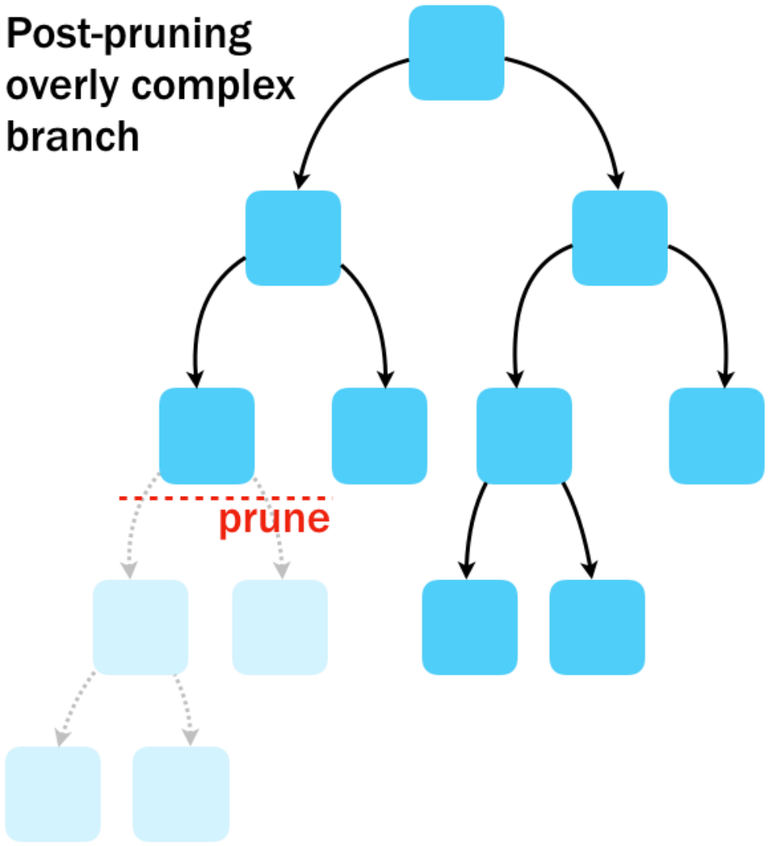
datacamp

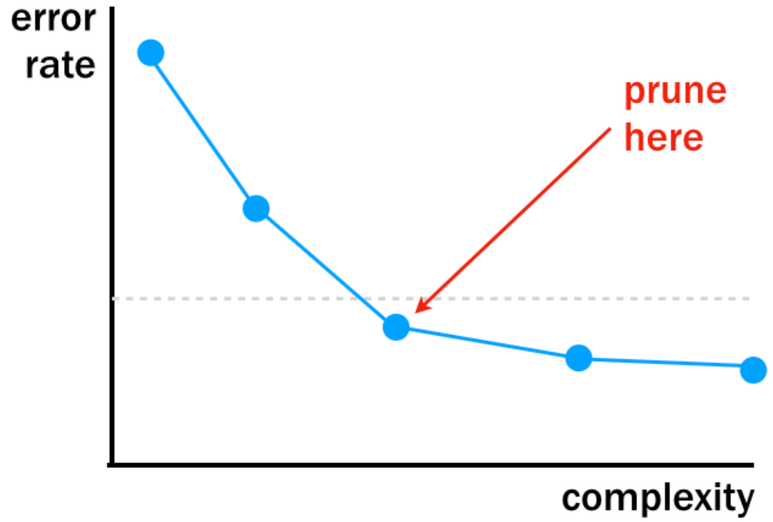
Pre-pruning


Maximum depth
of 3 levels
depth 0
depth 1
depth 2
depth 3
stop


Minimum 10
observations
to split
50
30
20
21
9
15
5
17
4
7
8
9
8
stop

Post-pruning





Pre- and post-pruning with R

```
# pre-pruning with rpart
library(rpart)
prune_control <- rpart.control(maxdepth = 30, minsplit = 20)

m <- rpart(repaid ~ credit_score + request_amt,
           data = loans,
           method = "class",
           control = prune_control)
```

```
# post-pruning with rpart
m <- rpart(repaid ~ credit_score + request_amt,
           data = loans,
           method = "class")

plotcp(m)

m_pruned <- prune(m, cp = 0.20)
```

Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION

Seeing the forest from the trees

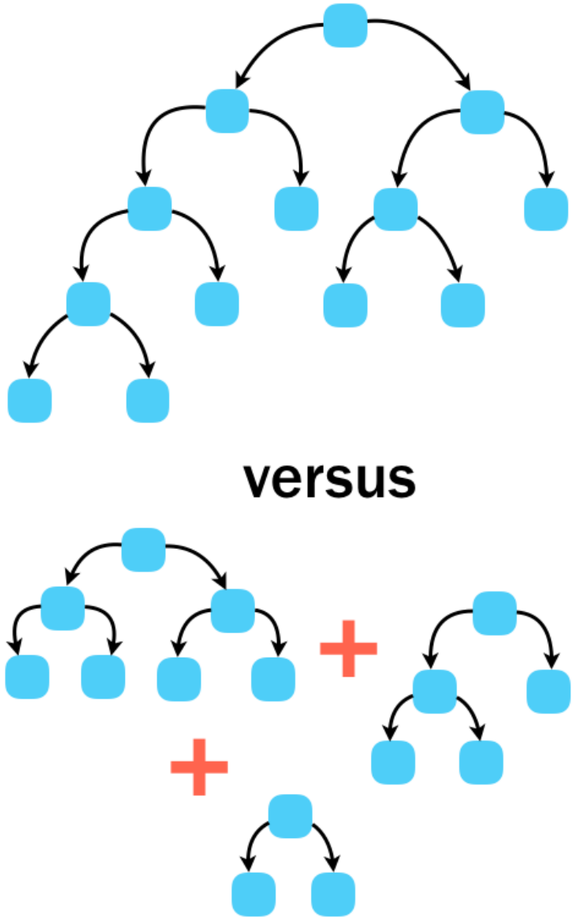
SUPERVISED LEARNING IN R: CLASSIFICATION

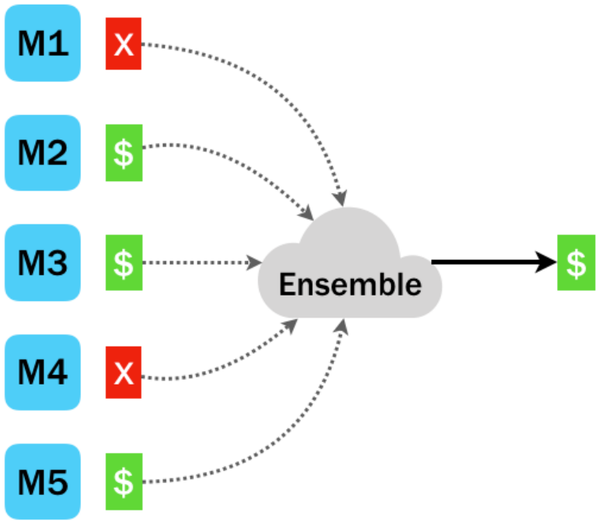
Brett Lantz
Instructor

datacamp

Understanding random forests

Making decisions as an ensemble





Random forests in R

```
# building a simple random forest
library(randomForest)
m <- randomForest(repaid ~ credit_score + request_amt, data = loans,
             ntree = 500,    # number of trees in the forest
             mtry = sqrt(p)) # number of predictors (p) per tree
```

```
# making predictions from a random forest
p <- predict(m, test_data)
```

Let's practice!

SUPERVISED LEARNING IN R: CLASSIFICATION