Tree-Based Methods

AU STAT627

Emil Hvitfeldt

2021-11-15

Overview

We will cover 4 new methods today

- Decision Trees
- Bagging
- Random Forrest
- boosting

Overview

We will cover 4 new methods today

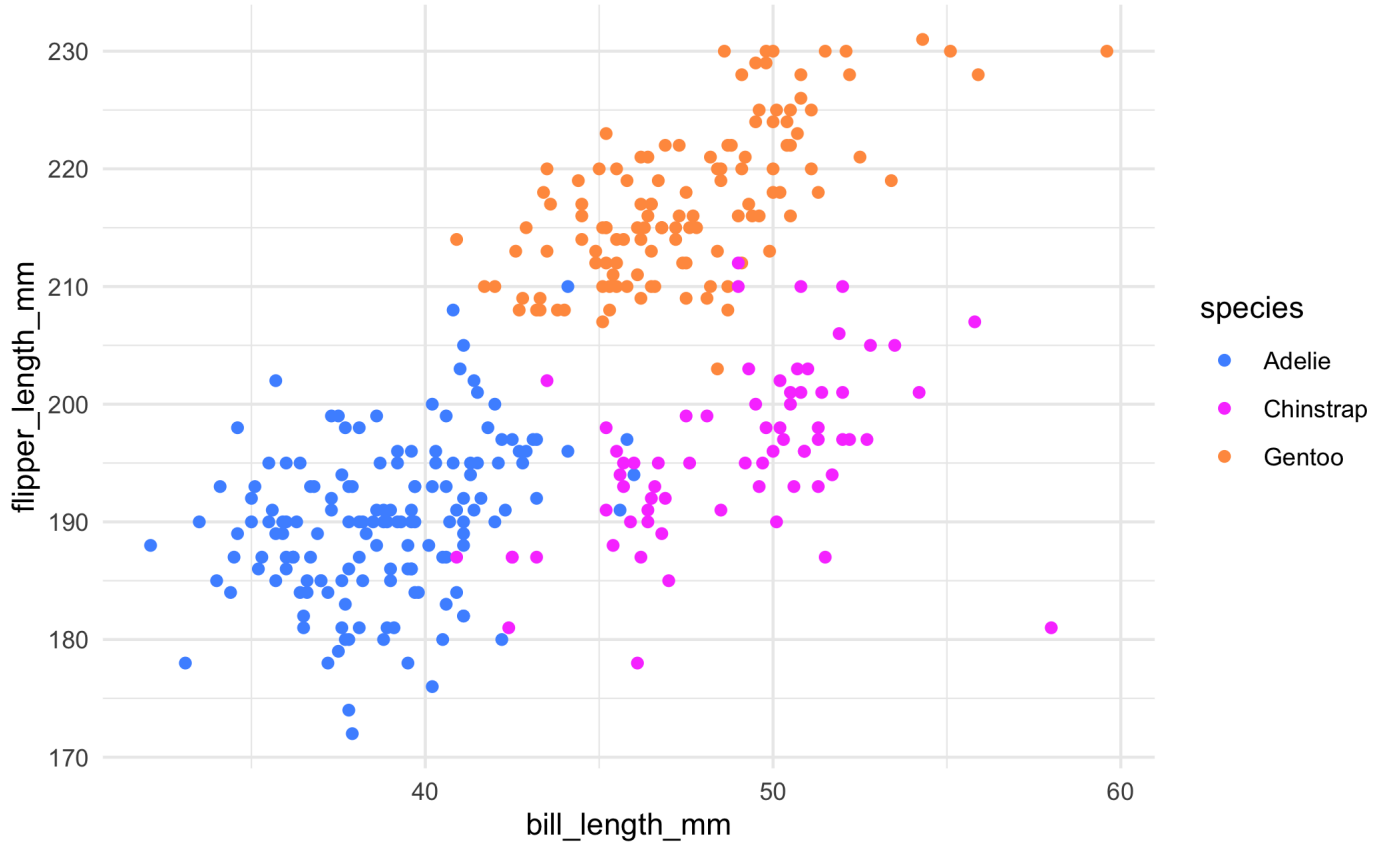
- Decision Trees
- Bagging
- Random Forrest
- boosting

Decision trees act as the building block for this chapter

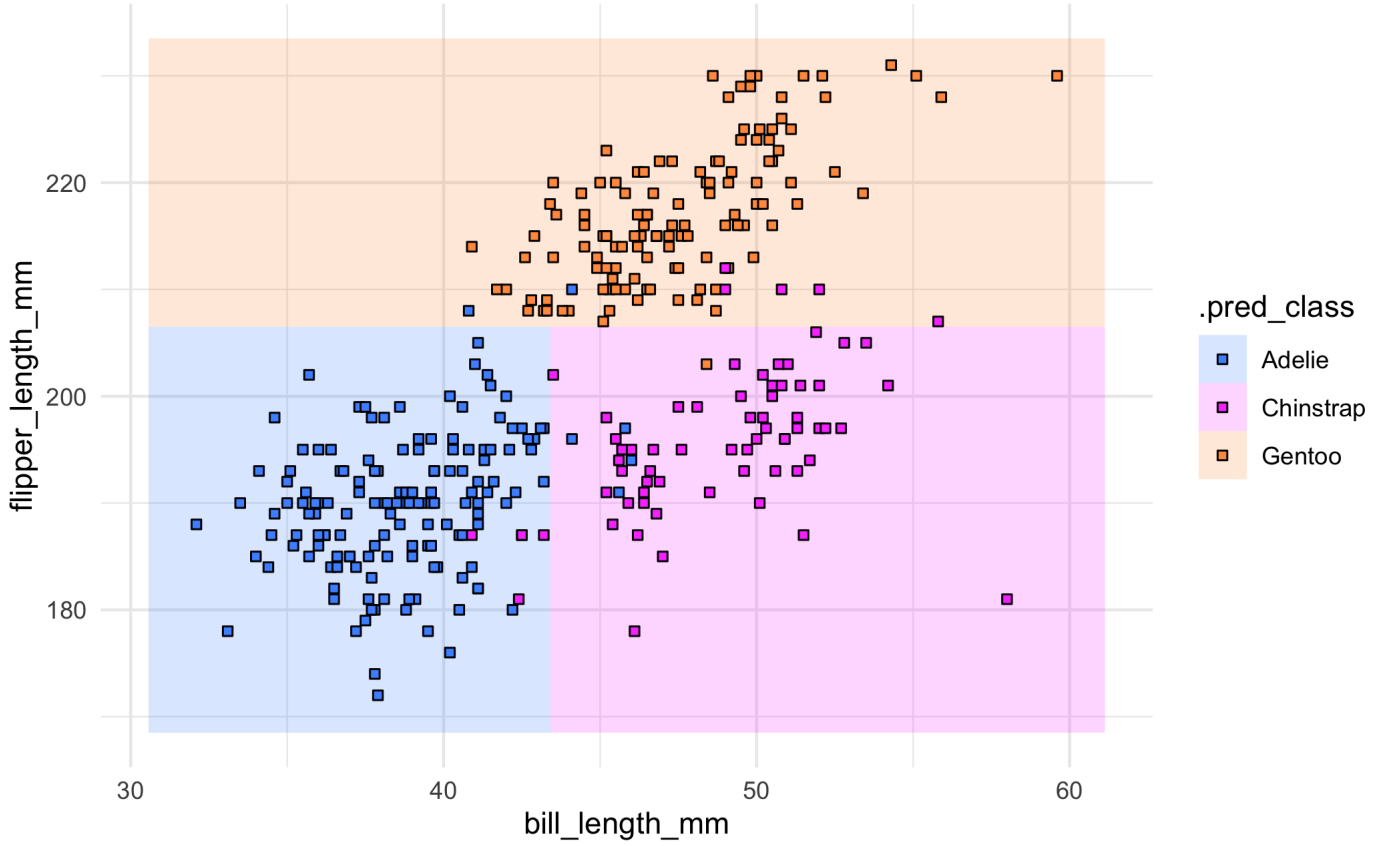
Decision Trees

Given a problem, give me a flow chart of if-else statements to find the answer

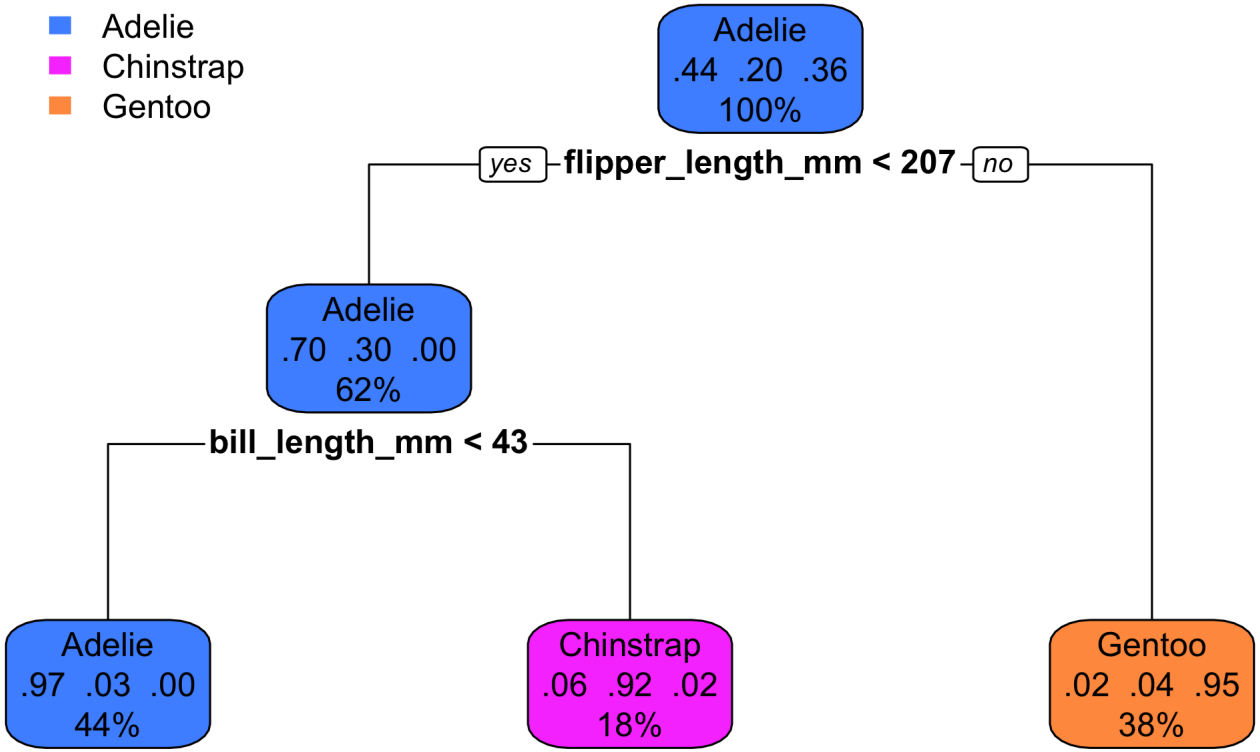
Penguins


230
220
210
200
190
180
170
flipper_length_mm
40
50
60
bill_length_mm
species
Adelie
Chinstrap
Gentoo

Penguins


flipper_length_mm
220
200
180
30
40
50
60
bill_length_mm
.pred_class
Adelie
Chinstrap
Gentoo

The flowchart



The rules

```
##         ..y  Ade Chi Gen
##      Adelie [.97 .03 .00] when flipper_length_mm <  207 & bill_length_mm <  43
##   Chinstrap [.06 .92 .02] when flipper_length_mm <  207 & bill_length_mm >= 43
##      Gentoo [.02 .04 .95] when flipper_length_mm >= 207
```

General setup

- We divide the predictor space into multiple non-overlapping regions ($R_1, R_2, \ldots, R_J$).

- Every observation that falls into a region will have the same prediction, and that prediction will be based on the observations in that region

    - Regression: mean value

    - Classification: Most common value

General setup

The shapes could in theory be any shape, but for simplicity, we are using rectangles/boxes to partition the space

The main goal is to build a partition that minimizes some loss such as RSS

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

General setup

It is generally computational unfeasible to calculate all possible partitions

We use a **recursive binary splitting** procedure to find the trees

This approach is **top-down** approach since we start at the top and split our way down

It is **greedy** because we select the best possible split each time

Details

How many times should we split?

If we continue to split we end up with each observation belonging to their region, giving us a wildly flexible model

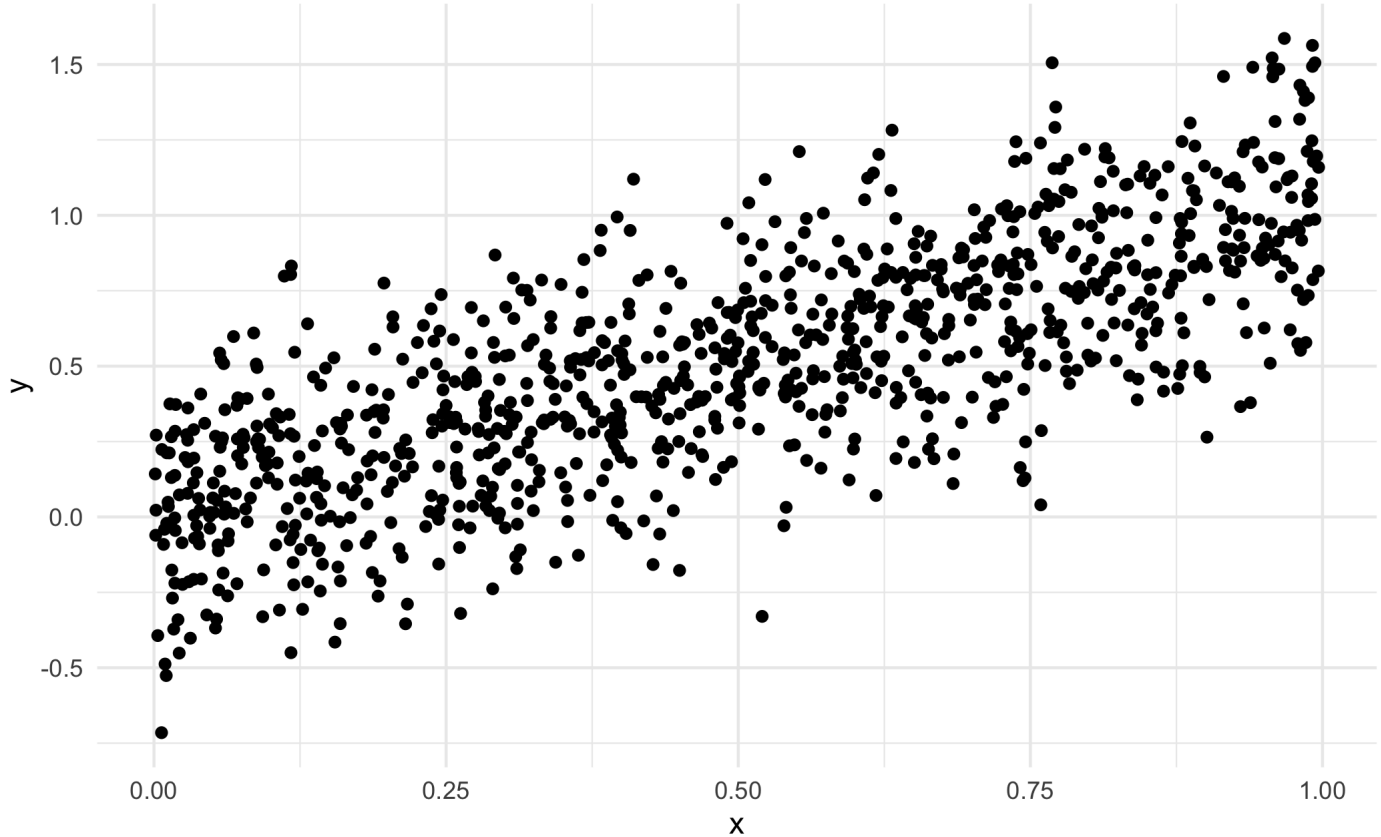
We can control several different things, simple ones are

- Tree depth, maximum depth of the tree
- minimum number of data points in a node that is required for the node to be split further

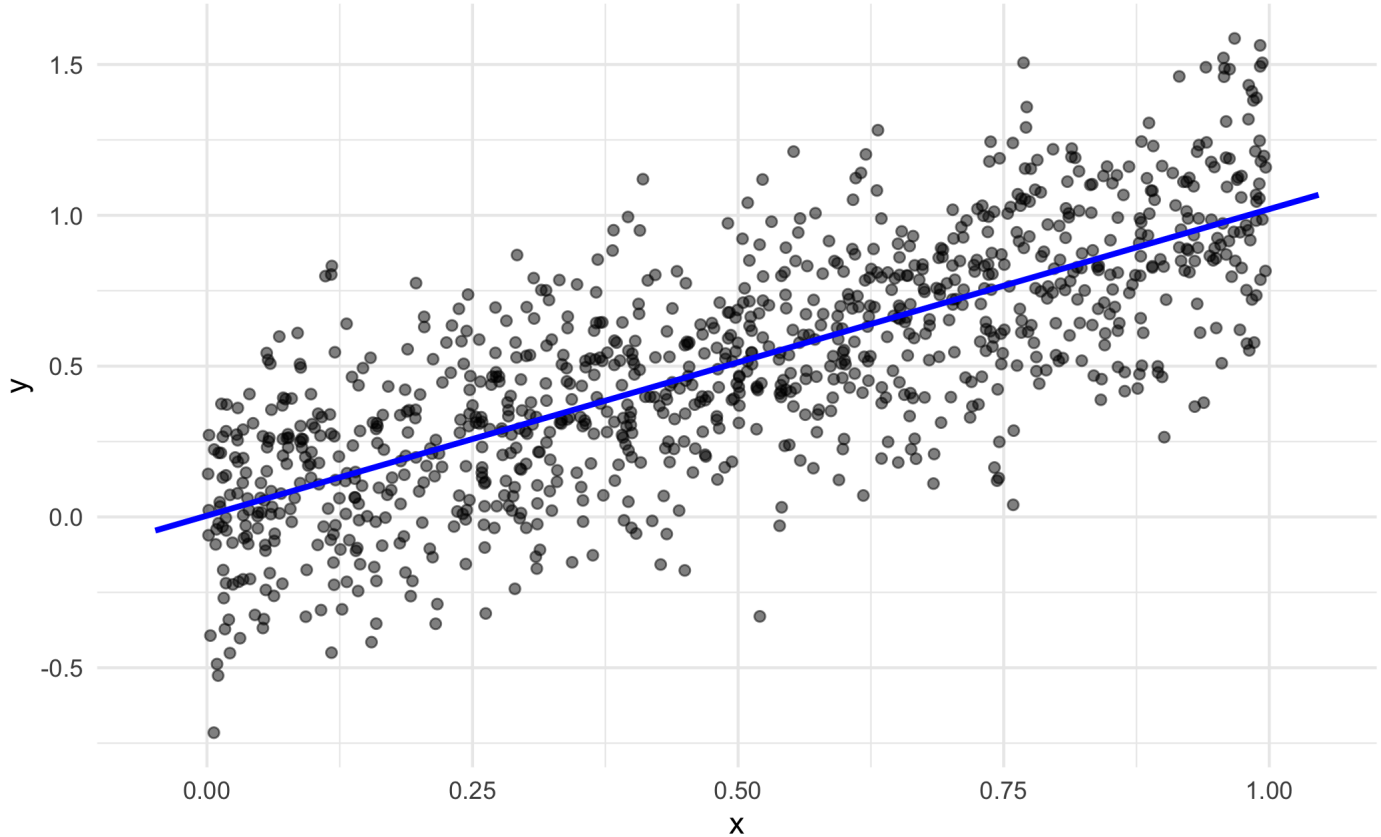
Tree Pruning

Due to the way decision trees are grown, it can be beneficial to grow larger trees and then go back and reduce the complexity of the tree after

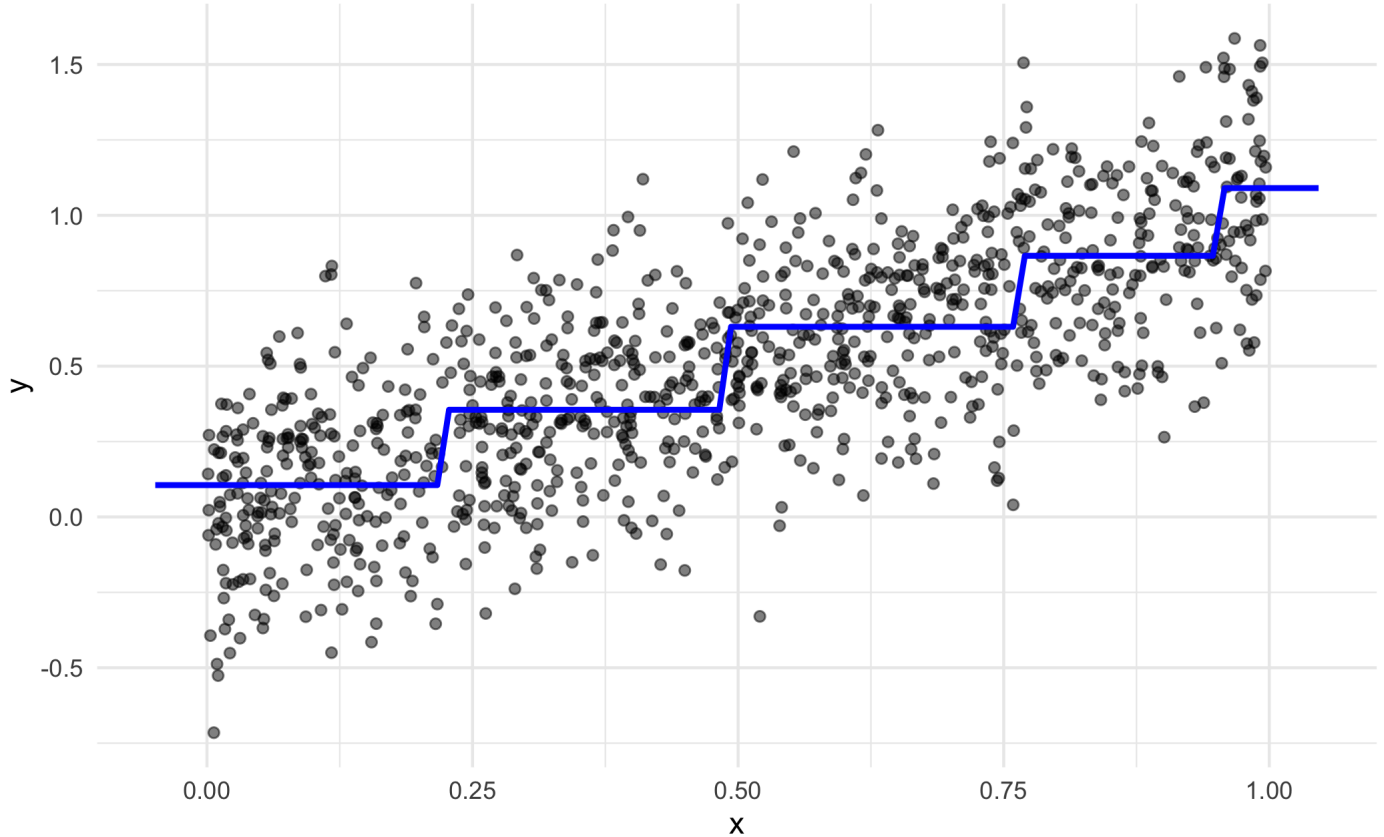
Regression "curves"



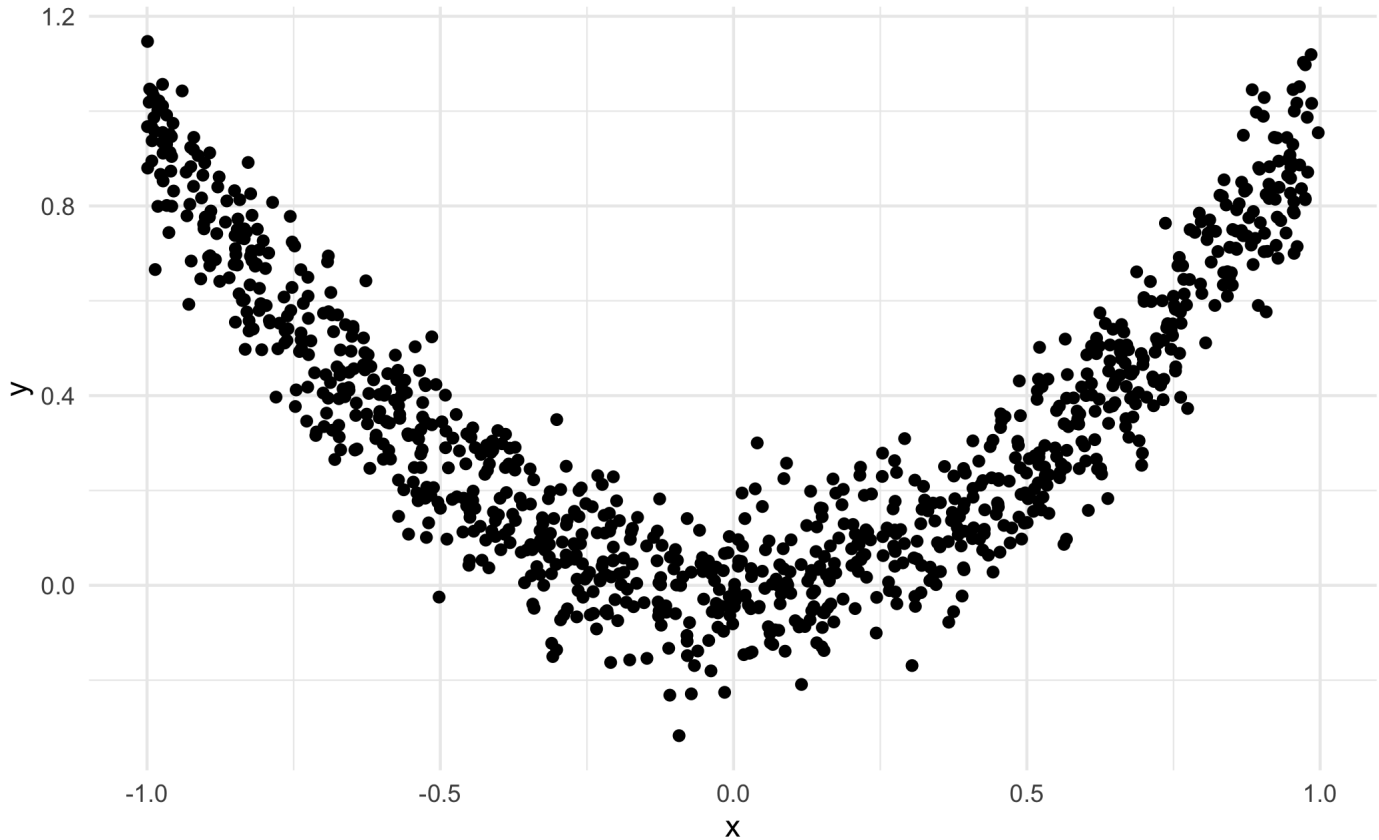
Regression "curves"



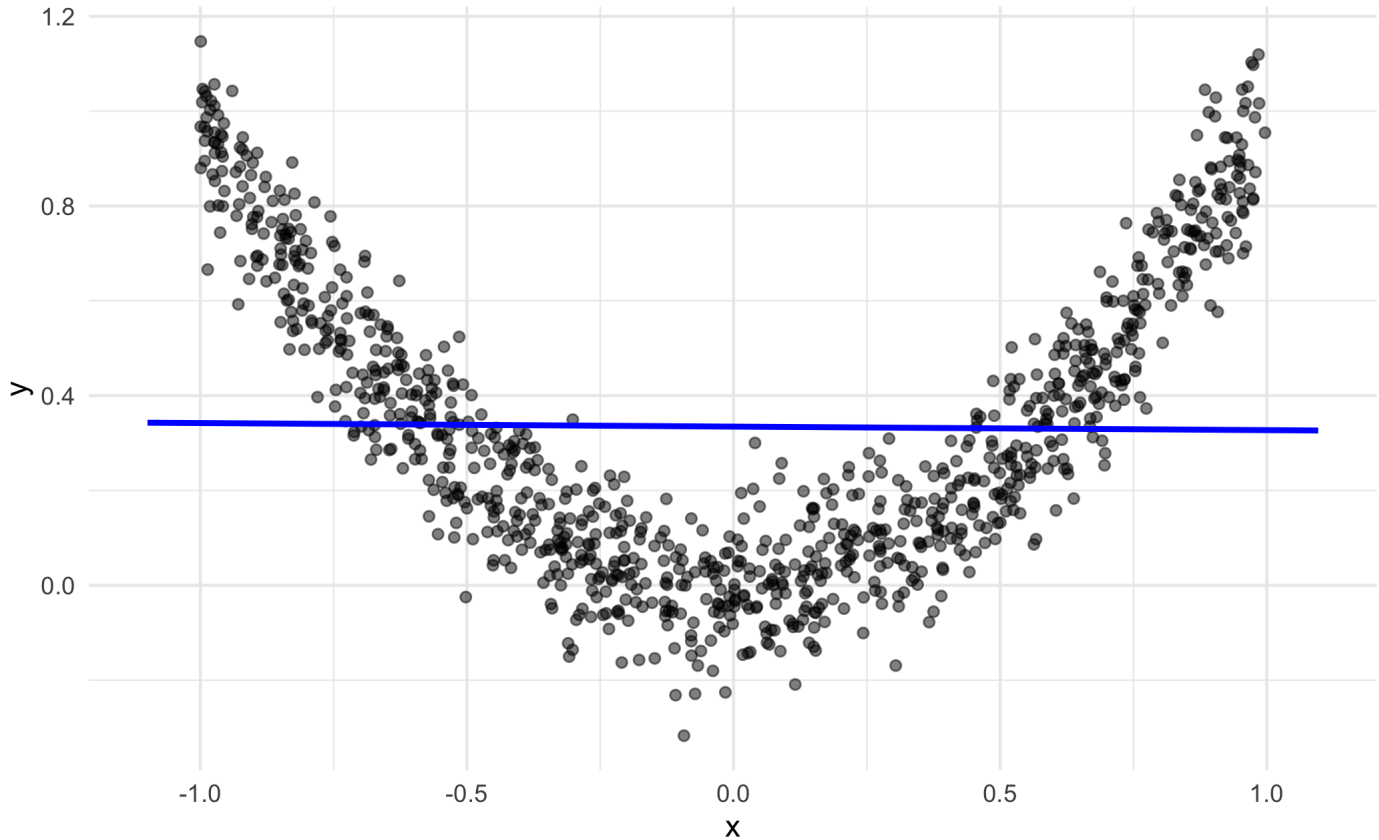
Regression "curves"



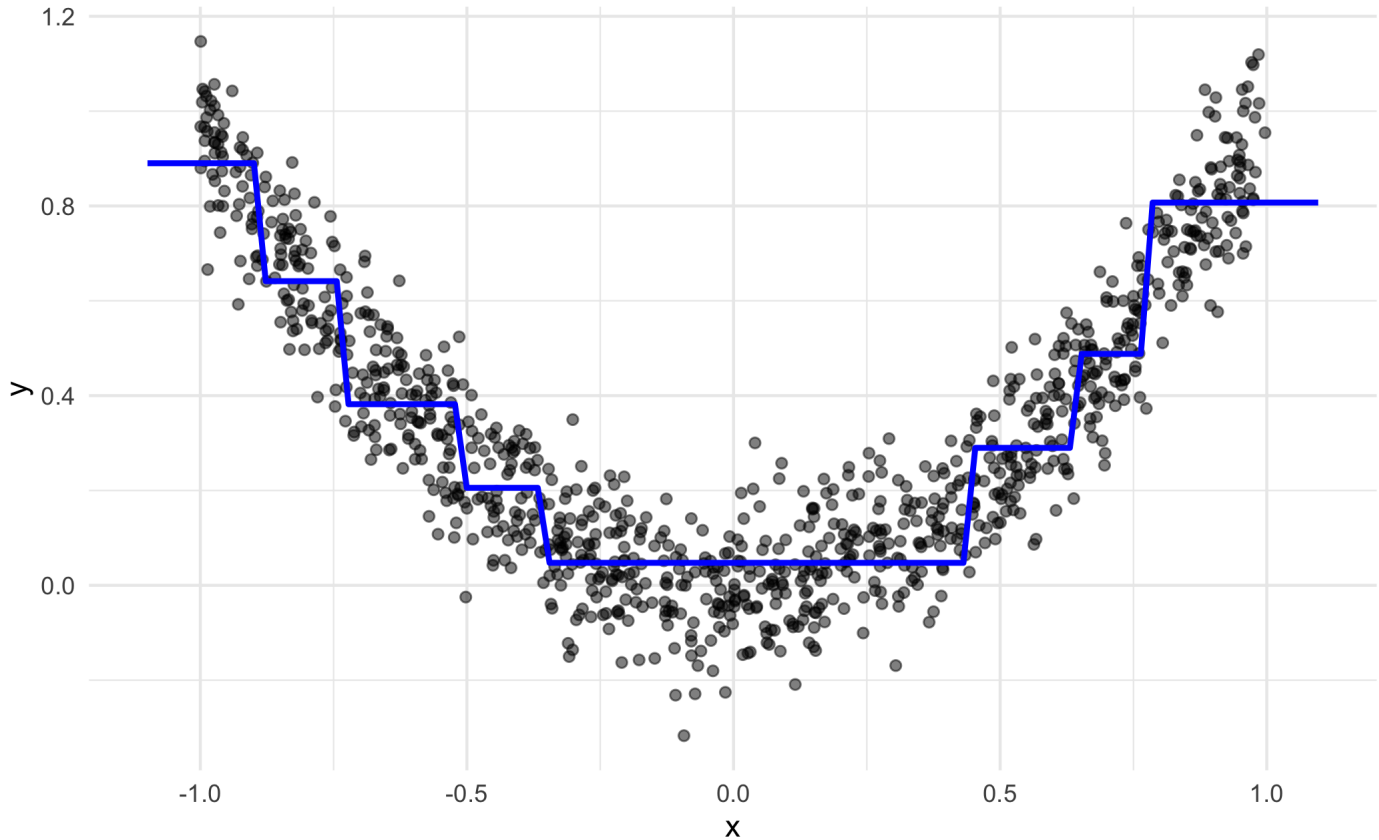
Regression "curves"



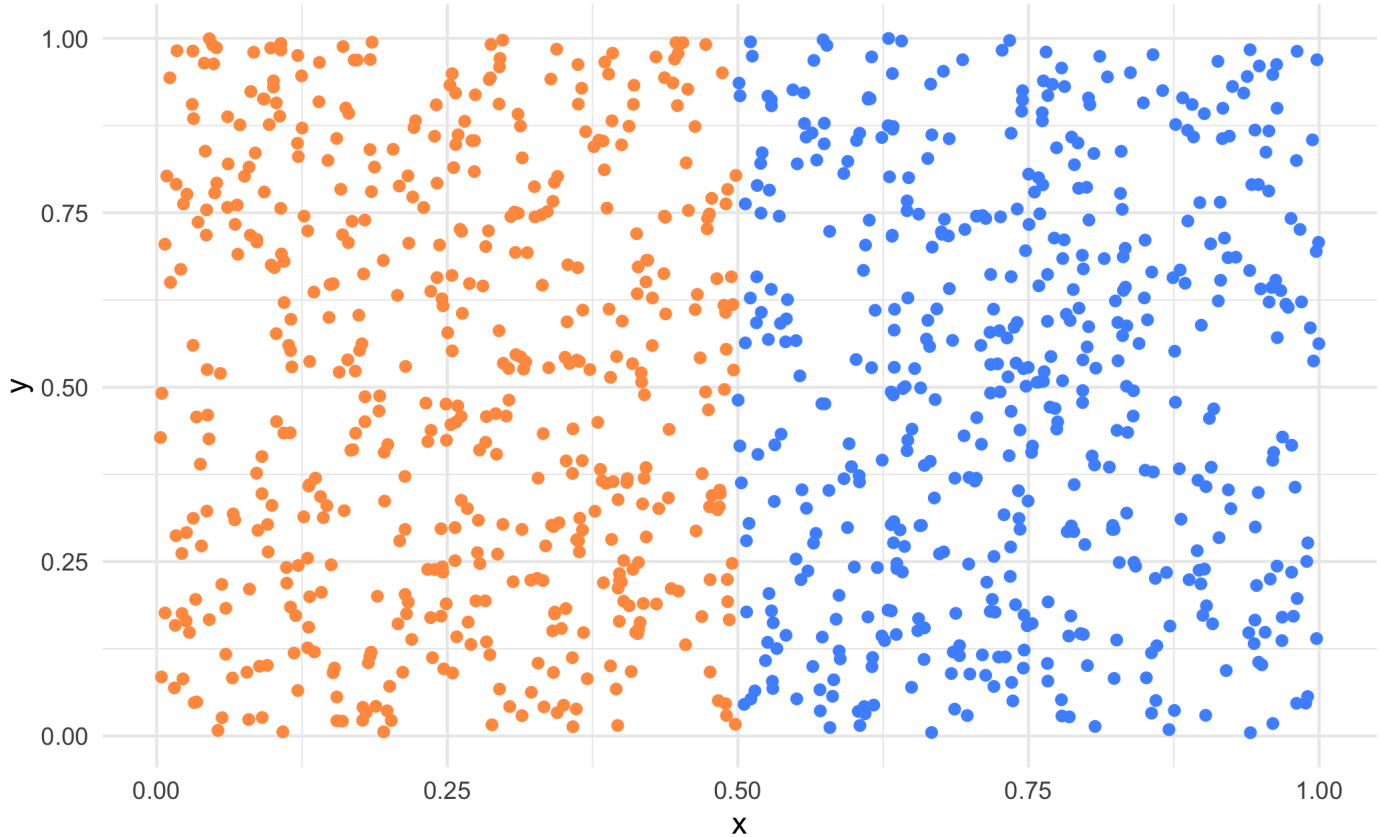
Regression "curves"



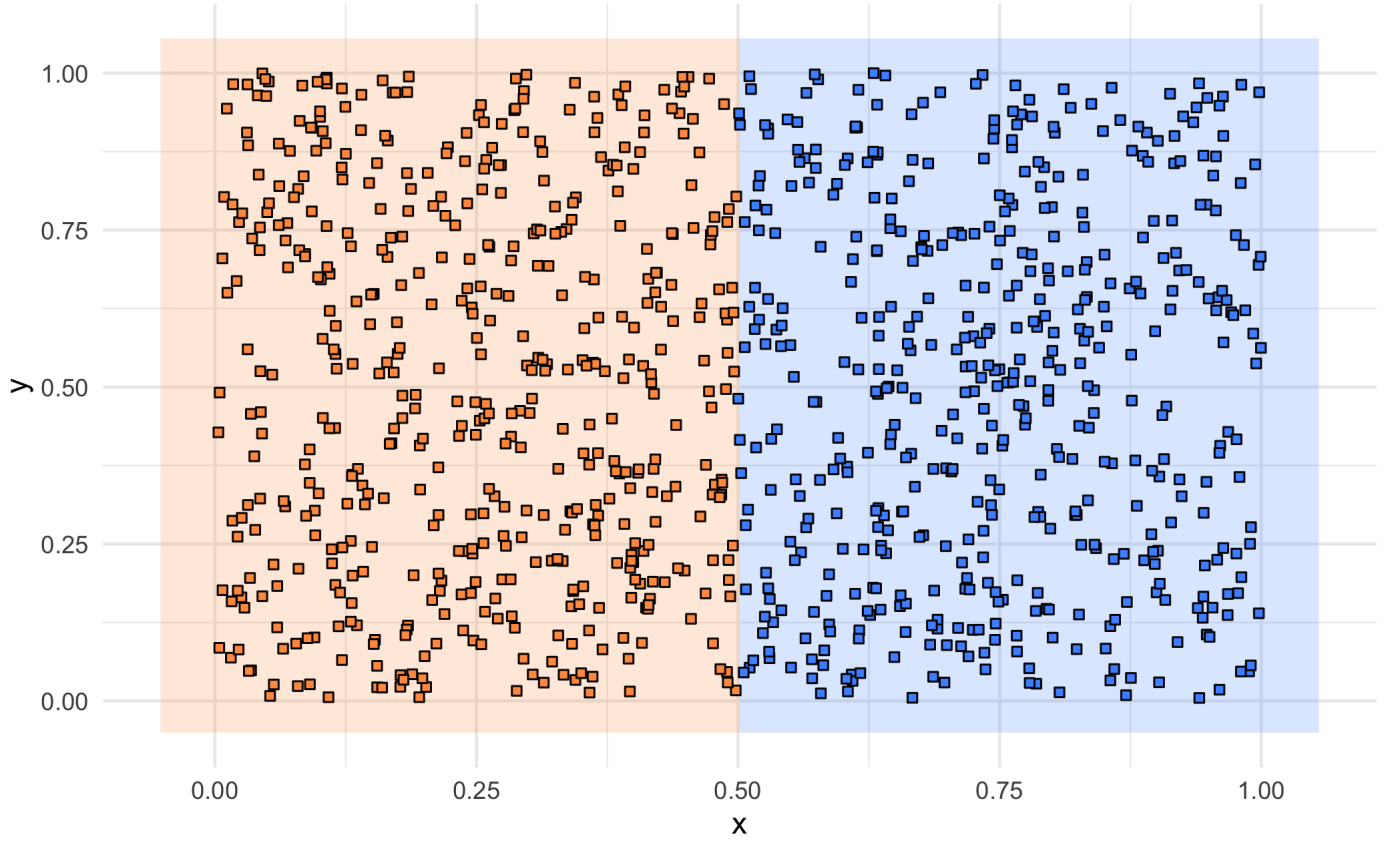
Regression "curves"



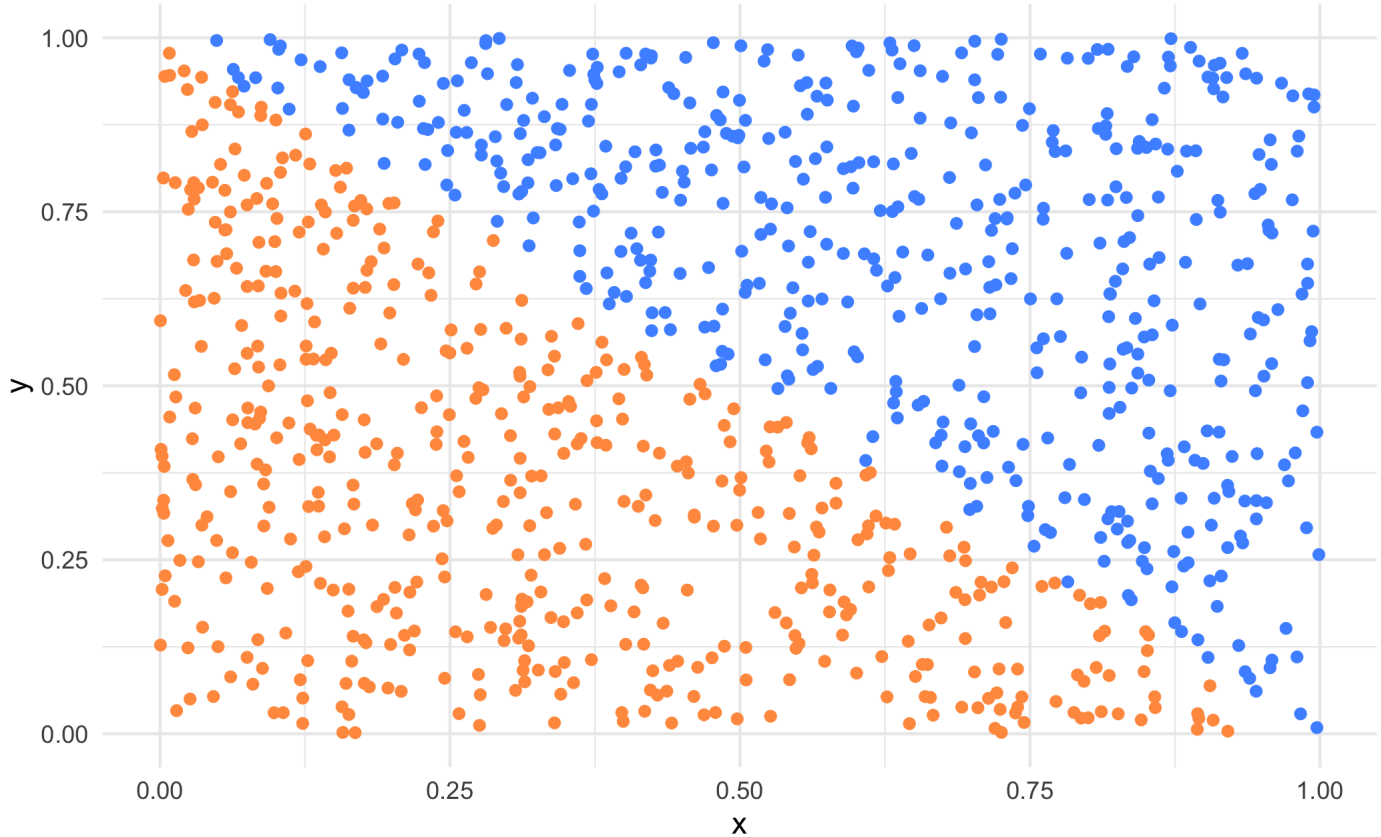
Decision boundary



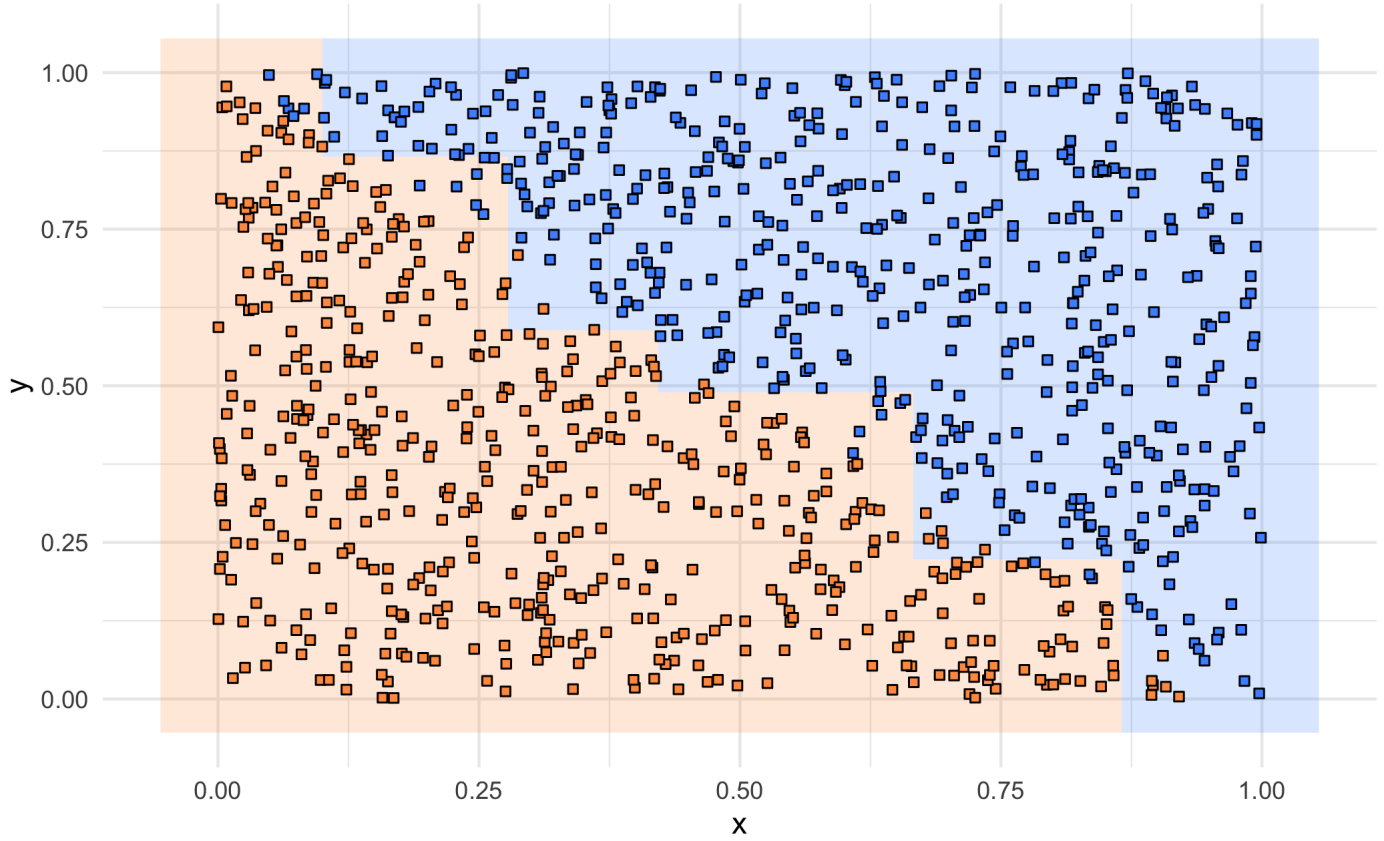
Decision boundary



Decision boundary



Decision boundary



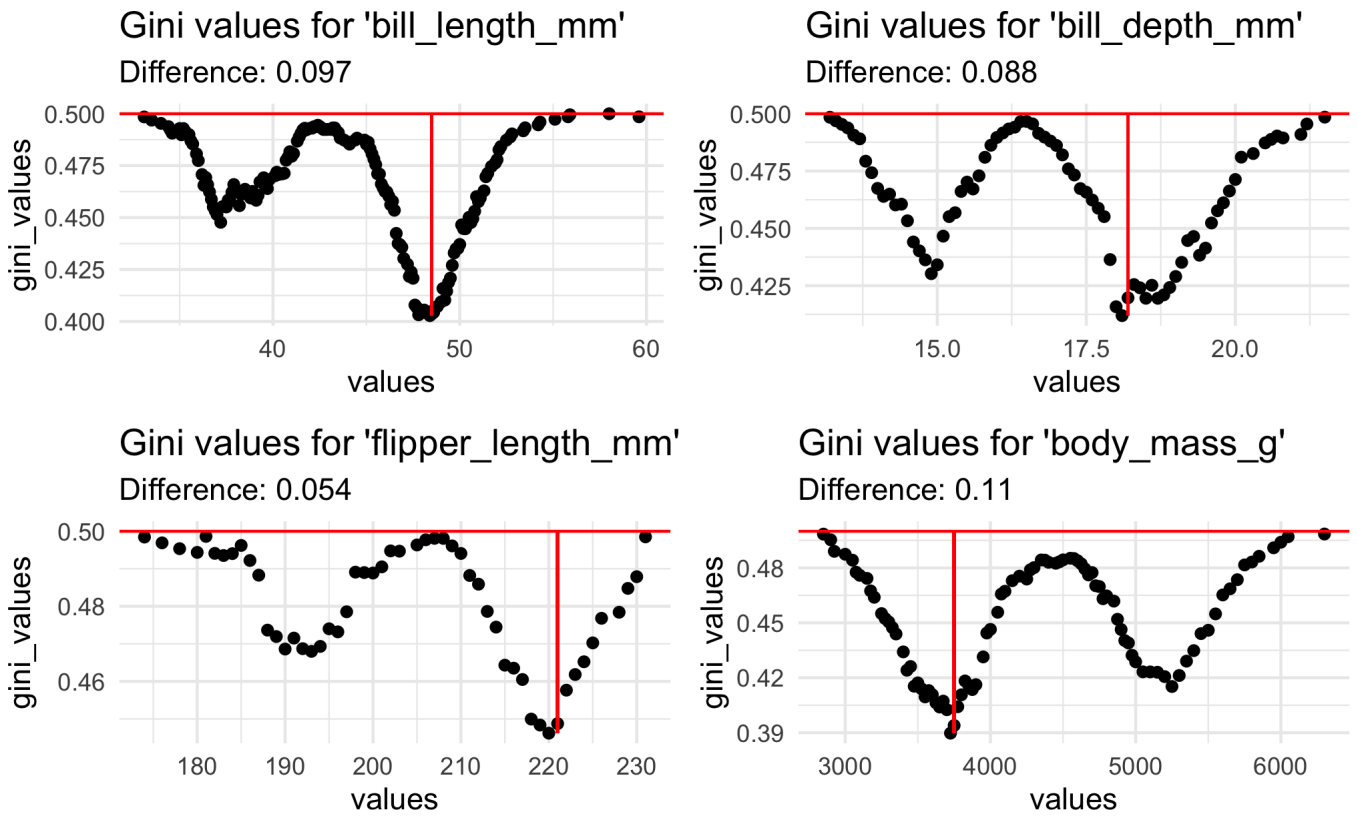
Pros and Cons

Pros

- Very easy to explain and reason about
- Can Handle qualitative predictors without the need for dummy variables

Cons

- Don't have great predictive power
- Non-robust, small changes in the data can give wildly different models


Gini values for 'bill_length_mm'
Difference: 0.097
gini_values
0.500
0.475
0.450
0.425
0.400
40
50
60
values
Gini values for 'bill_depth_mm'
Difference: 0.088
gini_values
0.500
0.475
0.450
0.425
15.0
17.5
20.0
values
Gini values for 'flipper_length_mm'
Difference: 0.054
gini_values
0.50
0.48
0.46
180
190
200
210
220
230
values
Gini values for 'body_mass_g'
Difference: 0.11
gini_values
0.48
0.45
0.42
0.39
3000
4000
5000
6000
values

Next Steps

Individual decision trees don't offer great predictive performance due to their simple nature

Bagging, Random Forests and Boosting uses multiple decision trees together to get better performance with a trade-off of more complexity

Bagging

Decision trees suffer for high variance

We saw in week 3 how bootstrapping could be used to reduce the variance of a statistical learning method

We will use bootstrapping again with decision trees to reduce the variance. We can feasible do this since individual decision trees are fast to train

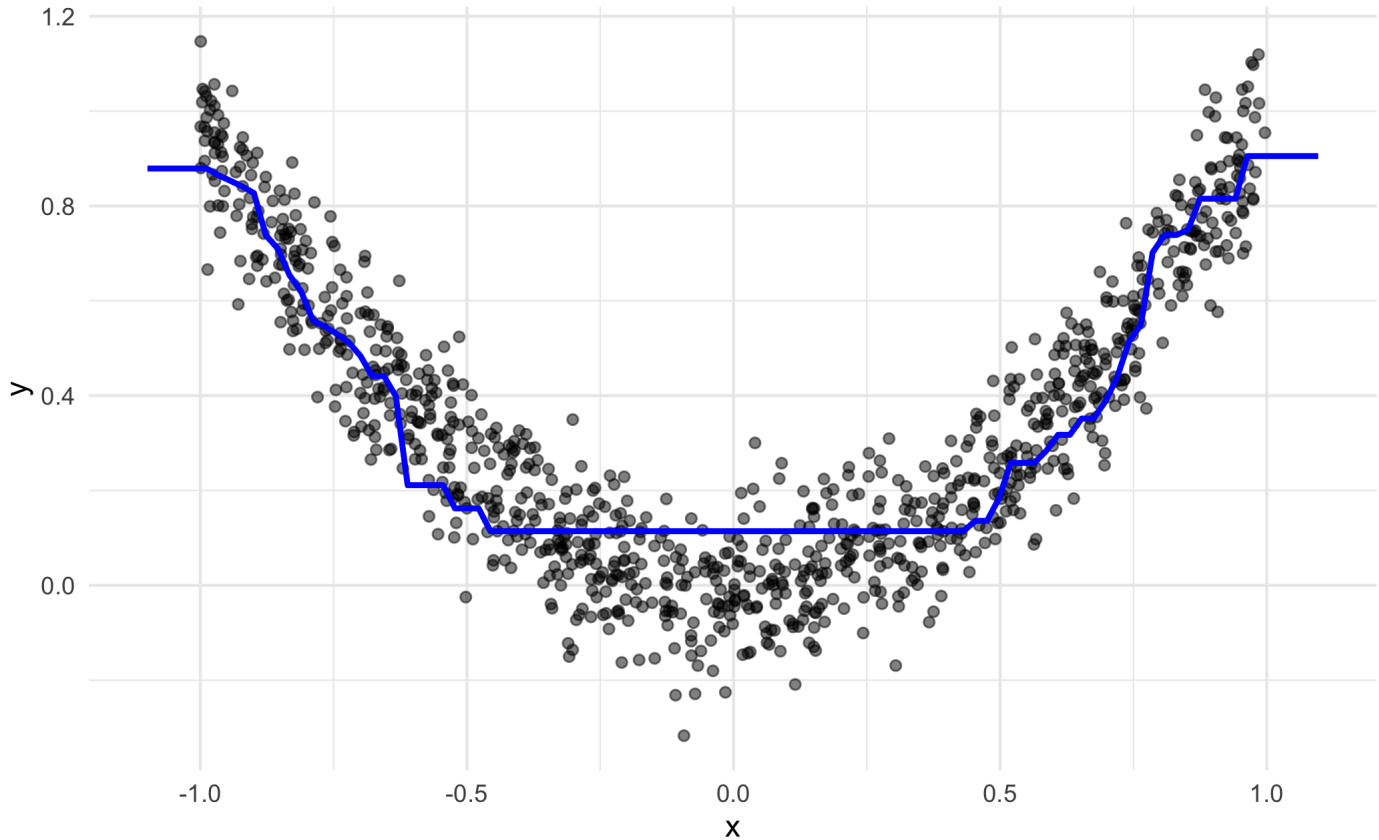
Bagging

"Algorithm"

- Generate B different bootstrapped training data sets
- Fit a decision tree on each of the bootstraps to get $\hat{f}^{*b}(x)$
- Take the average of all the estimates t get your final estimate

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

Bagging



Bagging Notes

The number of bootstraps is not very important here, you just need to use a value of B that is large enough to have the error settled down, ~100 seems to work well

You do not overfit by increasing B, just increase the run-time

Bagged trees offer quite low interpretability since it is a mixture of multiple models

We can obtain a summary of the variable importance of our model by looking at the average amount of RSS a given predictor has decreased due to splits to a given variables

Random Forest

The Random Forest method offers an improvement over Bagged trees

One of the main downsides to Bagged Trees is that the trees become quite correlated with each other

When fitting a Random forest, we start the same way as a Bagged tree with multiple bootstrapped data sets

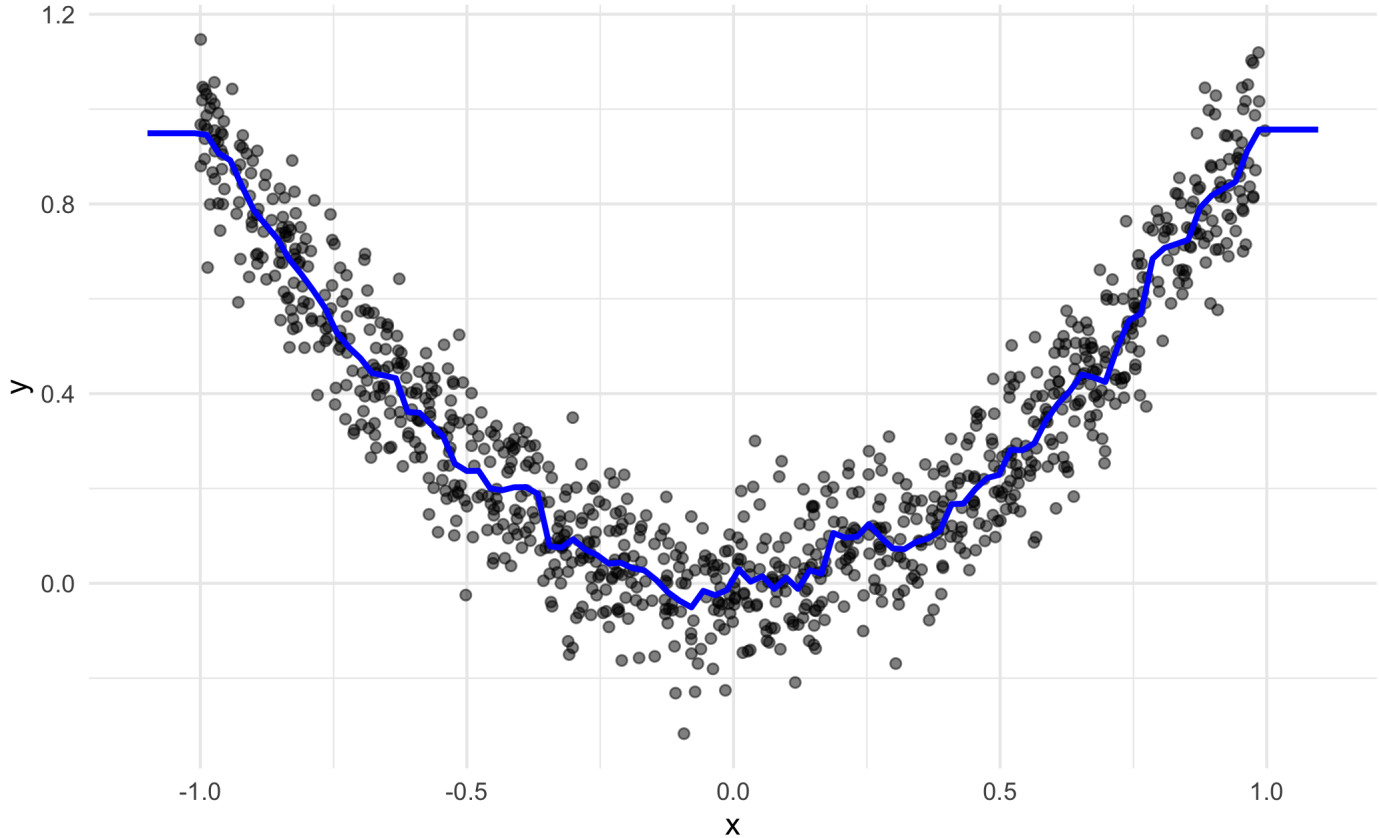
but each time a split in a tree is considered, only a random sample of the predictors can be chosen

Random Forest

The sample is typically $m = \sqrt{p}$ with p predictors

But this value is tuneable as well, along with everything tuneable from the decision tree

Random Forest



Boosting

Boosting is a general approach that can be used with many statistical machine learning methods

In bagging, we fit multiple decision trees side by side

In Boosting we fit multiple decision trees back to back

Boosting

Algorithm

- Fit a tree $\hat{f}^b$ to the model
- Update the final fit using a shrunken version of the tree
- Update the residuals using a shrunken version of the tree
- repeat B times

Final model

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Boosting

Large values of B can result in overfitting

The shrinkage parameter λ typically takes a small value but will need to be tuned

The number of splits d will need to be tuned as well, typically very small trees are fit during boosting