NYU FRE 7773 - Week 5

Machine Learning in Financial Engineering

~~Ethan Rosenthal~~ Jacopo Tagliabue & Friends!

Today's Agenda

- Team and project reviews
  - Reminder: by Oct 14, you should come up with practical project ideas and list them in the google spreadsheet.
- Tree and ensemble methods (slides + code)
- Intro to MLOps with Chip!
- TA homework / methodology review
- Intro to metrics

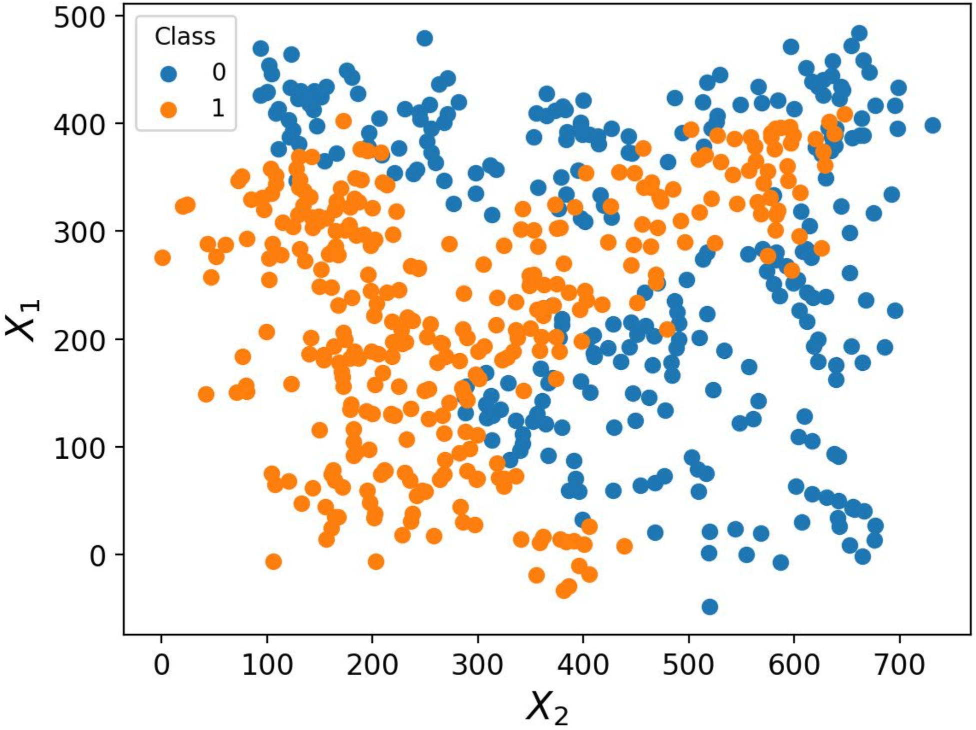
Trees and Ensemble Models

Machine Learning in Financial Engineering

Ethan Rosenthal

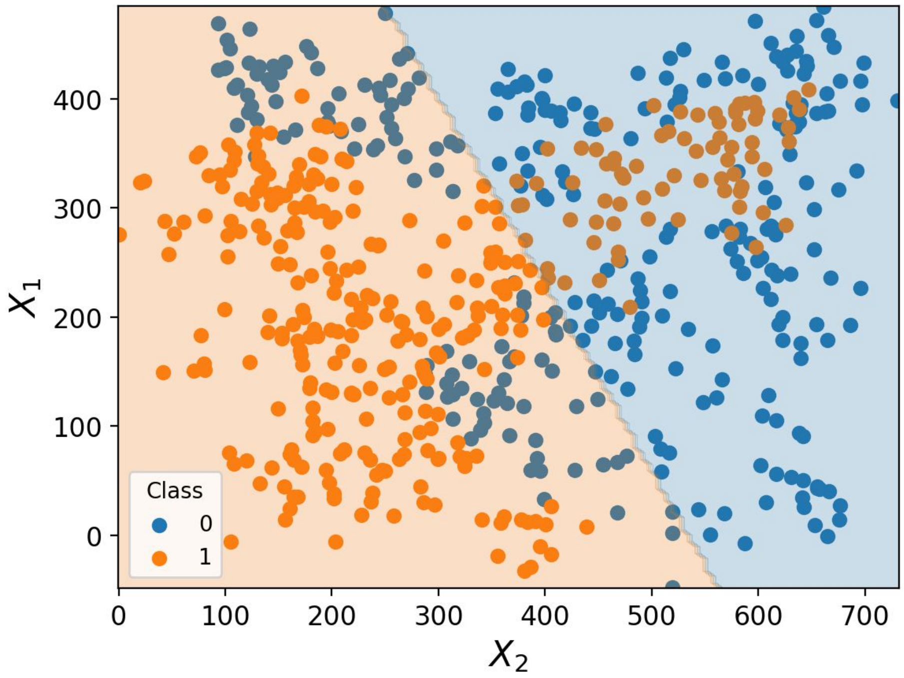
Decision Trees

Limits of Linear Classification



Limits of Linear Classification

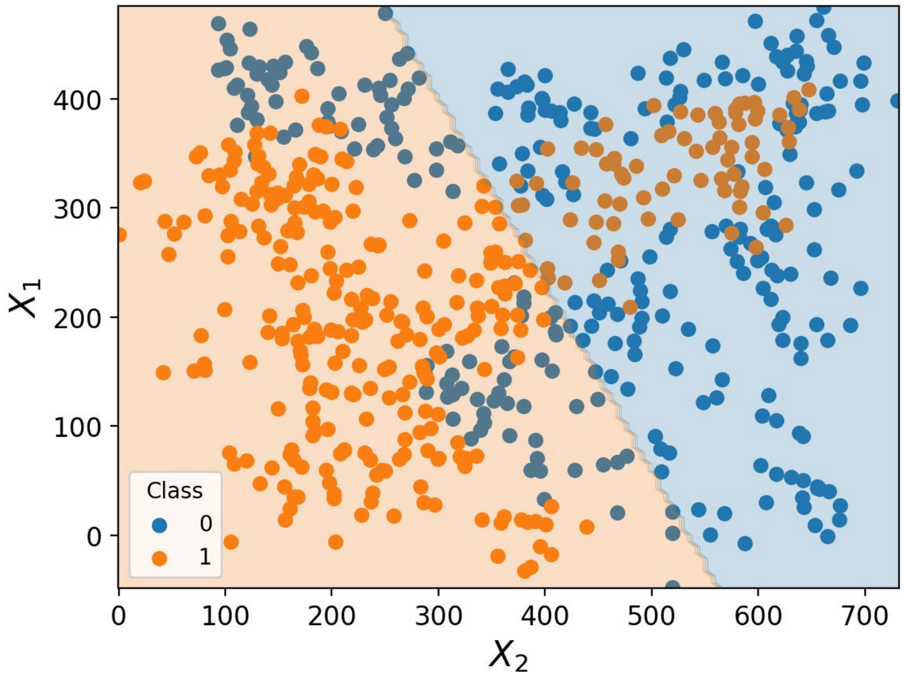
Linear models make linear decision boundaries



Limits of Linear Classification

Linear models make linear decision boundaries

But what if we combined lots of them together?

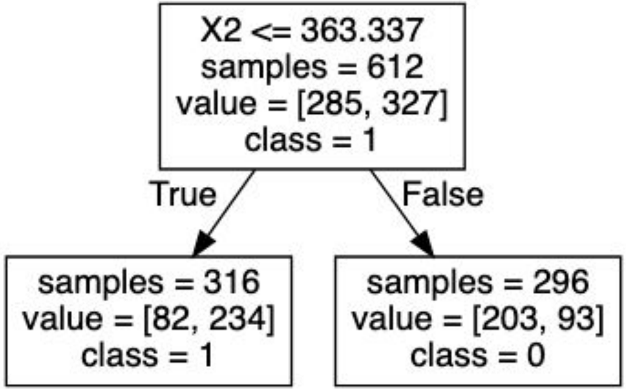


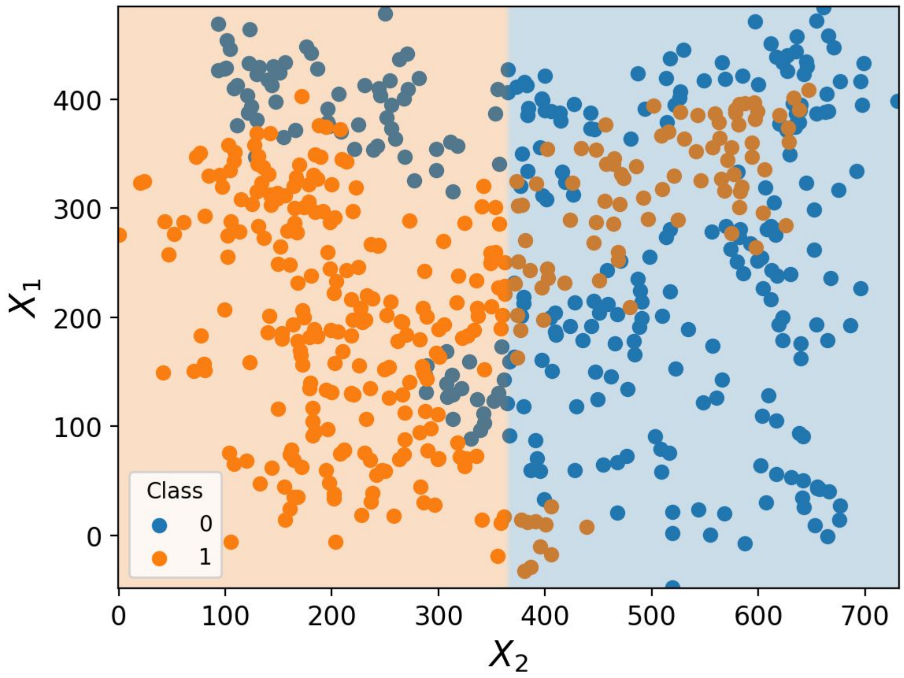
Decision Trees

Start with a very simple "rule":

If X2 <= 363.337,

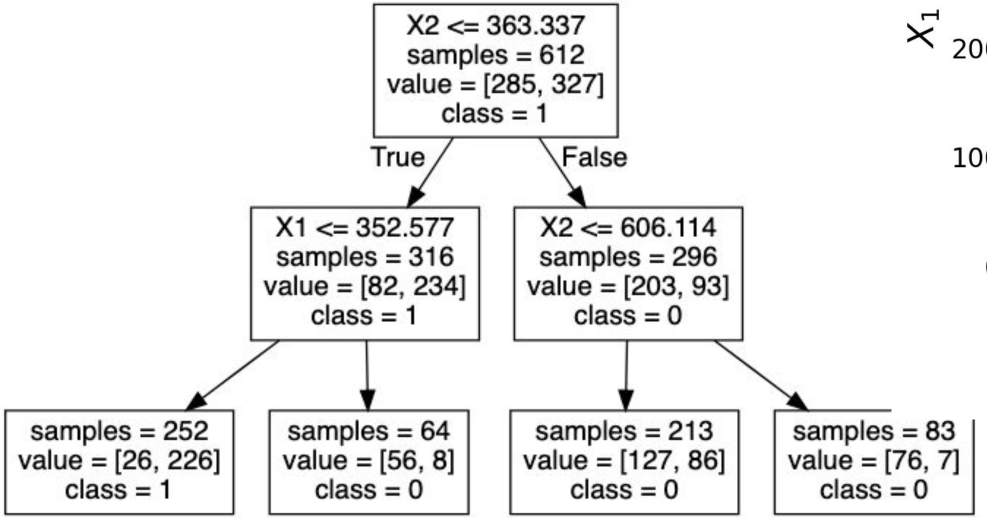
then Class 1, else Class 0

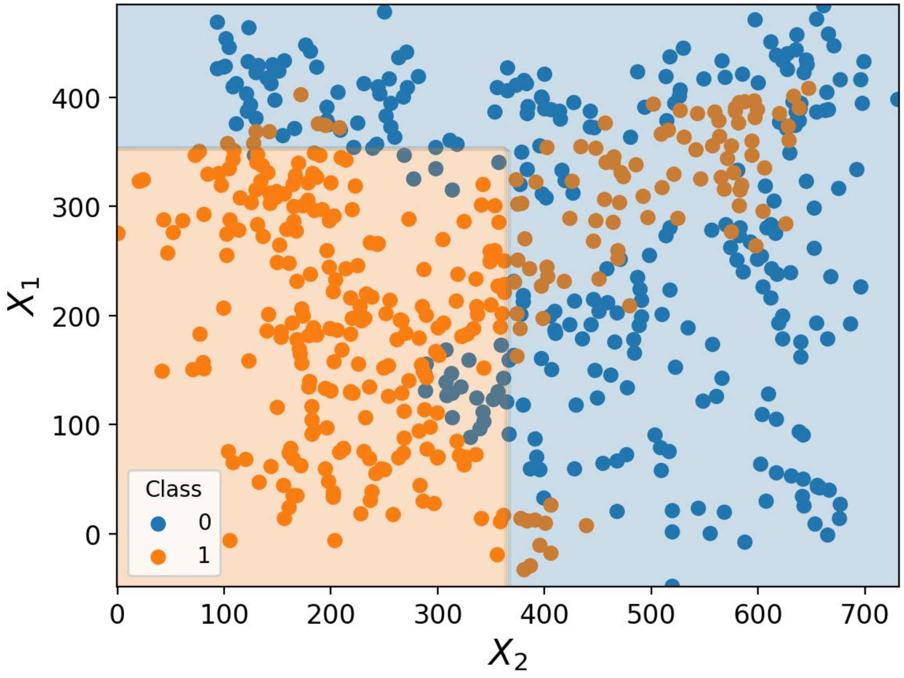




Decision Trees

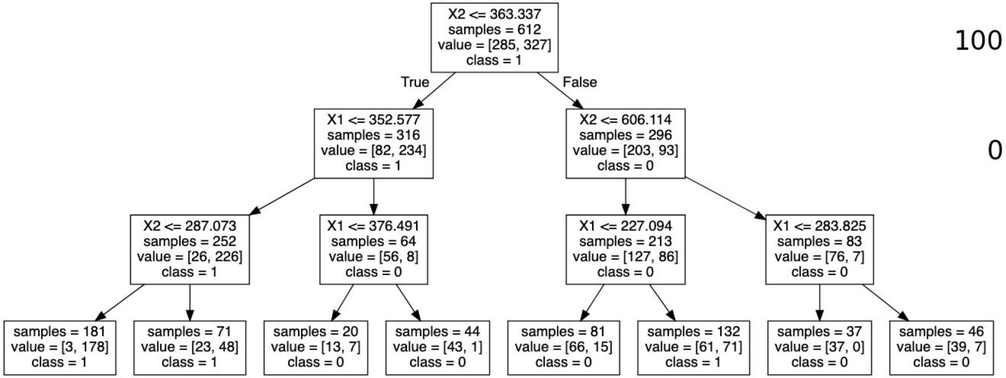
Add more if/else statements, get more sides to the decision boundary.

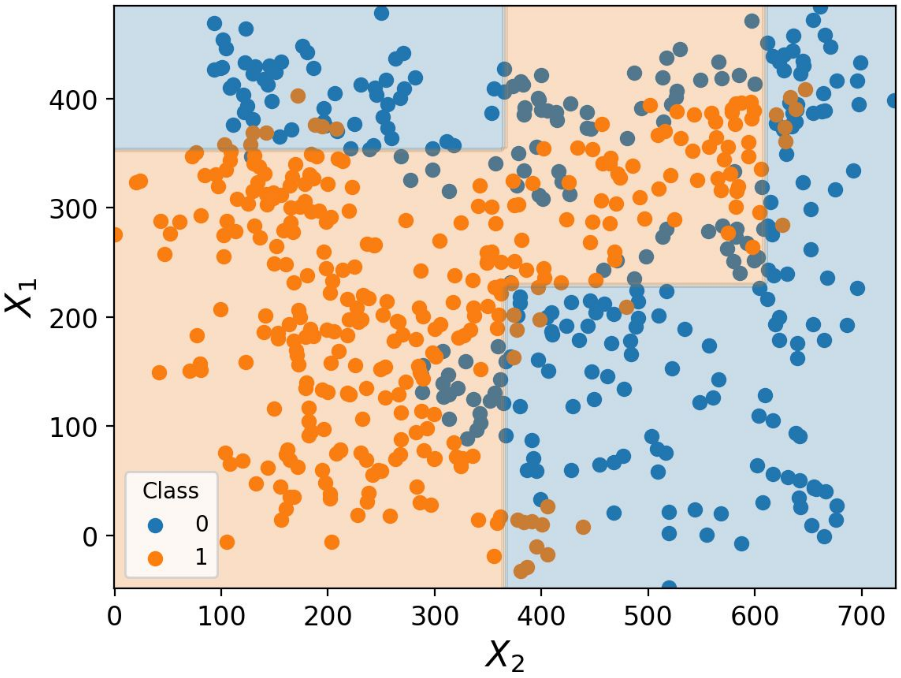




Decision Trees

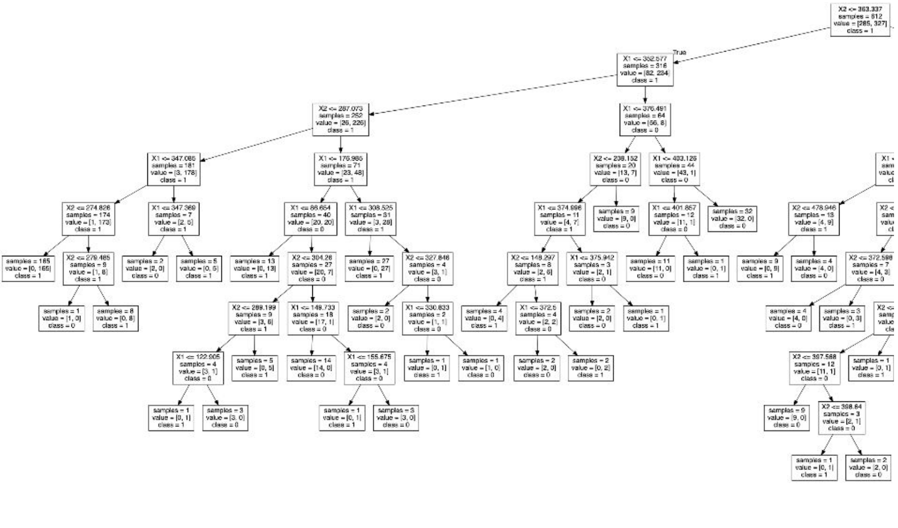
Add more if/else statements, get more sides to the decision boundary.

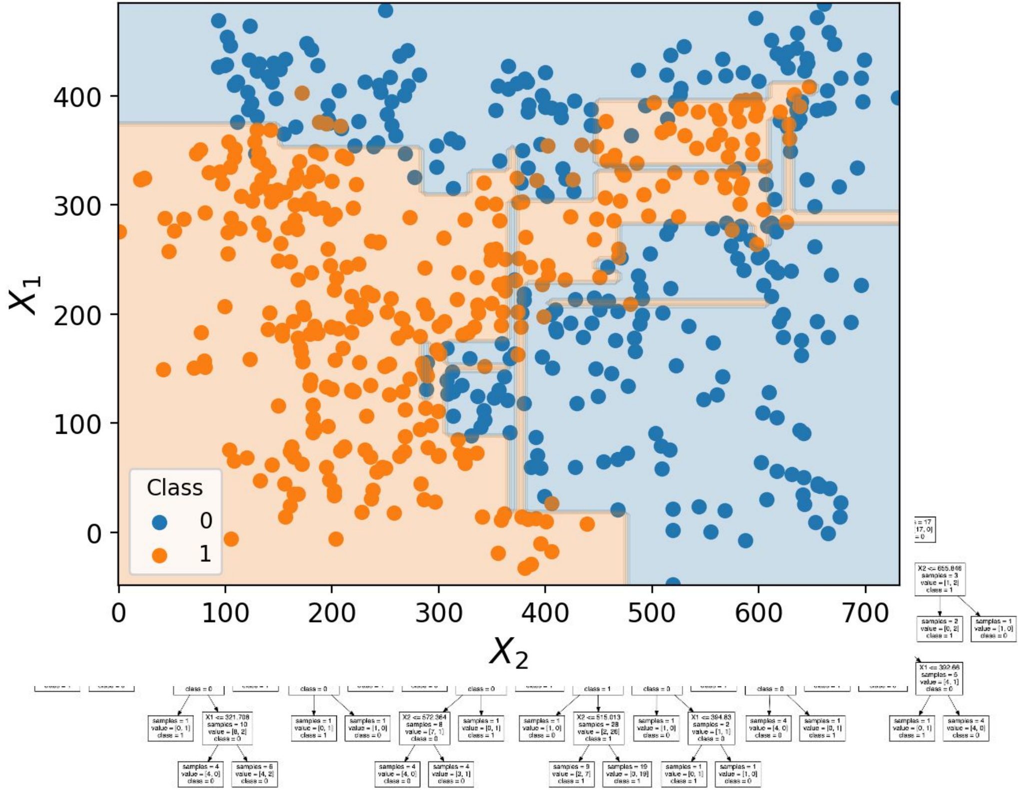




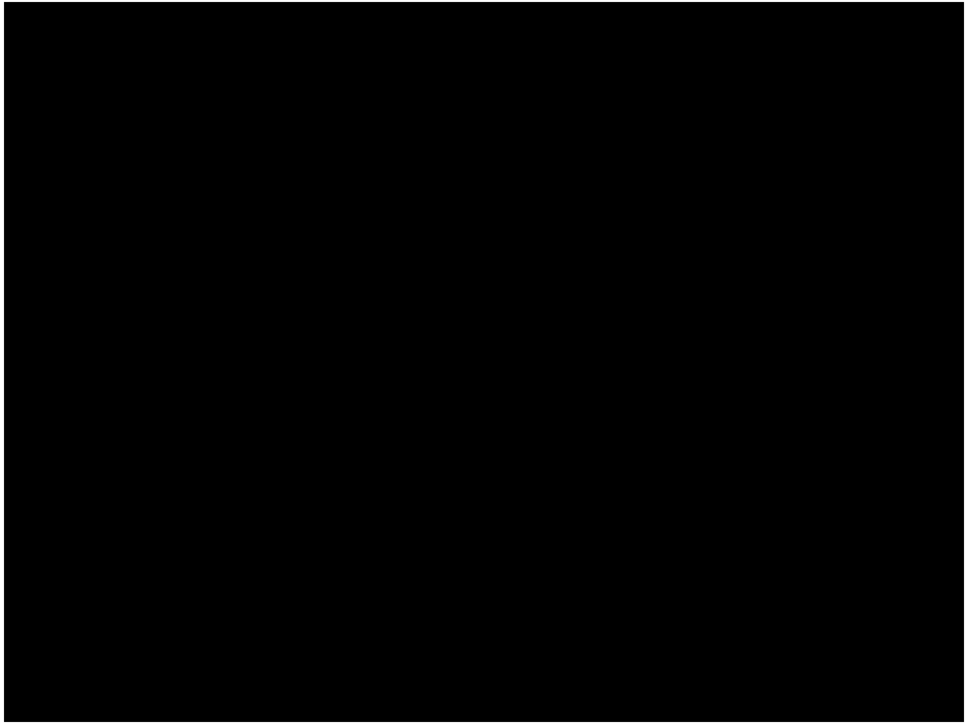
Decision Trees

Add more if/else statements, get more sides to the decision boundary.



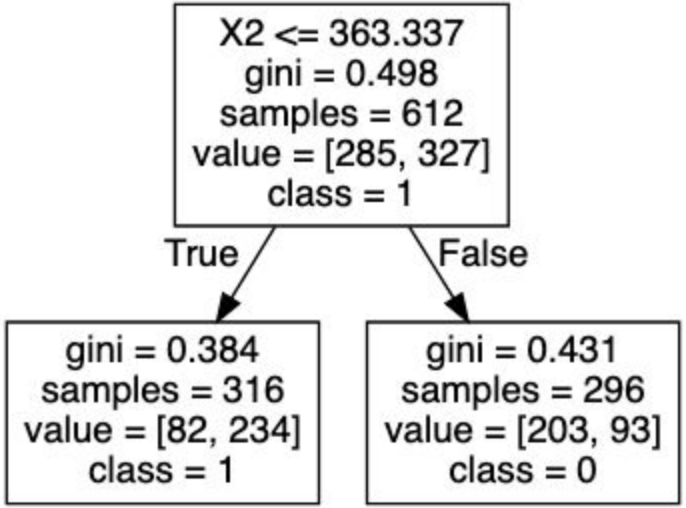


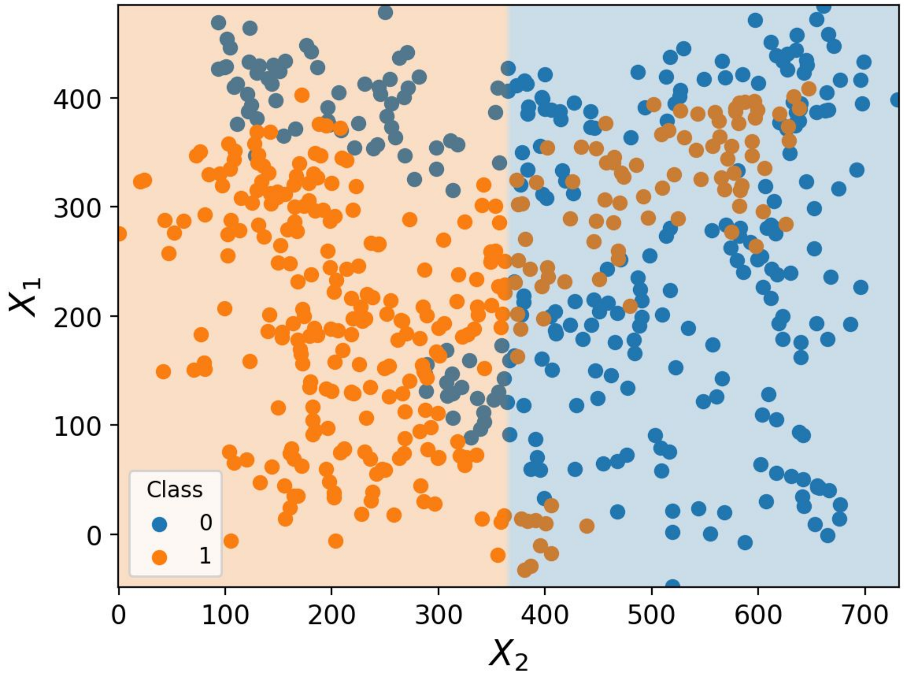
Decision Trees



Decision Trees - How do they Grow?

At each node, goal is to find the feature and threshold that maximally splits the classes.



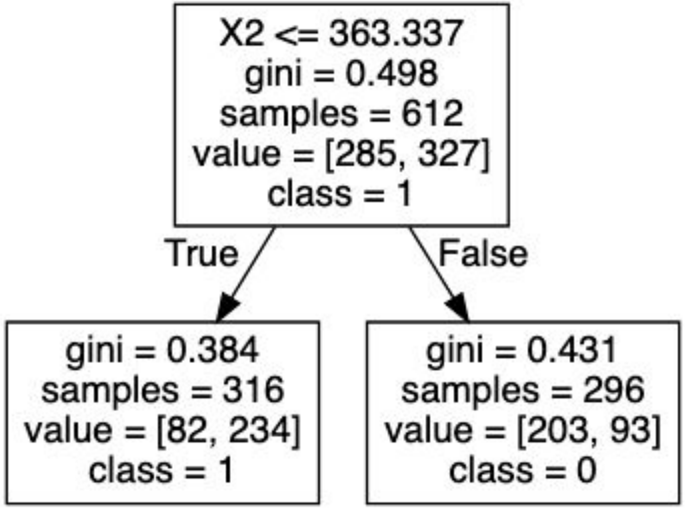


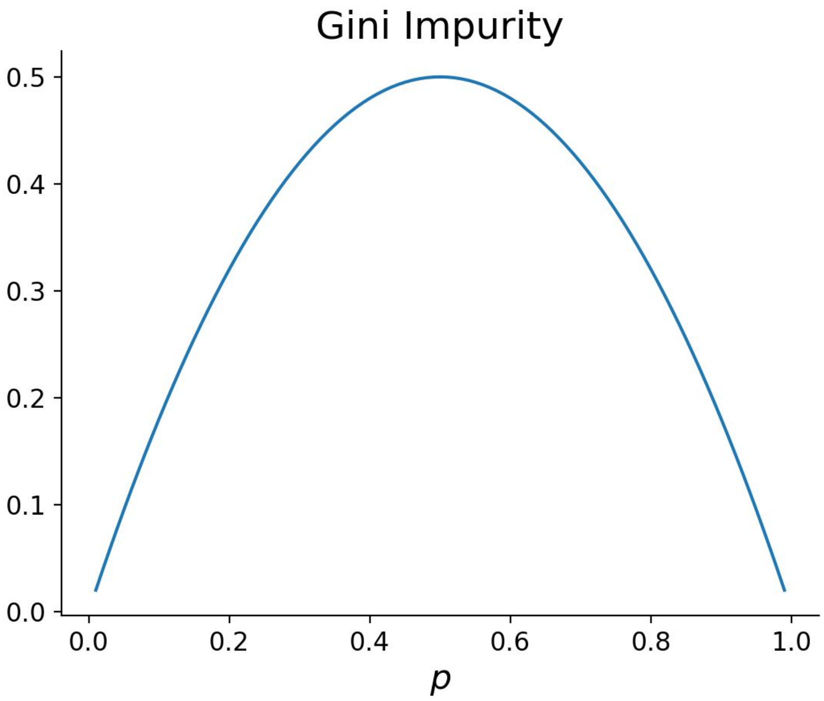
Proportion of samples in the positive class after the node

Decision Trees - How do they Grow?

$$Gini = 2p(1-p)$$

The Gini Impurity is one measure of how well split the classes are.





Decision Trees - How do they Grow?

Various other
“impurity” measures

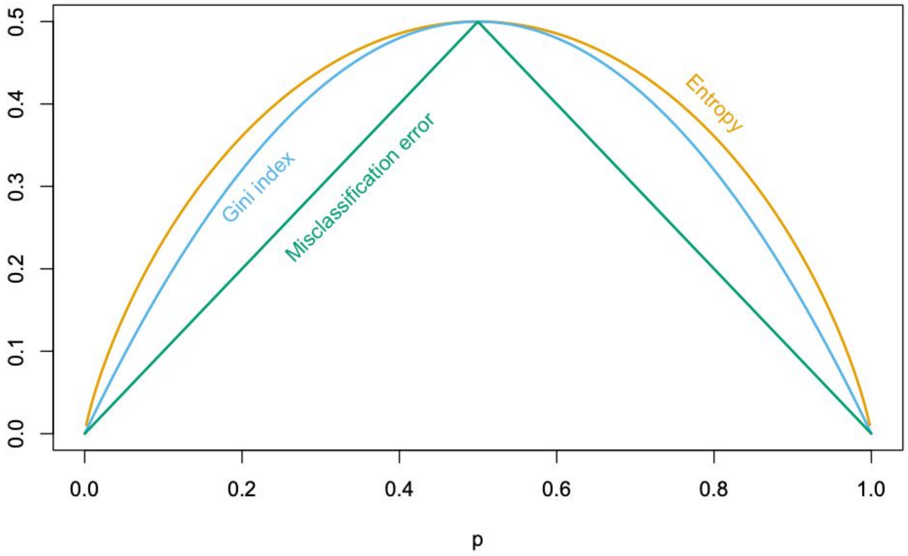


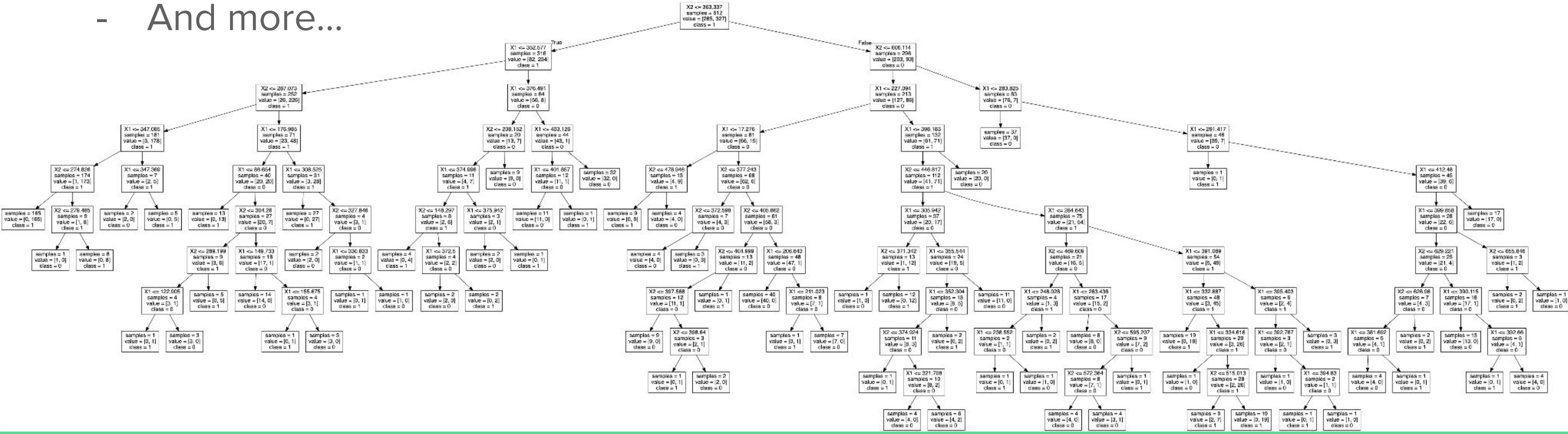
FIGURE 9.3. *Node impurity measures for two-class classification, as a function of the proportion p in class 2. Cross-entropy has been scaled to pass through* (0.5, 0.5).

Hastie et. al., Elements of Statistical Learning
https://hastie.su.domains/ElemStatLearn/

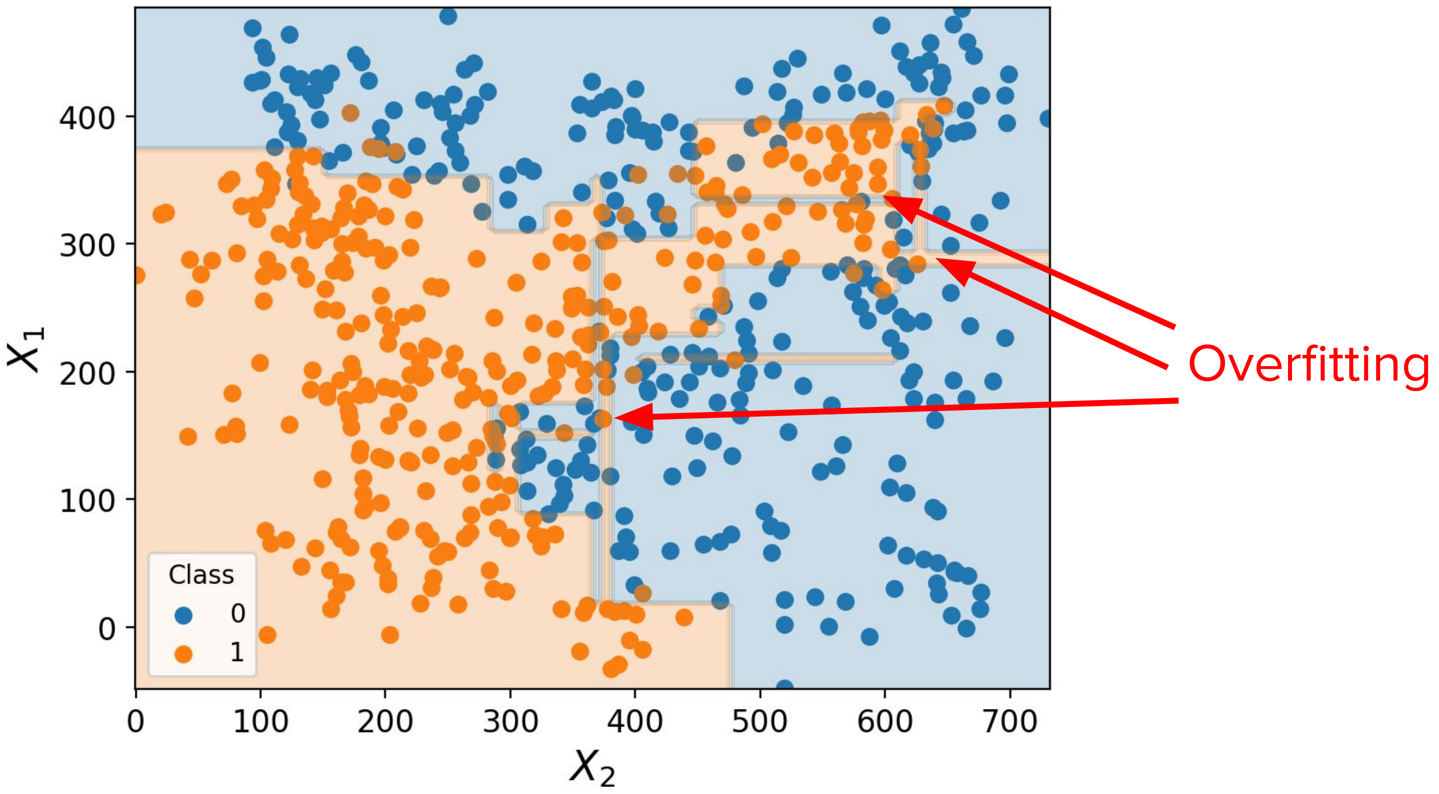
Decision Trees - How do they Grow?

Keep growing the tree until some cutoff criteria:

- Max depth
- Min samples in leaf nodes reached
- Min impurity decrease
- And more...



Limits of Decision Trees



Random Forests

Random Forests - Seeing the Forest for the Trees

Instead of a single decision tree, create many trees (a forest!).

But, induce randomness.

For each tree:

- Generate a *bootstrap* sample of the dataset (i.e. sample with replacement).
- For each node, only consider a subset of the features when deciding what feature to split on.
- The prediction *score* for each class is the fraction of trees that classify the sample into that class.

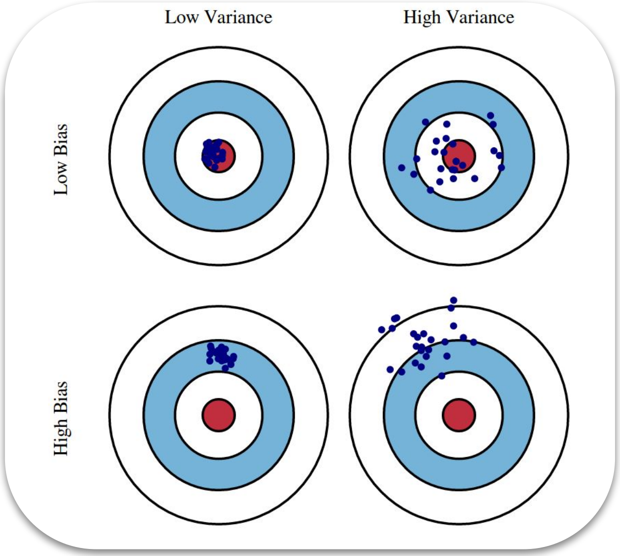
Random Forests - Seeing the Forest for the Trees

- Each tree is not as good at predicting as a single decision tree due induced randomness.
- But, the forest helps prevent overfitting.
- This overfitting prevention is often more powerful than the weakness of each tree, leading to a better overall model.
- (Out of scope, but this is a manifestation of the bias-variance tradeoff).

Random Forests - Seeing the Forest for the Trees

- **(Out of scope, but this is a manifestation of the bias-variance tradeoff).**

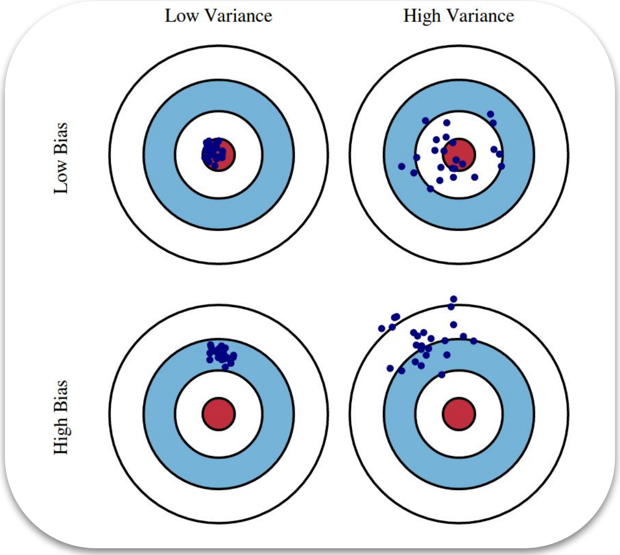
- Each prediction error is a result of:
  - Bias Error: due to our assumptions about the target function.
  - Variance Error: due to the specifics of the dataset.
  - Irreducible Error: nothing we can do here!
- Examples:
  - Regression has low/high bias but a low/high variance.
  - Decision trees have low/high bias but a low/high variance.
  - Random forests have low/high bias but a low/high variance.



The typical bias/variance image in all blog posts on the web!

Random Forests - Seeing the Forest for the Trees

- **(Out of scope, but this is a manifestation of the bias-variance tradeoff).**
- Each prediction error is a result of:
  - Bias Error: due to our assumptions about the target function.
  - Variance Error: due to the specifics of the dataset.
  - Irreducible Error: nothing we can do here!
- Examples:
  - Regression has low/**high** bias but a **low**/high variance.
  - Decision trees have **low**/high bias but a low/**high** variance (**overfitting**).
  - Random forests have **low**/high bias but a low/**high** variance (less than single trees).



The typical bias/variance image in all blog posts on the web!

Random Forests - Why use them?

- Naturally handle nonlinear relationships in the data.
- Quick to fit.
- Robust (but not immune) to overfitting.
- You don’t have to scale your data.
- They can be (kind of) interpretable.
  - See feature_importances_ which measures the total Gini reduction brought by each feature.
- They just work really well!

Guest Speaker: Chip Huyen

MLOps with Chip

- In the introductory lecture, we discussed the importance of going from “your laptop” to “the world”: if you ML model stays on your laptop, **it cannot have much impact**!
- The second part of the course will focus on “ML Operations” (MLOps):
    - today we have *one of the world leading figure* on the topic providing a first look at MLOps.
- **Chip Huyen** is a co-founder of Claypot AI, a platform for real-time machine learning. Previously, she was with Snorkel AI and NVIDIA. She teaches *Machine Learning Systems Design* at Stanford, and she likes to hang out with both your professors, even if she is much cooler than us!

Chip (some years ago)