Decision Trees: CART

Machine Learning

Classification and Regression

- Decision Trees can be used for both

X1	X2	Y
0.268	0.266	Bad
0.219	0.372	Bad
0.517	0.573	Bad
0.269	0.908	Good
0.181	0.202	Bad
0.519	0.898	Good
0.563	0.945	Bad
0.129	0.661	Bad

- <u>Classification</u>
    - Spam / not Spam
    - Admit to ICU /not
    - Lend money / deny
    - Intrusion detections

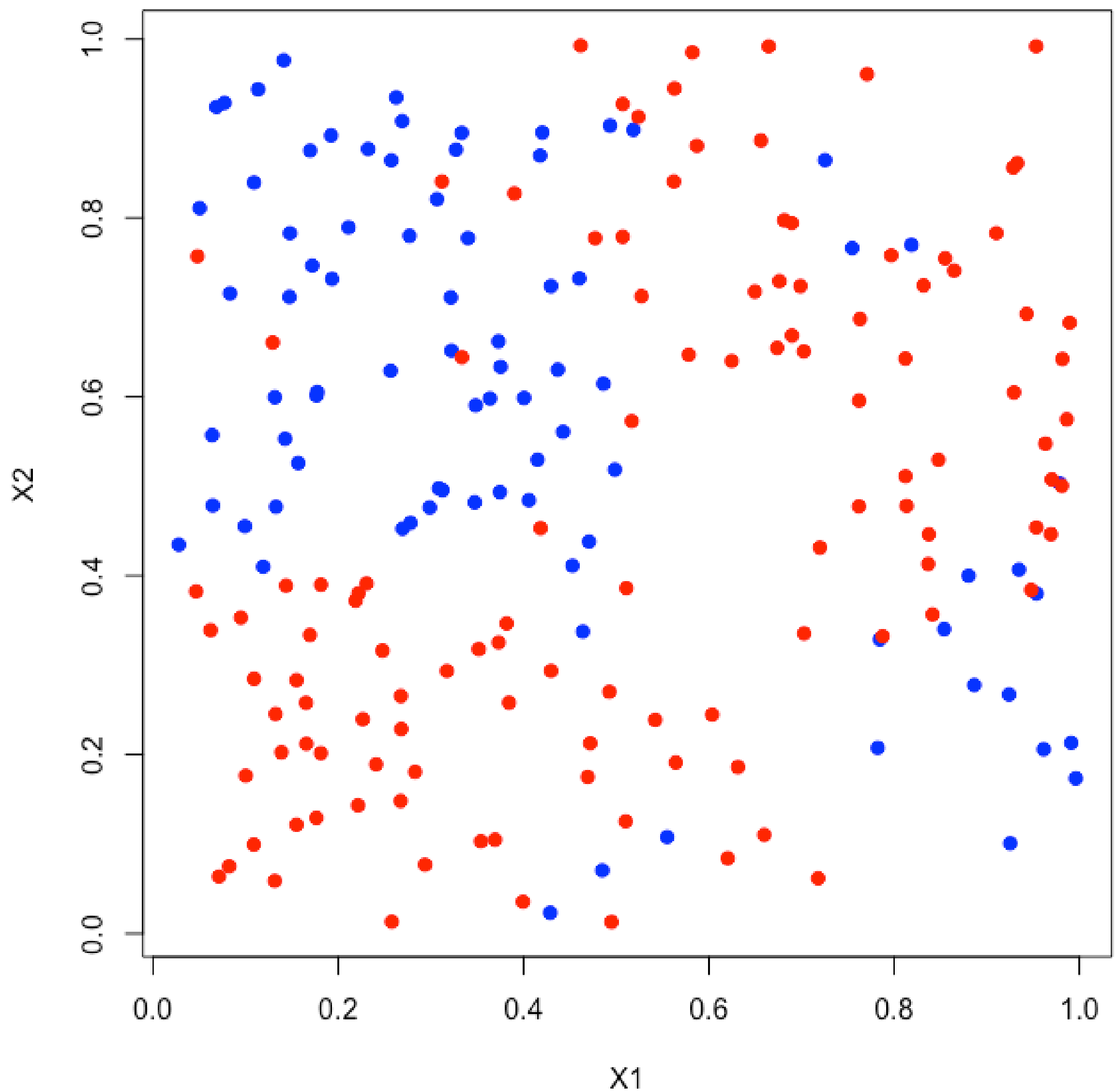
X1	X2	Y
0.268	0.266	64.41
0.219	0.372	28.08
0.517	0.573	95.76
0.269	0.908	15.84
0.181	0.202	41.83
0.519	0.898	25.20
0.563	0.945	9.44
0.129	0.661	82.77

- <u>Regression</u>
    - Predict stock returns
    - Pricing a house or a car
    - Weather predictions (temp, rain fall etc)
    - Economic growth predictions
    - Predicting sports scores

Decision Trees

- The general idea is that we will segment the space into a number of simple regions.
- The segmentation can be illustrated as a tree
- The end nodes can have a category (classification) or a continuous number (regression)
- These methods, while quite simple are very powerful.

Visualizing Classification as a Tree



Metrics

- Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.
- Different algorithms use different metrics for measuring "best"
- These metrics measure how similar a region or a node is. They are said to measure the impurity of a region.
- Larger these impurity metrics the larger the "dissimilarity" of a nodes/regions data.
- Examples: Gini impurity, Entropy, Variance

Algorithms for building Decision Trees

- Popular ones include
  - CART (Classification And Regression Tree)
  - C4.5
  - CHAID (CHi-squared Automatic Interaction Detector)

- We will focus on CART, that uses the Gini impurity as its impurity measure.

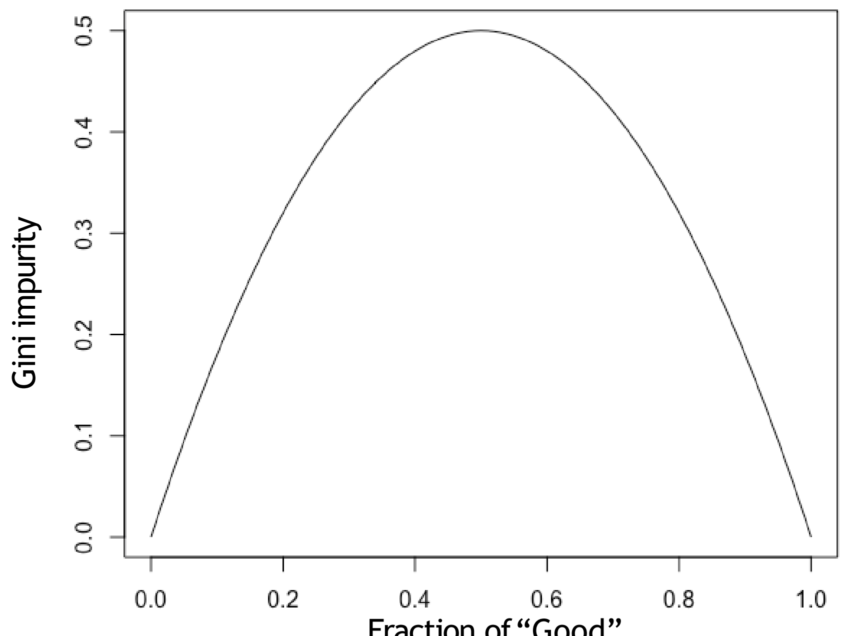
CART: An Example

Cust_ID	Gender	Occupation	Age	Target
1	M	Sal	22	1
2	M	Sal	22	0
3	M	Self-Emp	23	1
4	M	Self-Emp	23	0
5	M	Self-Emp	24	1
6	M	Self-Emp	24	0
7	F	Sal	25	1
8	F	Sal	25	0
9	F	Sal	26	0
10	F	Self-Emp	26	0

Gini impurity

- Used by the CART
- Is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.
- Can be computed by summing the probability of an item with label i being chosen (p_i), times the probability of a mistake ($1 - p_i$) in categorizing that item.
- Simplifying gives, the Gini impurity of a set:

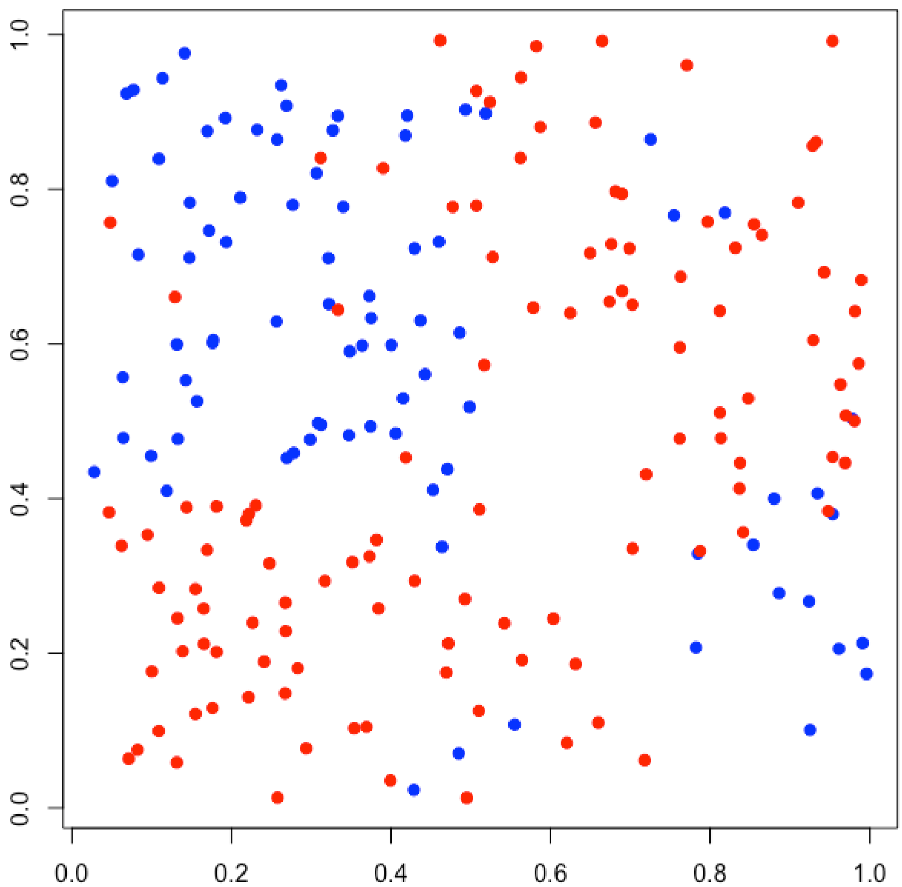
$$1 - \sum_{i=1}^{J} p_i^2$$

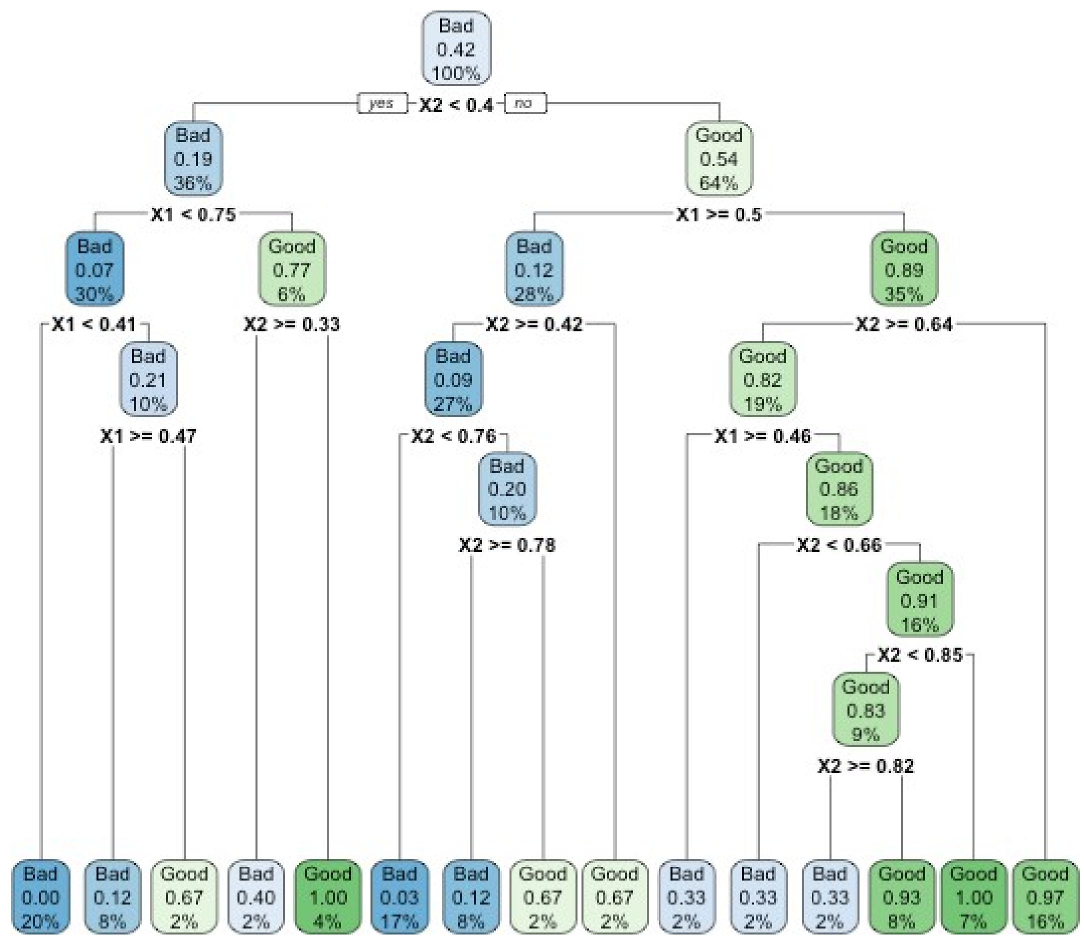


Splitting using Gini impurity

- When splitting, the Gini impurity of the two resulting nodes are combined using a weighted average.
- With weights being the fraction of data on each node.
- The CART algorithm simply chooses the right “split” by finding the split that maximizes the “decrease in Gini impurity” - also called the Gini Gain.

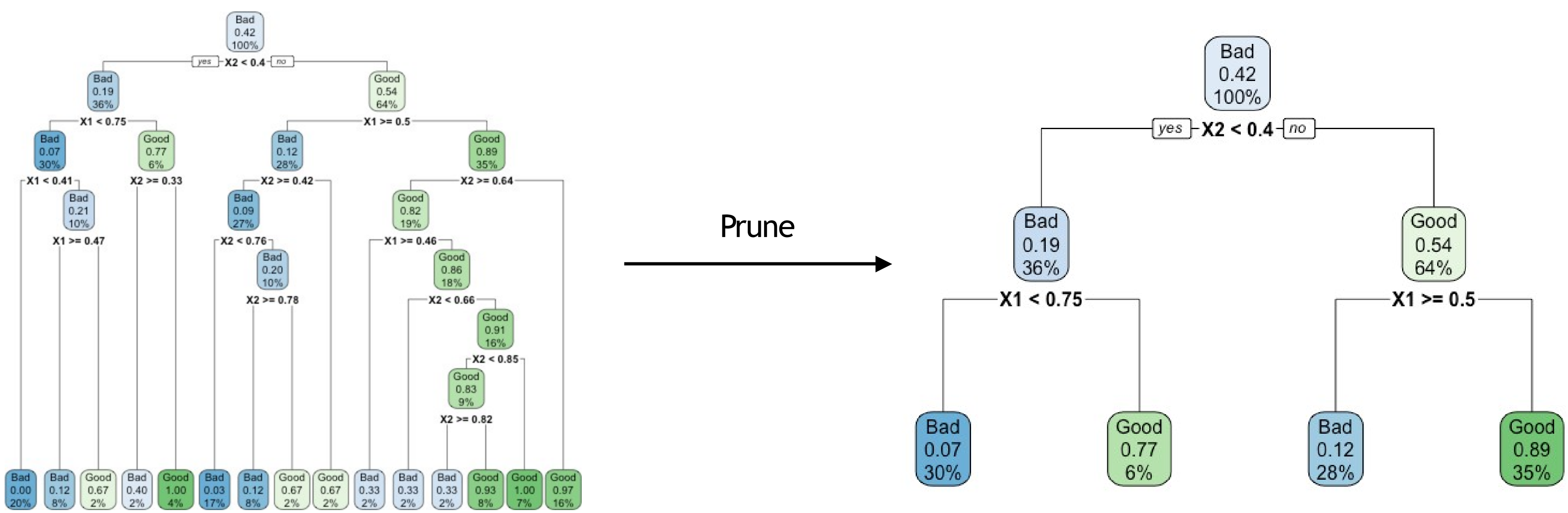
Overfitting in Decision Trees



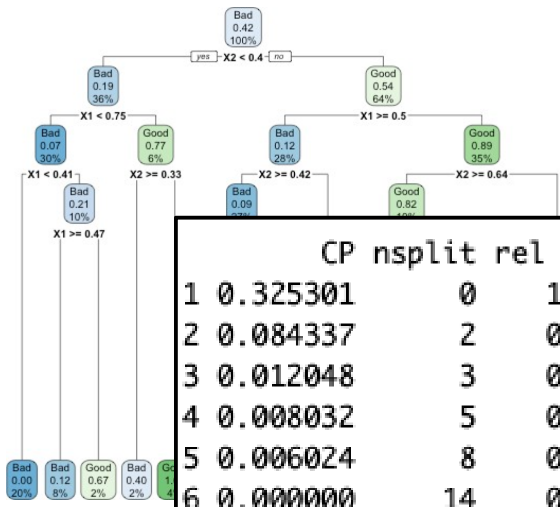


Pruning

- Ideally we would like a tree that does not over-fit the given data
- Pruning can be achieved by saying that each split needs to decreases error by at least amount.
- Cost complexity pruning is the most common at it chooses a CP parameter and requires each split to decrease relative error by at least that amount



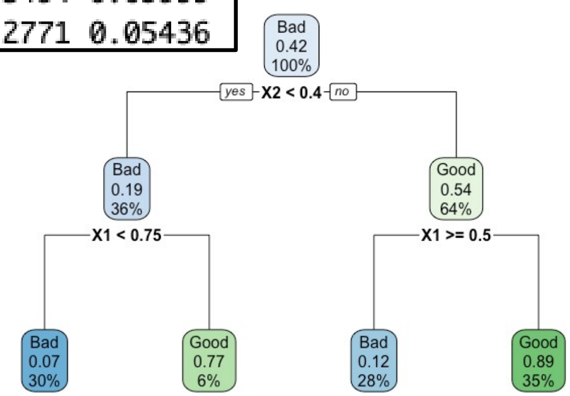
Pruning



	CP	nsplit	rel error	xerror	xstd
1	0.325301	0	1.0000	1.0000	0.08395
2	0.084337	2	0.3494	0.3494	0.05999
3	0.012048	3	0.2651	0.2771	0.05436
4	0.008032	5	0.2410	0.3133	0.05730
5	0.006024	8	0.2169	0.3133	0.05730
6	0.000000	14	0.1807	0.4217	0.06474

Prune with
α = 0.15

	CP	nsplit	rel error	xerror	xstd
1	0.325301	0	1.0000	1.0000	0.08395
2	0.084337	2	0.3494	0.3494	0.05999
3	0.012048	3	0.2651	0.2771	0.05436



Cross Validation

- Cross Validation is a common Machine Learning technique that splits the data into n non-overlapping groups, and runs n experiments:
  - In each experiment, n-1 groups are used to train a model and the model is tested on the left out group.
  - The results are summarized over the n experiments.
- It gives a mechanism that allows us to test a model repeatedly on data that was not used to build the model.
- For Decision Trees, a very common approach is simply to choose the tree with minimum cross validation error

	CP	nsplit	rel error	xerror	xstd
1	0.325301	0	1.0000	1.0000	0.08395
2	0.084337	2	0.3494	0.3494	0.05999
3	0.012048	3	0.2651	0.2771	0.05436
4	0.008032	5	0.2410	0.3133	0.05730
5	0.006024	8	0.2169	0.3133	0.05730
6	0.000000	14	0.1807	0.4217	0.06474

Regression Trees

	Price	Country	Reliability	Mileage	Type
Acura Integra 4	11950	Japan	Much better	NA	Small
Dodge Colt 4	6851	Japan	NA	NA	Small
Dodge Omni 4	6995	USA	Much worse	NA	Small
Eagle Summit 4	8895	USA	better	33	Small
Ford Escort 4	7402	USA	worse	33	Small
Ford Festiva 4	6319	Korea	better	37	Small
GEO Metro 3	6695	Japan	NA	NA	Small
GEO Prizm 4	10125	Japan/USA	Much better	NA	Small
Honda Civic 4	6635	Japan/USA	Much better	32	Small
Hyundai Excel 4	5899	Korea	worse	NA	Small
Mazda Protege 4	6599	Japan	Much better	32	Small