GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 6

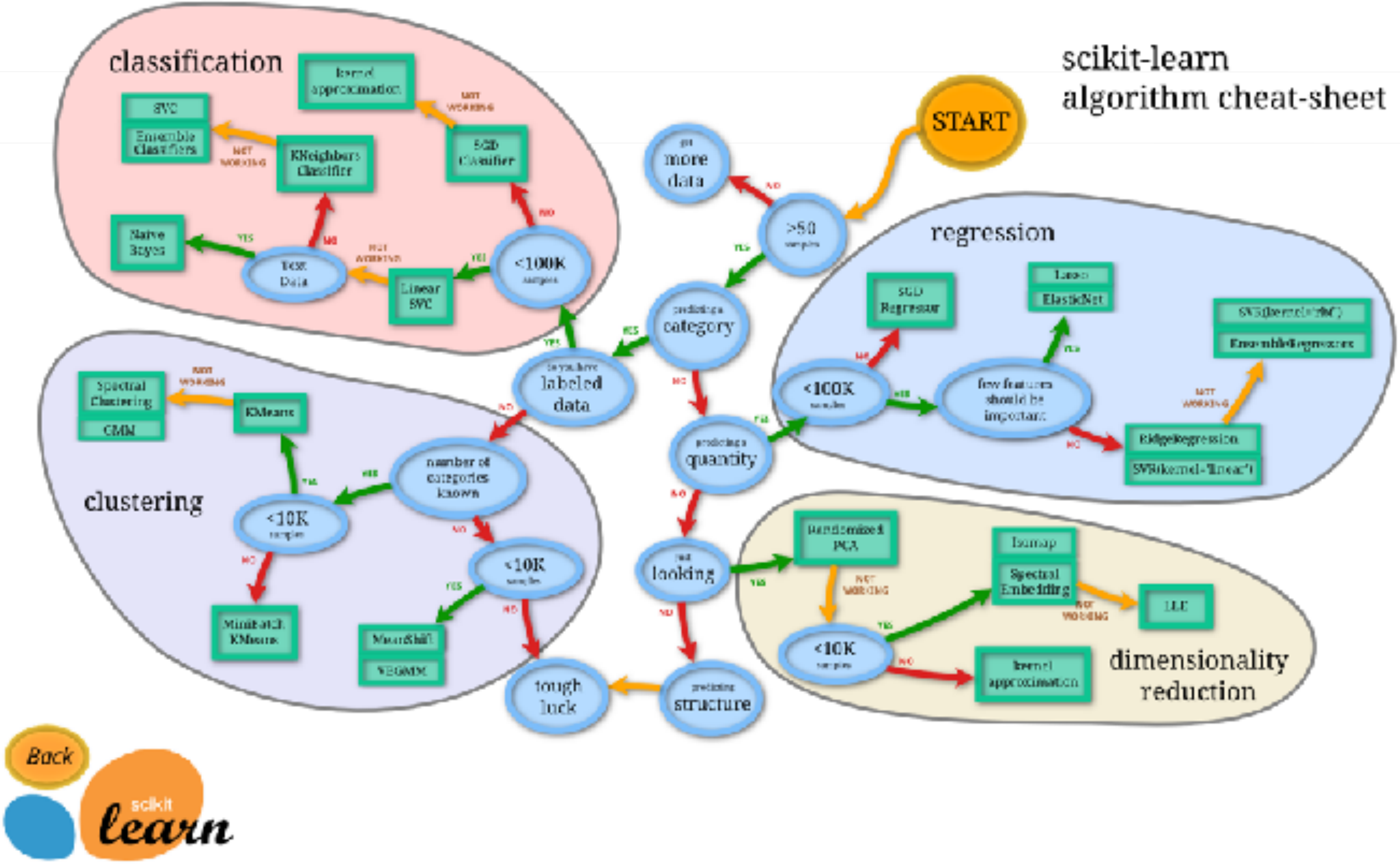
Week 5 - Decision Trees
Monday 7th November

AGENDA

1. What are decision trees?
2. What are decision trees useful for?
3. How decision trees work
4. Visual example on Titanic dataset
5. Lab
6. Talks
7. Discussion

DATA SCIENCE PART TIME COURSE

DECISION TREES



scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality reduction
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
kernel approximation
Naive Bayes
Text Data
Linear SVC
<100K samples
>50 samples
more data
category
labeled data
quantity
looking
structure
tough luck
SGD Regressor
Lasso
ElasticNet
<100K samples
few features should be important
RidgeRegression
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING
Back
scikit learn

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.

WHAT ARE DECISION TREES?

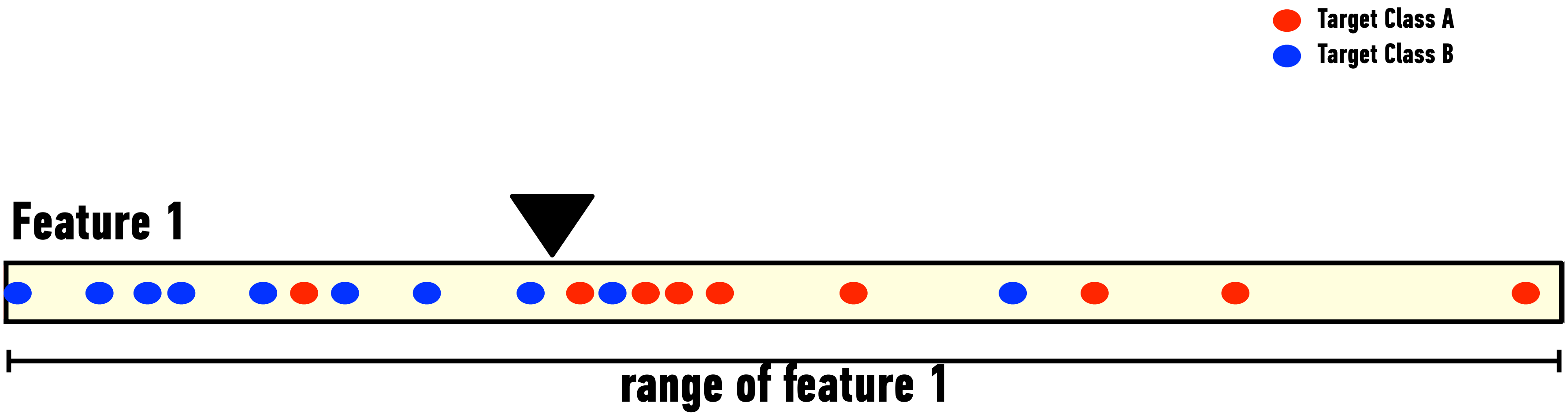
- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.
- Great for explaining to people!

- Prone to overfitting.

DECISION TREES - SHORTCOMINGS

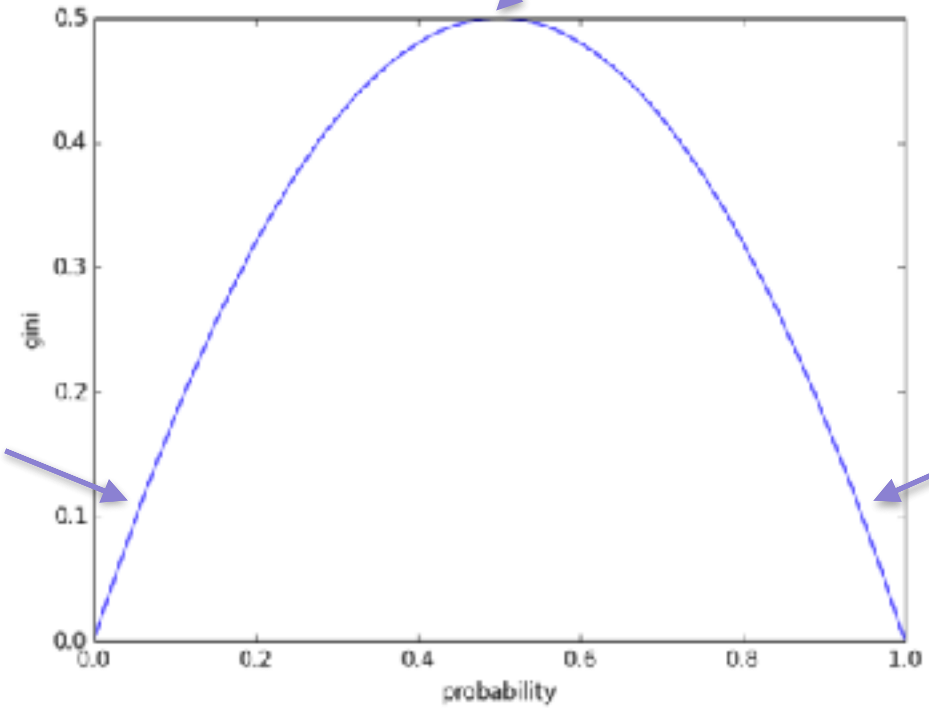
- Prone to overfitting.
- Predictive power is lower in comparison to many other modern techniques.

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.



The Gini Index

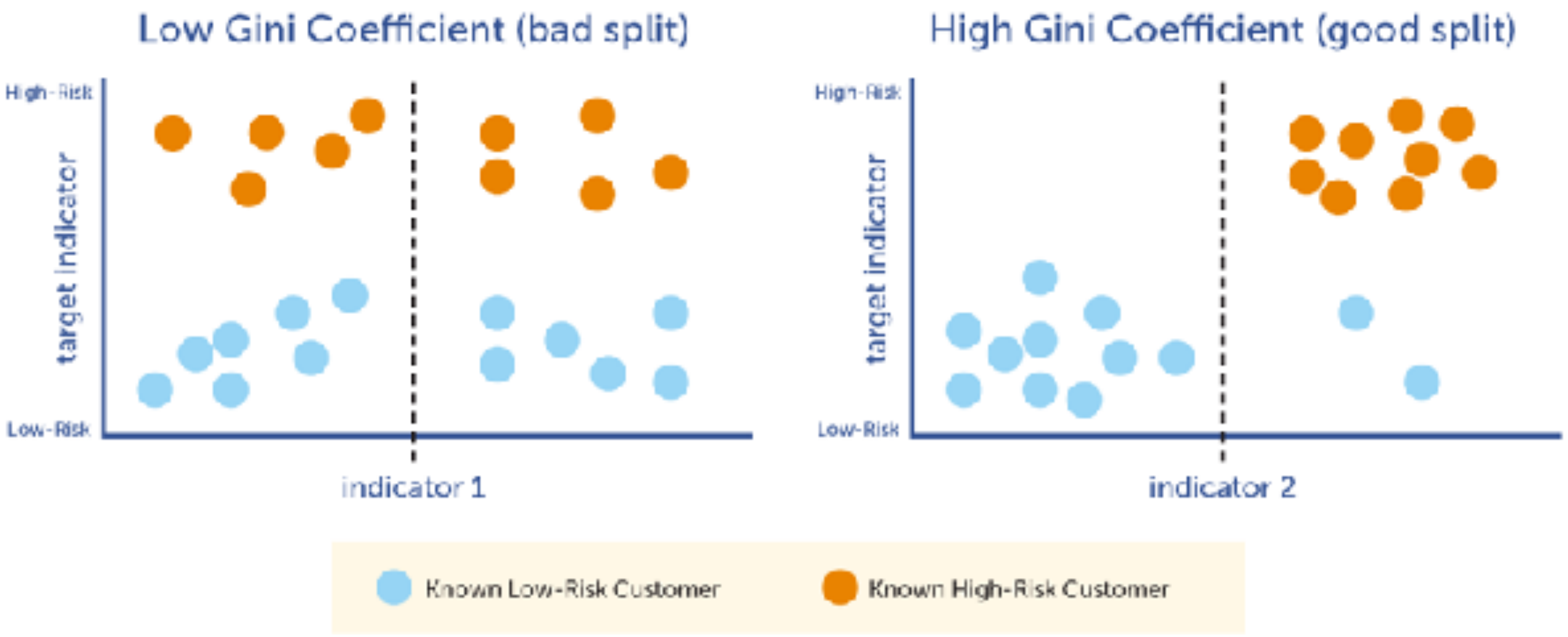
Equal ratio of target classes 50:50



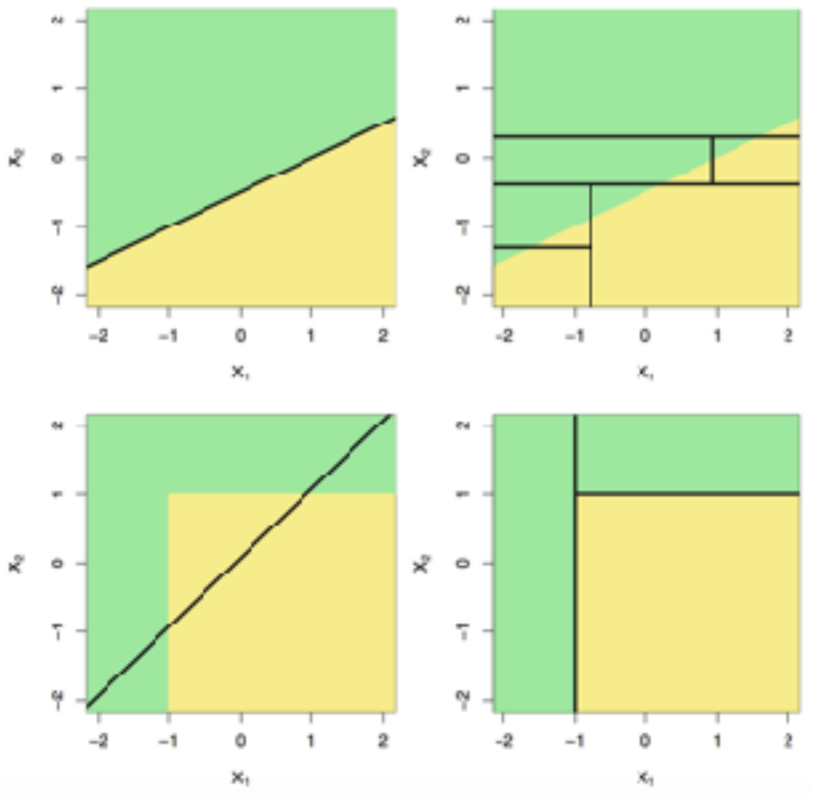
High purity of class 0

High purity of class 1

The Gini Index



- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear.



Linear decision boundary

Non-linear decision boundary

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.
- We naturally get combinations of features used for our prediction.

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

TITANIC DATA

Target

Features

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7
2	1	1	Cumings, Mrs. John Bradley (Florence Bri	female	38	1	0	PC 17599	71
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	8
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Pe	female	35	1	0	113803	53
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8
6	0	3	Moran, Mr. James	male		0	0	330877	8
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	52
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelm	female	27	0	2	347742	11
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30

In pairs, pick the two features from the titanic dataset that you believe will be the most predictive of survival.

Variable	Description
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin

Before Split	All
Survived	10
Died	15

$$1 - \sum \left(\frac{class_i}{total} \right)^2$$

Before Split	All
Survived	10
Died	15

$$1 - \sum \left(\frac{class_i}{total} \right)^2$$

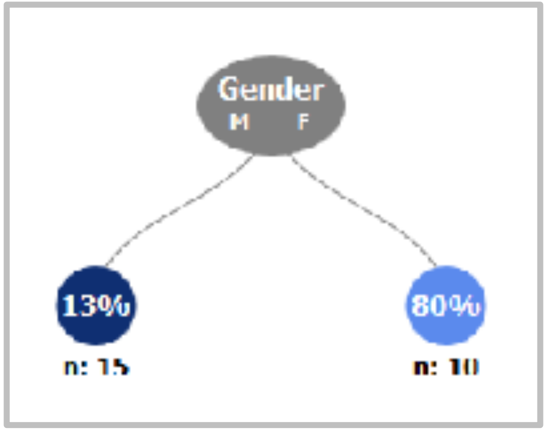
$$1 - \left(\frac{survived}{total} \right)^2 - \left(\frac{died}{total} \right)^2$$

Before Split	All
Survived	10
Died	15

$$1 - \left(\frac{survived}{total}\right)^2 - \left(\frac{died}{total}\right)^2$$

$$1 - \left(\frac{10}{25}\right)^2 - \left(\frac{15}{25}\right)^2 = 0.48$$

$Gini_O$
$= 0.48$

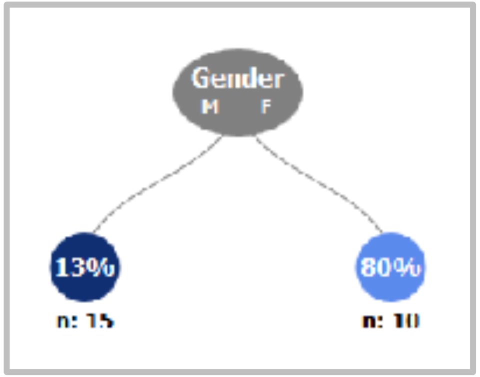


Gender	M
Survived	2
Died	13

Gender	F
Survived	8
Died	2

$Gini_O$

$= 0.48$



$Gini_M$

$= 0.23$

$Gini_F$

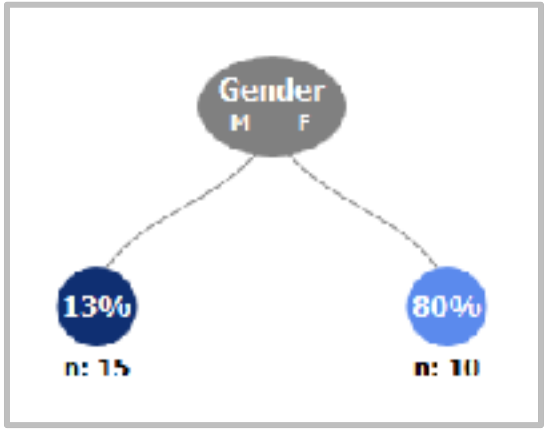
$= 0.32$

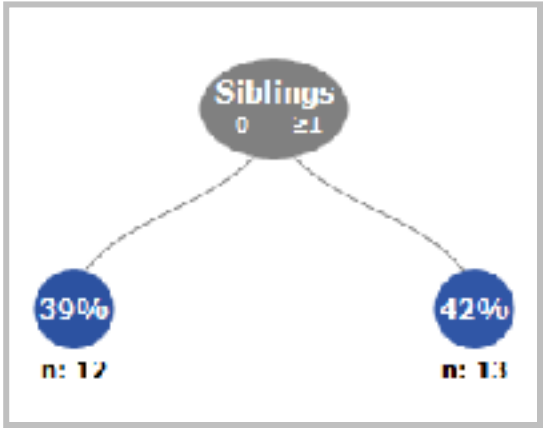
$Gini_C$

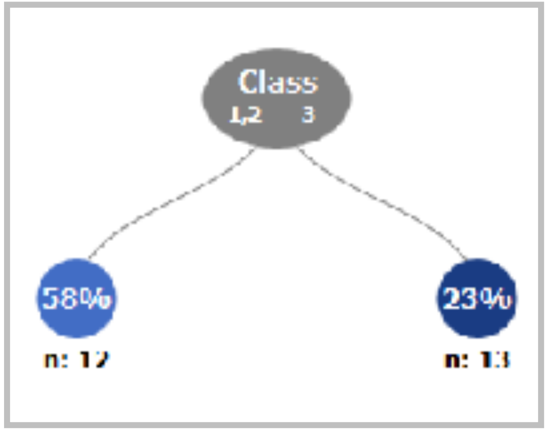
$$Gini_M\left(\frac{M}{M+F}\right) + Gini_F\left(\frac{F}{M+F}\right) =$$

$$0.23\left(\frac{15}{10+15}\right) + 0.32\left(\frac{10}{10+15}\right) = 0.27$$

SPLITTING – USING GINI INDEX – How a first split GINI is calculated





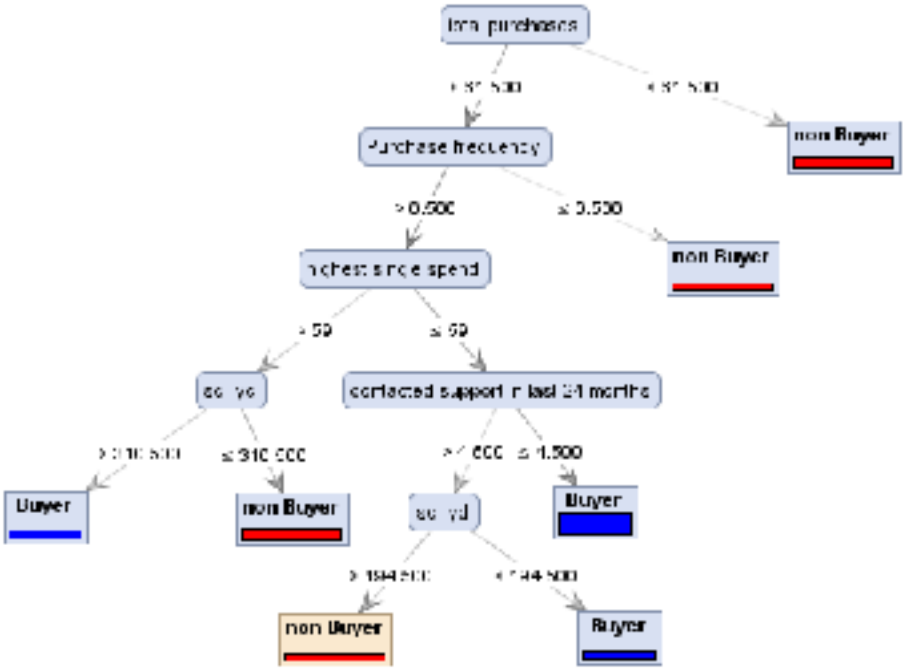


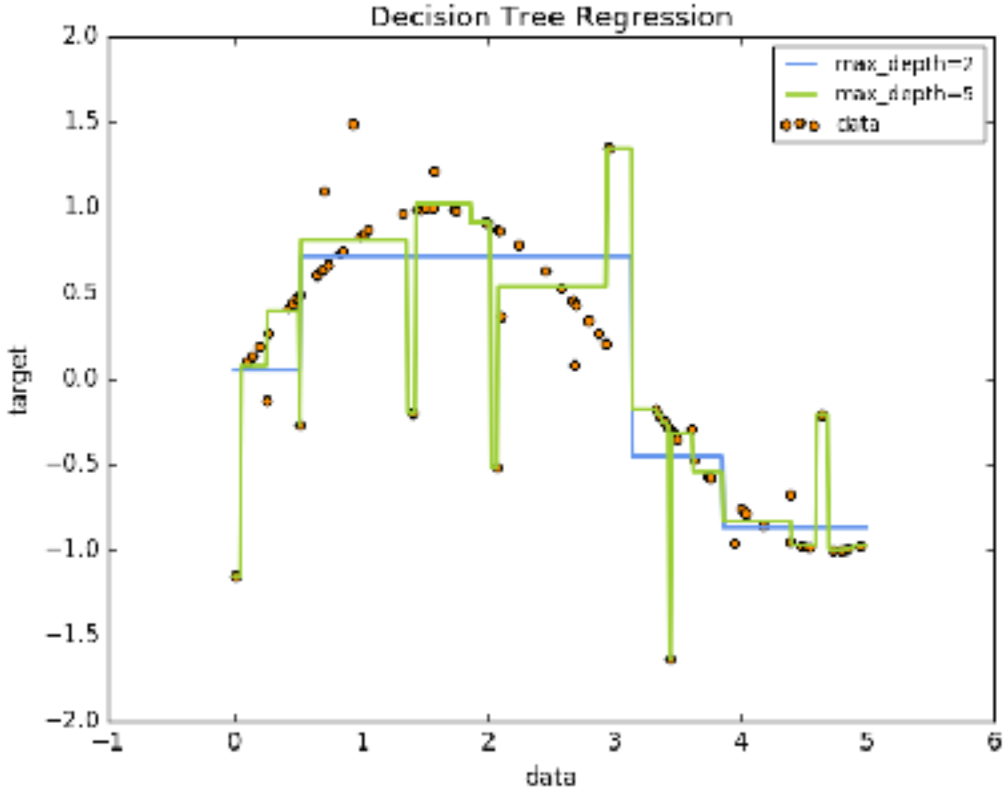
Gender	M	F
Survived	2	8
Died	13	2
$Gini_C$	0.27	

Siblings	0	≥1
Survived	5	5
Died	7	8
$Gini_C$	0.48	

Class	1,2	3
Survived	7	3
Died	5	10
$Gini_C$	0.42	

USES


Purchase frequency
> 0.500
≤ 0.500
non Buyer
non Buyer
> 59
≤ 59
≤ 1.500
Buyer
non Buyer
Buyer
non Buyer
Buyer


Decision Tree Regression
max_depth=2
max_depth=5
data
target
data
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3
4
5
6

ADVANTAGES AND DISADVANTAGES OF TREES

ADVANTAGES

- Trees are easy to explain to people.
- Trees can be displayed graphically, and are easily interpreted even by a non-expert (especially if they are small).
- Trees can easily handle qualitative predictors without the need to create dummy variables.

DISADVANTAGES

- Trees on their own generally do not have high predictive accuracy.

Using BigML to demonstrate a decision tree model on the Titanic dataset.

https://bigml.com/dashboard/datasets

BigML is a cloud based machine learning tool, designed to make machine learning more approachable.

LAB

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
   - git status
   - git add .
   - git commit -m “descriptive label for the commit”
   - git status
4. download new material from official course repo (upstream) and merge it
   - git checkout master (ensures you are in the master branch)
   - git fetch upstream
   - git merge upstream/master

HOMEWORK

Homework

- Homework 2 - Due Monday 14th of November

Read the following

- Chapter 8.1 of Introduction to Statistical Learning - The Basics of Decision Trees
- Chapter 8.2 of Introduction to Statistical Learning - Bagging, Random Forests, Boosting