
# Decision Trees

Jan 22st 2020

# Decision Trees

- What does it do?
- How does it look?
- Splitting criteria
- Stopping criteria
- Pros / Cons

//

# Choosing threshold

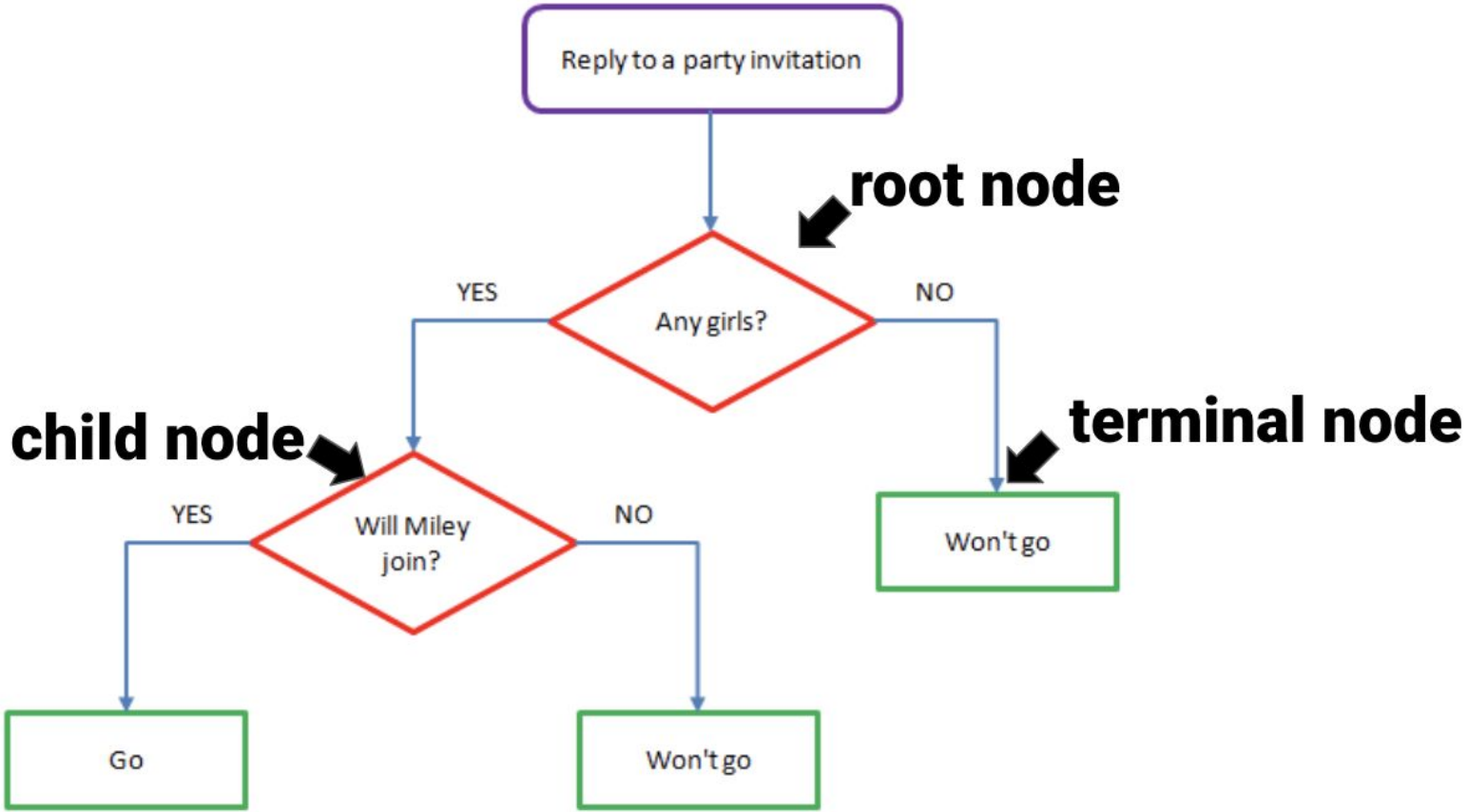
**Task:** Work with your neighbor to choose the best threshold:

- **Threshold 1:** TP = 45 TN = 30 FP = 10 FN = 10
- **Threshold 2:** TP = 35 TN = 15 FP = 5 FN = 5

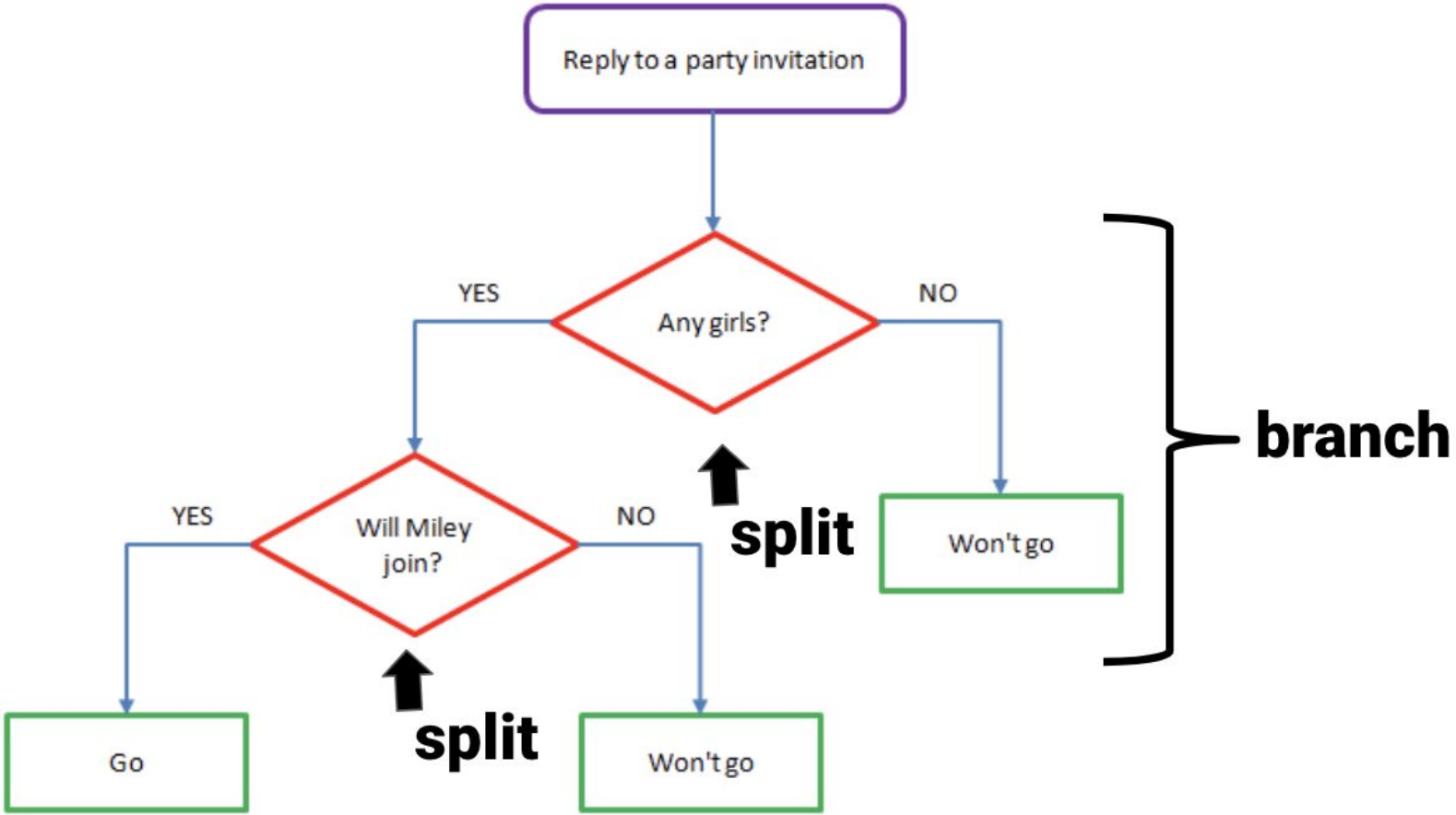
- Take into account the following prevalence and costs:
  - Prevalence: 70%
  - Cost(FP) - Cost(TN) = 10
  - Cost(FN) - Cost(TP) = 5

We will then come back to the large group and I'll pick some of you to share your answers with the rest of the class.

# Interpreting splits

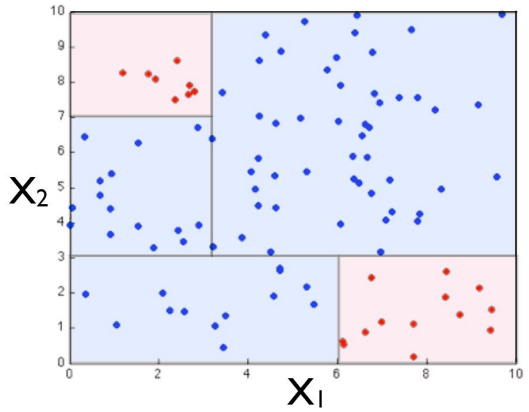


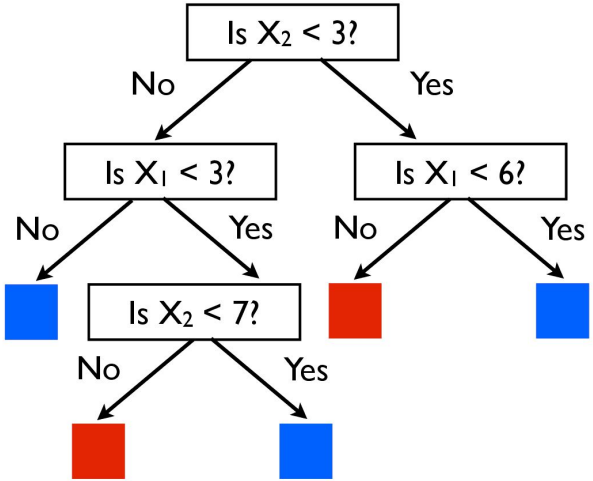
# Interpreting splits



# Interpreting splits

## Decision Tree Classifiers

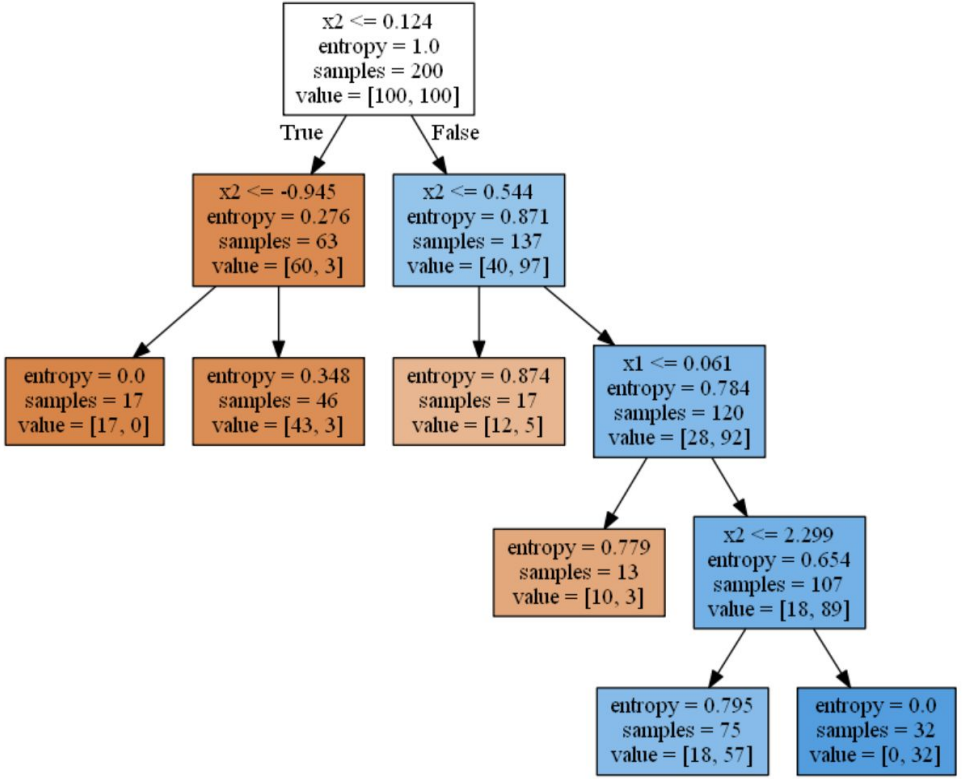


# Interpreting splits

# Decision Tree Classifiers

- Splits your data again and again
- Before each split DTs try:
  - Every predictor
  - Each of its values (sort of)
- The split that makes the resulting groups the most homogenous (concentration of one class) wins
- Each variable can be used as many times as necessary



# Choosing threshold

**Task:** Discuss with your partner:

- What problems are Decision Trees better suited for?
- How does a Decision Tree choose how to split the data?
- What is a leaf? What is the depth of a tree?

We will then come back to the large group and I'll pick some of you to share your answers with the rest of the class.

# Splitting rules

- If categorical predictor:
  - Get dummies (n-1 is variables enough)
  - For each dummy:
    - Split on <= 0.5
      - False means ‘value in column’s header’
      - True means ‘ not the value in the header’

- If float or integer:
  - For all pairs of contiguous values in the variable
    - Split on <= gap between them

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Choosing splitting points

- Let's imagine we only have marital status and refund as predictors of Cheat
- We are going to try Refund and see what splits it generates.
- Because we only have 2 values we have only one splitting point

- With Refund <= 0.5 True [3,4]
- With Refund <= 0.5 False [0,3]

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Splitting like a DT

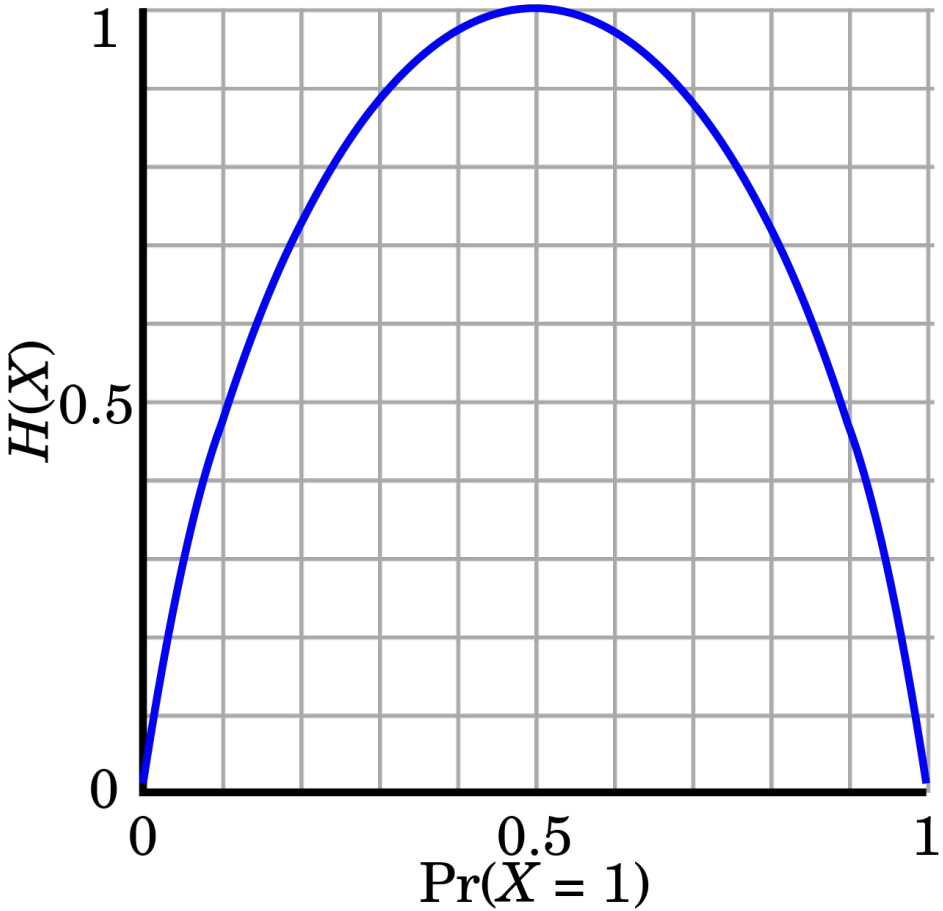
**Task:** Work with your neighbor to figure out the following:

- How many + and - do you have as a result of splitting on Marital Status?
- Which of the two columns that we have seen is preferable?

We will then come back to the large group and I'll pick some of you to share your answers with the rest of the class.

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Split metrics



# Entropy

- Entropy (disorder)

$$E = -\Sigma_i p_i \log_2(p_i),$$

$$-p\log_2(p) - q\log_2(q)$$

$$-0.4 * \log_2(0.4) - 0.6 * \log_2(0.6).$$

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Gini Impurity

- Gini (impurity)

$$I_G(p) = 1 - \sum_{i=1}^{J} {p_i}^2$$

$$= 1 - p^2 - q^2$$

$$= 1 - 0.4^2 - 0.6^2$$

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

# Splitting like a DT

**Task:** Work with your neighbor to figure out the following:

- What are the disorder and impurity values for the split resulting from Marital Status?
- Which of the two columns that we have seen is preferable?

We will then come back to the large group and I'll pick some of you to share your answers with the rest of the class.

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

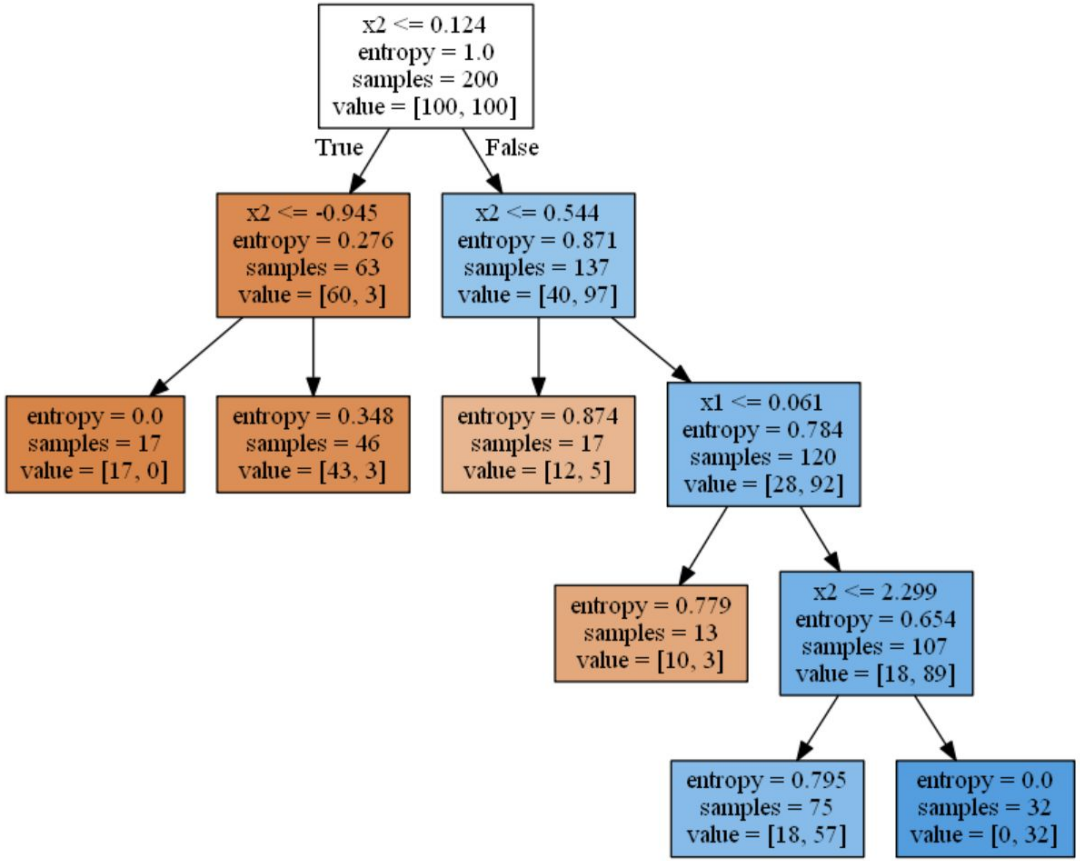
# Stopping Criteria

- There is a rumor out there that says that Decision Trees are very prone to overfit
- If you let them, they won't stop splitting until they achieve absolute purity/order
- With real life problems that will mean overfitting

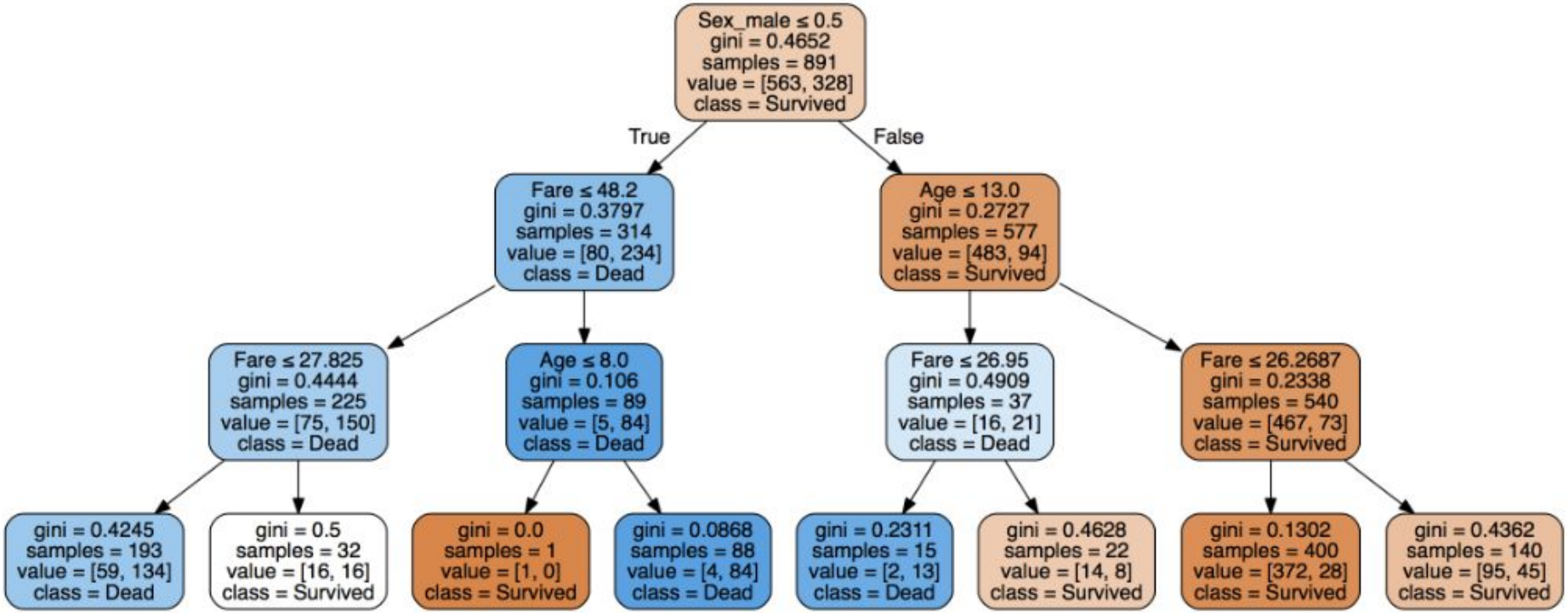
- Thankfully we have stopping criteria
- When the criteria is met the tree will stop splitting

- Because stopping criteria have a huge impact on performance they are hyperparameters that you want to tune (more on this soon)

# min_samples_leaf

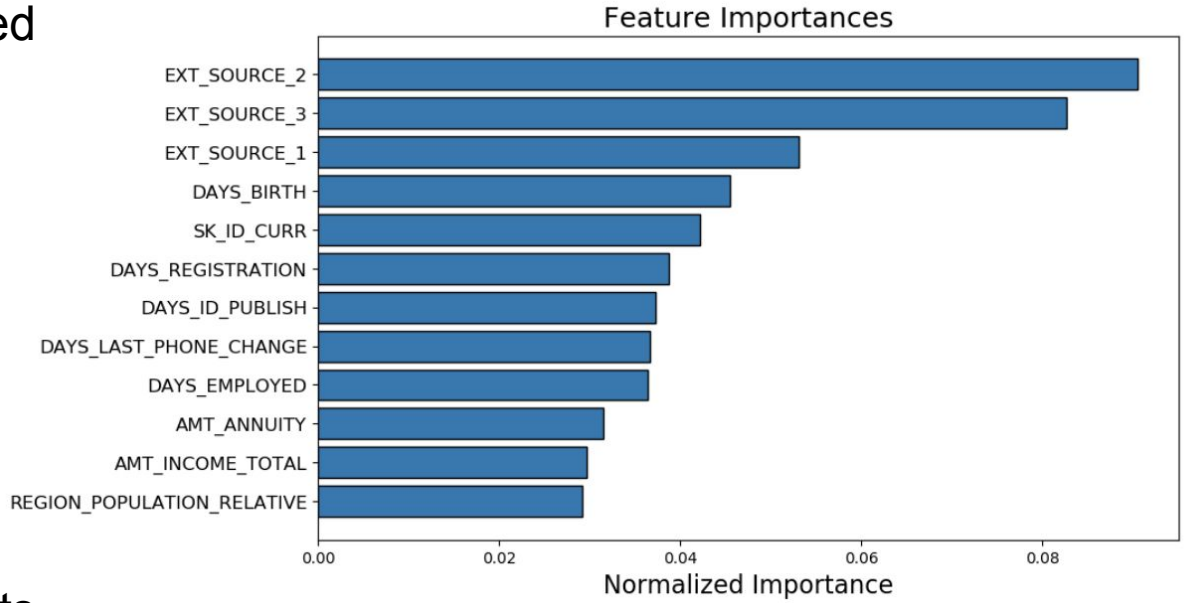


# max_depth


Sex_male ≤ 0.5
gini = 0.4652
samples = 891
value = [563, 328]
class = Survived
True
False
Fare ≤ 48.2
gini = 0.3797
samples = 314
value = [80, 234]
class = Dead
Age ≤ 13.0
gini = 0.2727
samples = 577
value = [483, 94]
class = Survived
Fare ≤ 27.825
gini = 0.4444
samples = 225
value = [75, 150]
class = Dead
Age ≤ 8.0
gini = 0.106
samples = 89
value = [5, 84]
class = Dead
Fare ≤ 26.95
gini = 0.4909
samples = 37
value = [16, 21]
class = Dead
Fare ≤ 26.2687
gini = 0.2338
samples = 540
value = [467, 73]
class = Survived
gini = 0.4245
samples = 193
value = [59, 134]
class = Dead
gini = 0.5
samples = 32
value = [16, 16]
class = Survived
gini = 0.0
samples = 1
value = [1, 0]
class = Survived
gini = 0.0868
samples = 88
value = [4, 84]
class = Dead
gini = 0.2311
samples = 15
value = [2, 13]
class = Dead
gini = 0.4628
samples = 22
value = [14, 8]
class = Survived
gini = 0.1302
samples = 400
value = [372, 28]
class = Survived
gini = 0.4362
samples = 140
value = [95, 45]
class = Survived

# Feature Importances

- Based on samples impacted
- They don’t indicate the direction of the impact
- Useful for EDA
- Linearises the non-linear tree




- Go beyond importances when interpreting the results of your model

# Pros / Cons

## Advantages

- Easy to Understand/Explain
- Useful in Data exploration
- Less data cleaning required

## Disadvantages

- Computational intensity of splitting on continuous variables
- Prone to overfit (but we know how to get around it)

# Delivering Value

Your job is not to create high performance models

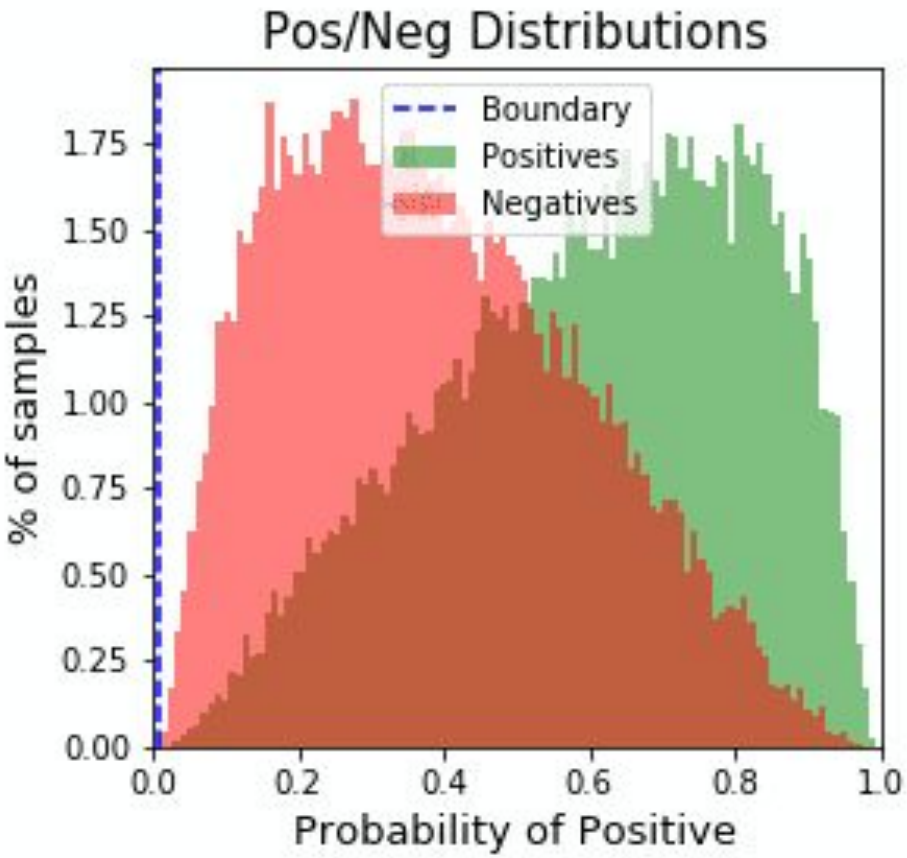
They pay you to **solve problems**

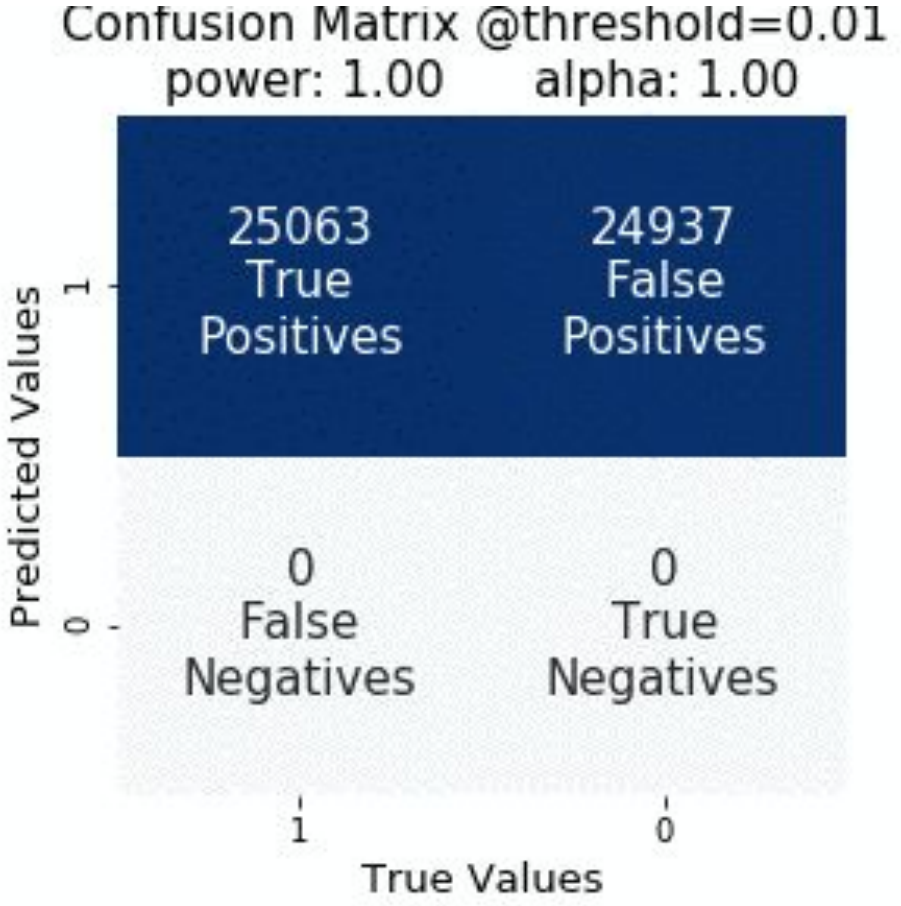
# Summary + Exit Ticket

Presented by Dan Sanz

FLATIRON SCHOOL

# Thresholds


Pos/Neg Distributions
Boundary
Positives
Negatives
% of samples
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0.0
0.2
0.4
0.6
0.8
1.0
Probability of Positive


Confusion Matrix @threshold=0.01
power: 1.00
alpha: 1.00
25063
True
Positives
24937
False
Positives
0
False
Negatives
0
True
Negatives
Predicted Values
1
0
1
0
True Values

# Thresholds to ROC curve