A Forest Full of Trees

Using Trees and Forests for Classification Tasks

Evan Simpson

p.evansimpson@gmail.com

EducatorsRLearners

Agenda

* Review of Machine Learning Types
* Decision Trees
* Random Forest

Types of Machine Learning

- Unsupervised?
- Supervised?
- Reinforcement?

Why?

SUPERVISED LEARNING

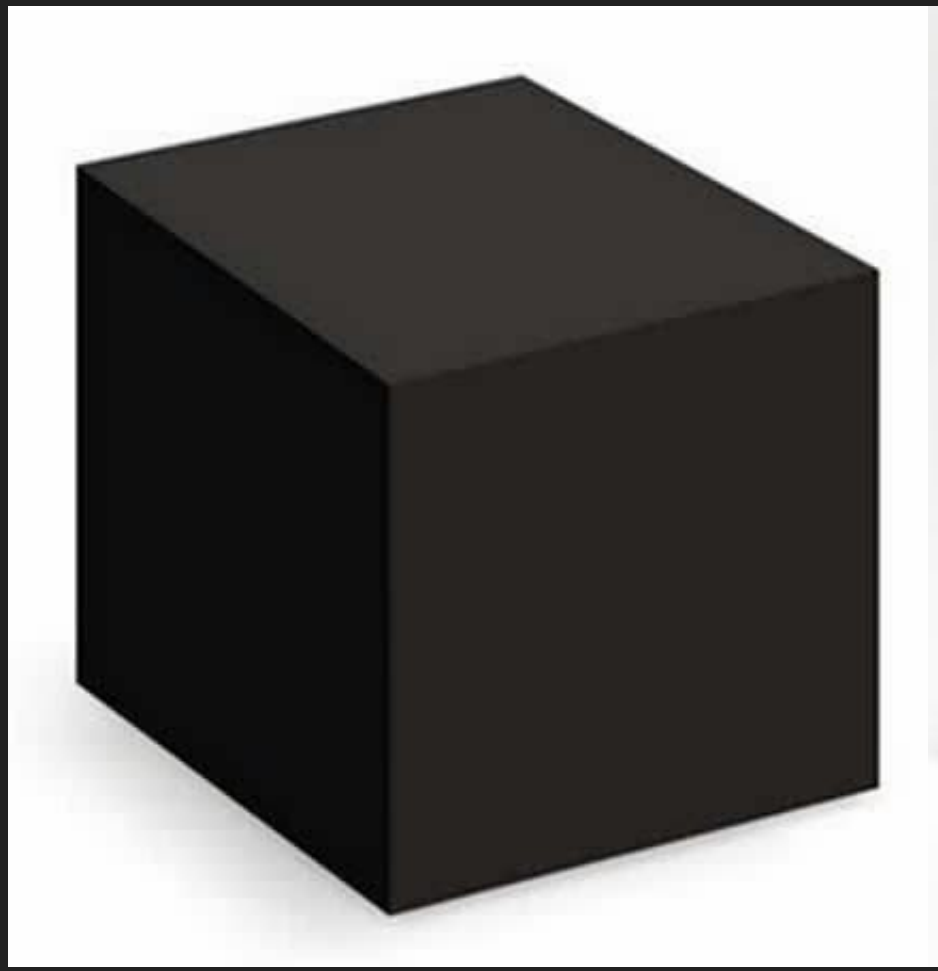
The sorting hat is provided with criteria which it uses categorize the students into one of four LABELED houses.

Hopefully, the sorting hat sees the outcome of its decisions. If it does, and uses those outcomes to improve its performance, then it is using REINFORCEMENT LEARNING.

But…..

100% Accurate?

Machine Learning Algorithms



Decision Trees

1. start with all observations in one group

2. identify a binary question, (i.e., "yes/no", "over/under", “present/absent”) which results in two groups that are as homogeneous as possible

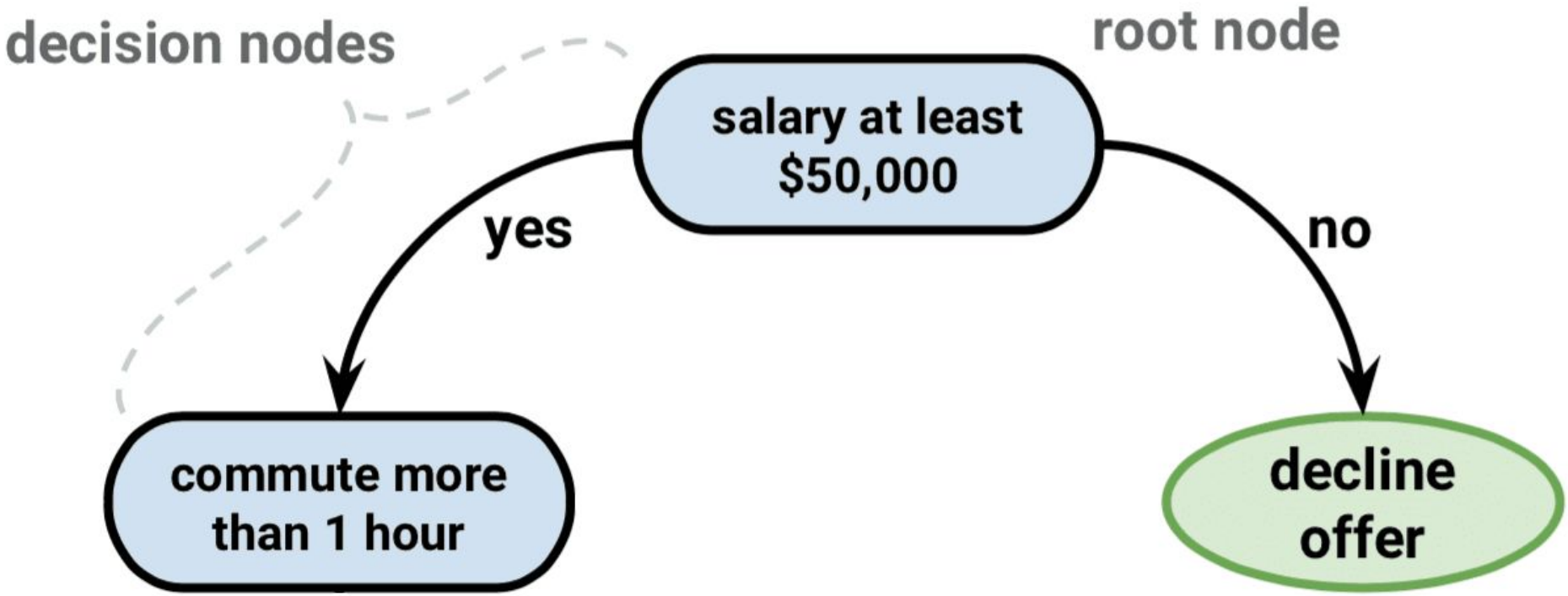
3. repeat step two until every subgroup is homogeneous or some other metric has been achieved

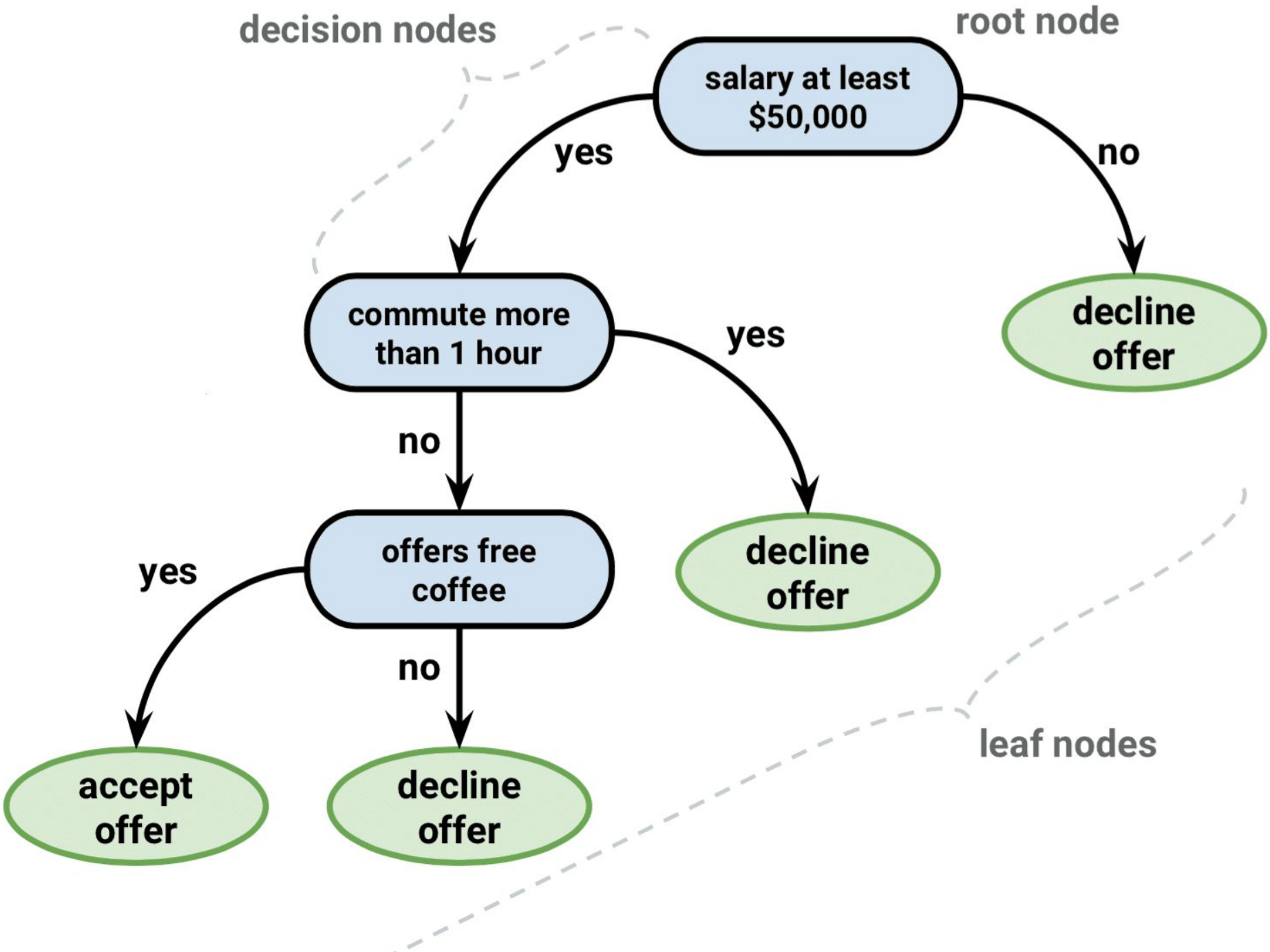
Decision Trees

Example: Job Offer

root node


root node
salary at least
$50,000
yes
no


decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
decline offer


decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
yes
decline offer
no
offers free coffee
decline offer
yes
no
accept offer
decline offer
leaf nodes

Task

Iris Setosa

Iris Versicolor

Iris Virginica

Sample Workflow

Frame the problem

Install and Load the Libraries

Collect/Load the Data

Conduct Exploratory Data Analysis (EDA)

Prepare the Data

Build and Evaluate the Model

Inspect the Prediction

Decision Trees

Advantage

-Easy to interpret and use

-Transparent

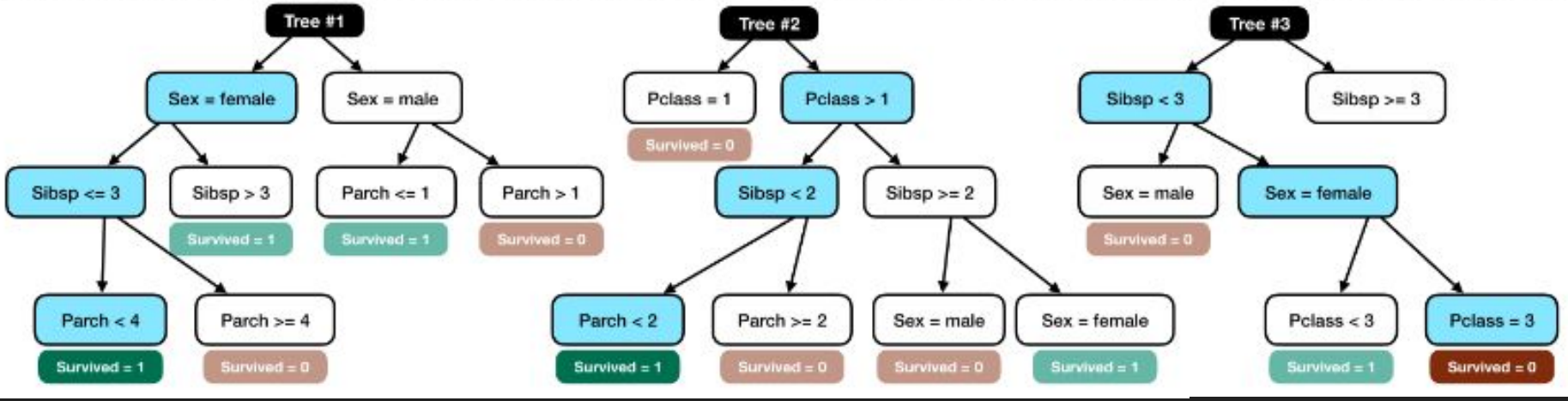
Disadvantage

- Non-linear Data

- Easy to overfit

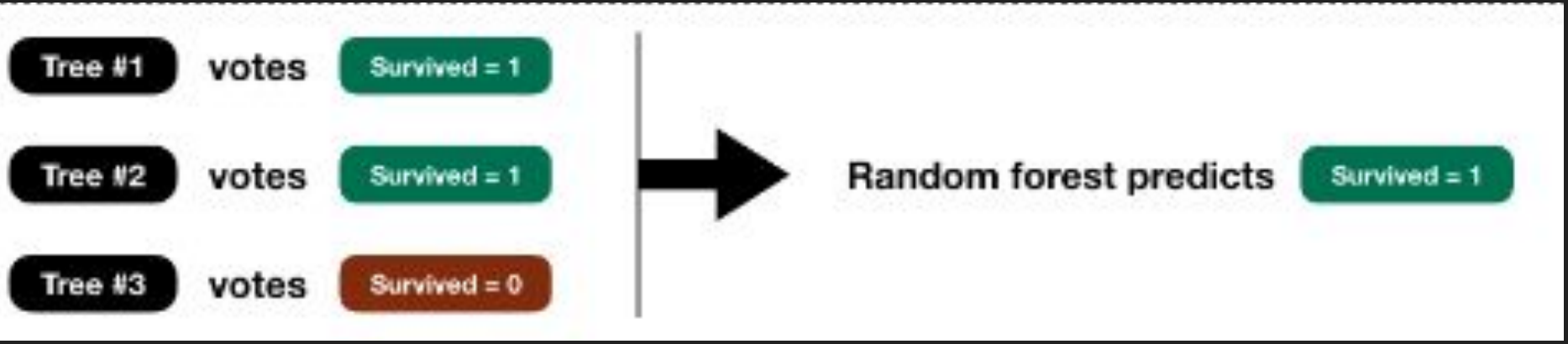
Ensemble Methods

Random Forest



Did the passenger survive?

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S



Task

Iris Setosa

Iris Versicolor

Iris Virginica

Random Forest

Advantage

- Incredibly Powerful

Disadvantage

- Black Box

Summary

Decision Trees

- Interpretability/Over-fitting

Random Forest

- Flexibility/Black Box

Resources

Kaggle Tutorial using Random Forest

Visualizing Decision Trees

DataCamp Decision Tree Tutorial

Hands On Machine Learning Chapters 6 and 7

Numsense! Data Science for the Layman