GENERAL ASSEMBLY

DATA SCIENCE

10 WEEK PART TIME COURSE

Week 6 - Decision Trees

Before we get started

- **Download GraphViz and PyDot**
  - `pip install pydot`
  - `pip install pydot2`
  - `http://graphviz.org/Download..php`

Yes, really.

AGENDA 3

1. Why this is the most awesome and interesting lesson
2. What are decision trees?
3. How decision trees work
4. Visual example on Titanic dataset
5. Lab
6. Talks
7. Discussion

DATA SCIENCE PART TIME COURSE

Why this lesson will change your life for the better*

[*] YMMV

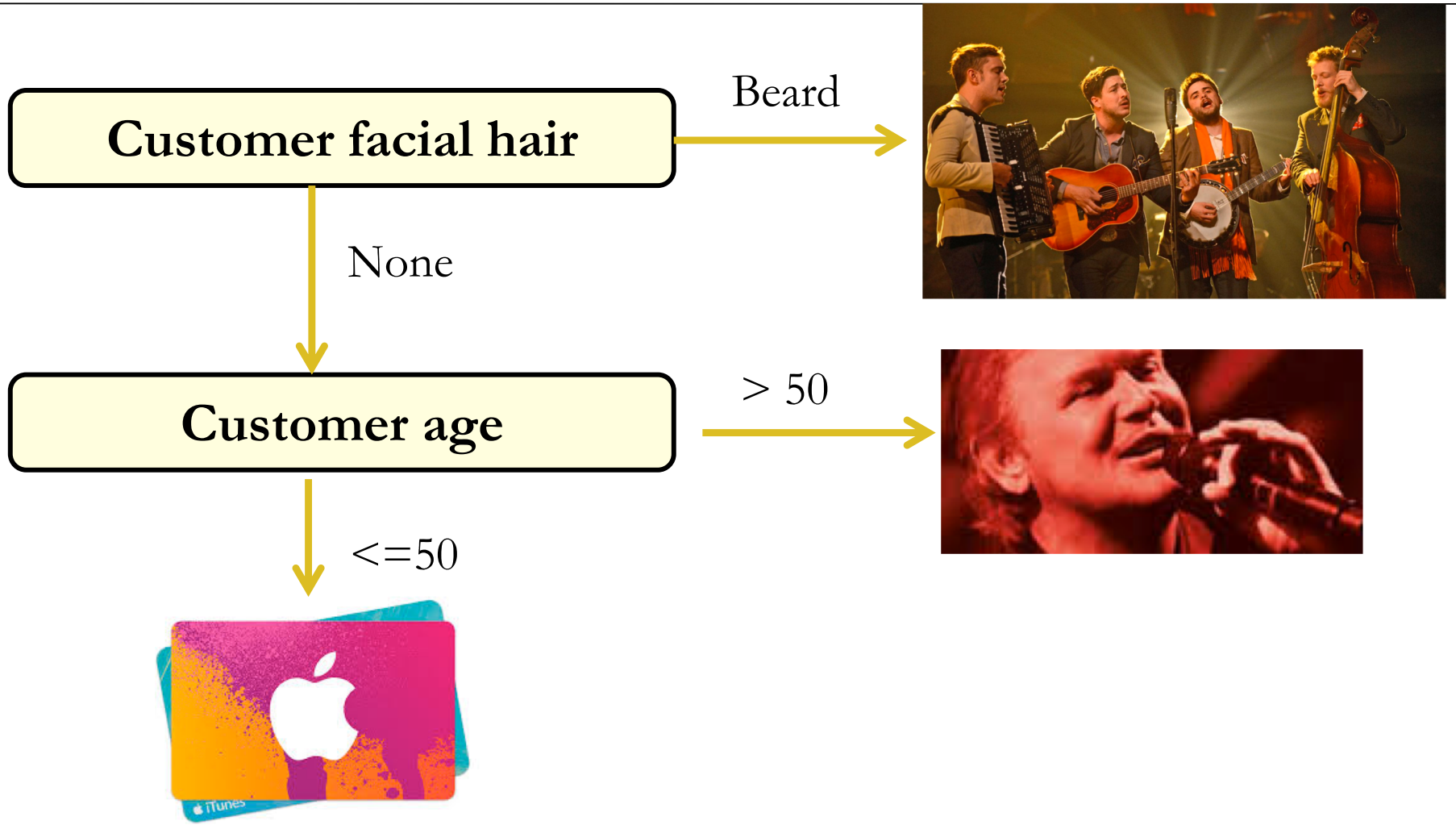
Decision trees

- Save lives
  - http://www.ncbi.nlm.nih.gov/pubmed/25160603
- Make data scientists look awesome
  - “If you follow this simple diagram, we make $$$$$”
- Can be used for some really smart stuff…

DATA SCIENCE PART TIME COURSE

DECISION TREES

Record store example



Yes, but how to make them?

- **Perfect decision trees are NP-complete**
  - (Computer scientist speak for "good luck with that")
- **Some techniques exist for creating quite good ones**
- **Danger of over-fitting**
  - 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

DATA SCIENCE PART TIME COURSE

Interactive Immersive Experience

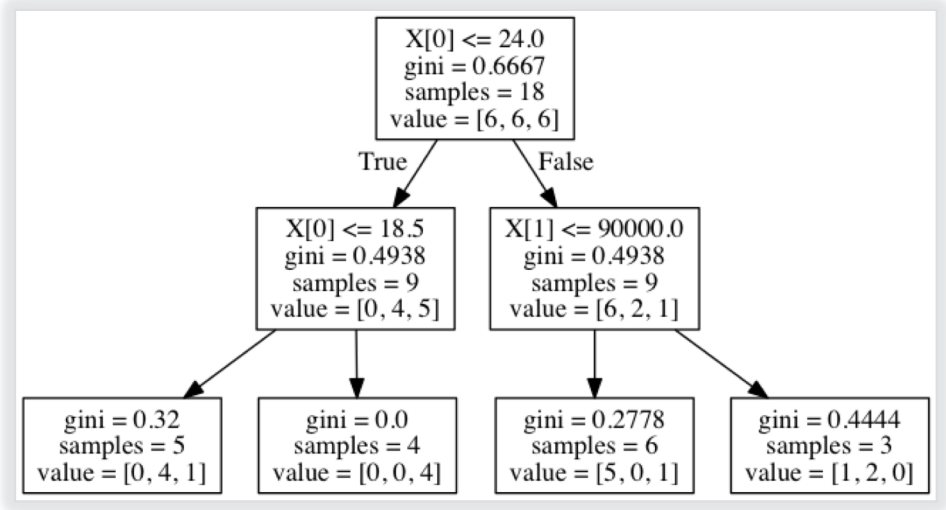
Phone case purchase data cards
Blu-tak

Using sklearn's Decision Tree API

```
import sklearn.tree
DF = …
Trying_to_predict = DF[['target_column']]
Source_data = DF[['column1','column2',…]]
dtc = sklearn.tree.DecisionTreeClassifier(criterion='gini',
                                          max_depth=1)

dtc.fit(Source_data,Trying_to_predict)


Predictions = dtc.predict(Source_data)
```

Making a quick diagram

```
try:
 from StringIO import StringIO
except ImportError:
 from io import StringIO

import sklearn.tree
import IPython.display
import pydot
File_obj = StringIO()
sklearn.tree.export_graphviz(dtc, out_file=File_obj)
Graph = pydot.graph_from_dot_data(File_obj.getvalue())
IPython.display.Image(Graph[0].create_png())
```



DATA SCIENCE PART TIME COURSE

Interactive Immersive Experience

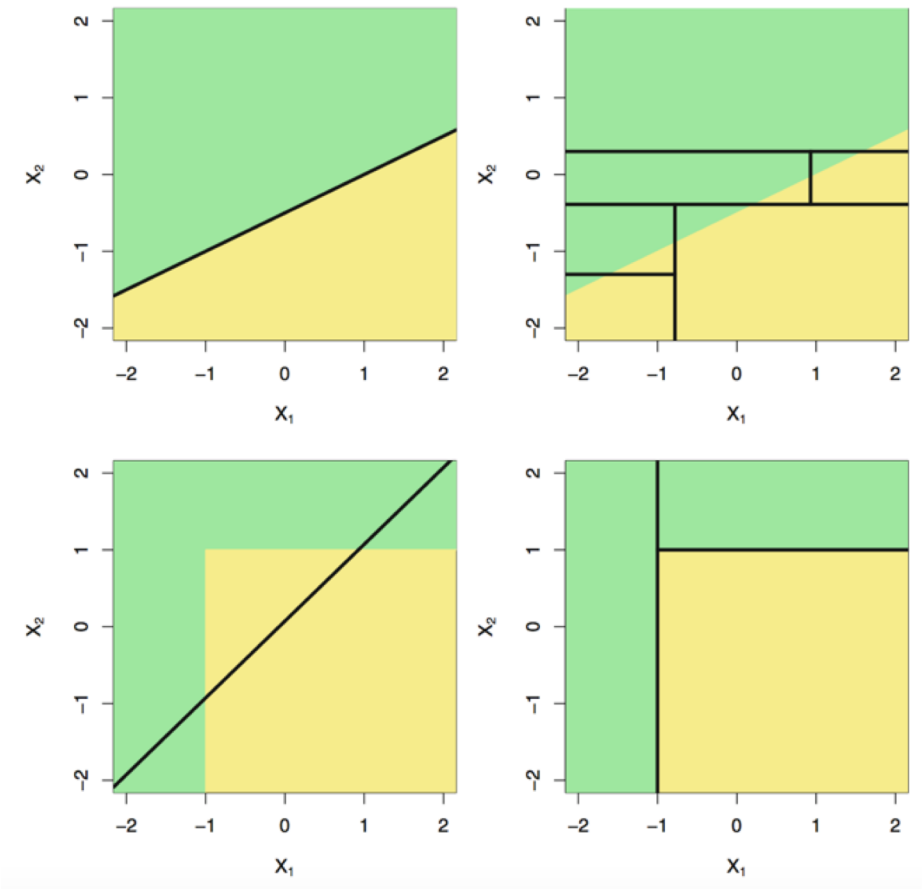
Phone case purchase data cards
Blu-tak

Repeat until satisfied

- Increase max_depth
- Try ‘entropy’ instead of ‘gini’ (matters less than 2% of the time)

Gini	Entropy (Information Gain)
Usually finds the largest class	Finds groups of classes that make ~50%
Minimise misclassification	Exploratory analysis
Continuous attributes	Attributes are classes

DECISION TREES - HOW IT WORKS



Linear decision boundary

Non-linear decision boundary

Simple ways to measure what it did

- **Gini importance**
  - **Which parameters mattered?**
  - `dtc.feature_importances_`

DECISION TREES - SHORTCOMINGS

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

DATA SCIENCE PART TIME COURSE

LAB