GENERAL ASSEMBLY

DECISION TREES AND RANDOM FORESTS

Sri Kanajan

LEARNING OBJECTIVES

- Understand and build decision tree models for classification
- Understand and build random forest models for classification
- Know how to extract the most important predictors in a random forest model

COURSE

PRE-WORK

PRE-WORK REVIEW

- Use Seaborn to create plots
- Explain the concepts of cross-validation, logistic regression, and overfitting
- Know how to build and evaluate *some* classification model in sckit-learn using cross-validation and AUC

OPENING

DECISION TREES AND RANDOM FORESTS

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

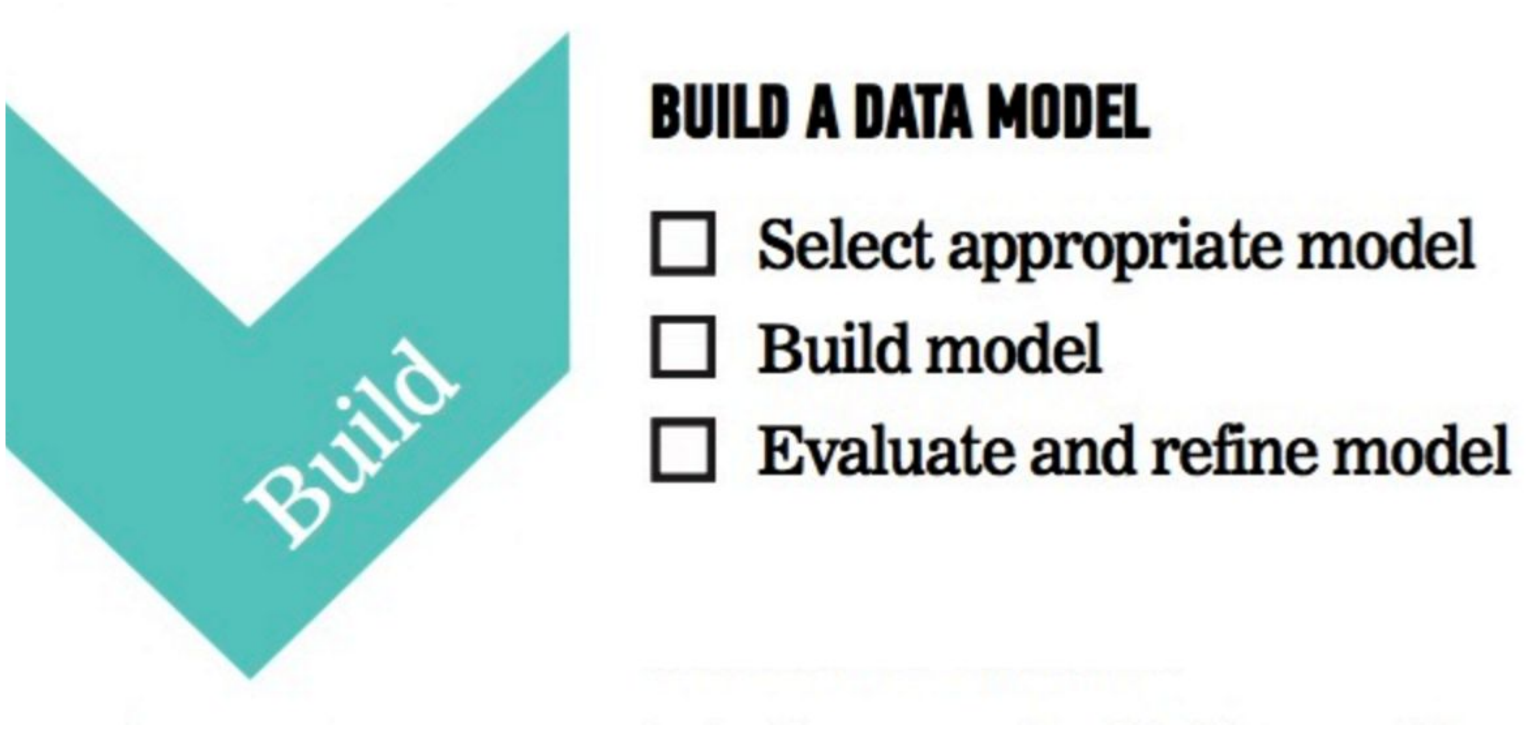
1. Define the difference between the precision and recall of a model.
2. What does the coefficients in Logistic Regression represent?

DELIVERABLE

Answers to the above questions

OVERVIEW OF THE DATA SCIENCE WORKFLOW

- In this lesson, we will focus on mining the dataset and building a model. We will focus on refining our model for the best predictive ability.



BUILD A DATA MODEL

□ Select appropriate model

□ Build model

□ Evaluate and refine model

GUIDED PRACTICE

EXPLORE THE DATASET

ACTIVITY: EXPLORE THE DATASET

EXERCISE

DIRECTIONS (25 minutes)

We will be using a dataset from StumbleUpon, a service that recommends webpages to users based upon their interests. They like to recommend “evergreen” sites, ones that are always relevant. This usually means websites that avoid topical content and focus on recipes, how-to guides, art projects, etc. We want to determine important characteristics for “evergreen” websites. Follow these prompts to get started:

1. Break into groups.
2. Prior to looking at the data, brainstorm 3-5 characteristics that would be useful for predicting evergreen websites.
3. After looking at the dataset, can you model or quantify any of the characteristics you wanted? See the Notebook for data dictionary and starter code.
4. Does being a news site affect evergreeness? Compute or plot the percent of evergreen news sites.

ACTIVITY: EXPLORE THE DATASET

EXERCISE

DIRECTIONS (25 minutes)

5. In general, does category affect evergreeness? Plot the rate of evergreen sites for all Alchemy categories.
6. How many articles are there per category?
7. Create a feature for the title containing “recipe”. Is the percentage of evergreen websites higher or lower on pages that have “recipe” in the title?

Check: Were you able to plot the requested features? Can you explain how you would approach this type of dataset?

DELIVERABLE

Requested features and answers to questions

INTRODUCTION

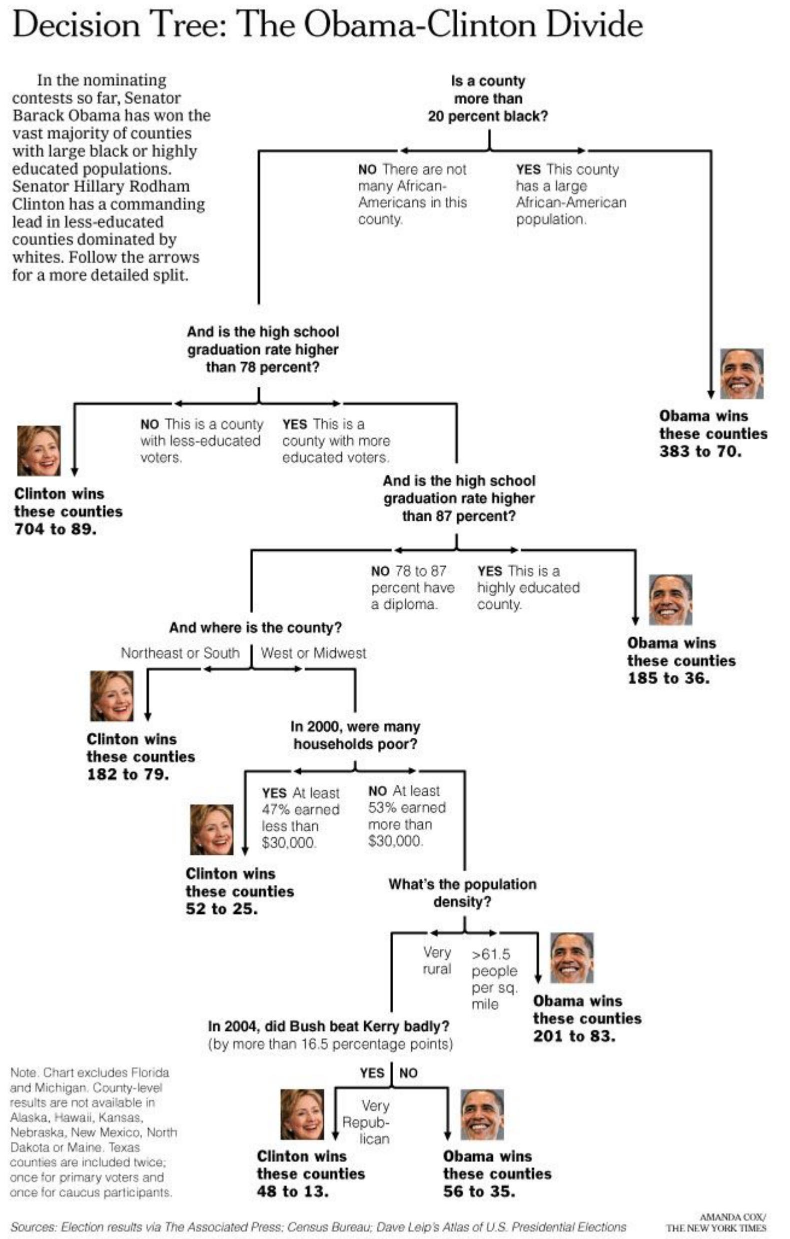
TRAINING DECISION TREES

INTUITION BEHIND DECISION TREES

- Decision trees are like the game “20 questions”. They make decision by answering a series of questions, most often binary questions (yes or no).

- We want the smallest set of questions to get to the right answer.

- Each questions should reduce the search space as much as possible.



Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

Is a county more than 20 percent black?
NO There are not many African-Americans in this county.
YES This county has a large African-American population.
Obama wins these counties 383 to 70.
And is the high school graduation rate higher than 78 percent?
NO This is a county with less-educated voters.
YES This is a county with more educated voters.
Clinton wins these counties 704 to 89.
And is the high school graduation rate higher than 87 percent?
NO 78 to 87 percent have a diploma.
YES This is a highly educated county.
Obama wins these counties 185 to 36.
And where is the county?
Northeast or South
West or Midwest
Clinton wins these counties 182 to 79.
In 2000, were many households poor?
YES At least 47% earned less than $30,000.
NO At least 53% earned more than $30,000.
Clinton wins these counties 52 to 25.
What’s the population density?
Very rural
>61.5 people per sq. mile
Obama wins these counties 201 to 83.
In 2004, did Bush beat Kerry badly? (by more than 16.5 percentage points)
YES
NO
Very Repub-lican
Clinton wins these counties 48 to 13.
Obama wins these counties 56 to 35.

Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

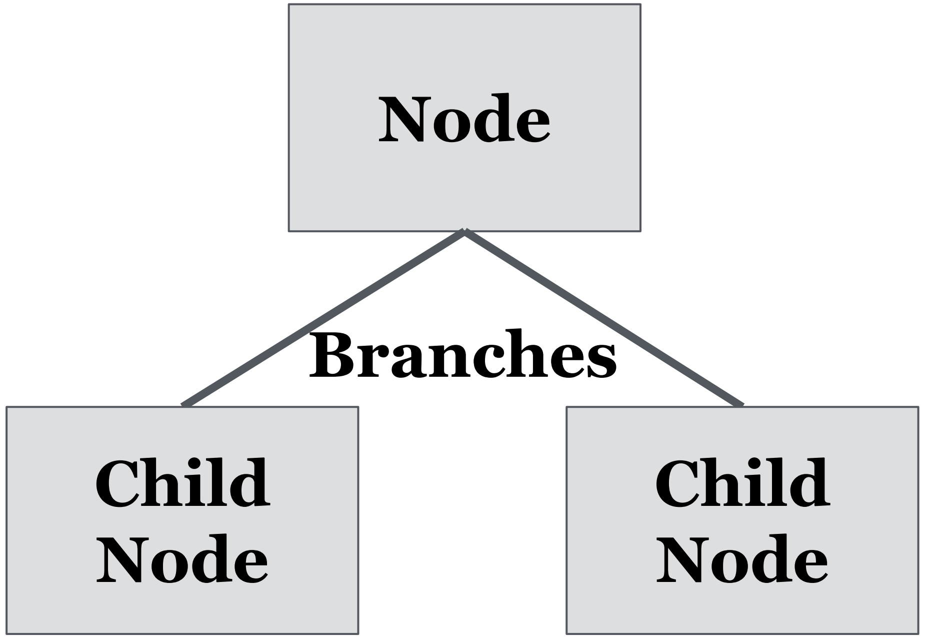
Sources: Election results via The Associated Press; Census Bureau; Dave Leip’s Atlas of U.S. Presidential Elections

AMANDA COX/
THE NEW YORK TIMES

TREES

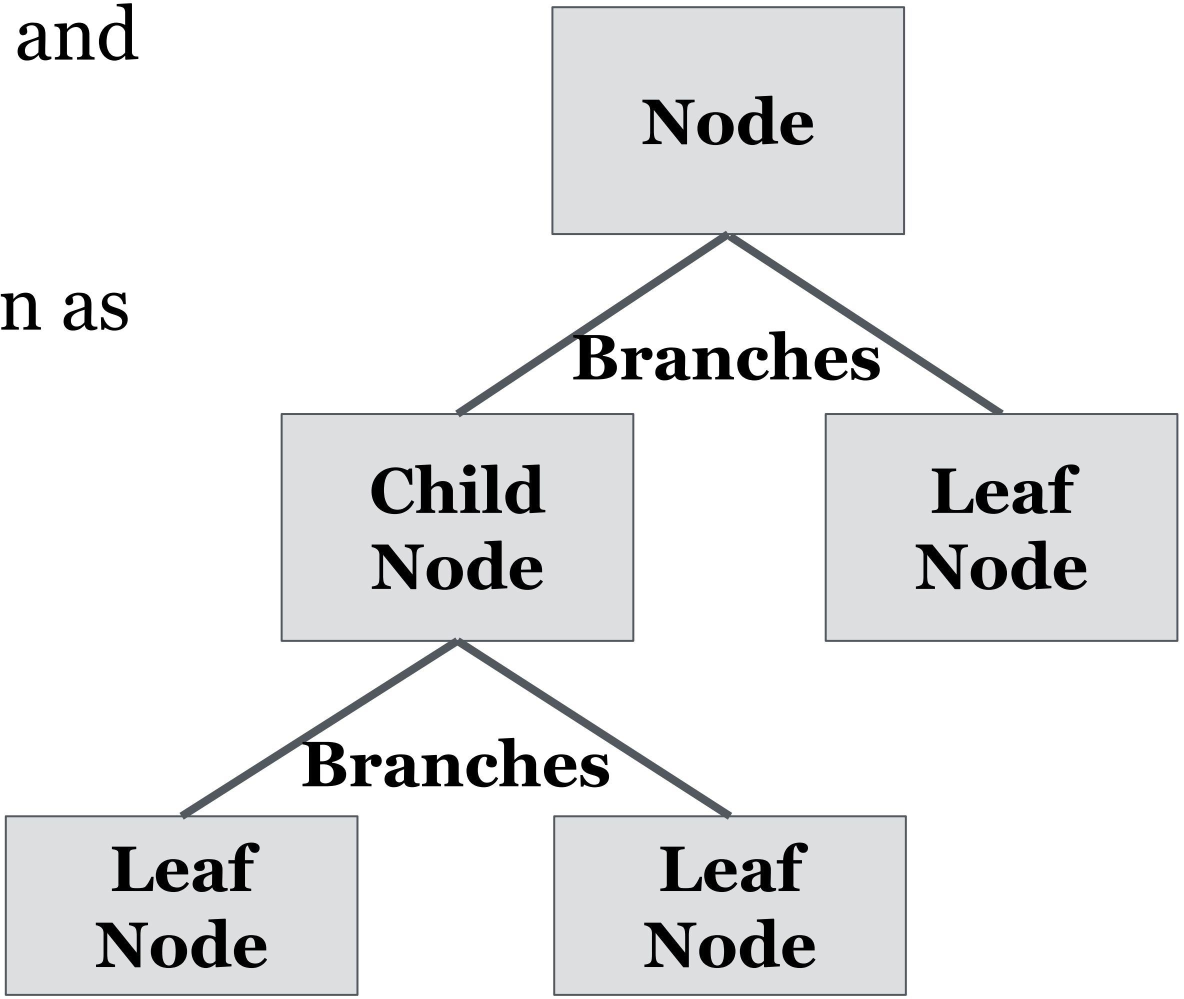
- Trees are a data structure made up of *nodes* and *branches*.

- Each node typically has two or more branches that connect it to its children.



TREES

- Each child is another node in the tree and contains its own *subtree*.
- Nodes without any children are known as *leaf* nodes.



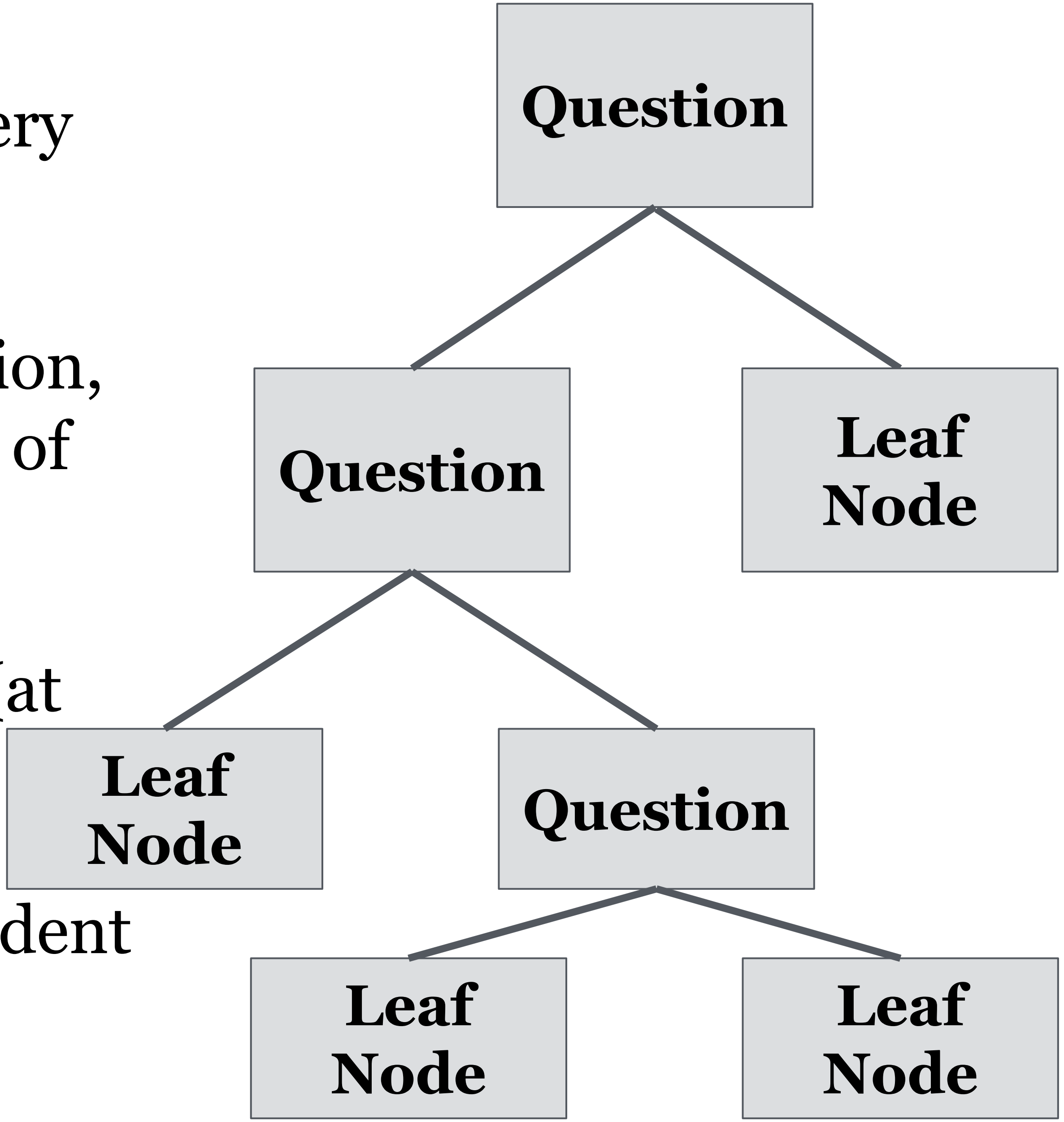
DECISION TREES

- A *decision tree* contains a question at every node.

- Depending upon the answer to the question, we proceed down the left or right branch of the tree and ask another question.

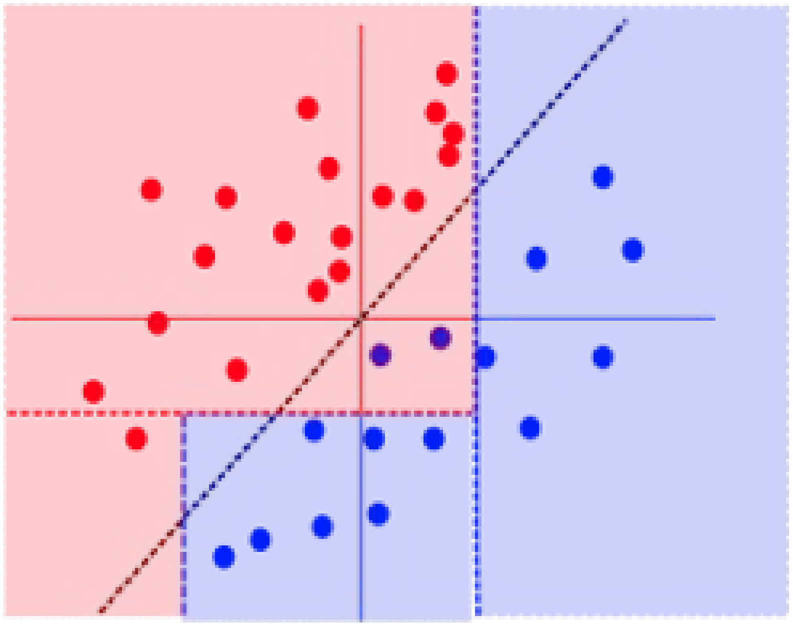
- Once we don't have any more questions (at the *leaf* nodes), we make a prediction.

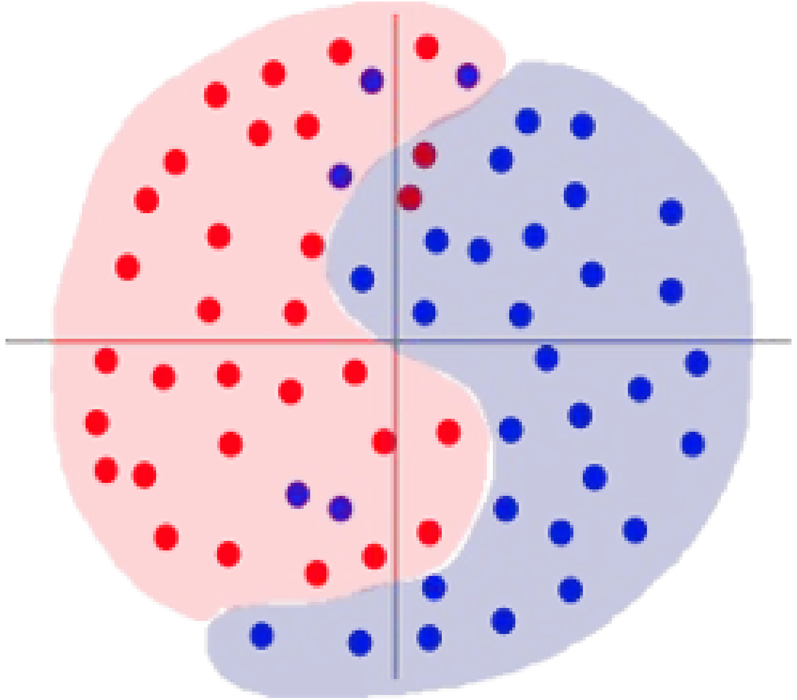
- Note: The next question is always dependent on the last.



COMPARISON TO PREVIOUS MODELS

- Decision trees are *non-linear*, an advantage over logistic regression.
- A *linear* model is one in which a change in an input variable has a constant change on the output variable.





TRAINING A DECISION TREE MODEL

- Training a decision model is deciding the best set of questions to ask.

- A good question will be one that best segregates the positive group from the negative group and then narrows in on the correct answer.

- For example, in our evergreen article decision tree, the best question is one that creates two groups, one that is mostly evergreen websites and one that is mostly non-evergreen websites.

TRAINING A DECISION TREE MODEL

- We can quantify the *purity* of the separation of groups using Classification Error, Entropy, or Gini Coefficient.

- We want to choose the question that gives us the best *change* in our purity measure. At each step, we can ask, “Given our current set of data points, which question will make the largest change in purity?”

- This is done *recursively* for each new set of two groups until we reach a stopping point.

TRAINING A DECISION TREE MODEL

- Algorithm used to generate the tree:
  - Evaluate every threshold within each feature relative to a “purity” metric and find the feature,threshold that provides the greatest increase in “purity”
  - Do this until an exit criteria such as depth of tree or purity of leaves is met
- Purity metrics
  - Entropy

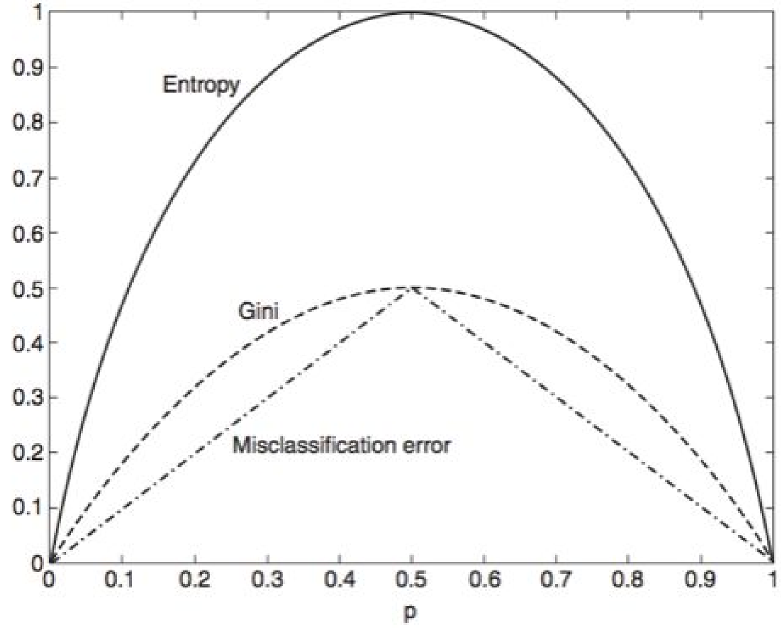
$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i)$$

  - Gini

$$\mathrm{Gini} = \sum_{i=1}^{classes} p_i(1-p_i) = 1-\sum_{i=1}^{classes} p(i \mid t)^2$$

- Calculate info gain

$$\Delta = I(\mathrm{parent}) - \sum_{\mathrm{children}} \frac{N_j}{N} I(\mathrm{child}_j)$$



MAKING PREDICTIONS FROM A DECISION TREE

- Predictions are made by answering each of the questions.

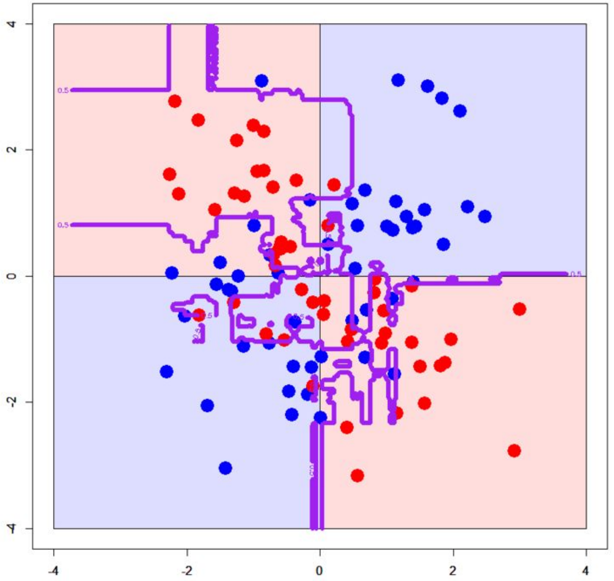
- Once we reach a leaf node, our prediction is made by taking the majority label of the training samples that fulfill the questions.

- In our sample tree, if we want to classify a new article, ask:
    - Does the article contain the word recipe?
    - If it doesn’t, does the article have a lot of images?
    - If it does, then 630 / 943 article are evergreen.
        - So we can assign a 0.67 probability for evergreen sites.

OVERFITTING IN DECISION TREES

- Decision trees tend to be weak models because they can easily memorize or overfit to a dataset.

- A model is *overfit* when it memorizes or bends to a few specific data points rather than picking up general trends in the data.



OVERFITTING IN DECISION TREES

- An unconstrained decision tree can learn an extreme tree (e.g. one feature for each word in a news article).

- We can limit our decision trees using a few methods.
  - Limiting the number of questions (nodes) a tree can have).
  - Limiting the number of samples in the leaf nodes.

GUIDED PRACTICE

DECISION TREES IN SCIKIT-LEARN

ACTIVITY: DECISION TREES IN SCIKIT-LEARN

EXERCISE

DIRECTIONS (15 minutes)

1. In the starter code notebook, work through the exercises titled "Decision Trees in scikit-learn".
2. In your groups from earlier, work on evaluating the decision tree using cross-validation methods.
3. What metrics would work best? Why?

Check: Are you able to evaluate the decision tree model using cross-validation methods?

DELIVERABLE

Completed exercises and answer to #3

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS

Let's work as a class to accomplish the following:

1. Using our StumbleUpon dataset, try to predict whether a given article is evergreen.
2. Build a decision tree to determine the above.
3. Explore different hyperparameters in the decision tree model

DELIVERABLE

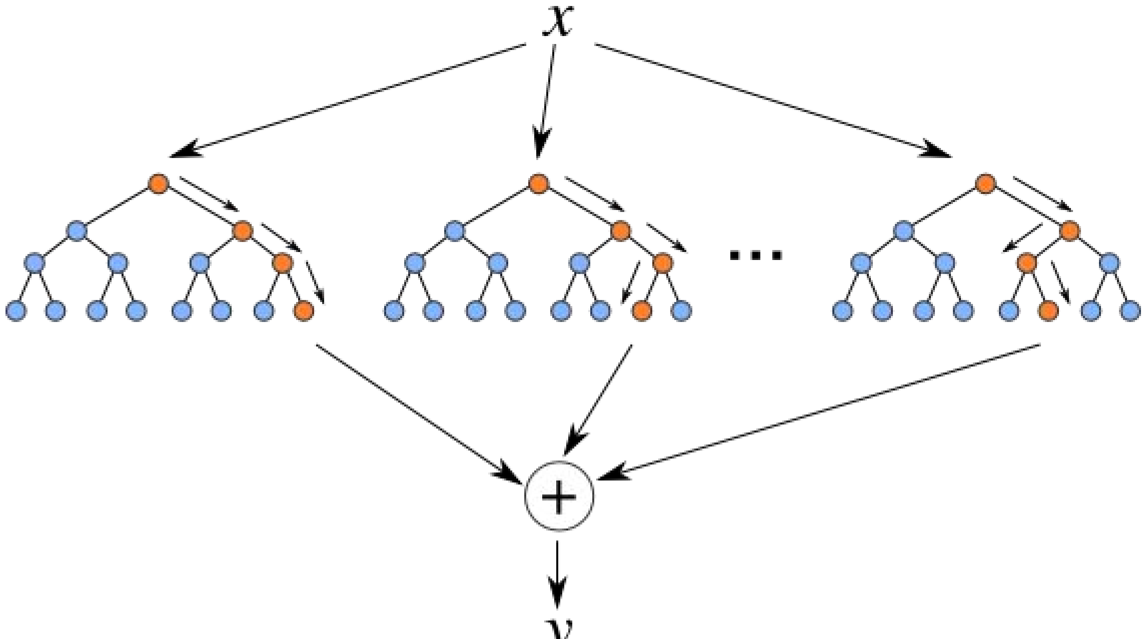
Our decision tree

INTRODUCTION

RUNNING THROUGH THE RANDOM FORESTS

RUNNING THROUGH THE RANDOM FORESTS

- Random forest models are one of the most widespread classifiers used.
- They are relatively simple to use and help avoid overfitting.
- Random Forests are an *ensemble* or collection of individual decision trees.



PROS AND CONS OF RANDOM FORESTS

- Advantages
  - Easy to tune
  - Built-in protection against overfitting
  - Non-linear
  - Built-in interaction effects

- Disadvantages
  - Slow
  - Black-box
  - No “coefficients”
  - Harder to explain

TRAINING A RANDOM FOREST

- Training a random forest model involves training many decision tree models.

- Since decision trees overfit easily, we use many decision trees together and randomize the way they are created.

TRAINING A RANDOM FOREST

- Random Forest Algorithm

  a. Take a bootstrap sample (sample with replacement) of the dataset.

  b. Train a decision tree on the bootstrap sample. For each split/feature selection, only evaluate a *limited* number of features to find the best one.

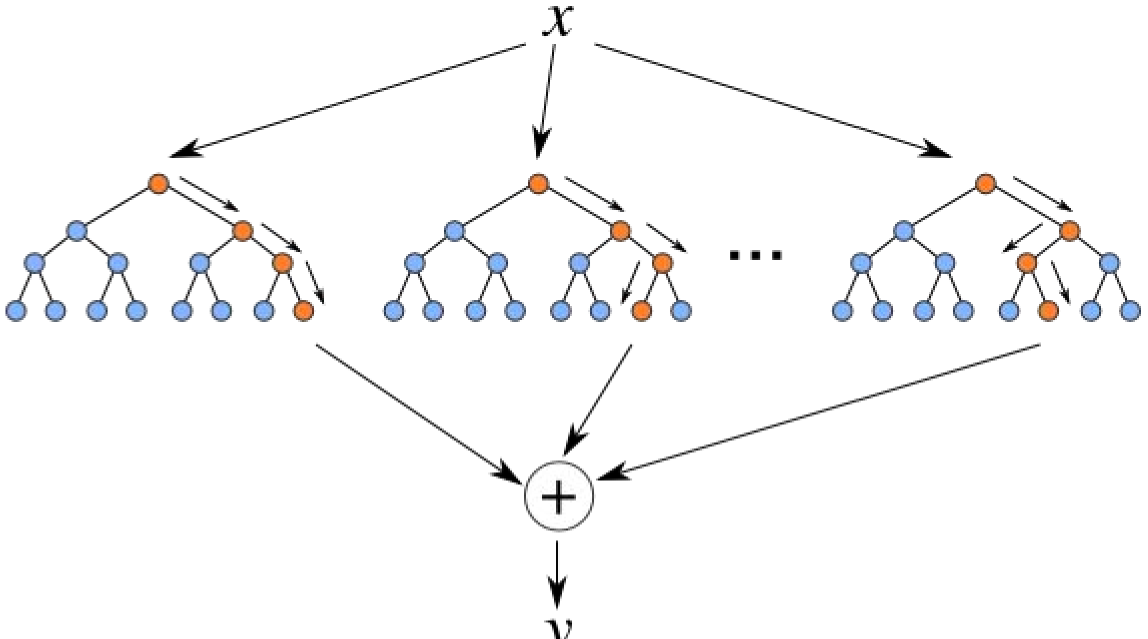
  c. Repeat this for *N* trees.

PREDICTIONS USING A RANDOM FOREST

- Predictions for a random forest model come from each decision tree.

- Make an individual prediction with each decision tree.

- Combine the individual predictions and take the majority vote.



INDEPENDENT PRACTICE

EVALUATE RANDOM FOREST USING CROSS-VALIDATION

ACTIVITY: RANDOM FORESTS

EXERCISE

DIRECTIONS (20 minutes)

1. Build a random forest model to predict the evergreeness of a website. Remember to use the parameter `n_estimators` to control the number of trees used in the model.
2. Take note of the most important features.

DELIVERABLE

The models mentioned above

ACTIVITY: EVALUATE RANDOM FOREST USING CROSS-VALIDATION

EXERCISE

DIRECTIONS (25 minutes)

1. Building upon the previous Guided Practice, add any input variables to the model that you think may be relevant.

2. For each feature:
   a. Evaluate the model for improved predictive performance using cross-validation.
   b. Evaluate the importance of the feature.

3. **Bonus**: Just like the ‘recipe’ feature, add in similar text features and evaluate their performance. Try to find the best possible model

DELIVERABLE

Newly created features and models

CONCLUSION

TOPIC REVIEW

REVIEW Q&A

- What are decision trees?
- What does training involve?
- What are some common problems with decision trees?
- What are random forests?
- What are some common problems with random forests?

COURSE

BEFORE NEXT CLASS

LESSON

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET