CSCAR CONSULTING FOR STATISTICS, COMPUTING & ANALYTICS RESEARCH
UNIVERSITY OF MICHIGAN

Supervised Machine Learning

Random Forests

Data Science Skills Series

Márcio Mourão

References

http://scikit-learn.org/

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

www.cs.kent.edu/~jin/DM07/ClassificationDecisionTree.ppt

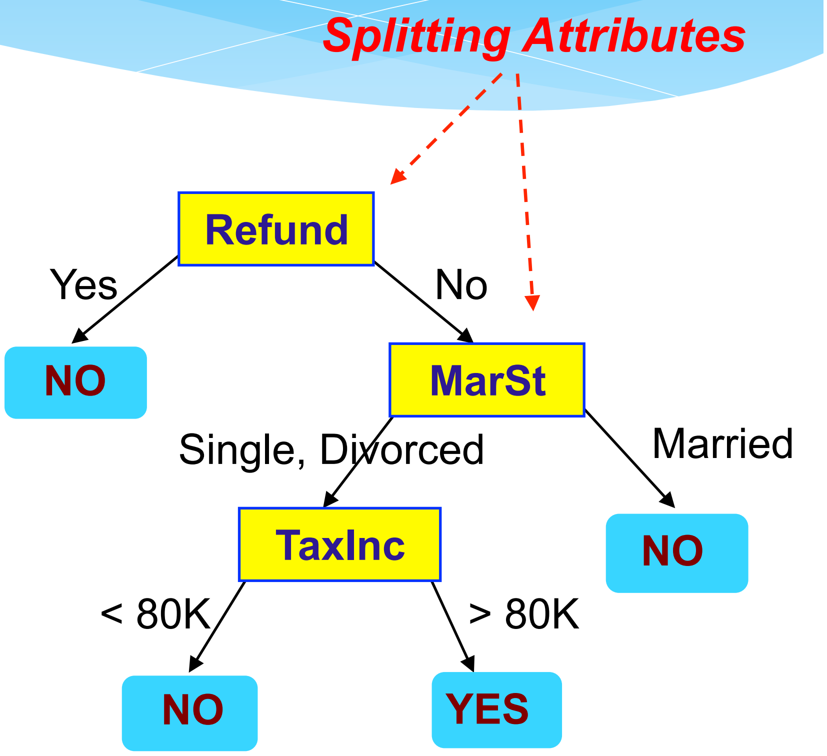
Decision trees

Create a decision tree from training data

categorical categorical continuous 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

Training Data

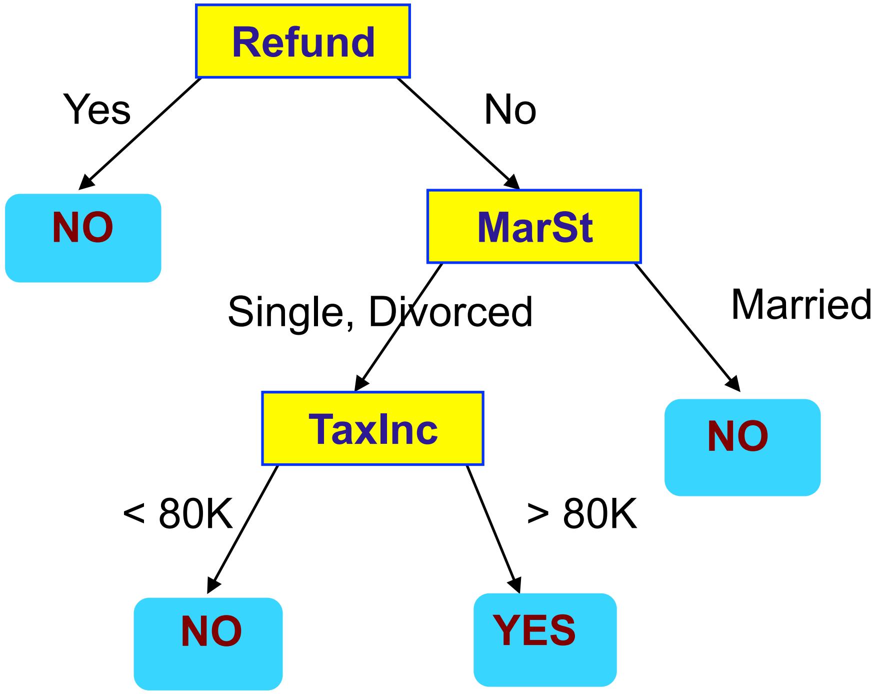


Model: Decision Tree

Apply model to test data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Decisions and algorithms

- Determine how to split
  - Specify attribute test condition
    - Attribute type: nominal, ordinal, continuous
    - Splitting: 2-way split, multi-way split

- Determine best split
  - Need a measurement of impurity: Gini Index, Entropy, Misclassification Error

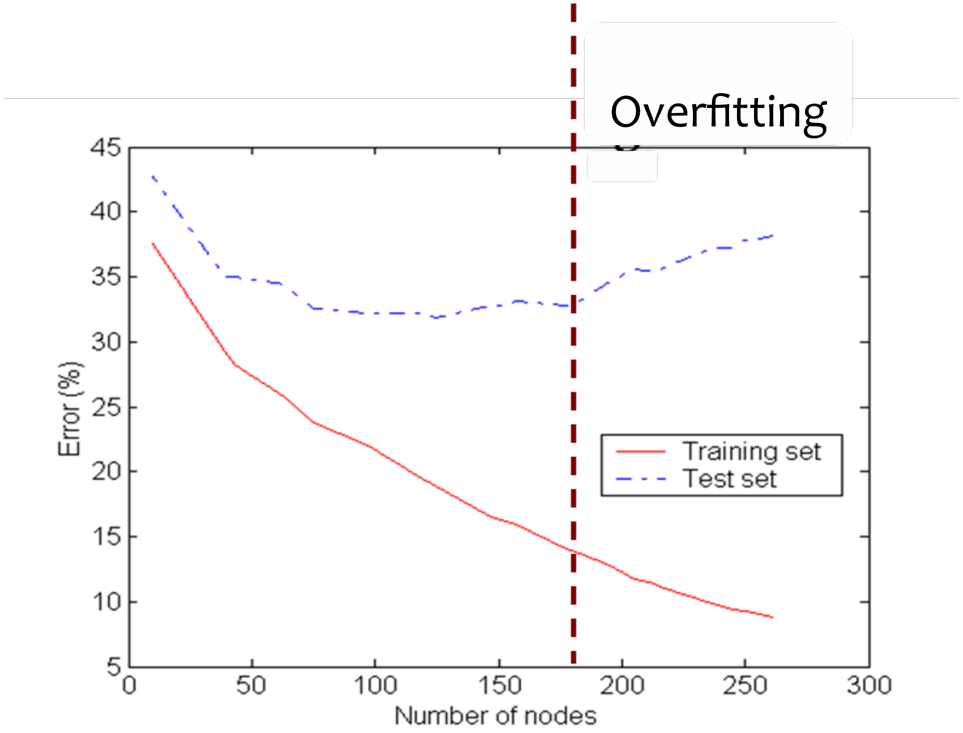
- There are many algorithms available
  - Hunt's, CART, ID3, C4.5, SLIQ, SPRINT

Advantages

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Easily handles mixed data
- Scalable
- Accuracy is comparable to other classification techniques for many simple data sets

Disadvantages

- Prone to produce decision trees that are more complex than necessary (overfitting)
- In overfitting, trees are highly sensitive to noise in the training set
- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records


Overfitting

Random Forests

Random Forests

- Random Forests correct decision's tree habit of overfitting

- Random Forests fall under a class of learners called "Ensemble Methods"

- For what problems can you apply this method?
  - Regression for Continuous Outcome Variables
  - Classification for Categorical Outcome Variables (2 or more levels)

Methodology

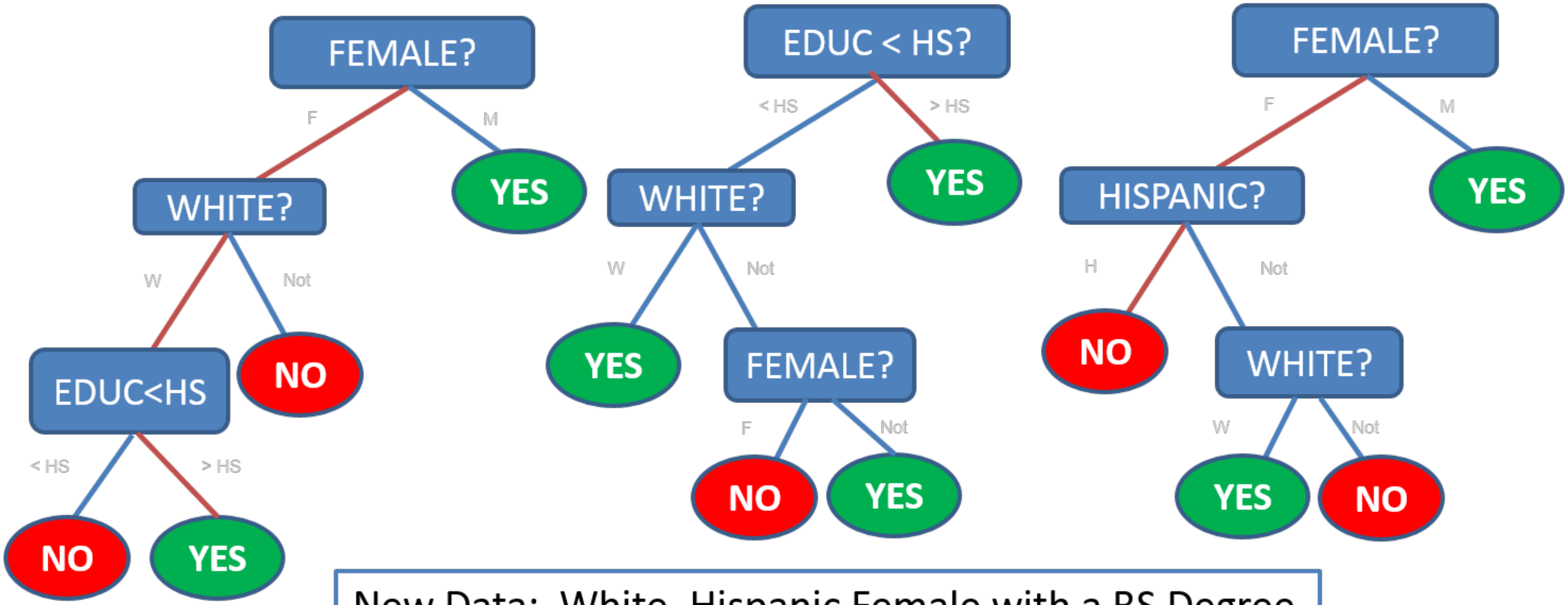
- **Step 1:** Establish the Forest (bagging)
  - Generate ntree bootstrap samples from the original data set, with replacement, and having the same size as the input data set.

- **Step 2:** Grow Trees
  - For each bootstrap sample from Step 1, grow a classification or regression tree to full size (with no pruning)
  - At each node, randomly select predictor variables to use for splitting
    - Splits will be based on the BEST variable from those randomly selected for a given node

Methodology

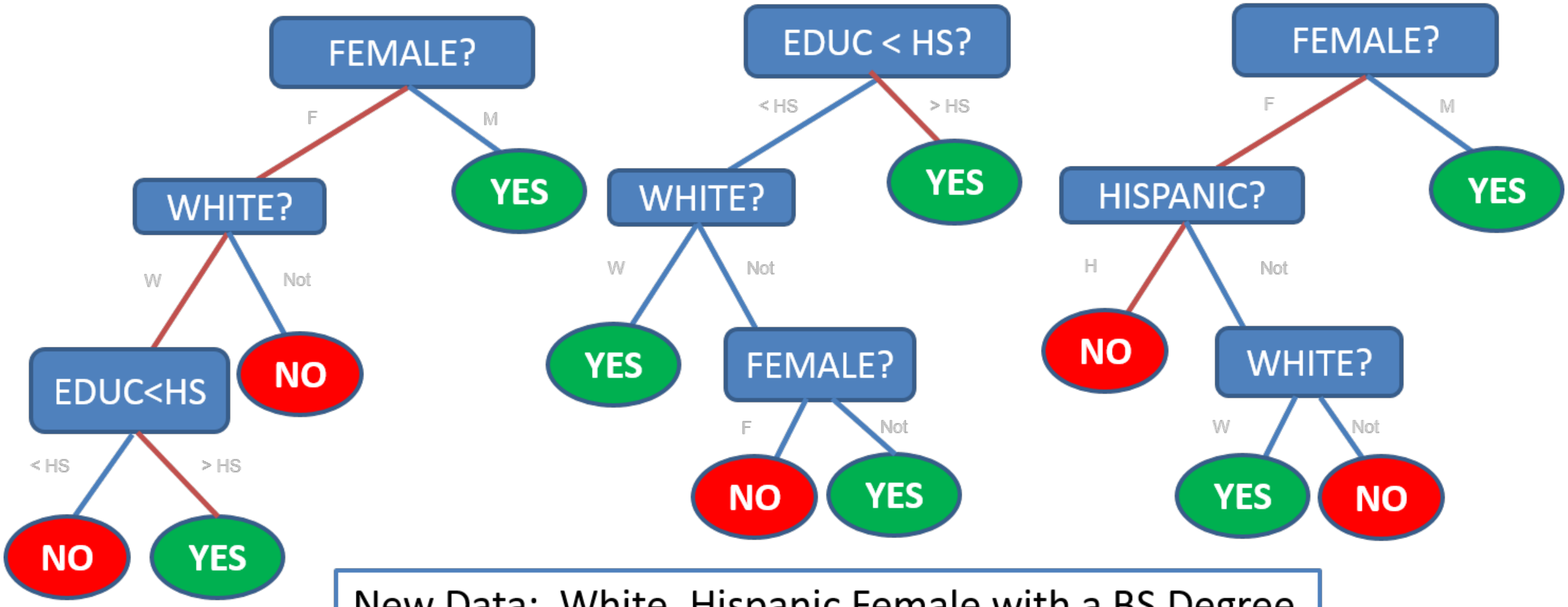
- **Step 3:** Prediction
  - After all trees in the forest have been grown, form a prediction for new data from each tree in the forest
  - Final predictions in classification problems are based on the majority
  - Final predictions in regression problems are means

Illustration



New Data: White, Hispanic Female with a BS Degree

Illustration



New Data: White, Hispanic Female with a BS Degree

Tree 1: YES

Tree 2: YES

Tree 3: NO

Performance: Error Rates

How are error rates assessed for Random Forests?

- Unlike other machine learning methods that require some type of external test sample to assess performance, Random Forests automatically generate an internally valid, and in many cases unbiased, assessment of error

- **Out of Bag Error Rate (OOB Error Rate)**

Performance: Error Rates

How is the OOB Error Rate Computed?

- **Step 1:** Each tree in the forest uses approximately 2/3 of the original data. For the remaining 1/3 of cases (called Out of Bag, OOB), use the given tree to make predictions. Call these “Out of Bag Predictions.”

- **Step 2:** For a given case, aggregate all “Out of Bag Predictions” across the roughly ntree/3 trees for which the case was OOB.

  - Means for regression applications
  - Majority vote for classification applications

Performance: Error Rates

- **Step 3:** Compare the aggregated estimate from Step 2 to the actual value of the outcome for each case in the data set
  - Correct/Incorrect for Classification
  - Residual for Regression (y_i – Step 2 Estimate).

The OOB Error is the:

- Proportion of “incorrect” values from Step 3 (classification)
- Average squared residual from Step 3 (regression)

Pros and Cons of Random Forests

Pros

- Can handle predictors that are continuous, categorical, skewed and sparse data
- Aptly suited for the "large p, small n" scenario
- Very effective for estimating outcomes that are a complex functions of predictors with many interactions or possibly non-linear functions of the parameters

Cons

- Random Forests can be **extremely computationally intensive**
- Unlike Trees, Random Forests are **not easily visualized**

Random Forests using Scikit-Learn in Python

```
#Import Library
from sklearn.ensemble import RandomForestClassifier

#Create a RandomForest classification object
model = RandomForestClassifier(n_estimators=10, criterion='gini’)

#Train the model using the training sets
model.fit(X, y)

#Predict Output
predicted= model.predict(x_test)
```