Decision Trees

- An Introduction

Overview

- Key Terminology and Structure of a Decision Tree
- Overview of Decision Tree Algorithms
- Information Gain
- Recursive Partitioning
- Pruning
- Advantages and Disadvantages of Decision Trees

Trees and Rules

Goal: Classify or predict an outcome based on a set of predictors

The output is a set of **rules**

Example:

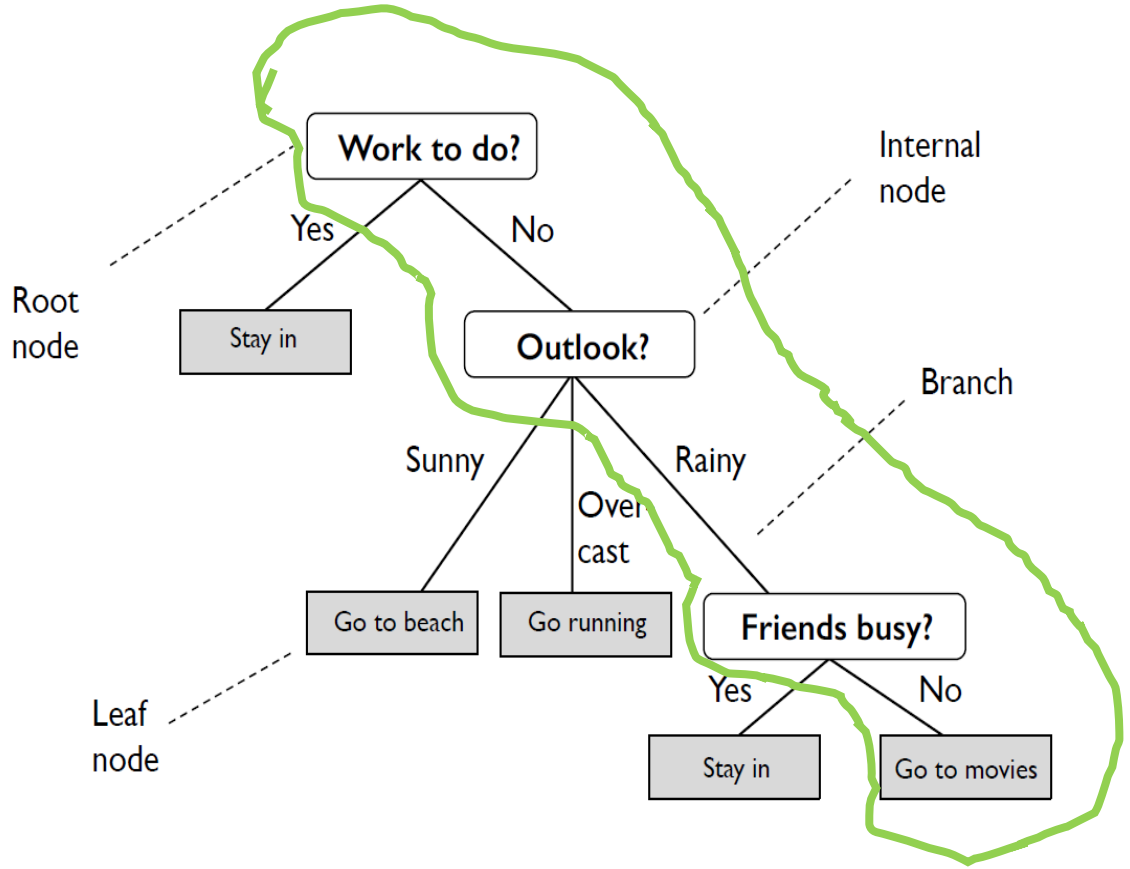
- Goal: Classify the response to a JOB offer as “Accept JOB offer” or “Reject JOB offer”, based on ‘Compensation Offered’ and ‘Location’
- Rule might be
  - “IF (**Offer** >= 10 Lakh) AND (**Location** = Bangalore)
    THEN **Accept Offer**
  - “If (**Offer** < 10 Lakh) AND (**Location** = Bangalore)
    THEN **Reject Offer**

  …..

- Such rules are best captured by tree diagrams

Key Terminology and Structure of a Decision Tree

- One **Root node**: no incoming edge, zero, or more outgoing edges.
- **Internal node**: one incoming edge, two (or more) outgoing edges.
- **Leaf node**: one incoming edge, no outgoing edge with each **leaf node** assigned a class label
- Each non-leaf node contains a test condition on one of the attributes



(Work to do? = No) ∩ (Outlook = Rainy?) ∩ (Friends busy? = No)
→ Go to movies

Source: http://pages.stat.wisc.edu/~sraschka/teaching/stat451-fs2020/

Decision Tree Algorithms

- ID3, C4.5, CART, CHAID, MARS, C5.0

- Algorithms differ in
  - Splitting criterion: information gain (Shannon Entropy, Gini impurity, misclassification error), use of statistical tests, objective function, etc.
  - Binary split vs. multi-way splits
  - Discrete vs. continuous variables
  - Pre vs. post-pruning

Information Gain

- The standard criterion that is being used for splitting in decision trees is information gain.
- We want to determine which predictor in a given set of predictors is the most useful for discriminating between the classes to be learned
- Information gain indicates the relative importance of a predictor amongst the given set of a predictors for splitting
  - Better the split, the higher the information gain.
- Information gain relies on the concept of mutual information: The reduction of the entropy of one variable by knowing the other.
  - Maximize mutual information when defining splitting criteria.

General Decision Tree Algorithm – Recursive Partitioning

- **Recursive partitioning:** Repeatedly split the records into two parts so as to achieve maximum homogeneity of outcome within each new part

- The process of growing a decision tree can be expressed as a recursive algorithm as follows:
  1. Pick a feature such that when parent node is split, it results in the largest information gain.
  2. Stop if child nodes are pure or no improvement in class purity can be made.
  3. Go back to step 1 for each of the two child nodes.

Example: Ownership of Lawn Mowers

- Goal: Classify 24 households as owning or not owning lawn mowers
- Predictors : Income, Lot Size
- Outcome/Response : Owner/Non-Owner

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

Riding Mowers Dataset – Ownership of Lawn Mowers

Income	Lot_Size	Ownership
60.0	18.4	owner
85.5	16.8	owner
64.8	21.6	owner
61.5	20.8	owner
87.0	23.6	owner
110.1	19.2	owner
108.0	17.6	owner
82.8	22.4	owner
69.0	20.0	owner
93.0	20.8	owner
51.0	22.0	owner
81.0	20.0	owner
75.0	19.6	non-owner
52.8	20.8	non-owner
64.8	17.2	non-owner
43.2	20.4	non-owner
84.0	17.6	non-owner
49.2	17.6	non-owner
59.4	16.0	non-owner
66.0	18.4	non-owner
47.4	16.4	non-owner
33.0	18.8	non-owner
51.0	14.0	non-owner
63.0	14.8	non-owner

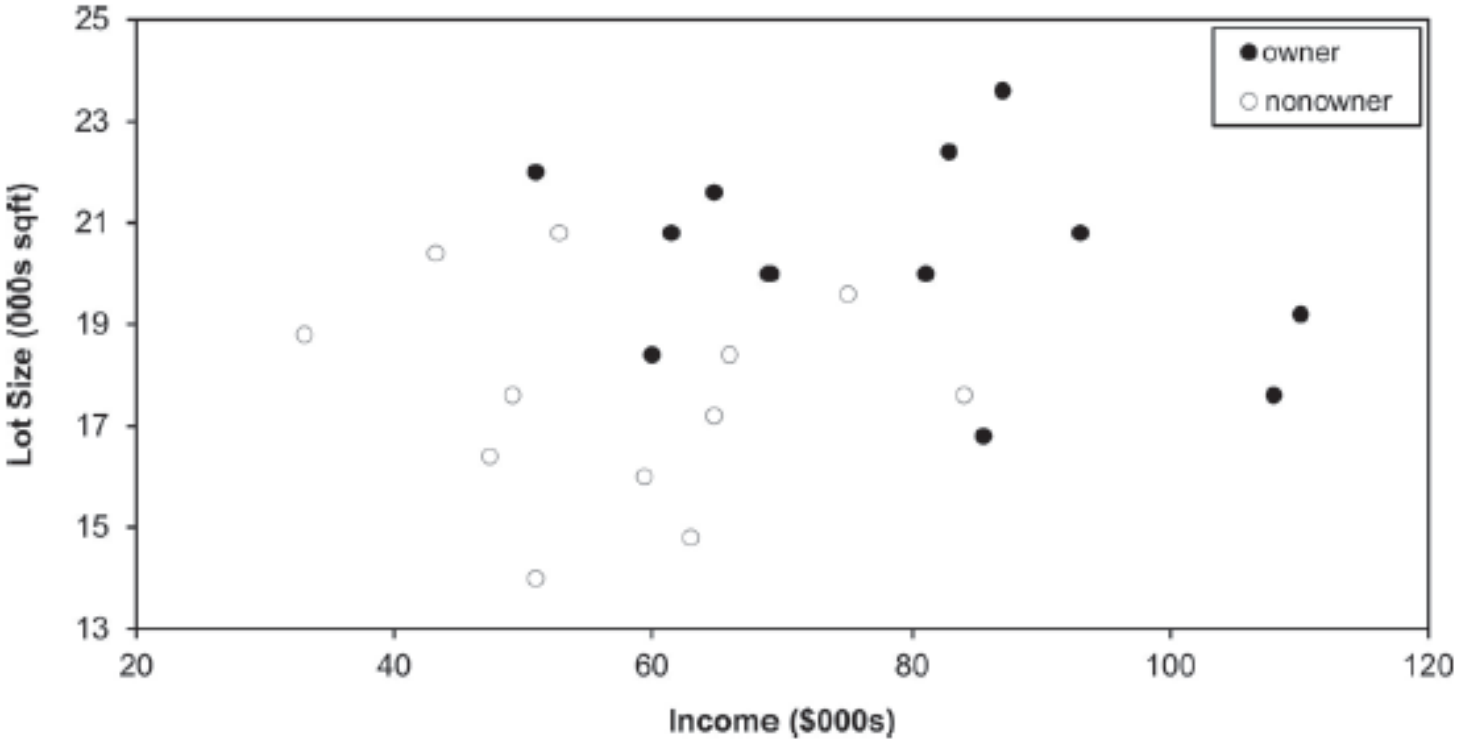


FIGURE 9.2 SCATTER PLOT OF LOT SIZE VS. INCOME FOR 24 OWNERS AND NONOWNERS OF RIDING MOWERS

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

After All Splits

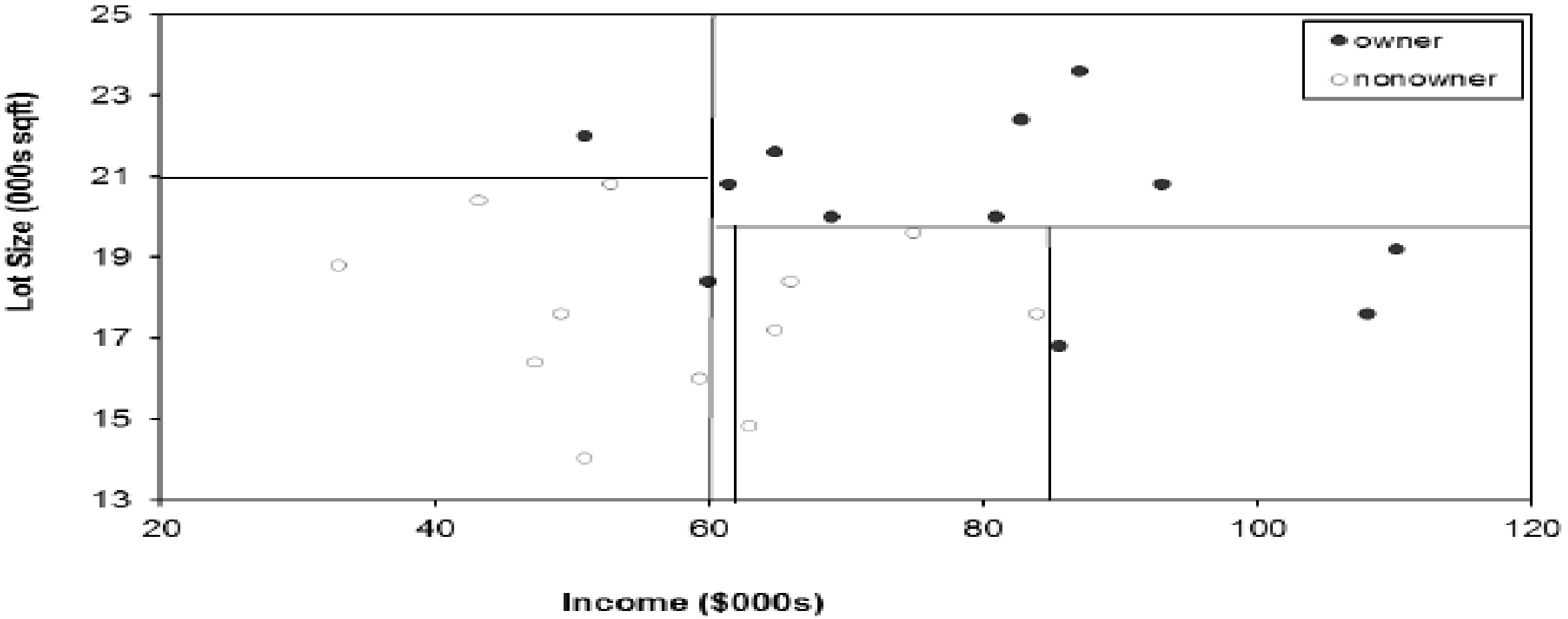
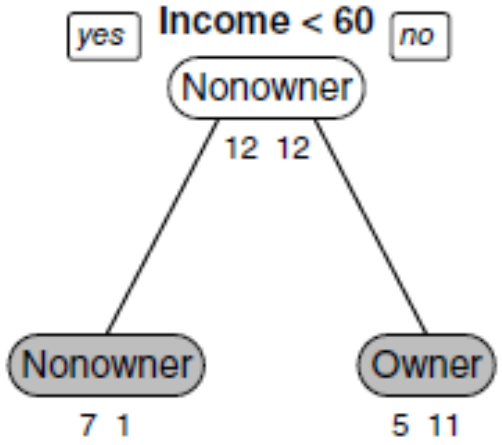


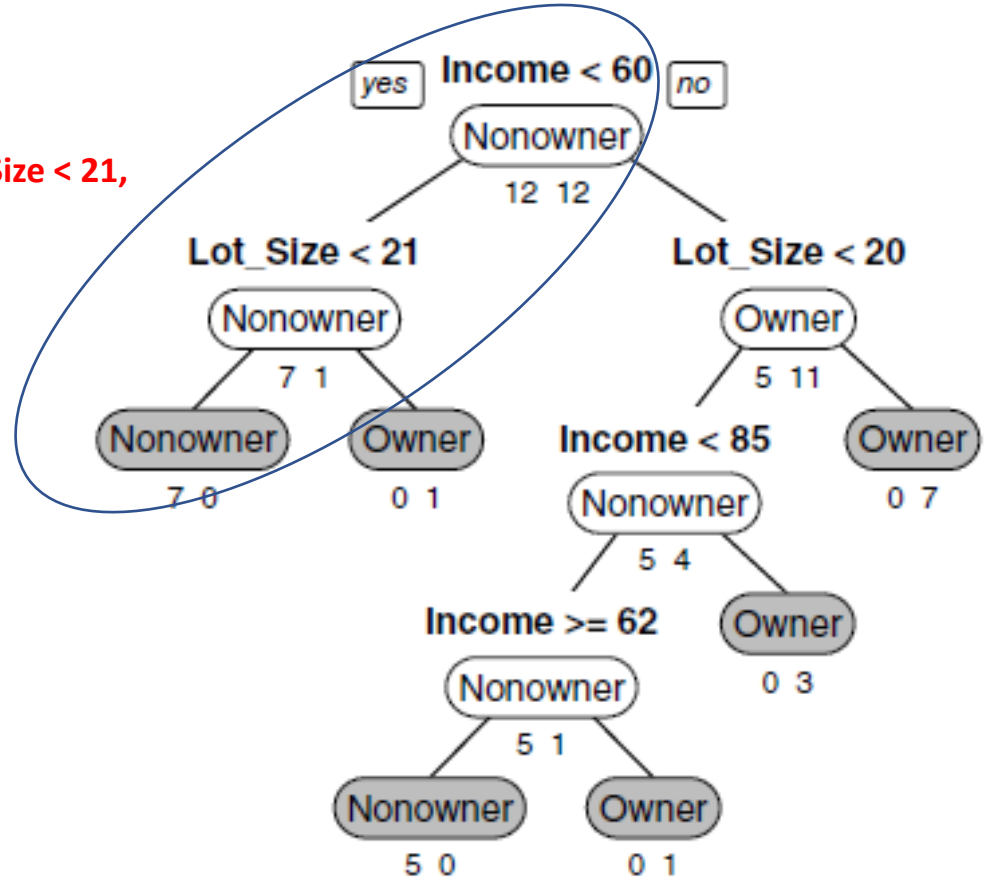
FIGURE 9.6 FINAL STAGE OF RECURSIVE PARTITIONING; EACH RECTANGLE CONSISTING OF A SINGLE CLASS (OWNERS OR NONOWNERS)

First Split of The Tree



Tree after all splits

If Income < 60 AND Lot Size < 21, classify as "Nonowner"



The first split is on Income, then the next split is on Lot Size for both the low income group (at lot size 21) and the high income split (at lot size 20)

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

Personal Loan Offer

Outcome variable: accept bank loan (0/1)

Predictors: Demographic info, and info about their bank relationship

- Age
- Experience
- Income
- Family
- Educational Qualifications
- CreditCard
- Fixed Deposits
- Online Banking
- Demat Account

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

Full trees are complex and overfit the data

- Natural end of process is 100% purity in each leaf
- This **overfits** the data, which end up fitting noise in the data
- For eg. the Loan Acceptance dataset with many more records and more variables than the Lawn Mower data generate a very complex full tree

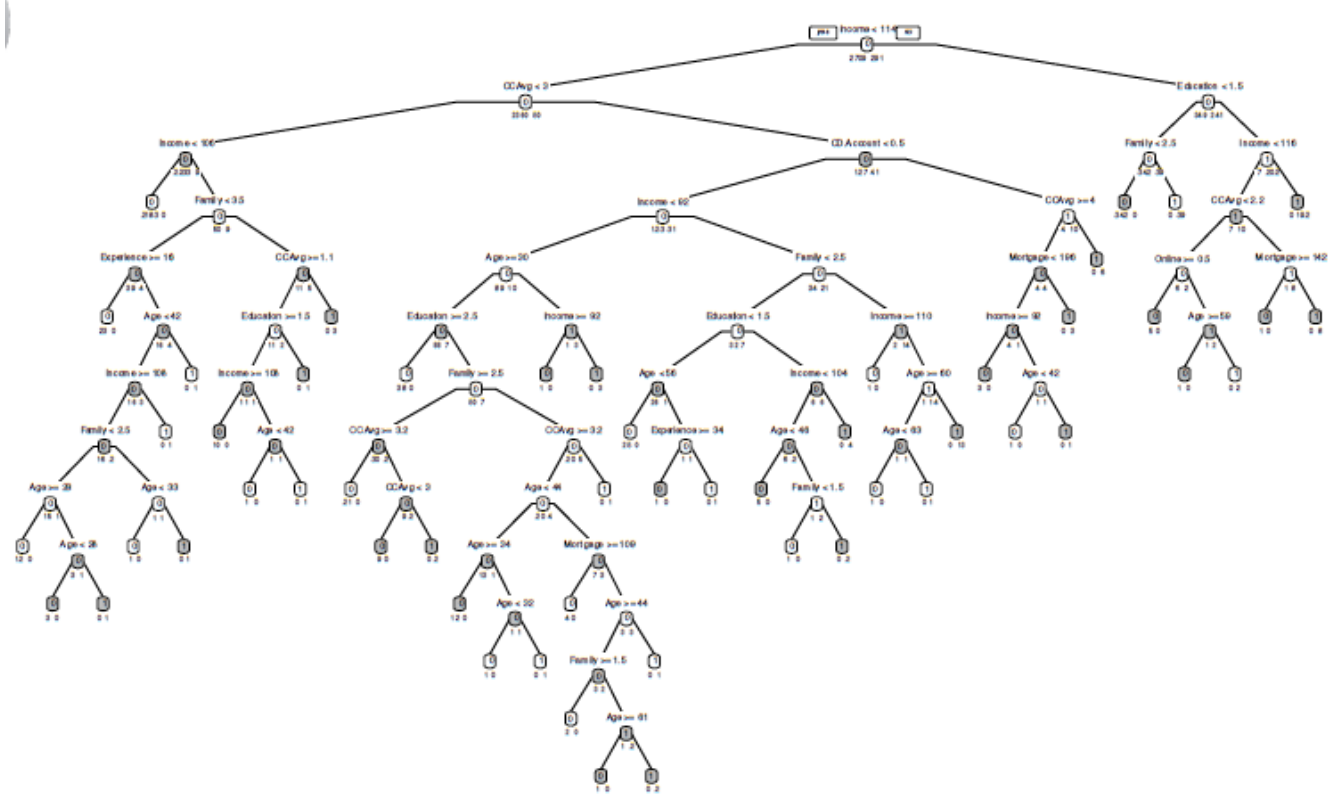


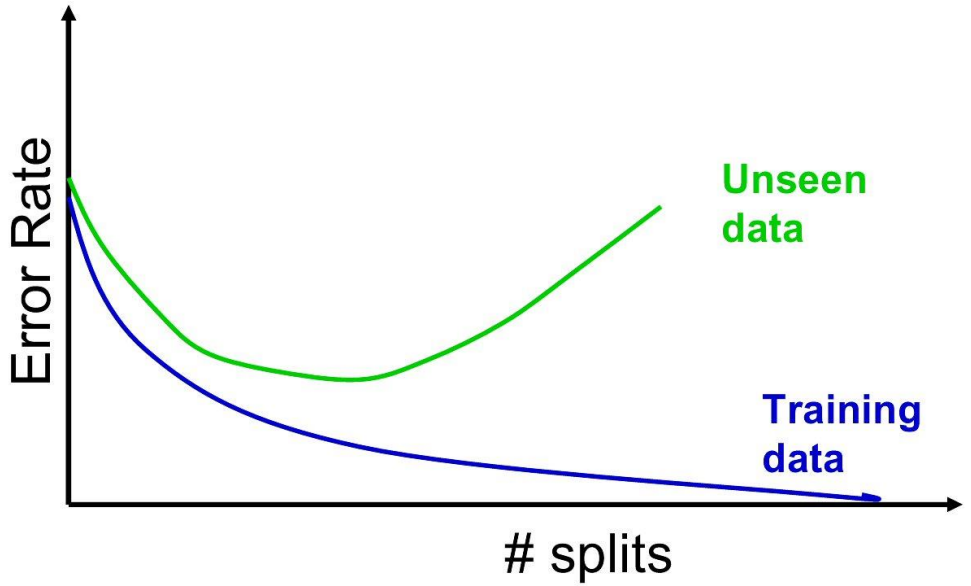
FIGURE 9.10 A FULL TREE FOR THE LOAN ACCEPTANCE DATA USING THE TRAINING SET (3000 RECORDS)

Ref: Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.

Overfitting

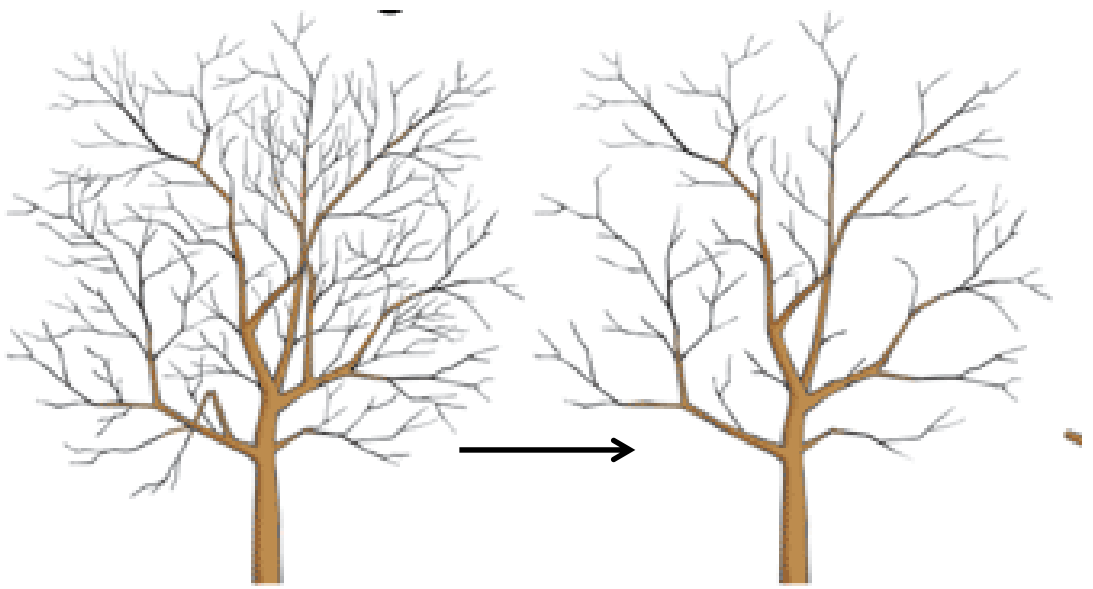
- ML models can produce highly complex explanations of relationships between variables, where the ‘fit’ maybe excellent, but when used with new data, models of great complexity do not do so well

  Overfitting produces poor predictive performance – past a certain point in tree complexity, the error rate on new data starts to increase



Pruning

- Pruning reduces the impact of overfitting
- Allowing the tree to grow to the full extent, then prune it back
- Generate successively smaller trees by pruning leaves
- At each pruning stage, multiple trees are possible
- Use *cost complexity* to choose the best tree at that stage



Regression Trees for Prediction

- Used for continuous (numerical) outcome variable
- Procedure similar to classification tree
- Many splits attempted, choose the one that minimizes impurity

- Prediction is computed as the **average** of numerical target variable in the rectangle (in Classification Tree it is majority vote)
- Impurity measured by **sum of squared deviations** from leaf mean
- Performance measured by RMSE (root mean squared error)

Advantages and Disadvantages of Decision trees

- Easy to use, understand
- Produce rules that are easy to interpret & implement
- Variable selection & reduction is automatic
- Does not require the assumptions of statistical models
- Can work without extensive handling of missing data

Disadvantage of single trees:

- instability and poor predictive performance
- Easy to overfit
- Require elaborate pruning
- Output range is bounded (in regression trees)

Summary

- Decision trees are an easily understandable and transparent method for predicting or classifying new records
- A single tree is a graphical representation of a set of rules
- Many different decision tree algorithms such as ID3, C4.5, CART, etc
- Splitting criterion: information gain (Shannon Entropy, Gini impurity, misclassification error), use of statistical tests, objective function, etc.
- Concept of recursive partitioning to repeatedly split the records into two parts so as to achieve maximum homogeneity of outcome within each new part
- Tree growth must be stopped to avoid overfitting of the training data – Pruning helps to reduce overfitting
- Easy to use and understand as they produce rules that are easy to interpret & implement, but often prone to overfitting

References

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- https://www.kdnuggets.com/2019/02/decision-trees-introduction.html
- https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/