Decision Trees
divide & conquer

Tom Breur
London, 23 April 2008
tombreur@xlntconsulting.com
www.xlntconsulting.com
+31-6-463 468 75

XLNT consulting

Agenda

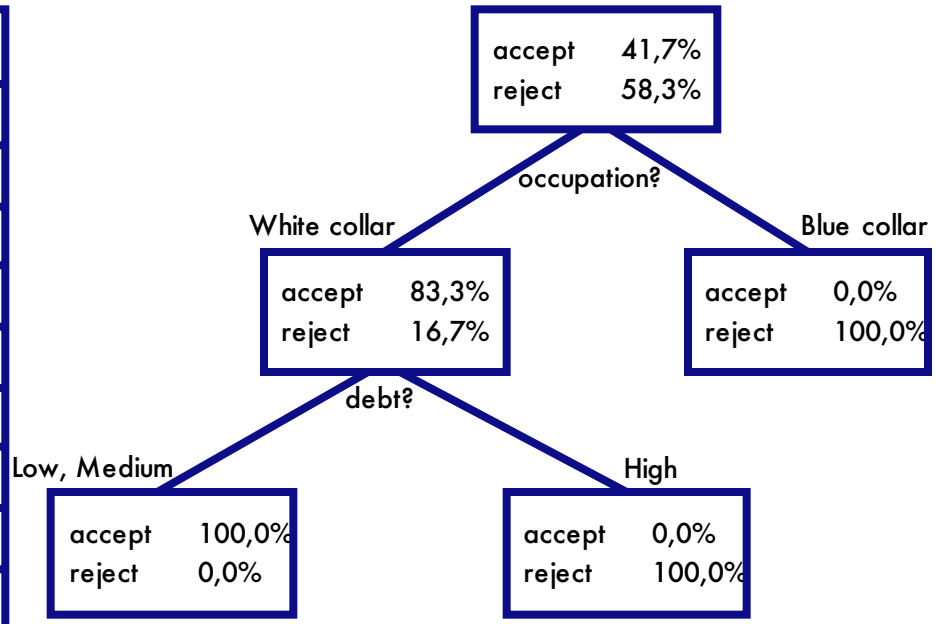
- Features of decision trees
- Overview algorithms
- Exploration and prediction
- Automatic ⇔ manual
- Pitfalls & best practices
- Decision trees ⇔ regression

Features of decision trees

- Symbolic analysis ⇒ recursive partitioning ⇒ decision trees (inductive learning)
- Split record set arriving at each node, using "best" variable ⇒ 'rinse & repeat'
- Stop when:
  - All records in leaf belong to same class
  - No variable can be found for splitting

Example decision tree

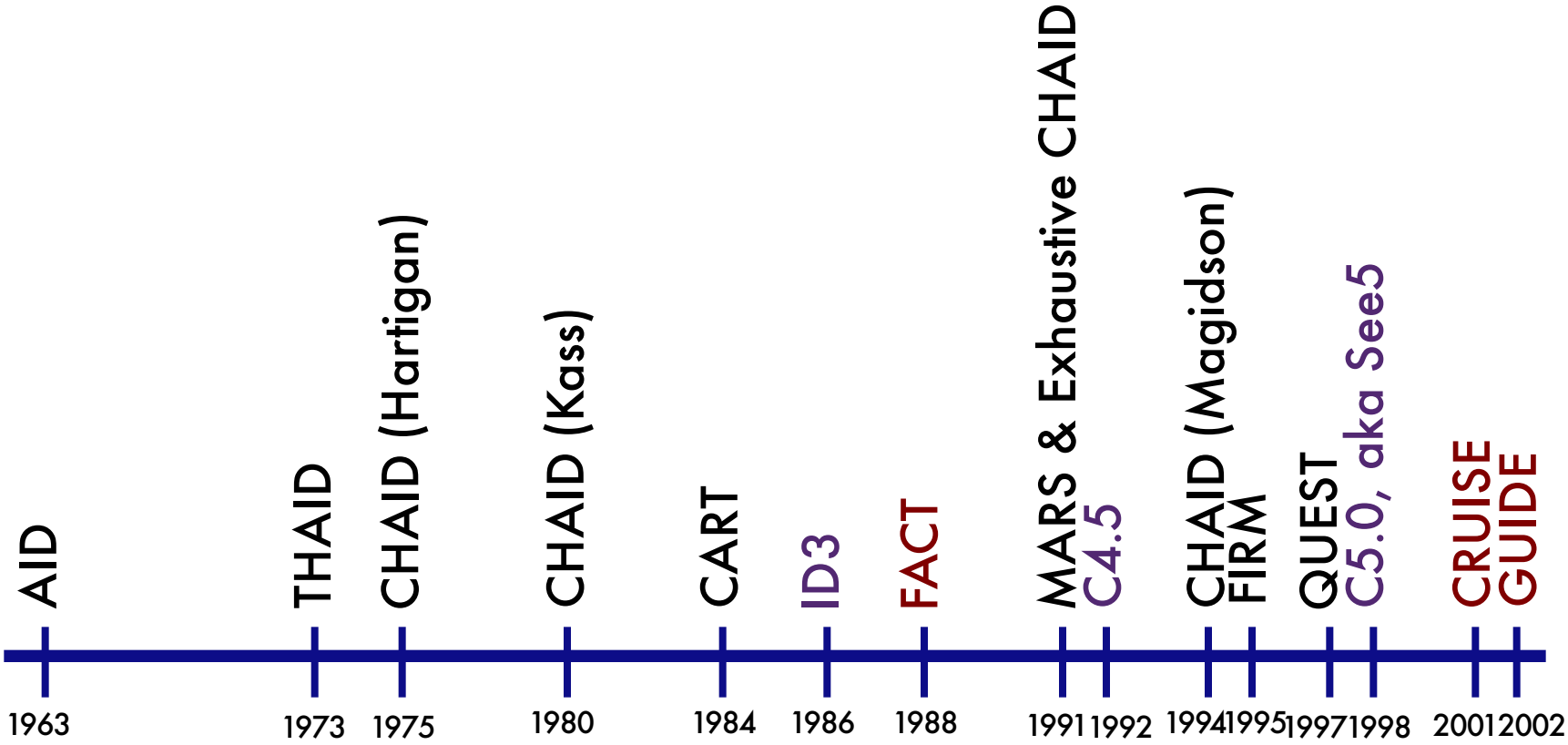
occupation	buroscore	debt	credit?
Blue collar	Low	Low	No
White collar	High	Low	Yes
Blue collar	Low	High	No
White collar	High	High	No
Blue collar	Low	High	No
White collar	High	Medium	Yes
White collar	High	Low	Yes
White collar	High	Medium	Yes
Blue collar	Low	High	No
Blue collar	High	High	No
Blue collar	Low	High	No
White collar	High	Medium	Yes



Overview algorithms (1)

- *Usually* **categorical target**, sometimes continuous (e.g.: CART❶)
- *Usually* **binary target**, sometimes multiple categories (e.g.: CHAID)
- *Usually* ***statistical*** **loss function**, sometimes information theory (e.g.: ID3, C4.5, C5.0)
- Binary or multiple splits; nominal, ordinal or 'continuous' predictors

Overview algorithms (2)



Exploration and prediction

- ***Segmented*** prediction
- Insight in complex structures
- Discover noteworthy interactions

Every predictive model ***must*** be accompanied by insight:

- Sanity check
- Foster adoption
- Spur data-driven business innovation

Automatic ⇔ manual

- Manual: apply domain expertise (aka "model engineering"):
  - Variable selection
  - Business problem
  - Implementation specification
- Short ⇔ long term lift characteristics:
  - "Transient"/behavioural ⇔ stable variables
- Manual tree building ***drives*** variable development

Pitfalls & best practices (1)

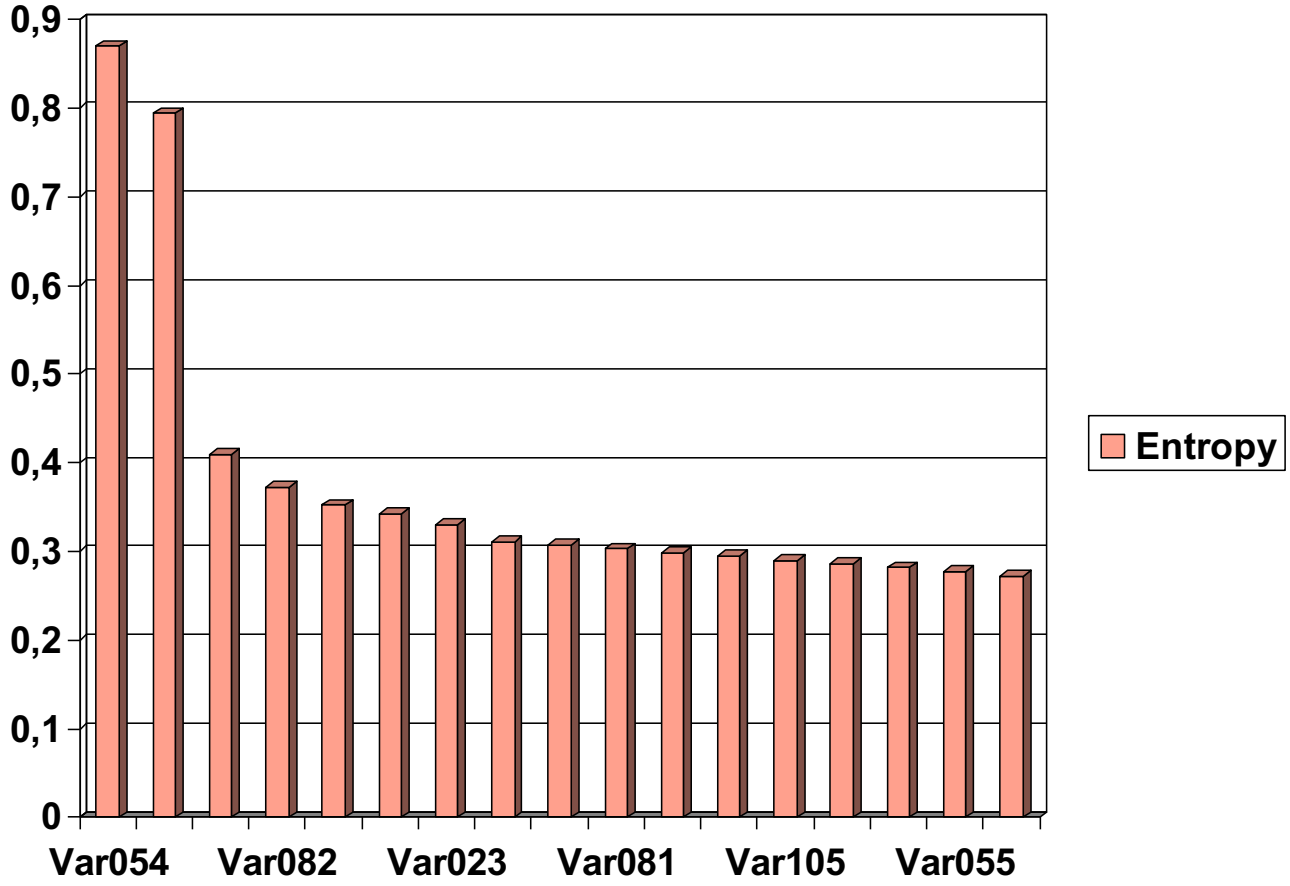
Pitfall #1

- "Leakers"❷/"Anachronistic variables"❸ – when the model looks too good to be true, it probably is…

Best practice #1

- Plot ***sorted*** univariate relation between input & output, look for a 'drop' (=suspect) ⇒ e.g.: 1st two variables next slide

Example best practice #1



Pitfalls & best practices (2)

Pitfall #2

- Never assume it is ***the*** tree, it is always ***a*** (possible) tree

Best practice #2

- Describe associations between most important input variables and target, *even if variables do not appear in the (eventual) tree*

Pitfalls & best practices (3)

Pitfall #3

- Overtraining, overly optimistic prognosis

Best practice #3

- Divide mining set into 3 parts[4]: training-, test-, and evaluation set (50-40-10% feels about right)

Pitfalls & best practices (4)

Pitfall #4

- Replace missing by constant (mean/mode)

Best practice #4

- Identify ***rightfully*** missing yes/no❸
- ***If*** replacing, append boolean: "was previously missing"
- Avoid ***adding bias***, "intelligent" imputation

Decision trees ⇔ regression

- Few sound comparative studies
- Most *familiar* technique works best
- 'On average' regression predicts more accurately

Alternative considerations:

- Spur development of variables
- Innovate business

Conclusion

- Decision trees are:
  - Flexible
  - Versatile
  - Gentle learning curve
- ***Superior*** insight drives:
  - Development of (better) predictive variables
  - Innovation of business
- Manual tree building enhances 'data learning'