Evaluating Decision Tree models for credit

Data Analysis
FGV – Winter School

Raphael Ferreira

FGV EAESP

Problem

Evaluating possible decision tree models for credit

Target: Creditability ("bad" or "good")

Database

- German Credit Data (Statlog)
- 1000 instances
- 20 attributes (13 categorical, others integer)
- Year: 1994
- Source: UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html)

Attributes/Variables
Duration in month
Credit history
Purpose
Credit amount
Savings account/bonds
Present employment since
Installment rate in percentage of disposable income
Personal status and sex
Other debtors / guarantors
Present residence since

Attributes/Variables
Property
Age in years
Other installment plans
Housing
Number of existing credits at this bank
Job
Number of people being liable to provide maintenance for
Telephone
foreign worker
Creditability

300 Bad
700 Good

Data Preparation

- Changing categorical to numerical
- Database split in 60% training and 40% testing, randomly sorted

Watson Suggestion

- Predictive strength was around **70%** for almost all variables, when it was in categorical format (Age was the first variable suggested)

- But predictive strength drop to **18%** and variables chosen changed, when was formatted as numerical

- However, using numerical format, accuracy improved in this model

- And data quality improved as well: from 69% to 74%

Models

Model 1: “All”

- “Supervised”
- Use all variables

Model 2: “Chosen”

- Supervised
- Use some variables
  - Duration in month
  - Credit Amount
  - Credit History
  - Present Employment Since
  - Job

Model 5: “Watson Analytics”

- “Supervised”
- Use some variables
  - Duration In month
  - Installment Rate
  - Credit Amount
  - Credit History
  - Provide Maintenance For

Model 4: “Random Forest”

- “Supervised”
- Use all variables

Results

Considering this Cost/Ben. Matrix

Cost/Ben. Matrix		Reference	
		Bad = 0	Good = 1
Prediction	Bad = 0	0	-1
	Good = 1	-5	3

Model 1: "All"

Confusion Matrix		Reference	
		Bad = 0	Good = 1
Prediction	Bad = 0	50	47
	Good = 1	62	241

# Nodes	17
Accuracy	0,7275
Balanced Accuracy	0,6416
AUC	0,7311
EV	366

Model 2: "Chosen"

Confusion Matrix		Reference	
		Bad = 0	Good = 1
Prediction	Bad = 0	43	42
	Good = 1	69	246

# Nodes	7
Accuracy	0,7225
Balanced Accuracy	0,6190
AUC	0,7165
EV	351

Model 3: "Watson"

Confusion Matrix		Reference	
		Bad = 0	Good = 1
Prediction	Bad = 0	38	27
	Good = 1	74	261

# Nodes	5
Accuracy	0,7475
Balanced Accuracy	0,6228
AUC	0,7165
EV	386

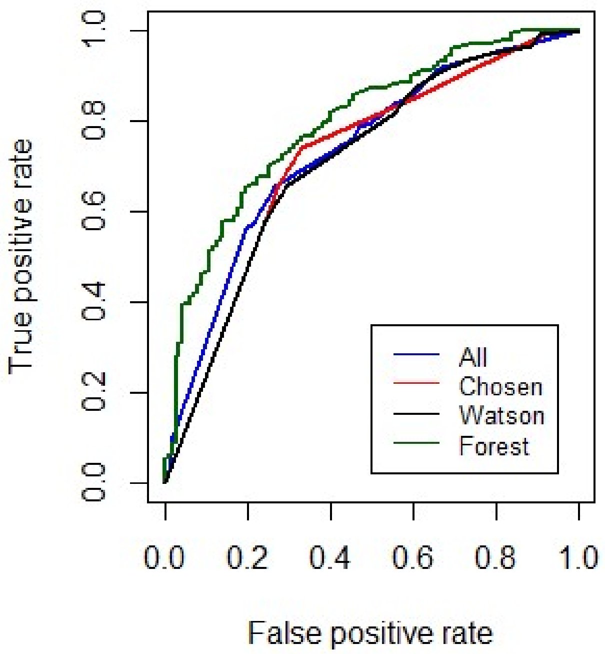
Model 4: "Random Forest"

Confusion Matrix		Reference	
		Bad = 0	Good = 1
Prediction	Bad = 0	45	31
	Good = 1	67	257

# Trees	500
Accuracy	0,7550
Balanced Accuracy	0,6471
AUC	0,7930
EV	405

Performance

- Random Forest presents the best performance in all indicators
- "Watson" is the second, and not so far from Random Forest (Accuracy, Balance Acc. and EV), and less complex
- "Chose" is the worst
- Important: EV depends on Cost/Benefit Matrix



Conclusions

- Although Random Forest is the best model, it uses all variables, and some of them could lead to accountability problems (expected legal costs)
- Board should take in account these potential legal costs and expected profit from this model, and if potential costs are too high then..
- **Watson Model** is recommended!