Predicting Ship Casualties

David Delgado & Michel van de Velden

Content

- Ship Casualties Data
- Methodology: Classification Trees
- Asymmetry
- Results

Ship casualties

- IMO Maritime Safety Committee
- Very serious
- Serious
- Less serious

Data

- Serious casualties
- About 41.000 observations, 7.8% casualties
- Basic Ship properties: age , size, type
- Deficiencies
- Flag

Methodology: Classification Trees

- Classification tree
- Overfitting
- Ensembles: Boosting and Random Forests
- Benchmark: Logit

Asymmetry: issues

- More than ten times as many non-casualties
- Casualties are more important

Asymmetry: solutions

- Under sampling
- Sampling bias corrections
- Misclassification costs

	Actual is 1	Actual is 0
Predicted is 1	0	Cost of false positive
Predicted is 0	Cost of false negative	0

Evaluation

- Hitrate
- Precision$= \frac{true\ positives}{true\ positives + false\ positives}$
- Recall $= \frac{true\ positives}{true\ positives + false\ negatives}$
- $F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$

Results

Method	Precision	Recall	Hitrate	F_1 score	F_2 score
Logit	0.74	0.28	0.94	0.40	0.32
Single Tree	0.21	0.72	0.77	0.33	0.49
Random Forest	0.27	0.77	0.82	0.40	0.56
GentleBoost	0.24	0.67	0.81	0.35	0.49
LogitBoost	0.20	0.76	0.75	0.33	0.50

Random Forests

Sampling	Correction	Precision	Recall	Hitrate	F_1 score	F_2 score
Regular sampling	None	0.81	0.39	0.95	0.47	0.41
	Costs	0.83	0.35	0.94	0.44	0.38
Undersampling	None	0.27	0.77	0.82	0.40	0.56
	Costs	0.28	0.75	0.83	0.40	0.54
	Simple	0.29	0.72	0.84	0.41	0.55
	Bag	0.81	0.03	0.92	0.06	0.04

Conclusion

- Random forest performs well
- Under sampling and different costs are effective