Predicting Caravan Insurance: A Bumpy Ride

Fabian Blasch, Gregor Steiner, Sophie Steininger, Jakob Zellmann

01/19/2022

Data

- Caravan Insurance Data Set based on real world business data
- Supplied by the Dutch datamining company Sentient Machine Research
- 86 variables containing data on
- demographic statistics
- product usage
- Unbalanced Data
- classification data set with skewed class proportions

Elastic Net GLMs

- Combination of LASSO and Ridge penalty:

$$C(\alpha;\beta) = \alpha||\beta||_1 + \frac{1-\alpha}{2}||\beta||_2^2.$$

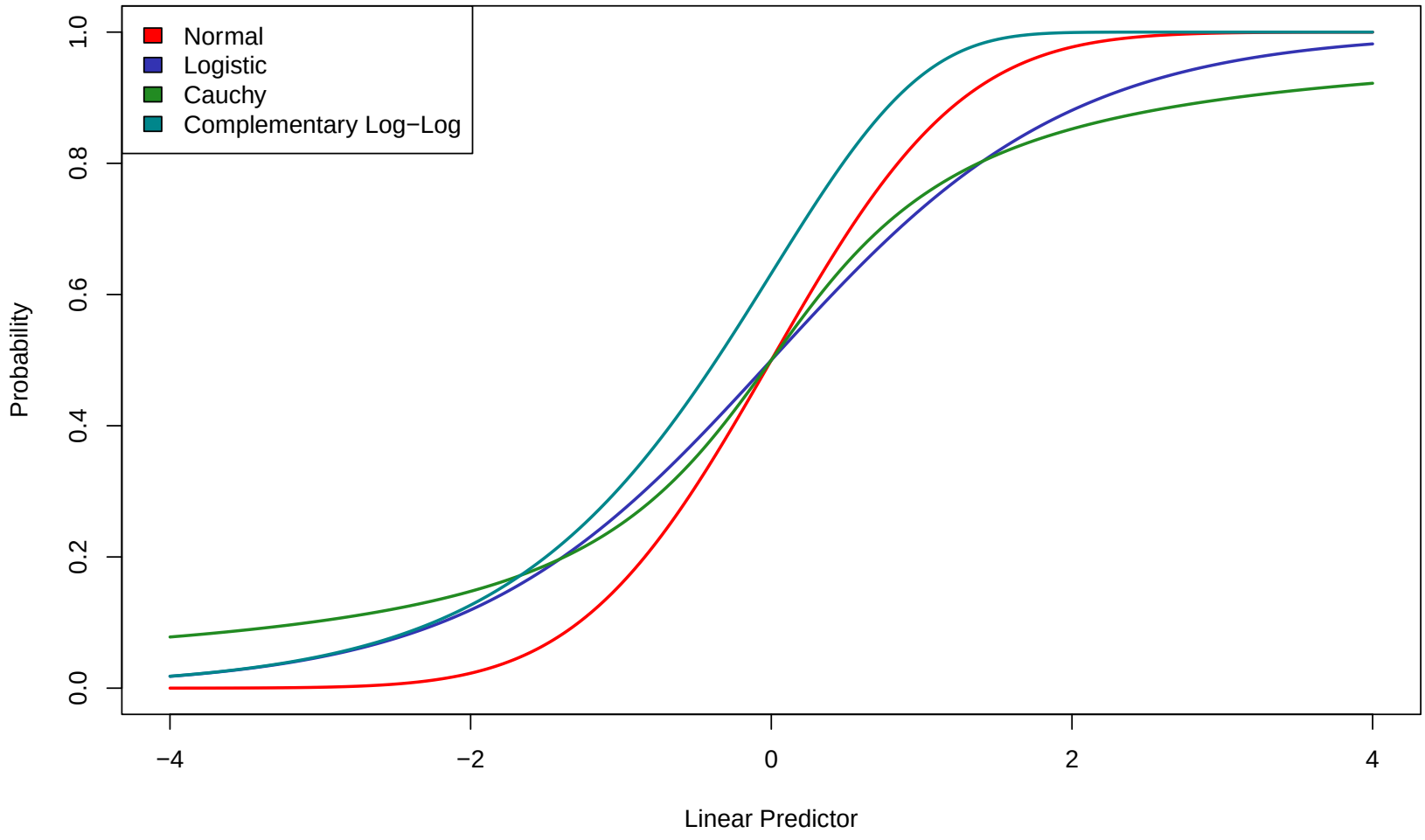
- GLM from the binomial family with different link functions
- The minimization problem is

$$\min_{\beta\in\mathbb{R}^k} -\frac{1}{n} l(y, X;\beta) + \lambda C(\alpha;\beta)$$

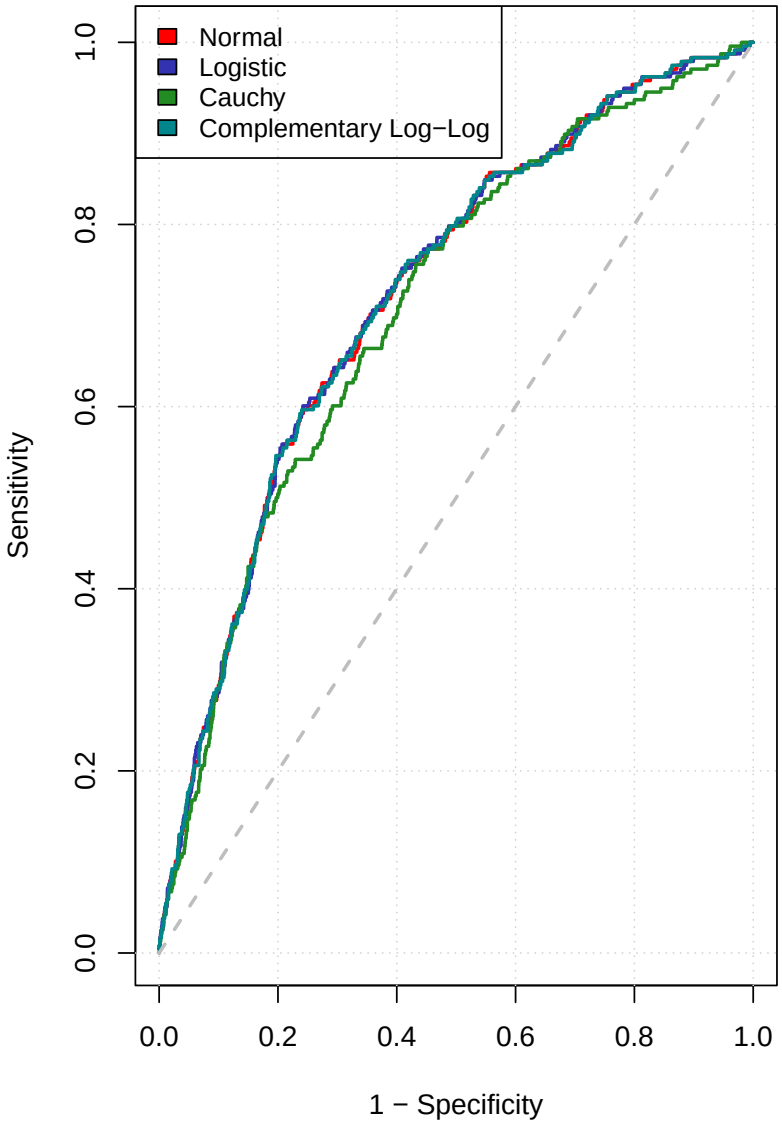
where $l(y, X;\beta)$ is the log-likelihood.

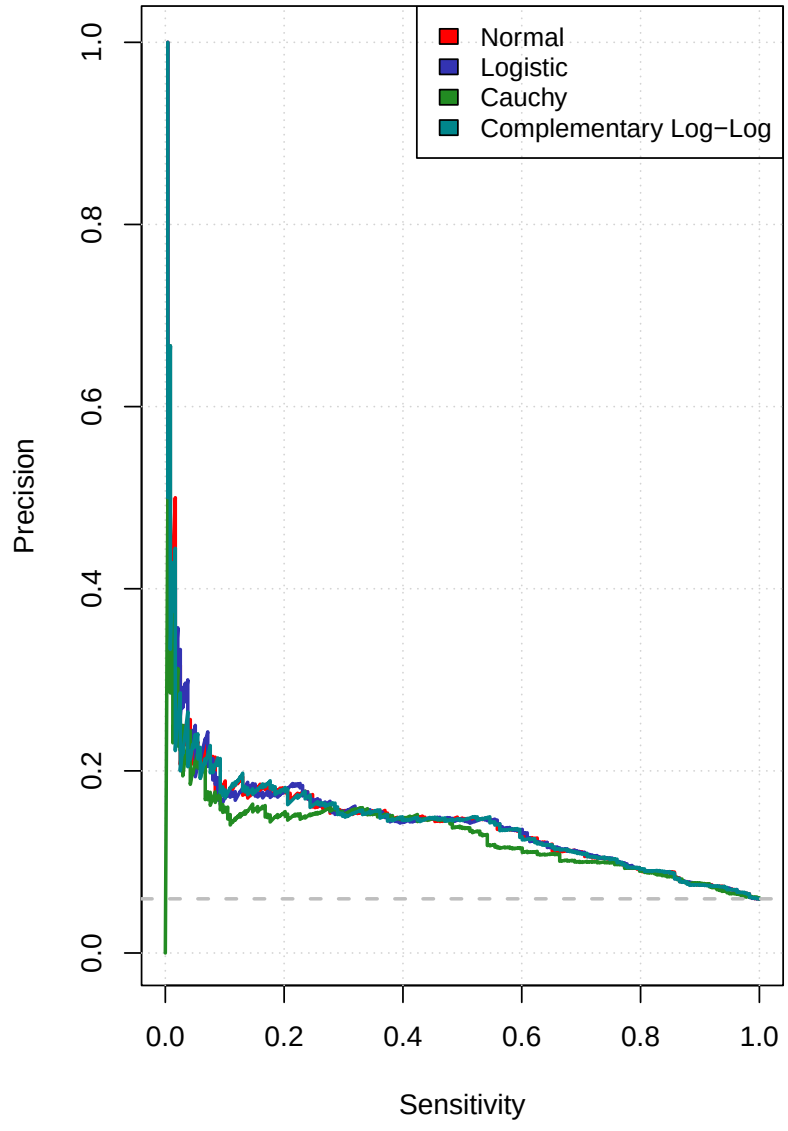
GLMs: Link Functions

- We try 4 different link functions: Normal cdf (Probit), Logistic cdf (Logit), Cauchy cdf, and complementary log-log



GLMs: Performance

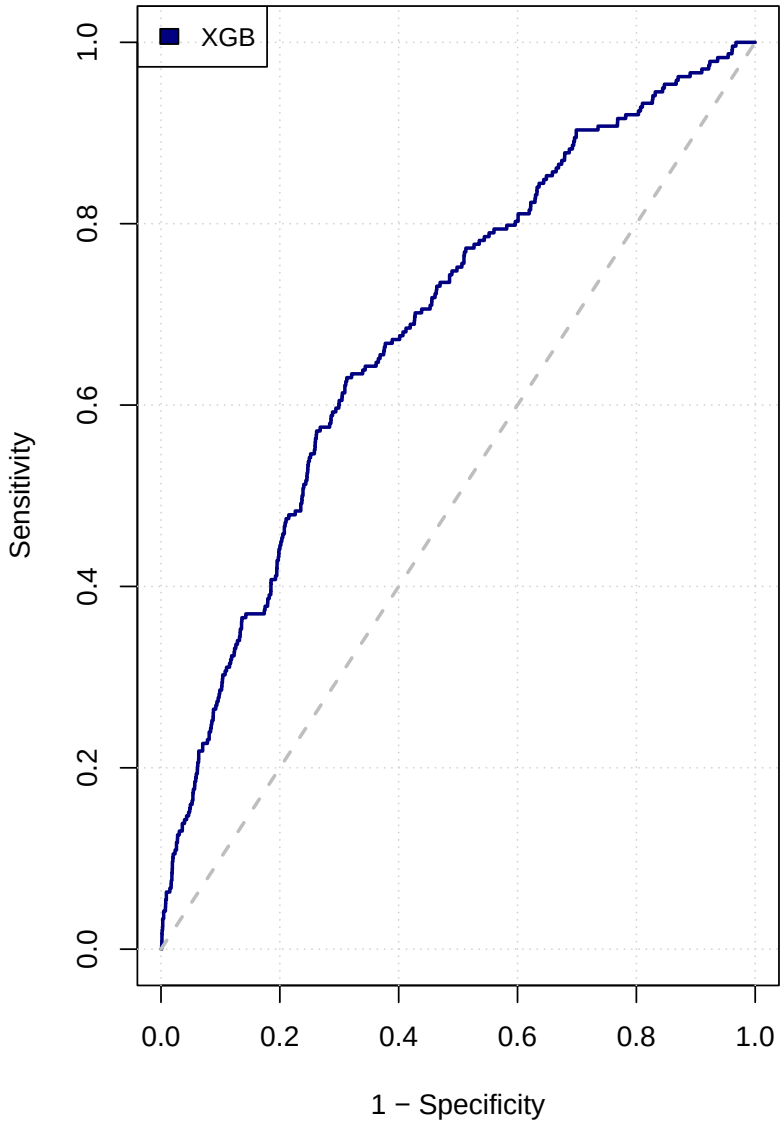


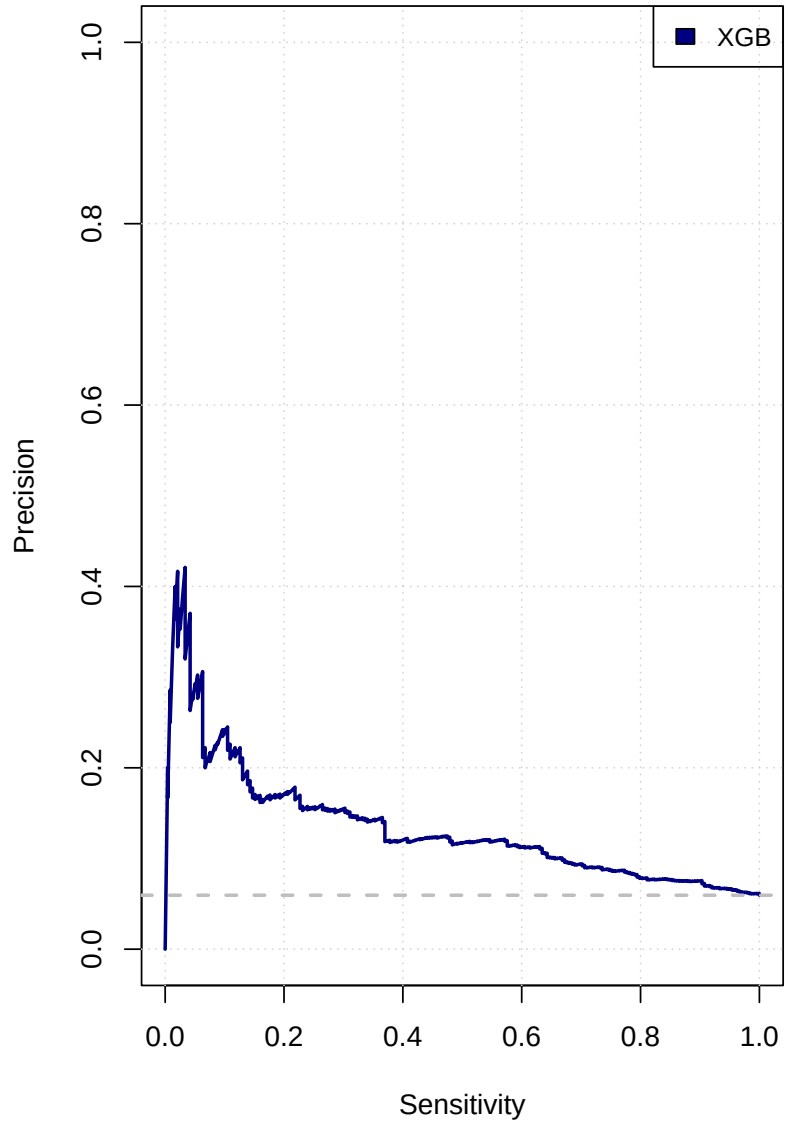


XGBoost

- the learning rate, $\eta \in [0.01, 0.6]$ (default: 0.3),
- the regularization parameters, $(\gamma, \lambda) \in [0, 1] \times [0.01, 2]$ (default: 0 and 1 respectively),
- the maximal depth of the trees, $max\_depth \in \{2, ..., 10\}$ (default: 6),
- the maximal number of single trees contained in one model, $nrounds \in [1, 1000]$,

XGBoost Performance

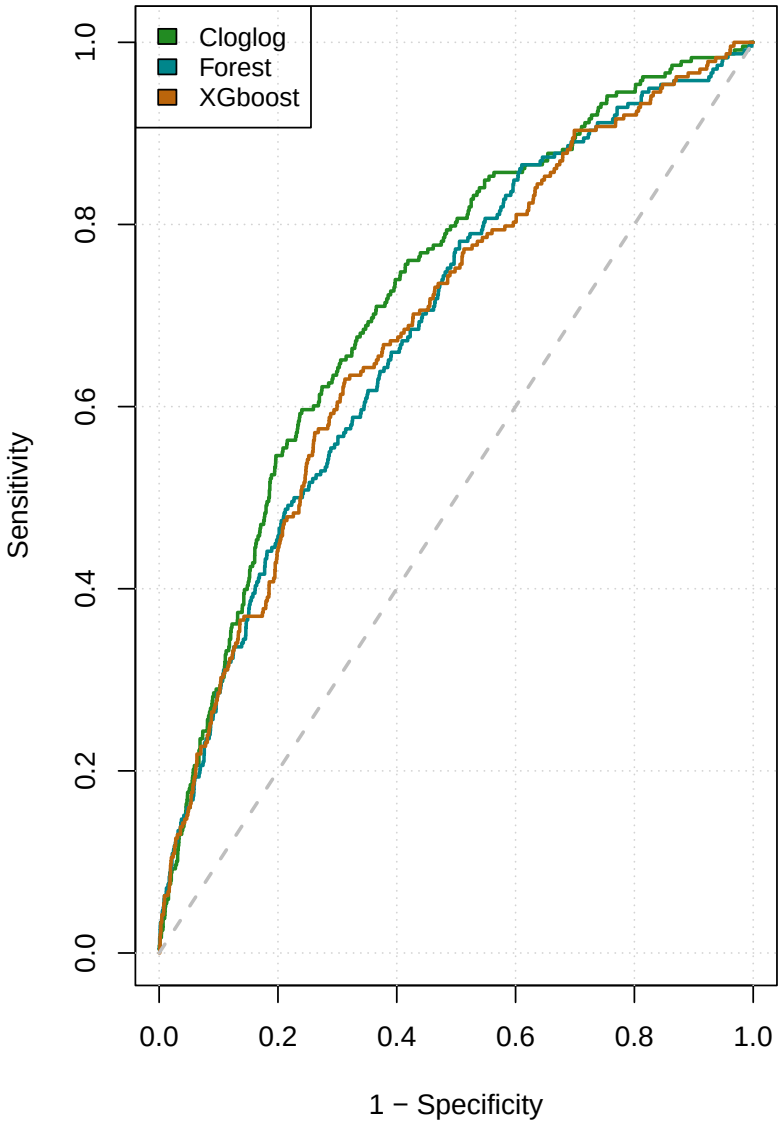


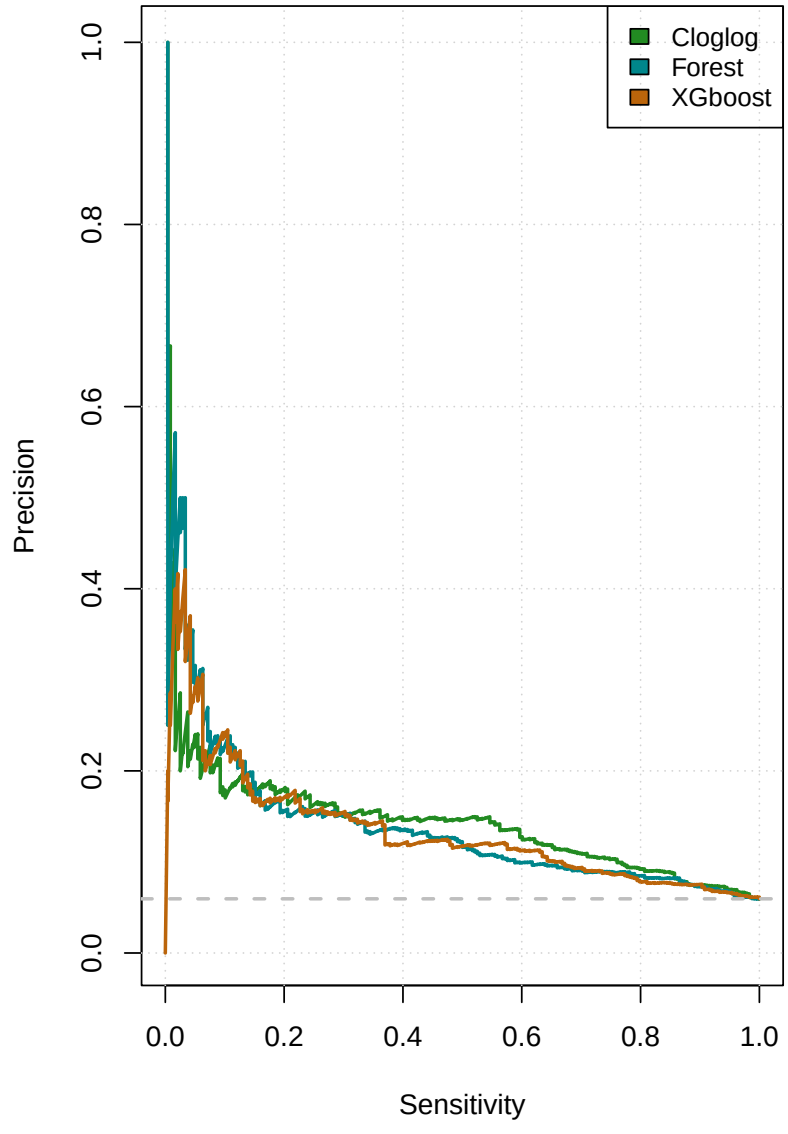


Random Forest

- mtry: number of variables to possibly split at in each node (85)
- min.node.size: minimal node size (8)
- splitrule: gini

Performance Comparison

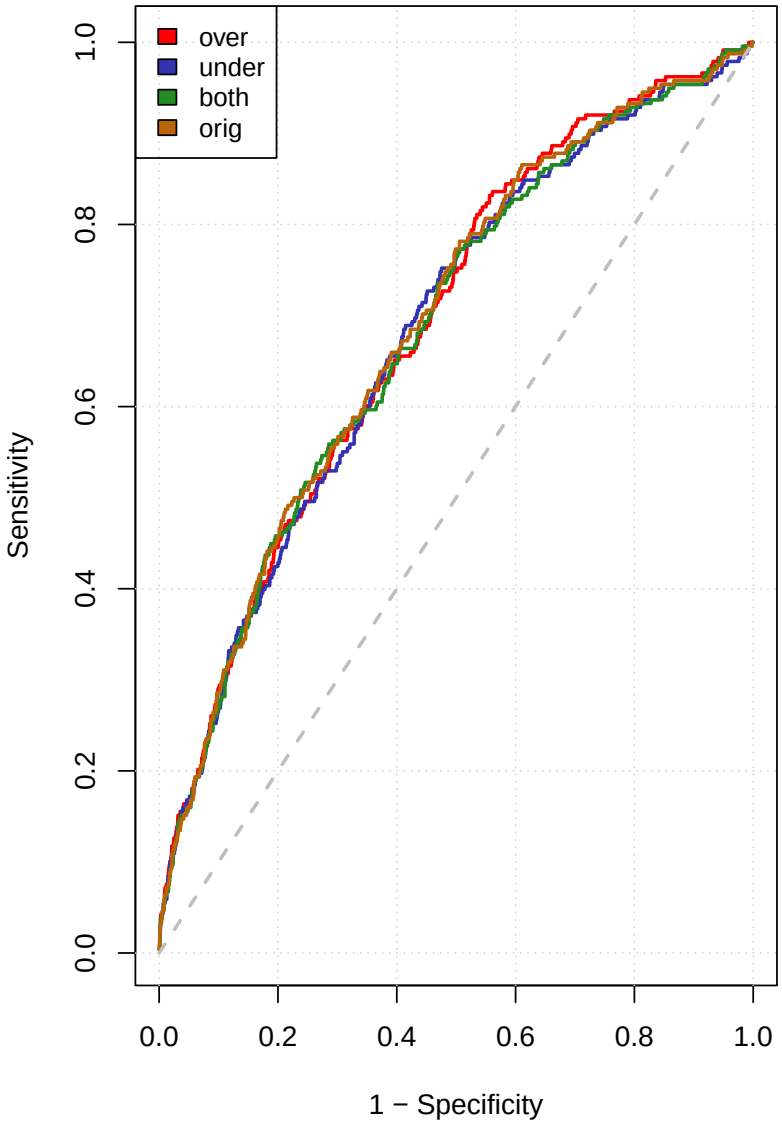


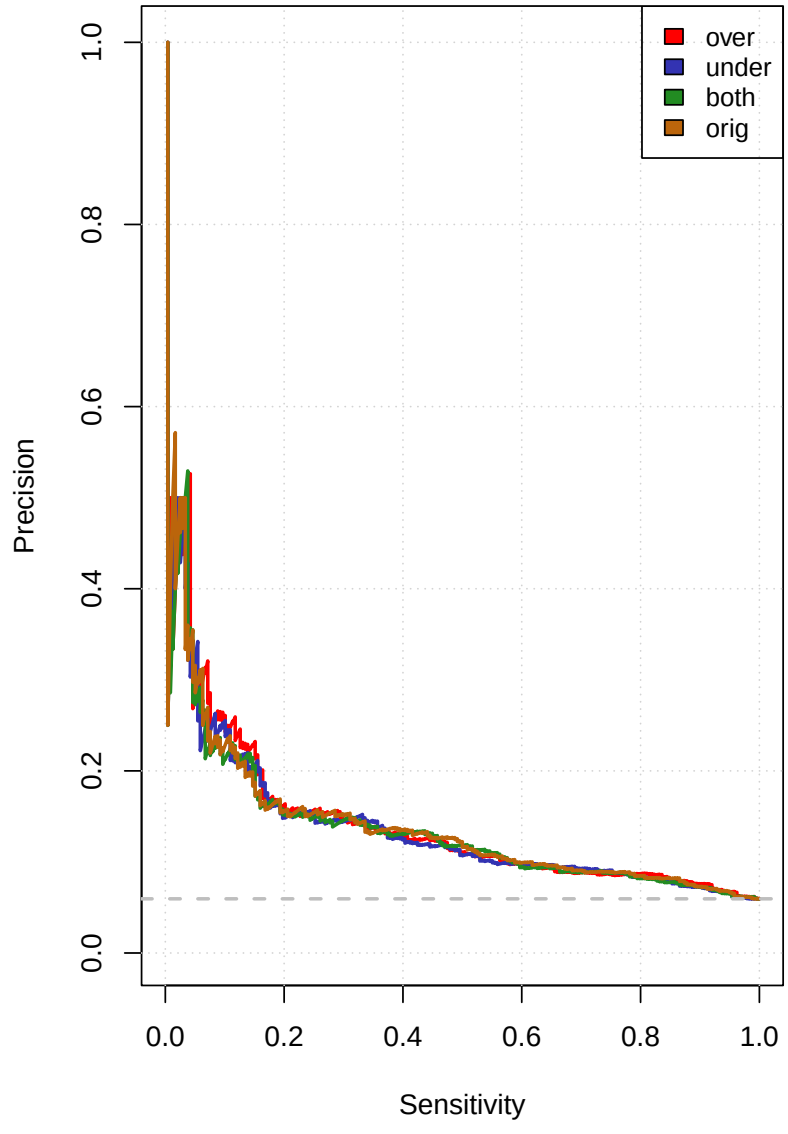


Sampling

- Under sampling: sampling from the majority class without replacement and leaving the minority class in tact
- Over sampling: sampling from the minority class with replacement and leaving the majority class in tact
- Both: sampling from the minority class with replacement and from the majority class without replacement

Forest Sampling Performance





Current Challenge Performance

Model	CoIL Performance
logit	118
probit	116
cauchit	114
cloglog	117
forest	105
over	100
under	98
both	104
XGB	97
1st Place	121

Outlook

- Cost function that penalizes false negatives
- Feature engineering

Citation

- P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000.