CSCC 2015
Edinburgh, 27/08/15

On a Systematic Hyperparameter Tuning Framework in R for Credit Scoring with Special Considerations to Class Imbalance Correction

Gero Szepannek
Head of Scoring & Rating Models
Santander Consumer Bank
Germany

Bernd Bischl, Tobias Kühn
Dept. Statistics
LMU München
Germany

Santander
CONSUMER BANK

Scope: Yet another ML Benchmark Study...

- Many ML benchmark studies published within last decade
- **Here:** focus on class imbalance
- ...publications often suffer from „academic bias“

- Class imbalance
- Systematic hyperparameter tuning
- Benchmark experiment on large DB
- Realistic evaluation

- Joint framework for classifier tuning and imbalancy correction
- Recent techniques like SMOTE and overbagging investigated
- Extension of SMOTE for categorical variables using Gower distance
- Iterated F-Racing instead of grid search for tuning within mlr
- Experiment on large data base – validation on credit scoring data sets
- Realistic evaluation on coarse classed data

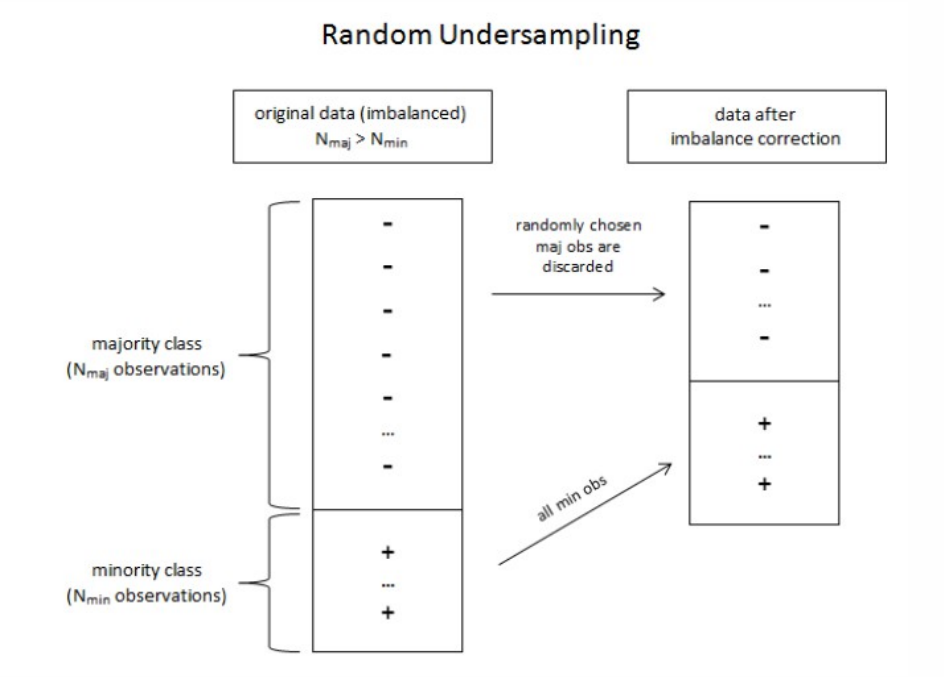
Class Imbalancy

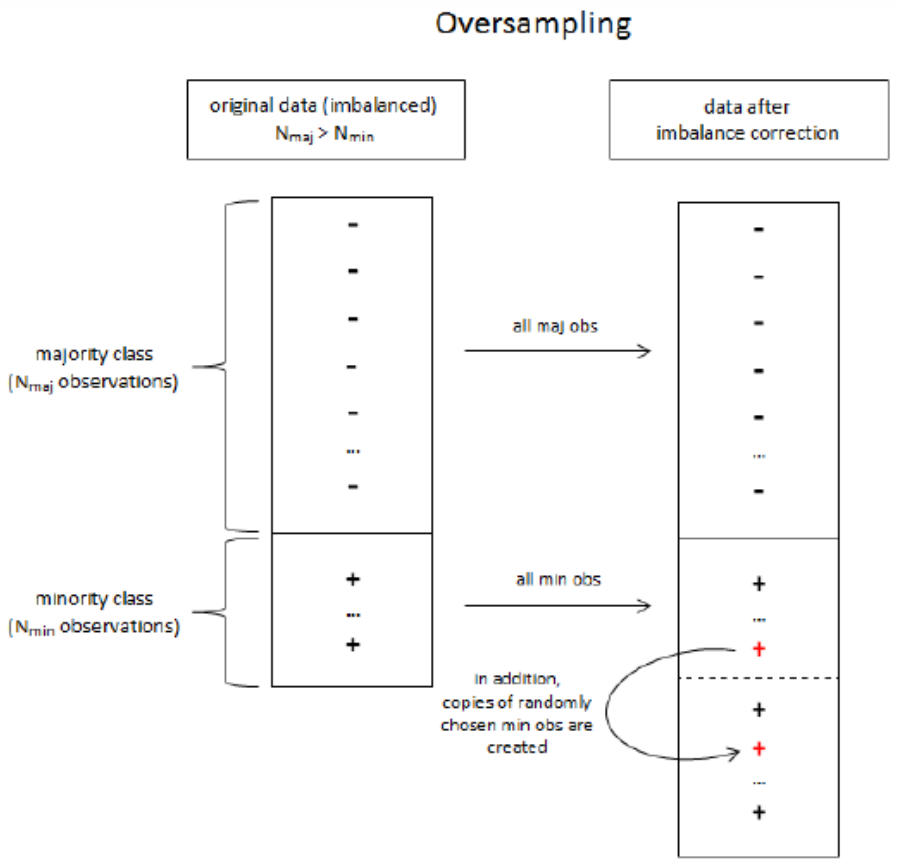
- **Class imbalance**
- Systematic hyperparameter tuning
- Benchmark experiment on large DB
- Realistic evaluation

- **Joint framework for classifier tuning and imbalancy correction**
- **Recent techniques like SMOTE and overbagging investigated**
- **Extension of SMOTE for categorical variables using Gower distance**
- Iterated F-Racing instead of grid search for tuning within mlr
- Experiment on large data base – validation on credit scoring data sets
- Realistic evaluation on coarse classed data

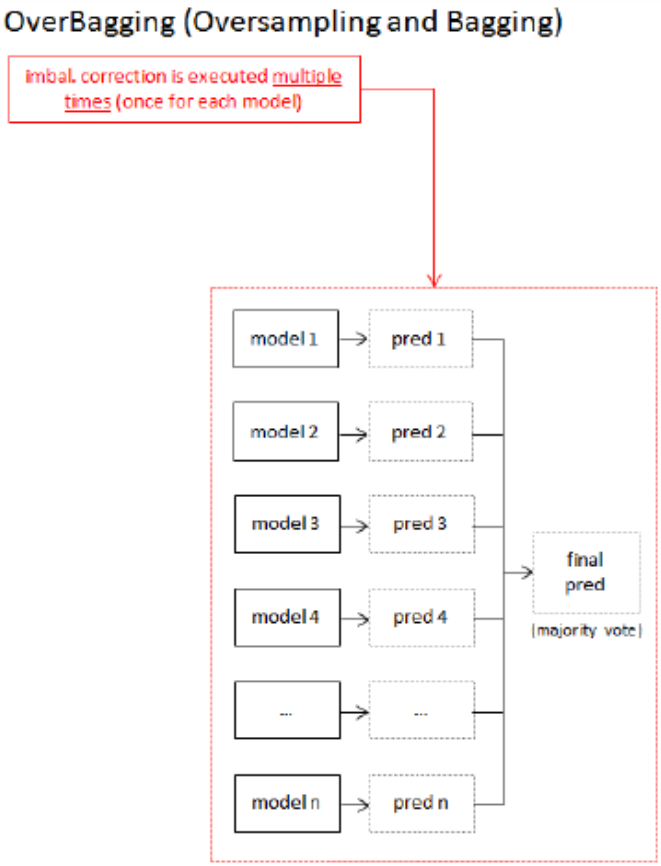
- In CS often highly imbalanced classification problems (i.e. P(default) << 0.5)
- Several techniques for class imbalancy correction (IC)
- Vincotti and Hand (2002): considerations w.r.t. misclassification costs
- Brown and Mues (2012) analyze effect of over-/undersampling
- Crone and Finlay (2012) investigate over-/undersampling w.r.t. effective class sizes
- From ML community: recent techniques like SMOTE / overbagging (Chawla et al., 2002)
- Data in CS typically contains (at least several) nominal predictors
- Integrated optimization of classifier and IC

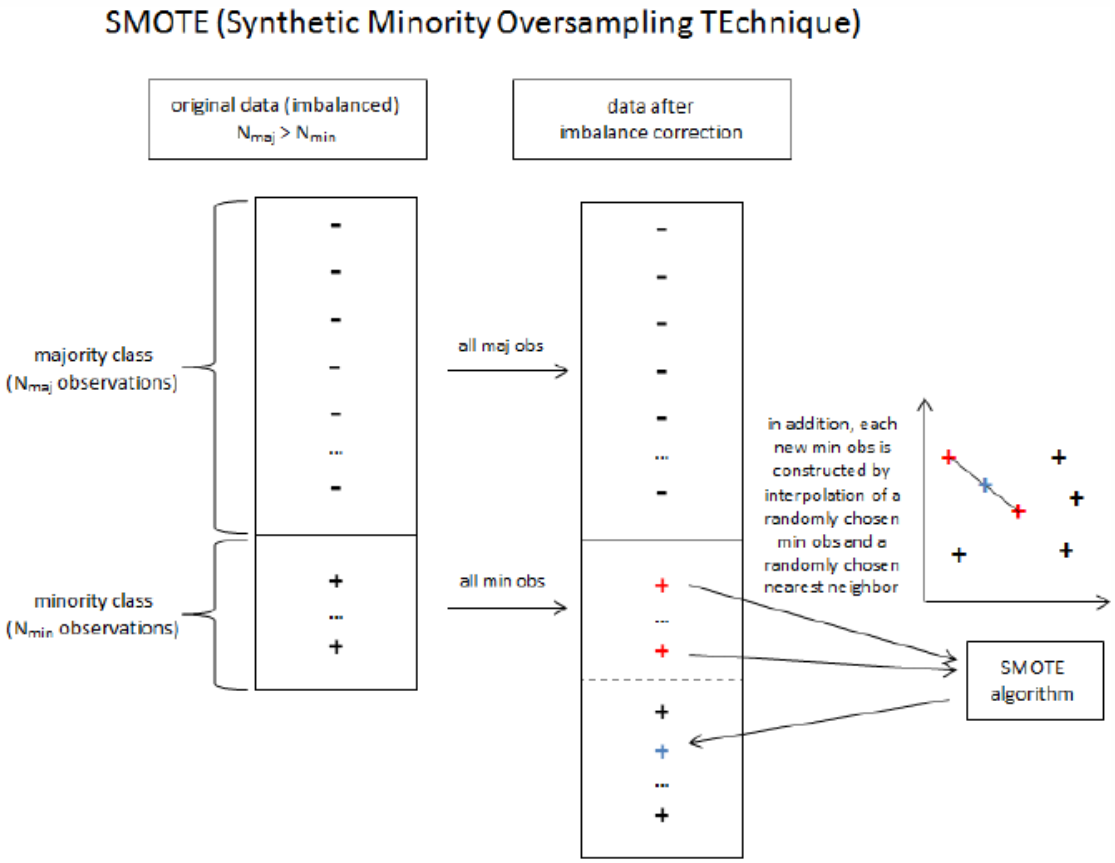
Over- and Undersampling





Overbagging and SMOTE





...alternatively, several classifiers allow for specification of observation weights.

Hyperparameter Tuning

- Class imbalance
- **Systematic hyperparameter tuning**
- Benchmark experiment on large DB
- Realistic evaluation

- Recent techniques like SMOTE and overbagging investigated
- Extension of SMOTE for categorical variables using Gower distance
- **Joint framework for classifier tuning and imbalance correction**
- **Iterated F-Racing instead of grid search for tuning within mlr**
- Experiment on large data base – validation on credit scoring data sets
- Realistic evaluation on coarse classed data

- Most classification algorithms depend on the choice of additional parameters
- Typical benchmark studies optimize algorithms w.r.t. these parameters (using validation sets)
- Often few attention is paid to the design of parameter optimization
- Mostly grid searches through parameter space are used

{mlr}: A Framework for Machine Learning in R

(Szepannek et al., 2010, Bischl et al., 2015)

- **Systematic optimization and evaluation of learning problems**
- **Implemented in open source software R**

Conceptual features:

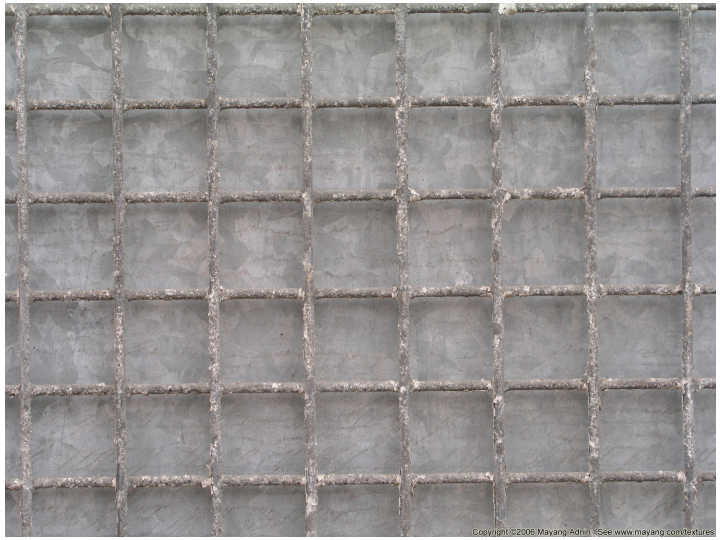
- Definition of Tasks
- Implementation of all standard (and even many non-standard) algorithms
- Allows for integrated optimization over all stages
- Object oriented: easy extension & customization

```
task  <- makeClassifTask(data = gcd, target = "Bad", positive = "1")
lrn   <- makeLearner("classif.ksvm")
ps    <- makeParamSet(makeNumericParam("C", lower = -12, upper = 12 )


mod   <- train(lrn, task = task)
pred  <- predict(mod, task = task)
performance(pred, measure = auc)

rdesc <- makeResampleDesc("CV", iters = 3)
res   <- resample(lrn, task, rdesc)
```

Grid Search over Parameter Space



- **Major drawbacks of grid search:**
  - computationally intensive: exponentially increasing with parameter dimension
  - ...typically optimization only w.r.t few parameters
  - only pre-specified parameters are evaluated

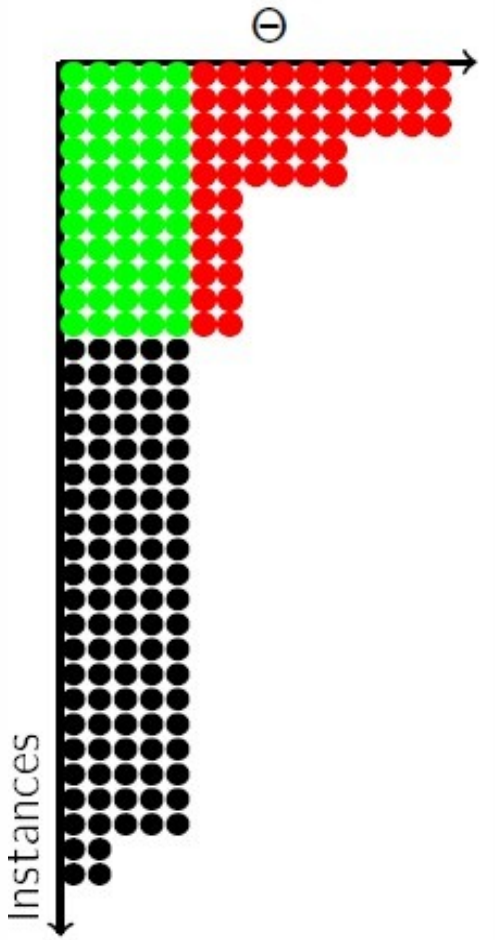
- **Better:**
  - Use of more intelligent search through parameter space like EA / Kriging /...
  - In {mlr} different intelligent optimization algorithms are implemented
  - ...here: Iterated F-Race (Lopez-Ibanez et al., 2011) is used

Iterated F-Racing

Algorithm

1. Initialize set of parameters θ
2. Assume stochastic Process i ~ P
3. ...where instance i = train/test split
4. Repeatedly evaluate f(i, θ)
5. ...until statistical test on E(f(i, θ)) allows removing some parameter candidates θ
6. Sample new parameters according to distributions centered at elite candidates
7. Reduces variance in later generations => convergence

...somewhat heuristic but often works well



Availability of Credit Scoring Data

- Class imbalance
- Systematic hyperparameter tuning
- **Benchmark experiment on large DB**
- Realistic evaluation

- Joint framework for classifier tuning and imbalancy correction
- Recent techniques like SMOTE and overbagging investigated
- Extension of SMOTE for categorical variables using Gower distance
- Iterated F-Racing instead of grid search for tuning within mlr
- **Experiment on large data base – validation on credit scoring data sets**
- Realistic evaluation on coarse classed data

- Real world CS data sets are often confidential
- ...thus most studies are limited to one or few data sets
- Popular studies of Baesens et al. (2003) / Lessmann et al. (2013) try to obtain generalized results using eight different real world scoring problems
- In order to obtain most generalizable results slightly different approach:
  - ...take large number of (unbalanced) real world problems (using data public data. e.g. from UCI, Bache and Lichman, 2014) to obtain valid results.
  - ...use several (three) real CS data sets to confirm results.

Description of the DB

Data	IR	N	Feat
vehicle	2.90	846	18
satellite image	9.28	6435	36
abalone7	9.68	4177	10
balance	11.76	625	20
us crime	12.29	1994	100
yeast ml8	12.58	2417	103
scene	12.60	2407	294
coil2000	15.76	9822	85
solar flare m0	19.43	1389	32
oil spill	21.85	937	48
yeast2vs8	23.10	482	8
wine quality4	25.77	4898	11
yeast uci me2	28.10	1484	8
ozone level	33.74	2536	72
yeast6	41.40	1484	8
mammography	42.01	11183	6
poker	58.40	1485	10
abalone19	129.53	4177	10

Data	IR	N	Feat
gcd (UCI)	2.33	1000	19
glc	11.91	28882	26
gmsc (kaggle)	13.96	150000	10

- 18 publicly available data sets with
  - ...different imbalance ratios
  - ...numbers of features
  - ...numbers of observations
- 3 credit scoring data sets:
  - German Credit Data (Hoffmann, 1994)
  - Kaggle GMSC competition
  - GfKl contest

Fair Comparison of Modern ML Techniques & Industry Standards

- Class imbalance
- Systematic hyperparameter tuning
- Benchmark experiment on large DB
- **Realistic evaluation**

- Joint framework for classifier tuning and imbalancy correction
- Recent techniques like SMOTE and overbagging investigated
- Extension of SMOTE for categorical variables using Gower distance
- Iterated F-Racing instead of grid search for tuning within mlr
- Experiment on large data base – validation on credit scoring data sets
- **Realistic evaluation on coarse classed data**

„Academic bias“:

- Within grids often some algorithms higher parameterized ~> better chance
- Sometimes even implicit bias by author's preferred algo
- Reproducible experiments: typically no separate preprocessing ~> advantage for black box algos
- Typically no integration of expert knowledge ~> would improve easier algorithms
- ...compare results from ML experiment to logistic regression on manually pre-binned data

Experiment

Experimental Design:

- Experiment on HPC
- 400 iterations of F-race
- Joint optimization of tuning / weighting / sampling strategies
- Optimization criterion: AUC (… no focus on comparison of difft performance measures)
- Outer 5 fold CV for performance evaluation
- Manual binning for Logistic Regression based on RPART decision trees using difft complexity parameters (Therneau and Atkinson, 1997)

Summary Learners

Learner	Abbr.	Tuning Parameters with (lower upper)
Gradient Boosting	gbm	n.trees (100, 5000) interaction.depth (1, 5) shrinkage (1e-05, 0.1) bag.fraction (0.7, 1)
RBF SVM	ksvm	C (2^{-12}, 2^{12}) sigma (2^{-12}, 2^{12})
Logistic Regression	logreg	-
Random Forest	RF	ntree (10, 500) mtry (1, 10)
Decision Tree	rpart	cp (0.0001, 0.1) minsplit (1, 50)

Results: Strongest Improvements

Data	Model	Base	Tuning	Data	Model	Tuning	Imbal	
balance	gbm	0.29	0.89	poker	rpart	0.47	0.76	sm
poker	gbm	0.53	1.00	abalone19	rpart	0.56	0.81	ob
balance	ksvm	0.68	0.92	balance	rpart	0.50	0.73	ob
mammogr.	gbm	0.71	0.94	balance	logreg	0.29	0.50	us
gmsc	gbm	0.66	0.87	solar fl. m0	ksvm	0.62	0.82	sm
satellite	gbm	0.78	0.97	ozone level	rpart	0.67	0.84	ob
abalone7	rpart	0.50	0.67	poker	logreg	0.34	0.52	us
vehicle	gbm	0.69	0.86	abalone7	rpart	0.67	0.83	os
coil2000	rpart	0.50	0.66	oil spill	rpart	0.70	0.85	ob
glc	rpart	0.70	0.85	balance	RF	0.36	0.50	ob

Results: Average and Maximum Improvement | Learner

Learner	Tuning	Imbal		Method
	Mean	Mean	Max	
gbm	0.14	0.02	0.04	ob
ksvm	0.05	0.04	0.06	us
logreg	0.00	0.05	0.13	us
RF	0.00	0.04	0.14	ob
rpart	0.04	0.13	0.29	sm

Results on DB

Data	Learner	Base	Tuning	Weights	Sampling	Method
vehicle	ksvm	0.868	0.926	0.926	**0.935**	os
satellite image	ksvm	0.935	0.967	0.965	**0.968**	sm
abalone7	ksvm	0.774	0.85	0.865	**0.87**	os
balance	ksvm	0.68	**0.917**	0.857	0.765	os
us crime	ksvm	0.87	0.927	0.925	**0.928**	sm
yeast ml8	ksvm	0.592	0.605	**0.619**	0.605	sm
scene	RF	0.763	0.783	0.804	**0.815**	os
coil2000	gbm	0.685	0.758	0.762	**0.762**	os
solar flare m0	ksvm	0.628	0.619	0.814	**0.818**	sm
oil spill	RF	0.93	0.923	0.907	**0.948**	sm
yeast2vs8	RF	0.927	0.847	0.906	**0.929**	os
wine quality4	RF	0.898	0.898	0.874	**0.907**	sm
yeast uci me2	RF	0.93	0.923	0.915	**0.934**	os
ozone level	ksvm	0.845	0.886	0.915	**0.916**	os
yeast6	gbm	0.903	0.945	0.944	**0.954**	sm
mammography	gbm	0.708	0.943	**0.953**	0.949	os
poker	gbm	0.525	0.998	**1**	1	os
abalone19	logreg	0.816	0.816	0.816	**0.842**	os

Results on CS Data Sets

Data	Learner	Base	Tuning	Weights	Sampling	Method
gcd	RF	**0.798**	0.792	0.781	0.787	os
gcd pre-binned	logreg	**0.787**	0.787	0.787	0.784	os
glc	gbm	0.788	0.922	**0.924**	0.922	os
glc pre-binned	logreg	**0.909**	0.909	0.909	0.909	os
gmsc	gbm	0.656	0.865	**0.866**	0.864	os
gmsc pre-binned	logreg	0.86	0.86	0.86	**0.861**	os

Conclusions

- Some strong improvements achieved, mostly by oversampling strategies
- Sometimes already tuning helps (w/o any additional imbalance correction)

- Tree based methods most sensible to imbalance correction
- Strongest effect of parameter choice for Boosting
- Traditional Logistic Regression quite insensitive to class imbalance (cf. also Hand, 1993, Brown and Mues, 2012)

- Only small improvements of modern ML compared to manually preprocessed Logistic Regression
- ...but: even small but significant improvements might have strong business impact!
- Modern techniques allow for automatic detection of nonlinear relationships also between explanatory variables (cf. e.g. Fahner, 2013, Sharma, 2013)

Disclaimer

The opinions expressed in this presentation are those of the author and do not necessarily reflect views of Santander Consumer Bank Germany.

Banco Santander S.A. and Santander Consumer Bank advise that this presentation contains statement on forecasts and estimates. Such forecasts and estimates appear in several sections of this document and include, among other things, comments on the development of future business and future returns. Whilst these forecasts and estimates reflect our judgement on future business expectations, if is possible that certain risks, uncertainties and other relevant factors may mean that the results are materially different from those expected. Such factors include: (1) the market situation, macroeconomic factors, regulatory and governmental guidelines; (2) movements in domestic and international stock markets, exchange rates and interest rates; (3) competitive pressure; (4) technological developments; (5) changes in financial position or credit value of our customers, debtors or counterparties. The risk factors and other basic factors we have indicated could adversely affect our business and the performance and results described and contained in our past reports or in those to be presented by us in the future, including those sent to the regulatory and supervisory authorities, including the U.S. Securities Exchange Commission.

N.B.- The information contained in this publication is not audited. However, the consolidated accounts have been drawn up according to generally-accepted accounting standards.

References

- Bache, K. and Lichmann, M (2013): UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences
- Baesens B., Van Gestel T., Viaene S., Stepanova M., Suykens J., Vanthienen J. (2003): Benchmarking state of the art classification algorithms for credit scoring, Journal of the Operational Research Society 54 (6), 627–635
- Bischl, B., Lang, M., Richter, J., Bossek, J, .Judt, L., Kuehn, T., Kotthoff, L., Jones, Z. (2015). mlr: Machine Learning in R., R package v.2.4.
- Brown I., Mues C. (2012): An experimental comparison of classification algorithms for imbalancedcredit scoring data sets, Expert Systems with Applications 39 (3), 3446–3453
- Chawla N.V., Bowyer K.W., Hall, L.O., Kegelmeyer,W.P. (2002): SMOTE: Synthetic minority oversamplingtechnique, Journal of Artificial Intelligence Research 16, 321–357
- Crone S., Finlay S. (2012): Instance Sampling in Credit Scoring: an empirical study of sample sizeand balancing, International Journal of Forecasting 28 (1), 224–238
- Fahner, G. (2013): Imposing Domain Knowledge on Algorithmic Learning - A New Approach to Construct Robust and Deployable Predictive Models, Talk at CSCC XIII.
- Hand, D., Henley, W. (1993): Can Reject Inference Ever Work? IMA 5(1) 45-55.
- Hastie, T., Tbishirani, R. and Friedman, J. (2001): The Elements of Statistical Learning. Springer, NY.
- Hoffmann, H. (1994): German Credit Data Set (Statlog), http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
- Lessmann S., Seow H.-V., Baesens, B., Thomas, L.C. (2013): Benchmarking state-of-theart classification algorithms for credit scoring: A ten-year update, http://www.business-school.ed.ac.uk/waf/crc_archive/2013/42.pdf
- Lopez-Ibanez, M., Dubois-Lacoste, J., Stützle, T., Birattari, M. (2011): The irace Package: iteratedracing for automatic algorithm configuration, Technical Report TR/IRIDIA/2011-004, IRIDIA, Bruxelles
- Sharma, D. (2013): Improving Logistic Regression/Credit Scorecards Using Random Forests, Talk at CSCC XIII.
- Szepannek, G., Gruhne, M., Bischl, B., Krey, S., Harczos, T. Klefenz, F. and Weihs, C. (2010): Perceptually Based Phoneme Recognition in Popular Music, in Locareck-Junge, H. and Weihs, C. (eds.): Classification as a Tool for Research, Springer, 751-758/offer/de/datamining_2011?page=download.
- Therneau, T., Atkinson, E. (1997): An introduction to recursive partitioning using the rpart routines. http://www. mayo. edu/hsr/techrpt/61.pdf.
- Vincotti T., Hand D. (2002): Scorecard construction with unbalanced class sizes, Journal of the Iranian Statistical Society 2, 189–205

Thank you!

Santander

CONSUMER BANK

Dow Jones Sustainability Indexes
Member 2008/09

FTSE4Good Index Series