ORIGINAL RESEARCH

A simple and robust scoring technique for binary classification

Charles Gomes *[1], Hicham Noçairi[1], Marie Thomas[1], Jean-Francois Collin[1], Gilbert Saporta[2]

[1] *L' Oréal, R and D, Aulnay sous Bois, France.*
[2] *Chaire de Statistique Appliquée and CEDRIC, CNAM, France.*

Received: November 21, 2013 **Accepted:** January 21, 2014 **Online Published:** February 26, 2014
DOI: 10.5430/air.v3n1p52 **URL:** http://dx.doi.org/10.5430/air.v3n1p52

Abstract

A new simple scoring technique is developed in a binary supervised classification context when only a few observations are available. It consists in two steps: in the first one partial scores are obtained, one for each predictor, either categorical or continuous. Each partial score is a discrete variable with 7 values ranging from -3 to 3, based upon an empirical comparison of the distributions for each class. In a second step the partial scores are added and standardised into a global score, which allows a decision rule.

This simple technique is successfully compared with classical supervised techniques for a classical benchmark and has been proved to be especially well fitted in an industrial problem.

Key Words: Binary outcome, Prediction, Scores, Small learning set

1 Introduction

A large number of prediction models has been developed by statisticians and researchers from the machine learning community, in order to predict a binary outcome: logistic regression, decision trees, bayesian networks, support vector machines (SVM) etc.[1] which have proved their efficiency in many cases. However most of them need parameter estimation and a large enough number of observations.

Indeed, in the industry the cost of one experience is often very expensive, so the number of data used for the construction of a prediction model is often small. Most of the above mentioned prediction models have poor performances on small samples.

The objective of this paper is to propose a new approach for the prediction of a binary outcome in the case of small samples and with a small number of variables.

Let Y be the binary response with categories A and B, and X_j, j=1, ...,p the predictors.

Our technique is based on an additive score, made of sub-scores: each sub-score associated to a single predictor X is a discrete variable with seven integer values from -3 till +3 depending on how well a predictor discriminates between the 2 outcomes of the response variable Y. In other words the sub or partial scores define (possibly non linear) transformations of the predictors. The set of values from -3 till +3 is inspired by the normal distribution and six sigma ideas. The way of coding predictors, either categorical or continuous, is generic and does not depend on a specific data set.

The paper is organized as follows: After a presentation of the way of obtaining sub-scores, several applications are made on a well known benchmark and compared with classical techniques. The method is then applied to a problem coming from the cosmetics industry.

2 A new "Score method"

2.1 Principle

Like Boosting where a set of weak learners may produce a single stronger learner, our method obtains a strong prediction rule based upon a linear combination of weak rules, each rule being associated to a single predictor by a simple scoring technique. Here the combination is an unweighted

* **Correspondence:** Charles Gomes; Email: cgomes@rd.loreal.com; Address: L'Oréal, 1 avenue Eugène Schueller, BP22, 93601 Aulnay sous bois, France.

mean of the partial scores, instead of a weighted mean, which is not an issue since a non discriminant variable will have partial scores concentrated around zero.

2.2 Scoring

Partial scores are obtained in different ways according to the nature (qualitative or quantitative) of each predictor. Let us denote by A and B the two categories of the response variable Y.

2.2.1 Scoring a qualitative predictor

For a qualitative variable X, a histogram is used to define the importance of each category.

The method is based on the following simple idea: if the frequency of "A" is larger than 3 times the frequency of "B" in a category M of X, we give to M the highest score, i.e. 3. A symmetric rule is applied for negative values.

More generally, the score given to a category M depends on the ratio of the frequencies of both classes of the outcome Y conditionally to this modality: $\frac{n_{AM}}{n_{BM}}$.

Actually, for each category M, we define a positive A-score SAM for class A and a negative B-score SBM for class B. However the confidence depends on the total number of observations, and we will be more severe if this number is low. For a sample size lower than 20, we use the scores given by table 1, and the following formulas for larger sizes. In case where the frequencies of the modalities would be less than 5, the score is put to 0.

The score S_{AM} of class A corresponds to:

-S_{AM}=0, if $\frac{N_{AM}}{N_{BM}} \leq 2$;

-S_{AM}=1, if 2<$\frac{N_{AM}}{N_{BM}} \leq 2.5$;

-S_{AM}=2, if 2.5<$\frac{N_{AM}}{N_{BM}} \leq 3.5$;

-S_{AM}=3, if 3.5<$\frac{N_{AM}}{N_{BM}}$

The score S_{BM} of the class B corresponds to:

-S_{BM}=0, if $\frac{N_{BM}}{N_{AM}} \leq 2$;

-S_{BM}=1, if 2<$\frac{N_{BM}}{N_{AM}} \leq 2.5$;

-S_{BM}=2, if 2.5<$\frac{N_{BM}}{N_{AM}} \leq 3.5$;

-S_{BM}=3, if 3.5<$\frac{N_{BM}}{N_{AM}}$

All previous values have been empirically validated for a large number of data sets.

Figure 1 shows the coherence between the table and the formulas. In Figure 1, the x-axis is the frequency of the category M and the y-axis is the ratio $\frac{n_{AM}}{n_{BM}}$.

The range of the scored variable reflects its discrimination performance. So, a variable where all its modalities are coded 0 will not be influential.

Example: let X be a variable with 3 categories. As it is shown in Figure 2, category 1 has a majority of B, with a ratio $\frac{n_{B1}}{n_{A1}}$ between 2 and 2.5 so the A-score is equal to zero and the B-score equal to -1. In category 2 of X, both frequencies are equal, hence the A-score and the B-score are both equal to zero since there is no discrimination, etc.

The final score for a statistical unit is the sum of its A- and B- scores.

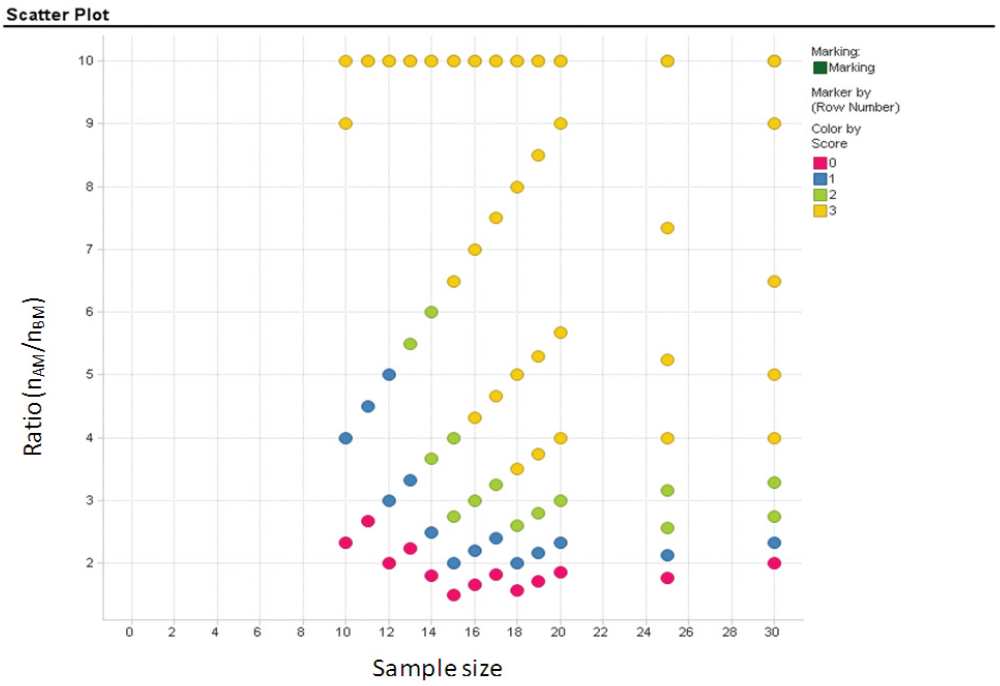


Figure 1: Scores according to frequency nM and ratio $\frac{n_{AM}}{n_{BM}}$

Table 1: Score Table.

Score\N	10	11	12	13	14	15	16	17	18	19	20
1	8/2	9/2	10/2	10/3	10/4	10/5	11/5	12/5	12/6	13/6	14/6
			9/3								
2	9/1	10/1	11/1	12/1	11/3	12/3	12/4	13/4	13/5	14/5	15/5
				11/2	12/2	11/4					
	10/0	11/0	12/0	13/0	14/0	15/0	16/0	17/0	18/0	19/0	20/0
					13/1	14/1	15/1	16/1	17/1	18/1	19/1
3						13/2	14/2	15/2	16/2	17/2	18/2
							13/3	14/3	15/3	16/3	17/3
									14/4	15/4	16/4

2.2.2 Scoring a continuous predictor

Like for qualitative variables, we give integer scores between [-3, 3], positive for class A and negative for class B. The score of a unit is the sum of its A- and B- scores. It is necessary to split the quantitative variable into classes: for this purpose we project the quantiles of the two boxplots corresponding to both classes of the response, as described in Figure 3.

Then, we obtain a new variable which is composed of seven levels: this transformation allows us to use the scoring methodology defined in 2.2.1 for each class.

One advantage of this transformation is that it allows to take into account non linear phenomena.

Remark: if some modalities have a null score, we may merge extreme adjacent modalities e.g. 1 and 2, or 6 and 7. In a similar way, if a category has less than 15 observations, it should be merged with adjacent ones, in order to have enough observations.

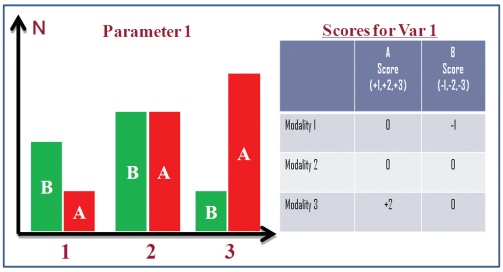
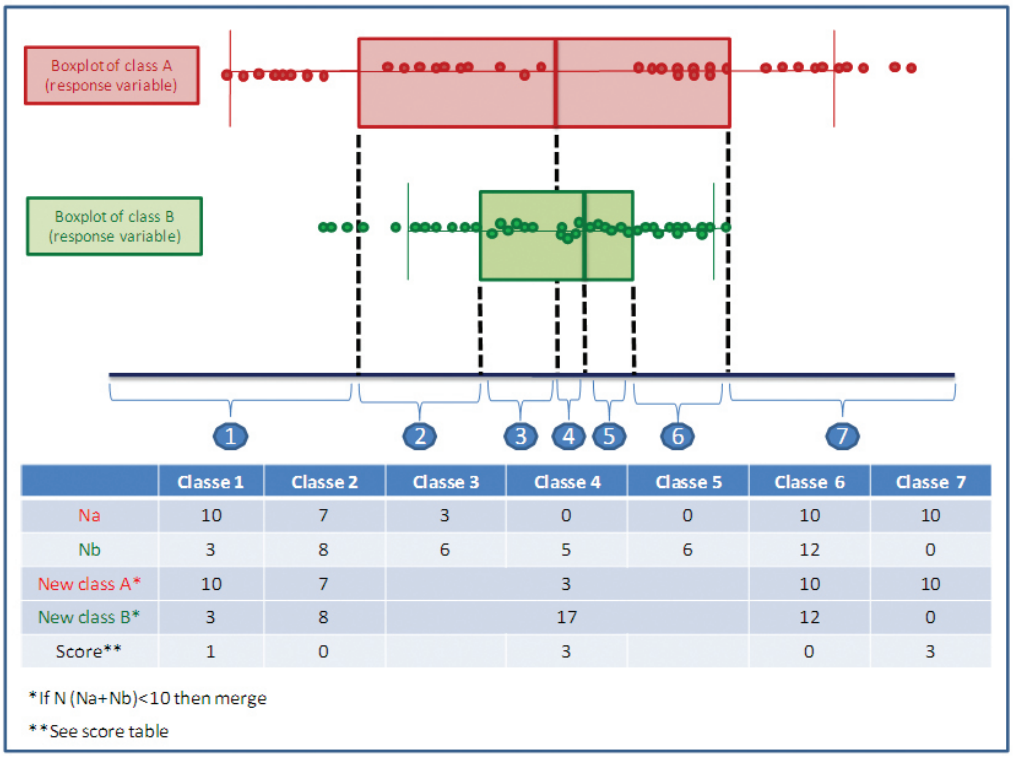


Figure 2: Example of a score for a categorical variable



	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6	Classe 7
Na	10	7	3	0	0	10	10
Nb	3	8	6	5	6	12	0
New class A*	10	7		3		10	10
New class B*	3	8		17		12	0
Score**	1	0		3		0	3

Figure 3: Example of a score for a continuous variable

2.3 Global score, thresholds

For each individual, all its sub-scores may be displayed in a radar chart and the global score is the sum of partial scores. The global score X is finally linearly standardised within the scale [0, 100] by the transformation $Y = 100\frac{X-a}{b-a}$, where a and b are respectively the maximal and the minimal value of X. This transformation is useful if we like to compare our score to scores or posterior probabilities obtained by other classification techniques. In the following, we will consider (abusively) our score as a probability multiplied by 100.

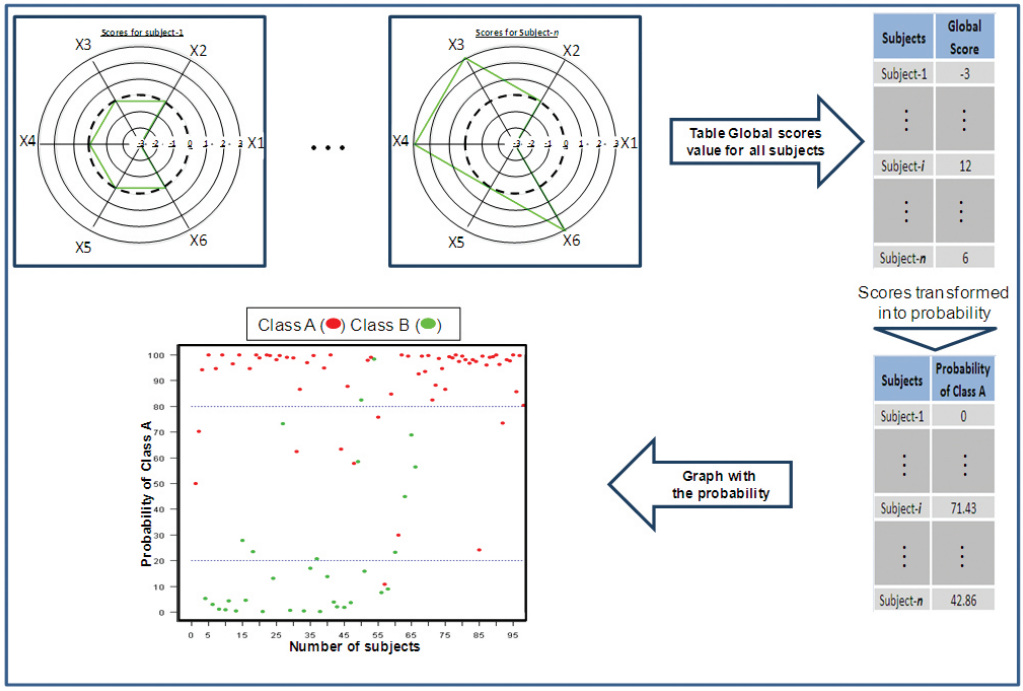


Figure 4: Methodology scheme

Another important point on the methodology is that we define a no-decision area. If for some unit we get a probability which lies between two thresholds (eg 40 and 55) we consider that we have not enough informations to conclude.

More precisely, the no-decision zone is obtained as follows: The first step is to begin in the central zone defined by [45 and 55] with a strong entropy (both groups are strongly mixed), then we increase the superior bound by steps of 5 until we approach a strong density zone and so a weak entropy (because we incorporate many individuals of the group A) of group A. In the same way we decrease the lower bound until we incorporate too many units from the group B.

If we observe an increase of the entropy linked to an increase of the sensibility (or of the specificity) we stop at the previous step. We do the same for class B: if we observe a strong increase of the specificity, we stop at the previous step.

Anyway, the percentage of inconclusive observations must not be larger than 40%.

Figure 4 summarizes the final steps of the methodology, once the subscores have been obtained for each predictor.

3 Application to a classical data set

We have applied our score methodology to the well-known data set "South African Coronary Heart Disease" used e.[1] This set is taken from a larger dataset, described in Rousseauw et al, 1983.[2] It is a retrospective sample of 462 males between 15 and 64 in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of Coronary Heart Disease (N=462), described by 9 attributes (8 quantitative and 1 binary). The response variable MI is the presence or absence of myocardial infarction.

As described in part 2, we have performed the score model on a training set (N=340) and evaluated it on the validation set (N=122). As shown in Figure 5, scores have been determined for 6 of the 9 attributes. For all the selected attributes, the seven classes have been merged into two classes, giving only one threshold per attribute.

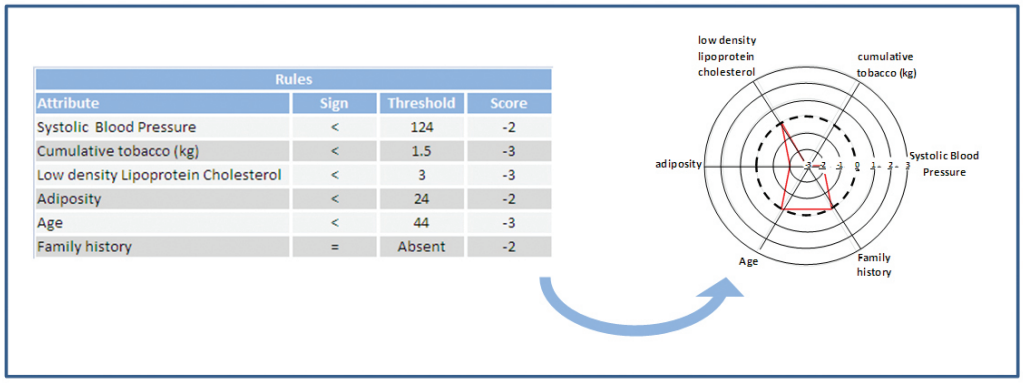


Figure 5: Details on score model, and example of a radar obtained for a patient

Note that only negative scores are present. It indicates that the model will be more confident for absence of CHD prediction than for presence of CHD.
Results for prediction are illustrated in Figure 6.

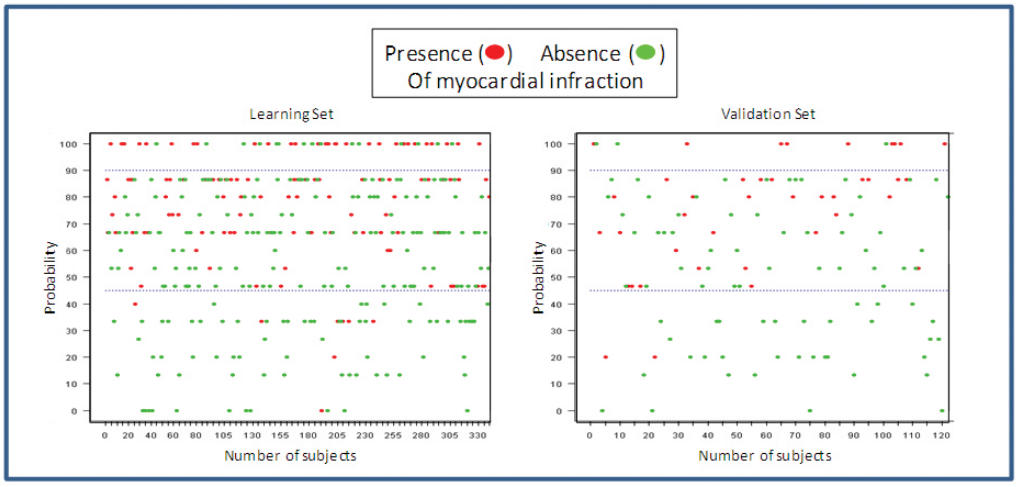


Figure 6: Probability prediction for learning set (left) and validation set (right)

On the learning set, we can determine confidence thresholds: as mentioned before, only negative scores are possible in this example. Taking into account this dissymmetry, the 45% threshold for absence of CHD is retained. At the opposite, only very high probability (higher than 90%) for presence of CHD will be retained.
To challenge this methodology, we have also used two other models: Sparse PLS -Discriminant analysis among linear models, and Support Vector Machines among non linear models. Sparse PLS-DA is a adaptation of the Sparse-PLS proposed by Chun, H. and Keles, S., (2010)[3] when the response is binary. SVM have been proposed in Cortes, C. and Vapnik, V. (1995).[4]
For these two methods, the 20% threshold for absence of CHD is retained. At the opposite, probability higher than 80% for presence of CHD will be retained.

In order to compare performances of the three methods, sensitivity (capacity to predict presence of CHD), specificity (capacity to predict absence of CHD), accuracy (proportion of true results) and kappa (percentage of the maximum agreement, corrected by what it would be under randomness) have been calculated for learning set and validation set, and are presented in Table 2.

As shown in Table 2, the model that provides the highest performance in terms of the kappa coefficient is our Score model.

Furthermore, they are complementary: Score model is the best one in terms of sensitivity, and SVM and sparse PLS-DA are excellent for specificity.

Table 2: Performance comparisons on the learning set (N=340) and validation set (N=122)

	Score Score		**SVM**		**Sparse PLS-DA**	
	Learning set	**Validation Set**	**Learning set**	**Validation Set**	**Learning set**	**Validation Set**
% predicted	37%	38.5%	33.5%	34.5%	36%	38.5%
True positive	30	**9**	10	3	10	3
False positive	15	3	2	**0**	2	**0**
False negative	8	2	10	4	13	5
True negative	72	33	92	35	98	**39**
Sensitivity	79%	**82%**	50%	43%	44%	37.5%
Specificity	83%	92%	98%	**100%**	98%	**100%**
Accuracy	82%	89%	90%	**90.5%**	88%	89%
Kappa	0.59	0.71	0.57	0.55	0.51	0.5

4 Application to the screening of Raw Material (RM) in cosmetics industry

4.1 Statutory context

When a screening test is expensive, in terms of money, time or resources, and thus has a low throughput, an upstream selection based on prediction of the screening results is highly beneficial. In other words, rationalization is required in order to test only the compounds that would have the best chance to give a positive result in the screening. In this context, our objective was to characterize the profile(s) of an ideal RM giving a favorable answer in the in vitro test, by combining data that can be generated inexpensively with virtually no limitation in the size of the set of compounds in which selection is required such as simple *in silico* models or *in vitro* HTS platforms.

4.2 Experiments on data sets

In our example, we considered a set of 74 RM characterized by:

– calculated physico-chemical data: 7 quantitative variables, 2 qualitative variables;

– preliminary efficiency measured in vitro:2 quantitative variables.

The list of these variables being industry confidential cannot be detailed here.

The 74 RM have been randomly split into a learning set of 49 R and a validation set of 25 RM.

4.3 Results

Like in the previous example, we have used successively SPARSE PLS-DA, SVM, and the Score method. Figures 7, 8, 9 show the threshold choices. Performances are summarized in Table 3.

Table 3: Performances of the 3 methods for learning (N=49) and validation (N=25) sets

	Score Score		**SVM**		**Sparse PLS-DA**	
	Learning set	**Validation Set**	**Learning set**	**Validation Set**	**Learning set**	**Validation Set**
% predicted	69%	80%	96%	100%	53%	64%
True positive	22	**6**	24	5	16	4
False positive	6	**1**	8	2	4	3
False negative	0	**2**	2	6	2	4
True negative	6	11	13	**12**	4	5
Sensitivity	100%	**75%**	92.3%	45.45%	88.8%	50%
Specificity	50%	**91.6%**	61.9%	85.7%	73.3%	62.5%
Accuracy	82%	**85%**	78.7%	68%	81.8%	56.25%
Kappa	0.56	**0.68**	0.56	0.32	0.63	0.125

The score method gives a prediction model with better performances on the validation set than the two other methods (Sparse PLS-DA & SVM), while keeping 80% of the data.

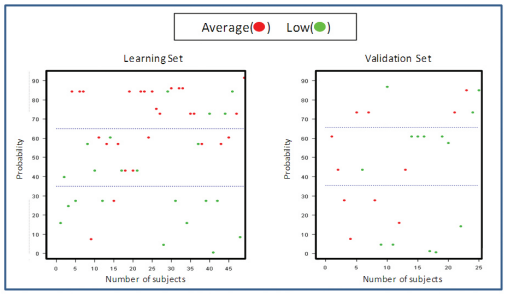


Figure 7: Sparse PLS DA

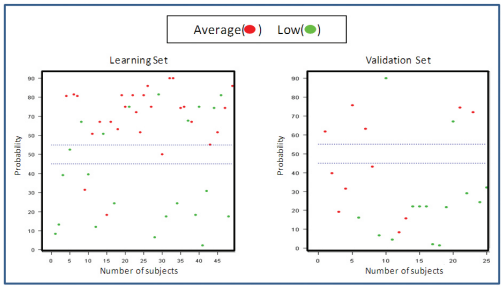


Figure 8: SVM

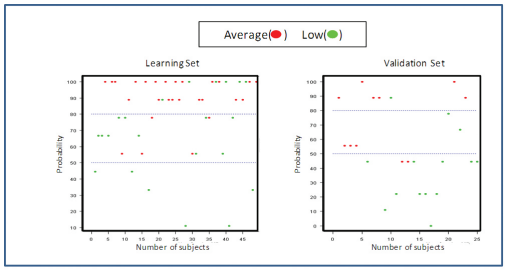


Figure 9: Score method

5 Conclusion and perspectives

We have presented a simple scoring technique which has good performances compared to more sophisticated techniques, especially in the case of small data sets. An advantage is that this model can treat as well qualitative as quantitative variables. However we do not claim that our method is the most efficient, but that its ratio= efficiency/simplicity is very high. Instead of being a competitor to other methods, our score method may be combined with other ones by ensemble techniques like model averaging or stacking to give improved predictions, as it has been verified in real applications.[5]

Acknowledgements

Fabian Ibanez from Keyrus company, took an important part in the practical implementation of this methodology.

References

[1] Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition, Springer. 2009.

[2] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. et al. Coronary risk factor screening in three rural communities, South African Medical Journal. 1983; 64: 430–436.

[3] McCullagh, P. Regression models for ordinal data, Journal of the Royal Statistical Society. 1980; 42(2): 109-142. http://dx.doi.org/10.1109/ICDAR.2001.953882

[4] Cortes, C. and Vapnik, V. Support-vector network. Machine Learning. 1995; 20(3): 273-297.

[5] Gomes, C., Noçairi, H., Thomas, M., Collin, J.F., Ibanez, F. , Saporta, G. Stacking prediction for a binary outcome, in COMPSTAT 2012, 20th International Conference on Computational Statistics, Limassol, Cyprus. 2012: 271-282.

[6] Cortes, C. and Vapnik, V. Support-vector network. Machine Learning. 1995; 20(3): 273-297.