First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4245

Weighting machine learning solutions by economic and institutional context for decision making

Alvarez-Jareño, Jose A.[a] and Pavía, Jose M.[b]

[a] Departament of Economics, Universitat Jaume I, Spain.

[b] Department of Applied Economics, Universitat de València, Spain.

Abstract

It is quite common that machine learning approaches reach high accuracy forecast rates in imbalanced datasets. However, the results in the category with few instances are usually low. This paper seeks to improve the results obtained applying different techniques (such as bagging, boosting or random forests) with the inclusion of cost matrices. We propose applying the actual costs incurred by the company for misclassification of instances as a cost matrix. This approach, along with an economic analysis of the different solutions, makes it possible to incorporate a business perspective in the decision making process. The approach is tested on a publicly available dataset. In our example, the best ratings are obtained by combining the cost matrix with random forests. However, our analysis shows that the best technical solution is not always the best economical solution available. A company cannot always implement the optimal solution, but has to adopt a solution constrained by its social, institutional and economic context. Once an economic analysis is carried out, it seems the final decision of the company will depend on its economic situation and its institutional policy.

Keywords: *imbalanced datasets, random forest, cost matrix, economic analysis, uplift modeling.*

The authors wish to thank M. Hodkinson for translating into the English the text of paper and the support of the Spanish Ministry of Economics and Competitiveness through grant CSO2013-43054-R.

1. Introduction

This paper takes as reference Moro et al. (2014) and proposes a set of methods of statistical learning (logistic regression [LG], decision trees [DT], neural networks [NN] and support vector machine [SVM]) to analyse and subsequently predict the response of clients of a bank to a telephone marketing campaign.

As suggested by Radcliffe and Surry (2011), three types of models, which have continued evolving over time, can be identified:

- Entry models, which aim to profile those clients that are consumers of the product. These models answer the question “Who is buying?”.
- Purchasing models, which aim to profile customers that have recently bought a product. As well as responding to the question “Who is buying?”, the model also looks at “When does the purchase take place?”.
- Response models, whose objective is to profile customers that have bought the product, ostensibly in response to a marketing campaing. These types of models look for answers to “Who is influenced by the marketing campaign in question?”.

The model used by Moro et al. (2014) is a purchasing model and as such uses a unique set of data without a control group. A control group would be necessary to be able to estimate an uplift model, as shown by Guelman et al. (2015).

The data used are available on the webpage of UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

2. The problem

The objective of the modelling is to forecast a dichotomous response variable, with a reliable degree of probability, by way of 20 predictors. The purpose of the forecast is to increase the success rate among the people contacted, avoiding contacting those clients who have a lower or null probability of opening a term deposit.

To be able to analyse the predictive capacity of the models the initial set of data was divided into two subsets, where the instances that belong to each set were randomly selected. The first subset (of learning or training) makes up 80% of the data and will be used to train the models.

Table 1 shows the predictive capacity of conventional processes analysed with default options. All the methods classify correctly in 90% of the instances, with the differences between them being minimal. The data set, however, is clearly imbalanced. The percentage of instances with “No” responses is 88.73%, so a naïve classifier would obtain similar percentages of overall success.

Table 1. Predictive capacity of the models.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**
Correctly Classified	90.75%	90.67%	90.17%	89.9%
Incorrectly Classified	9.25%	9.33%	9.83%	10.1%
True Positive (TP) Rate	0.973 (no)	0.957 (no)	0.966 (no)	0.979 (no)
	0.421 (yes)	0.533 (yes)	0.427 (yes)	0.303 (yes)
False Positive (FP) Rate	0.579 (no)	0.467 (no)	0.573 (no)	0.697 (no)
	0.027 (yes)	0.043 (yes)	0.0034 (yes)	0.021 (yes)

Source: Own elaboration from Weka ouputs.

The TP rate of the methods in the target category (yes) is, however, very low. On average, only 42.1% of the clients who contracted the product are properly targeted. The objective of these models should be to obtain a higher rate of true positives in the target category, even though the false positive (FP) may be high in the complementary category or the ROC area may be lower. From this perspective, the decision tree would be preferable to the other models.

Regarding the measures of goodness of fit, there is not consensus. Whilst decision trees obtain the best measure of the Kappa statistic, logistic regression reaches a higher value in the ROC area. The lowest average absolute error is obtained using support vector machine.

Table 2. Model assessments.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**
Kappa statistic	0.4715	0.5242	0.4555	0.3696
ROC Area	0.9330	0.8940	0.8750	0.6410
Mean Absolute Error	0.1256	0.1197	0.1031	0.1010

Source: Own elaboration from Weka ouputs.

3. Methodologies to improve the results

The approaches analysed so far have not taken into account the fact that there is an imbalance in the data. This can be seen in the poor results obtained in predicting the uptake by customers of the product. The main objective for constructing these models is to identify with greater accuracy those clients who would open a term deposit. In other words, which bank clients could be more easily encouraged to contract this product.

The imbalanced data sets need to introduce techniques which allow correct identification of the minority response, the solution to this problem being found in two distinct levels: at data level and algorithm level.

At the data level, the proposed solutions include different ways of model ensemble and resampling. Its aim is to ensure the predictions generated are robust. The different ensemble approaches are outlined by Rokach (2009). One of the first model assembly systems was bagging, proposed by Breiman (1996) and Buhlmann and Yu (2002) and implemented in R by Spanish researchers, Alfaro et al. (2013). Among the resampling techniques, we can found boosting, in particular, the algorithm AdaBoost M1, introduced by Freund and Schapire (1997) and extensively assessed in many studies and analyses, most notably by Eibl and Pfeiffer (2002) and Meir and Rätsch (2003). Random forests is another resampling method developed by Breiman (2001) as a variant of the bagging methodology using decision trees.

At the algorithm level, solutions include adjustments to the costs of various classes in order to counteract the imbalanced class: cost matrix. The aim is to adjust the probabilistic estimate of decision tree leaves (when working with decision trees), to adjust the decision threshold and to base learning on recognition rather than discrimination.

4. Results obtained with the proposed methodologies

After applying the different techniques described above, the following results were obtained:

- At data level, the different ensemble techniques (bagging, boosting and random forest) improve the robustness of the results, but without significantly improving the predictive capacity of the models.
- At algorithm level, the application of the cost matrix enables better identification of true positives. However, false positives are also increased, reducing the global accuracy of the model.

Tables 3 and 4 show the results obtained for the models in Table 1 with the application of the cost matrix. Table 4 also presents the evaluation of the model for the methodology of random forests with cost matrix. This approach is the one showing the greatest predictive capacity in the target category.

While the results of the quality of adjustment have worsened overall (see Table 4), the predictive capacity of all the models has improved for the target category (yes), even reaching levels of around 97% for Logistic Regression (see Table 3).

Table 3. Predictive capacity of the models with cost matrix.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**
Correctly Classified	76.57%	80.79%	88.02%	79.49%
Incorrectly Classified	23.23%	19.21%	11.98%	20.51%
True Positive (TP) Rate	0.738 (no) 0.969 (yes)	0.794 (no) 0.912 (yes)	0.903 (no) 0.709 (yes)	0.772% (no) 0.965 (yes)
False Positive (FP) Rate	0.031 (no) 0.262 (yes)	0.088 (no) 0.206 (yes)	0.291 (no) 0.097 (yes)	0.035 (no) 0.228 (yes)

Source: Own elaboration from Weka ouputs.

These results can be further improved by using random forests [RF] with cost matrix, achieving TP rates of 0.771 in the “no” category and 0.972 in the “yes” category.

Table 4. Assessments of models with cost matrix.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**	**Random Forest**
Kappa statistic	0.3875	0.435	0.5169	0.4296	0.4314
ROC Area	0.854	0.853	0.806	0.869	0.872
Mean Absolute Error	0.2343	0.192	0.1198	0.205	0.2051

Source: Own elaboration from Weka ouputs.

5. Economic evaluation of the results

With so many results, and with often very little difference between them, it is difficult to decide which model to use to select customers worth contacting. To simplify this decision making process, the economic criteria of income and costs can be used.

Assuming that the experiment was real, and under each prediction approach, the contacts would be reduced to those customers likely to say “yes” according to the model; and no contact would be made with customers identified by the model as unlikely to buy the financial product. Therefore, the cost of the campaign would be the sum of the second column of the confusion matrix multiplied by the cost of each contact (5 units), and the income would be obtained by multiplying the number of clients that enter into a contract (amongst those who were contacted) and the average income of the bank for each contract (100 units). The difference would be the financial gain. The results are summarised in Table 5.

Table 5. Costs, income and ratios (over costs) of the campaign using different models.

Selection model	Costs	Income	Gain	Inc/Cost	Gain/Cost	Variation
No Selection	41,190	97,900	56,710	2.38	1.38	-
LG	3,035	41,200	38,165	13.58	12.58	-18,545
DT	4,165	52,200	48,035	12.53	11.53	-8,675
NN	3,335	41,800	38,465	12.53	11.53	-18,245
SVM	2,235	29,700	27,465	13.29	12.29	-29,245
RF	3,510	46,400	42,890	13.22	12.22	-13,820
LG-wCM	14,245	94,900	80,655	6.66	5.66	23,945
DT-wCM	11,945	89,300	77,355	7.48	6.48	20,645
NN-wCM	6,980	69,400	62,420	9.94	8.94	5,710
SVM-wCM	13,000	94,500	81,500	7.27	6.27	24,790
RF-wCM	13,075	95,200	82,125	7.28	6.28	25,415

Source: Own elaboration.

If no selection were made, and all clients were called, as is the norm in many real campaings, the cost would be 41,190 units, the income 97,900 and, consequently, the gain would be 56,170 units.

If a selection of people to contact had been made using the traditional models as proposed by Moro et al. (2014), the costs of the marketing campaign would have been significantly reduced, to a tenth of the original costs. However, there would have been a substantial reduction in the income of the financial institution and, consequently, in the gains. The highest income and costs had been achieved with the decision tree model, which shows the smallest losses compared to the situation where no selection was made.

The results (relative to a general campaign) change to a positive sign when the cost matrix is applied, achieving greater gains in all cases. In this scenario, the marketing campaign costs had been higher than under the proposals of Moro et al. (2014), ranging from 6,980 units for model NN to 14,245 units for model LG, but with income increasing substantially. With respect to a universal campaign, the costs are reduced by up to a quarter and the outcome is a greater gain.

All financial institutions may not have the same target or the same restrictions. Some might want to minimise the costs of the campaign, others maximise the income or the gain. The strategy of the bank will determine which method is the best to use for the marketing campaign. In any case, from an economic perspective, using any of the models for the marketing campaign is still more efficient that not using one at all.

From the point of view of the strength of investment, we can compute the indices of income over costs or gains over costs to observe that the strategy of contacting all potential clients yields the lowest return on investment (1.38 units) with a cost almost even 13 times higher. Models without cost matrix would require an average investment of 3,200 monetary units, while models with cost matrix would need an average of 11,500 units. In other words, the necessary investment would be 3.6 times higher.

Although the task set for the models is the same in all cases, the best solution will depend on a series of financial variables and on the institutional policy adopted by the company. In the situation of the example studied, where capturing passive assets (by way of long term deposits) was one of the strategic objectives of Portuguese banks to improve the cash balances and to pass the European Central Bank stress tests, an increase in passive assets was favourable above a cost/gain analysis (Moro et al. 2014).

The selection of the model to be used for identifying the target audience of a marketing campaign will depend on the strategic objective set by the company, and could be very different for each situation. The best scientific solution is not always the best economic solution for a company, and having different options ensures the most appropriate strategic solution is selected.

6. Conclusions

The results obtained from the application of new models indicate that the inclusion of a cost matrix in the imbalanced sets significantly improves the classification of true positives to the detriment of true negatives. The other techniques used (boosting, bagging and random forest without cost matrix, and with or without using cross-validation, whose results have not been included due to lack of space), do not show any substantial improvement in the results obtained by Moro et al. (2014). Although it is true that they add robustness to the results, this does not always lead to an improvement and when there is, it is usually marginal.

From a scientific point of view, the best results are obtained combining the cost matrix with random forests. However, since the data is of an economic nature, the results should be approached from an economical-financial perspective. When entering costs and income as variables for decision making, we observe a variety of strategies to be taken by companies according to their economic situation and institutional policy.

The different technical solutions lead to different economic consequences, which in many cases need to fit in with the individual bank's circumstances. Faced with a set of economic restrictions not all the available technical solutions are viable and one should be chosen which either maximises or minimises a target within the capabilities of the company.

References

Alfaro, E., Gamez, M. & Garcia, N. (2013) adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, 54(2), 1-35.

Breiman, L. (1996). Bagging predictors. *Machine Learning*. 24(2), 123-140.

Breiman, L. (2001) Random Forests. *Machine Learning*. 45(1), 5-32.

Buhlmann, P., & Yu, B. (2002) Analyzing bagging. *Annals of Statistics,* 30, 927-961.

Eibl, G., & Pfeiffer, K. P. (2002). How to make AdaBoost. M1 work for weak base classifiers by changing only one line of the code. In *Machine Learning: ECML 2002* (pp. 72-83). Berlin-Heidelberg: Springer.

Freund, Y & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift random forests. *Cybernetics and Systems*, Vol. 46 (3-4), pp. 230-248.

Meir, R., & Rätsch, G. (2003). An Introduction to Boosting and Leveraging. In *Advanced Lectures on Machine Learning*, *LNAI 2600* (pp. 118-183). Springer.

Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, 62, 22-31.

Radcliffe, N., & Surry, P. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. *Stochastic Solutions Limited*. http://stochasticsolutions.com/pdf/sig-based-up-trees.pdf

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12), 4046-4072.