Post-stratification as a tool for enhancing the predictive power of classification methods

F.D. d'Ovidio, A.M. D'Uggento, R. Mancarella, E. Toma

1. Introduction

As is well known, any decision-making model involving classification algorithms often faces the problem of predictive or diagnostic power (sensitivity or specificity), which tends to decrease rapidly as the asymmetry of the target variable increases (Sonquist et al., 1973; Fielding 1977). For example, segmentation analyses with categorical target variables generally provide very little improvement in purity (or none at all) if the least represented category accounts for less than one-fourth of the cases of the most represented category. The same problem occurs with other theoretically more exhaustive techniques, such as artificial neural networks. In fact, the optimal situation for any classification analysis is the maximum uncertainty, namely the equal distribution of the target variable.

Certainly, some classification techniques are more robust, such as those based on a logit transformation of the target variable (Fabbris & Martini 2002), which is less sensitive to the distribution's shape. However, even this technique is affected by the distributive asymmetry of the target variable, as will be shown below.

Indeed, beginning from the results of a direct survey in which the target variable (binary) was highly asymmetric (12.3% versus 87.7%), the first analysis performed here shows that even logit models with very significant parameter estimates can have an insufficient fit and such low predictive power that they are useless in decision-making processes.

To address this prediction problem, we tested a post-stratification technique originally developed to solve classification problems by making a training sample that is artificially symmetrical in terms of the target variable's distribution.

In this way, a substantial increase in goodness of fit and predictive ability was achieved for both the symmetrized sample and, more importantly, for the original sample, whose probabilities of success are assessed by the parameters estimated by the model.

2. The case study

A sample of participants in a national survey on dietary habits was studied from December 2020 to the end of May 2021 (in continuation of similar surveys carried out since 2018), selecting only those who had regularly completed the proposed questionnaire, corresponding to 2,562 people residing or domiciled in Italy. One of the research topics was the tendency on the part of Italians to eat away from home, i.e, in restaurants or pizzerias, in view of the restrictions necessitated by the COVID-19 pandemic.

The target variable resulted from a question about the frequency with which subjects tended to eat outside home, distinguishing between sporadic customers (who did so, at most, occasionally) and those for whom eating at a restaurant was a usual habit. The percentage of the latter, which had never been very high in previous years, dropped sharply to zero during the pandemic period, but because the survey investigated (even retrospectively) the eating habits of respondents, the result was not quite so poor. However, considering that the pandemic had already affected the social habits of Italians prior to 2021, the response variable shows that

Francesco D. d'Ovidio, University of Bari Aldo Moro, Italy, francescodomenico.dovidio@uniba.it, 0000-0003-1641-039X
Angela Maria D'Uggento, University of Bari Aldo Moro, Italy, angelamaria.duggento@uniba.it, 0000-0001-9768-651X
Rossana Mancarella, ARTI, Agency for Technology and Innovation of Apulia, Italy, r.mancarella@arti.puglia.it, 0000-0001-8179-4970
Ernesto Toma, University of Bari Aldo Moro, Italy, ernesto.toma@uniba.it, 0000-0002-4817-7169

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

F.D. d'Ovidio, A.M. D'Uggento, R. Mancarella, E. Toma, *Post-stratification as a tool for enhancing the predictive power of classification methods*, pp. 125-130, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-461-8.24, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci (edited by), *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the on-site conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-461-8 (PDF), DOI 10.36253/978-88-5518-461-8

more than 87.7% of respondents (2,248 people) fall into the "non-customers" group and less than 12.3% (314 people) fall into the "restaurant lovers" group.

Despite this outstanding asymmetry, an investigation was conducted into the individual characteristics that were found to be related in some way to the target variable[1] and may explain the motivations for this tendency to eat away from home.

To this end, a common logistic regression model was first developed for exploratory purposes[2]. Such a model is not reported here, because it includes variables with insufficient or zero significance, although it has high statistical significance for the variables gender ($p<0.002$), work position ($p<0.001$) and food-delivery frequency ($p<0.001$); however, the model has minimal fit to the data (Cox-Snell $R^2 = 0.094$; Nagelkerke $R^2 = 0.179$) and minimal predictive power for the category of interest: only 28 cases were correctly identified as "habitual customers" (8.9% of actual cases).

The *correct identification* of "non-customers" (2,224 cases, that is, 98.9% of the subgroup), in contrast, is very relevant, but this result seems trivial.

The overall percentage of correctly identified cases is 87.9%, but it should be noted that, simply assigning the sampling mode "non-usual customer" to predict all cases would have resulted in 100% correct classifications of non-customers and, of course, no correct classifications of usual customers. In short, over 87.7% of cases could be correctly classified simply by assigning the mode value, without the need for complex statistical processing.

Estimating a more articulated logistic model, one that also included the more important interactions among the explanatory variables (but was made parsimonious by using the stepwise forward-deletion criterion, i.e., gradually inserting the most related variables and removing the non-significant ones), did not improve the result.

However, the final model is shown in Table 1, and it is interesting in its own way. The reference categories of the explanatory variables (referred to as *baselines* and shown in brackets after the names in the table) are generally identified with the first category, except for employment position.

This model (although better than the previous one, at least in terms of potential generalization due to the statistical significance found for many variables and items) is also affected by an overestimation of "non-habitual customers," as shown in Table 2. In fact, compared to the almost perfect classification of these (99.1%), few regular customers of the restaurant are also correctly classified, at only 27. Therefore, the overall correct classifications are almost entirely due to the predominance of "non-habitual customers" in the sample, for which the usefulness of the model for predictive and decision-making purposes remains very limited or practically zero.

[1] The following individual characteristics were considered: *Gender* (F, M), *Age group* (18 to 80 years), *Highest level of education attained* (from primary or lower secondary school to PhD, also including higher non-university studies), *Employment position* (entrepreneur, full-time employee, part-time employee, self-employed, unemployed, student, retired, other position), *Marital status* ("Married or Cohabiting" to "Single/never married", but also "I prefer not to say"), *Dietary habits* (omnivore, omnivore with reduced meat in diet, vegetarian, vegan), *Average time spent preparing meals at home* ("No time, do not cook at home", and ranging from "Less than 30 minutes" to "4 hours or more"), *Frequency of using food delivery*, *Frequency of buying sustainable food*, *Frequency of buying fresh food*, *Frequency of buying local food*, *Frequency of buying organic food*, *Frequency of buying food "Made in Italy"* (all frequency questions ranging from "Never" to "Always"), *Willingness to pay an extra fee for "sustainable" food* (scale from "definitely not" to "definitely yes"), *Willingness to pay an extra fee for "Made in Italy" food* (same scale as previous question), *Annual income class* ("Not specified", and ranging from "Less than 4,500€" to "Over 130,000€").

[2] The statistical tests used in the analysis are 1) the maximum likelihood ratio test, in terms of improving the fit of the model by adding or removing variables, and 2) the Wald test, which is used to assess the statistical significance of individual parameters.

Table 1. *Estimation of logit model's parameters related to respondents' propensity to consume meals at a restaurant or pizzeria.*

Characteristics of the respondents		B	Std. err.	Exp(B)	p	
Work position	(baseline: *Other position)*				<0.001	***
Entrepreneur		1.076	0.484	2.933	0.026	*
Full-time employee		0.692	0.417	1.997	0.097	°
Part-time employee		-0.039	0.496	0.962	0.937	
Self-employed		0.506	0.451	1.658	0.262	
Unemployed, seeking work		-0.847	0.544	0.429	0.119	
Retired		-0.406	0.744	0.666	0.585	
Student		-0.341	0.439	0.711	0.437	
Marital status	(baseline: *Married or cohabiting)*				0.005	**
Widowed, Separated or Divorced		-0.353	0.429	0.702	0.410	
Single never married		0.407	0.150	1.503	0.006	**
Would not like to provide information		0.700	0.248	2.013	0.005	**
Use of food delivery	(baseline: *Never)*				<0.001	***
Only sometimes		0.430	0.208	1,538	0.038	*
Often		2.322	0.280	10.200	<0.001	**
Very often or always		2.177	0.479	8.818	<0.001	**
Buying of sustainable food	(baseline: *Never)*				<0.001	***
Only sometimes		-1.184	0.280	0.306	<0.001	***
Thick		-1.098	0.279	0.334	<0.001	***
Very often		-1.020	0.276	0.360	<0.001	***
Always		-1.633	0.599	0.195	0.006	**
Gender*Use of food delivery	(baseline: *F*Never)*				0.023	*
M* Only sometimes		-0.390	0.312	0.677	0.211	
M * Thick		-1.241	0.463	0.289	0.007	**
M* Very often or always		-1.608	0.806	0.200	0.046	*
Gender*Purchasing of sustainable food	(baseline: *F*Never)*				<0.001	***
M* Only sometimes		1.120	0.325	3.064	0.001	**
M * Thick		1.117	0.352	3.055	0.002	**
M * Very often		0.585	0.393	1.795	0.137	
M * Always		2.813	0.755	16.656	<0.001	***
Constant		-2.119	0.452	0.120		

Significance of parameters: (°) 10%; (*) 5%; (**) 1%; (***) 1‰

Table 2. *Matrix of correct classification of the model.*

	Expected response		**Percentage of correct classification**
Observed response	Non-customer	Regular customer	
Not a restaurant customer	2,227	21	99.1
Regular restaurant customer	291	23	7.3
% correct overall classification			87.8

3. Post-stratification for symmetrisation of the target variable

The main reason for the poor performance described above is undoubtedly the extreme asymmetry of the alternatives investigated. Indeed, if about 90% of the observations have one of the two modalities, in practice, any analysis aimed at assessing the probability of the complementary modality will be able to use only a minimal fraction of the necessary information. This phenomenon, which is almost fatal in other statistical techniques based on the search for the best predictability (for example, in the analysis of segmentation, Fabbris, 1997; Fabbris & Martini, 2002), is less relevant in logit analysis, especially when the samples are quite numerous. However, it persists and, sometimes, makes any decision rule impossible or very difficult.

Therefore, here, it was appropriate to experiment with a "Deep Learning" technique that has previously shown excellent results in solving very heavy penalties for symmetry in segmentation analysis and, later, in artificial neural networks elaborated on the basis of

dichotomous response variables (d'Ovidio, Mancarella & Toma, 2016): the formulation of a *symmetric learning sample* constructed by randomly extracting, from the group of statistical units with the majority response, a subgroup of the same size as the one indicating the minority response. The combination of the two subgroups provides a (post-stratified) sample that is almost symmetric in terms of the target variable, although it is undoubtedly smaller in size. In fact, through the above procedure, in addition to the 314 surveyed customers of restaurants and pizzerias, 320 people were randomly selected who ate out only occasionally or never[3]. The corresponding percentages are 49.5% and 50.5%, and the almost perfect symmetry of the distribution of the responses should improve the predictive power of the model.

Table 3, which was elaborated with the same criteria as the previous Table 1, highlights some important differences. First, there is an absence of significant interactions, so *Gender*, whose effect was previously diluted in the interactions, assumes considerable and significant importance in its own right; both the variables *Average time devoted to cooking at home* (but not its specific modalities) and *Willingness to pay an extra fee for food "Made in Italy"* assume statistical relevance; in contrast, *Marital status* and *Frequency of buying "sustainable" food* lose all their relevance and do not appear in the model, while *Use of food delivery services* (which indeed replaced restaurants and pizzerias in terms of the habits of many Italians in the pandemic period) retains statistical significance and much of its relevance. The model fits better, even if sample size is smaller: Cox-Snell $R^2 = 0.157$; Nagelkerke $R^2 = 0.210$.

Finally, the predictive power of the model assumes acceptable values (Table 4), reaching almost two-thirds of correct predictions for the target variable (and surpassing this level in the correct classification of respondents who do not tend to have lunch or dinner outside the home), with 63.4% correct classifications of regular customers of restaurants and pizzerias.

Table 3. *Estimation of logit model's parameters related to respondents' propensity to consume meals at a restaurant or pizzeria, symmetrised sample.*

Characteristics of the respondents	B	Std. err.	Exp(B)	p	
Gender: M (baseline: *F)*	0.752	0.196	2.120	<0.001	***
Work position (baseline: *Other position)*				0.015	*
Entrepreneur	0.593	0.646	1.809	0.359	
Full-time employee	0.317	0.546	1.374	0.561	
Part-time employee	-0.091	0.630	0.913	0.885	
Self-employed	0.112	0.581	1.119	0.847	
Unemployed looking for work	-1.067	0.638	0.344	0.094	°
Retired	-1.022	0.849	0.360	0.229	
Student	-0.193	0.549	0.824	0.725	
Average time devoted to cooking at home (baseline: *No time, no one cooks at home)*				0.050	*
Less than 30 minutes	0.720	1.348	2.054	0.593	
30 min–1 hour	0.968	1.312	2.633	0.461	
1–2 hour	0.319	1.314	1.376	0.808	
2–4 hour	0.760	1.325	2.138	0.566	
4 hours or more	1.061	1.496	2.888	0.479	
Use of food delivery (baseline: *Never)*				<0.001	***
Only sometimes	0.195	0.209	1.215	0.351	
Often	1.785	0.355	5.957	<0.001	***
Very often or always	1.964	0.802	7.125	0.014	*
Willingness to pay extra fee for foods "Made in Italy" (baseline: *Definitely not)*				0.050	*
Probably not	0.200	0.547	1.222	0.714	
Maybe yes, maybe no	-0.844	0.323	0.430	0.009	**
Probably yes	-0.316	0.292	0.729	0.280	
Definitely yes	-0.348	0.317	0.706	0.273	
Constant	-0.892	1.435	0.410		

Significance of parameters: (°) 10%; (*) 5%; (**) 1%; (***) 1‰

[3]The number don't match perfectly between the two groups, because an unavoidable approximation of the computerised procedure of random extraction of the sample of the non-customers respondents.

Table 4. *Matrix of correct classification of the model, symmetrised sample.*

Observed response	Expected response		Percentage of correct classification
	Non-customer	Regular customer	
Not a restaurant customer	217	103	67.8
Regular restaurant customer	115	199	63.4
% correct overall classification			65.6

The striking difference in structure between the model shown in Table 3 and the previous model is obviously due to the different hierarchy of objectives. The model shown in Table 1, while it aimed to identify the characteristics of individuals who tend to eat outside home, necessarily identified, instead, only the variables that characterise individuals who are not accustomed to eating in restaurants. The present model, on the other hand, correctly identified the primary required characteristics, certainly not optimally but well enough for the purposes of the study.

To investigate the reproducibility of the results obtained, it is possible, to calculate the value that the probability of success p of each unit of the total sample assumes using the estimated coefficients, in a logit transformation, for the symmetrised sample (of course, by setting the *baseline category* coefficient to zero):

$$\text{logit}(p) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m.$$

For each subject, the following is then calculated:

$$p = \frac{\exp[\text{logit}(p)]}{1 + \exp[\text{logit}(p)]},$$

rounding the result to the value that identifies the target characteristic "habitual customer" if this probability is close to 1, as well as to the value of the reference characteristic "non-customer" if it is close to zero. In practice, the cut-off line is assumed to be 0.5, in accordance with the given threshold for statistical software.

Thus, once the "expected condition" has been identified (and assigned to a specific record) for each unit of the joint sample, the collected and "expected" data can be easily compared in a contingency table that plays the role of the correct classification matrix.

This transfer (to the totality of the data) of the results obtained with the model derived from the symmetrised sample, as shown in Table 5, provides (as in other experiments previously conducted) results that are fully comparable to those obtained thus far, that is, quite adequate but not optimal. Presumably, beginning from a larger sample and randomly selecting the units to make the modalities of the target variable symmetric would yield a post-stratified sample large enough to guarantee the power and representativeness of the procedure.

Table 5. Matrix of correct classification of the model, applied to the whole sample.

Observed response	Expected response		Percentage of correct classification
	Non-customer	Regular customer	
Not a restaurant customer	1,534	714	68.2
Regular restaurant customer	115	199	63.4
% correct overall classification			67.6

4. Final remarks

The deep learning post-stratification method was first shown to be useful in classification techniques such as segmentation analysis (or artificial neural networks) for symmetrising a categorical response (d'Ovidio, Mancarella & Toma; 2016). In that study, in which no inference was involved, the method provided optimal and robust results. In the first analysis, using the CRT technique, 84% of the minority responses were correctly classified, as compared to 79%

of the alternative, while the original sample analysis provided only 50% correct classification of the interest responses and 99% of the alternative). The same results were obtained by applying the classification rules to the entire dataset (well over one million cases).

In the above research, artificial neural networks, of course, provided better results in the learning and testing samples and were more stable in population reporting (84% to 88% of correct classifications).

The application here shown, thus, demonstrate that post-stratification into symmetric groups provides an effective solution to the problem of the correct representation of relationships by more complex analyses, such as logistic regression. Further applications (including multinomial response variables) could provide a better understanding of the advantages and limitations of this technique.

Information and Acknowledgements.

This paper is the result of joint research, conducted in compliance with statistical ethics, but F.D. d'Ovidio handled the final editing of Section 3, A.M. D'Uggento edited Section 2, R. Mancarella handled Section 4 and E. Toma edited Section 1.

The authors thank Prof. M. G. Onorati and the University of Pollenzo (Bra) for their kind permission to use the survey results described in Sections 2–3.

References

d'Ovidio, F.D., Mancarella, R., Toma, E. (2016). Multivariate data analysis techniques for healthcare organizational efficiency improvement. In: *Proceedings of 5th International Conference "From Challenges to Opportunities: Development of Transition Countries in the Globalization Era"* (Elbasan, AL, December 17). Shpresa Print, Elbasan, AL: 24-39. Book Chapter or Paper in Conference Proceedings.

Fabbris, L. (1997). *Statistica multivariata. Analisi esplorativa dei dati*, McGraw-Hill, Milano. Book.

Fabbris, L., Martini, M.C. (2002). Analisi di segmentazione con una variabile dipendente trasformata in *logit.* In: Carli Sardi L., Delvecchio F. (eds), *Indicatori e metodi per l'analisi dei percorsi universitari e post-universitari*, CLEUP, Padova. Book Chapter or Paper in Conference Proceedings.

Fielding, A. (1977). Binary segmentation: the Automatic Interaction Detector and related techniques for exploring data structure. In: O'Muircheartaigh C.A., Payne C. (eds) *The Analysis of Survey Data. Volume 1; Exploring Data Structures*, Wiley, London: 221-257.Book Chapter or Paper in Conference Proceedings.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 2 (4): 42-47. Journal Article.

Sonquist, J.A., Baker, E.L., Morgan, J.N. (1973). *Searching for Structure*, Institute for Social Research, Ann Arbor, Michigan. Book.