A Binarization Approach to Model Interactions Between Categorical Predictors in Generalized Linear Models

Carrizosa, Emilio; Restrepo, Marcela Galvis; Romero Morales, Dolores

Document Version
Final published version

Published in:
Applied Intelligence

DOI:
10.1007/s10489-024-05576-x

Publication date:
2024

License
CC BY

Citation for published version (APA):
Carrizosa, E., Restrepo, M. G., & Romero Morales, D. (2024). A Binarization Approach to Model Interactions Between Categorical Predictors in Generalized Linear Models. *Applied Intelligence*, *54*(17-18), 7969-7981. https://doi.org/10.1007/s10489-024-05576-x

Applied Intelligence
https://doi.org/10.1007/s10489-024-05576-x

A binarization approach to model interactions between categorical predictors in Generalized Linear Models

Emilio Carrizosa[1] · **Marcela Galvis Restrepo**[2] · **Dolores Romero Morales**[3]

Accepted: 30 May 2024

Abstract
In this paper, our goal is to enhance the interpretability of Generalized Linear Models by identifying the most relevant interactions between categorical predictors. Searching for interaction effects can quickly become a highly combinatorial, and thus computationally costly, problem when we have many categorical predictors or even a few of them but with many categories. Moreover, the estimation of coefficients requires large training samples with enough observations for each interaction between categories. To address these bottlenecks, we propose to find a reduced representation for each categorical predictor as a binary predictor, where categories are clustered based on a dissimilarity. We provide a collection of binarized representations for each categorical predictor, where the dissimilarity takes into account information from the main effects and the interactions. The choice of the binarized predictors representing the categorical predictors is made with a novel heuristic procedure that is guided by the accuracy of the so-called binarized model. We test our methodology on both real-world and simulated data, illustrating that, without damaging the out-of-sample accuracy, our approach trains sparse models including only the most relevant interactions between categorical predictors.

Keywords Generalized linear models · Interpretability · Categorical predictors · Interactions · Clustering of categories

1 Introduction

Modeling interactions between categorical predictors is standard practice in many empirical applications using linear models. For example, in randomized control trials it is common to include interactions between a treatment and a set of covariates to search for treatment effect heterogeneity [9, 14, 16]. Other types of studies on education, health, or labor market outcomes, also commonly include interactions between socioeconomic status and characteristics like race and ethnicity [4, 8, 10, 15]. There is a common approach to modeling categorical predictors and their interactions in linear models that involves encoding each category and each combination of categories using a single binary variable, commonly referred to as one-hot dummy encoding [2]. However, learning from a model with interactions becomes challenging when there are many categorical predictors and/or categories [12].

The simplest case of an interaction between categorical predictors is the one given by two binary predictors. As an illustration, consider the real-world *German* credit dataset used in our numerical section. The aim is to perform a supervised classification task, where we try to classify people according to a set of predictors as *good* or *bad* in terms of credit worthiness. This dataset contains 967 records. Consider two of its binary predictors, namely, *Telephone (in clients name)* and *Foreign worker*. The interaction between two binary predictors can be modeled by adding a new binary predictor which is the combination of both characteristics. In our example that would mean individuals with *Telephone (in clients name)* $=$ 1 and *Foreign worker* $=$ 1. Clearly, it is easy to interpret the role of both binary predictors and their

✉ Marcela Galvis Restrepo
mgalvisr@outlook.com

Emilio Carrizosa
ecarrizosa@us.es

Dolores Romero Morales
drm.eco@cbs.dk

[1] Instituto de Matemáticas de la Universidad de Sevilla, Sevilla, Spain

[2] DEAS Group, Copenhagen Metropolitan Area, Frederiksberg, Denmark

[3] Copenhagen Business School, Frederiksberg, Denmark

Published online: 19 June 2024

interaction, as this only involves looking at three coefficients. In our example, we would have one coefficient for the effect of *Telephone (in clients name)* $= 1$ compared to the *Telephone (in clients name)* $= 0$, another one for *Foreign worker* $= 1$ compared to *Foreign worker* $= 0$, and the last one for the interaction, i.e., for *Telephone (in clients name)* $= 1$ and *Foreign worker* $= 1$. Therefore, in this paper the goal is to binarize the categorical variables.

Continuing with the example above, consider now the case in which we have two categorical predictors, such as *Job* (with 4 categories) and *Purpose* (with 11 categories). To model the interactions between two categorical predictors, we need a coefficient for each possible combination of a category from the first predictor and another from the second one. Clearly, when interpreting these two categorical predictors and their interaction, we require (many) more coefficients. In our example we need to estimate 3 coefficients associated with the categories of predictor *Job*, 10 for the categories of predictor *Purpose* and $3 \cdot 10 = 30$ for the interaction terms. This means that we need to estimate a total of 43 coefficients to interpret the role of both categorical predictors and their interaction. Needless to say, the number of parameters to be estimated is even higher if we have more than 2 categorical predictors in the dataset. In our example, if we consider the pairwise interactions between all 13 categorical predictors in the *German* dataset, we would have to estimate 379 coefficients, after the deletion of the interactions for which we have no data. This makes the estimation of some coefficients imprecise and adds noise to the regression since we have too few records (967) with respect to the high number of parameters to be estimated. Our methodology aims at dramatically reducing this complexity.

In this paper, we propose to find a reduced representation of the categorical predictors as binary predictors to tackle the burden of having too many coefficients to estimate with possibly too few records. As an illustration, let us take the categorical predictor *Job* which includes categories *Unemployed/unskilled - non-resident, Unskilled - resident, Skilled employee/official, Management/self-employed/highly qualified employee/officer*. If some of these categories have a similar impact on the response variable, we could group them together. Say, for instance, *Unemployed/unskilled - non-resident* and *Unskilled - resident* are in one group and *Skilled employee/official* and *Management/self-employed/highly qualified employee/officer* in another group. Thus, instead of 4 binary variables associated with *Job*, we would have just one, indicating whether the individual shows any category of the first group. Similarly, instead of 11 binary variables for *Purpose*, after splitting the categories into two groups, we would have just one. Then, the interaction between *Job* and *Purpose* would be represented by just one coefficient. By doing so, and after the deletion of interactions for which we have no data, our approach reduces from 379 to 34 the number of coefficients associated with all categorical predictors and their interactions in the *German* dataset.

In this paper, we propose a novel methodology to binarize the categorical predictors in Generalized Linear Models (GLM) to model interactions. The goal is to split the categories associated with each categorical predictor into two groups, such that categories in the same group have a similar impact on the response variable. Thus, we make categories in the same group share the same coefficient in the GLM, with the hope that accuracy is not affected much while reducing the number of coefficients. We provide a collection of binarized representations for each categorical predictor, where the dissimilarity takes into account information from the main effects and the interactions. The choice of the binarized predictors representing the categorical predictors is made with a heuristic procedure that is guided by the accuracy of the so-called binarized model.

Our approach to binarizing the categorical predictors to model interactions offers several advantages. First, assuming that the samples of records associated with categories are homogeneous enough, by binarizing the categories we avoid having an over-parametrized model with a coefficient to be estimated per category. Second, we have just one coefficient for each categorical predictor and another one for each interaction between two categorical predictors. This is a step towards enhancing the interpretability of the Generalized Linear Model with interactions. Third, our methodology searches for groups of similar categories that have a similar impact on the response. This is in contrast to shrinkage methods like the version of group lasso proposed by [3, 12], where the goal is just to select relevant predictors and interactions. Fourth, since we are grouping together similar categories, with our approach we have more records to estimate each coefficient, which together with the homogeneity ensures lower standard errors as pointed by, e.g., [11] and [6].

The rest of the paper is organized as follows. Section 2 introduces the algorithm to binarize the categorical predictors using information from the main effects and the interactions. Section 3 illustrates the performance of our methodology on real-world and simulated data, compared to lasso and group lasso. Finally, conclusions and future research are collected in Section 4.

2 Methodology

In this section, we detail the methodology to find a reduced representation of categorical predictors as binary predictors. First, we introduce the notation for the Generalized Linear Model (GLM) with categorical predictors and their interactions. We then introduce a dissimilarity measure between categories of the same predictor based on the GLM coefficients. With this dissimilarity, we define an iterative

algorithm, where in each iteration we cluster the categories of a predictor into two groups to achieve a reduced representation as a binary variable. The binarized predictors will be used to train the so-called binarized GLM in which each categorical predictor is modeled using its reduced representation.

Let us first describe the required notation. We have J categorical predictors. Predictor j has K_j categories, which, when needed will be denoted with letters of the alphabet. In the GLM using the traditional one-hot encoding, a categorical predictor j with K_j categories is represented by $K_j - 1$ binary variables, one for each category, leaving one out for contrast. Therefore, for each categorical predictor, we will leave out one of its categories. We follow the notation in [12]. Consider a GLM where the outcome Y is related to X, comprising the predictors and their interactions, through a link function G:

$$\mathbb{E}[Y|X] = G\Big(\alpha + \sum_{j=1}^{J} X_j \cdot \beta_j + X_{j:l} \cdot \Theta_{j:l}\Big), \tag{1}$$

where α is the intercept, X_j is the vector of binary variables associated with the $K_j - 1$ categories of categorical predictor j, with corresponding parameter vector β_j. The term $X_{j:l}$ is the interaction between categorical predictors j and l, with the corresponding vector of model parameters $\Theta_{j:l}$, where $X_{j:l}$ is the Kronecker product between X_j and X_l. For example, for $K_j = 3$ and $K_l = 4$, we have

$$\begin{aligned} X_{j:l} &= \left(X_{jb}\ X_{jc}\right) * \left(X_{lb}\ X_{lc}\ X_{ld}\right) \\ &= \left(X_{jb:lb}\ X_{jb:lc}\ X_{jb:ld}\ X_{jc:lb}\ X_{jc:lc}\ X_{jc:ld}\right), \end{aligned}$$

where $X_{jb:lb}$ is the interaction between category b of predictor j and category b of predictor l, and $\Theta_{jb:lb}$ is its corresponding coefficient. The rest of the terms can be defined in a similar fashion.

A couple of remarks about the GLM in (1) are worth noting. First, for a binary response variable $Y \in \{0, 1\}$, a natural choice of link function G is the Logit, which we use in Section 3. The approach below can deal with other types of response variables, such as count data, as well as other link functions, such as the one in Poisson regression. Second, among the J categorical predictors we may have binary ones. These are already represented in the most compact form and therefore do not need to be binarized. Third, our methodology can also handle data containing continuous predictors, as in Section 3, but for the sake of notational simplicity, we have decided not to include them in (1).

We will now explain how the binarization of a given categorical predictor, say s, is performed. If s is an ordinal categorical predictor, we apply the approach in [5]. In this case, there is a natural order in the categories of s, which is used to define the so-called feasible clusterings of the categories. For a given threshold value τ, a feasible clustering is one in which the first τ categories of s compose the first cluster, and the remaining ones the second cluster. By changing τ appropriately, we obtain all possible feasible clusterings of the categories and the corresponding binarized representation of the ordinal variable s. Therefore, these ordinal predictors are not included in the discussion below.

In case s is a non-ordinal categorical predictor, we have a more complex relationship between the response variable and the predictors, with both marginal as well as interaction effects, and therefore the approach in [5] is not applicable. Thus, we will inspect the marginal effects and the interactions in (1) to build a dissimilarity matrix which can then be used in a clustering procedure to find two clusters for categorical predictor s.

Let us explain how we calculate the dissimilarity between the pair of categories b and c of predictor s. Category b is similar to category c if they affect the response variable in a similar way. We calculate this by estimating the GLM in (1) and comparing the marginal coefficients for b and c, as well as the coefficients associated with the interactions for these categories. Given the challenges of training a GLM with all possible interactions, where we would have, in general, an overparametrized model, we consider the interaction of s with another categorical predictor j. As we will see below, we iterate over all the possibilities j, having thus different dissimilarity matrices for s yielding a different binarization of s.

We are now ready to define the dissimilarity between the categories b and c of predictor s, when modeling the interaction between s and j:

$$\delta_s^{(j)}(b, c) = (1 - \lambda)\delta_s^{mar}(b, c) + \lambda\delta_s^{int}(b, c), \tag{2}$$

where $\delta_s^{mar}(b, c) = |\beta_b - \beta_c|$ is the difference, in absolute terms, between the pair of marginal coefficients for b and c, $\delta_s^{int}(b, c)$ is the ℓ_1 distance between the two interaction coefficient vectors, and $\lambda \in [0, 1]$. We place more weight on the information provided by the interaction coefficients the higher the value of λ. Note that even when $\lambda = 0$, in which case the interaction coefficients do not play a role in (2), the dissimilarity still contains information from the interactions through the marginal coefficients, since they have been estimated from a model including these interactions.

Let $\boldsymbol{\delta}_s^{(j)}$ denote the dissimilarity matrix, which contains the dissimilarities between all possible pairs of categories of predictor s, when modeling the interaction between s and j. With $\boldsymbol{\delta}_s^{(j)}$, and using a clustering procedure, we can cluster

Fig. 1 Binarization steps for categorical predictor *Job* from the *German* dataset, when interaction effects with categorical predictor *Housing* are considered

Category of *Job*	Marginal coefficients of *Job*	Interaction coefficients between *Job* and *Housing*		
		Housing = *Rent*	*Housing* = *Own*	*Housing* = *For free*
Unemployed	0.06	-2.17	0.00	0.82
Unskilled	0.47	-0.71	0.00	1.40
Skilled	0.00	0.00	0.00	0.00
Management	0.24	-0.68	0.00	-1.78

(a) GLM marginal coefficients of *Job*, as well as interaction coefficients between *Job* and *Housing*

$$\boldsymbol{\delta}_{Job}^{(Housing)} = \begin{pmatrix} & \textit{Unemployed} & \textit{Unskilled} & \textit{Skilled} & \textit{Management} \\ \textit{Unemployed} & 0.00 & 1.22 & 1.53 & 2.14 \\ \textit{Unskilled} & & 0.00 & 1.29 & 1.72 \\ \textit{Skilled} & & & 0.00 & 1.35 \\ \textit{Management} & & & & 0.00 \end{pmatrix}$$

(b) Dissimilarity between the categories of *Job*, using Equation (2) with $\lambda = 0.5$

Category	Cluster	Binary variable
Unemployed *Unskilled*	1	1
Skilled *Management*	2	0

(c) Binarization of *Job* using the dissimilarity $\boldsymbol{\delta}_{Job}^{(Housing)}$

the categories of s into two groups, such that categories in the same group affect the response variable in a similar way. These two groups yield a reduced representation of predictor s as a binary variable, where all categories in the same group now affect the response variable in the same way.

In Fig. 1 we illustrate this process when s is the categorical predictor *Job* from the *German* dataset, see Table 2 for the full list of predictors. For the sake of clarity, we have shortened the names of the categories of *Job* to their first word. We estimate the coefficients in the GLM in (1) with all marginal

1 **Initialization:** Let $\mathcal{L} \subseteq \{1, \ldots, J\}$ be the set of categorical predictors to binarize;
2 Let m be the repeats of the algorithm;
3 Let $\lambda \in [0, 1]$ be the weight parameter in Equation (2);
4 **for** $i \in \{1, \ldots, m\}$ **do**
5 Set $\mathcal{L}' = \mathcal{L}$;
6 **while** $\mathcal{L}' \neq \emptyset$ **do**
7 Randomly sample one predictor from $\mathcal{L}'$, s;
8 Let $\mathcal{V}_s$ be the set out-of-sample accuracies for the binarizations of s;
9 Set $\mathcal{V}_s = \emptyset$;
10 **for** $j \in \{1, \ldots, J\} \setminus \{s\}$ **do**
11 Estimate the GLM in Equation (1) which includes all marginal effects and the interaction between s and j;
12 Calculate the dissimilarity matrix for s, $\boldsymbol{\delta}_s^{(j)}$, using Equation (2);
13 Use $\boldsymbol{\delta}_s^{(j)}$ in a clustering procedure to split the categories of s into two groups;
14 Binarize categorical predictor s;
15 Estimate a GLM as in Line 11, considering s binarized;
16 Calculate its out-of-sample accuracy and add it to $\mathcal{V}_s$;
17 **end**
18 Choose the maximum out-of-sample accuracy in $\mathcal{V}_s$ and binarize s accordingly;
19 Eliminate s from $\mathcal{L}'$;
20 **end**
21 **Return:** GLM_i^B, the GLM in Equation (1) where the predictors in $\mathcal{L}$ are binarized;
22 **end**
23 **Return:** The binarized GLM, GLM_i^B with the highest out-of-sample accuracy.

Fig. 2 Pseudocode for the binarization algorithm of categorical predictors to model interactions

Table 1 Description of the datasets used to test the binarization algorithm

Dataset	N	Response distribution	J	P	$\sum_{j=1}^{J} K_j$	K_j
Coil-2000	5,822	94% − 6%	5	80	77	41,6,10,10,10
Bank marketing	4,119	89% − 11%	9	10	47	12,4,7,10,5,3,2,2,2
German	967	30% − 70%	13	7	51	4,5,7,5,5,4,3,4,3,3,4,2,2
Adult	32,561	24% − 76%	8	5	104	9,16,7,15,6,5,42,2,2
Simulated	12,000	45% − 55%	4	2	35	4, 4, 12, 15

effects and the interactions between the categories of *Job* and the ones from another predictor, namely *Housing*, with three categories, namely *Rent*, *Own* and *For free*. The coefficients can be found in Fig. 1a. Note that *Own* has been chosen as the reference category for *Housing*, having thus a coefficient equal to zero, which explains the column of zeroes in the table in Fig. 1a. This is the same for the category *Skilled* of *Job*, in this case explaining the row of zeroes.

Then, we calculate the dissimilarity matrix $\boldsymbol{\delta}_{Job}^{(Housing)}$ using (2) with $\lambda = 0.5$, see Fig. 1b. We apply a hierarchical clustering procedure with the resulting clusters shown in Fig. 1c. With this, we find a reduced representation of predictor *Job* as a binary variable that takes on value 1 if *Job* is equal to *Unemployed* or *Unskilled* and 0 otherwise.

Our goal is to try out different binarizations of predictor s in order to find a good one in terms of accuracy. The dissimilarity matrix $\boldsymbol{\delta}_s^{(j)}$ depends on which interactions are incorporated in (1). In our example above, if instead of interacting *Job* with *Housing*, we interact it, for instance, with *Status of existing checking account*, we would have had a different dissimilarity matrix. By doing this for all possible predictors $j \neq Job$, we would have $J-1$ dissimilarity matrices, $\boldsymbol{\delta}_{Job}^{(j)}$, $j = 1, \ldots, J-1$. Then, we would have $J-1$ different binarizations for the same categorical predictor that we could choose from, based on out-of-sample accuracy. After making the choice and binarizing the predictor using the corresponding clustering, we incorporate this reduced representation in the next decision to make, namely, the binarization of another categorical predictor.

The pseudocode of our algorithm to binarize categorical predictors to model interactions can be found in Fig. 2. In lines 1 to 3, we initialize the parameters of the algorithm. In lines 7 to 17, we choose randomly the next predictor to binarize and estimate the coefficients in the GLM in (1) which includes all marginal effects and the interactions between the categories of s and another categorical predictor at a time. Then, we calculate the dissimilarity matrices and apply a clustering procedure to find different binarizations of the categorical predictor. In line 18, we estimate the coefficients in the GLM, in a similar fashion as before, but here we have s binarized. The binarization of s that gives the highest out-of-sample accuracy will be chosen, and the categorical predictor will be considered binarized in this way for the steps to come. Once all predictors are binarized, we train, in line 21, the binarized GLM, GLM_i^B, including all binary predictors and we evaluate its performance in a validation set. Since the order in which we binarize predictors matters, we repeat the process m times and finally choose the final GLM_i^B that gives the

Table 2 *German* dataset: description of the categorical predictors

Categorical predictor	Name	K_j	Top counts	Ordinal
X_1	*Status of existing checking account*	4	A14: 394, A11: 274, A12: 269, A13: 63	Yes
X_2	*Credit history*	5	A32: 530, A34: 293, A33: 88, A31: 49	No
X_3	*Purpose*	7	A43: 280, A40: 234, A42: 181, A41: 103	No
X_4	*Savings accounts/bonds*	5	A61: 603, A65: 183, A62: 103, A63: 63	Yes
X_5	*Present employment since*	5	A73: 339, A75: 253, A74: 174, A72: 172	Yes
X_6	*Personal status and sex*	4	A93: 548, A92: 310, A94: 92, A91: 50	No
X_7	*Other debtors/guarantors*	3	A101: 907, A102: 52, A103: 41	No
X_8	*Property*	4	A123: 332, A121: 282, A122: 232, A124: 154	No
X_9	*Other installment plans*	3	A143: 814, A141: 139, A142: 47	No
X_{10}	*Housing*	3	A152: 713, A151: 179, A153: 108	No
X_{11}	*Job*	4	A173: 630, A172: 200, A174: 148, A171: 22	No
X_{12}	*Telephone (in clients name)*	2	A191: 596, A192: 404	No
X_{13}	*Foreign worker*	2	A201: 963, A202: 37	No

Fig. 3 *Simulated* dataset: coefficients in the data generating model

$$\beta_1 = \begin{pmatrix}\beta_{1b}\\ \beta_{1c}\\ \beta_{1d}\end{pmatrix} = \begin{pmatrix}2\\ 2\\ 0\end{pmatrix}, \beta_2 = \begin{pmatrix}\beta_{2b}\\ \beta_{2c}\\ \beta_{2d}\end{pmatrix} = \begin{pmatrix}2\\ 2\\ 0\end{pmatrix}, \beta_3 = \begin{pmatrix}\beta_{3b}\\ \beta_{3c}\\ \vdots\\ \beta_{3f}\\ \beta_{3g}\\ \vdots\\ \beta_{3k}\\ \beta_{3l}\end{pmatrix} = \begin{pmatrix}\text{-1}\\ \text{-1}\\ \vdots\\ \text{-1}\\ 0\\ \vdots\\ 0\\ 0\end{pmatrix}, \beta_4 = \begin{pmatrix}\beta_{4b}\\ \beta_{4c}\\ \beta_{4d}\\ \beta_{4e}\\ \vdots\\ \beta_{4n}\\ \beta_{4o}\end{pmatrix} = \begin{pmatrix}2\\ 2\\ 2\\ \text{-0.5}\\ \vdots\\ \text{-0.5}\\ \text{-0.5}\end{pmatrix},$$

$$\beta_5 = -0.3, \beta_6 = 0.3, \boldsymbol{\Theta_{1:2}} = \begin{pmatrix}\Theta_{1b:2b} & \Theta_{1b:2c} & \Theta_{1b:2d}\\ \Theta_{1c:2b} & \Theta_{1c:2c} & \Theta_{1c:2d}\\ \Theta_{1d:2b} & \Theta_{1d:2c} & \Theta_{1d:2d}\end{pmatrix} = \begin{pmatrix}\text{-6} & \text{-6} & 0\\ \text{-6} & \text{-6} & 0\\ 0 & 0 & 0\end{pmatrix}, \alpha = -0.5$$

highest out-of-sample accuracy.

3 Numerical illustrations

In this section, we illustrate the performance of our binarization methodology for categorical predictors to model interactions. We focus on supervised classification and use as a baseline the logistic regression (LR), where G in (1) is the logit function. The *binarized LR*, obtained with the algorithm in Fig. 2, is compared against *LR*, *lasso* and *group lasso*. These four models are trained with interactions between the categorical predictors. For completeness, we also include the LR in which only marginal effects are modeled, and refer to it as *LR without interactions*. As performance criteria, for each model we report its classification accuracy and its *relative complexity*, which is defined as the number of estimated coefficients for the categorical predictors and their interactions relative to the number of estimated ones for *LR*. With this, the lowest value of the *relative complexity* is equal to 0, when the categorical predictors do not play a role in the model.

Our message is twofold. First, we will illustrate that *LR* has a poor classification accuracy performance since the number of coefficients to estimate, $\sum_{j=1}^{J} K_j + \sum_{1\leq j<s\leq J} K_j \cdot K_s$, can be very large compared to the number of records available, while for some of the combinations of categories from j and s there may not be enough records. Second, we will show that the accuracy of the *binarized LR* is comparable to that of the benchmarks, while the *binarized LR* outperforms them in terms of *relative complexity*, which is our measure of interpretability.

The algorithm in Fig. 2 has two parameters, namely the number of iterations m and the weight λ. We chose $m = 200$ but, by looking at the output of all iterations, one could see that in these datasets a smaller number would have yielded almost the same results. As for λ, after performing a sensitivity analysis, we decided to set it to 0.5. In our benchmark datasets, other choices returned similar values. We perform ten-fold cross-validation to select the final binarization of the categorical predictors. The output of this algorithm is further simplified using a stepwise selection routine in which we select the relevant marginal and interaction effects. To make the comparison fair, we apply this selection to *LR* and *binarized LR*, guided by the Akaike Information Criterion (AIC) measure. For *lasso* and *group lasso*, we perform ten-fold cross-validation to select the shrinkage parameter. For *group lasso*, we implement the version in [12] that considers interactions, the categories associated with each categorical predictor are part of the same group. We coded our algorithm in R and conducted the experiments in a Workstation with an Intel ® Core™ i5-4460 processor with 8 GB of RAM.

Table 3 Real-world datasets: accuracy and relative complexity in the validation set, for the LR without interactions, binarized LR, lasso and group lasso models

Criterion	Model	*Coil2000*	*Bank marketing*	*Adult*	*German*
Accuracy (%)	LR without interactions	94.28	91.6	85.05	76.78
	LR	89.68	83.93	79.20	62.19
	Binarized LR	93.62	92.10	84.28	76.44
	Lasso	93.79	91.15	82.93	76.37
	Group lasso	93.74	90.24	80.41	73.32
Relative complexity (%)	LR without interactions	12.41	7.68	16.36	4.68
	Binarized LR	6.58	6.14	8.97	3.64
	Lasso	0.00	7.24	33.03	30.08
	Group lasso	59.24	29.61	100	88.39

The rest of this section is organized as follows. Section 3.1 describes the datasets, Section 3.2 is devoted to the analysis of the real-world datasets, and Section 3.3 to the simulated dataset.

3.1 Datasets

Our methodology is illustrated on four real-world datasets available at the UCI Machine Learning Repository [7] and one simulated dataset, see Table 1. In the first two columns, we report the name of the dataset and the total number of records (N). In the remaining columns, we report the response distribution, i.e., the percentage of observations with response $Y = 0$ and the percentage with $Y = 1$, the number of categorical predictors (J), which includes binary ones too, the number of continuous predictors (P), the total number of categories ($\sum_{j=1}^{J} K_j$), and the number of categories for each categorical predictor (K_j).

To illustrate the resulting *binarized LR*, we show the resulting coefficients and p-values for one of our real-world datasets, namely, the *German* dataset. In the *German* dataset, we try to classify people according to a set of predictors as *good* or *bad* in terms of creditworthiness. We have 967 records with 13 categorical predictors. Table 2 shows a summary of the categorical predictors in the dataset including the full name of the predictor, the number of categories, the four categories with the top counts, and whether the predictor is ordinal or not. Predictors X_{12} and X_{13} are already binary. Therefore, the first 11 categorical predictors are the ones that need to be binarized. The three ordinal predictors have been binarized using the methodology in [5]. The eight remaining ones are binarized using the algorithm in Fig. 2.

Now let us explain how we have designed the simulated experiment. We want to have clear groups of coefficients within each categorical predictor. The existence of clear groups would lead to an over-parametrized logistic regression if estimated using the one-hot dummy encoding. We

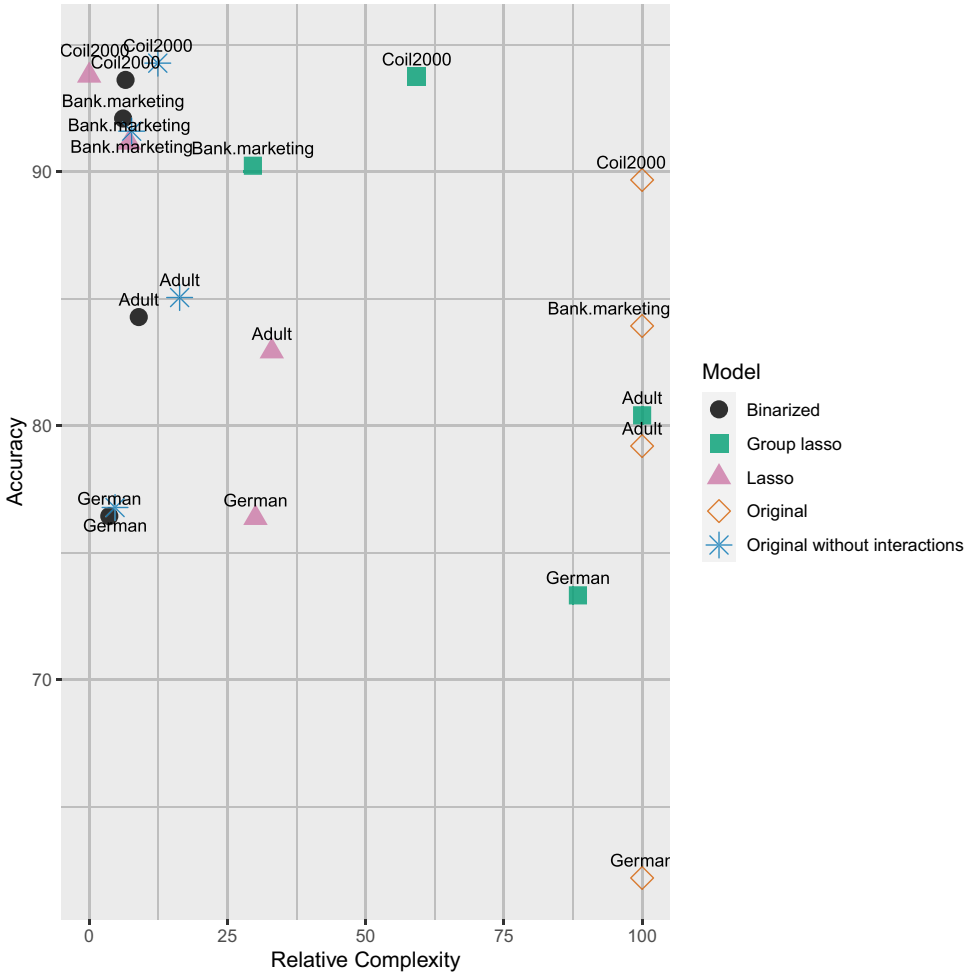

Fig. 4 Real-world datasets: accuracy and relative complexity in the validation set, for the LR without interactions, binarized LR, lasso and group lasso models

generate 12, 000 records of 4 categorical predictors, drawn from a multinomial distribution with equal probabilities for each category, and 2 continuous predictors from a normal distribution with mean 0 and standard deviation 1. Our generating model has only one interaction effect, namely, between the first two categorical predictors. The response $Y \in \{0, 1\}$ is generated from the binomial distribution with probabilities obtained by applying the logistic regression model, using the coefficients in Fig. 3. The groups of categories are clear when we observe these coefficients. For example, categories b and c of predictor X_1 share the same value of the coefficient, $\beta_{1b} = \beta_{1c} = 2$, while for a and d we have $\beta_{1a} = \beta_{1d} = 0$. In summary, there is an equivalent generating model where the four categorical predictors are binary, namely, B_1 with coefficient equal to 2, B_2 with 2, B_3 with -1 and B_4 with 2.5, and one relevant interaction, namely, $B_{1:2}$ with coefficient -6. In Section 3.3, we will show that our algorithm is able to recover this equivalent binary generating model.

3.2 Real-world datasets

In this section, we illustrate the performance of our binarization algorithm in four real-world datasets in terms of accuracy and relative complexity. These estimates are obtained as follows: the dataset is split into a training sample (70%), a test sample (15%), and a validation sample (15%). The model is built in the training sample, we choose the binarization of the categorical predictors using the out-of-sample performance in the test sample and we report its final accuracy in the validation sample. The process is repeated ten times and

Binarized predictor	Name	Category	Cluster
B_1	*Status of existing checking account*	A11 : ... <0 DM	1
		A12 : 0 <= ... <200 DM	2
		A13 : ... >= 200 DM/salary assignments at least 1 year	
		A14 : no checking account	
B_2	*Credit history*	A34 : critical account/ other credits existing	1
		A30 : no credits taken/ all credits paid back duly	2
		A31 : all credits at this bank paid back duly	
		A32 : existing credits paid back duly till now	
		A33 : delay in paying off in the past	
B_3	*Purpose*	A45 : repairs	1
		A49 : business	
		A40 : car (new)	2
		A41 : car (used)	
		A42 : furniture/ equipment	
		A43 : radio/television	
		A46 : education	
B_4	*Savings account/bonds*	A61 : ... <100 DM	1
		A62 : 100 <= ... <500 DM	2
		A63 : 500 <= ... <1000 DM	
		A64 : .. >= 1000 DM	
		A65 : unknown/ no savings account	
B_5	*Present employment since*	A71 : unemployed	1
		A72 : ... <1 year	2
		A73 : 1 <= ... <4 years	
		A74 : 4 <= ... <7 years	
		A75 : .. >= 7 years	
B_6	*Personal status and sex*	A92 : female : divorced/separated/married	1
		A93 : male : single	
		A94 : male : married/widowed	
		A91 : male : divorced/separated	2
B_7	*Other debtors / guarantors*	A101 : none	1
		A102 : co-applicant	
		A103 : guarantor	2
B_8	*Property*	A121 : real estate	1
		A122 : if not A121 : building society savings agreement	
		A123 : if not A121/A122 : car or other, not in attribute 6	
		A124 : unknown / no property	2
B_9	*Other installment plans*	A143 : none	1
		A142 : stores	2
		A141 : bank	
B_{10}	*Housing*	A151 : rent	1
		A152 : own	2
		A153 : for free	
B_{11}	*Job*	A171 : unemployed/ unskilled - non-resident	1
		A172 : unskilled - resident	
		A173 : skilled employee / official	
		A174 : management/ self-employed/ highly qualified employee/ officer	2

Fig. 5 *German* dataset: binarization of the categorical predictors. Note that X_{12} and X_{13} are already binary, i.e., $B_{12} = X_{12}$ and $B_{13} = X_{13}$, and therefore have not been included here

we report as an estimate the average out-of-sample accuracy. A similar process is used for the benchmarks, *LR without interactions*, *LR*, *lasso* and *group lasso*.

The accuracy and the relative complexity can be found in Table 3 and in Fig. 4, both measured as a percentage. We can see that for all datasets, *LR* gives a lower accuracy than the *binarized LR*. This is because in the real-world datasets the number of records associated with each category is not evenly distributed, and hence, some categories have few observations which lead to even fewer observations for the interactions. In some cases, this is exacerbated by the small absolute number of observations, like in the *German* dataset, where the ratio of the number of coefficients associated with the categorical predictors, after deleting those for which we do not have records (379/967) makes training this model very challenging. In this dataset, the accuracy goes from 62.19% for *LR* to 76.44% for the *binarized LR*. This outperformance of the *binarized LR* can be seen in the other three real-world datasets too.

The relative complexity of *binarized LR* is competitive not only against *LR* but also when compared with the *LR without interactions*, which has much fewer coefficients to be estimated. For the *German* dataset, the relative complexity of the *binarized LR* is 3.64%, while 4.68% for *LR without interactions*. This outperformance is even more pronounced in the *Coil2000* and *Adult* datasets where the *binarized LR* halves the relative complexity of the *LR without interactions*. In conclusion, we are able to model the interactions, as well as work with a much smaller model.

Comparing the *binarized LR* to *lasso* and *group lasso*, we find that our algorithm produces a model with higher accuracy for the datasets *Adult* and *Bank marketing*. For *Coil2000* and *German*, our method performs similarly to *lasso* and *group lasso* in terms of accuracy. In terms of relative complexity, in three out of four datasets, our method results in a smaller model.

For the *binarized LR* model in the *German* dataset, Fig. 5 reports for each categorical predictor the two clusters of categories yielding the reduced representation as a binary predictor. For instance, let us look at the first and the last categorical predictors to be binarized, namely X_1 defined as *Status of existing checking account* (with 4 categories) and X_{11} defined as *Job* (with 4 categories). For X_1, binarized as B_1, cluster 1 contains one category ($... < 0DM$) and cluster 2 the remaining three ($... < 200DM$, $... >= 200$ *DM/salary assignments at least 1 year, no checking account*). For X_{11}, binarized as B_{11}, cluster 1 contains three categories (*unemployed/unskilled - non-resident, unskilled - resident, skilled employee/official*) and cluster 2 the remaining category (*management/self-employed/highly qualified employee/officer*). As pointed out in the introduction, it is now easier to interpret the role of *Status of existing checking*

	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$	$B7$	$B8$	$B9$	$B10$	$B11$	$B12$	$B13$
$B1$	40.53	0.31	0.85	0.42	−0.61	0.26	0.31	−0.94	-0.93^{***}	−0.48	2.45	0.51	−41.05
$B2$		−1.04	-2.95^{**}	0.09	1.84	−1.81	2.97^{**}	−0.21	-1.533^{***}	0.00	0.35	0.44	0.53
$B3$			4.27	−1.40	-4.23^{**}	0.15	0.24	2.03^{**}	0.46	−1.08	1.99^{**}	0.42	−0.26
$B4$				−62.17	−0.58	20.00	2.31	0.60	0.63	0.65	0.10	0.70	40.35
$B5$					−17.30	19.99		2.39	0.03^{***}	−1.22	-2.42^{*}	-2.51^{*}	
$B6$						−126.80	17.06	−1.53	22.39	2.70^{*}	4.39^{**}	4.15^{**}	77.57
$B7$							−32.92	2.46	0.56	−1.60	−15.27	1.78	12.76
$B8$								−29.77	−0.75	−0.22	0.46	0.04	30.93
$B9$									−47.78	0.73	0.88	-0.84^{*}	25.32
$B10$										−28.41	0.72	−0.10	26.03
$B11$											-5.67^{**}	−0.52	
$B12$												15.70	−16.94
$B13$													−71.79

(a) Before the stepwise variable selection procedure has been applied

	$B1$	$B2$	$B3$	$B4$	$B5$	$B7$	$B8$	$B9$	$B10$	$B11$	$B12$	$B13$
$B1$	-0.83^{**}							−0.61		1.32^{***}		
$B2$		0.22	-1.44^{*}			1.98^{**}		-1.34^{***}				
$B3$			0.27				1.39^{*}					
$B4$				-0.75^{***}		1.53						
$B5$					−0.77		2.02^{***}			−0.57		
$B7$						-1.67^{***}						
$B8$							−1.18					
$B9$								0.07				
$B10$									-0.47^{**}			
$B11$										−0.55		
$B12$											0.39^{**}	
$B13$												1.51^{**}

(b) After the stepwise variable selection procedure has been applied

Fig. 6 *German* dataset: coefficients for the binarized model with interactions and their significance, where * indicates a p-value below 0.1, ** below 0.05, and *** below 0.01, before and after the stepwise variable selection procedure has been applied

Table 4 *Simulated* dataset: accuracy and relative complexity in the validation set, for the LR without interactions, binarized LR, lasso and group lasso models

Criterion	Model	Simulated dataset
Accuracy (%)	LR without interactions	78.49
	LR	78.53
	Binarized LR	78.45
	Lasso	77.83
	Group lasso	78.08
Relative complexity (%)	LR without interactions	11.59
	Binarized LR	1.45
	Lasso	17.15
	Group lasso	4.65

account and *Job* and their interaction, as this only involves looking at three coefficients, the ones of B_1, B_{11}, and $B_{1:11}$. Figure 6 helps us visualize these coefficients.

Figure 6 provides information about the coefficients of the *binarized LR* model before the stepwise variable selection procedure has been applied (Fig. 6a) and after (Fig. 6b). In the diagonal of each matrix we find the marginal coefficients for the binary predictors and outside the diagonal the coefficients for their interactions when both binary predictors are set to one. Looking at Fig. 6b, we can see that there are 12 marginal coefficients, 2 are significant at the 1% level, and 4 more at the 5%. As for the interactions, there are 9 coefficients after a stepwise selection has been performed. From those, 3 are significant at the 1% level, 1 more at the 5%, and 2 additional ones at the 10%. At the 1% level, *Savings account/bonds* (B_4) and *Other debtors / guarantors* (B_7) are significant, as well as the interactions between *Status of existing checking account* and *Job* ($B_{1:11}$), *Credit history* and *Other installment plans* ($B_{2:9}$), and *Present employment since* and *Property* ($B_{5:8}$).

3.3 Simulated dataset

In this section, we discuss the results for the simulated dataset. As before, we split the data into a training sample

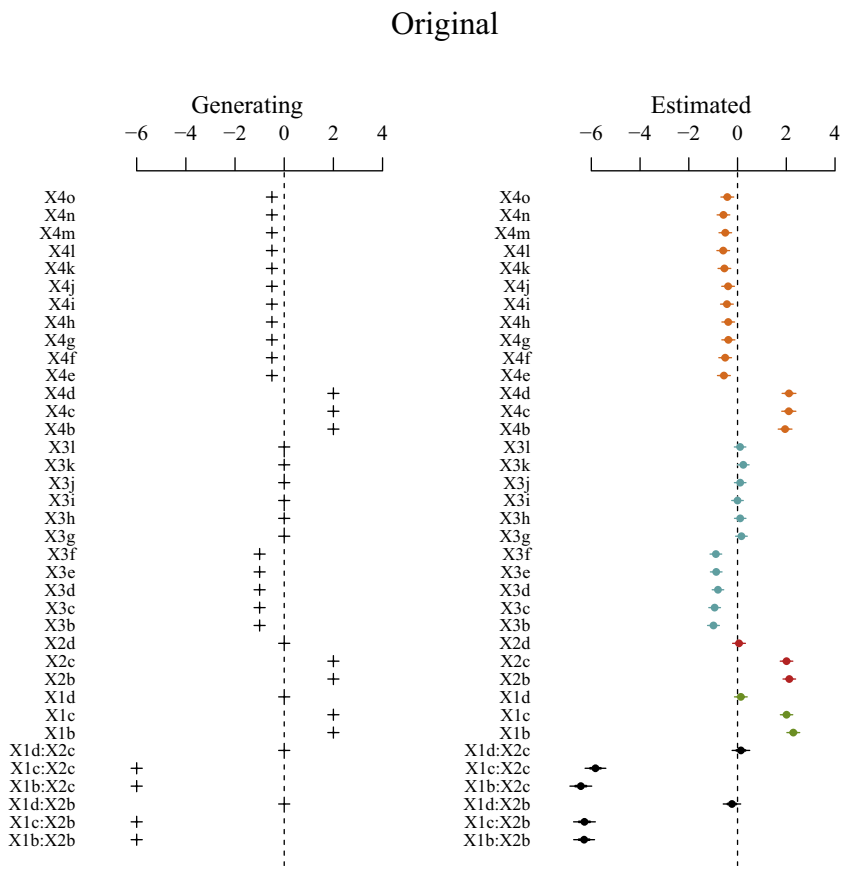

Fig. 7 *Simulated* dataset: generating model (left) and coefficients of original model with 95% confidence intervals (right)

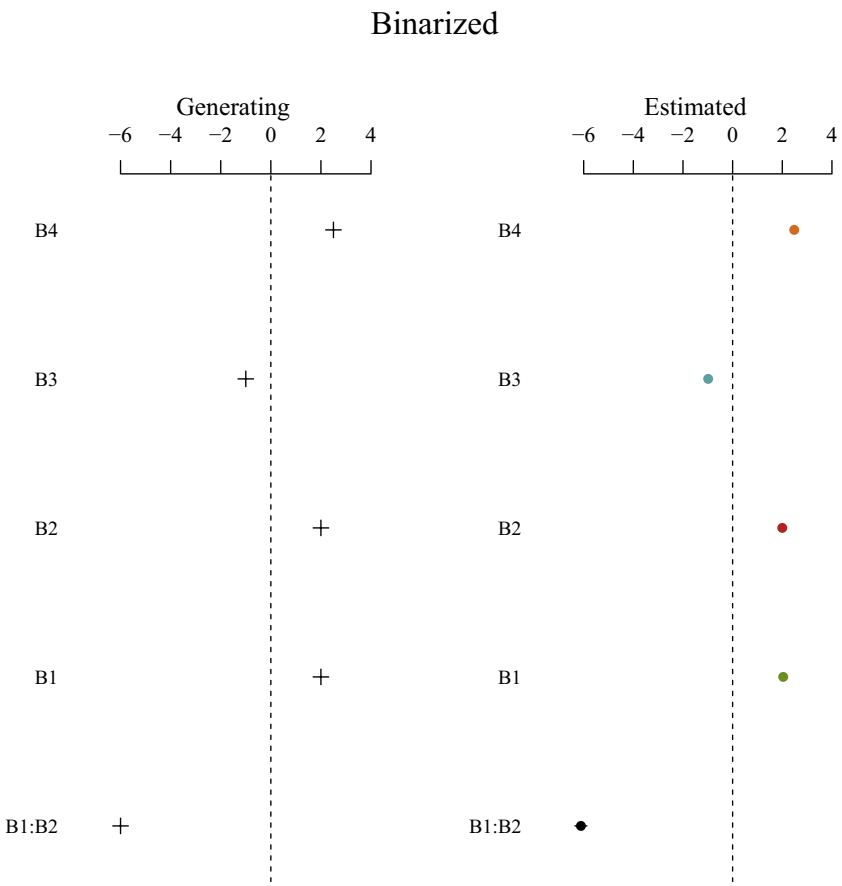

Fig. 8 *Simulated* dataset: equivalent binary generating model (left) and coefficients of binarized model with 95% confidence intervals (right)

(70%), a test sample (15%), and a validation sample (15%). The model is built in the training sample, the parameters are chosen using the test sample, and the performance is measured in the validation sample. We repeat the process ten times, and report average out-of-sample accuracy and *relative accuracy*.

Table 4 reports the accuracy and the relative complexity. The conclusions are similar, as before. While all models have a similar accuracy, the *binarized LR* outperforms the benchmarks in terms of relative performance. This means that our approach allows us to model the interactions, using a much smaller model. We end this section by illustrating how our binarization algorithm is able to recover the underlying generating model. On the right panel of Fig. 7, we plot the value of the coefficients and their 95% confidence intervals for the original model with interactions, while the left panel plots the values used by the data generating model. We plot similar information in Fig. 8 for the *binarized LR*. We see that we recover the generating model in both cases, while in the *binarized LR* the coefficients are estimated with a larger sample, resulting in smaller standard errors, as seen in the 95% confidence intervals around the coefficients.

4 Conclusions

In this paper, we have presented an approach to binarizing categorical predictors that enables working with interactions in Generalized Linear Models. Our approach offers several advantages. First, given that the samples of categories are homogeneous enough, by binarizing we can avoid having an over-parametrized model with a coefficient for each category. Second, we estimate just one coefficient for each categorical predictor and another one for each interaction. This gives a more interpretable model compared to having all the categories as binary variables. Third, by binarizing the categories we have more records to estimate each coefficient, which together with the homogeneity ensures lower standard errors. In the numerical section, we have used a simulated dataset and four real-world ones from supervised classification. In all these cases, our algorithm, in which the GLM with the logit function was used, considerably reduces the number of coefficients of the model, allowing the user to interpret and select interactions between the new binarized categorical predictors. We end by noting that, although for simplicity the methodology has been tested on supervised

classification with the logistic regression as baseline, the very same approach is applicable to classification and regression tasks, as long as they are based on GLM.

A fruitful line of future work is related to the use of categorical predictors that contain sensitive information. In the future, our clustering methodology could take into account not only the overall accuracy but also a fair treatment of the sensitive groups [1, 13, 17]. Another interesting line of future research is the pursuit of metaheuristics that can deal with large-scale datasets involving an extremely large number of categories.

Funding Open access funding provided by Copenhagen Business School. This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), FQM-329 (Junta de Andalucía), and PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain). This support is gratefully acknowledged.

Data Availibility Statement The raw data required to reproduce the above findings are available to download from https://archive.ics.uci.edu/ml/index.php.

Declarations

Competing interests The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing, arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

1. Aghaei S, Azizi M, Vayanos P (2019) Learning optimal and fair decision trees for non-discriminative decision-making. In Proceedings of the AAAI conference on artificial intelligence 33:1418–1426
2. Agresti A, Kateri M (2011) Categorical Data Analysis. Springer
3. Bien J, Taylor J, Tibshirani R (2013) A lasso for hierarchical interactions. Ann Stat 41(3):1111–1141
4. Busetta G, Campolo MG, Panarello D (2020) Weight-based discrimination in the Italian labor market: an analysis of the interaction with gender and ethnicity. J Econ Inequal 18(4):617–637
5. Carrizosa E, Galvis Restrepo M, Romero Morales D (2021) On clustering categories of categorical predictors in generalized linear models. Expert Syst Appl p 115245
6. Carrizosa E, Mortensen LH, Romero Morales D, Sillero-Denamiel MR (2022) The tree based linear regression model for hierarchical categorical variables. Expert Syst Appl 203:117423
7. Dua D, Graff C (2017) UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
8. Howard KA, Carlstrom AH, Katz AD, Chew AY, Ray GC, Laine L, Caulum D (2011) Career aspirations of youth: untangling race/ethnicity, SES, and gender. J Vocat Behav 79(1):98–109
9. Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 7(1):443–470
10. Kingston G, McGinnity F, O'Connell PJ (2015) Discrimination in the labour market: nationality, ethnicity and the recession. Work Employ Soc 29(2):213–232
11. LeBlanc M, Tibshirani R (1998) Monotone shrinkage of trees. J Comput Graph Stat 7(4):417–433
12. Lim M, Hastie T (2015) Learning interactions via hierarchical group-lasso regularization. J Comput Graph Stat 24(3):627–654
13. Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. Knowl Eng Rev 29(5):582–638
14. Seibold H, Zeileis A, Hothorn T (2016) Model-based recursive partitioning for subgroup analyses. Int J Biostat 12(1):45–63
15. Toutkoushian RK, Bellas ML, Moore JV (2007) The interaction effects of gender, race, and marital status on faculty salaries. J High Educ 78(5):572–601
16. Weisberg HI, Pontes VP (2015) Post hoc subgroups in clinical trials: Anathema or analytics? Clin Trials 12(4):357–364
17. Zafar M, Valera I, Gomez Rodriguez M, Gummadi K (2017) Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics, pp 962–970. Proceedings of Machine Learning Research

Emilio Carrizosa is Full Professor of Statistics and Operations Research in the University of Seville, Spain.

His research interests include: Industrial and Applied Mathematics; Data Science (Explainable and Fair Machine Learning, Supervised Classification and Regression), Mathematical Optimization and Operations Research (Mixed Integer Nonlinear Programming, Global Optimization, Vector Optimization).

He is President of math-in, the Spanish Network of Industrial Mathematics, and has served in the past as Director of IMUS, the Mathematical Institute of the University of Seville, President of SEIO, the Spanish Statistics and OR Society, and Editor-in-Chief of TOP, the Spanish OR journal.

He has +150 publications with +2,000 citations, yielding an h-index of 27.

He has papers in journals in the area Operations Research and Management Science: Operations Research, Mathematical Programming, Management Science, Mathematics of Operations Research, etc. Due to his interdisciplinary research, he has also published in disciplines beyond OR: Statistics and Probability, Energy, Chemical Engineering, and Hydrology.

He has supervised 15 PhD Theses (plus 6 ongoing), 5 of them awarded with various national and international prizes.

He is involved in Transfer of Knowledge activities, leading industrial projects and contracts in applications to different sectors: Energy, Health, Logistics, Information Technologies, Environment, and Smart cities.

He has an intense activity of outreach, participating in debates and interviews in tv, radio and newspapers on industrial mathematics and teaching mathematics.

Marcela Galvis Restrepo is a Data Scientist at DEAS Group in Denmark, she finished her PhD at Copenhagen Business School in 2022. Her thesis was on methods to improve the fairness and interpretability of machine learning algorithms in the presence of high-cardinality categorical predictors, under the supervision of Prof. Dolores Romero Morales. Her work develops methods using tools from mathematical optimization and data science. She is interested in applying these new methods to aid decision making in fields like educational data mining (dropout prediction), customer churn prediction and knowledge management.

She holds a B.Sc. from University of Antioquia (Colombia) from 2009, an MSc. in Economics with a Major in Innovation and Change from the Friedrich Schiller University Jena (Germany) from 2013. After her master studies she worked first in the Education Planning and Statistics Unit of Medellin City Hall where she was responsible for applied research in the economics of education, especially the measurement of education quality. In 2017 she worked as a Knowledge Management Consultant in the Inter-American Development Bank in Washington DC where she helped use data analytics techniques to improve knowledge management.

Dolores Romero Morales is a Professor in Operations Research at Copenhagen Business School. Her areas of expertise include Data Science, Supply Chain Optimization and Revenue Management. In Data Science she investigates explainability/interpretability, fairness and visualization matters. In Supply Chain Optimization she works on environmental issues and robustness. In Revenue Management she works on large-scale network models. Her work has appeared in a variety of leading scholarly journals, including European Journal of Operational Research, Management Science, Mathematical Programming and Operations Research, and has received various distinctions. Currently, she is Editor-in-Chief to TOP, the Operations Research journal of the Spanish Society of Statistics and Operations Research, and an Associate Editor of Journal of the Operational Research Society, and the INFORMS Journal on Data Science.

She has worked with and advised various companies on these topics, including IBM, SAS, KLM and Radisson Edwardian Hotels, as a result of which these companies managed to improve some of their practices. SAS named her an Honorary SAS Fellow and member of the SAS Academic Advisory Board. She currently leads the EU H2020-MSCA-RISE NeEDS project, which has a total of 15 participants and a budget of more than €1.000.000 for intersectoral and international mobility, with the aim to improve the state of the art in Data Driven Decision Making.

Dolores joined Copenhagen Business School in 2014. Prior to coming to Copenhagen Business School, she was a Full Professor at University of Oxford (2003–2014) and an Assistant Professor at Maastricht University (2000–2003). She has a BSc and an MSc in Mathematics from Universidad de Sevilla and a PhD in Operations Research from Erasmus University Rotterdam.