Feature Reduction for Classification with Mixed Data
An Algorithmic Approach

Restrepo, Marcela Galvis

COPENHAGEN BUSINESS SCHOOL
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DANMARK

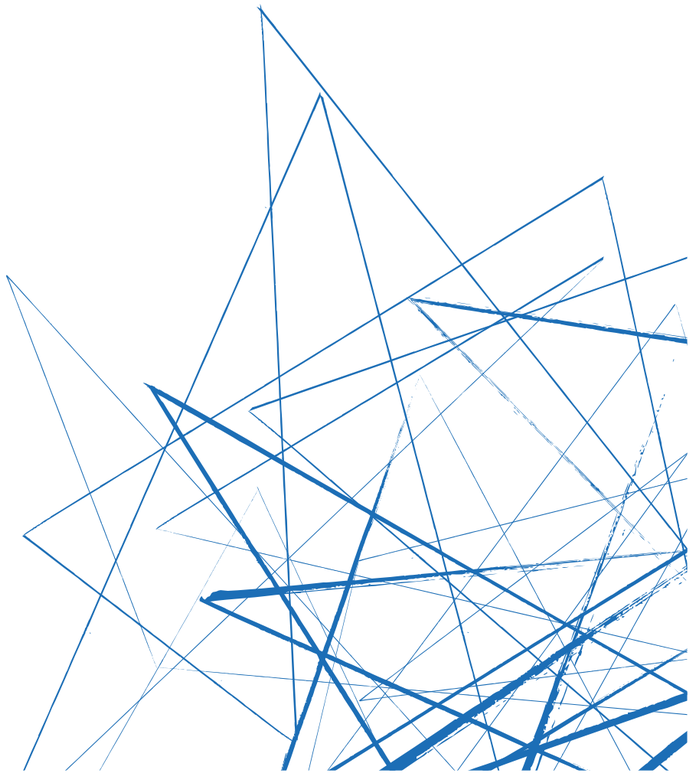
WWW.CBS.DK

ISSN 0906-6934

Print ISBN: 978-87-7568-123-5
Online ISBN: 978-87-7568-124-2

FEATURE REDUCTION FOR CLASSIFICATION WITH MIXED DATA: AN ALGORITHMIC APPROACH

PhD Series 35.2022



Marcela Galvis Restrepo

FEATURE REDUCTION FOR CLASSIFICATION WITH MIXED DATA: AN ALGORITHMIC APPROACH

CBS PhD School Economics

PhD Series 35.2022

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Feature reduction for classification with mixed data: an algorithmic approach

Marcela Galvis Restrepo

A thesis presented for the degree of
Doctor of Philosophy

Primary supervisor: Dolores Romero Morales
Secondary supervisors: Emilio Carrizosa, Fane Naja Groes

CBS PhD School
Copenhagen Business School

Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

First edition 2022
Ph.D. Serie 35.2022

ISSN 0906-6934

Print ISBN: 978-87-7568-123-5
Online ISBN: 978-87-7568-124-2

Preface

This thesis is a result of my PhD studies at the Department of Economics, Copenhagen Business School. I am grateful for the support provided by the Department during my studies. This thesis was written during strange times: midway through my Ph.D. a world pandemic broke out and we were no longer able to meet with friends and colleagues. It became lonely at times and I would not have been able to make it without the encouragement and support from my main supervisor Dolores Romero Morales, and my secondary supervisor Emilio Carrizosa. Dolores, your work ethic, passion, and perseverance have taught me that enjoying the journey can be as important as the final result, I would not have made it without you. Emilio, your encouragement and optimism have inspired me to keep going. The two of you make a dream team and I feel privileged to have had you as my supervisors. Thanks also to my secondary supervisor, Fane Groes, for your insightful comments and feedback on my work.

I would like to thank my many colleagues who have commented on my projects at seminars, conferences, and even in the hallways and during lunch breaks. A special thanks to my discussants during the closing seminar for helpful comments and suggestions.

Thanks to my friends, the old and the new, moving to a new country is never easy but finding a community is what kept me going. To my partner, Alejandro, the way in which you see every challenge as an opportunity is what ultimately allowed me to embark on this journey, and for that, I am forever grateful. To Canela for her companionship. Lastly, I would like to thank my parents and siblings for all their love and support, I have been very lucky to have my parents give me the opportunity to choose my own path in a country where not everyone has that chance. I dedicate this thesis to you.

English Abstract

This thesis consists of five chapters including the introduction. The chapters deal with feature reduction for classification with mixed data, with an application to the prediction of school dropout. The traditional way to incorporate categorical predictors in linear models is through one-hot encoding, where each category is represented by a dummy variable, this can be wasteful, difficult to interpret, and prone to overfitting, especially when dealing with high-cardinality categorical predictors.

In the second chapter, co-authored with Emilio Carrizosa and Dolores Romero Morales, we propose a method to find a reduced representation of the categorical predictors by clustering their categories. This is done through a numerical method that aims to preserve (or even, improve) accuracy while reducing the number of coefficients to be estimated for the categorical predictors. We illustrate the performance of our approach in real-world classification and count-data datasets where we see that clustering the categorical predictors reduces complexity substantially without harming accuracy.

The third chapter, co-authored with Emilio Carrizosa and Dolores Romero Morales, proposes a methodology to deal with Generalized Linear Models including interactions between categorical predictors. In the presence of categorical predictors, searching for interaction effects can quickly become a highly combinatorial problem when we have many of them or even a few high-cardinality categorical predictors. In these cases, including all potential interactions in the model is computationally time-consuming, if not intractable. To alleviate this and at the same time enhance model interpretability without compromising accuracy, we propose to find an alternative representation for each categorical predictor as a binarized predictor via a clustering procedure. We apply our methodology to both

real-world and simulated data demonstrating the usefulness of our approach.

In the fourth chapter, I apply Supervised Learning to address the problem of predicting school dropout in Colombia. School dropout is pervasive in many education systems around the world. Identifying students at a higher risk of dropping out can be an important tool to target interventions intended to reduce this phenomenon, especially in a developing setting where cost is an important consideration. In this chapter, I apply supervised machine learning methods to predict school dropout in rural Antioquia, Colombia, using individual-level administrative data. The results are comparable to other Latin American studies in terms of predictive performance. It is further demonstrated in the chapter that machine learning methods are more accurate in detecting false positives than targeting students in schools with high dropout rates, so using data-driven targeting based on statistical learning can save money in the long run.

The fifth chapter, co-authored with Emilio Carrizosa and Dolores Romero Morales, deals with unfairness in Supervised Learning. In recent years, Supervised Learning has been used to support or even replace human decisions in high-stakes domains such as pretrial risk assessment, police stop-and-frisk programs, credit scoring, insurance premiums, and healthcare access. The training of these algorithms uses historical data which might be biased against individuals with certain sensitive characteristics. The increasing concern over potential biases has motivated lawmakers to pass anti-discrimination laws which prohibit unfair treatment based on characteristics such as gender or race. In this chapter, we propose a methodology that enhances the trade-off between accuracy and unfairness in classification. We use a numerical method that shrinks the possible values the predictors can take. To do this, we optimize a linear combination of the accuracy and a measure of unfairness (the disparate mistreatment) of the shrunk model.

Danish Abstract

Denne afhandling består af seks kapitler, heriblandt en introduktion og en konklusion. Kapitlerne omhandler feature reduction til klassificering af blandet data og en anvendelse til forudsigelse af skolefrafald. Den traditionelle måde at inkorporere kategoriske prædiktorer i lineære modeller er gennem one-hot encoding, hvor hver kategori er repræsenteret af en dummy-variable. Dette kan medføre spild, er vanskeligt at fortolke og med tilbøjelighed til overfitting. Det er især udtalt, når man har at gøre med kategoriske prædiktorer med høj kardinalitet.

I det andet kapitel, som er skrevet sammen med Emilio Carrizosa og Dolores Romero Morales, foreslår vi en metode til at finde en reduceret repræsentation af de kategoriske prædiktorer ved at gruppere deres kategorier. Dette gøres gennem en numerisk metode, der har til formål at bevare (eller endda forbedre) præcision og samtidig reducere antallet af koefficienter, der skal estimeres for de kategoriske prædiktorer. Vi illustrerer ydeevnen af vores tilgang ved at bruge både et virkeligt klassfikations- og et countdata datasæt. Gruppering af de kategoriske prædiktorer reducerer kompleksiteten væsentligt uden at skade præcision.

Det tredje kapitel, som er skrevet sammen med Emilio Carrizosa og Dolores Romero Morales, foreslår en metode til at håndtere generaliserede lineære modeller, herunder interaktioner mellem kategoriske prædiktorer. I arbejdet med kategoriske prædiktorer kan søgning efter interaktionseffekter hurtigt blive et meget kombinatorisk problem. Dette gælder især hvis der er tale om mange prædiktorer eller endda nogle få kategoriske prædiktorer med høj kardinalitet. I disse tilfælde er det beregningsmæssigt tidskrævende at inkludere alle potentielle interaktioner i modellen – hvis ikke umuligt. For at afhjælpe dette og sam-

tidig forbedre modelfortolkningen uden at gå på kompromis med præcision, foreslår vi at finde en alternativ repræsentation for hver kategorisk prædiktor som en binær prædiktor via en clustering procedure. Vi anvender vores metode på både virkelig og simuleret data hvilket demonstrerer anvendeligheden af vores tilgang.

I det fjerde kapitel anvendes Supervised Learning til at løse problemet med at prædiktere skolefrafald i Colombia. Skolefrafald er gennemgående i mange uddannelsessystemer rundt om i verden. At identificere elever med en højere risiko for at droppe ud kan være et vigtigt værktøj til at målrette indsatser, der har til formål at reducere dette fænomen, især i et udviklingslande, hvor omkostninger er en vigtig overvejelse. I dette kapitel anvender jeg supervised machine learing til at prædiktere skolefrafald i Antioquia, Colombia, ved at bruge administrative data på individuelt niveau, der er tilgængelige for alle uddannelsesmyndigheder i Colombia. Disse metoder identificerer omkring 70% af eleverne med risiko for frafald med en falsk positiv rate på 25% uden for stikprøven (hvor den positive klasse er frafald). Resultaterne kan sammenlignes med undersøgelser af andre lande i Latinamerika og nogle regioner i USA. Hvis man antager forskellige effektivitetsniveauer for en intervention, sammenlignes disse prædiktioner med andre metoder til målretning mod elever, som f.eks. at fokusere på skoler med de højeste frafaldsrater. Brug af prædiktion kan føre til besparelser på 24% af omkostningerne forbundet med en intervention designet til at forhindre skolefrafald.

Det femte kapitel, som er skrevet sammen med Emilio Carrizosa og Dolores Romero Morales, omhandler unfairness i Supervised Learning. I de senere år er Supervised Learning blevet brugt til at understøtte eller endda erstatte menneskers beslutninger på områder med stor betydning, som f.eks. risikovurdering forud for retssager, politiets stop-and-frisk-programmer, kreditscoring, forsikringspræmier og adgang til sundhedsydelser. Træningen

af disse algoritmer bruger historiske data, som kan være biased mod individer med bestemte karakteristika. Den stigende bekymring over potentielle biases har motiveret lovgivere til at vedtage love mod forskelsbehandling som forbyder uretfærdig behandling baseret på karakteristika som køn eller race. I dette kapitel foreslår vi en metode, der forbedrer afvejningen mellem præcision og unfairness i klassificeringen. Vi bruger en numerisk metode, der formindsker de mulige værdier, som prædiktorerne kan tage. For at gøre dette optimerer vi en lineær kombination af præcision og et mål for unfairness (den disparate mistreatment) af den krympede model.

Contents

Chapter 1

Introduction

1.1 Supervised Learning with mixed data

Increasingly, approaches from the Supervised Learning literature have been incorporated into the econometric toolbox as important means to tackle a variety of problems (Athey, 2019). Supervised Learning takes a training dataset and estimates or "learns" parameters that can be used to make predictions on new data with the purpose of having a good performance out-of-sample using data-driven model selection (Athey, 2017). In recent years, several authors have explored the potential of using Supervised Learning to solve prediction problems relevant to policy domains (Varian, 2014; Kleinberg et al., 2015; Athey, 2017, 2019; Athey and Imbens, 2019). This is, for instance, the case in criminal justice systems, to make decisions about whether to retain or release arrestees (Kleinberg et al., 2018; Berk, 2019; Jung et al., 2020; Završnik, 2021), in labor market policy, predicting unemployment spell length to inform workers on optimal saving rates (Viljanen and Pahikkala, 2020), or in social policy, predicting students at-risk of dropping out of school to target interventions (Adelman et al., 2018; Chandler et al., 2011).

Supervised Learning requires the manipulation and analysis of data that is increasingly complex (Varian, 2014; Carrizosa et al., 2022b). Such complexity can come from the number of observations and/or the number of predictors, different types of predictors (numerical, categorical, network, process, text, and unstructured data), and extreme values, among others (Carrizosa et al., 2021c). In the case of categorical predictors, they are used in many social science applications with individual-level data, where they can represent group membership, like individuals with the same country of origin, attending the same school or being treated at the same hospital (Bonhomme and Manresa, 2015; Johannemann et al., 2019). Categorical predictors can also depict complex behavioral data related to, e.g., people's

demographics and preferences (Moeyersoms et al., 2016) and in labor economics, the mining of job posts on the internet requires the use of categorical variables associated to skills coded in text (Boselli et al., 2018; Pejic-Bach et al., 2020; Jensen, 2021)

No matter the application, before using mixed data, with both numerical and categorical predictors in any linear model, the categorical predictors must be transformed into real-valued vectors. It is typically done by representing each category by a 0-1 coded dummy variable and leaving one out for contrast, an approach referred to as *one-hot dummy encoding*. A number of other techniques can be used to encode categorical features, including *effect coding*, *hashing*, *weight of evidence analysis (WOE)*, *frequency encoding*, *PCA* adapted to categorical features, and *embeddings*. Hashing and effect encoding are similar to one-hot dummy encoding in that they create new variables associated with each category. In effect encoding, the categories can be represented by values in the -1,0,1 format instead of the 0-1 format, while hashing uses a reduced number of dummies that needs to be specified. In WOE and embeddings, categorical features are transformed into numerical features by replacing the categories with numerical values. For a particular category of a categorical feature, WOE is the natural logarithm of the percentage of cases in the positive class to the percentage of cases in the negative class. An embedding is a way to turn a categorical feature into a numerical feature by defining a distance measure for each pair of categories. As an example, if we have a dataset with numerical predictors like income and age and want to compare categories b and c of a categorical feature j, we calculate the average income and age for people in group b and the average of these features for people in group c and compare these with a distance

Our approach relies on one-hot dummy encoding since this is the most commonly used approach for representing the categorical features and it is often embedded in the way that

standard statistical software models them. The one-hot dummy encoding can have negative consequences in the case of many categorical predictors and/or many categories, such as the risk of overfitting, or may yield estimates of coefficients with high variances (Blanquero et al., 2019; Carrizosa et al., 2016, 2017, 2022b, 2021b). In order to address the challenges above, we propose numerical methods to cluster the categories of categorical predictors, finding a reduced representation that requires fewer dummy variables, and therefore coefficients to estimate and interpret.

1.2 Sparsity in Supervised Learning

Simple linear methods like *linear regression* and *logistic regression* can be used for prediction tasks. These methods are fundamental to data analysis since they offer the advantage of being interpretable when the number of predictors is not very large (Bertsimas et al., 2016; Hastie et al., 2017; Hazimeh and Mazumder, 2020). However, interpretability can be compromised when there are many predictors at hand. In that case, it is desirable to implement methods that ensure *sparsity*. A solution is said to be sparse if only a small subset of coefficients are non-zero. Best subset selection achieves sparser solutions (Garside, 1965; Bertsimas et al., 2016; Hastie et al., 2017; Hazimeh and Mazumder, 2020) but suffer from high variability and computational difficulties (Feng et al., 2015). In contrast, coefficient shrinkage methods like the *lasso* (Hastie et al., 2015) have gained popularity for being computational efficient even with more predictors than observations. As opposed to *linear regression*, which minimize squared error loss through linear combinations of predictors, *lasso* models combine least-squares loss with an ℓ_1-regularization term penalizing the sum of the absolute coefficient values multiplied by a parameter. This has the effect of shrinking the coefficients, setting some to zero when the parameter increases (Hastie et al., 2015).

In the presence of categorical predictors which are structurally grouped, the lasso does not perform so well, so group lasso (Yuan and Lin, 2006) was proposed. It combines ℓ_1 and ℓ_2 penalties to try to remove groups of predictors all at once.

Having many predictors is not the only issue with linear models. Linearity may be too simplistic in describing the associations and structural relationships in real data (Zhao and Hastie, 2019). Additive tree based methods like *random forest* (RF) and *gradient boosting machines* (GBM) allow to estimate nonlinear relationships in the data. RF is a bagging method that builds a large collection of de-correlated classification trees, and then aggregates them (Breiman, 2001; Hastie et al., 2015). Classification trees partition the predictor space into a series of regions and then assign a class to each region. In RF we build several of these trees and each tree uses a different training sample selected by bootstrapping. A subset of the available predictors is randomly selected at each node in the tree and the best split available within those predictors is selected for that node (Breiman, 2001). After a large number of trees is generated, they vote for the most popular class for each observation, and provide this as the predicted class. Like RF, GBM is a collection of trees, but while RF builds each tree independently, boosting tries to improve the prediction for misclassified cases in the previous trees and the prediction is made by weighting the sum of the predictions made by all the tree models. At the end of each iteration, GBM performs a steepest descent minimization for logistic likelihood.

1.3 From sparsity to interpretability

According to Hastie et al. (2015) there are two reasons to use regularized methods: one is that the maximum-likelihood estimate usually has low bias but large variance, so prediction accuracy can be improved by shrinking the coefficients; the second one is that, when

we have a large number of predictors, interpretability is improved if a smaller subset with the strongest effects is identified (Molnar et al., 2020; Murdoch et al., 2019; Rudin, 2019; Rudin et al., 2022). One of the most common questions in terms of interpretation of the results of a prediction model is the importance of a component of $\mathbf{x}$. In logistic regression, since the estimated function is $log(Pr[Y = 1|X = x]) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$, β_j is a measure of the importance of x_j given that the variable is properly normalized. When dealing with black-box methods like random forest and gradient boosting machines, the importance of a variable x_j can be measured by its contribution to predictive accuracy (Zhao and Hastie, 2019). Since the random forest method uses bootstrap to select the training samples used in each tree, there will always be some observations left out of each bootstrap sample that are called the out-of-bag observations. After each tree is constructed, the values of the pth variable in the out-of-bag observations are randomly permuted and the out-of-bag misclassification rate is calculated. This is repeated for all the predictors. Variable importance is the percent increase in misclassification rate as compared to the out-of-bag rate (with all variables intact).

Complementary to variable importance, visualization tools help understand the relationship between the output of a model and the important predictors. Once we have a measure of variable importance we can try to understand the functional relationship between a reduced set of important predictors and the probability of the positive class. The Partial Dependence Plot (PDP) developed by Friedman (2001) is a tool to analyze that relationship. The partial dependence of G over a subset of variables $\mathbf{x}_S$ is defined as the expectation of G over the marginal distribution of all variables, except $\mathbf{x}_S$ (Zhao and Hastie, 2019). Several authors have shown the use of these tools to better understand the predictions of black-box models like random forest and gradient boosting machines (Goldstein et al., 2015; Zhao and Hastie,

2019; Buckmann et al., 2021). In this thesis, the fourth chapter deals with an application of Supervised Learning to the prediction of school dropout in a developing country. Variable importance and partial dependence plots are shown to be powerful tools to identify relevant predictors and relationships in the data.

1.4 Outline of the thesis

The rest of the thesis is organized as follows Chapter 2, *On Clustering Categories of Categorical Predictors in Generalized Linear Models*, based on Carrizosa et al. (2021b), proposes a methodology to cluster the categories of categorical predictors in Generalized Linear Models (GLM). The goal is to split the categories of each categorical predictor into a number of clusters, such that categories in the same cluster have a similar impact in the model and thus can be grouped together. This clustering of the categories yields to a reduced representation of the categorical predictor in which we have as many dummy variables as clusters. Our approach achieves a more compact representation of the categorical predictors and therefore reduces the number of coefficients to estimate and interpret without hurting accuracy. The proposed methodology has the following advantages. First, having fewer coefficients to interpret gives a less complex model which is a step towards enhancing the interpretability of the GLM with categorical predictors. Second, the choice of reduced representation of categorical predictors is guided by out-of-sample accuracy, with the aim that the clustered GLM preserves (or even improves) the accuracy of the original model, as illustrated in our numerical experiments. Third, we have a data driven approach to measure and visualize the similarity between categories based on a large collection of clustered models. Fourth, given that we are clustering together categories that have a similar impact in the model, we have more observations to estimate each coefficient in the clustered GLM,

which ensures lower standard errors. Fifth, our approach can also be applied to regularized linear models, i.e., instead of using GLM as the base model we could have used, e.g., a lasso (Van de Geer, 2008; Hastie et al., 2015) or a group lasso (Meier et al., 2008; Detmer et al., 2020).

Chapter 3, *A Binarization Approach to Model Interactions Between Categorical Predictors in Generalized Linear Models*, based on Carrizosa et al. (2021a), proposes a methodology to binarize the categorical predictors in Generalized Linear Models (GLM) considering that interactions may play a role in the GLM. Modelling 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 (Imai and Ratkovic, 2013; Weisberg and Pontes, 2015; Seibold et al., 2016). Other types of studies on education, health, or labour market outcomes, also commonly include interactions between socioeconomic status and characteristics like race and ethnicity (Busetta et al., 2020; Kingston et al., 2015; Howard et al., 2011; Toutkoushian et al., 2007). 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. The binarization is guided by the out-of-sample accuracy to ensure that performance is not affected, while reducing considerably the number of coefficients to estimate for main effects and interactions.

Chapter 4, *Applications of Supervised Learning to the prediction of school dropout in Colombia* uses some of the methods from Supervised Learning explored in this introduction to predict student's probability of dropout from school, using individual-level administrative data available to every education authority in Colombia. Different prediction methods

are compared in terms of their out-of-sample performance. The methods implemented are logistic regression, random forest and gradient boosting machines. To handle the categorical predictors, the method developed in Chapter 2 is applied. After the predictive performance is assessed, these predictions can be used in a real-world policy context to target students at risk of dropping out (Adelman et al., 2018), and compare it to other methods of targeting currently used by public authorities across the country. Assuming three scenarios, I compare the costs of a hypothetical intervention under a targeting approach informed by the student's dropout probability and show that this might be a cost-effective way of targeting a hypothetical program for preventing school dropout. Finally, I analyze variable importance and PDP to show important correlations in the data.

Chapter 5, *Improving fairness of Generalized Linear Models by feature shrinkage*, based on Carrizosa et al. (2022a) considers that Supervised Learning models might be trained using data that is biased against a certain group with a sensitive attribute. We propose a methodology that enhances the trade-off between accuracy and unfairness in classification by shrinking the values the predictors can take. This shrinkage depends on whether the variable is categorical or numerical. Instead of having the original representation of a categorical feature, we propose to find a reduced representation, which in the extreme case shrinks to just one dummy variable. This is achieved by a data driven clustering of the categories into a number of clusters guided by a linear combination of accuracy and unfairness. For numerical predictors, we reduce the values they can take by collapsing the tails of the empirical distribution, according to a percentile that is chosen again using a linear combination of accuracy and unfairness. In our numerical illustrations, the shrunk model, with the chosen reduced representation for all predictors in the dataset, shows a similar accuracy to the original model independently of the weight we use in the linear combination of ac-

curacy and unfairness, while fairness improves when more weight is given to it. In other words, with our approach we avoid giving the classifier information which is unnecessary for classification and might be harmful to sensitive groups.

Chapter 2

On Clustering Categories of Categorical Predictors in Generalized Linear Models

2.1 Introduction

Categorical predictors are increasingly present in classification and regression applications. In linear models, the usual way to treat them is through *one-hot dummy encoding*. The main downside when using the *one-hot dummy encoding* is that understanding the effect of each categorical predictor on the linear model requires the interpretation of the coefficients of the corresponding collection of dummy variables. In the presence of high-cardinality categorical predictors, this can be a cumbersome task. To see this, consider the *Adult* dataset that we use in our experimental section. For one of the 11 categorical predictors in the dataset, namely *Country*, we have 41 categories. This means that to explain the effect of *Country* in the model requires the interpretation of 40 coefficients, each of them with its own magnitude and sign. It would have been much easier to interpret one single coefficient, instead, as it is the case for binary predictors. There are other potential negative consequences of the one-hot dummy encoding in the presence of many categorical predictors and/or many categories, such as the risk of overfitting or having estimates of the coefficients with high uncertainty (LeBlanc and Tibshirani, 1998). To address the challenges above, we propose to cluster the categories of categorical predictors (Carrizosa et al., 2017), finding thus a reduced representation of the categorical predictors that requires fewer dummy variables, and thus coefficients.

To illustrate this, let us go back to our example from the *Adult* dataset. The predictor *Type of job* includes categories *State-gov, Federal-gov, Local-gov, Self-emp, Private, Missing*. If *State-gov* and *Local-gov* are clustered together and the rest are in another cluster, the 5 dummy variables representation for *Type of job* would be reduced to just one dummy variable. Our methodology is an iterative process, in which we cluster a categorical predictor

in each iteration. We propose a numerical method that in each iteration chooses the predictor to cluster and its best clustering. These decisions are made guided by out-of-sample accuracy of GLM in which the categorical predictors clustered in previous iterations are represented in their reduced form, an additional predictor is clustered, and the remaining ones stay as originally. Once all categorical predictors have been clustered, we train the clustered GLM. This process is repeated a number of times to obtain a series of clustered models, and the one with the best out-of-sample accuracy in a testing set is chosen, then, we report the final accuracy in an independent validation set. This collection of clustered models offers a way to measure and visualize the proximity between categories of a predictor as the percentage of these clustered GLM where the categories are together in the same cluster

The remainder of this chapter is structured as follows. The next section introduces the algorithm to cluster the categories of categorical predictors in GLM. Section 2.3 illustrates the performance of our method for a collection real-world datasets, in terms of accuracy and complexity, compared to the original one-hot encoding and regularization techniques (lasso and group lasso). Finally, we show the proximity measure between categories of a categorical predictor derived from our method. Conclusions and future research are collected in Section 2.4.

2.2 Methodology

In this section, we present our approach to clustering the categories of categorical predictors in Generalized Linear Models (GLM). We first introduce the notation for the GLM with categorical predictors. We then introduce the algorithm for clustering the categories, yielding a representation of each categorical predictor j with K' dummy variables. With

this, we build the so-called clustered GLM in which each categorical predictor is modeled using its reduced representation. We end the section by discussing how to get insights into the relationship between categories of a categorical predictor using a measure of proximity that stems from our numerical method.

We are given a training sample of size N. We have J categorical as well as P continuous predictors. Categorical predictor j has K_j categories, $j = 1, \ldots, J$. In the GLM using the traditional one-hot encoding, a categorical predictor j with K_j categories is represented by $K_j - 1$ dummy variables, one for each category, leaving one out for contrast. We denote by **d** the vector of dummy variables associated with the categorical predictors, while **x** denotes the vector of continuous predictors. Consider a GLM where the outcome y is related to **d** and **x** through a link function G, namely,

$$\mathbb{E}\left[y|\mathbf{d}, \mathbf{x}\right] = G(\beta_0 + (\boldsymbol{\beta})^T\mathbf{d} + (\tilde{\boldsymbol{\beta}})^T\mathbf{x}), \tag{2.1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is the set of model parameters for the dummy variables and $\tilde{\boldsymbol{\beta}}$ the one for the continuous predictors. For a binary response variable $y \in \{0, 1\}$, a natural choice of link the function G is the Logit, while for a count response variable $y \in \{0, 1, 2, \ldots\}$, it would be the Log Link function. Both link functions will be illustrated in Section 2.3, but our approach can handle any other link function.

We will now explain how the clustering of categories for a given categorical predictor will be performed. To reduce the possible clusterings of categories for a categorical predictor, we will assume that the categories are ordered. For instance, for ordinal categorical predictors like the level of *Education* from the *Adult* dataset in Section 2.3, we could take the *natural* order of the predictor as the order of the categories (see Table 2.1). For non-ordinal categorical predictors, we could use the coefficients from the GLM with one-

	Feasible clusterings														
1st-4th	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5th-6th	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7th-8th	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
9th	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
10th	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
11th	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12th	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
HS-grad	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Some-college	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Assoc-voc	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
Assoc-acdm	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Prof-school	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
Bachelors	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Masters	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
Doctorate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 2.1: Feasible clusterings for the predictor *Education* in the *Adult* dataset with $K' = 2$

hot encoding to order the categories.We do this because the coefficients indicate how each category affects the response, so categories that are more similar will be closer together.

Now let us define the concept of feasible clustering for categorical predictor j: we will say that a clustering of the categories of j into K' clusters is feasible if each of the clusters consists of consecutive categories. For the *Education* level predictor, with $K_j = 15$, there are 15 feasible clusterings with $K' = 2$ clusters (see Table 2.1). The first clustering corresponds to having the lowest level of education (*1st-4th*) in one group, and the remaining levels in another one. The second clustering corresponds to having the first and second level of education (*1st-4th*, *5th-6th*) in one group and the remaining levels in another one. We successively move to include higher levels of education one by one in the first cluster, until we reach the last level, where all the categories are together in the first cluster and the second cluster is empty. This last clustering is equivalent to removing predictor j entirely from the model.

Once we have a feasible clustering, we can obtain the reduced representation of the categorical predictor with K' dummy variables, where we have a dummy variable for each of the K' clusters. For $K' = 2$, the reduced representation consists of one single dummy variable indicating whether the category belongs to the first cluster or not. This is illustrated in Table 2.1 for the categorical predictor *Education.*

We next argue the necessity to use a randomized numerical method to decide which feasible clustering will be used in our clustered GLM for each categorical predictor. We might decide that the best choice is that one with which the GLM achieves the highest accuracy in a test set. However, clusterings that differ little may yield a very similar accuracy. In the presence of multiple predictors to be clustered, it may be desirable to choose a clustering with good out-of-sample accuracy, but not necessarily the best one. Therefore, we design a Greedy Randomized Adaptive Search Procedure (GRASP) that chooses randomly between the feasible clusterings with the highest out-of-sample accuracies, see Figure 2.1.

GRASP is a class of numerical methods that have been successfully applied to a number of optimization problems (Resende and Ribeiro, 2016). GRASP deals with optimization problems in which a collection of decisions needs to be made, and for each decision, one has a number of feasible actions. The decisions have an associated payoff, and the goal is to maximize the payoff. GRASP makes in each step a new decision in a random fashion, choosing from the top $h\%$ payoffs. GRASP is repeated m times, and the solution with the best payoff across the m iterations is returned. In our case, GRASP needs to decide in each step which categorical predictor to cluster, the feasible actions are the feasible clusterings for the corresponding predictor, while the payoff is the out-of-sample accuracy of the GLM where the categorical predictor at hand is represented by the dummy variables associated with the feasible clustering. Once the feasible clustering is chosen, the predictor is clustered

Initialization: Let $\mathcal{L} \subseteq \{1, 2, \ldots, J\}$ be the set of categorical predictors to be clustered. Let **d** be the vector of dummy variables in the one-hot-encoding and **x** the vector of continuous predictors. Let $K' < K_j - 1$, for all $j \in \mathcal{L}$;
for $i \in \{1, \ldots, m\}$ **do**
  Set $\mathcal{L}' = \mathcal{L}$;
  while $\mathcal{L}' \neq \emptyset$ **do**
    for $j \in \mathcal{L}'$ **do**
      for *each feasible clustering of predictor* j *with* K' *clusters* **do**
        Estimate the GLM in (2.1) where predictors in $(\mathcal{L} \setminus \mathcal{L}') \cup \{j\}$ are clustered;
        Calculate its out-of-sample accuracy;
      end
      Return: $\mathcal{V}_j$, the set of out-of-sample accuracies;
    end
    Let $\mathcal{V} = \cup_{j \in \mathcal{L}'} \mathcal{V}_j$ be the merged set of out-of-sample accuracies;
    Sort $\mathcal{V}$ from max to min;
    Choose randomly one accuracy from the top h% ones in $\mathcal{V}$. This accuracy is linked to $s \in \mathcal{L}'$ and a reduced representation with K' dummy variables;
    Replace the $K_s - 1$ dummy variables of s by the new K' dummy variables and eliminate s from $\mathcal{L}'$;
  end
  Return: GLM_i^C, the GLM in (2.1) where predictors in $\mathcal{L}$ are clustered;
end
Return: The clustered GLM, GLM_i^C with the highest out-of-sample accuracy

Figure 2.1: Pseudocode for the GRASP algorithm

for the steps to come. In this way, after all categorical predictors have been clustered into K' clusters, we can build the corresponding clustered GLM. These steps are repeated m times, and GRASP returns the best clustered GLM across the m iterations performed in terms of out-of-sample accuracy in the test set. Then we can report the accuracy in an independent validation set.

We will now describe how from the m iterations of GRASP, we do not only obtain a less complex model, which hopefully has a similar or even better accuracy, but we can also

derive a proximity measure for the categories of each categorical predictor. For a given categorical predictor j, we can exploit the m clustered GLMs built across all iterations to measure the proximity between categories c and d . We define the proximity between categories c and d as the percentage of these clustered GLMs where c and d are together in the same cluster. The higher this proximity, the closer is their impact on the GLMs, and the more plausible to cluster them together. In the next section, we will visualize these proximities to better understand the underlying structure between the categories.

We end by noting that the methodology proposed in this section can be easily extended to discretize continuous predictors (Carrizosa et al., 2010) or to have as a base model a regularized linear model, such as the lasso or the group lasso.

2.3 Numerical Illustrations

In this section, we illustrate how our methodology performs in several real-world datasets. Our aim is to empirically analyse the effect of clustering categories in a baseline classification or regression method in terms of accuracy and relative complexity. As baseline procedures for classification we have chosen logistic regression, lasso logistic regression, and group lasso logistic regression. We cluster all our categorical predictors into $K' = 2$ clusters. Accuracy is measured by the correct classification rate and relative complexity is the number of estimated coefficients for the categorical predictors compared to the number of estimated coefficients in the original model. In other words, the relative complexity of the clustered model is $\frac{J}{\sum_{j=1}^{J} K_j - J} \cdot 100\%$, note that the numerator will be smaller if some of the coefficients are equal to zero. Accuracy estimates are obtained as follows: the data set is split into training sample (70%) a testing sample (15%) and a validation sample (15%). The model is built in the training sample, we choose a clustering using the out-of-sample

performance in the testing sample and we report its final accuracy in the validation sample. The process is repeated ten times and we report the average out-of-sample accuracy. We have also empirically analysed the effect of clustering categories in a baseline regression method, namely Poisson regression for count data, in terms of accuracy (measured as Root Mean Square Error) and relative complexity.

Our method uses a GRASP algorithm, as described in Figure 2.1. In our experiments, the GRASP parameters are set to $m = 100$ iterations, and selection is done out of the top 3 accuracies for categorical predictors with more than 5 categories and out of the top 2 otherwise. For lasso and group lasso, we perform ten-fold cross-validation to select the shrinkage parameter. We report accuracy in the same validation set that we used for our clustered model. For group lasso, the categories associated with each categorical predictor are part of the same group. We coded our method in R and conducted our experiments in a Workstation with an Intel® Core™ i5-4460 processor with 8 Gb of RAM.

The rest of the section is organized as follows. The datasets are described in Section 2.3.1, performance in terms of accuracy and relative complexity is discussed in Section 2.3.2, and proximity graphs are shown in Section 2.3.3.

2.3.1 Datasets

We use eight real-world classification datasets to illustrate the method, which are available in the UCI Machine Learning Repository (Dua and Graff, 2017). We also use one count-data dataset, which is available in Deb and Trivedi (1997). The datasets are described in Table 2.2, in the first two columns we report the name and the total number of records in the dataset (N). In columns three to seven, we report the class split in percentage, the number of categorical (J) and continuous predictors (P), the total number of categories ($\sum_{j=1}^{J} K_j$)

Name	N	Class-split	J	P	$\sum_{j=1}^{J} K_j$	K_j
Solar	1066	83/17	5	5	23	7,6,4,3,3
Coil-2000	5822	94/6	5	80	77	41,6,10,10,10
Nursery	12960	33/66	7	1	25	3,5,4,4,3,3,3
Mushrooms	8124	48/52	17	4	111	6,4,10,9,4,3,12,4,4,9,9,4,3,8,9,6,7
Bank marketing	4119	89/11	6	14	42	12,4,8,10,5,3
Car evaluation	1728	30/70	6	0	21	4,4,4,3,3,3
Adult	32561	24/76	11	3	117	5,8,5,16,5,7,14,6,5,5,41
German	1000	30/70	11	9	52	4,5,11,5,5,5,3,4,3,3,4
DebTrivedi	4406	—	5	5	25	3,5,7,4,6

Table 2.2: Description of the classification datasets (first eight ones) and regression dataset (last one)

across all categorical predictors, and the number of categories for each categorical predictor K_j, respectively. Please note that *Coil-2000*, *Car Evaluation*, *Solar* and *Nursery* are converted into two-class problems using the majority class against the rest.

2.3.2 Performance in terms of accuracy and relative complexity

In this section, we discuss the accuracy and the relative complexity. We start with the classification task. We will say that the original and the clustered models give comparable results in terms of accuracy if the difference is below 1 percentage point (p.p.). Table 2.3 and Figure 2.2 report the mean validation accuracy across the ten reshuffles for accuracy as well as the relative complexity of clustered model with respect to the original one, we do not include a column for relative complexity of the original model since this is by definition equal to one.

Let us begin with the accuracy criterion. The accuracy of the clustered model is comparable to that of the original model for six datasets (*Solar, Coil-2000, Nursery, Mushrooms, Bank marketing, Adult*). The original model outperforms the clustered model for

Dataset	Accuracy (%)				Relative Complexity (%)		
	Original	Clustered	Group lasso	Lasso	Clustered	Group lasso	Lasso
Solar	83.69	84.00	83.63	83.63	6.25	0.00	18.75
Coil2000	94.28	94.36	94.67	94.67	6.12	0.00	0.00
Nursery	100.00	100.00	100.00	100.00	38.89	11.11	11.11
Mushrooms	99.98	99.42	99.53	99.89	18.92	89.19	25.68
Bank Marketing	91.60	91.89	90.57	91.28	8.57	68.57	8.57
Car evaluation	95.12	92.96	94.98	94.83	40.00	100.00	100.00
Adult	85.05	84.37	83.80	84.74	11.29	101.61	59.68
German	76.78	75.13	72.19	75.00	19.44	69.44	41.67

Table 2.3: Accuracy and Relative complexity (the number of non-zero coefficients estimated for the categorical variables relative to the total number of estimated coefficients in the original model) in the validation set, for the original, clustered, lasso, and group lasso models

two datasets (*Car evaluation* and *German*), but only by 2.16 p.p. and 1.3 p.p. respectively. Compared to regularization methods, our clustered model outperforms group lasso in *Bank Marketing* and *German* and is similar to lasso in seven out of eight datasets (*Solar, Coil-2000, Nursery, Mushrooms, Bank marketing, Adult, German*). In the *Car evaluation* dataset, group lasso and lasso outperform our clustering method by 2.02 p.p. and 1.86 p.p. respectively.

We will now examine the relative complexity of the clustered model in comparison to the regularization methods. On five of eight datasets, our clustered model has lower complexity than either lasso or group lasso (*Mushrooms, Bank, Car evaluation, Adult, German*). The complexity of the clustered model is greater for three datasets (*Solar, Coil2000, and Nursery*) compared to lasso and group lasso. In Figure 2.2 we see that, in general, our clustered models can compress information more efficiently (lower relative complexity) than traditional regularization techniques, without sacrificing accuracy.

Finally, our method can be used for other link functions for the GLM. We show how it performs for a count data example. The baseline method is poisson regression, we compare

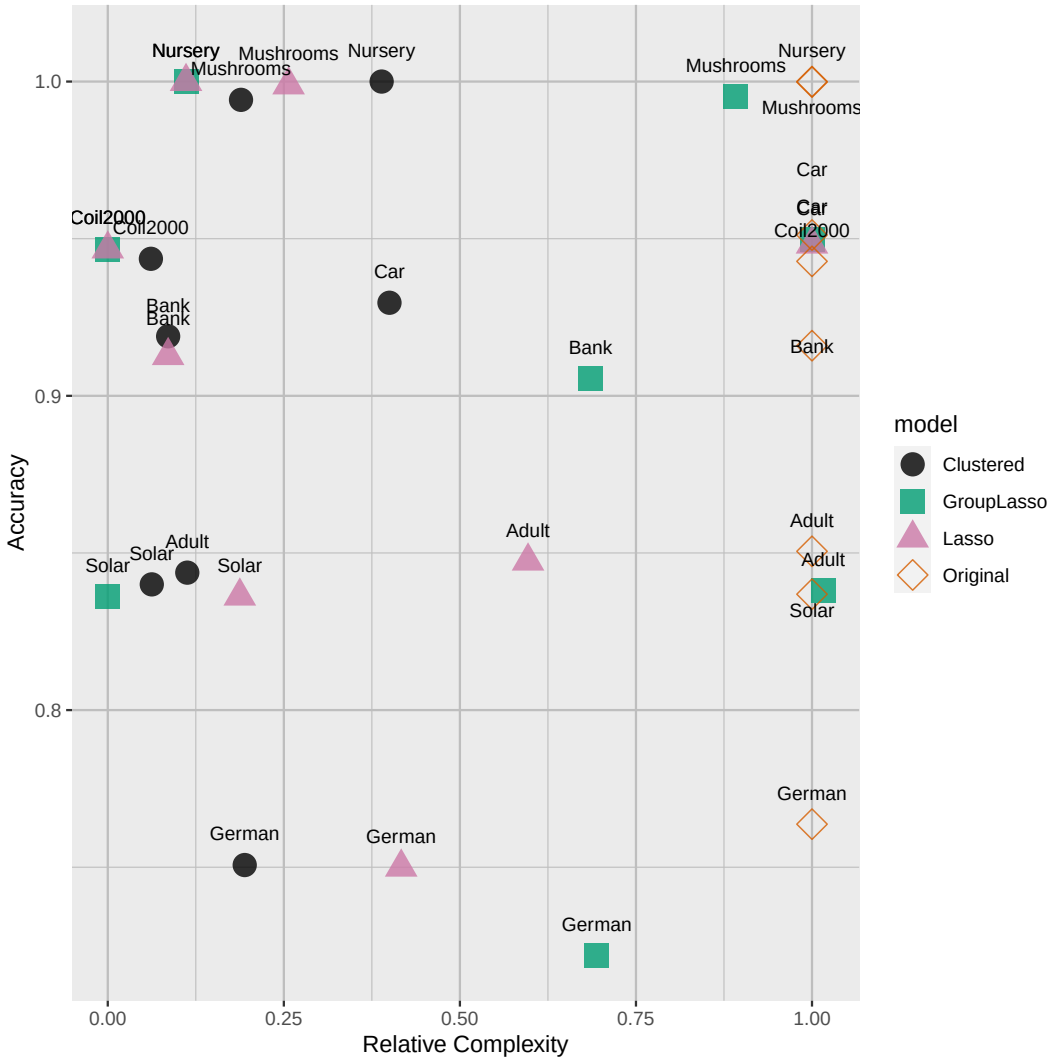


Figure 2.2: Accuracy and Relative complexity (the number of non-zero coefficients estimated for the categorical variables relative to the total number of estimated coefficients in the original model) in the validation set, for the original, clustered, lasso and group lasso models

Name	RMSE			Relative Complexity (%)	
	Original	Clustered	Lasso	Clustered	Lasso
DebTrivedi	4.78	4.84	4.81	20	80

Table 2.4: RMSE and Relative complexity (the number of non-zero coefficients estimated for the categorical variables relative to the total number of estimated coefficients in the original model) in the validation set, for the original, clustered, and lasso

the original model with the clustered model and lasso poisson regression, group lasso is not implemented. In this case, we measure accuracy by the Root Mean Square Error (RMSE), the squared root of the mean of the squares of the differences between observed and predicted values. We see there is a very small worsening of the RMSE from 4.78 to 4.84, while the number of estimated coefficients is 20% of the ones we would have to estimate with the original model, compared to lasso, complexity is reduced more by the clustered model. Hence, as in the previous results for classification, for the *DebTrivedi* dataset using Poisson regression, we achieve a lower level of complexity with a slight increase in error given that the number of categories/categorical predictors is small.

2.3.3 Proximity graphs

In this section, we illustrate the results of the proximity measure between categories of a categorical predictor presented in Section 2. The GRASP algorithm is repeated m times, out of these m repeats we take only the one that gives the highest accuracy on the test set and that will be our final clustered GLM. Here we dig deeper into the information from all iterations to recover valuable insights from the data, which reflects the underlying structure of the categories and their relationship in terms of proximity. We show the proximity graphs for the predictors *Education*, *Occupation*, and *Type of Employer* from the *Adult* dataset. Recall that we measure proximity as the percentage of the m clustered GLMs where two categories

are together in the same cluster. In the graphs, each category is a node and the thickness of the edges represents the proximity between categories.

Figure 2.3 shows the proximity graph for the predictor *Education* in the *Adult* dataset. Recall from Section 2, that this predictor is ordinal, so the categories are sorted according to its natural order: lower levels of education first and higher levels last. It is interesting to see that the cut between the two levels of education established by GRASP is well defined, with categories that represent finished university education and beyond on one cluster (left side of the graph), and those that represent unfinished and no university at all in the other (right side of the graph). In the middle, we find two categories (*Associate vocational* and *Associate-acdm*), which are deemed as tertiary education in international classifications but have proportionally more edges on the right-hand side.

Figure 2.4 shows the predictor *Occupation*, for which the categories were ordered according to the coefficients of the Logistic Regression with one-hot encoding. For this predictor, we can see more edges between categories, but in general, the cut is established between lower paying occupations on the upper part of the graph (*Farming-fishing, Priv-housing-serv, Handlers-cleaners, Transport-moving, Machine-op-insp*) and higher-paying occupations (*Exec-managerial, Protective-serv, Tech-support, Sales, Craft-repair, Prof. specialty*). Some occupations like *Admn-clerical* and *Transport-moving* have similar proximity to some of the lower-paying occupations as well as the high-paying occupations.

In Figure 2.5 we can see the predictor *Type of employer*. The order for the categories also comes from the coefficients of the Logistic Regression with one-hot encoding. The cut is less clear in terms of establishing “high” or “low” paying types of employer, but we can see that categories *Federal, Self-employed, Private and Local Government* are closer to each other, and *State government* and *Self-empl-not-inc* are closer with *misLevel*.

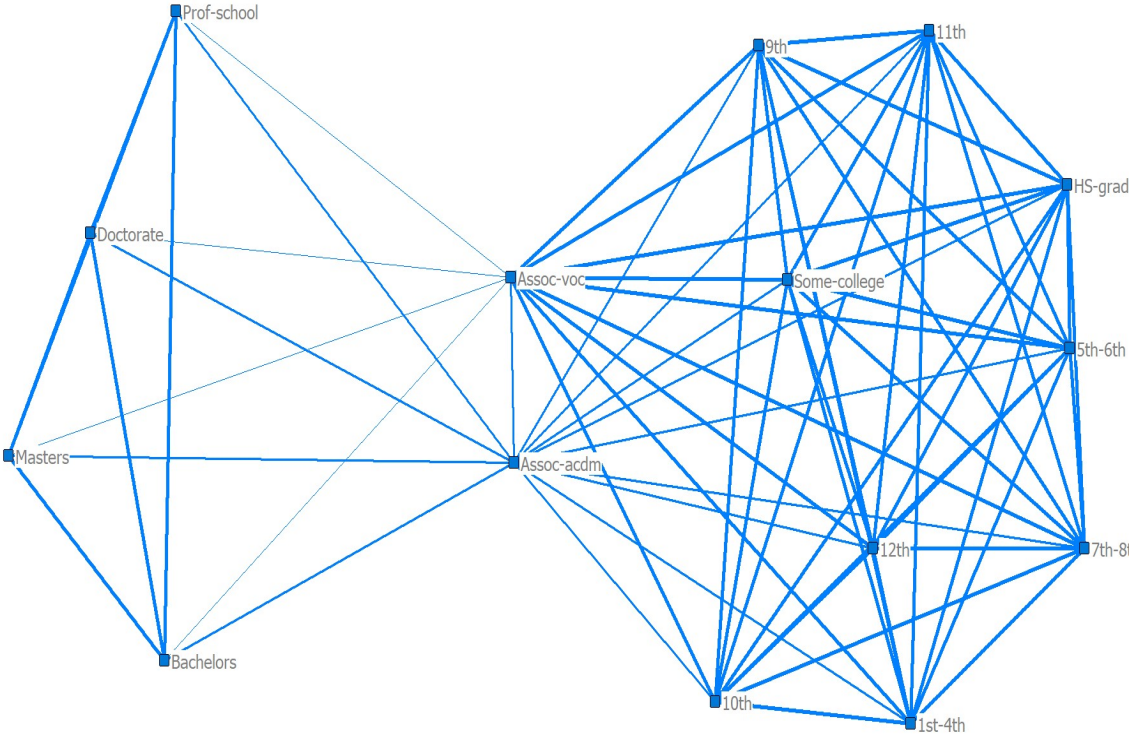


Figure 2.3: Proximity Graph for the predictor *Education* in the *Adult* dataset

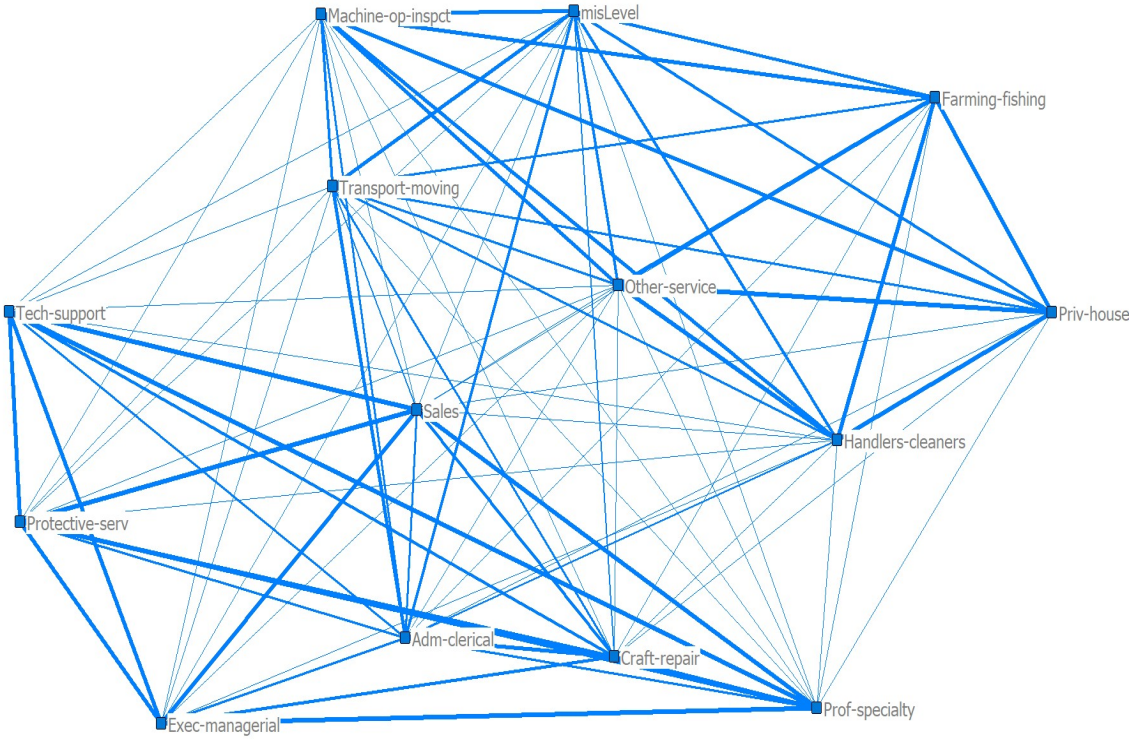


Figure 2.4: Proximity Graph for the predictor *Occupation* in the *Adult* dataset

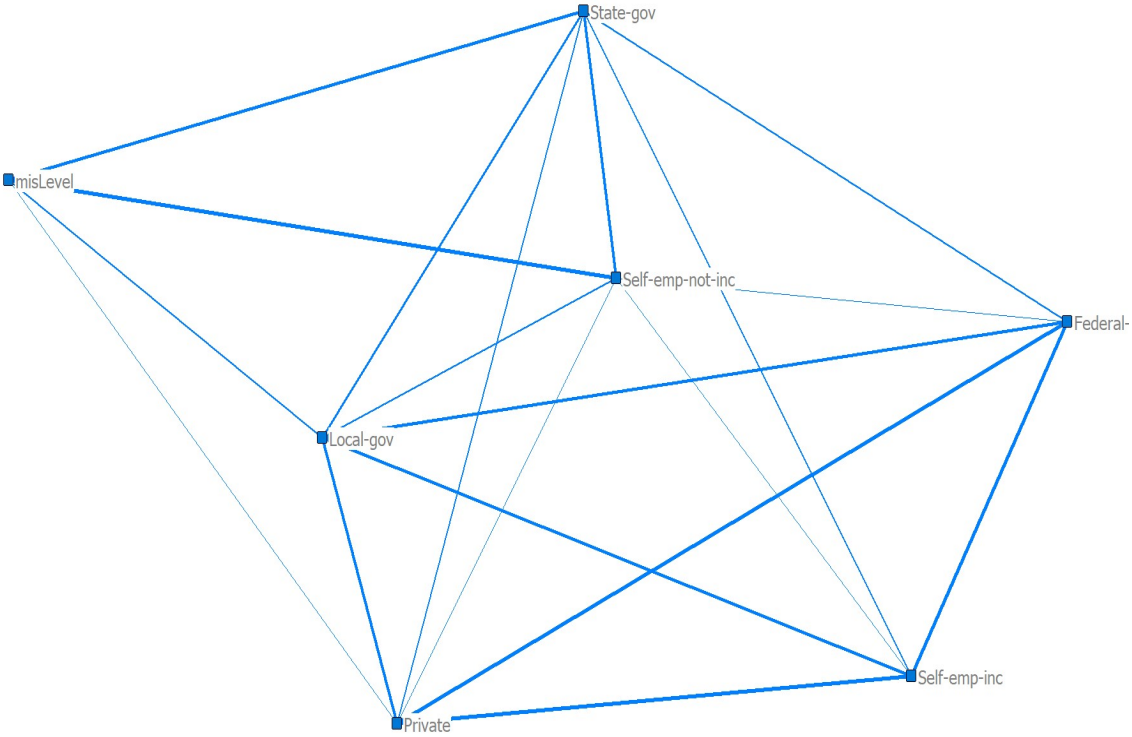


Figure 2.5: Proximity Graph for the predictor *Type of employer* in the *Adult* dataset

2.4 Conclusions

In this chapter, we developed a numerical method to reduce the complexity of Generalized Linear Models in the presence of categorical predictors, by clustering their categories into K' clusters. Our approach has two main advantages. First, the Clustered GLM has fewer coefficients to be estimated, and therefore it is easier to interpret the impact of each categorical predictor, especially when $K' = 2$ and each categorical predictor is represented by one dummy variable. Second, after clustering categories, the number of observations for each cluster is higher than for the original categories, allowing for a better estimation of the coefficients. Our numerical results illustrate that it is possible to cluster categorical predictors, with a corresponding reduction in complexity, without compromising accuracy and that our method is competitive with coefficient shrinkage methodologies like lasso and group lasso.

In terms of future research, it might be interesting to explore interactions (Carrizosa et al., 2011) between categorical predictors which in the case of high-cardinality yields a highly combinatorial problem that needs attention in the future. In addition, other definitions of feasible clusterings of the categories of a categorical predictor can be useful, especially for cases where some categories might have too few observations to obtain accurate estimates of their coefficients, or in the presence of user-defined constraints on the shape of these clusters.

Chapter 3

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

3.1 Introduction

Modelling interactions between categorical predictors becomes challenging (Lim and Hastie, 2015), and can hinder interpretability (Blanquero et al., 2019; Carrizosa et al., 2022b, 2016, 2017) when the number of categorical predictors and/or categories is large. The simplest case of an interaction is the one given by two binary predictors. Consider the real-world *German* credit dataset used in our numerical section, 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 that 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 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* $= 1$.

Consider now that instead, we have two categorical predictors. Using the example above, consider *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 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.

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 dummy variables associated with *Job*, we would have just one, indicating whether the individual shows any category of the first group. Similarly, instead of 11 dummy 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, we reduce from 379 to 34 the number of coefficients associated with all categorical predictors and their interactions in the *German* dataset.

In this chapter, we propose a methodology to binarize the categorical predictors in Generalized Linear Models (GLM) considering that interactions may play a role in the GLM.

Without interactions, in Chapter 2 we proposed a methodology to cluster categories using the natural order of the categories in ordinal predictors or, otherwise, the order defined by the coefficients of a GLM. This approach is not applicable here since we want to consider the role of interactions, which would mean we estimate a GLM in the presence of both marginal as well as interaction effects, and hence we have a more complex relationship between the categories and the response. Given the difficulties in considering a model with all possible interactions, we propose an iterative algorithm that will consider only the interactions between a pair of categorical predictors at a time.

Our approach to binarizing the categorical predictors 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 Bien et al. (2013); Lim and Hastie (2015), 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 LeBlanc and Tibshirani (1998) and Carrizosa et al. (2022b).

The rest of the chapter is organized as follows. Section 3.2 introduces the algorithm to binarize the categorical predictors using information from the main effects and the interactions. Section 3.3 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 3.4.

3.2 Methodology

In this section, we detail the methodology to find a reduced representation of categorical predictors as binarized 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 dummy variable. The binarized predictors will be used to build 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 $K_j - 1$ dummy 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 Lim and Hastie (2015). Consider a GLM where the outcome Y is related to X, comprising the predictors and their interactions, through a link function G:

$$\mathbb{E}\left[Y|X\right] = G\Big(\alpha + \sum_{j=1}^{J} X_j \cdot \beta_j + X_{j:l} \cdot \Theta_{j:l}\Big), \tag{3.1}$$

where α is the intercept, X_j is the vector of dummy 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

$$X_{j:l} = \begin{pmatrix} X_{jb} & X_{jc} \end{pmatrix} * \begin{pmatrix} X_{lb} & X_{lc} & X_{ld} \end{pmatrix} = \begin{pmatrix} X_{jb:lb} & X_{jb:lc} & X_{jb:ld} & X_{jc:lb} & X_{jc:lc} & X_{jc:ld} \end{pmatrix},$$

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 (3.1) 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 3.3. 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 3.3, but for the sake of notational simplicity, we have decided not to include them in (3.1).

We will now explain how the binarization of a given categorical predictor, say s, is performed. For ordinal categorical predictors, we apply the approach in Carrizosa et al. (2021b) since there is a natural order in the categories. For non-ordinal categorical predictors, in the presence of interactions, we have a more complex relationship between the response variable and the predictors. Thus, we will inspect the marginal effects and the interactions in (3.1) to build a dissimilarity matrix which can then be used in a clustering procedure to find two clusters for each categorical predictor.

Now let us define 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 a GLM with (3.1) and comparing the marginal coefficients for b and c, as well as the coefficients associated with the interactions for these categories. To be more precise, we define the dissimilarity between b and c as follows:

$$\delta_s(b,c) = (1-\lambda)\delta_s^{mar}(b,c) + \lambda\delta_s^{int}(b,c), \tag{3.2}$$

where $\delta_s^{mar}(b,c) = |\beta_{jb} - \beta_{jc}|$ is the difference, in absolute terms, between the pair of marginal coefficients for *b* and *c*, and $\delta_s^{int}(b,c)$ is the ℓ_1 distance between the two interaction coefficient vectors. We place more weight on the information provided by the interaction coefficients the higher the value of $\lambda \in [0,1]$.

We can define the dissimilarity between other pairs of categories of s in an analogous way, denoting by $\boldsymbol{\delta}_s$ the corresponding dissimilarity matrix, which contains the dissimilarities for all pairs of categories in s. In Figure 3.1 we show an illustration of how to binarize categorical predictor *Job* from the *German* dataset, see Table 3.2 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 a GLM using (3.1) with all marginal 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 Panel 3.1a. Then, we calculate the dissimilarity matrix $\boldsymbol{\delta}_{Job}^{(1)}$ using (3.2) with $\lambda = 0.5$, see Panel 3.1b. We apply a hierarchical clustering procedure with the resulting clusters shown in Panel 3.1c. With this, we find a reduced representation of predictor *Job* as a dummy variable that takes on value 1 if *Job* is equal to *Unemployed* or *Unskilled* and 0 otherwise.

With $\boldsymbol{\delta}_s$, and using a clustering procedure, we can cluster the categories of s into two

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*

$$\boldsymbol{\delta}^{(1)}_{Job} = \begin{pmatrix} & \textit{Unemployed} & \textit{Unskilled} & \textit{Skilled} & \textit{Management} \\ \textit{Unemployed} & 0.00 & 1.22 & 1.53 & 2.14 \\ \textit{Unskilled} & & 0.00 & 1.29 & 1.72 \\ \textit{Skilled} & & & 0.00 & 1.35 \\ \textit{Management} & & & & 0.00 \end{pmatrix}$$

(b) Dissimilarity between the categories of *Job*, using (3.2) with $\lambda = 0.5$

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

(c) Binarization of *Job* using the dissimilarity $\boldsymbol{\delta}^{(1)}_{Job}$.

Figure 3.1: Binarization steps for categorical predictor *Job* from the *German* dataset

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 dummy variable, where all categories in the same group now affect the response variable in the same way.

Our goal is to try out different binarizations of predictor s in order to find a good one. The dissimilarity matrix $\boldsymbol{\delta}_s$ depends on which interactions are incorporated in (3.1). 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, $\boldsymbol{\delta}_{Job}^{(j)}$. Then, we would have $J-1$ different binarizations for the same categorical predictor that we could choose from, based on accuracy in a test set. After making the choice and binarizing the predictor using the corresponding clustering, we incorporate this reduced representation in the next decision to make.

The pseudocode of our approach can be found in Figure 3.2. 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 GLM with (3.1) which includes all marginal effects and the interactions between the categories of s and one other 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 step 18, we estimate the GLM, in a similar fashion as before, but here we have s binarized. The binarization of s that gives the highest accuracy in a test set 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 build, in line 21, the binarized GLM, GLM_i^B, including all binary predictors and 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 highest out-of-sample

Initialization: Let $\mathcal{L} \subseteq \{1, \ldots, J\}$ be the set of categorical predictors to binarize;
Let m be the repeats of the algorithm;
Let $\lambda \in [0, 1]$ be the weight parameter in (3.2);
for $i \in \{1, \ldots, m\}$ **do**
    Set $\mathcal{L} = \mathcal{L}$;
    while $\mathcal{L} \neq \emptyset$ **do**
        Randomly sample one predictor from $\mathcal{L}$, s;
        Let $\mathcal{V}_s$ be the set out-of-sample accuracies for the binarizations of s;
        Set $\mathcal{V}_s = \emptyset$;
        for $j \in \{1, \ldots, J\} \setminus \{s\}$ **do**
            Estimate a GLM with (3.1) which includes all marginal effects and the interaction between s and j;
            Calculate the dissimilarity matrix for s, $\boldsymbol{\delta}_s^{(j)}$, using (3.2);
            Use $\boldsymbol{\delta}_s^{(j)}$ in a clustering procedure to split the categories of s into two groups;
            Binarize categorical predictor s;
            Estimate a GLM as in Step 11, considering s binarized;
            Calculate its out-of-sample accuracy and add it to $\mathcal{V}_s$;
        end
        Choose the maximum out-of-sample accuracy in $\mathcal{V}_s$ and binarize s accordingly;
        Eliminate s from $\mathcal{L}$;
    end
    Return: GLM_i^B, the GLM in (3.1) where the predictors in $\mathcal{L}$ are binarized;
end
Return: The binarized GLM, GLM_i^B with the highest out-of-sample accuracy.

Figure 3.2: Pseudocode for the binarization algorithm of categorical predictors, considering interactions

accuracy.

3.3 Numerical illustrations

In this section, we illustrate how our binarization methodology for categorical predictors performs in terms of accuracy and interpretability. We report these two performance metrics for the original model with interactions and the binarized model with interactions. We use as a baseline the logistic regression, and measure final accuracy as the correct classification

rate in a validation sample. Accuracy 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 a clustering using the out-of-sample performance in the testing sample and we report its final accuracy in the validation sample. The process is repeated ten times and we report as an estimate the average out-of-sample accuracy. In terms of interpretability, we report the relative complexity as the number of estimated coefficients for the categorical predictors and their interactions compared to the number of estimated coefficients for the categorical predictors and their interactions in the original model. We end with the discussion of the binarized predictors for the German dataset and their coefficients in the binarized GLM with interactions.

We use four real-world datasets available at the UCI Machine Learning Repository (Dua and Graff, 2017). We will show that the original model with interactions has a poor accuracy performance since the number of coefficients to estimate is very large compared to the number of records available. However, once the predictors are binarized with the algorithm in Figure 3.2, we are able to train the model with interactions. This much smaller model with fewer coefficients can be seen as the final result, or even as the starting model, to be 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 the original and the binarized models, guided by AIC.

We also use a simulated dataset, with the purpose to illustrate that our procedure is able to discover the relevant interactions in the generating model.

The algorithm in Figure 3.2 has two parameters, namely the number of iterations m and the weight λ. We set those to 200 and 0.5, respectively, after running a sensitivity analysis. For lasso and group lasso, we perform ten-fold cross-validation to select the shrinkage

Dataset	N	Class split (%)	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

Table 3.1: Description of the datasets used to test the binarization algorithm

parameter. We report accuracy in the same validation set that we used for our binarized model. For group lasso, we implement the version in Lim and Hastie (2015) 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.

The rest of this section is organized as follows. Section 3.3.1 describes the datasets, Section 3.3.2 is devoted to the analysis of the real-world dataset, and Section 3.3.3 to the simulated dataset.

3.3.1 Datasets

Table 3.1 shows the descriptive statistics for both the simulated and the real-world datasets. 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 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 model with interactions, 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

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

Table 3.2: *German* dataset: Description of the categorical predictors

terms of creditworthiness. We have 967 records with 13 categorical predictors. Table 3.2 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 need binarization. The three ordinal predictors have been binarized using the methodology in Carrizosa et al. (2021b). The eight remaining ones are binarized using the algorithm in Figure 3.2.

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 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 Figure 3.3. The groups of categories are clear when we observe these coefficients. For example, categories

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

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

Figure 3.3: *Simulated* dataset: Coefficients in the data generating model

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. We will show that our algorithm is able to recover this equivalent binary generating model.

3.3.2 Real-world datasets

In this section, we illustrate the performance of our binarization procedure for the real-world datasets. The accuracy and the number of coefficients associated with the categorical predictors can be found in Table 3.3. We can see that for all datasets, having the original model with interactions gives a lower accuracy than the binarized model with interactions. 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

Measure	Model	*Coil2000*	*Bank marketing*	*Adult*	*German*
Accuracy	Original without interactions	94.28	91.6	85.05	76.78
	Original	89.68	83.93	79.20	62.19
	Binarized	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	Original without interactions	12.41	7.68	16.36	4.68
	Binarized	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

Table 3.3: Real-world datasets: Accuracy and Relative complexity (the number of non-zero coefficients estimated for the categorical variables relative to the total number of estimated coefficients in the original model with interactions) in the validation set, for the original, binarized, lasso and group lasso models

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 learning from this model very challenging. We can see that accuracy goes from 89.68% to 93.62% for *Coil2000*, from 83.93% to 92.10% for *Bank marketing*, from 62.19% to 76.44% for *German* and from 79.20% to 84.28% for the *Adult* dataset, after we include the interactions. Therefore, with our binarization procedure, we are able to model the interactions, as well as work with a much smaller model with less than 10% coefficients associated with the categorical predictors and their interactions, again after deleting those for which we do not have data. Comparing our algorithm 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 complexity, in three out of four datasets, our method results in a smaller model. We can see these results graphically in Figure 3.4.

For the binarized model with interactions in the German dataset, Figure 3.5 reports for

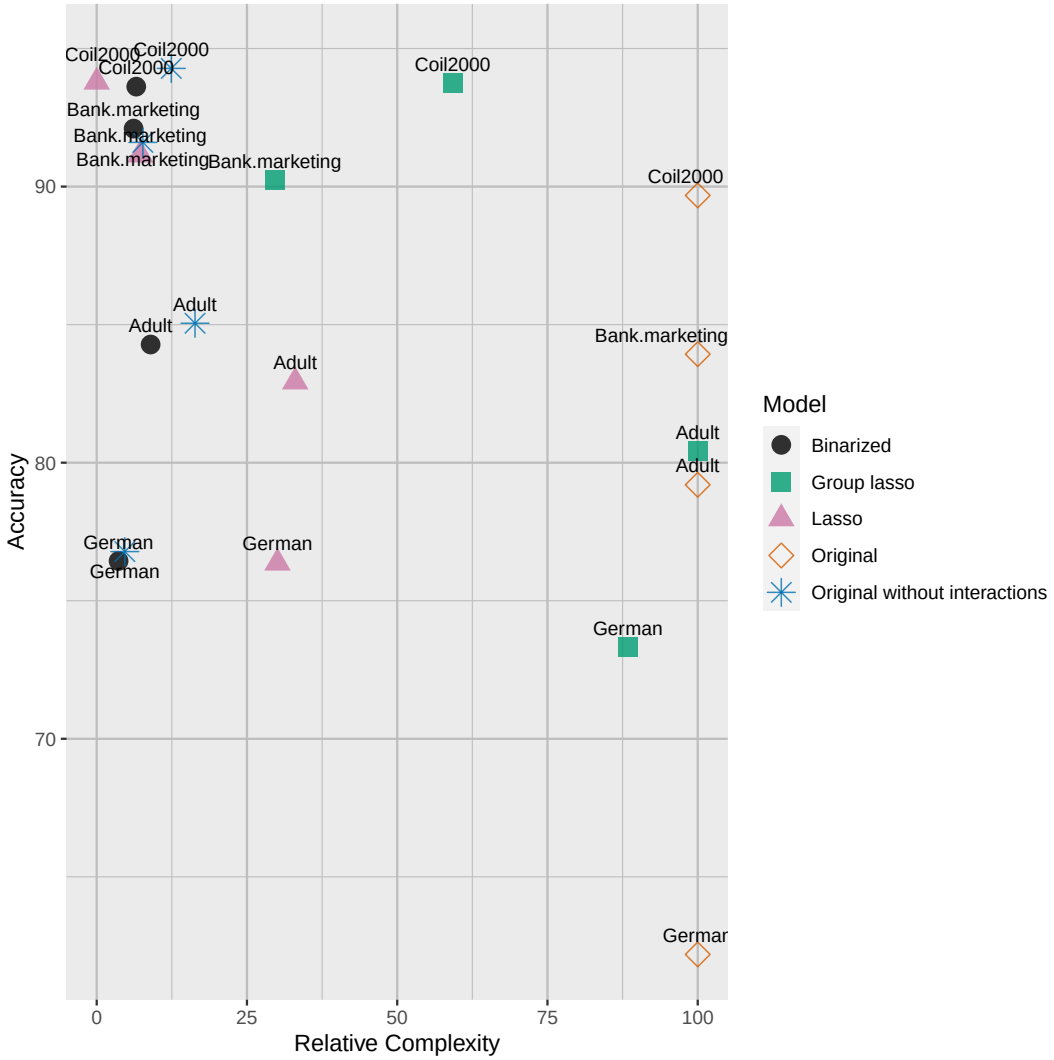


Figure 3.4: Accuracy and Relative complexity (the number of non-zero coefficients estimated for the categorical variables and their interactions relative to the total number of estimated coefficients for these variables in the original model) in the validation set, for the original, binarized, lasso and group lasso models, the original model without interactions is added for reference

each categorical predictor the two clusters of categories yielding the alternative representation as a binarized predictor. In Figure 3.5, we can see how the algorithm has clustered each category into two clusters. For instance, let us look at the first (X_1) and the last (X_{11}) categorical predictors to be binarized, namely *Status of existing checking account* (with 4 categories) and *Job* (with 4 categories). For X_1, binarized as B_1, cluster 1 contains one category (... *<0 DM*) and cluster 2 the remaining three (... *<200 DM, ... >= 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 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 3.6 helps us visualize these coefficients.

Figure 3.6 provides information about the coefficients of the binarized model with all marginal and interaction effects before the stepwise variable selection procedure has been applied (Figure 3.6a) and after (Figure 3.6b). 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. Clearly, with the much smaller model with few coefficients obtained with our algorithm, we are able to select the relevant marginal and interaction effects. Looking at Figure 3.6b, 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%. Getting back to the role of *Status of existing checking account* and *Job* and their interaction,

Measure	Model	Simulated dataset
Accuracy	Original without interactions	78.49
	Original	78.53
	Binarized	78.45
	Lasso	77.83
	Group lasso	78.08
Complexity	Original without interactions	11.59
	Binarized	1.45
	Lasso	17.15
	Group lasso	4.65

Table 3.4: *Simulated* dataset: Accuracy and Relative complexity (the number of non-zero coefficients estimated for the categorical variables relative to the total number of estimated coefficients in the original model with interactions) in the validation set, for the original, binarized, lasso and group lasso models

we can see that B_1 is significant at the 5% level, B_{11} is not significant even at the 10% level, while their interaction $B_{1:11}$ is significant at the 1% level.

3.3.3 Simulated dataset

In this section, we discuss the results for the simulated dataset.

Table 3.4 reports the accuracy and the relative complexity for the simulated dataset. The binarized model with interactions has similar accuracy to that of the original model with interactions, with a dramatic reduction in the number of coefficients. This means that our approach allows us to model the interactions, using a much smaller model. Lasso and group Lasso also reduce the relative complexity.

We end this section by illustrating how our binarization approach is able to recover the underlying generating model. On the right panel of Figure 3.7, 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 Figure 3.8 for the binarized model. We see that we recover the generating

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

Figure 3.5: *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

	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$	$B7$	$B8$	$B9$	$B10$	$B11$	$B12$	$B13$
$B1$	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

Figure 3.6: *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

model in both cases, while in the binarized model the coefficients are estimated with a larger sample, resulting in smaller standard errors, as seen in the 95% confidence intervals around the coefficients.

3.4 Conclusions

In this chapter, 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 dummy variables. Third, by binarizing the categories we have more records to estimate each coefficient, which together with the homogeneity ensures lower standard errors. We used a simulated dataset and four real-world datasets, and in all cases our algorithm considerably reduces the number of coefficients of the model, allowing the user to interpret and select interactions between the new binarized categorical predictors.

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 (Aghaei et al., 2019; Romei and Ruggieri, 2014; Zafar et al., 2017a). 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.

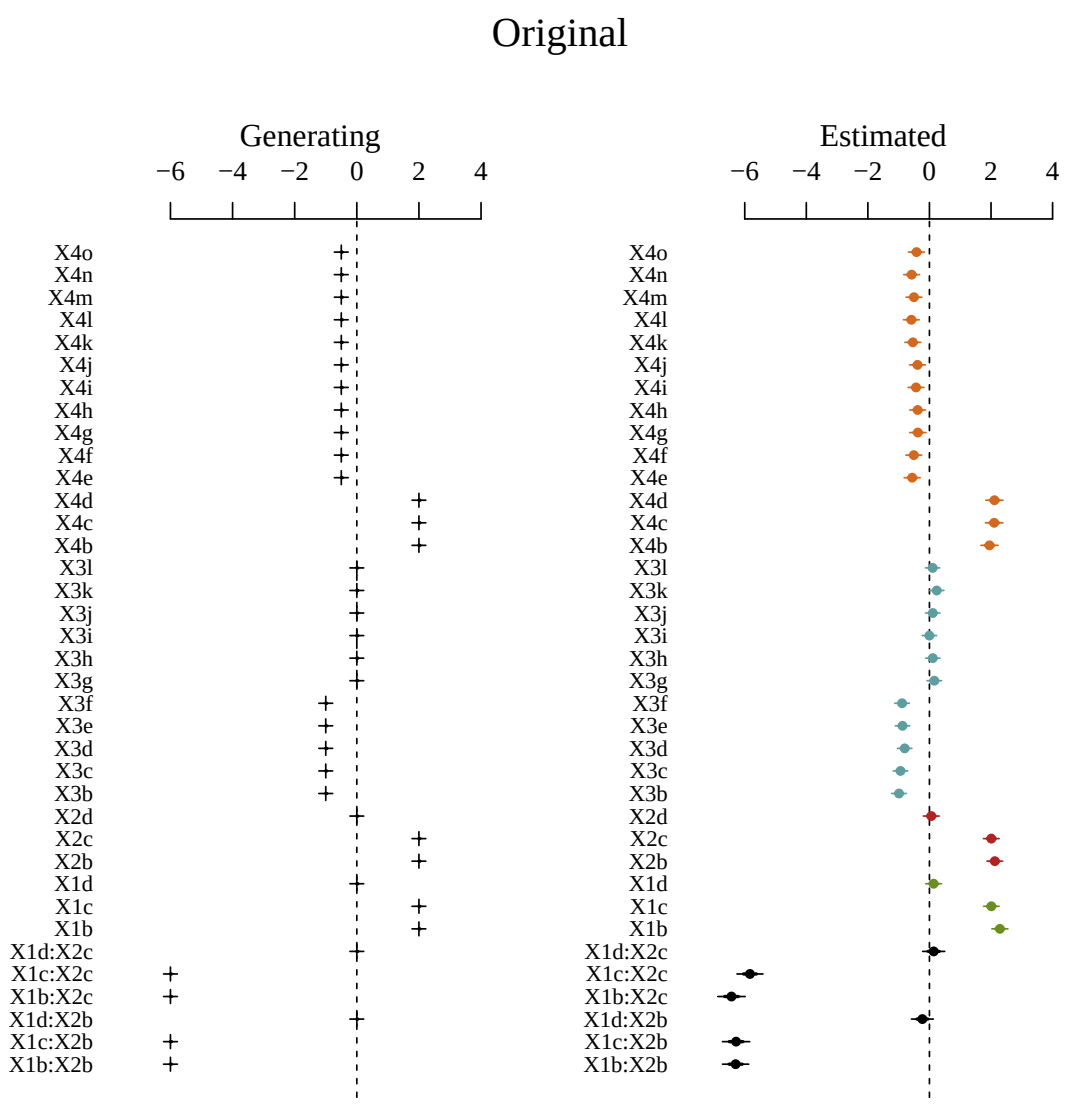


Figure 3.7: *Simulated* dataset: Generating model (left) and coefficients of original model with 95% confidence intervals (right)

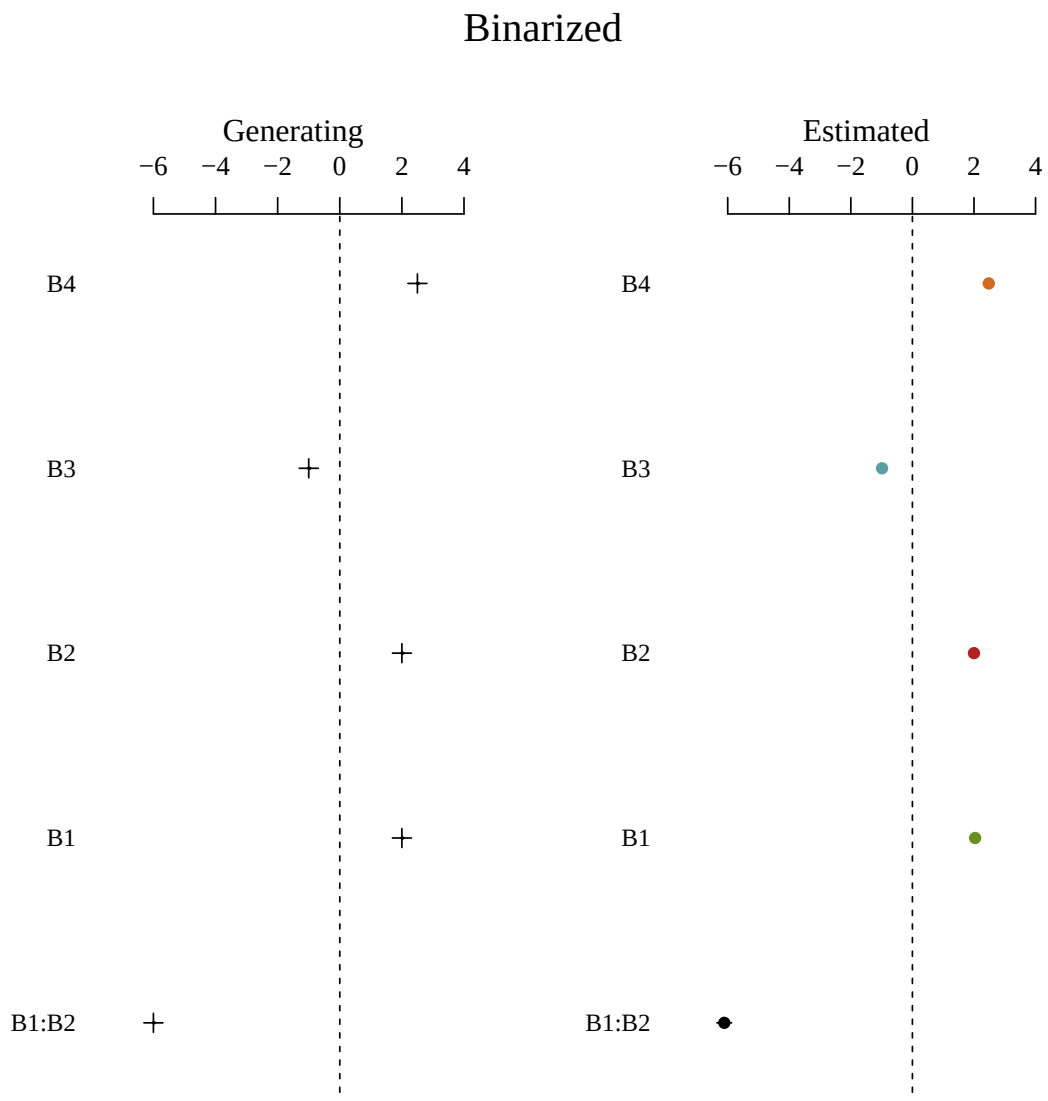


Figure 3.8: *Simulated* dataset: Equivalent binary generating model (left) and coefficients of binarized model with 95% confidence intervals (right)

Chapter 4

Prediction of school dropout in rural Antioquia, Colombia, using Machine Learning: improving targeting and identifying important predictors

4.1 Introduction

Despite the rapid expansion in school enrolment in developing countries in the last decades, completion rates remain low (Adelman et al., 2018). In Latin America, the high rates of enrolment in primary education are followed by a decline throughout secondary education, due to students dropping out before graduation (Bassi et al., 2015). In Colombia, the focus in this chapter, within a year, around 3% of the students leave the school system (considering all school grades from 1 to 11). This number compounds along the whole education cycle: it translates into a survival rate of about 44%, meaning that out of 10 students starting Grade 1, 4 will reach Grade 11 on time (being the right age for the grade).

Students who leave the school system early can expect to have fewer economic opportunities (Rumberger and Lamb, 2003; Psacharopoulos and Patrinos, 2018), lower life expectancy, lifetime wealth, health quality and even happiness (Oreopoulos, 2007; Oreopoulos and Salvanes, 2011; Bentaouet Kattan and Székely, 2015). Consequently, dropout prevention is high on the agenda of governments and society in general (Simon et al., 2007; García et al., 2018).

In education policy making, several states in the United States and countries in Europe have designed early warning systems for students at-risk of dropout using rich administrative data (O'Cummings and Therriault, 2015; Knowles, 2015; Sara et al., 2015) where the goal is to identify most students at risk while keeping a low false-positive rate (where the positive class corresponds to dropout). In Latin America, the recent improvements in setting up information management systems on student enrolment allow tracking individual students over time with unique identifiers which makes it easier to implement these types of approaches (Adelman et al., 2018).

This chapter uses administrative data from rural areas in Antioquia, Colombia, to train prediction models of school dropouts using state-of-the-art machine learning methods. To the best of my knowledge, this is the first study to use enrolment data from Colombia to build and compare different machine learning methods in predicting dropout and to propose the use of these predictions as a targeting strategy to select students who need preventive interventions.

In order for an intervention to be effective and efficient, accurate targeting is essential (McBride and Nichols, 2018). This chapter studies the usefulness of building machine learning-based predictions for targeting school dropout prevention interventions. In rural Colombia, household surveys indicate that lack of transportation is one of the main causes of school dropout (Departamento Administrativo Nacional de Estadística, 2021). As well as providing transportation to students in remote areas, the government is designing strategies to reduce the barriers to accessing schools for students who live relatively close, but far enough to consider lack of transportation to be a problem. Through the program "En mi Bici a la Escuela", students ages 10 to 17 will receive bikes in an effort to reduce dropout rates (Secretaria de Educación de Antioquia, 2022) and promote the bike as a sustainable means of transport. Typically, these types of interventions are targeted at schools with a high dropout rate. Since machine learning methods are better at detecting false positives than targeting students in schools with high dropout rate, I will demonstrate how data-driven targeting based on dropout probability can save money in the long run. There has been previous work for Guatemala and Honduras showing similar results regarding potential savings for a hypothetical program (Gaebel et al., 2012; Herrera Prada, 2021), and I take this a step further by presenting a realistic scenario that looks at a specific program designed in Antioquia in 2022 (Secretaria de Educación de Antioquia, 2022).

In addition, this study contributes to the identification of important factors that influence school dropout in Antioquia, Colombia, indicating areas that the public administration can intervene. The causal effects of specific variables on dropout probability have been extensively studied in economics in order to devise interventions to reduce it (Behrman et al., 2014). The human capital theory is used to examine the determinants of dropout, in which the decision to remain in or dropout of the school system is made by weighing the marginal costs of continuing to invest in education against the marginal expected benefits of getting more schooling (Becker, 2009). This decision can be explained by a range of individual, family, school, and community-related factors like family's income and wealth, school quality, and labor market conditions which shape its costs and benefits (Adelman et al., 2018). A reliable way to estimate the causal effect of a specific factor on dropout is through a Randomized Control Trial (RCT). An RCT must, however, be designed after determining which variables should be considered. This chapter explores predictors that can guide the design of RCTs. Even for cases where we know the determinants, as in the case of lack of transportation in Antioquia, it may be difficult to observe such as with the bike program previously described. In this study, we learn more about all these issues.

The data used in this chapter comes from the Ministry of Education of Colombia which tracks school enrolment through the Sistema de Matricula -SIMAT- platform (Ministerio de Educación Nacional, 2021). It contains information on monthly enrolment in primary to upper-secondary education between 2016 and 2019 in schools in the Antioquia region, the second-largest in Colombia in terms of population. Each month, there are about 450,000 students enrolled in the 740 schools in the region. The schools are located in 9 subregions, including schools in some municipalities within the metropolitan area of Medellin, the capital. However, the data does not include the city of Medellin and the biggest municipalities

around it (Itagui, Envigado, Bello, Rionegro), since they all have their own systems, and the Secretary of Education of Antioquia only handles the smaller and more rural municipalities which are also the ones that have the highest dropout rates (Ministerio de Educación Nacional, 2019).

Regarding the processing of the data before running the prediction models, the class imbalance is an important consideration. Out of the total number of students who start a school year, 3.7% leave the school system at the end of the year. The relative rarity of the dropout phenomenon within a year makes learning from it challenging. Methods that try to predict with very imbalanced datasets might have a very low performance for the minority class which is precisely the one we are interested in (dropouts). To counter this problem, alternative measures of performance, changing the prediction thresholds and sampling techniques (under or over-sampling of the majority class) have been proposed in the literature (Burez and Van den Poel, 2009; Goorbergh et al., 2022).

The chapter compares different prediction models in terms of their performance out-of-sample. The methods implemented are Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Machines (GBM). The predictive performance of the models is compared among themselves and to a benchmark study of other Latin American countries using the same type of data for the context of secondary education (Adelman et al., 2018).

The chapter is organized as follows. Section 4.2 describes the data sources, the predictors, and the process of including other external predictors. Section 4.3 presents the methodology, including the learning methods used for the predictions (Section 4.3.1), the handling of imbalanced data (Section 4.3.2), the methods for measuring variable importance (Section 4.3.3), as well as the pre-processing of the data (Section 4.3.4). Section 4.4 includes the descriptive statistics in Section 4.4.1, the results in terms of performance of

the different methods (Section 4.4.2), the graphical analysis of variable importance and the partial dependence plots (Section 4.4.3) and the comparison of different targeting strategies (Section 4.4.4). Section 4.5 concludes with policy recommendations and ways forward in the design of predictive models of school dropout and the targeting of students using these tools.

4.2 Description of the dataset and the context

The available data for this chapter corresponds to rural areas in Antioquia, Colombia. This is justified since previous studies have shown that the dropout rate in Antioquia and Colombia is higher in rural areas than in urban areas (Goel et al., 2016). Therefore, school dropout prediction models are developed using student-level administrative data from rural areas in Antioquia, Colombia.

The education system in Colombia is made up of 5 levels: early childhood education, preschool education, basic education (five grades of primary and four grades of secondary), high school education (two grades and ends with a bachelor's degree), and higher education (Ministerio de Educación Nacional, 2022a). The basic, mandatory cycle goes from grades 1 to 9 but I focus on students in basic education and high school (Grades 1 to 11) since finishing grade 11 is an implicit goal and what the job market requires even for non-skilled labor. The models are trained for each grade separately to be able to compare how the models perform in each of them and whether there are differences. The school system is organized into schools that can have different sections or satellite schools (smaller schools spread around the territory), many satellite schools are also characterized by offering only some grades (usually from 1 to 5 or 1 to 9), so this distinction is important since often students have to change schools in the transition from 5 to 6 and from 9 to 10. Each school

belongs to a Secretary ("Secretaría de Educación" in Spanish). The Secretary is responsible for the education system, guaranteeing the education service, providing schools with teachers and resources, and ultimately implementing actions to guarantee quality and successful completion of the education cycle. The system allows students to delay (retention is allowed), this leads to many students having to repeat one or several grades, which as we will see, is an important predictor of dropout.

Table 4.1 shows the predictors in the dataset. They can be categorized into Academic, Demographic, Family-peers, School, socioeconomic, and Trajectory. The main data source is the SIMAT platform which contains mainly socio-economic and demographic predictors for the student and their households as well as school-level data. SIMAT is a system to follow each school's enrolment, and as such it provides a picture of enrolment at one point in time. However, one advantage of SIMAT is that it has a unique identifier for each student which allows linking each monthly dataset to follow the student's trajectory in the school system which can be useful in identifying students at risk of dropping out. In addition to the trajectory, I add predictors that capture the school and municipality environment in which the student is involved, and some related to the family. In the following lines, I describe the predictors and the data sources.

The School directory information -DUE in Spanish- (Ministerio de Educación Nacional, 2022b) is the school registry that identifies the school, address, and all contact information related to the school. Since many of the satellite schools are located in rural areas (away from the municipality's downtown area) I use the DUE to calculate the distance of the student's school to the municipality's downtown area. This is a numerical indicator of how rural the context is. Municipality-level indicators of socioeconomic level are added in variables GDPpercap and Dependencyrt and they come from the Anuario Estadistico of

Antioquia (Departamento de Antioquia, 2018a). The individual socio-economic level is represented by predictor SocEcoIndex, and it was obtained from the SISBEN database for each student (Castañeda and Fernández, 2005). Selecting Beneficiaries of Social Spending (SISBEN in Spanish) is a method to target social spending in Colombia, it is a proxy means-tested system that is based on an assessment of the living conditions of individual families. The estimated index is composed of four factors, as follows: 1) housing quality and possession of durables; 2) public utility services, 3) human capital (education) levels, and 4) family demographics, unemployment, dependency ratio, and income per capita. It is ultimately an index of the socio-economic conditions of the family the student belongs to. Hence, this database allows to link siblings in the school system and to add predictors like the Family size (using the number of siblings in the system as a proxy) and Family position (within the siblings in the school system).

Trajectory predictors are related to the student's trajectory in the school system and they are built from the historic data of SIMAT, they include information on whether the student has not been enrolled at some point in the past two years (Months.t.1, Months.t.2), or they have transferred from a different school in the same region in the previous year (sch_cg), or if they dropped out the previous year (Change(t-1)). Finally, there are few academic performance predictors, among them, whether the student has exceptional abilities -Exceptional- (diagnosed with a test and available in the SIMAT data), their academic status the previous year (whether they failed or passed) (Acad_prev_yr) and whether they are repeating the current grade if they failed.

Let me comment on the data limitations. Dropouts are students who leave (cannot be found in Antioquia's enrolment system) during the school year (March-December) and do not re-enroll during that period. Two factors could bias this dropout rate. First, I do not have

information on the enrolment systems of other Secretaries. If the student moves to another administrative area, they will be counted as a dropout in my data. This would mean that the dropout rate is overestimated. At the same time, there is a risk that students leave and are not retired from the system until later. This means that there is also a risk of underestimating the dropout rate.

I do two things to make sure that the estimates are as precise as possible given the limitations. First, I gained access to data for the city of Medellin (the capital of Antioquia). This should cover the students who leave for the city. Thus, to build my outcome variable I search for students also in Medellin. Current estimates of inter-region migration are hard to obtain but several authors have quantified inter-region migration in the past and Census (Silva Arias and González Román, 2009; Blanco, 2014). They found that Antioquia has a vigorous within-region movement, mainly from the rural area in Antioquia to the metropolitan area of Medellin. At the same time, emigration to other regions is the lowest in the whole country and it is a net receiver of migrants (Departamento de Antioquia, 2018b). This suggests that the bias in my definition of dropout should not be very large. Second, I check that the dropout rates calculated using the current dataset, roughly coincide with the ones calculated using the nationwide dataset, this can be seen in the Appendix. In Table 4.7, the dropout rate is similar for primary and high school and it is slightly underestimated for secondary school.

In the future it will be necessary to access data for the whole country to avoid this, however, this is not an issue of Antioquia alone but all administrative areas in the country since the Ministry only hands in the enrolment data corresponding to schools in each administrative area so the regions do not have the whole picture. One last limitation of the data is that schools might have an incentive to keep a student enrolled on paper even when the student

already left (the school is paid a fee per student by the government). The Ministry is aware of this and there is a yearly audit of the enrolment records to identify "ghost" students. To the best of my knowledge, the system works relatively well and the audits ensure the good quality of the data.

4.3 Methodology

I now introduce the prediction models for the detection of school dropout and the post-prediction interpretation of the results. Section 4.3.1 describes the three models used in the numerical section to produce a prediction of the probability of dropout. Section 4.3.2 explains how class imbalance is handled. Section 4.3.3 explains the use of variable importance and partial dependence plots to interpret the relationship between the predictions and a subset of the predictors in the dataset. Finally, Section 4.3.4 explains the pre-processing of the dataset to make it ready for prediction.

4.3.1 Prediction models

Let me start by introducing the predictive learning problem. There is an outcome variable $y \in \{0, 1\}$, where class 1 indicates a dropout and 0 a non dropout, and a set of input predictors $\mathbf{x} = (x_1, \ldots, x_P)$. Using a randomly selected training sample size N $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$, we want to get a predictive model $\hat{y} = G(\mathbf{x}) \in \{0, 1\}$. The first method to consider is the logistic regression. The outcome y is related to $\mathbf{x}$ through a logit link function, namely,

$$log\frac{Pr(y = 1|X = \mathbf{x})}{Pr(y = 0|X = \mathbf{x})} = \beta_0 + \beta^T\mathbf{x}, \tag{4.1}$$

Type of predictor	Name	Description
Academic	Exceptional	1 Exceptional abilities, 0 otherwise
	Late.t.4.	1 if student was delayed in 2015, 0 otherwise
	Repeating	1 Student is repeating the grade, 0 otherwise
Demographic	Delay	Difference between Age and Normal age for the grade
	Displaced	1 Student was displaced by violence, 0 otherwise
	Dissability	1 Student has a disability, 0 otherwise
	EthnicBackg	9 ethnic backgrounds
	Female	1 Female, 0 Male
	Indig_reserv	1 if belongs to indigenous group, 0 otherwise
	Overage	1 if student is three or more years delayed, 0 otherwise
	Region	10 Antioquia regions
	RelAge	Student's age over average age in student's class
	Victim	1 Victim, 0 Not a Victim [1]
Family-peers	Attendance	% of the school age population not attending school in municipality
	FamPos	1 to 18 birth order calculated with siblings in the system
	FamSize	1 to 18 number of sibling in school system
	Illiteracy	Illiteracy rate in the student's municipality
School	BoardingSch	1 Boarding school, 0 Otherwise
	Calendar	1 if school is A calendar 0 if it is B calendar[2]
	ClassSize	Number of students in the student's classroom
	Distance	Distance from school to municipality center in km
	DropoutSchool	School's dropout rate in previous year
	Grade	11 grades from 1 to 11
	Language5	Average standardized test level of language in grade 5
	Language9	Average standardized test level of language in grade 9
	Methodology	44 educational methodologies
	Schedule	0 Morning, 1 Afternoon
	SchMain	1 Main School, 0 Satellite school
	SchoolSocioec	Average socio-economic index at the school
	SchRural	1 school rural, 0 otherwise
	Sector	1 Public, 0 Private school
Socio-economic	BenHeadHous	1 beneficiary to the mother head of household program, 0 otherwise
	BenHerNac	1 beneficiary to the heroes program, 0 otherwise
	BenMilitForc	1 beneficiary of veteran program, 0 otherwise
	Dependencyrt	Dependency rate in municipality
	GDPpercap	GDP per capita in municipality
	Head_household	1 student is head of household, 0 otherwise
	Lives_rural	1 Student lives in rural area, 0 otherwise
	SocEcoIndex	Socioeconomic vulnerability index
	Stratum	Household Socio-economic level (1 to 6)
Trajectory	Acad_prev_yr	The student passed, failed, dropout, did not study in t-1
	Change(t-1)	Dropout, changed schools, other in t-1
	Medellin.t.2.	The student was enrolled in Medellin in t-2
	Medellin.t.1.	The student was enrolled in Medellin in t-1
	Months.t.1.	Number of months the student was enrolled in year t-1
	Months.t.2.	Number of months the student was enrolled in year t-2
	OtherMun	1 Student comes from different municipality, 0 otherwise
	sch_cg	1 Student changed schools in t-1, 0 otherwise

Source: SIMAT Antioquia, 2019, SISBEN, 2019, Directorio de establecimiento y sedes Antioquia, 2019, Anuario Estadístico de Antioquia, 2019. [2]

[1] The law of victims of the Colombian armed conflict (Law 1448 of 2011) established that a victim is: "those persons who, individually or collectively, have suffered damage due to events that occurred as of January 1, 1985, as a consequence of infractions of International Humanitarian Law or of serious and manifest violations of international Human Rights norms that occurred during the internal armed conflict"

[2] Schools in Colombia can be Calendar A: the school year goes from January to November, or Calendar B: the school year goes from August to June the following year

Table 4.1: Description of predictors of school dropout

where β_0 is the intercept and β is the vector of model parameters for the predictors **x**, which determine the effect of their effect on the response. For new observations, we use Equation (4.1) with the coefficients estimated using the training sample, to predict the probability that the student will dropout.

The second prediction method is Random Forests. Random Forests is a bagging method that builds a large collection of de-correlated classification trees, and then uses majority vote to decide the predicted class (Hastie et al., 2015). Classification trees partition the predictor space into a series of regions and then assign a class to each region. RF builds several of these trees and each tree uses a different training sample, selected by bootstrapping. A subset of the available predictors is randomly selected at each node in the tree and the best split available within those predictors is selected for that node. This method uses bagging to create the training set of observations for each individual tree, and the number of predictors selected randomly at each node is a parameter to be tuned (Breiman, 2001). After a large number of trees is generated, they vote for the most popular class for each observation, and provide this as the predicted class.

The third prediction method is Gradient Boosting Machines (Friedman, 2001). Like RF, GBM is a collection of trees, but while RF builds each tree independently, boosting tries to improve the prediction for misclassified cases in the previous trees by generating new trees and reweighing all of them. The prediction is made by weighing the sum of the predictions made by all the tree models. For our two-class problem, in each iteration, GBM performs a steepest descent minimization for logistic likelihood.

4.3.2 Handling class imbalance in the response variable

Recall that the outcome variable consists of class 1, dropout, and class 0, non-dropout. In a particular year, 3.7% of the students leave the school system. Therefore, we observe an imbalance between these two classes, i.e., the number of observations associated to class 1 is much smaller than the number of observations associated to class 0. This is known in the literature as a class imbalance. In this section, I discuss approaches to handle a class imbalance in the Colombian school dropout data.

According to Weiss and Provost (2003) there are several categories of problems that arise when learning from an imbalanced dataset. The first issue is that *accuracy* is an improper evaluating metric of the performance of the classifier. This is because, even if the classifier does a very poor job in the minority class, it can still have a very high accuracy given the low number of dropout instances in the data relative to the non-dropout. Hence, using alternative measures of performance is a better way of assessing classifiers when there is class imbalance.

We can then focus on three measures of the classifier's performance: Sensitivity, Specificity, and AUC. Sensitivity is the true positive rate (the rate of correct classification for the dropouts), and Specificity is the true negative rate (the rate of correct classification for the non-dropout). AUC provides an aggregate measure of performance across all possible classification thresholds. The AUC can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. A random classifier (tossing up a coin) will give an $AUC = 0.5$, while a perfect classifier will give $AUC = 1$, ideally, we would want to be closer to 1. In this chapter, AUC is used in the cross-validation stage and then the results are presented in terms of Sensitivity, the false positive rate, and

AUC.

Using these alternative measures will give a better idea about the balance in the rates of classification for the positive and the negative class, but it does not ensure that the classifier will do good for the positive class when the number of positive cases is low compared to negative cases. For that, we need to use different techniques (Burez and Van den Poel, 2009). Sampling is one of the techniques to deal with rarity. The idea is to alter the distribution of classes in the training sample to reduce rarity. The basic sampling methods are under-sampling and over-sampling. Under-sampling eliminates majority-class observations from the training sample, while over-sampling duplicates minority-class observations. In both cases, we make the rare class less rare in the training phase.

In this dataset a large number of observations in the majority class in the training sample (290,740) allows to perform under-sampling comfortably -see Table 4.2-. Now let me explain the approach to finding an appropriate balance between the classes. For each classifier, I will perform a grid search of different ratios between the classes. The ratio that gives high average value between Sensitivity and Specificity for the three methods will be chosen and used for the final training of each classifier.

4.3.3 Variable importance and partial dependence plots

This section describes the methodology to interpret the output from the prediction models of school dropout. In Random Forests, variable importance is measured with the method proposed by Breiman (2001). Since the RF method uses bootstrap to select the training samples used in each tree, there will always be some observations left out of each bootstrap sample that are called the out-of-bag observations. As stated in the methodology, we have P predictors in the dataset. After each tree is constructed, the values of the pth variable in the

out-of-bag observations are randomly permuted and the out-of-bag misclassification rate is calculated. This is repeated for $p = 1, 2, \ldots, P$. Variable importance is the percent increase in misclassification rate as compared to the out-of-bag rate (with all variables intact).

Once we have a measure of variable importance we can try to understand the functional relationship between a reduced set of important predictors and the probability of dropout. The Partial Dependence Plot -PDP- developed by Friedman (2001) is a tool to analyze that relationship. The PDP shows the marginal effect one or two features have on the predicted outcome of a machine learning model. Let $S \subset \{1, \ldots, P\}$ and let C be the complement set of S. Here S and C index subsets of predictors, hence, if $S = \{1, 2\}$, then $\mathbf{x}_S$ is the 2 dimensional vector containing the values of the first two coordinates of $\mathbf{x}$. Given the output from a machine learning model $G(\mathbf{x})$ (the estimated probability of the positive class), the partial dependence of G over a subset of $\mathbf{x}_S$, denoted by $\hat{G}_S(\mathbf{x}_S)$, is defined as the expectation of G over the marginal distribution of all variables in $\mathbf{x}_C$ (Zhao and Hastie, 2019). In practice, the partial function $\hat{G}_S(\mathbf{x}_S)$ is estimated by averaging over the training data:

$$\hat{G}_S(\mathbf{x}_S) = \frac{1}{N} \sum_{i=1}^{N} G(\mathbf{x}_S, \mathbf{x}_{C_i}) \tag{4.2}$$

where $\{\mathbf{x}_{C_1}, \ldots, \mathbf{x}_{C_N}\}$ represent the different values of $\mathbf{x}_C$ that are observed in the training data. Several authors have shown how the use of these tools to better understand the predictions of black-box models like RF and GBM (Goldstein et al., 2015; Zhao and Hastie, 2019; Buckmann et al., 2021).

4.3.4 Pre-processing of the data

This section explains how the data is processed before using it for prediction. As explained in the introduction, the data corresponds to a cross-section of students enrolled in schools in Antioquia, Colombia, in 2019. The response variable is the status of the students by the beginning of the following year (dropout or non-dropout). For the numerical predictors with missing values, new dummies indicating whether there is a missing value are added and then the missing values are imputed using the mean for each grade. For the categorical predictors, observations with missing values were assigned category unknown ("NA").

In what follows I discuss the approach followed to model the categorical predictors in the dataset. As mentioned in previous chapters, for linear models like logistic regression, the categorical predictors need to be transformed into real-valued vectors and this is usually done by representing each category by one dummy variable, leaving one out for contrasts. Having many high-cardinality categorical predictors can be wasteful, hard to interpret, and prone to overfitting (Wager and Athey, 2018; Carrizosa et al., 2021b). In this chapter, I use the method proposed by Carrizosa et al. (2021b) to cluster the categories of each categorical predictor into just one dummy, see Chapter 2. After finding a single dummy for each categorical predictor, I train all my models (LR, RF, GBM) with this reduced representation (including all other predictors as they are), reducing the number of coefficients to estimate for the logistic regression model and the number of splits that tree-based methods have to consider. The clustering is done for all categorical predictors with more than 5 categories, except for the predictors *Region* and *Grade* for which it is interesting to analyze each region separately, the grades will be used to split the data to make predictions for each grade separately.

4.4 Results

This section shows the results. First, I introduce the descriptive statistics for the data in Section 4.4.1. Then, the results in terms of performance of the different models in Section 4.4.2 and the the graphical analysis of variable importance and the partial dependence plots in Section 4.4.3. Finally, the analysis of different targeting strategies in Section 4.4.4.

4.4.1 Descriptive statistics

This section shows the descriptive statistics for the data used for the prediction of school dropout in the region of Antioquia in Colombia. The data corresponds to 431,371 students enrolled in the year 2019 in Grade 1 to 11 in schools in the region of Antioquia. The yearly dropout rate for the whole sample according to the previous definition was 3.7% or 16,027 students in 2019 (Table 4.2). The highest dropout rates are seen in the first grades of a cycle (Grade 1 in primary, Grade 6 in Secondary, and Grade 10 in Upper secondary). Students from Grade 11 dropped out at the lowest rate of 2.1%. This is consistent with economic theory suggesting that students in the final grades are less likely to dropout due to a higher opportunity cost of dropping out (they are a few months away from graduating with a degree) (Fiske, 1998; Pinzón Hernández, 2018). However, it could also be attributed to selection bias since the students with the highest probability of dropping out already did in previous years. I will build the predictions separately for each grade to be able to discern patterns for each grade separately.

Table 4.3 describes the categorical predictors including the binary ones and the ones that have been transformed into binary using the method in Carrizosa et al. (2021b). Table 4.4 describes the mean, standard deviation, and quartiles for the numerical predictors in

Cycle	Grade	Dropout status (y) 0	1	% Dropout
Primary	1	39,938	1,522	3.7%
	2	39,326	1,295	3.2%
	3	40,391	1,282	3.1%
	4	41,910	1,275	3.0%
	5	43,086	1,320	3.0%
Secondary	6	48,113	2,783	5.5%
	7	43,586	1,973	4.3%
	8	37,387	1,663	4.3%
	9	31,745	1,290	3.9%
Upper secondary	10	26,919	1,136	4.0%
	11	22,943	488	2.1%
All		415,344	16,027	3.7%

Source: Author calculations based on SIMAT Antioquia 2019

Table 4.2: School dropout rate by grade in 2019, Antioquia, Colombia

the dataset. Half of the students are female, the average socio-economic index is 34 out of 100, which means they come from lower socio-economic background families, most of the students (95%) were enrolled in public schools, and they are delayed on average by 1.2 years, about 10% were 3 or more years delayed (Overage), half of them live in a rural area, the rest of the predictors can be seen in Tables 4.3-4.4.

4.4.2 Predictive performance

This section illustrates the predictive performance in terms of Sensitivity (True positive rate), False positive rate, and AUC of three models: Logistic Regression, Random Forests and Gradient Boosting Machines for 11 education grades. Training is computationally costly and I will follow the recommendation by Dietterich (1998) to use just one training split (70% of the data selected randomly) to train the models, a validation split (15%) to select the parameters for the models that require parameter tuning, and several independent

Name	Categories	Top counts
Region	10	5: 76865, 0: 55507, 7: 51413, 9: 51222
Grade	11	6: 50896, 7: 45559, 5: 44406, 4: 43185
Acad_prev_yr	4	1: 393487, 2: 21854, 3: 11649, 0: 4381
Change(t-1)	4	3: 419722, 1: 6173, 0: 3454, 2: 2022
Stratum	2	0: 378107, 1: 53264
EthnicBackg	2	0: 430747, 1: 624
Schedule	2	1: 295899, 0: 135472
Methodology	2	1: 426321, 0: 5050
Late.t.4.	2	0: 292537, 1: 138834
Sector	2	1: 409717, 0: 21654
Calendar	2	1: 429990, 0: 1381
Female	2	0: 218956, 1: 212415
OtherMun	2	0: 421252, 1: 10119
Dissability	2	0: 417189, 1: 14182
Exceptional	2	0: 430901, 1: 470
Indig_reserv	2	0: 425205, 1: 6166
Repeating	2	0: 414266, 1: 17105
Lives_rural	2	0: 219519, 1: 211852
Head_household	2	0: 430149, 1: 1222
BenHeadHous	2	0: 417728, 1: 13643
BenMilitForc	2	0: 431261, 1: 110
BenHerNac	2	0: 431206, 1: 165
SchRural	2	0: 250563, 1: 180808
Displaced	2	0: 352866, 1: 78505
Victim	2	1: 352370, 0: 79001
Overage	2	0: 389507, 1: 41864
Medellin.t.2.	2	0: 424894, 1: 6477
Medellin.t.1.	2	0: 427484, 1: 3887
BoardingSch	2	0: 414761, 1: 16610
sch_cg	2	0: 387559, 1: 43812
SchMain	2	1: 266118, 0: 165253

Table 4.3: Number of categories and top counts for the categorical and binary predictors in the school dropout dataset, 2019

Name	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
SocEcoIndex	34.372	10.404	5.390	27.801	33.508	41.123	66.581
Delay	1.191	1.386	−8.816	0.299	0.775	1.756	52.721
ClassSize	28.184	13.599	1	19	31	38	92
Months.t.2.	8.263	3.568	0	10	10	10	10
Months.t.1.	9.232	2.351	0	10	10	10	10
Dependencyrt	0.569	0.071	0.417	0.525	0.564	0.603	0.793
GDPpercap	16.859	8.443	7.102	11.201	14.501	20.309	67.496
Illiteracy	0.223	0.097	0.066	0.126	0.232	0.304	0.704
Attendance	0.109	0.039	0.028	0.080	0.105	0.136	0.298
Distance	15.383	42.941	0.000	0.000	0.486	5.179	359.290
FamPos	1.388	0.561	1.000	1.000	1.388	1.388	18.000
FamSize	1.769	0.789	1.000	1.000	1.769	2.000	18.000
RelAge	1.000	0.106	0.279	0.938	0.983	1.042	5.041
Language5	0.353	0.159	0.000	0.322	0.353	0.372	1.000
Language9	0.405	0.125	0.000	0.405	0.405	0.405	1.000
SchoolSocioec	34.372	10.404	5.390	27.801	33.508	41.123	66.581
DropoutSchool	0.037	0.023	0.000	0.023	0.033	0.047	0.250

Table 4.4: Descriptive statistics for the numerical predictors in the school dropout dataset, 2019

testing splits (15%) to test its performance. I build one prediction model for each of the 11 grades separately. Parameter tuning for Random Forests and Gradient Boosting Machine is performed with 10-fold cross-validation. All the predictive models are trained using R and the package Caret in a Workstation with an Intel® Core™ i5-4460 processor with 8 Gb of RAM.

Table 4.5 shows the out-of-sample performance for each of the predictive models and each education grade. We can see that the models identify correctly between 60% and 80% of the students who will dropout in the following year, with a false positive rate between 23% and 32%, AUC is shown in Figure 4.1. The models are able to predict better for the middle grades (Secondary education), with AUC above 80% and to a lesser degree for students in the early stages (Grades 1 and 2) and later stages (Grades 10 and 11). We can see that GBM tends to perform best, especially for Secondary education (Grades 6 to 9).

Regarding the poorer performance in some grades, let me attempt to analyze it in line with the literature and economic theory. Students in the later stages (Grades 10 and 11) are almost finished with their education. As seen in the data, the percentage who dropout in this stage is much smaller since the opportunity cost of dropping out is much higher for these students and the fact that the students with the highest probability of dropping out already did in previous years. However, for those who do dropout the reason might be beyond what our data can capture: they might get sick, have a family loss, unemployment, or other external cause that is hard to capture in this dataset. These factors might play a role for students in all grades but the literature has found that since students in high school are older, the unobserved factors pulling them out of school might be stronger for them than for younger students (teenage pregnancy, drug abuse, working instead of being in school) (Allensworth, 2005; De Witte et al., 2013). As for students in primary education

a similar story might be true, since the opportunity cost of dropping out is much lower for them, they might be more vulnerable to dropping out if something external happens in their lives compared to students in Secondary education (and the decision might be made by the parents more often than by the student). All these hypotheses need further exploration in future research.

Grade	Sensitivity			False positive rate		
	LR	RF	GBM	LR	RF	GBM
Grade1	0.65	0.67	0.69	0.28	0.31	0.31
Grade2	0.67	0.71	0.67	0.27	0.29	0.27
Grade3	0.67	0.66	0.67	0.24	0.23	0.23
Grade4	0.72	0.68	0.72	0.32	0.24	0.25
Grade5	0.74	0.71	0.75	0.23	0.24	0.24
Grade6	0.74	0.74	0.76	0.26	0.27	0.27
Grade7	0.70	0.74	0.74	0.26	0.29	0.28
Grade8	0.75	0.72	0.78	0.27	0.26	0.27
Grade9	0.72	0.72	0.77	0.30	0.26	0.31
Grade10	0.66	0.68	0.72	0.29	0.28	0.32
Grade11	0.62	0.70	0.73	0.27	0.27	0.32

Table 4.5: Out-of-sample performance of prediction models of school dropout -LR, RF, GBM- by grade, 2019

I now turn to the question of evaluating these predictions compared to other similar exercises. Comparisons with other contexts are difficult since the performance of such a system is heavily dependent on the dataset, the levels of education considered, the prediction goals, the sources of information about the students, and different evaluation procedures (Sara et al., 2015). Adelman et al. (2018) analyze system-level data from two Latin American countries, Guatemala and Honduras. They have access to other data not available in the present study, however, they do build one model including only individual demographic variables, they find 67.8% Sensitivity and 32.8% false positive rate for Guatemala and it can go up to 80% Sensitivity and 21.2% false positive rate when more data is available

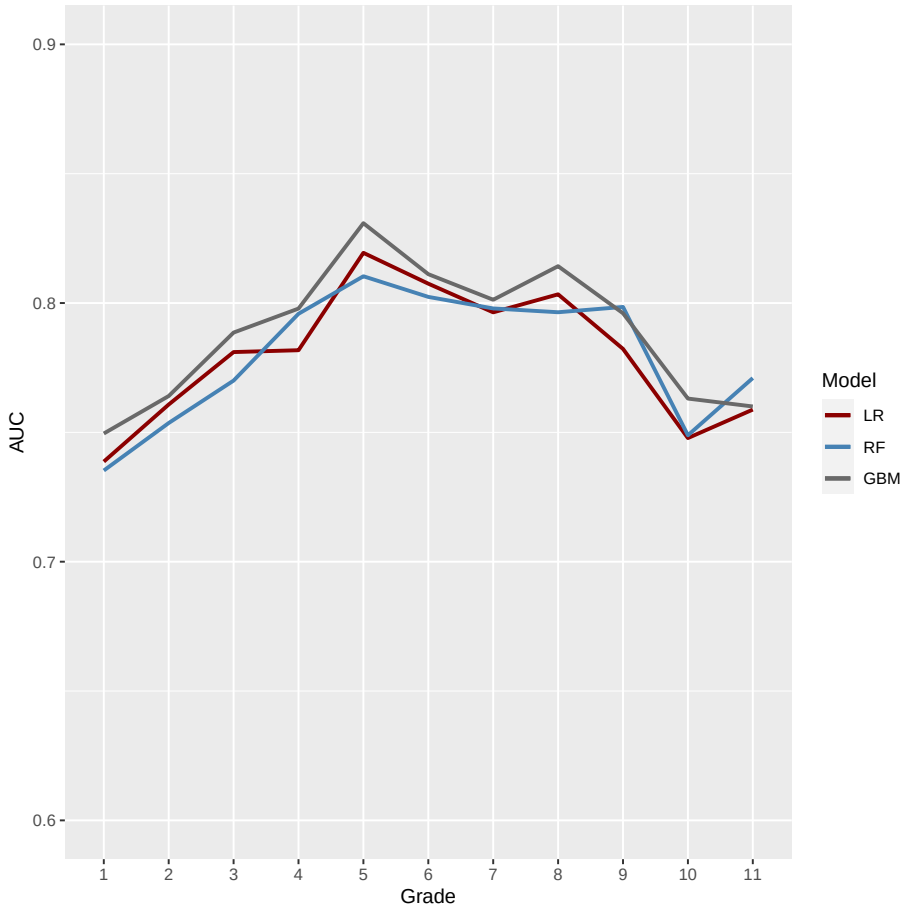


Figure 4.1: Out-of-sample AUC of the prediction models of school dropout -LR, RF, GBM- by grade, 2019

(Parents education, Parents involvement in children´s education, Books at home). Adding household-level predictors such as parents' maximum educational attainment, whether the student has helped with homework, and the number of books, the authors are able to increase performance. Hence, it would be valuable to add this information which is not currently available in the present context. In conclusion, with the data available at hand the performance of the prediction models presented here is comparable to that of Adelman et al. (2018) when they do not include the household-level predictors.

Finally, in this section, we have focused on the out-of-sample prediction of dropout for the year 2019. In the Appendix, I show predictions for the year 2018, with similar results to the ones presented here.

4.4.3 Variable importance and partial dependence

In this section, I analyze the variable importance measure calculated using Random Forest. GBM also provides variable importance measures but the results are very similar and therefore I focus on RF. Figure 4.2 shows a heatmap of the variable importance measures for each grade of education, the most important predictors are shown in dark red and the least important in pale yellow. To simplify the visualization the heatmap is organized according to the most important predictors in Grade 6 and only the first 20 predictors are shown. We see that age-related predictors (Delay -number of years the student is delayed compared to the normal age for the grade and RelAge - age relative to peers) are amongst the most important, especially when the student reaches the 5th grade. Many studies suggest that being past the typical age in a grade significantly increases the hazard of leaving school early (Rumberger, 2004; Plank et al., 2005; Entwisle et al., 2004, 2005), an effect also found in Colombia (Pardo Pinzón and Sorzano Montaña, 2004). In part, this is due to the perception

that being retained entails being unintelligent, failing, and lagging behind (De Witte et al., 2013). According to a meta-analysis of studies of grade repetition's effects, students who repeat a grade have worse academic, social, and behavioral outcomes than those who don't repeat (Jimerson et al., 1997). Also, teenage pregnancy and work might pull someone out of school as they get older (Allensworth, 2005). The Latin American region has especially high rates of teenage pregnancy and, according to Daniels (2015), within Latin America, Colombia shows surprisingly high rates of teenage pregnancy given its development. Furthermore, non-causal studies have shown that teenage pregnancy is associated with school dropout in Colombia (Gómez-Restrepo et al., 2016; Dávila Ramírez et al., 2016). With regards to work, the evidence suggests that working more than a certain number of hours is negatively related to school success in Colombia, Peru, Ecuador and Chile (Post, 2011).

Similarly, the dropout rate in the school where the student is enrolled is an important predictor of dropout, social researchers have been interested for decades in the phenomenon of how an individual's behavior is partly explained by the behavior of others (Morales, 2015) and peer effects are documented on previous studies on dropout causes (De Witte et al., 2013). For Bogotá, Colombia, a study found peer effects related to post-secondary decisions of high school students, however they are very small (Bonilla Mejia, 2016) and hence this connection needs to be further explored.

Furthermore, dropping out in previous years, even if the student returned later, is also associated with a higher probability of dropping out in the current period (Months predictors), especially for the first grades (Grades 2-5). Changing schools is important in primary but not so much in secondary (sch_cg). Previous studies in the united States have found that student mobility is associated with higher dropout rates and lower school performance (Rumberger and Larson, 1998). Since many rural schools only have some levels, it is likely

that students who have to change schools will leave the system altogether because it is difficult to access a school close by. In Colombia, school mobility is high, especially in rural areas.

Delay starts to be an important predictor in the higher grades of primary, which makes sense since the students start to accumulate learning deficits and delays. Socioeconomic status is important, especially in the later grades, which is consistent with the theory that students in the later grades are more vulnerable to external circumstances like a family unemployment event, which is hard to capture in the current data (De Witte et al., 2013; Pinzón Hernández, 2018).

Finally, the following lines discuss the results of the Partial Dependence Plots for some of the important predictors mentioned previously. Figure 4.3 shows the PDP for the number of years of Delay (left) and the Previous enrolment history (Months.t.1) (right), $yhat$ represents the partial dependence of G over a single predictor (Equation (4.2)). Being delayed by more than two years increases the probability of dropout from less than 40% to more than 50%, when all the other predictors are at their mean and/or their most prevalent value. The enrolment history of the student in the previous year also has an important impact on the probability of dropout, which decreases from more than 50% when the student was enrolled for less than 2 months to around 40% for students who were enrolled most of the year. This points out the fact that dropout seems to be the culmination of a process of dis-engagement of the student from the school system as pointed by De Witte et al. (2013). Students with low socio-economic status are more likely to dropout than students with high socio-economic status. Changing schools in the previous year affects greatly the chances that the student will dropout. Finally, students from municipalities with a high dependency rate and a high illiteracy rate are also more likely to dropout pointing again to some kind

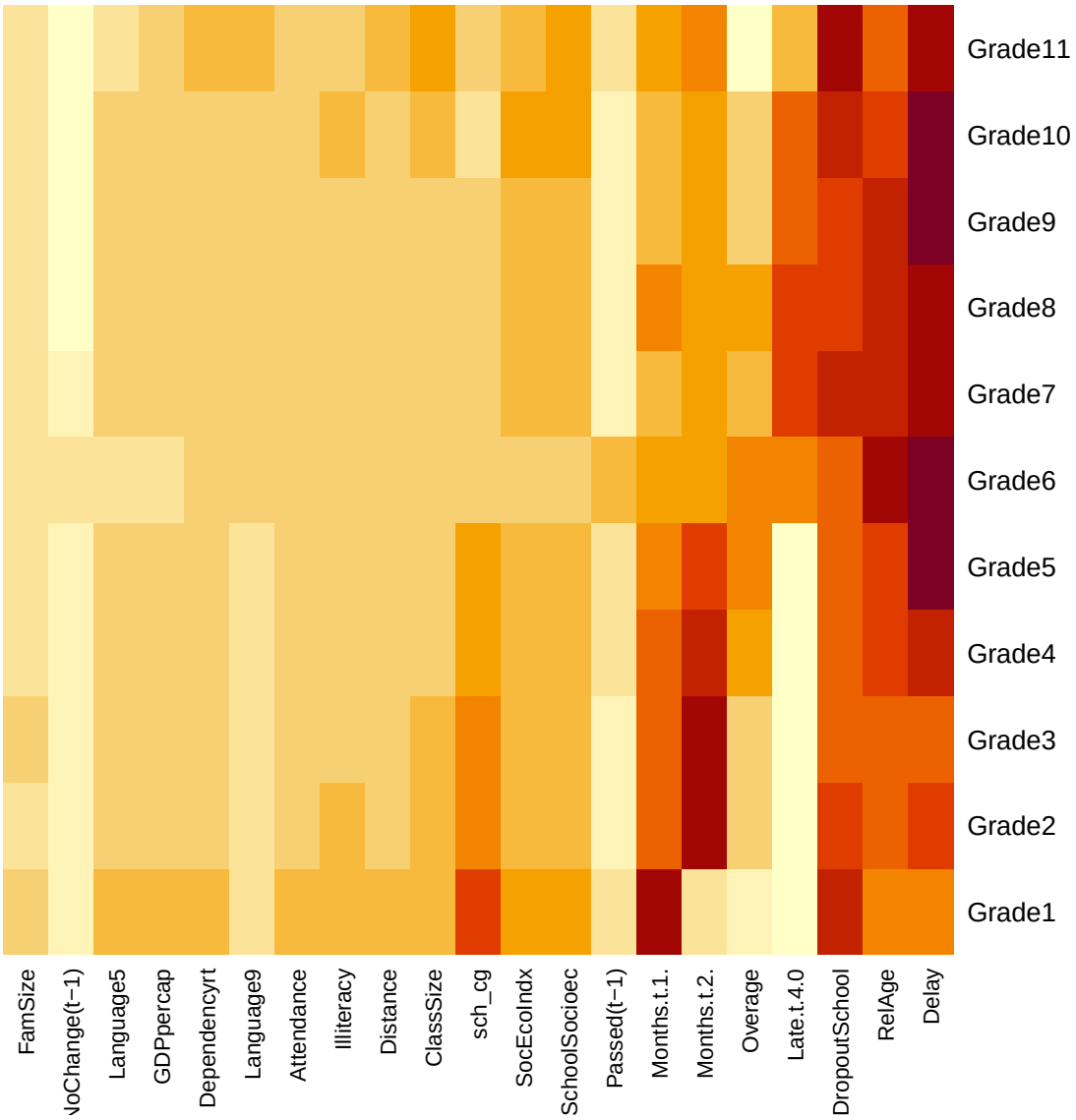


Figure 4.2: Variable importance (top 20 predictors) for school dropout using Random Forest by grade, 2019

of peer effects or at least the possibility that students who are involved in an environment where education is not that important might be more likely to dropout.

4.4.4 Targeting students at risk of dropping out

In this section, I examine how the predictions of these machine learning models can be used to target students at risk of dropping out. Without loss of generality, we focus on the Gradient Boosting Machine predictions.

The GBM is trained in 2018 to predict for students in the year 2019, which would be a realistic way of working with this type of data where we will usually have data on students the previous year and we would like to have predictions at the beginning of the following year. In the Appendix, we show the out-of-time performance of this model for all grades.

Firstly, we assume a hypothetical intervention that is 100% effective and is not limited in budget, and hence it will be aimed at the 50,896 students in Grade 6 in schools in Antioquia in the year 2019. Since Grade 6 has the highest dropout rate (5.5%), public authorities often focus on these students. Secondly, I move on to analyze a specific intervention, namely "En Bici a la Escuela", which has a budget of 1 million USD and will target 3,000 students ages 10 to 17 in schools within 50 municipalities in Antioquia (Secretaria de Educación de Antioquia, 2022).

Targeting of the entire population of students in Grade 6 for a hypothetical intervention with an unlimited budget

Currently, the government targets students in schools with high dropout rates to receive dropout prevention programs. Consider three targeting strategies assuming that the hypothetical intervention is 100% effective in retaining students.

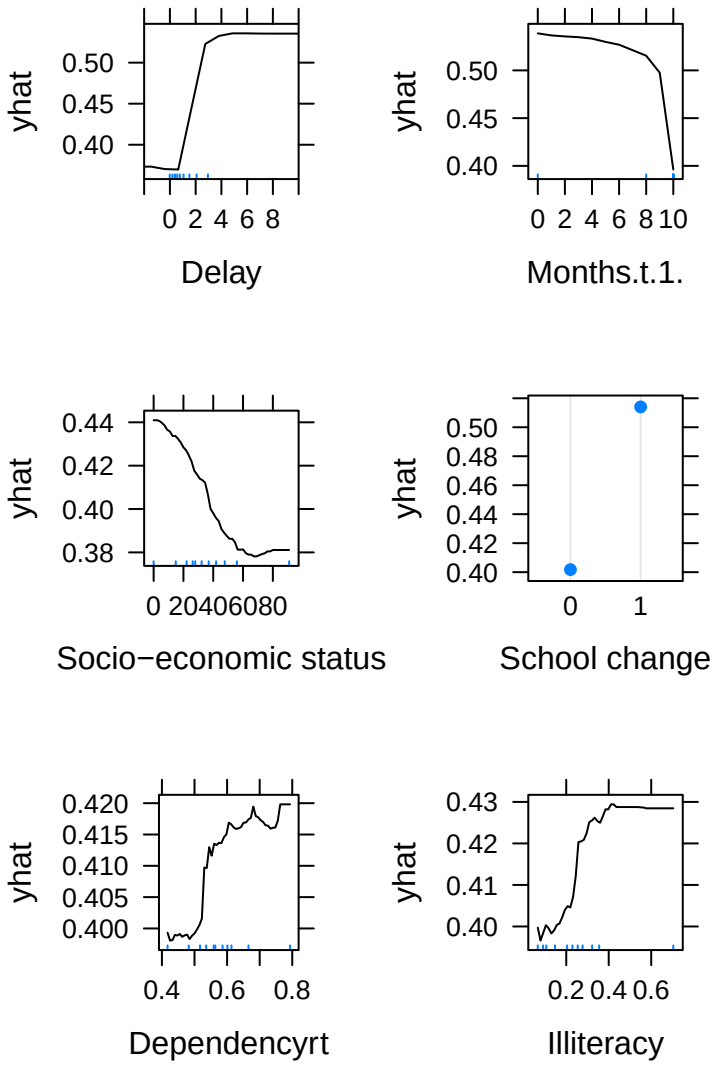


Figure 4.3: Partial Dependency Plots for school dropout using RF - Delay, Previous enrolment history (months), Socio-economic status (SocEcoIndex), School change (sch_cg), Dependency rate (Dependencyrt), Illiteracy rate (Illiteracyrt) - for the full training sample where all grades are together, 2019

- **Targeting strategy 1:** target students in the highest dropout schools
- **Targeting strategy 2:** target students with the highest probability of dropping out according to the GBM
- **Targeting strategy 3:** within the schools in scenario 1, target students who have a high probability of dropping out as estimated with the GBM

By choosing a threshold above which students will be selected, these three targeting strategies can be converted into classifiers. The Receiver Operating Curve (ROC) is a way to visualize the ability of a classifier to discriminate between positive (dropout) and negative (non-dropout) classes based on different thresholds for what we consider to be a dropout. Figure 4.4 shows the ROC curves for each of the scenarios. The 45 degree line represents a trivial discrimination power (tossing a coin). The curves that are placed closer to the 45-degree line represent a worse performing classification method. We see that strategy 2 (using the probabilities calculated by GBM to select the students to intervene) gives the best results, followed closely by strategy 3 (selecting the students with the highest probability within the schools with high dropout rates), whereas strategy 1 (selecting students from high dropout schools) underperforms the other two. We conclude that the machine learning approach can improve targeting since it has a higher ability to discriminate between the positive and negative classes.

Targeting students for "En Bici a la Escuela" with a limited budget

Now, let me examine how the GBM predictions can be used to target students at risk of dropping out to participate in the program "En Bici a la Escuela" (Secretaria de Educación de Antioquia, 2022). The budget is around 1 million USD, and it will target 3,000 students

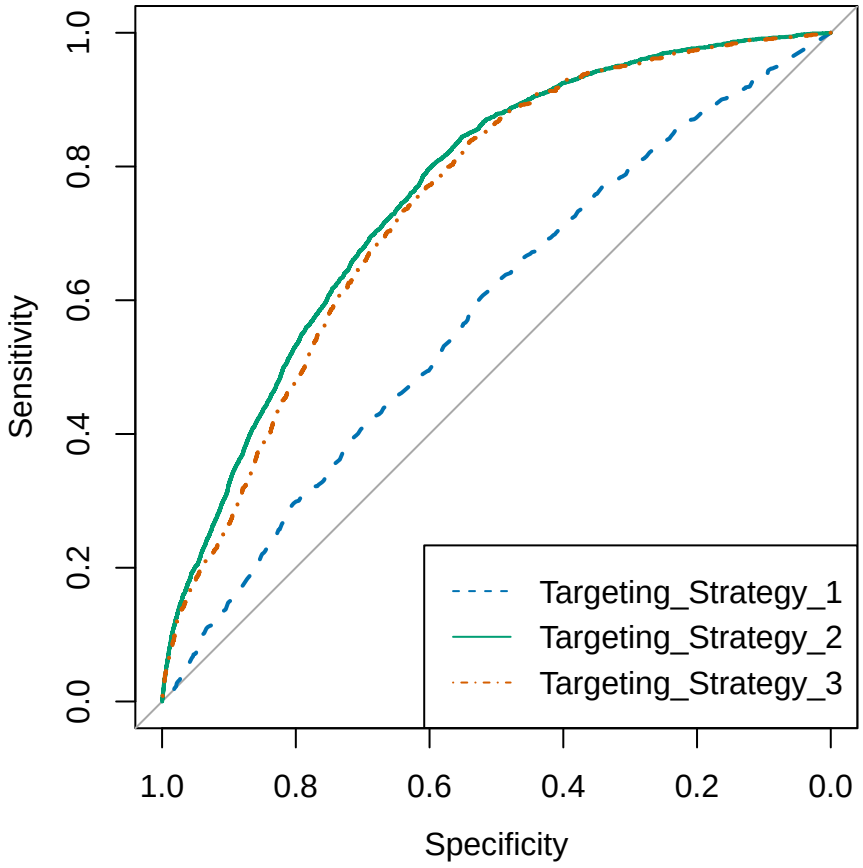


Figure 4.4: ROC curve for three strategies considered to target students of Grade 6 at risk of dropping out by changing the threshold, assuming 100% effectiveness and unlimited budget

ages 10 to 17, in 50 municipalities. The cost per student is 333 USD. It will assign bikes, safety equipment, and training on road safety. At first, the municipalities will submit an application, which will list 4 schools eligible for the program. At the time of writing this thesis, it has not been announced which municipalities and schools will participate in the program, nor how the students will be selected. Therefore, we examine two plausible targeting strategies in the following:

- **Targeting strategy A:** select the 50 municipalities with the highest dropout rates and, within those, select the 4 schools with the highest dropout rate, randomly assign 15 students within those schools to the program
- **Targeting strategy B:** select the 50 municipalities with the highest average dropout probability according to the GBM and, within those, select the 4 schools with the highest average dropout probability, assign the 15 students with the highest dropout probability within those schools to the program

To begin with, let us assume that the program is 100% effective in retaining students. We then compare different levels of effectiveness and the cost per student retained. Based on each targeting strategy, Table 4.6 shows the number of false positives and true positives. With strategy A, 2,873 students are incorrectly targeted as dropouts, while 2,533 are incorrectly targeted with strategy B. Additionally, with strategy A, only 127 true dropouts are targeted, whereas, with strategy B, 467 true dropouts are correctly targeted.

Now let us examine how differences in the number of true positives translate into differences in costs per student. We will assume that the false positives are students who didn't need the intervention since they were staying anyway, while the true positives are students who needed it because they would dropout otherwise. Assuming all the true positives re-

	Method of allocation	
	Targeting Strategy A	Targeting Strategy B
Falsely identified as dropout	2,873	2,533
True dropouts identified	127	467
Cost per student retained	7,874 USD	2,141 USD

Table 4.6: Performance of the two strategies considered to target students of age 10 to 17 at risk of dropping out with the "En Bici a la Escuela", assuming 100% effectiveness and limited budget

main after the intervention, the cost per retained student is calculated as (Total cost/True positives)·100%. Strategy A costs 7,874 USD per retained student, more than four times as much as strategy B at 2,141 USD (Table 4.6).

In order to make a complete comparison, it is important to take into account the benefits associated with the program and analyze different levels of effectiveness. I will assume that the effectiveness of the intervention does not depend on the predictors characterizing the student. The benefit is calculated based on the additional lifetime earnings of secondary school graduates in Colombia compared to primary school graduates. According to the School-to-Work Transition Survey (SWTS) from the International Labor Organization (Fundación Corona et al., 2020), high school graduates can expect to earn about 20% more than primary school graduates. This comes to additional lifetime earnings of around 10,603 USD. The net benefit can be then calculated by comparing this figure to the cost per retained student. Based on varying levels of effectiveness in the x axis (from 20% to 90%), Figure 4.5 shows the net benefit per retained student for each targeting strategy. The net benefit of using targeting strategy B is 4.4 times the one for targeting strategy A when the effectiveness is 90% (8,266 USD compared to 1,863 USD). When the effectiveness decreases, the gap between the net benefits for B and A increases, to the point where low effectiveness (30%) yields a negative net benefit for A while remaining positive for B. Thus, machine learning

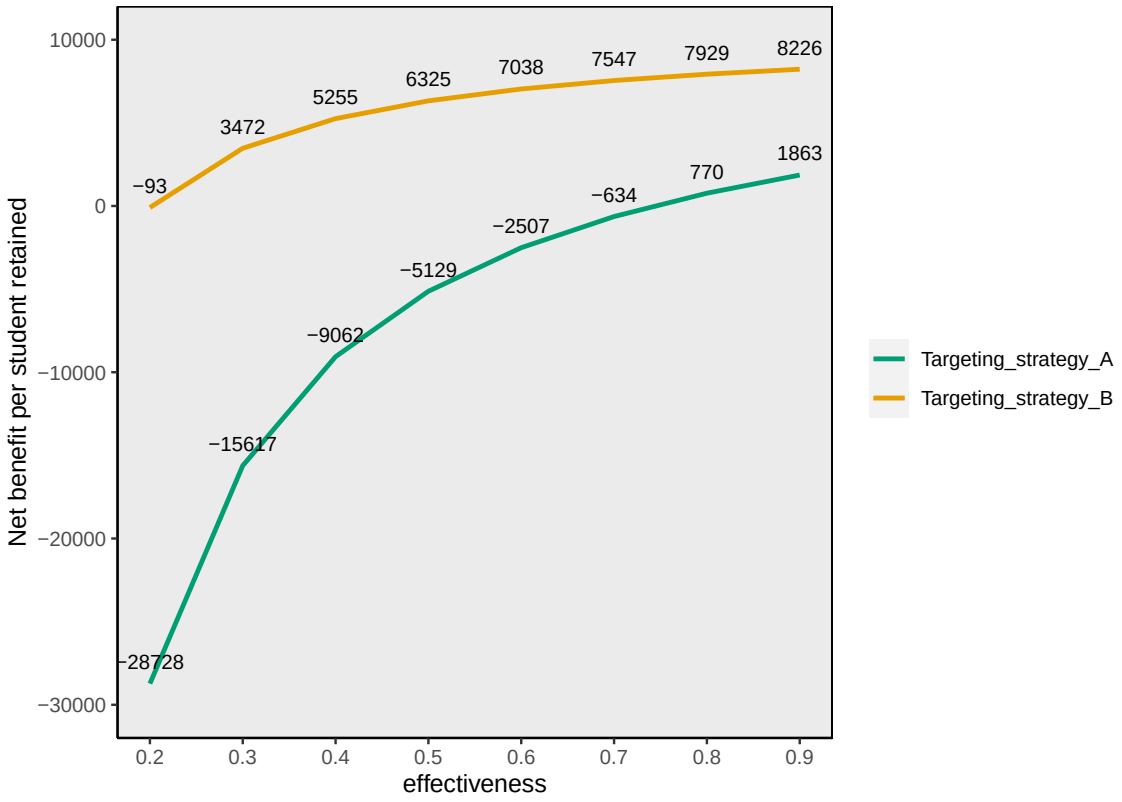


Figure 4.5: Net benefit per student retained of "En Bici a la Escuela" to prevent school dropout using different targeting strategies and levels of effectiveness

can offer substantial benefits over other targeting methods, especially if the program has a lower level of effectiveness.

Let me conclude with a comment regarding future extensions and improvements that are required to make these scenarios more realistic. We assume that the government would like to target students who have a high probability of dropping out within the next year. However, students who will likely dropout later should also be taken into account. Therefore, it would be appropriate to use predictions that factor in the probability of dropping out until the end of the cycle (Grade 11), as it is likely that some of the targeted students who don't dropout in the short run will have a very high probability of dropping out in the

future, which could result in lower costs per student retained. In addition, the predictions do not tell us anything about who is more likely to respond to an intervention like "En Bici a la Escuela" (Olaya et al., 2020). We are assuming that all students targeted would stay after participating in the program, but it is likely that some students would respond more to the program than others. It is necessary to design a Randomized Control Trial to estimate the effect of the program. Randomizing high probability students within high probability schools might be a good first step.

4.5 Conclusions and future research

This chapter has trained and compared different methods of Supervised Learning to the prediction of school dropout in Antioquia, Colombia. It shows how these methods can be used to target students at risk of dropping out and the most important predictors and partial dependence plots. The predictive performance of the models presented here is on-par with other prediction models of the Latin American region (Adelman et al., 2018). The administrative data used here is easily available for every education authority in Colombia and it could be leveraged to monitor the student's individual probability of dropout. However, more work is needed on adding other sources of data that allow monitoring a student's academic performance and attendance which the literature has shown are important predictors that could significantly improve the performance of these types of models (Sara et al., 2015; Adelman et al., 2018; Berens et al., 2018).

The prediction models presented here can become an important tool to target students for an intervention and they can reduce costs compared to targeting high dropout schools. As pointed out in this chapter, focusing on students with a high probability of dropout could be a first step, but more research is needed to identify groups of students who would benefit

from an intervention like “En Bici a la Escuela”. The targeting can be improved when the type of intervention designed is well defined and there is an opportunity to randomize the assignment of the program in order to estimate the causal effect it might have. This is a line for future research that could be pursued once the program is implemented and it would require designing an empirical strategy to evaluate heterogeneous treatment effects. Other programs could also be evaluated, for example, Colombia has been implementing the program “Familias en Acción” and one of the goals is to reduce school dropout, by giving families with children of school age cash transfers to keep students in the school system. Merging the data on school enrolment with data on “Familias en Acción” could be leveraged to answer the additional question of who is more likely to benefit from such a conditional cash transfer program and could improve the targeting of future programs. There are other studies in higher education that have attempted this (Olaya et al., 2020) using causal inference and machine learning to estimate heterogeneous treatment effects.

Finally, given that I find that delay is one of the most important predictors, one important conclusion is that retention strategies should start early to avoid delay once students reach secondary education. Students delayed by more than 2 years have an extreme probability of dropout as seen in the PDP. Students between 1 and 2 years of delay might be a good group to intervene. The probability of dropout for students in the last grades of secondary school reacts more to socio-economic status, more research is needed on whether monetary incentives might work better for these students.

Appendix

The purpose of this Appendix is twofold. I first show that the dropout rates calculated using the dataset described in Section 4.4.2 roughly coincide with the ones calculated using the

Level	Dropout rate (source: SIMAT Nacional)	Dropout rate (source: SIMAT Antioquia)
Primary	3.4%	3.2%
Secondary	5.4%	4.6%
High school	3.2%	3.2%

Source: Ministerio de Educación Nacional Colombia Ministerio de Educación Nacional (2019)

Table 4.7: Dropout rates by level of education in Antioquia (ETC) in 2019, estimated by the Ministry using the nationwide SIMAT dataset, compared to the one estimated using just the Antioquia SIMAT dataset

nationwide dataset. These results can be found in Table 4.7.

Secondly, I show that the out-of-sample performance of the prediction models LR, RF and GBM built for the year 2018 are similar to the ones presented in Section 4.4.2 for 2019. These results can be found in Figure 4.6, AUC goes from 75% to 82% for the different grades.

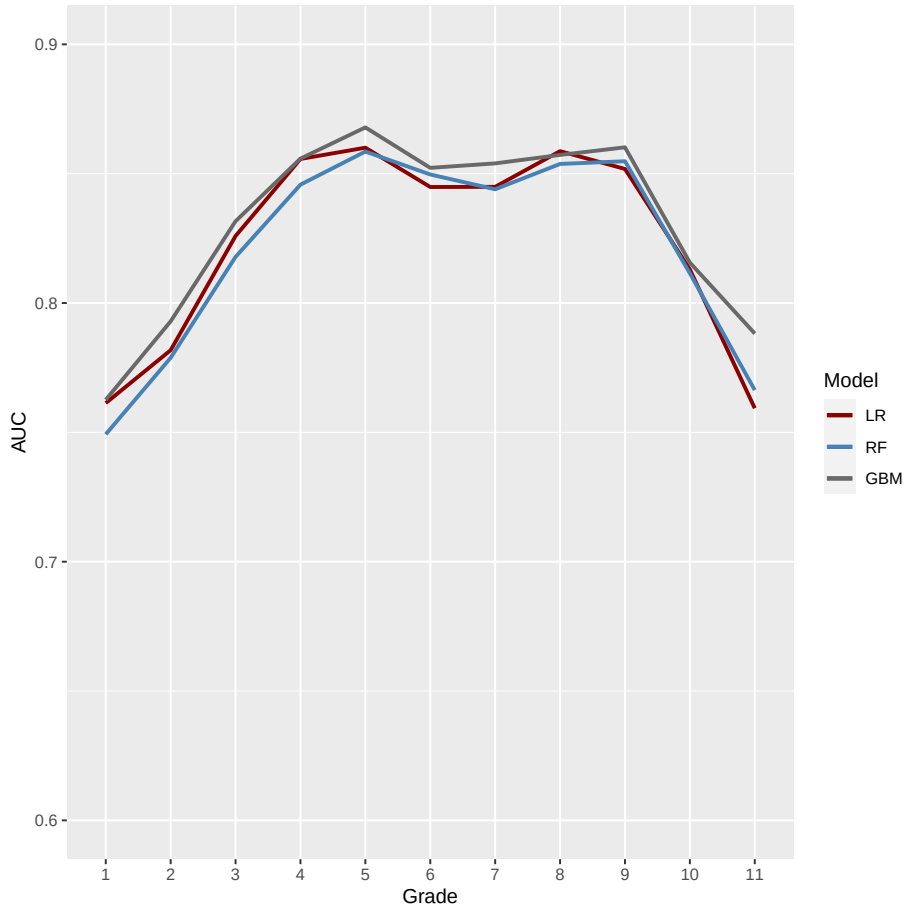


Figure 4.6: Out-of-sample AUC of the prediction models of school dropout -LR,RF,GBM- by grade, 2018

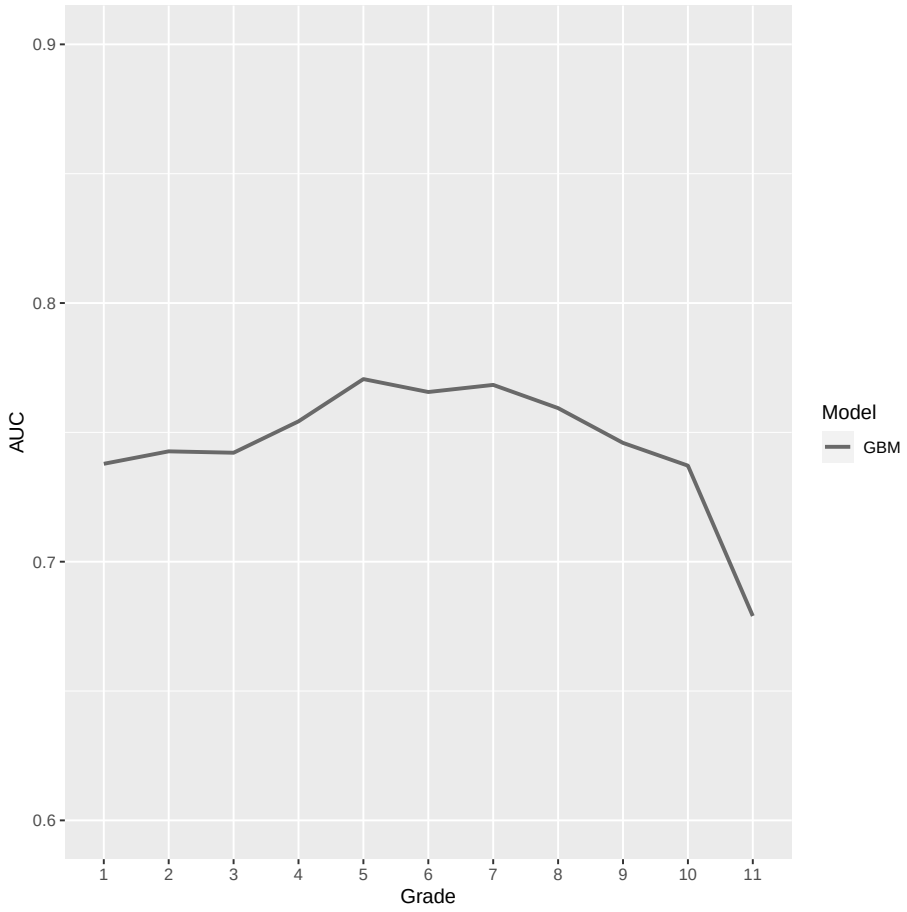


Figure 4.7: AUC of the prediction model of school dropout using GBM by grade, training with data from 2018 and predictions for 2019

Chapter 5

Improving the fairness of Generalized Linear Models by feature shrinkage

5.1 Introduction

In recent years, supervised classification has been used to support or even replace human decisions in high-stakes domains such as pre-trial risk assessment to decide who can be released while awaiting trial (Kleinberg et al., 2018; Berk, 2019; Jung et al., 2020; Završnik, 2021), police stop-and-frisk programs to decide who should be temporarily detained, questioned, and at times searched (Benbouzid, 2019), credit scoring to decide who gets a loan (Dastile et al., 2020), insurance premiums to decide the amount a client should pay (Henckaerts et al., 2021) and healthcare access to decide who is in need for extra care (Obermeyer et al., 2019). The training of these algorithms uses historical data which might be biased against individuals with sensitive characteristics, with consequences in terms of fairness and the potential for harming an individual's chances of fully participating in economic activity and society in general.

The increasing concern over biases has motivated lawmakers to pass anti-discrimination laws which prohibit unfair treatment based on specific characteristics such as gender or race (Goodman and Flaxman, 2017). Here we discuss the three main notions of unfairness found in the literature (Zemel et al., 2013; Feldman et al., 2015; Zafar et al., 2017a,b; Aghaei et al., 2019): *disparate treatment*, *disparate impact*, and *disparate mistreatment*. Disparate treatment arises when a decision-making system provides different outputs for groups of people with different values of the sensitive features, even if the value of their non-sensitive features are similar. For example, in a hypothetical scenario where we are classifying individuals according to creditworthiness (*good* or *bad*), disparate treatment with respect to the sensitive attribute *sex* would mean that even when *males* and *females* have similar attributes (other than *sex*), they have different probabilities of being classified as *good* or *bad*

payers. Disparate impact arises when the decision disproportionately benefits or hurts a specific group. In our hypothetical case, if we consider being classified as *good* in terms of creditworthiness as beneficial for the individual, disparate impact would arise when the probability of *good* differs for *men* and females. Note that, for discrimination to happen under the previous definitions, it is only necessary to know the decision made and not whether that decision was correct or not. Since the correctness of the decision is important in supervised classification, Zafar et al. (2017b) propose a third notion: *disparate mistreatment* considers the misclassification rates for the different groups. More specifically, a classifier suffers from disparate mistreatment with respect to a given sensitive attribute, if the misclassification rates differ for groups of people having different values of that sensitive attribute.

In this chapter, we have chosen *disparate mistreatment* as our notion of unfairness. In particular, we are interested in disparate mistreatment with respect to the false-negative rates since we define our positive class to reflect a beneficial outcome for the individual (e.g. in the case of the *German* dataset the individual gets the loan or in the case of the *COMPAS* dataset he/she walks out of jail). We define the protected group as individuals with the sensitive characteristic and the non-protected group as individuals without the sensitive characteristic. Hence, our measure of unfairness is the difference between the false-negative rates (FNR) for the protected and non-protected groups.

Throughout this section, we consider the following running example for illustrative purposes. In the *Adult* dataset, used in the numerical section, we are trying to classify individuals as having *high* (>=50k in income) or *low* income. Only about a third of the observations are *female*, which means that the number of *females* in the dataset is much smaller than *males*. In addition, only 11% of *females* are *high* income compared to 30%

of *males*. A classifier built on this data will tend to misclassify *females* as *low* income more often than *males*, as it can be seen in Figure 5.2 in the experimental results section ($FNR_{female} = 0.462 > FNR_{male} = 0.396$).

The bias does not disappear by simply eliminating the sensitive feature from the training process, since it may be highly correlated with other non-sensitive features in the dataset (Kamishima et al., 2012), thus indirectly affecting the response. For example, one ordinal categorical feature in the *Adult* dataset describes the *education* level. *Females* in the dataset tend to be underrepresented in all education levels, but especially in the highest ones, which are also the ones that have a higher likelihood of earning a higher salary. The same happens with numerical features in the dataset like the *number of hours worked*. *Males* tend to work more hours than *females* in this dataset and hence make a higher salary which means that even without the sensitive feature, the classifier would be able to distinguish between these groups using other correlated predictors.

Several approaches have been proposed in the literature to reduce unfairness by i) modifying the values of the sensitive attribute or the class labels in the training data (preprocessing approaches) (Dwork et al., 2012; Feldman et al., 2015), ii) directly modifying the classifier to incorporate fairness constraints (Kamishima et al., 2012; Zafar et al., 2017a,b; Carrizosa et al., 2021c; Blanquero et al., 2022), iii) post-processing the probability estimates of an unfair classifier to learn different decision thresholds for different sensitive attribute value groups (Hardt et al., 2016), and iv) a combination of i) and ii) where a fair representation of the data and the model parameters are learned jointly (Zemel et al., 2013). Like in iv), we mean to learn a fairer representation of the data while encoding the features in a way that preserves as much information about their relationship with the response.

In this chapter, we propose a methodology that enhances the trade-off between accu-

racy and unfairness in classification by shrinking the values the predictors can take. This shrinkage depends on whether the variable is categorical or numerical. Instead of having the original representation of a categorical feature, we propose to find a reduced representation, in the extreme case, as a dummy variable. This is achieved by a data-driven clustering of the categories into a number of clusters guided by a linear combination of accuracy and unfairness. A possible shrinkage of *education*, with 16 levels, would be a dummy variable that is equal to 1 for categories *bachelor*, *masters*, *prof-school*, *doctorate*, and 0 otherwise. By doing this, the percentage of *females* compared to *males* where this dummy variable is equal to 1, i.e., in the Higher education group, is similar to that when the dummy variable is equal to 0 (29% compared to 34%). This is much closer compared to the original representation of the data, where there was much more variation in the percentage of *females*, which ranged from less than 20% to more than 40%. For numerical predictors, we reduce the values they can take by collapsing the tails of the empirical distribution, according to a percentile that is chosen again using a linear combination of accuracy and unfairness. For instance, for the predictor *number of hours worked*, we shrink the values above 40 to this value. By doing this, the percentage of *females* compared to *males* working more than 40 hours ranges from 27% to 33%, while without the shrinkage this percentage varied from 14% to 33%.

In our numerical illustrations, the shrunk model, with the chosen reduced representation for all predictors in the dataset, shows a similar accuracy to the original model independently of the weight we use in the linear combination of accuracy and unfairness, while fairness improves when more weight is given to it. In other words, we avoid giving the classifier unnecessary information that might be harmful to especially sensitive groups.

The remainder of this chapter is structured as follows. The next section (Section 5.2)

introduces the algorithm to shrink predictors to enhance the trade-off between accuracy and unfairness in classification. Section 5.3 illustrates the performance of our method for a collection of real-world datasets, in terms of accuracy and unfairness. Finally, conclusions and future research are collected in Section 5.4.

5.2 Methodology

In this section, we present our approach to finding a shrunk representation of predictors with the aim of enhancing the trade-off between accuracy and unfairness in classification. We illustrate the method for Generalized Linear Models (GLM) with the *disparate mistreatment*, but our approach is also applicable to other classification methods and other measures of unfairness. In what follows, we first introduce the notation for the GLM and the notion of unfairness that we tackle. Then we explain how predictors are shrunk. We end the section with the pseudocode of the algorithm that chooses the shrunk representation of all predictors guided by a linear combination of the accuracy and the unfairness of the so-called shrunk GLM.

We are given a training sample of size N. We have J categorical predictors, P continuous predictors, as well as a sensitive attribute $z \in \{0, 1\}$. Categorical predictor j has K_j categories, $j = 1, \ldots, J$. In the GLM using the traditional one-hot encoding, a categorical predictor j with K_j categories is represented by $K_j - 1$ dummy variables, one for each category, leaving one out for contrast. We denote by **d** the vector of dummy variables associated with the categorical predictors, while **x** denotes the vector of continuous predictors. Consider a GLM where the outcome y is related to **d** and **x** through a link function G, namely,

$$\mathbb{E}\left[y|\mathbf{d},\mathbf{x}\right] = G(\beta_0 + (\boldsymbol{\beta})^T\mathbf{d} + (\tilde{\boldsymbol{\beta}})^T\mathbf{x}), \tag{5.1}$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is the vector of model parameters for the dummy variables and $\tilde{\boldsymbol{\beta}}$ the one for the continuous predictors. For a binary response variable $y \in \{0, 1\}$, a natural choice of link the function G is the Logit. This link function will be illustrated in Section 5.3, but our approach can handle any other link function in the family of GLMs.

As mentioned in the introduction, following Zafar et al. (2017b), we use the notion of *disparate mistreatment*. Without loss of generality, we will define the positive class ($y = 1$) as a beneficial outcome for the individual, and hence we would like to have low false-negative rates. An unfair classification will arise when individuals in the protected group, $z = 1$, have a higher false-negative rate than individuals in the non-protected group, $z = 0$. Hence, our measure of unfairness is the difference between the false-negative rate of the protected group and that of the non-protected group ($\Delta_{FNR} = FNR_{prot} - FNR_{non-prot}$).

We want to guarantee that our previously defined measure of unfairness Δ_{FNR} is as small as possible, while at the same time we want to get good overall accuracy. Then, we can combine these two criteria in the following equation:

$$V = \alpha \cdot Accuracy - (1 - \alpha) \cdot \Delta_{FNR}, \tag{5.2}$$

where $\alpha \in [0, 1]$ is a trade-off parameter that determines the weight given to accuracy. Hence, V is our objective function which we want to maximize.

Now let us explain how we shrink the categorical predictors in **d** first and then for the numerical predictors in **x**. We use the concept of feasible clusterings developed by Carrizosa et al. (2021b): we will say that a clustering of the categories of j into K' clusters is feasible

if each of the clusters consists of consecutive categories. We will assume that the categories are ordered. For instance, for ordinal categorical predictors, we could take the *natural* order of the predictor as the order of the categories. For non-ordinal categorical predictors, we could use the coefficients from the GLM with one-hot encoding to order the categories. The first clustering corresponds to having the first category in the predefined order in one group and the remaining categories in another one. The second clustering corresponds to having the first and second categories in one group and the remaining levels in another one. We successively move to include categories one by one in the first cluster, until we reach the last category, where all the categories are together in the first cluster and the second cluster is empty. This last clustering is equivalent to removing predictor j entirely from the model.

Once we have a feasible clustering, we can obtain the shrunk representation of the categorical predictor with $K' < K_j$ dummy variables, where we have a dummy variable for each of the K' clusters and the categories in the same cluster share the same coefficient in the GLM. For $K' = 2$, the shrunk representation consists of one single dummy variable indicating whether the category belongs to the first cluster or not. For a given categorical predictor, the different shrunk representations will yield, in general, different trade-offs between accuracy and our notion of unfairness Δ_{FNR}. This shrinking can be performed for each of the categorical predictors in **d**. Carrizosa et al. (2021b) claim that this reduced representation of the categorical features makes the model easier to interpret since it has fewer coefficients, reduces the possibility of overfitting, and can lower the standard errors since we have more data to estimate each coefficient.

Now, for the numerical predictors in **x**, we will also look at a set of shrunk representations. We consider that the protected and non-protected groups might have different distributions of the numerical predictors. For individuals in the protected group $z = 1$, we

calculate different percentiles for predictor x_p. Then, we collapse the right tail of the distribution using the calculated percentile. We do the same for individuals in the non-protected group $z = 0$, with their respective values of each percentile. Then, we have as many shrunk representations of x_p as percentiles in both groups, the protected and the non-protected.

The trade-off expressed in V should be measured for different shrunk representations of each of the predictors and the choice of one of these options will be guided by the maximization of V in Equation (5.2). We next argue the necessity to use a randomized numerical method to decide which feasible shrunk representation will be used in our reduced GLM for each predictor. We might decide that the best choice is that one with which the GLM achieves the highest V. However, representations that differ little may yield a very similar value of V. In the presence of multiple predictors to be reduced, it may be desirable to choose one with a good V, but not necessarily the best one. Therefore, we design a numerical method that chooses randomly between the shrunk representations with the highest V, see Figure 5.1. Once the shrunk representation is chosen for each predictor, we build the corresponding reduced GLM. These steps are repeated m times, and the algorithm returns the best GLM across the m iterations performed in terms of the out-of-sample linear combination of accuracy and unfairness in Equation (5.2).

5.3 Experimental results

In this section, we illustrate how our methodology performs in several real-world datasets. Our aim is to empirically analyze the effect of shrinking the predictors in a baseline classification method in terms of accuracy and unfairness. As the baseline procedure, we have chosen logistic regression. Accuracy is measured by the correct classification rate and unfairness is the difference between the false-negative rates for the protected and non-protected

Initialization: Let $\mathcal{L} = \{1, 2, \ldots, J\}$ be the set of predictors to be shrunk;
Let $V = \alpha \cdot Accuracy - (1 - \alpha) \cdot \Delta_{FNR}$ be the out-of-sample objective function;
Let m be the repeats of the algorithm;
for $j \in \mathcal{L}$ **do**
    Set $\mathcal{V}_j = \emptyset$;
    for *each shrunk representation of predictor* j **do**
        Estimate the GLM in Equation (5.1) where j is shrunk and the rest of the predictors are in their original form;
        Calculate V and add it to $\mathcal{V}_j$;
    end
    Sort $\mathcal{V}_j$ from max to min;
end
for $i \in \{1, \ldots, m\}$ **do**
    for $j \in \mathcal{L}$ **do**
        Choose randomly one value from the top h ones in $\mathcal{V}_j$. Replace j by its shrunk representation;
    end
    Estimate the GLM_i^S, the GLM in Equation (5.1) where all predictors in $\mathcal{L}$ are shrunk;
end
Return: GLM^S, the GLM_i^S with the highest V

Figure 5.1: Pseudocode of the algorithm to shrink predictors to enhance the trade-off between accuracy and unfairness

groups as defined in Section 5.1. Accuracy and unfairness 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 a shrinkage using the out-of-sample performance of the linear combination of the accuracy and the unfairness in Equation (5.2) in the testing sample and we report its final accuracy and unfairness in the validation sample. The process is repeated ten times and we report as an estimate the average accuracy and unfairness across the validation sets.

Our method uses an iterative algorithm, as described in Figure 5.1, where we aim to maximize the linear combination of accuracy and unfairness V in Equation (5.2). We shrink all our categorical predictors into $K' = 2$ clusters. For the numerical predictors we consider the following percentiles: $0, 75, 80, 85, 90, 95$. Percentile zero is equivalent to fitting a constant. In our experiments the parameters are set to $m = 100$ iterations, and selection in line 14 of pseudocode of Algorithm 5.1 is done out of the top $h = 3$ values of V in $\mathcal{V}_j$ for predictors with more than 5 categories and out of the top $h = 2$ otherwise. For the numerical predictors, we select from the top $h = 3$ percentiles. We consider values of $\alpha = 0, 0.1, 0.2, \ldots, 0.9$ in Equation (5.2). We coded our method in R and conducted our experiments in a Workstation with an Intel® Core™ i5-4460 processor with 8 Gb of RAM.

We use three real-world classification datasets to illustrate the method, the *Adult* dataset and the *German* credit dataset are available in the UCI Machine Learning Repository (Dua and Graff, 2017), and the *COMPAS* dataset is available in ProPublica's GitHub (Angwin et al., 2016). The datasets are described in Table 5.1, in the first two columns we report the name and the total number of records in the dataset (N). In columns three to eight, we report the class split in percentage, the number of protected attributes (Z), the number of categorical (J) and continuous predictors (P), the total number of categories ($\sum_{j=1}^{J} K_j$)

Name	N	Class-split	Z	J	P	$\sum_{j=1}^{J} K_j$	K_j
Adult	32561	24/76	2	10	2	117	8,5,16,5,7,14,6,5,5,41
German	1000	30/70	1	11	9	52	4,5,11,5,5,5,3,4,3,3,4
COMPAS	6150	46/54	1	0	4	–	–

Table 5.1: Description of the datasets to illustrate accuracy and unfairness with our methodology

across all categorical predictors, and the number of categories for each categorical predictor K_j, respectively. The *Adult* dataset has two attributes that can be considered sensitive, we apply our method with each sensitive attribute separately.

We now discuss the average out-of-sample performance of our method by comparing the results of the *original* and the *shrunk* model for different values of the parameter α, with the support of Figures 5.2–5.5. For each value of α, the figures display accuracy, unfairness, and false negative rates for the protected and non-protected groups, respectively, and on the right is similar information, but for the original model.

In Figure 5.2 we show the results for the *Adult* dataset, considering *sex* as the protected attribute. The *original* model has accuracy of 85.0% with a $FNR_{prot} = 46.2\%$ and $FNR_{non-prot} = 39.6\%$ and therefore an unfairness of 6.6%. We see that for the *shrunk* model, built with the proposed methodology in Section 5.2), with $\alpha = 0.9$, i.e., the smallest weight assigned to reducing unfairness, we are already able to lower unfairness from 6.6% to 2.5% compared to the *original* model, with a mild decrease in accuracy, namely, from 85% to 83.2%. If we move further to the left in the plot, unfairness can be reduced, even more, by giving a smaller weight to accuracy and therefore a larger one to reducing unfairness. It is even possible to obtain a model that has parity between the groups at $\alpha = 0.4$ while still having very similar accuracy to the original model (84.4% compared to 85.0%).

Figure 5.3 shows results for the *Adult* dataset when we consider that the protected at-

tribute is *race*: *black, ameri-indian-eskimo, asian-pac-islander, other* and the non-protected group is *race*: *white*. The *original* model has accuracy of 85.0% with a $FNR_{prot} = 45.6\%$, $FNR_{non-prot} = 39.7\%$, and therefore an unfairness of 5.9%. We see that the *shrunk* model that gives the highest weight to the accuracy of all the ones tested, namely $\alpha = 0.9$, is less fair than the original one, with the unfairness increased to 8.3%. At $\alpha = 0.7$ we reach the lowest unfairness (1.9%), with a slight reduction in accuracy (84.1% compared to 85.0%).

Figure 5.4 shows the analysis for the *German* dataset with *sex* as the protected attribute. The *original* model has accuracy of 73.7% with a $FNR_{prot} = 19.5\%$, $FNR_{non-prot} = 14.1\%$, and therefore an unfairness 5.4%. For $\alpha = 0.6$ we can achieve lower unfairness, 1.6%, and higher accuracy, 75.3%, compared to the original model. This increase in accuracy is probably due to the fact that this dataset only contains 1000 observations, and the number of coefficients to estimate for the *original* model with all categorical predictors as dummies, is relatively large, namely 73, and imbalanced (see Table 5.1). The shrunk model has the same number of observations with 20 coefficients to estimate. Since we have grouped similar categories, the number of observations to estimate each coefficient is larger and hence we have a more stable model (see Carrizosa et al. (2021a,b) for a discussion) which achieves higher accuracy.

Figure 5.5 shows the results for the *COMPAS* dataset with *race* as the protected attribute. The *original* model has accuracy of 67.5% with $FNR_{prot} = 33.8\%$, $FNR_{non-prot} = 14.3\%$ and therefore an unfairness of 19.6%. For $\alpha = 0.9$ we can reduce unfairness to 12.2% without damaging accuracy. At $\alpha = 0.5$ we reach the lowest unfairness (11.3%) with slightly less accuracy compared to the *original* model (66.5% compared to 67.5%).

For *Adult race* and *German*, unfairness was reduced, even with a reduction in the FNR for both groups. In *Adult sex*, the improvement in unfairness was reached both by increasing

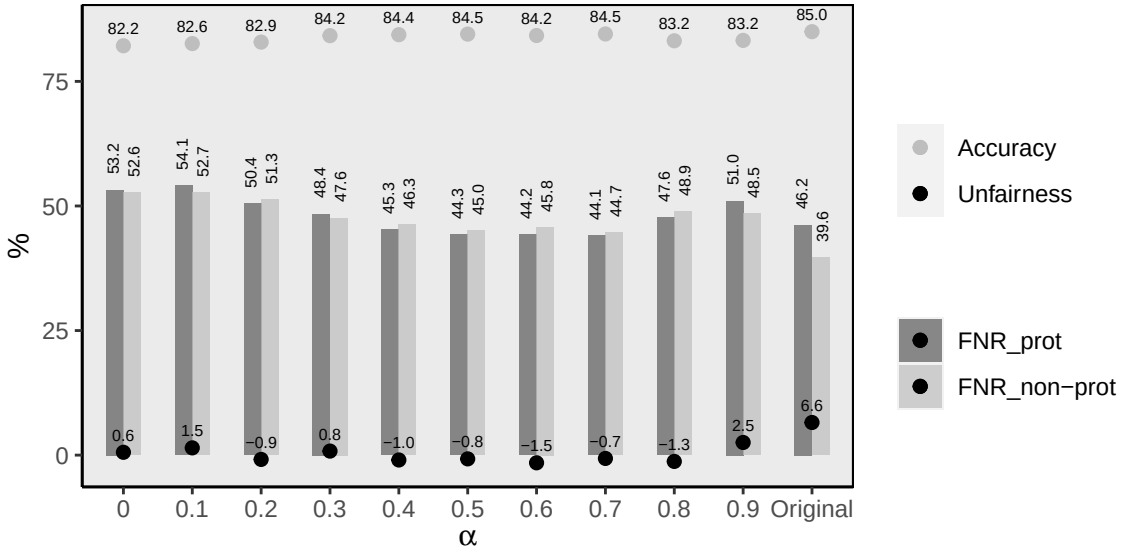


Figure 5.2: Accuracy, Unfairness and False Negative Rates for the protected (*sex: female*) and non-protected (*sex: male*) group for the proposed feature shrinkage methodology in the *Adult* dataset using different values of the parameter α, compared to the original model

the FNR of the non-protected group, as well as by decreasing the FNR for the protected group. Although *COMPAS* has a much higher degree of unfairness than the other datasets, our method can still reduce it without compromising accuracy, at the cost of a higher FNR for both groups.

5.4 Conclusions

In this chapter, we proposed a methodology that enhances the trade-off between accuracy and unfairness in classification with Generalized Linear Models, by finding a shrunk representation of the features. Our numerical results illustrate that with our approach we can find a less biased representation of the data, in some cases even reducing disparate mistreatment to zero, without harming accuracy.

In terms of future research, it would be interesting to incorporate several notions of un-

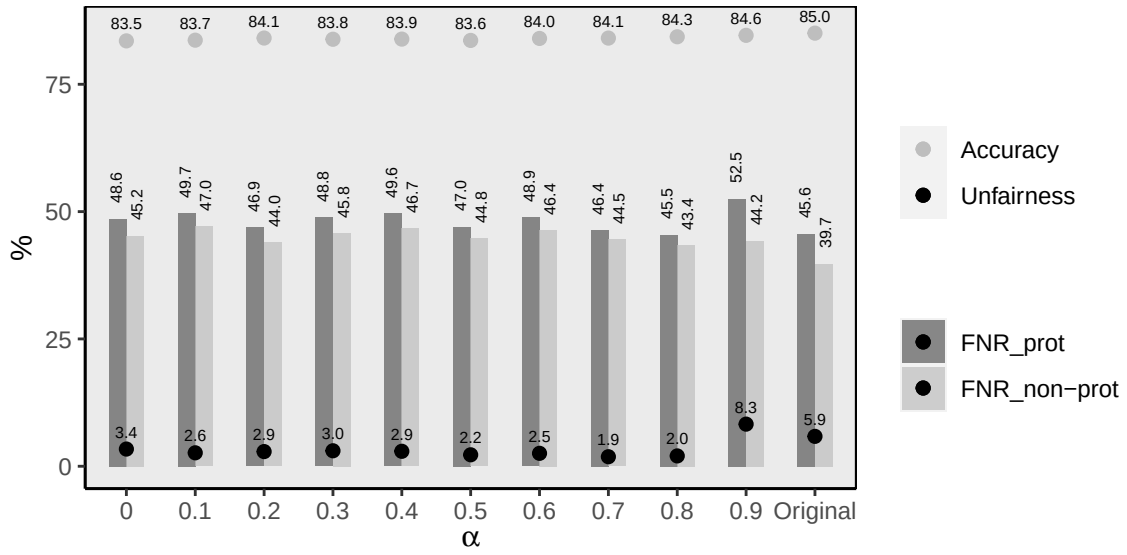


Figure 5.3: Accuracy, Unfairness and False Negative Rates for the protected (*race: black, ameri-indian-eskimo, asian-pac-islander, other*) and non-protected (*race: white*) group for the proposed feature shrinkage methodology in the *Adult* dataset using different values of the parameter α, compared to the original model

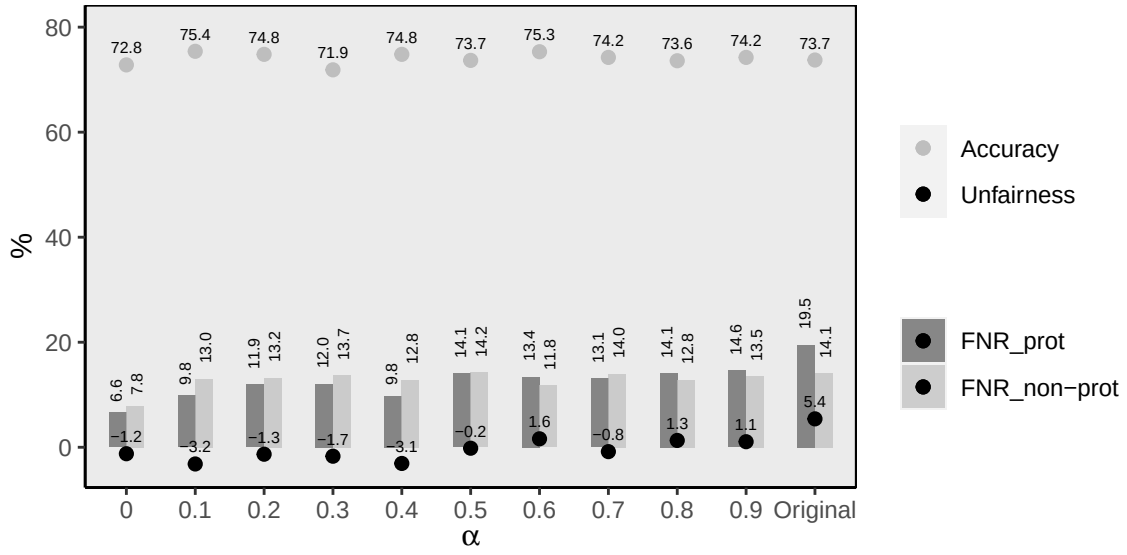


Figure 5.4: Accuracy, Unfairness and False Negative Rates for the protected (*sex: female*) and non-protected (*sex: male*) group for the proposed feature shrinkage methodology in the *German* dataset using different values of the parameter α, compared to the original model

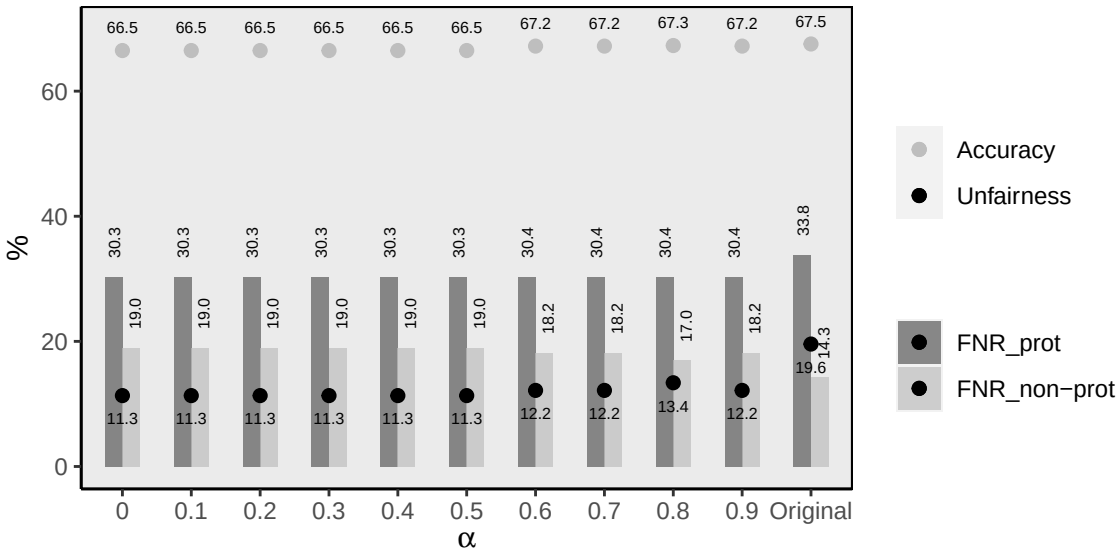


Figure 5.5: Accuracy, Unfairness and False Negative Rates for the protected (*race: African-American*) and non-protected (*race: white*) group for the proposed feature shrinkage methodology in the *COMPAS* dataset using different values of the parameter α, compared to the original model

fairness. Other research has shown that some of these unfairness criteria might be mutually incompatible (Kleinberg et al., 2016; Chouldechova, 2017; Jung et al., 2020). In this case, one would need to model the trade-off between accuracy and the multiple unfairness criteria in consideration, yielding a more complex objective function than the one in (5.2), with more trade-off parameters. Another interesting line of future research relates to the number of sensitive features and thus the number of groups to be protected from unfairness. This would yield again multiple unfairness criteria and thus a higher dimensional trade-off parameter space.

List of Figures

List of Tables

Bibliography

Adelman, M., Haimovich, F., Ham, A., and Vazquez, E. (2018). Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. *Education Economics*, 26(4):356–372.

Aghaei, S., Azizi, M., and Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426.

Allensworth, E. M. (2005). Dropout rates after high-stakes testing in elementary school: A study of the contradictory effects of Chicago's efforts to end social promotion. *Educational Evaluation and Policy Analysis*, 27(4):341–364.

Angwin, J., Kirchner, L., Larson, J., and Mattu, S. (2016). How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm#:~:text=We\%20compared\%20the\%20recidivism\%20risk,predictions\%20of\%20violent\%20recidivism\%2020. Accessed: 2021-11-10.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Athey, S. (2019). The impact of machine learning on economics. In *The Economics of Artificial Intelligence*, pages 507–552. University of Chicago Press.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Bassi, M., Busso, M., and Muñoz, J. S. (2015). Is the glass half empty or half full? enrolment, graduation, and dropout rates in Latin America. *Economía*, 16(1):113–156.

Becker, G. S. (2009). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

Behrman, J. R., De Hoyos Navarro, R. E., and Székely, M. (2014). Out of school and out of work: a conceptual framework for investigating "ninis" in Latin America and the Caribbean. Technical report, The World Bank, `https://openknowledge.worldbank.org/handle/10986/23835`.

Benbouzid, B. (2019). To predict and to manage. predictive policing in the United States. *Big Data & Society*, 6(1):2053951719861703.

Bentaouet Kattan, R. and Székely, M. (2015). Patterns, consequences, and possible causes of dropout in upper secondary education in Mexico. *Education Research International*, 2015.

Berens, J., Schneider, K., Görtz, S., Oster, S., and Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. Technical report, CESifo. `https://ideas.repec.org/p/ces/ceswps/_7259.html`.

Berk, R. (2019). *Machine learning risk assessments in criminal justice settings*. Springer.

Bertsimas, D., King, A., Mazumder, R., et al. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141.

Blanco, D. (2014). La migración interna contemporánea en antioquia desde la perspectiva de la teoría de sistemas. *Revista Virajes*, 16:298–327.

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., and Martín-Barragán, B. (2019). Variable selection in classification for multivariate functional data. *Information Sciences*, 481:445–462.

Blanquero, R., Carrizosa, E., Molero-Río, C., and Romero Morales, D. (2022). On sparse optimal regression trees. *European Journal of Operational Research*, 299(3):1045–1054.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bonilla Mejia, L. (2016). *Three essays on education decisions in Colombia*. PhD thesis, University of Illinois at Urbana-Champaign.

Boselli, R., Cesarini, M., Mercorio, F., and Mezzanzanica, M. (2018). Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, 86:319–328.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Buckmann, M., Joseph, A., and Robertson, H. (2021). An interpretable machine learning workflow with an application to economic forecasting. Technical report, Bank of England, `https://www.bankofengland.co.uk/working-paper/2022/an-interpretable-machine-learning-workflow-with-an-application-to-economic-forecasting`.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636.

Busetta, G., Campolo, M. G., and Panarello, D. (2020). Weight-based discrimination in the Italian labor market: an analysis of the interaction with gender and ethnicity. *The Journal of Economic Inequality*, 18(4):617–637.

Carrizosa, E., Galvis Restrepo, M., and Romero Morales, D. (2021a). A binarization approach to model interactions between categorical predictors. Technical report, Copenhagen Business School, Frederiksberg, Denmark, `https://www.researchgate.net/publication/350755054_A_binarization_approach_to_model_interactions_between_categorical_predictors_in_Generalized_Linear_Models`.

Carrizosa, E., Galvis Restrepo, M., and Romero Morales, D. (2021b). On clustering categories of categorical predictors in generalized linear models. *Expert Systems with Applications*, page 115245.

Carrizosa, E., Galvis Restrepo, M., and Romero Morales, D. (2022a). Improving fairness of generalized linear models by feature shrinkage. Technical report, Copenhagen Business School, Frederiksberg, Denmark, `https://www.researchgate.net/publication/358614960_Improving_fairness_of_Generalized_Linear_Models_by_feature_shrinkage`.

Carrizosa, E., Martín-Barragán, B., and Romero Morales, D. (2010). Binarized support vector machines. *INFORMS Journal on Computing*, 22(1):154–167.

Carrizosa, E., Martín-Barragán, B., and Romero Morales, D. (2011). Detecting relevant variables and interactions in supervised classification. *European Journal of Operational Research*, 213(1):260–269.

Carrizosa, E., Molero-Río, C., and Romero Morales, D. (2021c). Mathematical optimization in classification and regression trees. *TOP*, 29(1):5–33.

Carrizosa, E., Mortensen, L. H., Romero Morales, D., and Sillero-Denamiel, M. R. (2022b). The tree based linear regression model for hierarchical categorical variables. *Expert Systems with Applications*, 203:117423.

Carrizosa, E., Nogales-Gómez, A., and Romero Morales, D. (2016). Strongly agree or strongly disagree?: Rating features in Support Vector Machines. *Information Sciences*, 329:256–273.

Carrizosa, E., Nogales-Gómez, A., and Romero Morales, D. (2017). Clustering categories in support vector machines. *Omega*, 66:28–37.

Castañeda, T. and Fernández, L. (2005). Targeting social spending to the poor with proxy-means testing: Colombia's SISBEN system. *World Bank Human Development Network - Social Protection Unit Discussion Paper*, 529.

Chandler, D., Levitt, S. D., and List, J. A. (2011). Predicting and preventing shootings among at-risk youth. *American Economic Review*, 101(3):288–92.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Daniels, J. P. (2015). Tackling teenage pregnancy in Colombia. *The Lancet*, 385(9977):1495–1496.

Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263.

Dávila Ramírez, F. A., Fajardo Granados, D. E., Jiménez Cruz, C. A., Florido Pérez, C., and Vergara Castellón, K. C. (2016). Factores de riesgo psicosocial para embarazo temprano y deserción escolar en mujeres adolescentes. *Revista Ciencias de la Salud*, 14(1):93–101.

De Witte, K., Cabus, S., Thyssen, G., Groot, W., and van Den Brink, H. M. (2013). A critical review of the literature on school dropout. *Educational Research Review*, 10:13–28.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336.

Departamento Administrativo Nacional de Estadística, C. (2021). GEIH gran encuesta integrada de hogares. `https://www.dane.gov.co/index.php/178-english/sociales/cultura/2921-gran-encuesta-integrada-de-hogares`. Accessed: 2022-06-09.

Departamento de Antioquia, C. (2018a). Anuario Estadístico de Antioquia. `https://www.antioquiadatos.gov.co/index.php/anuario-estadistico-home`. Accessed: 2020-05-11.

Departamento de Antioquia, C. (2018b). Anuario Estadístico de Antioquia. `http://www.antioquiadatos.gov.co/index.php/poblacion-348`. Accessed: 2020-05-11.

Detmer, F. J., Cebral, J., and Slawski, M. (2020). A note on coding and standardization of categorical variables in (sparse) group lasso regression. *Journal of Statistical Planning and Inference*, 206:1–11.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Dua, D. and Graff, C. (2017). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226.

Entwisle, D. R., Alexander, K. L., and Olson, L. S. (2004). Temporary as compared to permanent high school dropout. *Social Forces*, 82(3):1181–1205.

Entwisle, D. R., Alexander, K. L., and Olson, L. S. (2005). Urban teenagers: Work and dropout. *Youth & Society*, 37(1):3–32.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268.

Feng, G., Guo, J., Jing, B.-Y., and Sun, T. (2015). Feature subset selection using naive bayes for text classification. *Pattern Recognition Letters*, 65:109–115.

Fiske, E. B. (1998). Wasted opportunities: When schools fail. Repetition and drop-out in primary schools. In *Education for All: Status and Trends. UNESCO, Paris*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Fundación Corona, Fundación ANDI, Programa de Alianzas para la Reconciliación de US-AID, and ACDI/VOCA (2020). Informe Nacional de Empleo Inclusivo INEI 2018-2019.

https://www.fundacioncorona.org/es/biblioteca/documentos-tecnicos/informe-nacional-de-empleo-inclusivo-inei-2018-2019. Accessed: 2022-01-26.

Gaebel, M., Hauschildt, K., Mühleck, K., and Smidt, H. (2012). *Tracking learners' and graduates' progression paths TRACKIT*. European University Association Brussels, Belgium.

García, I. J., González, L. A., Salgado, M. Q., Vásquez, J., and González, N. (2018). Lineamientos de política para la permanencia y graduación estudiantil. http://200.13.244.221:8080/SGI/Acreditaci\%C3\%B3n/Documentos\%20institucionales/Inventario/Documento\%20pol\%C3\%ADtica\%20de\%20permanencia\%20y\%20graduaci\%C3\%B3n\%20estudiantil.pdf. Accessed: 2021-05-10.

Garside, M. (1965). The best sub-set in multiple regression analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 14(2-3):196–200.

Goel, S., Rao, J. M., Shroff, R., et al. (2016). Precinct or prejudice? understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Gómez-Restrepo, C., Muñoz, A. P., and Rincón, C. J. (2016). Deserción escolar de ado-

lescentes a partir de un estudio de corte transversal: Encuesta nacional de salud mental Colombia 2015. *Revista Colombiana de Psiquiatría*, 45:105–112.

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Goorbergh, R. v. d., van Smeden, M., Timmerman, D., and Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *arXiv preprint arXiv:2202.09101*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hazimeh, H. and Mazumder, R. (2020). Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537.

Henckaerts, R., Côté, M.-P., Antonio, K., and Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2):255–285.

Herrera Prada, L. O. (2021). Ending the musical chairs game in higher education: How a software dashboard (spadies) improved information flow and educational outcomes

in Colombia. Technical report, IAI Discussion Papers, https://www.econstor.eu/handle/10419/233065.

Howard, K. A., Carlstrom, A. H., Katz, A. D., Chew, A. Y., Ray, G. C., Laine, L., and Caulum, D. (2011). Career aspirations of youth: Untangling race/ethnicity, SES, and gender. *Journal of Vocational Behavior*, 79(1):98–109.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Jensen, M. F. (2021). *Essays on Gender and Skills in the Labour Market*. PhD thesis, Copenhagen Business School.

Jimerson, S., Carlson, E., Rotert, M., Egeland, B., and Sroufe, L. A. (1997). A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology*, 35(1):3–25.

Johannemann, J., Hadad, V., Athey, S., and Wager, S. (2019). Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*.

Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., and Vohra, R. (2020). Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 677–678.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.

Kingston, G., McGinnity, F., and O'Connell, P. J. (2015). Discrimination in the labour market: nationality, ethnicity and the recession. *Work, Employment and Society*, 29(2):213–232.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Knowles, J. E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3):18–67.

LeBlanc, M. and Tibshirani, R. (1998). Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, 7(4):417–433.

Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

McBride, L. and Nichols, A. (2018). Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Ministerio de Educación Nacional, C. (2019). Estadísticas en educación preescolar, básica y media Colombia. https://www.datos.gov.co/Educaci-n/MEN_ESTADISTICAS_EN_EDUCACION_EN_PREESCOLAR-B-SICA/sras-4t5p. Accessed: 2022-01-24.

Ministerio de Educación Nacional, C. (2021). SIMAT sistema de matricula. https://www.mineducacion.gov.co/1759/w3-article-168883.html?_noredirect=1. Accessed: 2021-10-11.

Ministerio de Educación Nacional, C. (2022a). Colombian Education System. https://www.mineducacion.gov.co/1759/w3-article-355502.html?_noredirect=1#:~:text=The\%20{C}olombian\%20education\%20system\%20is,degree)\%2C\%20and\%20higher\%20education. Accessed: 2022-01-24.

Ministerio de Educación Nacional, C. (2022b). Directorio Único de Establecimientos. https://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-296670.html?_noredirect=1. Accessed: 2022-01-24.

Moeyersoms, J., d'Alessandro, B., Provost, F., and Martens, D. (2016). Explaining classification models built on high-dimensional sparse data. *arXiv preprint arXiv:1607.06280*.

Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning – a brief history, state-of-the-art and challenges. *arXiv preprint arXiv:2010.09337*.

Morales, L. F. (2015). Peer effects on a fertility decision: An application for Medellín, Colombia. *Economía*, 15(2):119–159.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

O'Cummings, M. and Therriault, S. B. (2015). From accountability to prevention: Early warning systems put data to work for struggling students. Technical report, American Institutes for Research, `https://files.eric.ed.gov/fulltext/ED576665.pdf`.

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., and Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decision Support Systems*, 134:113320.

Oreopoulos, P. (2007). Do dropouts drop out too soon?: Wealth, health and happiness from compulsory schooling. *Journal of Public Economics*, 91(11-12):2213–2229.

Oreopoulos, P. and Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, 25(1):159–84.

Pardo Pinzón, R. and Sorzano Montaña, O. (2004). Determinantes de la asistencia y la deserción escolar en primaria y secundaria. *Determinants of School Attendance and school dropout in primary and secondary school. Cuadernos PNUD/MPS, Investigaciones sobre desarrollo social en Colombia*, 3:17–81.

Pejic-Bach, M., Bertoncel, T., Meško, M., and Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, 50:416–431.

Pinzón Hernández, D. A. (2018). Reprobación y desempeño académico: evidencia de la implementación de la promoción automática en Colombia. Technical report, Universidad de los Andes, Facultad de Economía, Serie de Documentos CEDE, `https://repositorio.uniandes.edu.co/handle/1992/7862`.

Plank, S., DeLuca, S., and Estacion, A. (2005). Dropping out of high school and the place of career and technical education: A survival analysis of surviving high school. Technical report, National Research Center for Career and Technical Education, `https://files.eric.ed.gov/fulltext/ED497348.pdf`.

Post, D. (2011). Primary school student employment and academic achievement in Chile, Colombia, Ecuador and Peru. *International Labour Review*, 150(3-4):255–278.

Psacharopoulos, G. and Patrinos, H. A. (2018). Returns to investment in education: a decennial review of the global literature. *Education Economics*, 26(5):445–458.

Resende, M. G. and Ribeiro, C. C. (2016). *Optimization by GRASP*. Springer.

Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

Rumberger, R. W. (2004). Why students drop out of school. In *Conference: Dropouts in America: Confronting the graduation rate crisis*, pages 131–155. Harvard Education Press.

Rumberger, R. W. and Lamb, S. P. (2003). The early employment and further education

experiences of high school dropouts: A comparative study of the United States and Australia. *Economics of Education Review*, 22(4):353–366.

Rumberger, R. W. and Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education*, 107(1):1–35.

Sara, N.-B., Halland, R., Igel, C., and Alstrup, S. (2015). High-school dropout prediction using machine learning: A danish large-scale study. In *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, pages 319–24.

Secretaria de Educación de Antioquia, C. (2022). Concurso en mi bici a la escuela. `https://www.seduca.gov.co/sala-de-prensa/archivo-de-prensa/item/6208-concurso-en-mi-bici-a-la-escuela`. Accessed: 2022-06-09.

Seibold, H., Zeileis, A., and Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics*, 12(1):45–63.

Silva Arias, A. C. and González Román, P. (2009). Un análisis espacial de las migraciones internas en Colombia (2000-2005). *Revista Facultad de Ciencias Económicas: Investigación y Reflexión*, 17(1):123–144.

Simon, F., Małgorzata, K., and Beatriz, P. (2007). *Education and training policy no more failures: Ten steps to equity in education*. OECD Publishing.

Toutkoushian, R. K., Bellas, M. L., and Moore, J. V. (2007). The interaction effects of gender, race, and marital status on faculty salaries. *The Journal of Higher Education*, 78(5):572–601.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Viljanen, M. and Pahikkala, T. (2020). Predicting unemployment with machine learning based on registry data. In *International Conference on Research Challenges in Information Science*, pages 352–368. Springer.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Weisberg, H. I. and Pontes, V. P. (2015). Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical Trials*, 12(4):357–364.

Weiss, G. M. and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zafar, M., Valera, I., Gomez Rodriguez, M., and Gummadi, K. (2017a). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. Proceedings of Machine Learning Research.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017b). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology*, 18(5):623–642.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333. Proceedings of Machine Learning Research.

Zhao, Q. and Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):1–10.

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or, How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning – A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus – Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstilskrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynamiske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing performances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle læringsmiljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under
1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organizational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimensions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som interaktionsform mellem kunde og medarbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter. En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitetsarbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin
– en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management - A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
CSR som noget særligt
Et casestudie om styring og meningsskabelse i relation til CSR ud fra en intern optik

11. Michael Tell
Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C

12. Morten Holm
Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts

13. Katja Joo Dyppel
Beskatning af derivater
En analyse af dansk skatteret

14. Esben Anton Schultz
Essays in Labor Economics
Evidence from Danish Micro Data

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation

17. Frans Bévort
Making sense of management with logics
An ethnographic study of accountants who become managers

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008

22. Jacob Holm Hansen
Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
Promises of Reflexivity
Managing and Researching Inclusive Schools

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
Auditor's going-concern reporting
Reporting decision and content of the report

28. John Bang Mathiasen
Learning within a product development working practice:
- an understanding anchored in pragmatism

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være? Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement
Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sensemaking in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a Partly-Digital World
Comparing Digital Piracy and Legal Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work
An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in organizational decision-making

11. Mia Kaspersen
The construction of social and environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming
The prohibition against misleading names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED!
Kan giver-idealtyper forklare støtte til velgørenhed og understøtte relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning
I et internationalt og komparativt perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor
En kortlægning af styringsmæssig kontekst, strategisk tilgang, samt anvendte redskaber og teknologier for udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation
An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization of Emerging Market Firms:
A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside":
The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural Production
The case of travel guidebook production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie i det 19. århundrede
Omtydningen af melankolien og manien som bipolære stemningslidelser i dansk sammenhæng under hensyn til dannelsen af det moderne følelseslivs relative autonomi.
En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process
Circulating References and Matters of Concern

24. Esben Alfort
The Expression of a Need
Understanding search

25. Trine Pallesen
Assembling Markets for Wind Power
An Inquiry into the Making of Market Devices

26. Anders Koed Madsen
Web-Visions
Repurposing digital traces to organize social attention

27. Lærke Højgaard Christiansen
BREWING ORGANIZATIONAL RESPONSES TO INSTITUTIONAL LOGICS

28. Tommy Kjær Lassen
EGENTLIG SELVLEDELSE
En ledelsesfilosofisk afhandling om selvledelsens paradoksale dynamik og eksistentielle engagement

29. Morten Rossing
Local Adaption and Meaning Creation in Performance Appraisal

30. Søren Obed Madsen
Lederen som oversætter
Et oversættelsesteoretisk perspektiv på strategisk arbejde

31. Thomas Høgenhaven
Open Government Communities
Does Design Affect Participation?

32. Kirstine Zinck Pedersen
Failsafe Organizing?
A Pragmatic Stance on Patient Safety

33. Anne Petersen
Hverdagslogikker i psykiatrisk arbejde
En institutionsetnografisk undersøgelse af hverdagen i psykiatriske organisationer

34. Didde Maria Humle
Fortællinger om arbejde

35. Mark Holst-Mikkelsen
Strategieksekvering i praksis
– barrierer og muligheder!

36. Malek Maalouf
Sustaining lean
Strategies for dealing with organizational paradoxes

37. Nicolaj Tofte Brenneche
Systemic Innovation In The Making
The Social Productivity of Cartographic Crisis and Transitions in the Case of SEEIT

38. Morten Gylling
The Structure of Discourse
A Corpus-Based Cross-Linguistic Study

39. Binzhang YANG
Urban Green Spaces for Quality Life
- Case Study: the landscape architecture for people in Copenhagen

40. Michael Friis Pedersen
Finance and Organization:
The Implications for Whole Farm Risk Management

41. Even Fallan
Issues on supply and demand for environmental accounting information

42. Ather Nawaz
Website user experience
A cross-cultural study of the relation between users´ cognitive style, context of use, and information architecture of local websites

43. Karin Beukel
The Determinants for Creating Valuable Inventions

44. Arjan Markus
External Knowledge Sourcing and Firm Innovation
Essays on the Micro-Foundations of Firms' Search for Innovation

2014

1. Solon Moreira
Four Essays on Technology Licensing and Firm Innovation

2. Karin Strzeletz Ivertsen
Partnership Drift in Innovation Processes
A study of the Think City electric car development

3. Kathrine Hoffmann Pii
Responsibility Flows in Patient-centred Prevention

4. Jane Bjørn Vedel
Managing Strategic Research
An empirical analysis of science-industry collaboration in a pharmaceutical company

5. Martin Gylling
Processuel strategi i organisationer
Monografi om dobbeltheden i tænkning af strategi, dels som vidensfelt i organisationsteori, dels som kunstnerisk tilgang til at skabe i erhvervsmæssig innovation

6. Linne Marie Lauesen
Corporate Social Responsibility in the Water Sector: How Material Practices and their Symbolic and Physical Meanings Form a Colonising Logic

7. Maggie Qiuzhu Mei
LEARNING TO INNOVATE: The role of ambidexterity, standard, and decision process

8. Inger Høedt-Rasmussen
Developing Identity for Lawyers Towards Sustainable Lawyering

9. Sebastian Fux
Essays on Return Predictability and Term Structure Modelling

10. Thorbjørn N. M. Lund-Poulsen
Essays on Value Based Management

11. Oana Brindusa Albu
Transparency in Organizing: A Performative Approach

12. Lena Olaison
Entrepreneurship at the limits

13. Hanne Sørum
DRESSED FOR WEB SUCCESS? An Empirical Study of Website Quality in the Public Sector

14. Lasse Folke Henriksen
Knowing networks How experts shape transnational governance

15. Maria Halbinger
Entrepreneurial Individuals Empirical Investigations into Entrepreneurial Activities of Hackers and Makers

16. Robert Spliid
Kapitalfondenes metoder og kompetencer

17. Christiane Stelling
Public-private partnerships & the need, development and management of trusting A processual and embedded exploration

18. Marta Gasparin
Management of design as a translation process

19. Kåre Moberg
Assessing the Impact of Entrepreneurship Education From ABC to PhD

20. Alexander Cole
Distant neighbors Collective learning beyond the cluster

21. Martin Møller Boje Rasmussen
Is Competitiveness a Question of Being Alike? How the United Kingdom, Germany and Denmark Came to Compete through their Knowledge Regimes from 1993 to 2007

22. Anders Ravn Sørensen
Studies in central bank legitimacy, currency and national identity Four cases from Danish monetary history

23. Nina Bellak
Can Language be Managed in International Business? Insights into Language Choice from a Case Study of Danish and Austrian Multinational Corporations (MNCs)

24. Rikke Kristine Nielsen
Global Mindset as Managerial Meta-competence and Organizational Capability: Boundary-crossing Leadership Cooperation in the MNC The Case of 'Group Mindset' in Solar A/S.

25. Rasmus Koss Hartmann
User Innovation inside government Towards a critically performative foundation for inquiry

26. Kristian Gylling Olesen
Flertydig og emergerende ledelse i folkeskolen
Et aktør-netværksteoretisk ledelses-studie af politiske evalueringsreformers betydning for ledelse i den danske folkeskole

27. Troels Riis Larsen
Kampen om Danmarks omdømme 1945-2010
Omdømmearbejde og omdømmepolitik

28. Klaus Majgaard
Jagten på autenticitet i offentlig styring

29. Ming Hua Li
Institutional Transition and Organizational Diversity:
Differentiated internationalization strategies of emerging market state-owned enterprises

30. Sofie Blinkenberg Federspiel
IT, organisation og digitalisering:
Institutionelt arbejde i den kommunale digitaliseringsproces

31. Elvi Weinreich
Hvilke offentlige ledere er der brug for når velfærdstænkningen flytter sig – er Diplomuddannelsens lederprofil svaret?

32. Ellen Mølgaard Korsager
Self-conception and image of context in the growth of the firm – A Penrosian History of Fiberline Composites

33. Else Skjold
The Daily Selection

34. Marie Louise Conradsen
The Cancer Centre That Never Was
The Organisation of Danish Cancer Research 1949-1992

35. Virgilio Failla
Three Essays on the Dynamics of Entrepreneurs in the Labor Market

36. Nicky Nedergaard
Brand-Based Innovation
Relational Perspectives on Brand Logics and Design Innovation Strategies and Implementation

37. Mads Gjedsted Nielsen
Essays in Real Estate Finance

38. Kristin Martina Brandl
Process Perspectives on Service Offshoring

39. Mia Rosa Koss Hartmann
In the gray zone
With police in making space for creativity

40. Karen Ingerslev
Healthcare Innovation under The Microscope
Framing Boundaries of Wicked Problems

41. Tim Neerup Themsen
Risk Management in large Danish public capital investment programmes

2015

1. Jakob Ion Wille
Film som design
Design af levende billeder i film og tv-serier

2. Christiane Mossin
Interzones of Law and Metaphysics
Hierarchies, Logics and Foundations of Social Order seen through the Prism of EU Social Rights

3. Thomas Tøth
TRUSTWORTHINESS: ENABLING GLOBAL COLLABORATION
An Ethnographic Study of Trust, Distance, Control, Culture and Boundary Spanning within Offshore Outsourcing of IT Services

4. Steven Højlund
Evaluation Use in Evaluation Systems – The Case of the European Commission

5. Julia Kirch Kirkegaard
AMBIGUOUS WINDS OF CHANGE – OR FIGHTING AGAINST WINDMILLS IN CHINESE WIND POWER
A CONSTRUCTIVIST INQUIRY INTO CHINA'S PRAGMATICS OF GREEN MARKETISATION MAPPING CONTROVERSIES OVER A POTENTIAL TURN TO QUALITY IN CHINESE WIND POWER

6. Michelle Carol Antero
A Multi-case Analysis of the Development of Enterprise Resource Planning Systems (ERP) Business Practices

Morten Friis-Olivarius
The Associative Nature of Creativity

7. Mathew Abraham
New Cooperativism:
A study of emerging producer organisations in India

8. Stine Hedegaard
Sustainability-Focused Identity: Identity work performed to manage, negotiate and resolve barriers and tensions that arise in the process of constructing or ganizational identity in a sustainability context

9. Cecilie Glerup
Organizing Science in Society – the conduct and justification of resposible research

10. Allan Salling Pedersen
Implementering af ITIL® IT-governance - når best practice konflikter med kulturen Løsning af implementerings-problemer gennem anvendelse af kendte CSF i et aktionsforskningsforløb.

11. Nihat Misir
A Real Options Approach to Determining Power Prices

12. Mamdouh Medhat
MEASURING AND PRICING THE RISK OF CORPORATE FAILURES

13. Rina Hansen
Toward a Digital Strategy for Omnichannel Retailing

14. Eva Pallesen
In the rhythm of welfare creation
A relational processual investigation moving beyond the conceptual horizon of welfare management

15. Gouya Harirchi
In Search of Opportunities: Three Essays on Global Linkages for Innovation

16. Lotte Holck
Embedded Diversity: A critical ethnographic study of the structural tensions of organizing diversity

17. Jose Daniel Balarezo
Learning through Scenario Planning

18. Louise Pram Nielsen
Knowledge dissemination based on terminological ontologies. Using eye tracking to further user interface design.

19. Sofie Dam
PUBLIC-PRIVATE PARTNERSHIPS FOR INNOVATION AND SUSTAINABILITY TRANSFORMATION
An embedded, comparative case study of municipal waste management in England and Denmark

20. Ulrik Hartmyer Christiansen
Follwoing the Content of Reported Risk Across the Organization

21. Guro Refsum Sanden
Language strategies in multinational corporations. A cross-sector study of financial service companies and manufacturing companies.

22. Linn Gevoll
Designing performance management for operational level
- A closer look on the role of design choices in framing coordination and motivation

23. Frederik Larsen
Objects and Social Actions – on Second-hand Valuation Practices

24. Thorhildur Hansdottir Jetzek
The Sustainable Value of Open Government Data
Uncovering the Generative Mechanisms of Open Data through a Mixed Methods Approach

25. Gustav Toppenberg
Innovation-based M&A – Technological-Integration Challenges – The Case of Digital-Technology Companies

26. Mie Plotnikof
Challenges of Collaborative Governance
An Organizational Discourse Study of Public Managers' Struggles with Collaboration across the Daycare Area

27. Christian Garmann Johnsen
Who Are the Post-Bureaucrats? A Philosophical Examination of the Creative Manager, the Authentic Leader and the Entrepreneur

28. Jacob Brogaard-Kay
Constituting Performance Management A field study of a pharmaceutical company

29. Rasmus Ploug Jenle
Engineering Markets for Control: Integrating Wind Power into the Danish Electricity System

30. Morten Lindholst
Complex Business Negotiation: Understanding Preparation and Planning

31. Morten Grynings
TRUST AND TRANSPARENCY FROM AN ALIGNMENT PERSPECTIVE

32. Peter Andreas Norn
Byregimer og styringsevne: Politisk lederskab af store byudviklingsprojekter

33. Milan Miric
Essays on Competition, Innovation and Firm Strategy in Digital Markets

34. Sanne K. Hjordrup
The Value of Talent Management Rethinking practice, problems and possibilities

35. Johanna Sax
Strategic Risk Management – Analyzing Antecedents and Contingencies for Value Creation

36. Pernille Rydén
Strategic Cognition of Social Media

37. Mimmi Sjöklint
The Measurable Me - The Influence of Self-tracking on the User Experience

38. Juan Ignacio Staricco
Towards a Fair Global Economic Regime? A critical assessment of Fair Trade through the examination of the Argentinean wine industry

39. Marie Henriette Madsen
Emerging and temporary connections in Quality work

40. Yangfeng CAO
Toward a Process Framework of Business Model Innovation in the Global Context
Entrepreneurship-Enabled Dynamic Capability of Medium-Sized Multinational Enterprises

41. Carsten Scheibye
Enactment of the Organizational Cost Structure in Value Chain Configuration A Contribution to Strategic Cost Management

2016

1. Signe Sofie Dyrby
Enterprise Social Media at Work

2. Dorte Boesby Dahl
The making of the public parking attendant
Dirt, aesthetics and inclusion in public service work

3. Verena Girschik
Realizing Corporate Responsibility Positioning and Framing in Nascent Institutional Change

4. Anders Ørding Olsen
IN SEARCH OF SOLUTIONS
Inertia, Knowledge Sources and Diversity in Collaborative Problem-solving

5. Pernille Steen Pedersen
Udkast til et nyt copingbegreb
En kvalifikation af ledelsesmuligheder for at forebygge sygefravær ved psykiske problemer.

6. Kerli Kant Hvass
Weaving a Path from Waste to Value: Exploring fashion industry business models and the circular economy

7. Kasper Lindskow
Exploring Digital News Publishing Business Models – a production network approach

8. Mikkel Mouritz Marfelt
The chameleon workforce: Assembling and negotiating the content of a workforce

9. Marianne Bertelsen
Aesthetic encounters
Rethinking autonomy, space & time in today's world of art

10. Louise Hauberg Wilhelmsen
EU PERSPECTIVES ON INTERNATIONAL COMMERCIAL ARBITRATION

11. Abid Hussain
On the Design, Development and Use of the Social Data Analytics Tool (SODATO): Design Propositions, Patterns, and Principles for Big Social Data Analytics

12. Mark Bruun
Essays on Earnings Predictability

13. Tor Bøe-Lillegraven
BUSINESS PARADOXES, BLACK BOXES, AND BIG DATA: BEYOND ORGANIZATIONAL AMBIDEXTERITY

14. Hadis Khonsary-Atighi
ECONOMIC DETERMINANTS OF DOMESTIC INVESTMENT IN AN OIL-BASED ECONOMY: THE CASE OF IRAN (1965-2010)

15. Maj Lervad Grasten
Rule of Law or Rule by Lawyers?
On the Politics of Translation in Global Governance

16. Lene Granzau Juel-Jacobsen
SUPERMARKEDETS MODUS OPERANDI – en hverdagssociologisk undersøgelse af forholdet mellem rum og handlen og understøtte relationsopbygning?

17. Christine Thalsgård Henriques
In search of entrepreneurial learning – Towards a relational perspective on incubating practices?

18. Patrick Bennett
Essays in Education, Crime, and Job Displacement

19. Søren Korsgaard
Payments and Central Bank Policy

20. Marie Kruse Skibsted
Empirical Essays in Economics of Education and Labor

21. Elizabeth Benedict Christensen
The Constantly Contingent Sense of Belonging of the 1.5 Generation Undocumented Youth
An Everyday Perspective

22. Lasse J. Jessen
Essays on Discounting Behavior and Gambling Behavior

23. Kalle Johannes Rose
Når stifterviljen dør…
Et retsøkonomisk bidrag til 200 års juridisk konflikt om ejendomsretten

24. Andreas Søeborg Kirkedal
Danish Stød and Automatic Speech Recognition

25. Ida Lunde Jørgensen
Institutions and Legitimations in Finance for the Arts

26. Olga Rykov Ibsen
An empirical cross-linguistic study of directives: A semiotic approach to the sentence forms chosen by British, Danish and Russian speakers in native and ELF contexts

27. Desi Volker
Understanding Interest Rate Volatility

28. Angeli Elizabeth Weller
Practice at the Boundaries of Business Ethics & Corporate Social Responsibility

29. Ida Danneskiold-Samsøe
Levende læring i kunstneriske organisationer
En undersøgelse af læringsprocesser mellem projekt og organisation på Aarhus Teater

30. Leif Christensen
Quality of information – The role of internal controls and materiality

31. Olga Zarzecka
Tie Content in Professional Networks

32. Henrik Mahncke
De store gaver
- Filantropiens gensidighedsrelationer i teori og praksis

33. Carsten Lund Pedersen
Using the Collective Wisdom of Frontline Employees in Strategic Issue Management

34. Yun Liu
Essays on Market Design

35. Denitsa Hazarbassanova Blagoeva
The Internationalisation of Service Firms

36. Manya Jaura Lind
Capability development in an off-shoring context: How, why and by whom

37. Luis R. Boscán F.
Essays on the Design of Contracts and Markets for Power System Flexibility

38. Andreas Philipp Distel
Capabilities for Strategic Adaptation: Micro-Foundations, Organizational Conditions, and Performance Implications

39. Lavinia Bleoca
The Usefulness of Innovation and Intellectual Capital in Business Performance: The Financial Effects of Knowledge Management vs. Disclosure

40. Henrik Jensen
Economic Organization and Imperfect Managerial Knowledge: A Study of the Role of Managerial Meta-Knowledge in the Management of Distributed Knowledge

41. Stine Mosekjær
The Understanding of English Emotion Words by Chinese and Japanese Speakers of English as a Lingua Franca An Empirical Study

42. Hallur Tor Sigurdarson
The Ministry of Desire - Anxiety and entrepreneurship in a bureaucracy

43. Kätlin Pulk
Making Time While Being in Time
A study of the temporality of organizational processes

44. Valeria Giacomin
Contextualizing the cluster Palm oil in Southeast Asia in global perspective (1880s–1970s)

45. Jeanette Willert
Managers' use of multiple Management Control Systems: The role and interplay of management control systems and company performance

46. Mads Vestergaard Jensen
Financial Frictions: Implications for Early Option Exercise and Realized Volatility

47. Mikael Reimer Jensen
Interbank Markets and Frictions

48. Benjamin Faigen
Essays on Employee Ownership

49. Adela Michea
Enacting Business Models
An Ethnographic Study of an Emerging Business Model Innovation within the Frame of a Manufacturing Company.

50. Iben Sandal Stjerne
Transcending organization in temporary systems
Aesthetics' organizing work and employment in Creative Industries

51. Simon Krogh
Anticipating Organizational Change

52. Sarah Netter
Exploring the Sharing Economy

53. Lene Tolstrup Christensen
State-owned enterprises as institutional market actors in the marketization of public service provision:
A comparative case study of Danish and Swedish passenger rail 1990–2015

54. Kyoung(Kay) Sun Park
Three Essays on Financial Economics

2017

1. Mari Bjerck
Apparel at work. Work uniforms and women in male-dominated manual occupations.

2. Christoph H. Flöthmann
Who Manages Our Supply Chains? Backgrounds, Competencies and Contributions of Human Resources in Supply Chain Management

3. Aleksandra Anna Rzeźnik
Essays in Empirical Asset Pricing

4. Claes Bäckman
Essays on Housing Markets

5. Kirsti Reitan Andersen
Stabilizing Sustainability in the Textile and Fashion Industry

6. Kira Hoffmann
Cost Behavior: An Empirical Analysis of Determinants and Consequences of Asymmetries

7. Tobin Hanspal
Essays in Household Finance

8. Nina Lange
Correlation in Energy Markets

9. Anjum Fayyaz
Donor Interventions and SME Networking in Industrial Clusters in Punjab Province, Pakistan

10. Magnus Paulsen Hansen
Trying the unemployed. Justification and critique, emancipation and coercion towards the 'active society'. A study of contemporary reforms in France and Denmark

11. Sameer Azizi
Corporate Social Responsibility in Afghanistan
– a critical case study of the mobile telecommunications industry

12. Malene Myhre
The internationalization of small and medium-sized enterprises: A qualitative study

13. Thomas Presskorn-Thygesen
The Significance of Normativity – Studies in Post-Kantian Philosophy and Social Theory

14. Federico Clementi
Essays on multinational production and international trade

15. Lara Anne Hale
Experimental Standards in Sustainability Transitions: Insights from the Building Sector

16. Richard Pucci
Accounting for Financial Instruments in an Uncertain World Controversies in IFRS in the Aftermath of the 2008 Financial Crisis

17. Sarah Maria Denta
Kommunale offentlige private partnerskaber Regulering I skyggen af Farumsagen

18. Christian Östlund
Design for e-training

19. Amalie Martinus Hauge
Organizing Valuations – a pragmatic inquiry

20. Tim Holst Celik
Tension-filled Governance? Exploring the Emergence, Consolidation and Reconfiguration of Legitimatory and Fiscal State-crafting

21. Christian Bason
Leading Public Design: How managers engage with design to transform public governance

22. Davide Tomio
Essays on Arbitrage and Market Liquidity

23. Simone Stæhr
Financial Analysts' Forecasts Behavioral Aspects and the Impact of Personal Characteristics

24. Mikkel Godt Gregersen
Management Control, Intrinsic Motivation and Creativity – How Can They Coexist

25. Kristjan Johannes Suse Jespersen
Advancing the Payments for Ecosystem Service Discourse Through Institutional Theory

26. Kristian Bondo Hansen
Crowds and Speculation: A study of crowd phenomena in the U.S. financial markets 1890 to 1940

27. Lars Balslev
Actors and practices – An institutional study on management accounting change in Air Greenland

28. Sven Klingler
Essays on Asset Pricing with Financial Frictions

29. Klement Ahrensbach Rasmussen
Business Model Innovation The Role of Organizational Design

30. Giulio Zichella
Entrepreneurial Cognition. Three essays on entrepreneurial behavior and cognition under risk and uncertainty

31. Richard Ledborg Hansen
En forkærlighed til det eksisterende – mellemlederens oplevelse af forandringsmodstand i organisatoriske forandringer

32. Vilhelm Stefan Holsting
Militært chefvirke: Kritik og retfærdiggørelse mellem politik og profession

33. Thomas Jensen
Shipping Information Pipeline: An information infrastructure to improve international containerized shipping

34. Dzmitry Bartalevich
Do economic theories inform policy? Analysis of the influence of the Chicago School on European Union competition policy

35. Kristian Roed Nielsen
Crowdfunding for Sustainability: A study on the potential of reward-based crowdfunding in supporting sustainable entrepreneurship

36. Emil Husted
There is always an alternative: A study of control and commitment in political organization

37. Anders Ludvig Sevelsted
Interpreting Bonds and Boundaries of Obligation. A genealogy of the emergence and development of Protestant voluntary social work in Denmark as shown through the cases of the Copenhagen Home Mission and the Blue Cross (1850 – 1950)

38. Niklas Kohl
Essays on Stock Issuance

39. Maya Christiane Flensborg Jensen
BOUNDARIES OF PROFESSIONALIZATION AT WORK An ethnography-inspired study of care workers' dilemmas at the margin

40. Andreas Kamstrup
Crowdsourcing and the Architectural Competition as Organisational Technologies

41. Louise Lyngfeldt Gorm Hansen
Triggering Earthquakes in Science, Politics and Chinese Hydropower - A Controversy Study

2018

1. Vishv Priya Kohli
Combatting Falsifi cation and Counterfeiting of Medicinal Products in the E uropean Union – A Legal Analysis

2. Helle Haurum
Customer Engagement Behavior in the context of Continuous Service Relationships

3. Nis Grünberg
The Party -state order: Essays on China's political organization and political economic institutions

4. Jesper Christensen
A Behavioral Theory of Human Capital Integration

5. Poula Marie Helth
Learning in practice

6. Rasmus Vendler Toft-Kehler
Entrepreneurship as a career? An investigation of the relationship between entrepreneurial experience and entrepreneurial outcome

7. Szymon Furtak
Sensing the Future: Designing sensor-based predictive information systems for forecasting spare part demand for diesel engines

8. Mette Brehm Johansen
Organizing patient involvement. An ethnographic study

9. Iwona Sulinska
Complexities of Social Capital in Boards of Directors

10. Cecilie Fanøe Petersen
Award of public contracts as a means to conferring State aid: A legal analysis of the interface between public procurement law and State aid law

11. Ahmad Ahmad Barirani
Three Experimental Studies on Entrepreneurship

12. Carsten Allerslev Olsen
Financial Reporting Enforcement: Impact and Consequences

13. Irene Christensen
New product fumbles – Organizing for the Ramp-up process

14. Jacob Taarup-Esbensen
Managing communities – Mining MNEs' community risk management practices

15. Lester Allan Lasrado
Set-Theoretic approach to maturity models

16. Mia B. Münster
Intention vs. Perception of Designed Atmospheres in Fashion Stores

17. Anne Sluhan
Non-Financial Dimensions of Family Firm Ownership: How Socioemotional Wealth and Familiness Influence Internationalization

18. Henrik Yde Andersen
Essays on Debt and Pensions

19. Fabian Heinrich Müller
Valuation Reversed – When Valuators are Valuated. An Analysis of the Perception of and Reaction to Reviewers in Fine-Dining

20. Martin Jarmatz
Organizing for Pricing

21. Niels Joachim Christfort Gormsen
Essays on Empirical Asset Pricing

22. Diego Zunino
Socio-Cognitive Perspectives in Business Venturing

23. Benjamin Asmussen
Networks and Faces between Copenhagen and Canton, 1730-1840

24. Dalia Bagdziunaite
Brains at Brand Touchpoints A Consumer Neuroscience Study of Information Processing of Brand Advertisements and the Store Environment in Compulsive Buying

25. Erol Kazan
Towards a Disruptive Digital Platform Model

26. Andreas Bang Nielsen
Essays on Foreign Exchange and Credit Risk

27. Anne Krebs
Accountable, Operable Knowledge Toward Value Representations of Individual Knowledge in Accounting

28. Matilde Fogh Kirkegaard
A firm- and demand-side perspective on behavioral strategy for value creation: Insights from the hearing aid industry

29. Agnieszka Nowinska
SHIPS AND RELATION-SHIPS Tie formation in the sector of shipping intermediaries in shipping

30. Stine Evald Bentsen
The Comprehension of English Texts by Native Speakers of English and Japanese, Chinese and Russian Speakers of English as a Lingua Franca. An Empirical Study.

31. Stine Louise Daetz
Essays on Financial Frictions in Lending Markets

32. *Christian Skov Jensen*
Essays on Asset Pricing

33. Anders Kryger
Aligning future employee action and corporate strategy in a resource-scarce environment

34. Maitane Elorriaga-Rubio
The behavioral foundations of strategic decision-making: A contextual perspective

35. Roddy Walker
Leadership Development as Organisational Rehabilitation: Shaping Middle-Managers as Double Agents

36. Jinsun Bae
Producing Garments for Global Markets Corporate social responsibility (CSR) in Myanmar's export garment industry 2011–2015

37. Queralt Prat-i-Pubill
Axiological knowledge in a knowledge driven world. Considerations for organizations.

38. Pia Mølgaard
Essays on Corporate Loans and Credit Risk

39. Marzia Aricò
Service Design as a Transformative Force: Introduction and Adoption in an Organizational Context

40. Christian Dyrlund Wåhlin-Jacobsen
Constructing change initiatives in workplace voice activities Studies from a social interaction perspective

41. Peter Kalum Schou
Institutional Logics in Entrepreneurial Ventures: How Competing Logics arise and shape organizational processes and outcomes during scale-up

42. Per Henriksen
Enterprise Risk Management Rationaler og paradokser i en moderne ledelsesteknologi

43. Maximilian Schellmann
The Politics of Organizing Refugee Camps

44. Jacob Halvas Bjerre
Excluding the Jews: The Aryanization of Danish-German Trade and German Anti-Jewish Policy in Denmark 1937-1943

45. Ida Schrøder
Hybridising accounting and caring: A symmetrical study of how costs and needs are connected in Danish child protection work

46. Katrine Kunst
Electronic Word of Behavior: Transforming digital traces of consumer behaviors into communicative content in product design

47. Viktor Avlonitis
Essays on the role of modularity in management: Towards a unified perspective of modular and integral design

48. Anne Sofie Fischer
Negotiating Spaces of Everyday Politics: -An ethnographic study of organizing for social transformation for women in urban poverty, Delhi, India

2019

1. Shihan Du
   ESSAYS IN EMPIRICAL STUDIES BASED ON ADMINISTRATIVE LABOUR MARKET DATA

2. Mart Laatsit
   Policy learning in innovation policy: A comparative analysis of European Union member states

3. Peter J. Wynne
   Proactively Building Capabilities for the Post-Acquisition Integration of Information Systems

4. Kalina S. Staykova
   Generative Mechanisms for Digital Platform Ecosystem Evolution

5. Ieva Linkeviciute
   Essays on the Demand-Side Management in Electricity Markets

6. Jonatan Echebarria Fernández
   Jurisdiction and Arbitration Agreements in Contracts for the Carriage of Goods by Sea – Limitations on Party Autonomy

7. Louise Thorn Bøttkjær
   Votes for sale. Essays on clientelism in new democracies.

8. Ditte Vilstrup Holm
   The Poetics of Participation: the organizing of participation in contemporary art

9. Philip Rosenbaum
   Essays in Labor Markets – Gender, Fertility and Education

10. Mia Olsen
    Mobile Betalinger - Succesfaktorer og Adfærdsmæssige Konsekvenser

11. Adrián Luis Mérida Gutiérrez
    Entrepreneurial Careers: Determinants, Trajectories, and Outcomes

12. Frederik Regli
    Essays on Crude Oil Tanker Markets

13. Cancan Wang
    Becoming Adaptive through Social Media: Transforming Governance and Organizational Form in Collaborative E-government

14. Lena Lindbjerg Sperling
    Economic and Cultural Development: Empirical Studies of Micro-level Data

15. Xia Zhang
    Obligation, face and facework: An empirical study of the communicative act of cancellation of an obligation by Chinese, Danish and British business professionals in both L1 and ELF contexts

16. Stefan Kirkegaard Sløk-Madsen
    Entrepreneurial Judgment and Commercialization

17. Erin Leitheiser
    The Comparative Dynamics of Private Governance
    The case of the Bangladesh Ready-Made Garment Industry

18. Lone Christensen
    STRATEGIIMPLEMENTERING: STYRINGSBESTRÆBELSER, IDENTITET OG AFFEKT

19. Thomas Kjær Poulsen
    Essays on Asset Pricing with Financial Frictions

20. Maria Lundberg
    Trust and self-trust in leadership identity constructions: A qualitative exploration of narrative ecology in the discursive aftermath of heroic discourse

21. Tina Joanes
Sufficiency for sustainability Determinants and strategies for reducing clothing consumption

22. Benjamin Johannes Flesch
Social Set Visualizer (SoSeVi): Design, Development and Evaluation of a Visual Analytics Tool for Computational Set Analysis of Big Social Data

23. Henriette Sophia Groskopff
Tvede Schleimann
Creating innovation through collaboration – Partnering in the maritime sector

24. Kristian Steensen Nielsen
The Role of Self-Regulation in Environmental Behavior Change

25. Lydia L. Jørgensen
Moving Organizational Atmospheres

26. Theodor Lucian Vladasel
Embracing Heterogeneity: Essays in Entrepreneurship and Human Capital

27. Seidi Suurmets
Contextual Effects in Consumer Research: An Investigation of Consumer Information Processing and Behavior via the Applicati on of Eye-tracking Methodology

28. Marie Sundby Palle Nickelsen
Reformer mellem integritet og innovation: Reform af reformens form i den danske centraladministration fra 1920 til 2019

29. Vibeke Kristine Scheller
The temporal organizing of same-day discharge: A tempography of a Cardiac Day Unit

30. Qian Sun
Adopting Artificial Intelligence in Healthcare in the Digital Age: Perceived Challenges, Frame Incongruence, and Social Power

31. Dorthe Thorning Mejlhede
Artful change agency and organizing for innovation – the case of a Nordic fintech cooperative

32. Benjamin Christoffersen
Corporate Default Models: Empirical Evidence and Methodical Contributions

33. Filipe Antonio Bonito Vieira
Essays on Pensions and Fiscal Sustainability

34. Morten Nicklas Bigler Jensen
Earnings Management in Private Firms: An Empirical Analysis of Determinants and Consequences of Earnings Management in Private Firms

2020

1. Christian Hendriksen
Inside the Blue Box: Explaining industry influence in the International Maritime Organization

2. Vasileios Kosmas
Environmental and social issues in global supply chains: Emission reduction in the maritime transport industry and maritime search and rescue operational response to migration

3. Thorben Peter Simonsen
The spatial organization of psychiatric practice: A situated inquiry into 'healing architecture'

4. Signe Bruskin
The infinite storm: An ethnographic study of organizational change in a bank

5. Rasmus Corlin Christensen
Politics and Professionals: Transnational Struggles to Change International Taxation

6. Robert Lorenz Törmer
The Architectural Enablement of a Digital Platform Strategy

7. Anna Kirkebæk Johansson Gosovic
Ethics as Practice: An ethnographic study of business ethics in a multinational biopharmaceutical company

8. Frank Meier
Making up leaders in leadership development

9. Kai Basner
Servitization at work: On proliferation and containment

10. Anestis Keremis
Anti-corruption in action: How is anti-corruption practiced in multinational companies?

11. Marie Larsen Ryberg
Governing Interdisciolinarity: Stakes and translations of interdisciplinarity in Danish high school education.

12. Jannick Friis Christensen
Queering organisation(s): Norm-critical orientations to organising and researching diversity

13. Thorsteinn Sigurdur Sveinsson
Essays on Macroeconomic Implications of Demographic Change

14. Catherine Casler
Reconstruction in strategy and organization: For a pragmatic stance

15. Luisa Murphy
Revisiting the standard organization of multi-stakeholder initiatives (MSIs): The case of a meta-MSI in Southeast Asia

16. Friedrich Bergmann
Essays on International Trade

17. Nicholas Haagensen
European Legal Networks in Crisis: The Legal Construction of Economic Policy

18. Charlotte Biil
Samskabelse med en sommerfugle-model: Hybrid ret i forbindelse med et partnerskabsprojekt mellem 100 selvejende daginstitutioner, deres paraplyorganisation, tre kommuner og CBS

19. Andreas Dimmelmeier
The Role of Economic Ideas in Sustainable Finance: From Paradigms to Policy

20. Maibrith Kempka Jensen
Ledelse og autoritet i interaktion - En interaktionsbaseret undersøgelse af autoritet i ledelse i praksis

21. Thomas Burø
LAND OF LIGHT: Assembling the Ecology of Culture in Odsherred 2000-2018

22. Prins Marcus Valiant Lantz
Timely Emotion: The Rhetorical Framing of Strategic Decision Making

23. Thorbjørn Vittenhof Fejerskov
Fra værdi til invitationer - offentlig værdiskabelse gennem affekt, potentialitet og begivenhed

24. Lea Acre Foverskov
Demographic Change and Employment: Path dependencies and institutional logics in the European Commission

25. Anirudh Agrawal
A Doctoral Dissertation

26. Julie Marx
Households in the housing market

27. Hadar Gafni
Alternative Digital Methods of Providing Entrepreneurial Finance

28. Mathilde Hjerrild Carlsen
Ledelse af engagementer: En undersøgelse af samarbejde mellem folkeskoler og virksomheder i Danmark

29. Suen Wang
Essays on the Gendered Origins and Implications of Social Policies in the Developing World

30. Stine Hald Larsen
The Story of the Relative: A Systems-Theoretical Analysis of the Role of the Relative in Danish Eldercare Policy from 1930 to 2020

31. Christian Casper Hofma
Immersive technologies and organizational routines: When head-mounted displays meet organizational routines

32. Jonathan Feddersen
The temporal emergence of social relations: An event-based perspective of organising

33. Nageswaran Vaidyanathan
ENRICHING RETAIL CUSTOMER EXPERIENCE USING AUGMENTED REALITY

2021

1. Vanya Rusinova
The Determinants of Firms' Engagement in Corporate Social Responsibility: Evidence from Natural Experiments

2. Lívia Lopes Barakat
Knowledge management mechanisms at MNCs:
The enhancing effect of absorptive capacity and its effects on performance and innovation

3. Søren Bundgaard Brøgger
Essays on Modern Derivatives Markets

4. Martin Friis Nielsen
Consuming Memory: Towards a conceptualization of social media platforms as organizational technologies of consumption

05. Fei Liu
Emergent Technology Use in Consumer Decision Journeys: A Process-as-Propensity Approach

06. Jakob Rømer Barfod
Ledelse i militære højrisikoteams

07. Elham Shafiei Gol
Creative Crowdwork Arrangements

08. Árni Jóhan Petersen
Collective Imaginary as (Residual) Fantasy: A Case Study of the Faroese Oil Bonanza

09. Søren Bering
"Manufacturing, Forward Integration and Governance Strategy"

10. Lars Oehler
Technological Change and the Decomposition of Innovation:
Choices and Consequences for Latecomer Firm Upgrading: The Case of China's Wind Energy Sector

11. Lise Dahl Arvedsen
Leadership in interaction in a virtual context:
A study of the role of leadership processes in a complex context, and how such processes are accomplished in practice

12. Jacob Emil Jeppesen
Essays on Knowledge networks, scientific impact and new knowledge adoption

13. Kasper Ingeman Beck
Essays on Chinese State-Owned Enterprises:
Reform, Corporate Governance and Subnational Diversity

14. Sönnich Dahl Sönnichsen
Exploring the interface between public demand and private supply for implementation of circular economy principles

15. Benjamin Knox
Essays on Financial Markets and Monetary Policy

16. Anita Eskesen
Essays on Utility Regulation: Evaluating Negotiation-Based Approaches inthe Context of Danish Utility Regulation

17. Agnes Guenther
Essays on Firm Strategy and Human Capital

18. Sophie Marie Cappelen
Walking on Eggshells: The balancing act of temporal work in a setting of culinary change

19. Manar Saleh Alnamlah
About Gender Gaps in Entrepreneurial Finance

20. Kirsten Tangaa Nielsen
Essays on the Value of CEOs and Directors

21. Renée Ridgway
Re:search - the Personalised Subject vs. the Anonymous User

22. Codrina Ana Maria Lauth
IMPACT Industrial Hackathons: Findings from a longitudinal case study on short-term vs long-term IMPACT implementations from industrial hackathons within Grundfos

23. Wolf-Hendrik Uhlbach
Scientist Mobility: Essays on knowledge production and innovation

24. Tomaz Sedej
Blockchain technology and inter-organizational relationships

25. Lasse Bundgaard
Public Private Innovation Partnerships: Creating Public Value & Scaling Up Sustainable City Solutions

26. Dimitra Makri Andersen
Walking through Temporal Walls: Rethinking NGO Organizing for Sustainability through a Temporal Lens on NGO-Business Partnerships

27. Louise Fjord Kjærsgaard
Allocation of the Right to Tax Income from Digital Products and Services: A legal analysis of international tax treaty law

28. Sara Dahlman
Marginal alternativity: Organizing for sustainable investing

29. Henrik Gundelach
Performance determinants: An Investigation of the Relationship between Resources, Experience and Performance in Challenging Business Environments

30. Tom Wraight
Confronting the Developmental State: American Trade Policy in the Neoliberal Era

31. Mathias Fjællegaard Jensen
Essays on Gender and Skills in the Labour Market

32. Daniel Lundgaard
Using Social Media to Discuss Global Challenges: Case Studies of the Climate Change Debate on Twitter

33. Jonas Sveistrup Søgaard
Designs for Accounting Information Systems using Distributed Ledger Technology

34. Sarosh Asad
CEO narcissism and board composition: Implications for firm strategy and performance

35. Johann Ole Willers
Experts and Markets in Cybersecurity On Definitional Power and the Organization of Cyber Risks

36. Alexander Kronies
Opportunities and Risks in Alternative Investments

37. Niels Fuglsang
The Politics of Economic Models: An inquiry into the possibilities and limits concerning the rise of macroeconomic forecasting models and what this means for policymaking

38. David Howoldt
Policy Instruments and Policy Mixes for Innovation: Analysing Their Relation to Grand Challenges, Entrepreneurship and Innovation Capability with Natural Language Processing and Latent Variable Methods

2022

01. Ditte Thøgersen
Managing Public Innovation on the Frontline

02. Rasmus Jørgensen
Essays on Empirical Asset Pricing and Private Equity

03. Nicola Giommetti
Essays on Private Equity

04. Laila Starr
When Is Health Innovation Worth It? Essays On New Approaches To Value Creation In Health

05. Maria Krysfeldt Rasmussen
Den transformative ledelsesbyrde – etnografisk studie af en religionsinspireret ledelsesfilosofi i en dansk modevirksomhed

06. Rikke Sejer Nielsen
Mortgage Decisions of Households: Consequences for Consumption and Savings

07. Myriam Noémy Marending
Essays on development challenges of low income countries: Evidence from conflict, pest and credit

08. Selorm Agbleze
A BEHAVIORAL THEORY OF FIRM FORMALIZATION

09. Rasmus Arler Bogetoft
Rettighedshavers faktisk lidte tab i immaterialretssager: Studier af dansk ret med støtte i økonomisk teori og metode

10. Franz Maximilian Buchmann
Driving the Green Transition of the Maritime Industry through Clean Technology Adoption and Environmental Policies

11. Ivan Olav Vulchanov
The role of English as an organisational language in international workplaces

12. Anne Agerbak Bilde
TRANSFORMATIONER AF SKOLELEDELSE - en systemteoretisk analyse af hvordan betingelser for skoleledelse forandres med læring som genstand i perioden 1958-2020

13. JUAN JOSE PRICE ELTON
EFFICIENCY AND PRODUCTIVITY ANALYSIS: TWO EMPIRICAL APPLICATIONS AND A METHODOLOGICAL CONTRIBUTION

14. Catarina Pessanha Gomes
The Art of Occupying: Romanticism as Political Culture in French Prefigurative politics

15. Mark Ørberg
Fondsretten og den levende vedtægt

16. Majbritt Greve
Maersk's Role in Economic Development: A Study of Shipping and Logistics Foreign Direct Investment in Global Trade

17. Sille Julie J. Abildgaard
Doing-Being Creative: Empirical Studies of Interaction in Design Work

18. Jette Sandager
Glitter, Glamour, and the Future of (More) Girls in STEM: Gendered Formations of STEM Aspirations

19. Casper Hein Winther
Inside the innovation lab - How paradoxical tensions persist in ambidextrous organizations over time

20. Nikola Kostić
Collaborative governance of inter-organizational relationships: The effects of management controls, blockchain technology, and industry standards

21. Saila Naomi Stausholm
Maximum capital, minimum tax: Enablers and facilitators of corporate tax minimization

22. Robin Porsfelt
Seeing through Signs: On Economic Imagination and Semiotic Speculation

23. Michael Herburger
Supply chain resilience – a concept for coping with cyber risks

24. Katharina Christiane Nielsen Jeschke
Balancing safety in everyday work - A case study of construction managers' dynamic safety practices

25. Jakob Ahm Sørensen
Financial Markets with Frictions and Belief Distortions

26. Jakob Laage-Thomsen
Nudging Leviathan, Protecting Demos - A Comparative Sociology of Public Administration and Expertise in the Nordics

27. Kathrine Søs Jacobsen Cesko
Collaboration between Economic Operators in the Competition for Public Contracts: A Legal and Economic Analysis of Grey Zones between EU Public Procurement Law and EU Competition Law

28. Mette Nelund
Den nye jord – Et feltstudie af et bæredygtigt virke på Farendløse Mosteri

29. Benjamin Cedric Larsen
Governing Artificial Intelligence – Lessons from the United States and China

30. Anders Brøndum Klein
Kollektiv meningsdannelse iblandt heterogene aktører i eksperimentelle samskabelsesprocesser

31. Stefano Tripodi
Essays on Development Economicis

32. Katrine Maria Lumbye
Internationalization of European Electricity Multinationals in Times of Transition

33. Xiaochun Guo
Dynamic Roles of Digital Currency – An Exploration from Interactive Processes: Difference, Time, and Perspective

34. Louise Lindbjerg
Three Essays on Firm Innovation

35. Marcela Galvis Restrepo
Feature reduction for classification with mixed data: an algorithmic approach

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don't Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens