MOX-Report No. 36/2020

Generalized Mixed-Effects Random Forest: a flexible approach to predict university student dropout

Pellagatti, M.; Masci, C.; Ieva, F.; Paganoni A.M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Generalized Mixed Effects Random Forest: a flexible approach to predict university student dropout

Massimo Pellagatti*
MOX - Department of Mathematics, Politecnico di Milano

Chiara Masci
MOX - Department of Mathematics, Politecnico di Milano

Francesca Ieva
MOX - Department of Mathematics, Politecnico di Milano
and
Anna Maria Paganoni
MOX - Department of Mathematics, Politecnico di Milano

May 6, 2020

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

Abstract

We propose a new statistical method, called Generalized Mixed-Effects Random Forest (GMERF), that extends the use of random forest to the analysis of hierarchical data, for any type of response variable in the exponential family, considering both continuous and discrete covariates and without assuming a closed form in the association between the response and the fixed-effects covariates. At the same time GMERF takes into consideration the nested structure of hierarchical data, modelling the latent grouping structure that exists in the higher level of the hierarchy and allowing statistical inference on this structure. In the case study, we apply GMERF to Higher Education data to analyse the university students dropout; in particular, we are interested in predicting students dropout probability given students-level information and considering the degree program they are enrolled in as the grouping factor.

Keywords: Hierarchical data; Generalized models; Random forest, University students dropout.

1 Introduction

In today's *Big data* era there is often the need of analysing big amounts of complex data. The focus of the analyst is twofold: to reach a good accuracy in the prediction of a given phenomenon and to understand the complexity of the underlying structure that data have; in this sense the analyst has often to find a compromise between the interpretability of the model, high in a simple model, and its accuracy, which usually increases together with the model complexity.

To this purpose tree-based methods were introduced by Breiman et al. (1984) and they are now raising in popularity for how easy they are to fit and to understand; however their high variability is often an issue, resulting in loose predictions (Hastie et al. (2009)). Hence, new methods using trees as building blocks, called tree-based ensemble methods, started being developed to improve the predictive performance of trees (James et al. (2013)). An example of such methods is the Random Forest (RF), described in Breiman (2001), which is a bootstrap aggregation method that combines the predictions of a large number of trees. In recent years, part of the statistical literature focused on extending the use of tree-based methods to the analysis of nested data, i.e. data with a hierarchical structure, embedding them into mixed-effects models (Pinheiro and Bates (2006)). However the development of such methods is still at its beginning. One way in which tree-based methods for nested data are being employed is integrating them with Linear Mixed-effects Models (LMMs), with the aim of solving their low-flexibility issue, due to the parametric assumptions.

LMMs (Pinheiro and Bates (2006)) are used to model situations in which statistical units naturally have a hierarchical structure and this structure is worth to be taken into account for several reasons: (a) nested data are not i.i.d., as classical regression models assume, but their distribution depends on their grouping structure; (b) neglecting the hierarchical structure could result in a loss of a valuable piece of data information; (c) disentangling the effects given to each level of the hierarchy allows to understand and investigate the latent structure that is present in the higher level of the hierarchy and this might be exactly the focus of the interest.

The first step of this integration, called Mixed Effects Regression Tree (MERT), is pre-

sented in Hajjem et al. (2011) and it uses a regression tree to estimate the fixed effects part of a LMM. An analogous approach, but with a different estimation procedure, is presented in Sela and Simonoff (2012) with the name of Random Effects Expectation Maximization tree (RE-EM tree) and it is able to deal with both multilevel data and longitudinal data. With the aim of improving the accuracy in predictions, regression trees are replaced by a RF in the work of Hajjem et al. (2014), where the authors develop a method called Mixed Effects Random Forest (MERF).

However, all such methods deal with a Gaussian response variable and they are not suitable to classification problems. In Hajjem et al. (2017) the MERT approach is extended to non-gaussian data and a Generalized Mixed Effects Regression Tree (GMERT) is proposed. This algorithm is basically the Penalized Quasi Likelihood (PQL) algorithm used to fit Generalized Linear Mixed Models (GLMMs) where the weighted linear mixed effect pseudo-model is replaced by a weighted MERT pseudo-model. Another extension to a classification problem is the Generalized Mixed Effects Tree (GMET), presented in Fontana et al. (2018), which is in line with the approach of Sela and Simonoff (2012) since it uses the tree leaves as indicator variables, rather than using the tree predictions as the MERT approach does. Lastly, the most recent work is proposed in Speiser et al. (2018), where the authors develop a decision tree method for modelling clustered and longitudinal binary outcomes using a Bayesian setting.

In this work, we develop a novel model called Generalized Mixed Effects Random Forest (GMERF), that extends the GMET model for considering a RF instead of a standard tree in the fixed-effects part of the mixed-effects model. This work can then be considered as a further step in the literature consisting in tree-based mixed-effects models as Table 1 illustrates. Following the GMET approach, GMERF is based on a GLMM in which the estimation of fixed effects part is performed with a RF, with the aim of handling interactions among the different covariates and dealing with highly non linear effects. This new method is the first one in the literature able to model hierarchical data with a random forest, which is a flexible and robust method, for a non-gaussian response variable. Indeed GMERF, as all GLMs, is able to deal with different types of responses, as long as their distribution

mixed-effects models	**Regression**	**Classification**
Simple tree	MERT (Hajjem et al. (2017)) RE-EM trees (Sela and Simonoff (2012))	GMERT (Hajjem et al. (2017)) GMET (Fontana et al. (2018))
Random forest	MERF Hajjem et al. (2014)	**GMERF**

Table 1: Tree-based mixed-effects models in the literature

belongs to the exponential family; this is not true for the bayesian approach of Speiser et al. (2018), which works only with binary responses. The strength of this method is that it satisfies the flexibility and the predictive power typical of random forest, maintaining the ability of modelling hierarchical data, for different types of response variable in the exponential family.

After describing the GMERF, providing a pseudo-code for the estimation procedure, we show a simulation study, comparing its performance to other existing methods and then we apply it to a case study. We apply GMERF to a real dataset, that Politecnico di Milano selected for the Student Profile of Enhancing Tutoring Engineering (SPEET) project (`https://www.speet-project.com/`). SPEET is a project aimed at determining and categorizing different profiles of Engineering students across Europe. SPEET consortium is composed by six European universities: Universitat Autnoma de Barcelona (UAB) - Barcelona, Spain; Instituto Politecnico de Braganca (IPB) - Braganca, Portugal; Opole University of Technology - Opole, Poland; Politecnico di Milano (PoliMi) - Milano, Italy; Universidad de Len - Len, Spain; University of Galati *Dunarea de Jos* - Galati, Romania. The essence of SPEET project is to apply data mining algorithms in order to extract information about students and to profile students. A student profile is a set of categories to which a student belongs, that give an insight about how the student is approaching and dealing with his/her studies. Some examples of student profiles are: students that finish degree on time or students that are blocked on a certain set of subjects. Comparisons among different partner institutions will be done in order to establish correlations and get a more complete European-level picture. The role of Politecnico di Milano in the SPEET project is to describe why students leave their studies at the university before accomplish-

ing the degree and to produce a classification method that automatically identifies such students who are likely to drop their studies; from now on we refer to this abandonment as *dropout*. The importance of this task is motivated by the fact that, across all SPEET partners, almost a student out of two leaves his/her Engineering studies before obtaining the BSc degree.

In the last decades, the analysis of university students dropout is receiving particular attention in the educational context. Many studies focus on predicting which are the students at risk in the perspective of identifying the determinants of the dropout and of helping those students (see among the others Goldschmidt and Wang (1999), Chiandotto and Giusti (2005), Barbu et al. (2017), Romero and Ventura (2010)). If it was possible to know as soon as possible to which profile a student belongs, it would be of valuable help for tutors to improve their guiding actions.

We apply GMERF method to Politecnico di Milano data for predicting students dropout probability by means of student-level characteristics and considering the grouping structure of students within engineering degree programs. In our analysis students are the statistical units, which are considered nested based on the degree-program they are enrolled in; as student-level covariates we consider both their performances at Politecnico di Milano (during the first semester of the first year, in the perspective of providing an early warning system) and their collateral data, such as gender or nationality. It turns out that the dropout is correlated much more with the early performances of the student rather than with other student-level variables; also, with the information at our disposal, we are able to predict the dropout in the 90% of cases.

The paper is organised as follows: in Section 2 we present the GMERF method, in Section 3 we perform a simulation study to investigate the strengths and weaknesses of our method, Section 4 reports the case study, i.e. the application of GMERF to Politecnico di Milano data to predict students dropout probability and finally in Section 5 we draw our conclusions.

2 Methods

In this section, after a brief introduction about Generalized Mixed Models (Subsection 2.1), we present the Generalized Mixed Effects Random Forest (GMERF) model with the algorithm for the estimation of its parameters (Subsection 2.2).

2.1 Generalized Mixed Models

We start by considering a generic Generalized Linear Mixed Model (GLMM), described in Pinheiro and Bates (2006). This model is an extension of the Generalized Linear Model (GLM) Nelder and Wedderburn (1972) that includes both fixed and random effects in the linear predictor. Therefore GLMMs handle a wide range of response distributions and a wide range of scenarios where observations have a hierarchical structure which means they are grouped at different levels and so the independence assumption is no more valid.

For a GLMM with a two-level hierarchy, each observation j, for $j = 1, ..., n_i$, is nested within a group i, for $i = 1, ..., I$. Let $\underline{y}_i = (y_{i1}, ..., y_{in_i})$ be the n_i-dimensional response vector for observations in the i-th group. Conditionally on random effects denoted by $\underline{b}_i$, a GLMM assumes that the elements of $\underline{y}_i$ are independent, with density function f_i from the exponential family, of the form

$$f_i(y_{ij}|\underline{b}_i) = \exp\{\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} + c(y_{ij}, \phi)\}$$

where a and c are specified functions, η is the natural parameter and ϕ is the dispersion parameter. In addition, we have

$$E[y_{ij}|\underline{b}_i] = a'(\eta_{ij}) = \mu_{ij}$$
$$Var[y_{ij}|\underline{b}_i] = \psi a''(\eta_{ij})$$

A monotonic, differentiable link function g specifies the function of the mean that the model equates to the systematic component. Usually, the canonical link function is used, i.e., $g = (a')^{-1}$. From now on, without loss of generality the canonical link function is used.

In this case, the model takes the following form:

$$\begin{aligned} \underline{\mu}_i &= E[\underline{y}_i|\underline{b}_i] \qquad i = 1, ..., I \\ g(\underline{\mu}_i) &= \underline{\eta}_i \\ \underline{\eta}_i &= X_i\underline{\beta} + Z_i\underline{b}_i \\ \underline{b}_i &\sim \mathcal{N}_Q(0, \Psi) \end{aligned} \tag{1}$$

where i is the group index, I is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{I} n_i = J$, $\underline{\eta}_i$ is the n_i-dimensional linear predictor vector. In addition, X_i is the $n_i \times P$ matrix of fixed-effects regressors of observations in group i, $\underline{\beta}$ is the P-dimensional vector of their coefficients (including the fixed intercept), Z_i is the $n_i \times Q$ matrix of regressors for the random effects, $\underline{b}_i$ is the Q-dimensional vector of their coefficients (including the random intercept) and Ψ is the $Q \times Q$ within-group covariance matrix of the random effects. Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters.
GLMMs parameters are estimated through Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML), as described in Patterson and Thompson (1971); such estimation methods, for models of this type, do not have closed form solutions; optimal parameters are found numerically, for example with Gaussian Quadrature (GQ) or Penalized Quasi-Likelihood (PQL, Rodríguez (2008)) in order to estimate the integrals to evaluate the likelihood, which is then maximized through an iterative method.

2.2 Generalized Mixed-Effects Random Forest

Our proposed Generalized Mixed-Effects Random Forest (GMERF) embeds the use of tree-based methods for different classes of response variables in the exponential family. At the same time the method can deal with the grouped data structure as GLMMs do.
We basically relax the linear assumptions of the fixed effects of a GLMM and we substitute it with a tree-based structure, allowing the model to be more flexible. The matrix formulation

of the GMERF model is the following:

$$\begin{aligned} \underline{\mu}_i &= E[\underline{y}_i|\underline{b}_i] \qquad i = 1, ..., I \\ g(\underline{\mu}_i) &= \underline{\eta}_i \\ \underline{\eta}_i &= f(X_i) + Z_i\underline{b}_i \\ \underline{b}_i &\sim \mathcal{N}_Q(0, \Psi) \end{aligned} \tag{2}$$

with the same notation of Equation (1).

The fixed part $f(X_i)$ is not forced to be linear any more, but it is assumed to have a tree-structure. In particular, we estimate the effects of this part by means of a Random Forest (RF), which is a tree-based ensemble method (Breiman (2001)). The basic idea of a RF is to train a large number of trees, each one using a different dataset built from the original one by bootstrap and by testing only some of the available covariates; the prediction of the forest is a suitable aggregation of the prediction of each tree.

As in a GLMM, $\underline{b}_i$ and $\underline{b}_{i'}$ are independent for $i \neq i'$. Fixed effects are identified by a non-parametric RF model associated to the entire population, while random ones are identified by group-specific parameters.

To fit this kind of model we have to decouple the estimation of the fixed effects part of the model from the random effects one. To this purpose, we can note that, if random effects were known, the model implies that we could fit a random forest to estimate f using $\eta_{ij} - \underline{Z}_{ij}^T\underline{b}_i$ as dependent variable. Similarly, if the population-level effects f were known, then we could estimate the random effects using a traditional mixed-effects linear model with response corresponding to $\eta_{ij} - f(\underline{X}_{ij})$. Since neither the random effects nor the fixed effects are known, we implement an iterative method that alternates, until convergence, the estimation of the RF with the estimation of the random effects. A second issue that needs to be faced is that $\underline{\eta}_i$ is not known and cannot be directly deduced from data. The solution that we propose, which is in line with the one proposed in Fontana et al. (2018), is estimating it by means of a standard GLM model using as covariates the fixed effects

covariates.

The pseudo-code of the estimation procedure is shown in Algorithm 1.

Algorithm 1 GMERF model estimation procedure

Input:

y- vector with responses y_{ij}
cov- data frame with all covariates
gr- vector with the grouping variable for each observation
$znam$- vector with names of covariates to be used as random effects
$xnam$- vector with names of covariates to be used as fixed effects
fam- distribution of y (must be part of the exponential family)
b_0- optional matrix of initial values for each $\underline{b}_i$
$toll$ threshold to decide whether our estimation converged or not
$itmax$ maximum number of iterations

$Z \leftarrow (1; cov[znam])$ {to include also the random intercept}
Initialize b to a matrix of zero (if b_0 is not given) {Each column $b[i,]$ of b will be the i-th random coefficients $\underline{b}_i$}
$all.b[0] = b$
fit a GLM model using y as response and cov as matrix of covariates
$eta \leftarrow$ estimated η_{ij} by the GLM model
$it \leftarrow 1$
while $it < itmax$ **and not** $conv$ **do**
  $targ \leftarrow eta - Z \times b$
  fit a random forest model using $targ$ as target and cov as predictor matrix
  $fx \leftarrow$ fitted values of the forest model
  fit the GLMM $\eta_{ij} - f(\underline{x}_{ij}) = \underline{z}_{ij}^T \times \underline{b}_i$
  $all.b[it] \leftarrow b \leftarrow$ the estimated b from the model
  $M \leftarrow \max(abs(b - all.b[it-1]))$
  $(i, j) \leftarrow \text{argmax}(abs(b - all.b[it-1]))$
  $tr \leftarrow M/all.b[it-1](i, j)$
  if $tr < toll$ **then**
    $conv \leftarrow$ **true**
  else
    $conv \leftarrow$ **false**
  end if
  $it++$
end while
if not $conv$ **then**
  give a warning
end if

Output:

the final GLMM fitted
the final forest model fitted
b, the final estimation of the random coefficients
it, the number of iterations

The RF is fitted using the R package *randomForest* (Liaw and Wiener (2002)) which implements the original algorithm of (Breiman (2001)). The mixed model is fitted using the function *glmer* from the R package *lme4* (Bates et al. (2011)). To predict a new observation $[\underline{X}_{ij}; \underline{z}_i]$ we use the formula

$$\hat{\eta_{ij}} = \hat{f}(\underline{x}_{ij}) + \underline{z}_{ij}^T \hat{\underline{b}}_i \tag{3}$$

where $\hat{f}$ is the random forest estimated by the algorithm, $\hat{\underline{b}}_i$ is the vector of the random effects coefficients related to the i-th group. Then the prediction on $\hat{\mu}_{ij}$ is obtained by applying to the corresponding $\hat{\eta}_{ij}$ the inverse link function g^{-1}. In the application of GMERF model to binary outcomes we will use the canonical link function *logit*:

$$g(x) = logit(x) = log\left(\frac{x}{1-x}\right).$$

3 Simulation study

In this section we compare the performance of the proposed GMERF method to similar classification methods on different simulated datasets, with the aim of evaluating the strengths and weaknesses of our method.

3.1 Simulation design

Without loss of generality we simulate our response from a Bernoulli distribution, but any distribution from the exponential family could have been used. The Data Generating Process (DGP) of binary data is based on the following equations:

$$\begin{aligned} \eta_{ij} &= f(\underline{X}_{ij}) + \sum_{q=1}^{Q} b_{iq} z_{ijq} \\ \mu_{ij} &= logit^{-1}(\eta_{ij}) \\ y_{ij} &\sim Bernoulli(\mu_{ij}) \end{aligned} \tag{4}$$

where f is the fixed effect part and $\underline{X}_{ij}$ is the $P-$dimensional vector of fixed effects covariates, $\sum_{q=1}^{Q} b_{iq} z_{ijq}$ is the random effect part of the model, which will change in different simulations. As far as the fixed effect part is concerned we choose to have a good (but not too high) number P of covariates and we design f to include both a linear part and a tree-like part, as well as interactions among covariates; in this way we have a very diverse structure that will test the flexibility of our method; so we choose $P = 7$ and we design f like this:

$$f(x_1, ..., x_7) = \alpha(x_1^2 - 3x_2 - x_2 x_3^2) + \beta tree(x_4, x_5, x_6), \tag{5}$$

where α and β are two parameters used to control the variability of f; $tree(x_4, x_5, x_6)$ is a function with a tree-like structure, described in Figure 1. The last variable X_7 by construction has no significance so that we can test if the algorithm is misled by it. The covariates are randomly generated according to the following distributions: X_1, $X_2 \sim U(-1, 1)$, $X_3 \sim Weibull(3)$, $X_4 \sim U(-3, 3)$, $X_5 \sim U(-6, 6)$, $X_6 \sim U(-5, 5)$, $X_7 \sim U(-4, 4)$.

For the random effects part we generate $N = 10$ groups, each one with $n_i = 40$ observations [1] (for a total of 400 units) by sampling from a normal distribution, according to the assumption of the GLMM. For the random-effects generation, we simulate two cases:

- Random intercept: $\sum_{q=1}^{Q} b_q z_{qj} = b_{i0} \sim \mathcal{N}(0, \gamma^2)$ so there is just one scalar random effect; γ regulates the variability of the random effect;
- Random intercept and slope: $\sum_{q=1}^{Q} b_{iq} z_{ijq} = b_{i0} + b_{i1} x_{ij1}$, where x_{1ij} is the first fixed covariate and the random coefficients is $\underline{b} \sim \mathcal{N}_2(0, \Sigma)$, with $\Sigma = diag(\gamma^2; \delta^2)$; b_{0i} and b_{1i} are independent for any value of i; δ is a variance-regulation parameter as well.

Given the presented four parameters to regulate the variability, we select their values so that probability μ_{ij} of each unit is not too close to 0 or 1 (except for a small number of observations). We perform a total of 8 simulation cases in which we change the value of each coefficient to have a low or high variance for the corresponding component of the model; the cases and values of the coefficients are summarized in Table 2.
The models that we test in order to compare their performances with the GMERF's ones

[1] The size does not need to be the same one for all groups.

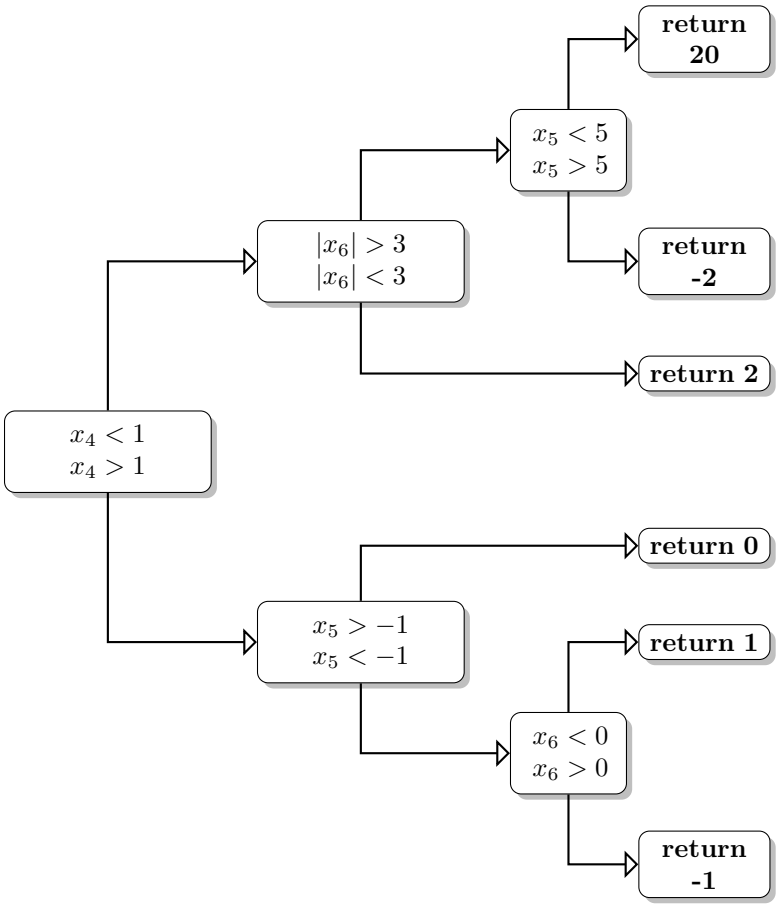


Figure 1: How $tree(x_4, x_5, x_6)$ in Equation (5) is computed.

Random effects	**Fixed effects variability**	α	β	**Random effects variability**	γ^2	δ^2
Intercept only	Small	0.4	0.25	Small	0.5	0
Intercept only	High	0.7	0.6	Small	0.5	0
Intercept only	Small	0.4	0.25	High	2	0
Intercept only	High	0.7	0.6	High	2	0
Intercept and slope	Small	0.4	0.25	Small	0.3	0.5
Intercept and slope	High	0.7	0.6	Small	0.3	0.5
Intercept and slope	Small	0.4	0.25	High	1.4	1.4
Intercept and slope	High	0.7	0.6	High	1.4	1.4

Table 2: Simulation parameters in Equation 5 for the simulation data process.

are: GLM, which can be fit with any version of R, GLMER, which fits the GLMM and is part of the R package *lme4* (Bates et al. (2011)), Random Forest (RF), which can be fitted using the R package *randomForest* (Liaw and Wiener (2002)), GMET, described in Fontana et al. (2018).

3.2 Simulation results

For each of the eight combinations described in Table 2 and each of the five models we simulate 100 times the dataset and analyse thir results, so that we get a good estimation of the performances. In order to have a good estimation of the predictive performances we generate, together with each dataset, also a test set, consisting of 50 observations for each group (500 in total), which will be used for model evaluation.

To evaluate the quality of the predictions we use two indexes: Predictive Mean Absolute Deviation (PMAD) and Predictive MisClassification Rate (PMCR), which are defined as

$$PMAD = \frac{1}{N_{test}} \sum_{i=1}^{I} \sum_{j=1}^{n_i} |\mu_{ij} - \hat{\mu_{ij}}| \tag{6}$$

$$PMCR = \frac{1}{N_{test}} \sum_{i=1}^{I} \sum_{j=1}^{n_i} |y_{ij} - \hat{y_{ij}}|$$

where $N_{test} = 500$ and $n_i = 50 \quad \forall i = 1, ..., 10$; μ_{ij} are the actual probabilities of the simulation generated by the DGP in (4), $\hat{\mu}_{ij}$ are the probabilities predicted by the model, y_{ij} are the actual values of the response and $\hat{y}_{ij}$ are the responses predicted by the model. Note that the RF algorithm does not produce probabilities as output, but just the actual responses, so the PMAD is not available for it.

Results of the simulation just described are shown in Table 3. Overall the prediction of the probabilities is quite good, since the average PMAD is usually a little greather than 0.1, while the misclassified samples are roughly 1 every 5.

First of all we notice that the best mean performances, both on PMAD and on PMCR, are the ones of GLM and GLMER models; GMET model performs always worse than those, while GMERF sometimes is comparable to them, some other times is even worse than GMET; as for RF, it is in two cases the best performing one (in terms of mean), all other times is the worst one. The two cases in which RF is the best one are the the ones with large-variability fixed effects and small-variability random effects; this means that RF is very good at identifying fixed effects, but, when the hierarchical structure becomes relevant, it struggles because on input is given just the label i; this result is quite expected because the effect of the group structure on the response is linear in the simulated dataset and RF is the only algorithm (among the ones we test) in which such effect is not assumed linear.

GMERF algorithm's performances seem to follow by RF ones, since its worst values of PMCR correspond to the cases where also the PMCR of RF is particularly bad; that is also reasonable, since GMERF is built with a RF. That being said, the performances of all algorithms are comparable, expecially in the average PMAD value, which almost never differs more than 0.02 between two different algorithms.

As for the variances of the estimations GMET and GMERF are at the opposites: the former is often the one having the largest variance (expecially for PMAD, where we do not take into account RF), while the latter is the one with the smallest variance (expecially in PMAD, where this happens 6 times out of 8). This is a big upside of the algorithm, which

Model	**Fixed effects variability**	**Random effects variability**	**Algorithm used**		**Empirical mean of PMAD**	**Sample variance of PMAD**	**Empirical mean of PMCR**	**Sample variance of PMCR**
Int	small	small	GLM		0.1129	0.1098	**0.201**	1.2302
Int	small	small	GLMER		**0.1089**	0.0844	0.2039	1.1796
Int	small	small	RF		NA	NA	0.2470	0.8260
Int	small	small	GMET		0.1204	0.1501	0.2199	1.2500
Int	small	small	GMERF		0.1115	0.0765	0.2026	0.8371
Int	large	small	GLM		0.1695	0.1004	0.228	0.7415
Int	large	small	GLMER		**0.1633**	0.0980	0.2177	0.6372
Int	large	small	RF		NA	NA	**0.1966**	0.6936
Int	large	small	GMET		0.1773	0.0758	0.2311	0.6164
Int	large	small	GMERF		0.1721	0.0882	0.2190	0.5244
Int	small	large	GLM		0.0912	0.1754	**0.1196**	1.2777
Int	small	large	GLMER		**0.0907**	0.1478	0.1201	1.1769
Int	small	large	RF		NA	NA	0.2118	2.6451
Int	small	large	GMET		0.1006	0.1820	0.1296	1.5699
Int	small	large	GMERF		0.1038	0.1215	0.1339	1.3794
Int	large	large	GLM		0.1331	0.2087	**0.1595**	0.8296
Int	large	large	GLMER		**0.1364**	0.1689	0.1644	0.6291
Int	large	large	RF		NA	NA	0.2166	0.9118
Int	large	large	GMET		0.1477	0.1967	0.1758	0.8511
Int	large	large	GMERF		0.1560	0.1179	0.1742	0.7673
Int+Slope	small	small	GLM		0.1124	0.1217	0.2019	1.0409
Int+Slope	small	small	GLMER		**0.1106**	0.1206	0.2009	1.1424
Int+Slope	small	small	RF		NA	NA	0.2451	1.2247
Int+Slope	small	small	GMET		0.1219	0.1629	0.2184	1.2822
Int+Slope	small	small	GMERF		0.1121	0.1192	**0.1999**	1.0417
Int+Slope	large	small	GLM		0.1636	0.0763	0.2171	0.4859
Int+Slope	large	small	GLMER		**0.1620**	0.0753	0.2142	0.6465
Int+Slope	large	small	RF		NA	NA	**0.1953**	0.4411
Int+Slope	large	small	GMET		0.1759	0.0539	0.2338	0.4713
Int+Slope	large	small	GMERF		0.1724	0.0681	0.2235	0.6432
Int+Slope	small	large	GLM		0.0897	0.1561	0.1178	1.3974
Int+Slope	small	large	GLMER		**0.0908**	0.1437	**0.1173**	1.3041
Int+Slope	small	large	RF		NA	NA	0.2116	2.4989
Int+Slope	small	large	GMET		0.1012	0.1831	0.1239	1.4359
Int+Slope	small	large	GMERF		0.1056	0.1297	0.1320	1.504
Int+Slope	small	large	GLM		0.1331	0.1969	**0.1581**	0.9307
Int+Slope	small	large	GLMER		**0.1368**	0.2644	0.1645	1.1237
Int+Slope	small	large	RF		NA	NA	0.2162	0.7322
Int+Slope	small	large	GMET		0.1472	0.2967	0.1737	1.3997
Int+Slope	small	large	GMERF		0.1570	0.2059	0.1778	1.239

Table 3: Prediction performances of each model in each of the 8 simulation cases listed in Table 2

proves to be the most robust one; this robustness is probably due to the iterative nature of the algorithm, which stabilizes the estimates.
This justifies our improvement of the algorithm by replacing the tree estimate with a forest.

In conclusion, GMERF algorithm performs comparably to GLMER and GLM, particularly where fixed effects are larger than random effects, but it is more robust than those in the estimates; this can be seen in the same way as ridge regression versus classical linear regression: ridge is biased, but its estimates have lower variance and so in some cases it is preferable to its unbiased alternative.

4 Case study

In this section we present a real life application of our GMERF model, by giving our contribution to the SPEET project in predicting students dropout, as anticipated in Section 1. The aim of this study is to apply GMERF to predict the university student dropout probability considering students information - including demographics, previous studies and the beginning of their academic career - and the engineering degree programs they are enrolled in. Our case study is inspired by the one proposed in Fontana et al. (2018), where the authors apply GMERT to the same dataset: in our application we aim to compare our results with GMET ones.

4.1 The dataset

The data for our analysis comes from Politecnico di Milano database and it consists of 41,098 engineering careers in Bachelor of Science (BSc) that began between A.Y. 2010/2011 and 2015/2016. Politecnico di Milano has $I = 23$ different engineering degree programmes and, in our sample, students are structurally nested within those programs. A descriptive analysis shows that a high percentage of students (27% , more than one out of four) leaves Politecnico di Milano before obtaining the degree. Therefore, our goal is to find out which student-level indicators could discriminate between two different profles: *dropout* and

graduate students. The dataset at our disposal contains a huge amount of information. We select the student-level covariates to include in the model that we think could be more informative as well as the university career information. In this regard, we include in the model only the career information of the first semester of the first year, in the perspective of predicting the student dropout probability as soon as possible. A similar approach has already been used in other articles, such as Goldschmidt and Wang (1999) and Belloc et al. (2010). Standing on the previous literature (Arulampalam et al. (2004)), there are typically three macro-areas of student-level information that result to be significant in student dropout prediction: student collateral data(i.e. general personal information about students who enrolled in the university), student previous studies (i.e. information about the studies of each student before enrolling at the university), student career data (i.e.: everything about the careers of each student in the university, including exams and mobilities). Taking this prior knowledge into account, after some explorative analysis we decided to include in our final dataset the covariates shown in Table 4. Since students are naturally nested in their degree programs, we choose to include in the model a random intercept given to the degree program in which students are enrolled in, in order to take into account this source of dependence among students and to investigate possible differences in the dropout phenomenon across degree programs. Variable *Avg1.1* has a peculiarity, in the sense that it takes values from 18 to 30 (the minimum and maximum score to pass an exam) plus a point mass at 0, representing students who passed no exams at all.

Variable name	**Type of variable**	**Domain**	**Description**
Status	Factor	$\{'0','1'\}$	The response variable: did the career end with a degree or with a dropout? ('1'=Dropout, '0'=Graduate)
Sex	Factor	$\{'M','F'\}$	Gender of the student
Nationality	Factor	$\{'I','F'\}$	Nationality of the student ('I'=Italian, 'F'=Foreign)
Previous studies	Factor	4 levels	Type of studies before university: $'Scientifica'$, $'Classica'$, $'Tecnica'$, $'Altro'$
Avg 1.1	Numeric	$\{0\} \cup [18; 30]$	Weighted average score obtained in exams of the first semester of the first year
Attempts 1.1	Numeric	$[0; 10]$	Average number of attempts per exam of the first semester of the first year
Credits 1.1	Integer	$\{1, ..., 40\}$	Total number of Crediti Formativi Universitari (CFU) obtained by the student after the first semester of the first year.
Degree program	Factor	23 levels	Degree program the student is enrolled in (it is the grouping variable)

Table 4: Covariates used for the analysis of SPEET data with GMERF model

We excluded from the dataset four degree programs having few students enrolled (less than 200), so the final number of degree programs considered is $I = 19$. The statistical units are the concluded (either graduated or dropout) careers of students in the university enrolled in the degree programs listed above. The final dataset has 24,736 statistical units nested in 19 degree programs and it is the same one which is used in Fontana et al. (2018), with the only difference that variable *Previous studies* here has 4 levels, while in Fontana et al. (2018) it has just 3 levels ('Classica' and 'Altro' are considered together); this has a minor impact on the final results of the analysis, so a comparison between the two of them is still possible.

We randomly split the dataset into training and test subsets, with a ratio of 70% for model fitting and 30% for evaluation (which we will refer to as test set). We then split again the model fitting set into a training set and a validation set using a proportion of, respectively, 80% and 20%; validation set will be used to select the best value of the treshold α for the prediction

4.2 Model results

The model we implement has, as random effect, just the intercept b_0. We apply our model using $toll = 0.02$ and $itmax = 30$; it converges after 8 iterations, so it reaches stability in a short time. Estimates of random intercepts together with their confidence intervals are shown in Figure 2. We can see that 10 intercepts are not significantly different from 0 (with 95% confidence), being in line with the average. Five programs increase the log-odds of dropout, while four programs decrease those. The Variance Partition Coefficient (VPC) is a possible measure of intraclass correlation introduced in Goldstein et al. (2002); it is equal to the percentage of variation that is found at the higher level of a hierarchical model over the total variance. It is defined as

$$VPC = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{lat}^2} \tag{7}$$

where σ_m^2 is the estimated variance of random effects, while σ_{lat}^2 is the residual variability that can neither be explained by fixed effects, nor through the group features that are

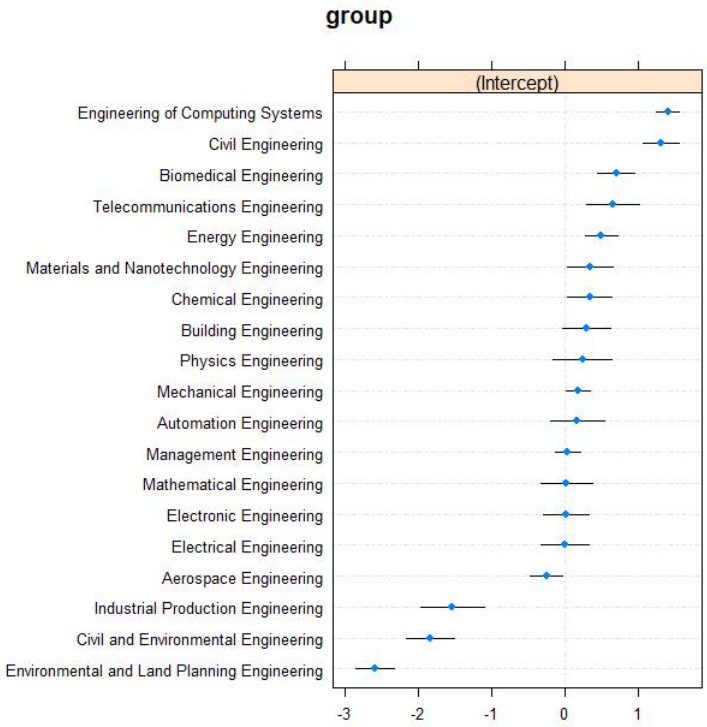


Figure 2: Random intercepts of the GMERF model with their confidence interval.

represented by the random intercept. Since the variance of the standard logistic distribution is $\pi^2/3 \simeq 3.29$, the VPC can be estimated as:

$$VPC = \frac{\sigma_m^2}{\sigma_m^2 + \pi^2/3} = 0.2261.$$

This means that roughly 23% of unexplained variation in the response is attributable to the nested structure; this is a good indicator of the significance of the hierarchical structure. Regarding the fixed-effects part, RF model gives us the importance of each covariate (measured as the increase of the Residual Sum of Squares (RSS) when the values of the corresponding variable are randomly permuted in the training dataset) in explaining the response and the partial effect of each covariate (that can be displayed using partial dependence plots). To this purpose we can look at Figure 3. Our GMERF model considers as most important variables *Avg 1.1* and *Credits 1.1*. In particular the three covariates associated with performance of the student during his/her first semester career are all more important than the collateral information; this suggests that what influences the choice of

leaving the studies is the university performance of the student, more than its background when enrolling.

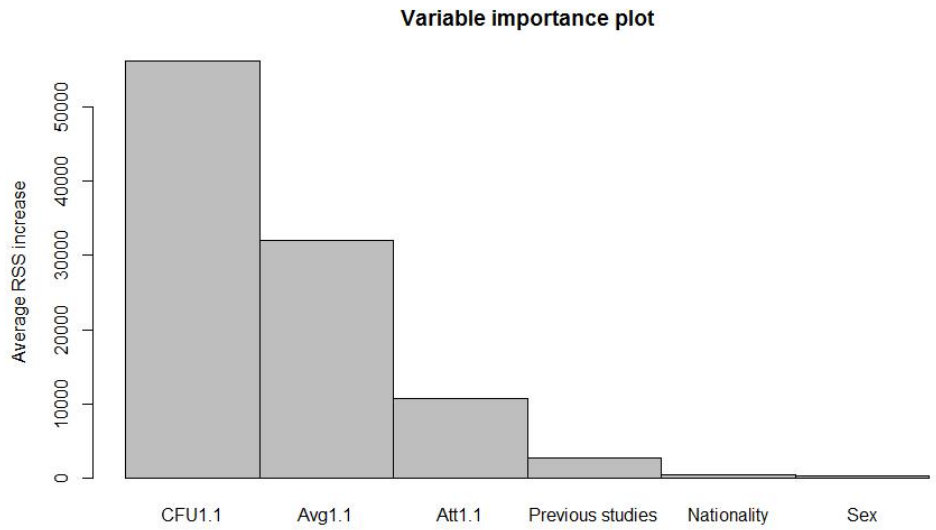


Figure 3: Plot of GMERF's variable importance; the height of the bar is the increase of the Residual Sum of Squares (RSS) when the values of the corresponding variable are randomly permuted.

Using partial plots we can highlight the effect of each variable with respect to the response; in Figures 4 and 5 the partial plot for the most important fixed effects covariates are shown; in particular, for variable *Avg1.1*, we show two different plots, Figures 4a and 4b: the former shows the plot with respect to all values of *Avg1.1*, while the latter focuses just on the values from 18 to 30; the jump after 0 in the first one is motivated by the fact that there are no values of this variable in the interval $(0; 18)$.

Looking at Figures 4a and 4d we can see an inverse proportional association between the probability of dropout and the variables *Avg1.1* and *Attempts1.1*, suggesting us that students trying less exams and not passing them at the first semester tend to drop their studies. This pattern repeats in Figure 4c, even if not in the same straightforward way; from this figure we can also note that students who obtain 30 credits after the first semester (which means that the student passes all exams of that semester) has almost null probability of dropout; this strongly suggests that a student likely to dropout can be identified already

after a semester of studies. Finally, Figure 4b shows that the probability of dropout decreases linearly with variable *Avg1.1*.
Regarding the previous studies, Figure 5 shows that there is not a significant difference in the dropout probability of students who attended scientific, classic or other schools (after adjusting for the other characteristics), while students who attended technical schools are more likely to dropout.

GMERF model returns also the probability that a student drops his/her studies. To evaluate the quality of the predictions we use four indexes: Accuracy A, that is the percentage of correctly classified units; Sensibility SN that is, out of all the positive units, the proportion of those found by the algorithm; Specificity SP that is, out of all the positive-predicted units, the percentage of those who actually are; F1-measure, which combines Sensitivity and Specificity as

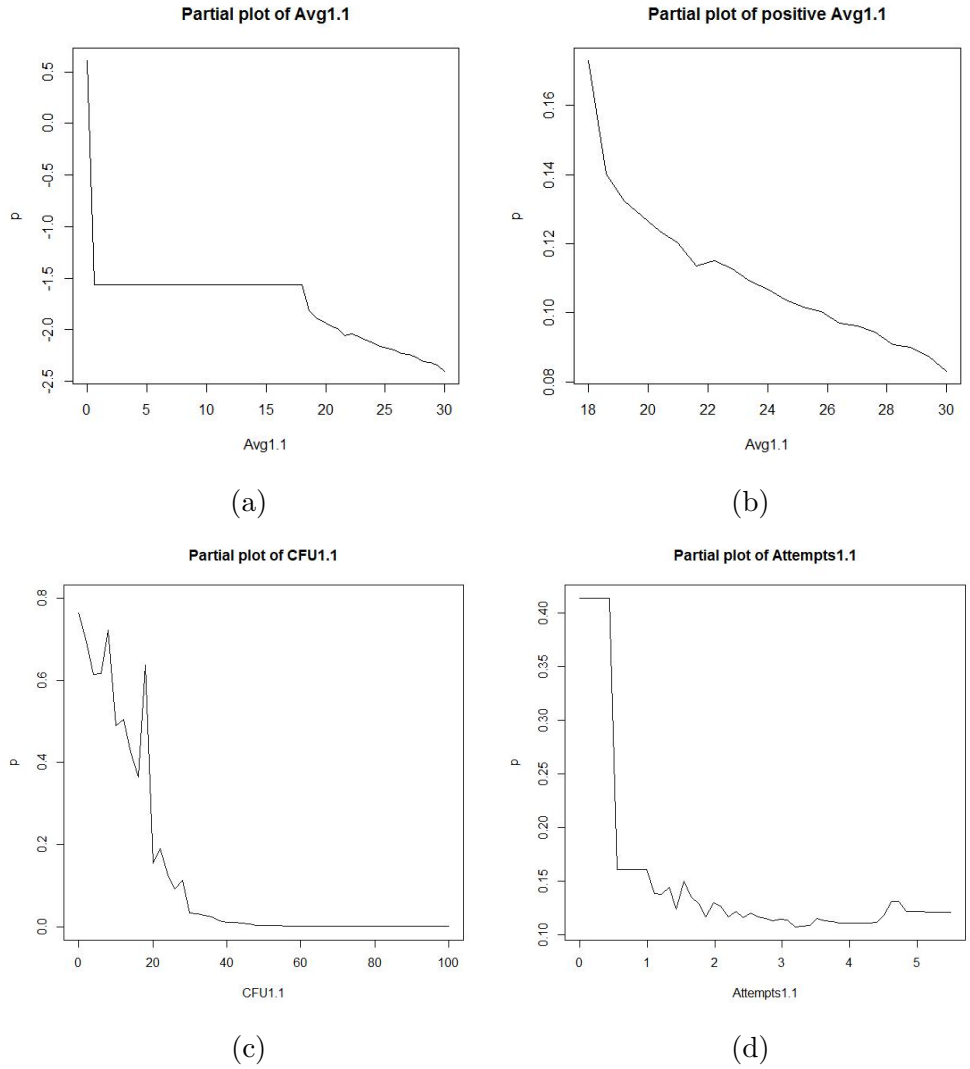
$$F1 = \frac{2 \cdot SN \cdot SP}{SN + SP}. \tag{8}$$

We use the validation set to choose the optimal treshold α for prediction, by looking at the prediction accuracy and at the ROC curve (we denote Specificity with SP and Sensitivity with SN) that we build with this set (Agresti and Kateri (2011)). In Figure 6 the complete ROC curve Sensitivity-Specificity is shown. The optimal value turns out to be $\alpha = 0.4$, both in terms of Accuracy ($A = 0.9082$) and F1-measure ($F1 = 0.8305$); the other indexes values are $SN = 0.8102$ and $SP = 0.8495$, while misclassification table relative to this value is

	$y = 0$	$y = 1$
$\hat{y} = 0$	2366	180
$\hat{y} = 1$	138	779

Overall these results show that the model is very good at predicting our validation set; we can then apply these results to the test set.
The contingency table of the predictions on the test set is

	$y = 0$	$y = 1$
$\hat{y} = 0$	5138	406
$\hat{y} = 1$	273	1603



(a) (b)

(c) (d)

Figure 4: Partial plots of dropout probability with respect to continuous variables: variable *Avg1.1* on the entire range in panel (a), variable *Avg1.1* on the range $(18; 30)$ in panel (b); variable *Cfu1.1* in panel (c) and variable *Attempts1.1* in panel (d).

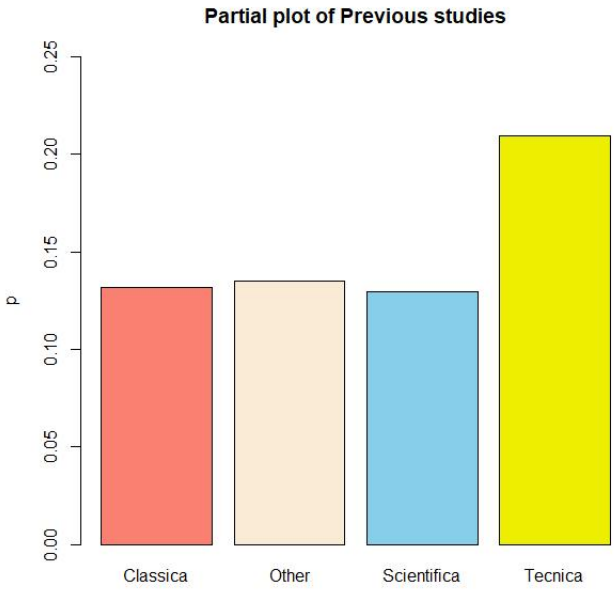


Figure 5: Partial plot of the student dropout probability with respect to variable *PrevStudies*

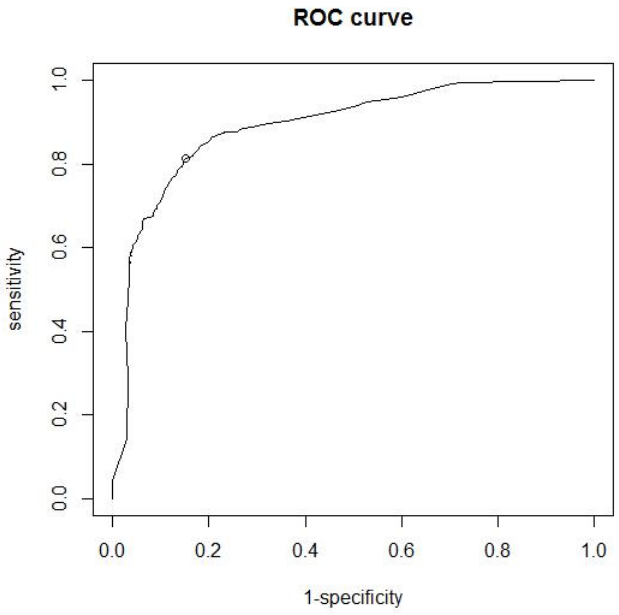


Figure 6: ROC curve obtained from the validation set; the point highlighted is the one corresponding to the optimal value of α found with the validation process.

and the value of the indexes are:

$$A = 0.9085$$
$$SP = 0.8544$$
$$SN = 0.7979$$
$$F1 = 0.8252.$$

So, overall, our model gives the right prediction 91% of times, 80% of students who will drop their studies are correctly identified and 85% of students predicted as *dropout* actually are; these are very good values, especially the indices SN and SP, because they show that our model is very accurate, but at the same time is not too sensitive in predicting students to drop their studies.

We can now compare our results with the ones found in Fontana et al. (2018) by using the GMET model on the same dataset. Both models identified in variables *CFU1.1* and *Avg1.1* the two most important variables to predict a dropout; on the other end variable *Sex* is not considered significant by either of them. As far as random effects are concerned, both models identify *Environmental* and *Land planning* engineering as the ones associated with the lowest dropout rate and they also both associate *Computer* and *Civil* engineering with a higher dropout probability. The major differences between the fixed effects estimates by GMET and GMERF are the following:

- Variable *Attempts 1.1* is considered important by GMERF, but it does not appear in GMET as splitting node; this may happen because the effect of this variable is masked, in GMET, by the first split based on variable *CFU 1.1*; the two variables, at least for very small values, are naturally correlated (people attempting no exams do not pass exams and therefore do not get any CFU); however, the random forest used in GMERF uses different variables in different trees and is then able to identify the effect of both variables, which is one of the main advantages of a RF over a classification tree;

- *Nationality* is considered very important by GMET, being the second split, while in GMERF it has almost null importance.

As for the estimation of random effects, the major differences between the estimations of

the two models (GMET and GMERF) are three:

- *Management Engineering*, which our model considers in line with the average, is estimated by GMET to decrease the dropout odds;
- *Biomedical Engineering* and *Telecommunications Engineering*, which in our model are associated with a positive coefficient, in GMET they are associated with a null coefficient.

Finally, comparing the predictive power of the two models on the test set, we can see that GMERF brought a slight improvement to the accuracy, from GMET's 0.878 to 0.908; therefore 3% more of the students are correctly classified, which confirms our initial expectation. Overall we can say that the two models highlighted similar dynamics, which is an evidence on the robustness of the two of them; the major difference is the higher precision with which our GMERF model classified students and showed the effects of each covariate on the dropout probability.

5 Conclusion

In this work, we present a model called Generalized Mixed Effects Random Forest (GMERF), which consists in a novel method that extends the use of random forest to the analysis of hierarchical data, for a non-gaussian response variable. GMERF modelling substitutes the linear combination of the fixed-effects covariates of a GLMM with a random forest. This new method contributes to the statistical literature about mixed-effects models and tree-based method, taking advantage of the flexibility and the predictive power of a random forest, but maintaining the structure of mixed-effects models. Moreover, although our study focuses on the binary response case, this approach can handle any type of response variable in the exponential family. Using suitable link functions, we can model different outcomes such as counts data, as well as the particular case of a Gaussian response. GMERF can be considered the missing piece of a class of models which combine tree-based methods with Linear Mixed Models. The simulation study shows that GMERF has prediction performances comparable to models like GLM and GLMM, with the advantage that its estimates

are less variable then the ones of these models; moreover it has the added benefit of having no assumption or structure for the fixed effect part; finally, as the RF algorithm, it can deal with heterogeneous covariates (discrete and continuous) at the same time, which is a very big advantage in terms of flexibility. In the case study we give a contribution to the SPEET project, by providing a robust method to classify students as *dropout* or *graduate* that is successful in the 90% of cases. These results might be useful in the perspective of defining new tutoring systems to help students at risk. Our study results in an improvement in the prediction accuracy over the GMET model, which was applied on the same dataset; this is one of the goals we expect to achieve when using GMERF, since the two models have the same formulation, but GMERF uses a RF to estimate the fixed effects, which is an algorithms that improves the regression tree used in GMET. When applied to a complex real data problem, GMERF proves to be a powerful and an easily interpretable method.

References

Agresti, A. and M. Kateri (2011). *Categorical data analysis*. Springer.

Arulampalam, W., R. Naylor, and J. Smith (2004). Factors affecting the probability of first year medical student dropout in the uk: a logistic analysis for the intake cohorts of 1980–92. *Medical Education 38*(5), 492–503.

Barbu, M., R. Vilanova, J. Lopez Vicario, M. J. Pereira, P. Alves, M. Podpora, M. Ángel Prada, A. Morán, A. Torreburno, S. Marin, et al. (2017). Data mining tool for academic data exploitation: literature review and first architecture proposal. *Projecto SPEET-Student Profile for Enhancing Engineering Tutoring*.

Bates, D., M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, and G. Grothendieck (2011). Package lme4. *Linear mixed-effects models using S4 classes. R package version*, 1–1.

Belloc, F., A. Maruotti, and L. Petrella (2010). University drop-out: an italian experience. *Higher education 60*(2), 127–138.

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

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Chiandotto, B. and C. Giusti (2005). Labbandono degli studi universitari. *Modelli statistici per lanalisi della transizione università-lavoro*, 1–22.

Fontana, L., C. Masci, F. Ieva, and A. M. Paganoni (2018). Performing learning analytics via generalized mixed-effects trees. *Mox-report*.

Goldschmidt, P. and J. Wang (1999). When can schools affect dropout behavior? a longitudinal multilevel analysis. *American Educational Research Journal 36*(4), 715–738.

Goldstein, H., W. Browne, and J. Rasbash (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences 1*(4), 223–231.

Hajjem, A., F. Bellavance, and D. Larocque (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters 81*(4), 451–459.

Hajjem, A., F. Bellavance, and D. Larocque (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation 84*(6), 1313–1328.

Hajjem, A., D. Larocque, and F. Bellavance (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters 126*, 114–118.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*(3), 18–22.

Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General) 135*(3), 370–384.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Pinheiro, J. and D. Bates (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Rodríguez, G. (2008). Multilevel generalized linear models. In *Handbook of multilevel analysis*, pp. 335–376. Springer.

Romero, C. and S. Ventura (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40*(6), 601–618.

Sela, R. J. and J. S. Simonoff (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning 86*(2), 169–207.

Speiser, J. L., B. J. Wolf, D. Chung, C. J. Karvellas, D. G. Koch, and V. L. Durkalski (2018). Bimm tree: a decision tree method for modeling clustered and longitudinal binary outcomes. *Communications in Statistics-Simulation and Computation*, 1–20.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

35/2020 Morbiducci, U.; Mazzi, V.; Domanin, M.; De Nisco, G.; Vergara, C.; Steinman, D.A.; Gallo, D
Wall shear stress topological skeleton independently predicts long-term restenosis after carotid bifurcation endarterectomy

34/2020 Antonietti, P.F.; Mazzieri, I.; Nati Poltri, S.
A high-order discontinuous Galerkin method for the poro-elasto-acoustic problem on polygonal and polyhedral grids

33/2020 Centofanti, F.; Fontana, M.; Lepore, A.; Vantini, S.
Smooth LASSO Estimator for the Function-on-Function Linear Regression Model

32/2020 Menafoglio, A.; Sgobba, S.; Lanzano, G.; Pacor, F.
Simulation of seismic ground motion fields via object-oriented spatial statistics: a case study in Northern Italy

31/2020 Bernardi, M.S.; Africa, P.C.; de Falco, C.; Formaggia, L.; Menafoglio, A.; Vantini, S.
On the Use of Interfeometric Synthetic Aperture Radar Data for Monitoring and Forecasting Natural Hazards

30/2020 Massi, M.C., Gasperoni, F., Ieva, F., Paganoni, A.M., Zunino, P., Manzoni, A., Franco, N.R., e
A deep learning approach validates genetic risk factors for late toxicity after prostate cancer radiotherapy in a REQUITE multinational cohort

29/2020 Piersanti, R.; Africa, P.C.; Fedele, M.; Vergara, C.; Dede', L.; Corno, A.F.; Quarteroni, A.
Modeling cardiac muscle fibers in ventricular and atrial electrophysiology simulations

26/2020 Zonca, S.; Antonietti, P.F.; Vergara, C.
A Polygonal Discontinuous Galerkin formulation for contact mechanics in fluid-structure interaction problems

28/2020 Regazzoni, F.; Dedè, L.; Quarteroni, A.
Biophysically detailed mathematical models of multiscale cardiac active mechanics

27/2020 Spreafico, M.; Ieva, F.; Fiocco, M.
Modelling dynamic covariates effect on survival via Functional Data Analysis: application to the MRC BO06 trial in osteosarcoma