ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

Dropout Prediction at University of Genoa: a Privacy Preserving Data Driven Approach

Luca Oneto[1]*, Anna Siri[2], Gianvittorio Luria[2], and Davide Anguita[1]

1 - DIBRIS - University of Genova
Via Opera Pia 13, I-16145 Genova - Italy

2 - DIMA - University of Genova
Via Dodecaneso 35, I-16146 Genova - Italy

Abstract. Nowadays many educational institutions crucially need to understand the dynamics at the basis of the university dropout (UD) phenomenon. However, the most informative educational data are personal and subject to strict privacy constraints. The challenge is therefore to develop a data driven system which accurately predicts students dropouts while preserving the privacy of individual data instances. In the present paper we investigate this problem, making use of data collected at University of Genoa as a case study.

1 Introduction

According to [1], by 2025, most job opportunities in the EU will require a high-level qualification.

In 2015, the average unemployment rate across OECD countries was 4.9% for people aged 25 to 64 with tertiary education, compared to 7.3% for adults with upper secondary or post–secondary non–tertiary education and 12.4% for adults with no upper secondary education [2].

In the area of tertiary education, the “Europe 2020 Strategy” has set, as headline target, the achievement of a tertiary or equivalent qualification by 2020 by at least the 40% of the population aged 30 to 34 [3]. The EU target is reflected into individual national goals which range from 26% of the population for Italy to 66% for Luxembourg. In 2015, 13 countries have already attained their own national target, while Italy, Poland and Romania were less than 2% away from its achievement [4].

In order to make higher education more effective in providing highly qualified graduates in line with the needs of the labour market, and thus increasing the efficiency of public investments, EU states have identified three main challenges: (i) broadening participation to higher education, (ii) improving the quality of teaching and learning, and (iii) reducing dropout rates and the time taken in completing a degree. Regarding this last point, 41% of students who enter a tertiary or equivalent programme graduate in time, while 69% of them take up to three times the standard minimum duration to successfully complete their studies [2]. The need to face up the tertiary educational system failure is therefore crucial.

*This work was supported by the Genoa Athenaeum through the project Drop@UNIGE (Machine Learning & Data Mining for Education: Students Dropout Prediction at University of Genoa) with the grants SEED 2015

Two main cultures to reach conclusions from data in education exist. Historically, the first one consists in applying theories borrowed from disciplines such as psychology, sociology, economics, and organization. Researchers usually suggest theories, and models closely related to them, as being made up of variables definition, a domain, a set of relationships between the factors and predictors. However, this kind of approach imposes a sort of straitjacket that limits the deep understanding of the complexity of the educational problems under study.

To overcome this issue, in the last decade, data–driven approaches making use of minimal prior knowledge about the problem have raised more and more interest.

The research field concerned with exploiting sophisticated data driven techniques and advanced statistics for discovering patterns and automatically extracting or predicting trends from highly complex educational datasets is called *Educational Data Mining* (EDM) [5, 6]. EDM makes use of standard data mining techniques such as Artificial Neural Networks, Decision Trees, Support Vector Machines, Random Forests (RF), etc [7].

In [8], Baker and Yacef identified the following main goals of EDM: (i) predicting students' future behavior, (ii) discovering or improving already existing models, (iii) investigating the effects of educational support and advice/counseling that can be achieved through new learning systems, and (iv) advancing scientific knowledge about students.

Due to its proven effectiveness, EDM has become very popular. As a consequence, the public has become aware of how much data is being collected about students and started to concern about students' privacy. Also EU is changing its policy in data protection with respect to two main issues: (i) protecting the disaggregated source by itself, eventually thinking how aggregated data could be exploited instead and (ii) protecting the data used for the model. Unfortunately, none of the above studies considered this fact and we will focus on the second issue.

In this work we investigate the problem of predicting students dropout in a privacy–preserving framework. In particular, we make use of the University of Genoa (UNIGE) as a case study of our research.

2 A Privacy Preserving Data Driven Approach

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines [9, 10, 11]. One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context [11]. DP addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population. In particular, a procedure is called *differentially private* if and only if its output is almost independent from any of the individual observations. In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken over the noise introduced by the procedure. DP allowed to reach a milestone result by connecting the field of

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1. Available from http://www.i6doc.com/en/.

privacy preserving data analysis and the generalization capability of a learning algorithm [12].

Let us consider the multiclass classification problem where we have an input space $\mathcal{X}$ and an output space $\mathcal{Y} = \{1, \cdots, c\}$. From $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ we observe a series of n i.i.d. samples $s = \{z_1, \cdots, z_n\}$ distributed according to μ. We denote with $\dot{s}$ the neighborhood dataset of s such that $\dot{s} = \{z_1, \cdots, z_{i-1}, \dot{z}_i, z_{i+1}, \cdots, z_n\}$ where i may assume any value in $\{1, \cdots, n\}$ and $\dot{z}_i$ i.i.d. with z_i. Let us define with $f : \mathcal{X} \to \mathcal{Y}$ a function in a space $\mathcal{F}$ of all the possible functions. A randomized algorithm $\mathscr{A}$ maps a dataset s in a function $f \in \mathcal{F}$ with a non-deterministic rule. The accuracy of $f \in \mathcal{F}$ in representing μ is measured with reference to the hard loss function $\ell : \mathcal{F} \times \mathcal{Z} \to \{0, 1\}$ which counts the number of misclassified examples. Hence, we can define the true risk of f, namely generalization error, as $L(f) = \mathbb{E}_z \ell(f, z)$. Since μ is unknown, $L(f)$ cannot be computed. Therefore, we have to resort to its empirical estimators, the empirical error $\widehat{L}^s_n(f) = {}^1\!/_n \sum_{i=1}^{n} \ell(f, z_i)$. Let us recall the definition of DP.

Definitions 1 ([11])**.** *$\mathscr{A}$ is ϵ–differentially private (ϵ–DP) if $\forall f, s$ we have that $\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) = f\} \leq e^{\epsilon} \mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) = f\}$.*

The milestone result of [12] is that an ϵ–DP algorithm generalizes. In particular, it is possible to show that the empirical error of a function chosen with an ϵ–DP algorithm is concentrated around its generalization error.

Theorem 1 ([12])**.** *Let $\mathscr{A}$ be an ϵ-DP, then $\mathbb{P}_{s,f=\mathscr{A}(s)}\{L(f) \geq \widehat{L}^s_n(f) + \epsilon\} \leq 3e^{-n\epsilon^2}$.*

Note that it is not possible to set the parameter ϵ and the confidence of the statement of Theorem 1 independently. In other words, fixing the confidence of the statement fixes also the accuracy in estimating the generalization error and the required privacy of the algorithm.

Corollary 1. *Let $\mathscr{A}$ be an ϵ–DP. Let us suppose that for $\mathscr{A}$ it is possible to set $\epsilon = \sqrt{\ln(3/\delta)/n}$, then $\mathbb{P}_{s,f=\mathscr{A}(s)}\{L(f) \geq \widehat{L}^s_n(f) + \sqrt{\ln(3/\delta)/n}\} \leq \delta$.*

Consequently, in order to be able to perform a privacy preserving analysis of the data we need to use a ϵ–DP learning algorithm where ϵ can be set by the user.

In the last years, many state-of-the-art learning algorithms, both in the supervised setting [13] and in the unsupervised one [14], have been privatized.

In the non-private setting it is well known that combining the output of several classifiers often results in a much better performance than using any one of them alone. In [15] Breiman proposed the RF of tree classifiers, one of the state-of-the-art algorithms for classification, and recently improved in [16], which has shown to be one of the most effective tools in this context [17].

In the last years, many ϵ–DP versions of the RF have been developed [18, 19, 20] by adding noise to the leaf nodes of each tree whose magnitude is scaled up with the number of trees in the ensemble. This results in high noise in individual trees. Therefore, the utility of such ensembles remains poor. For this reason in [21] a new ϵ–DP RF has been developed which proposes a new noise injection method which produces less noisy trees and results in better final performances.

We shall now compare the performances of the state-of-the-art private [21] and non-private [15, 16] RF on the problem of predicting UD at the UNIGE. The purpose is to understand how much the privacy constraints affect our ability of building an effective data driven model.

3 Case Study: the University of Genoa

Founded in 1481, UNIGE is a large University in Italy which counts approximately 40000 students.

In the last three years, UNIGE has been suffering from a high dropout rate of around 15% (with peaks of 30% in social and humanistic degree courses). A large number of students leaving university, with particular emphasis to dropouts at the beginning of the career, is interpreted by the Italian Ministry of Education as a clear warning signal of a possible malfunctioning in the educational system. Universities are highly encouraged to invest money in order to stem UD, also in the light of the fact that the share of students who do not enroll in the second year is one of the indicators determining their evaluation and their funding.

At UNIGE, all kind of information regarding students socio-demographic variables, previous school performance, university achievements and other factors affecting students careers is collected and controlled by the IT department. All this data are protected by the Italian privacy law and can be published only in an aggregate form or with privacy guarantees.

In this study we take into consideration students who enrolled, in the academic year (a.y.) 2008/2009, in the first year of a healthcare degree program. These students accepted to provide their personal information stored in the UNIGE databases, provided that no data regarding a single student is made public.

The dataset contains 810 instances, each one described by 49 attributes (both numeric and categorical) about ethnicity, gender, financial status, previous school experience, and quality of taught courses[1].

Instances have been divided into three classes, based on academic achievements. The identification of the three groups is based on the ministerial definition of regular students as those students who enrolled for a number of years not exceeding the normal duration of the course. In detail, classes are defined as follows.

- **Regular (RS):** students who regularly enrolled every year, who never changed course, and who have gained more than 70% of the university course credits (CFU) required by the degree program (including those obtained with the graduation thesis) within 36 months after the beginning of their learning process (October 2011). The dataset contains 375 RS instances.
- **Irregular (IS):** students who regularly enrolled every year and who got less than 70% of the CFU required by the degree program (including those

[1] The complete list can be found in `https://www.dropbox.com/s/a1ub2cflqmr7eky/InputVariables.pdf`

obtained with the graduation thesis) by October 2011. The dataset contains 283 IS instances.

- **Dropped out (DS):** students who formally gave up their studies by April 2010 or who did not renew their enrollment within two years. This category contains also those students who formalized a change of the course of study. The dataset contains 152 DS instances, composed by 138 abandonments and 14 changes of courses.

The dataset has been randomly split $n_S = 100$ times into a learning set (LS) and a test set (TS). The LS contains 90% of the 810 instances, while the TS is composed by the remaining 10% of the samples. For each one of the split procedure we trained three different kinds of RF, each with a number of trees equal to $n_T = 500$:

- the original RF proposed in [15] (RFB);
- the newly RF proposed in [16] (RFR) which introduces a random feature rotation;
- the ϵ-DP RF proposed in [21] (DPRF) with $\epsilon = \sqrt{\ln(3/\delta)/n}$ and $\delta = 0.05$, being n the number of samples used for training the forest, according to what presented in Section 2.

In Table 1 we reported the confusion matrixes over the TS, averaged over the n_S splits of RFB, RFR, and DPRF trained with the LS.

Results indicate that dropped out students can be distinguished from the non–dropped out ones with high accuracy (between 80% and 90%).

On the other hand, all algorithms make a high error in discriminating between regular and irregular students. This is not surprising since the motivations that make a student irregular may be very different in nature and vary from family problems to diseases, temporary economic issues, etc. In order to be able to correctly predict this phenomenon, a better characterization of the instances should be done. Unfortunately, the additional information needed is not stored within the university databases.

Comparing the performances of the three algorithms, it can be easily seen that, in the non–private framework, RFR shows a slightly greater accuracy than RFB.

(a) RFB

$n_T = 500$

	RS	IS	DS
RS	44.0	7.9	0.0
IS	11.4	23.0	1.0
DS	0.0	0.9	11.8

(b) RFR

$n_T = 500$

	RS	IS	DS
RS	46.1	6.7	0.0
IS	8.3	25.2	0.8
DS	0.0	0.9	12.0

(c) DPRF

n_T=500, ϵ=$\sqrt{\ln(3/\delta)/n}$

	RS	IS	DS
RS	40.0	10.9	0.0
IS	13.4	22.0	1.1
DS	0.0	1.2	11.4

Table 1: Confusion matrixes (in percentage) over the TS, averaged over the n_S splits of RFB, RFR, and DPRF.

Moreover, even if DPRF is outperformed by both RFB and RFR, the quality of the prediction system does not significantly decrease with the privacy constraints. This result is of paramount importance since it means that the data of a larger population of students can be accessed, without requiring any signed consent form.

In conclusion, we can state that a privacy preserving data driven approach to the prediction of the university student dropouts is a promising tool which needs to be tested on a larger population of students, both over a wider temporal scale and over a larger number of degree programs.

References

[1] CEDEFOP. Europe's uneven return to job growth. In *Briefing Note*, 2015.

[2] Organisation for Economic Co-operation and Development. *Education at a Glance 2016: OECD Indicators*. OECD, 2016.

[3] European Commission. A strategy for smart, sustainable and inclusive growth. In *Communication From The Commission*, 2010.

[4] European Commission. Eurostat online data code. In *Europe 2020 indicators - education*, 2016.

[5] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos, and C. P. Rosé. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4):333–353, 2015.

[6] W. Mason, J. W. Vaughan, and H. Wallach. Computational social science and social computing. *Machine Learning*, 95(3):257, 2014.

[7] Z. K. Papamitsiou and Anastasios A Economides. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

[8] R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17, 2009.

[9] V. S Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[10] S. Greengard. Privacy matters. *Communication ACM*, 51(9):17–18, 2008.

[11] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):1–277, 2014.

[12] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Symposium on Theory of Computing*, 2015.

[13] P. Jain and A. Thakurta. Differentially private learning with kernels. *ICML*, 2013.

[14] K. Chaudhuri, A. Sarwate, and K. Sinha. Near-optimal differentially private principal components. In *NIPS*, 2012.

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

[16] R. Blaser and P. Fryzlewicz. Random rotation ensembles. *JMLR*, 17(4):1–26, 2016.

[17] M. Wainberg, B. Alipanahi, and B. J. Frey. Are random forests truly the best classifiers? *JMLR*, 17(110):1–5, 2016.

[18] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright. A practical differentially private random decision tree classifier. In *ICDM*, 2009.

[19] M. Bojarski, A. Choromanska, K. Choromanski, and Y. LeCun. Differentially-and non-differentially-private random decision trees. In *arXiv preprint arXiv:1410.6973*, 2014.

[20] A. Patil and S. Singh. Differential private random forest. In *ICACCI*, 2014.

[21] S. Rana, S. K. Gupta, and S. Venkatesh. Differentially private random forest with high utility. In *ICDM*, 2015.