Data Augmentation for Enlarging Student Feature Space and Improving Random Forest Success Prediction

Timothy Bell, Christel Dartigues-Pallez, Florent Jaillet, Christophe Genolini

Data Augmentation for Enlarging Student Feature Space and Improving Random Forest Success Prediction

Chapter · June 2021

DOI: 10.1007/978-3-030-78270-2_14

Christel Dartigues-Pallez
University of Nice Sophia Antipolis

Christophe Genolini
Université Paris Nanterre

Data Augmentation for Enlarging Student Feature Space and Improving Random Forest Success Prediction

Timothy H. Bell(✉)[1,2][0000−0001−8547−6232], Christel Dartigues-Pallez[1][0000−0001−5727−5142], Florent Jaillet[1][0000−0003−3781−9514], and Christophe Genolini[2][0000−0002−3321−6364]

[1] Université Côte d'Azur, CNRS, I3S, France {Timothy.BELL, Christel.DARTIGUES-PALLEZ,Florent.JAILLET,}@univ-cotedazur.fr
[2] R++, 5 place Jean Deschamps, 31100 Toulouse, France cg@rplusplus.com

Abstract. One of the main problems encountered when predicting student success, as a tool to aid students, is the lack of data used to model each student. This lack of data is due in part to the small number of students in each university course and also, the limited number of features that describe the educational background for each student. In this article, we introduce new features by augmenting the student feature space to obtain an improved model. These features are divided into several groups, namely, external added data, metric and counter data, and evolutive data. We will then assess the quality of the augmented data to classify at-risk students in their first year of university. For this article, the classifiers are built using Random Forests. As this learning method measures variable importance, we can enquire on the relevance of the augmented data, as well as the data groups that allow a more significant collection of features.

Keywords: Student Success · Random Forest · Data Augmentation · Educational Data Mining · Student Metrics.

1 Introduction

In France among the students in their first year of university one in two will either repeat the year, change studies, or stop the course mid-year [14]. In 2017 only 29% got their first cycle degree without repeating or changing course. Many approaches to predict student success have been investigated through means of grade prediction or dropout prevention [2,3,7,8]. Generic data such as secondary education grades but also sociodemographic indicators [6,8,10] are used to predict student outcome. This ends in having a small amount of features usable by learning algorithms to output predictions. For this article, we will augment our initial set of data by performing operations on the existing features to obtain ratios or time-series coefficients. We also have metrics on the various high-schools. To classify at-risk students we are using Random Forests [4]. We first introduce the data to train the model, then our method for augmenting the given data, and lastly, before concluding, we discuss the obtained results.

2 Data

2.1 Initial Data

The data used comes from students studying in a University Institute of Technology. Students enter UITs after completing secondary studies. This particular set of data is pooled from first-year students of 18 different courses. All the data is thoroughly anonymized beforehand to respect student privacy within the General Data Protection Regulation [1]. Among the different courses, classes vary in size and display a very heterogeneous distribution of students. The particular set we are working on is of the year 2019 with a high of 169 students in one promotion and as low as 10 for another. All the data used for training is taken from the students' curriculum during his secondary education at high-school.

The French high-school system is divided into 3 years, and each school year is divided into 3 trimesters. For each student we have data from the first trimester of the second year up to the second trimester of the third year. In total five trimesters. We also have the results for the Baccalauréat (the end of high-school exam). Each year, a variety of subjects are taught, some common core courses (e.g., Physical Education) and some speciality courses (e.g., Economy). For every subject we have (see table 1) the student's grade (Stu), the class's mean (Avg), the class's highest grade (Max), and the class's lowest grade (Min). Most of the augmented data derive from these features.

Table 1. Stored information for each subject.

Stu	Avg	Min	Max
13	8.5	7	17
...	...	...	...
16	10	6	17

Additional data consist of: professor comments for every subject and each trimester, a cover letter, the student's high-school name, and comments from the high-school on the student's potential for succeeding in further studies. For this article only numeric data is used, disregarding all non-ordinal or non-categorical textual data. Therefore the professor comment, high-school comments, and cover letter are omitted in this work.

Lastly since optional courses can be taken at school, we get rid of features with a high number of missing data ($> 70\%$) during a pre-processing step.

To train the model we are doing supervised learning, and the label for each student is whether the student passed or failed. This is done by discretizing their weighted mean grade in the first university semester. This weighted mean attributes more weight to more important courses depending on the department and the field of study.

2.2 Augmented Data

We separated the augmented data into 3 groups to attribute changes in the models' outcomes to the different data. The augmented data is divided as such: pre-processed initial data(PPD), external data(G1), metrics and counters(G2), evolving data(G3).

Although most of the features are numeric, i.e., grades, some are nominal such as chosen language courses and some ordinal (Good, Very Good, etc.). These features are encoded respectively by one-hot encoding and ordinal encoding.

The first group of external data (G1) consists of various metrics for French high-schools: the percentage of students that repeat years, the percentage of graduated-with-honours students, the percentage of students that pass the final examination, and lastly, the added-value which indicates how well the high-school performed given its sociodemographic context.

The metrics and counters group (G2) holds features obtained from simple calculations: Stu-Avg, the student's highest grade - lowest grade for any given trimester, Stu-Min, Stu-Max. It also has the number of: repeated years, top marks, lowest marks, times Stu<Avg.

For the last group of data (G3), we apply linear regression, by k-combinations of all trimesters, to extract the regression coefficients β. These coefficients depict the evolution of G1 and G2 data. For instance, the evolution of: the student's grades, number of top marks, the difference Stu-Avg.

3 Methodology

For this article we chose Random Forest (RF) due to its high classification accuracy rates seen in [9,12,13,15]. To assess the efficiency of the applied methodology we use 'Zero-Rule' [5] as a baseline for this classification task. This predicts the class as the majority class, in this case the majority will always be students that have passed the first semester. Therefore, in a class of 100 students where 20 failed, we hope to achieve at least 80% in accuracy. This objective also avoids using false model accuracy due to class imbalance mentioned in [3]. When running our algorithms, departments are trained separately for this paper as certain features (e.g., French and mathematics grades) vary in relevancy depending on the chosen field of studies. Additionally, Random Forests' built-in Gini importance will be used to score each feature and their importance. The Gini importance will allow us to assess if our augmentation creates any relevant features.

For each configuration and its corresponding RF model we run the model 10 times with 10-fold cross validation to test the performance. The metrics used are the accuracy in classification and the F1-score.

4 Results

For the courses with less than 50 students, the results were inconclusive. The prediction didn't, or barely, perform better than the baseline. This was expected as the sample population is too low.

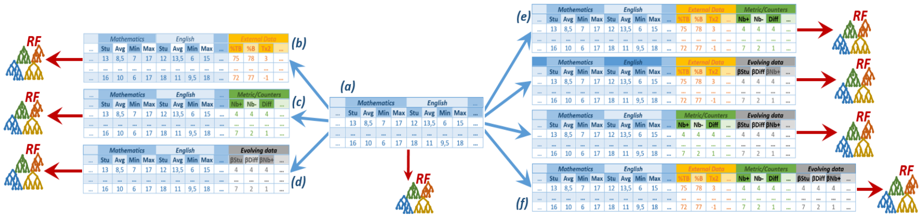

Fig. 1. Data group evaluation process. Blue: PPD. Yellow: G1. Green: G2. Grey: G3.

The higher scores on average were obtained with a combination of all groups, (f) in figure 1.

Table 2. Resulting classification scores for each model on the Computer Science course.

Scores	PPD	PPD+G1	PPD+G2	PPD+G3	PPD+G1+G2	PPD+G1+G3	PPD+G2+G3	PPD+G1+G2+G3
Accuracy	0.76	0.77	0.85	0.78	0.88	0.77	0.85	**0.89**
F1-score	0.73	0.69	0.79	0.75	0.81	0.70	0.79	**0.84**

For the particular course in table 2, we notice that G1 and G3 only marginally improve the classification. Whereas G2 improves it by quite a lot. But some features in both G1 and G3 can have substantial importance regarding the classification, therefore it might be interesting to perform feature selection [11] on all the augmented features. There were only 350 features before augmentation, and 1500 features total after augmentation. Some examples of augmented features that figure in the top 10 most relevant features are: Stu-Min for 3rd trimester French (G2), Stu-Min for 3rd trimester French (G2), regression on the 2nd, 3rd, and 5th trimesters in mathematics (G3), regression of Stu-Avg on trimesters 2 and 3 in English (G3). Interestingly some other courses such as the statistics course holds all 10 top features as augmented features from groups G2 and G3 with mostly regressions on the student's relative grades to the class's highest grades in mathematics.

5 Conclusion

This work sought to extend the feature space to improve student failure prediction, allowing a better understanding of what features may best represent students. Data augmentation improved prediction with courses of more than 50 students. It can also be used as a tool for Random Forest Feature Selection prior to inputting this into any learning model.

In future works, extra textual data could be exploited. Our dataset also provides for each subject and trimester professor comments usually in a dozen words. These comments usually hold information such as regular absenteeism and class disruption. The next step will be to incorporate these comments in the model as well, and further increase the prediction accuracy of our model.

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), http://data.europa.eu/eli/reg/2016/679/2016-05-04
2. Balakrishan, G.: Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. Ph.D. thesis, EECS Department, University of California, Berkeley (May 2013)
3. Barros, T.M., Souza Neto, P.A., Silva, I., Guedes, L.A.: Predictive Models for Imbalanced Data: A School Dropout Perspective. Education Sciences **9**(4), 275 (Dec 2019). https://doi.org/10.3390/educsci9040275, number: 4 Publisher: Multidisciplinary Digital Publishing Institute
4. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/A:1010933404324
5. Choudhary, R., Gianey, H.K.: Comprehensive Review On Supervised Machine Learning Algorithms. 2017 International Conference on Machine Learning and Data Science (MLDS) (2017). https://doi.org/10.1109/MLDS.2017.11
6. Cortez, P., Silva, A.: Using data mining to predict secondary school student performance. EUROSIS (Jan 2008)
7. Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P.: Student Dropout Prediction. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) Artificial Intelligence in Education. pp. 129–140. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_11
8. Gonçalves, O., Beltrame, W.: Socioeconomic Data Mining and Student Dropout: Analyzing a Higher Education Course in Brazil. International Journal for Innovation Education and Research **8**, 505–518 (Aug 2020). https://doi.org/10.31686/ijier.vol8.iss8.2554
9. Hussain, S., Dahan, N.A., Ba-Alwi, F.M., Ribata, N.: Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. Indonesian Journal of Electrical Engineering and Computer Science **9**(2), 447–459 (Feb 2018). https://doi.org/10.11591/ijeecs.v9.i2.pp447-459
10. J. Kovacic, Z.: Early Prediction of Student Success: Mining Students Enrolment Data. pp. 647–665 (2010). https://doi.org/10.28945/1281
11. Ma, C., Yao, B., Ge, F., Pan, Y., Guo, Y.: Improving Prediction of Student Performance based on Multiple Feature Selection Approaches. In: Proceedings of the 2017 International Conference on E-Education, E-Business and E-Technology. pp. 36–41. ICEBT 2017, Association for Computing Machinery, New York, NY, USA (Sep 2017). https://doi.org/10.1145/3141151.3141160
12. Mahboob, T., Irfan, S., Karamat, A.: A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms. In: 2016 19th International Multi-Topic Conference (INMIC). pp. 1–8 (Dec 2016). https://doi.org/10.1109/INMIC.2016.7840094
13. Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A.: Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems **115**, 36–51 (Nov 2018). https://doi.org/10.1016/j.dss.2018.09.001

14. Razafindratsima, N.: État de l'Enseignement supérieur, de la Recherche et de l'Innovation en France. Les parcours et la réussite en Licence, Licence professionnelle et Master à l'université - État de l'Enseignement supérieur, de la Recherche et de l'Innovation en France **n°13**, 50–51 (2020), https://publication.enseignementsup-recherche.gouv.fr/eesr/FR/T149/les_parcours_et_la_reussite_en_licence_licence_professionnelle_et_master_a_l_universite/
15. Sorour, S.E., Mine, T.: Building an Interpretable Model of Predicting Student Performance Using Comment Data Mining. In: 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). pp. 285–291 (Jul 2016). https://doi.org/10.1109/IIAI-AAI.2016.114