Early Prediction of student's Performance in Higher Education: A Case Study

Mónica V. Martins[1(✉)], Daniel Tolledo[1], Jorge Machado[1], Luís M. T. Baptista[1], and Valentim Realinho[1,2]

[1] Polytechnic Institute of Portalegre (IPP), Portalegre, Portugal
mvmartins@ipportelegre.pt
[2] VALORIZA - Research Center for Endogenous Resource Valorization, Portalegre, Portugal

Abstract. This work aims to contribute to the reduction of academic failure at higher education, by using machine learning techniques to identify students at risk of failure at an early stage of their academic path, so that strategies to support them can be put into place. A dataset from a higher education institution is used to build classification models to predict academic performance of students. The dataset includes information known at the time of student's enrollment – academic path, demographics and social-economic factors. The problem is formulated as a three category classification task, in which there's a strong imbalance towards one of the classes. Algorithms to promote class balancing with synthetic oversampling are tested, and classification models are trained and evaluated, both with standard machine learning algorithms and state of the art boosting algorithms. Our results show that boosting algorithms respond better to the specific classification task than standard methods. However, even these state of the art algorithms fall short in correctly identifying the majority of cases in one of the minority classes. Future directions of this study include the addition of information regarding student's first year performance, such as academic grades from the first academic semesters.

Keywords: Academic performance · Machine learning in education · Imbalanced classes · Multi-class classification · Boosting methods

1 Introduction

One of the challenges that higher education institutions around the globe face is to effectively deal with the different student's learning styles and different academic performances, as a means to promote student's learning experience and institution's formative efficiency. The ability to predict and anticipates student's potential difficulties is of interest for the institutions that aim to build strategies to provide support and guidance to students that might be at risk of academic failure or dropout. At the same time, large amount of data is collected each year by the institutions, including information regarding the academic path of the student, as well as demographics and socio-economic factors. The two combined factors make this a fertile ground for the contribution of machine learning approaches to predict student's performance.

Á. Rocha et al. (Eds.): WorldCIST 2021, AISC 1365, pp. 166–175, 2021.
https://doi.org/10.1007/978-3-030-72657-7_16

In this study, data from Polytechnic Institute of Portalegre (IPP), Portugal, is used to build machine learning classification models to predict students that might be at risk of failing to succeed in finishing their degrees in due time. The main goal is to provide a system that allows to identify, at a very initial stage, students with potential difficulties in their academic path, so that strategies to support the students can be put into place.

Some of the aspects of the work here presented that differ from a number of similar works found in the literature are as follows: (i) it doesn't focus on any specific field of study, because the focus is to build a system that generalizes to any course of IPP. Therefore, the dataset includes information from students enrolled in the several courses of the four different schools belonging to IPP; (ii) it only relies on information available at the moment of enrollment, because the focus is to develop a system that helps to segment students as soon as possible from the beginning of their path at higher education. This means that no information regarding academic performance after enrolment is used; (iii) differently from the usual approach of restricting the model categories to failure / success, it's also used a third intermediate class (relative success), because the kind of interventions for academic support and guidance might be quite different for students who are in moderate risk from those who are at high risk of being unsuccessful. As a result, methods to deal with the unbalanced nature of the resulting classes must be considered; (iv) besides using standard classification models, it also uses state of the art boosting algorithms to build the classification models.

The remainder of this paper is structured as follows: Sect. 2 presents a brief review of the literature; Sect. 3 describes the methodology, including a description of the data, the methods used to deal with the unbalanced dataset and the procedures for training and evaluation of the classification models; Sect. 4 presents and analyses the results and Sect. 5 indicates some directions for future work.

2 Related Work

The task of predicting academic success in higher education is not a new one, and many researchers have tried different approaches, using different models and different information. Mostly, researchers and higher education institutions are interested in being able to predict if a student is at risk of not completing his program, or of dropping out, because this information might be valuable for putting into place strategies to help those students move forward. In reviews such as [1–4] one can find extensive information regarding the different approaches that have been used to study this issue. Here, we review a few very recent works (published in the last 4 years) that find similarities with the work presented in this paper.

Beaulac and Roosenthal [5] analyze a large data set (38 842 students) from a large university in Canada to predict academic success using Random Forests (RF). The authors use the first few courses attempted and grades obtained by students in order to predict whether the student will complete their program; and if yes, which major they will complete. For the prediction of program completion, an overall 79% accuracy is obtained, with 91% for the class of students who completed their program, and 53% for the students who didn't. Regarding the prediction of the major, 47% accuracy was obtained.

Hoffait and Schyns [6] use a dataset of 6845 students and standard classification methods (RF, Logistic Regression (LR) and Artificial Neural Networks(ANN)) to identify freshmen's profiles likely to face major difficulties to complete their first academic year. The obtained accuracy for the majority class is about 70%, and for the minority class less than 60%, regardless of the algorithm used. They then use RF to develop a strategy to improve the accuracy of the prediction for some classes of major interest. The developed approach does not always lead to an increase in the identification of the number of students at risk.

Miguéis et al. [7] use the information available at the end of the first year of students' academic path to predict their overall academic performance, inferring academic success both from the average grade achieved and the time taken to conclude the degree. Their prediction models use information regarding demographics and social factors as well as academic measures, including assessments from first year courses. They use a dataset of 2459 students from a European Engineering School to build several models using Support Vector Machines, Naïve Bayes, Decision Trees (DT), RF, Bagging Decision Trees and Adaptive Boosting Decision Trees, obtaining the higher scores (overall accuracy of 96%) with Random Forests and Adaptive Boosting Decision Trees.

One of the common problems in classification student's success or dropout is class imbalance. Class imbalance happens when one or more of the classification categories have significant lower number of records than a majority class. This might result in a high prediction accuracy driven by the majority class at the expense of very poor performance on the minority classes. In the case of student's performance, this is a common problem because only a minority of students will underperform or drop out. Nevertheless, these are the classes of students that researchers aim to best identify.

In [8] Thammasiri et al. conduct a study using different class balancing strategies and several standard classification methods to predict dropout in a dataset with 21654 students. They compare class balancing techniques based on random under sampling, random oversampling, or synthetic oversampling. Best results are obtained with this last approach, named synthetic minority over-sampling technique (SMOTE) [9]. This approach has indeed shown to successfully tackle the imbalanced classification issue in different domains [10] and will be further explained in Sect. 3.2.

3 Methodology

In this section we present the dataset, the methods used to deal with the imbalanced nature of the data, and the methodology used to build and evaluate the classification models.

3.1 Data

In this study we use institutional data (acquired from several disjoint databases) related to students enrolled in undergraduate courses of Polytechnic Institute of Portalegre, Portugal. The data refers to records of students enrolled between academic years 2008/09–2018/2019 and from different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service and technologies.

We performed a rigorous data preprocessing to handle data from anomalies, unexplainable outliers and missing values, and dropped records that couldn't be classified as explained below (last 3 or 4 academic years, depending on the course duration). The final dataset consisted of 3623 records and 25 independent variables.

The data contains variables related to demographic factors (age at enrollment, gender, marital status, nationality, address code, special needs) socio-economic factors (student-worker, parent's habilitations, parent's professions, parent's employment situation, student grant, student's debt) and student's academic path (admission grade, retention years at high school, order of choice for enrolled course, type of course at high school). We limit the academic information to factors observable prior to registration, excluding internal assessments after enrollment.

Each record was classified as Success, Relative Success and Failure, depending on the time that the student took to obtain her degree. Success means that the student obtained the degree in due time; Relative Success means that the student took until three extra years to obtain the degree; Failure means that the student took more than three extra years to obtain the degree or doesn't obtain the degree at all. This somehow corresponds to three levels of risk: 'low-risk' students with high probability of succeeding; 'medium-risk' students, for whom the measures taken by the institution might contribute to success; and the 'high-risk' students, who have a high probability of failing. The distribution of the records among the three categories is shown in Fig. 1.

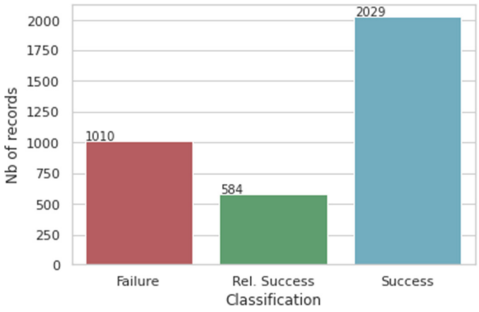

Fig. 1. Distribution of student's records among the three categories considered for academic success.

The distribution of the records among the three categories is imbalanced, with two minority classes, "Failure" and "Relative Success". "Failure" accounts for 28% of total records, and "Relative Success" accounts for 16% of total records, while the majority class, "Success", accounts for 56% of the records. The classes that we most aim to correctly identify are the minority ones, since the students from these classes are the ones that might benefit from planned interventions for academic support and guidance. How we tackled this multi-class imbalanced classification task will be explained in the next sections.

3.2 Data Sampling Techniques

Sampling strategies are often used to overcome the class imbalance problem, either eliminating some data from the majority class (under-sampling) or duplicating data of the minority class (over-sampling) or adding some artificially generated data to the minority class. The under-sampling techniques have the disadvantage of reducing the size of the data set; the over-sampling by data duplication approaches have the disadvantage of adding no new information to the models. The data augmentation approaches by synthesis of new data from the minority class have shown to produce very good results in imbalanced classification.

We used two strategies for data augmentation with SMOTE based sampling methods [9] applied to the two minority classes in our dataset: the plain SMOTE algorithm and the Adaptive Synthetic (ADASYN) [11] algorithm.

The SMOTE algorithm works by finding neighbor examples from the minority class in the feature space and synthetizing a new example in the space between the neighbors. This procedure is used as many times as needed to create a balance between the number of samples in the classes. The ADASYSN algorithm is derived from SMOTE, and features one important difference that has to do with how it chooses the points for synthetizing new examples, biasing its choice towards non homogeneous neighborhoods.

We used the implementations available at the *imbalanced-learn* module in *scikit-learn*, in Python [12], which allow to deal with multiple minority classes, such as the present case. For comparison and evaluation purposes, a Logistic Regression model was built with the original dataset. Then SMOTE and ADASYN were applied to the dataset, and a Logistic Regression model was built for each case. Therefore, for this part of the work, three different models were built.

3.3 Classification Models

Regarding the algorithms used for building the classification models, some of the standard algorithms often reported in the literature for student's performance classification were used as a first approach: Logistic Regression (LR) [13], Support Vector Machines (SVM) [14], Decision Trees (DT) [15] and an ensemble method, Random Forests (RF) [16]. To train these models the implementations available at the Scikit-learn library in Python [17] were used.

Then, a second stage went a step further and boosting methods were also exploited. Albeit being underrepresented in the educational context, boosting methods reportedly work well with imbalanced data classification, including multi-classification tasks [10, 18]. Boosting methods are a class of ensemble methods that build a strong model from the sequential training of weak models. There are many boosting schemes available, but all are variations of a general boosting scheme. Starting with an initial prediction model, in each boosting round a weak classifier is produced that aims to reduce the errors of the previous classifier. After a defined number of rounds, these sequentially built weak classifiers give origin to a single strong prediction model that is much more accurate than any of the previous weak learning models. Following some of the most interesting results reported in the literature for multiclass imbalanced classification [10, 18], we used

four general boosting methods that are applicable to multi-class classification: Gradient Boosting [19], Extreme Gradient Boosting [20], CatBoost [21] and LogitBoost [22].

Details on the model training procedure and evaluation are given in the next section.

3.4 Model Training, Evaluation Metrics and Hyperparameter Tuning

Following the usual procedure, data was divided into training set (80%) and test set (20%). Then, for each model, a 5-fold cross validation procedure was used to avoid overfitting. This means that the training data set was divided into 5 blocks, and the training of each model was done with 4 of the blocks, with the remaining one being used for validation purposes. The process was repeated 5 times, once for each block, thus enabling the maximization of the total number of observations used for validation while avoiding overfitting. The best average cross-validation estimator score was elected. This methodology also included a procedure to assure that every class was well represented in every fold. Then, the overall performance of each elected model was assessed with the test set.

Due to the imbalanced nature of our dataset, accuracy isn't the most adequate measure for model performance, since it's an overall metric that might result in high values based on a good performance solely for the majority class. For imbalanced data, single class metrics are more adequate [18]. In this work we use the F1 measure, which accounts for the trade-off between precision and recall. F1 scores were computed for each class, and the average F1 score for the three classes was also computed. This was the metric used for hyperparameter tuning, as will be explained next. For the optimized model, accuracy was also computed as an overall metric.

All the models went through a process of hyperparameter optimization. One way of tuning the hyperparameters is to perform a grid search, a very exhaustive way of testing many configuration and picking the one that performs better with cross-validation. This approach was used for LG, SVM and DT, using the Grid Search method available at Scikit-learn, using F1-score as the metric to be maximized. For the remaining models, the parameter space was much wider, and a Randomized Grid Search was used instead, where a set of parameter values and combinations is randomly chosen, allowing the control of the number of parameter combinations that are attempted.

4 Results

In this section we present the performance of the proposed models.

4.1 Data Sampling Techniques

Table 1 shows the performance metrics obtained with the test set for the logistic regression models build without correction for the minority classes, or used together either with SMOTE or with ADASYN.

These results show that the LR model without data sampling results in very low F1-score for the "Relative Success" class, the one with fewer samples, albeit resulting in the highest accuracy. This expresses the fact that, in imbalanced classes tasks, accuracy

Table 1. Classification performance without and with data sampling

	Logistic Regression	SMOTE + Logistic Regression	ADASYN + Logistic Regression
F1-score Failure	0.61	0.64	0.63
F1-score Rel.Success	0.06	0.41	0.38
F1-score Success	0.77	0.69	0.69
Average F1-score	0.49	0.58	0.56
Accuracy	0.68	0.61	0.60

alone is not a good performance metric. LR with SMOTE or with ADASYN result in better discrimination for the minority class, although still a low value for F1. SMOTE and ADASYN produce similar results for this dataset. The use of SMOTE leads to the highest F1-scores, either overall and for the individual classes. Therefore, for the remainder of this work, SMOTE was applied to the dataset prior to model training.

4.2 Standard Models

Table 2 presents the metrics obtained for the standard methods used in this work, after hyperparameter optimization. Random Forest lead to the best overall results, which is in line with some of the results reported in the literature [7]. On the other hand, SVM is the worst performer, contrary to what other researchers have obtained [8]. This is not unusual, though, because, depending on the dataset and on the formulation of the problem, any machine learning technique can achieve superior results, prompting an experimental approach to identify the best learner for each task.

Table 2. Classification performance for the standard models

	Logistic Regression	Support Vector Machine	Decision Tree	Random Forest
F1-score Failure	0.63	0.53	0.63	0.66
F1-score Rel.Success	0.41	0.31	0.39	0.37
F1-score Success	0.69	0.71	0.75	0.82
Average F1-score	0.58	0.52	0.59	**0.62**
Accuracy	0.61	0.60	0.65	**0.72**

4.3 Boosting Models

Table 3 presents the classification performance for the boosting methods. In general, the models built with the boosting methods outperform the models built with the standard methods, both on the in-class metrics and on the overall metrics. Among these, Extreme Gradient Boosting is the best classifier, although very similar to Gradient Boosting.

In both cases the lowest F1-score among the three classes is obtained for the "Relative Success" class, whereas the highest F1-score is obtained for the majority class.

Table 3. Classification performance for boosting models

	Gradient Boosting	Extreme Gradient Boosting	Logit Boost	CatBoost
F1-score Failure	0.68	0.68	0.69	0.69
F1-score Rel.Success	0.44	0.44	0.41	0.35
F1-score Sucess	0.81	0.83	0.82	0.82
Average F1-score	0.65	**0.65**	0.64	0.62
Accuracy	0.72	**0.73**	0.72	0.73

The fact that gradient boosting models outperform most standard machine learning models is in agreement with results obtained in other fields of study. In our case, however, even those models fail in identifying most of the students belonging to the most critical, minority class. This only confirms that the classification of an imbalanced dataset, on a multi class frame is a complicated task [18].

5 Conclusions and Future Work

In this work we use dataset from a higher education institution to build classification models to early predict academic performance of students. The data set includes information known at the time of enrollment – demographics, academic performance prior to enrollment, social-economics - but none information regarding performance after enrollment. We addressed the problem as a three category classification task, in which there's a strong imbalance towards one of the classes. The minority classes are also the target classes for this work. We used algorithms to promote class balancing with synthetic oversampling and built classification models both with standard machine learning algorithms and boosting algorithms. Our results show that boosting algorithms respond better to the specific classification task than standard methods, but even them fail in correctly classifying the minority classes. As a means to improve these results, information regarding the academic performance of students during the first academic semesters will also be included in the dataset for model building.

Acknowledgments. This research is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191.

References

1. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **40**, 601–618 (2010). https://doi.org/10.1109/TSMCC.2010.2053532
2. Mduma, N., Kalegele, K., Machuve, D.: A survey of machine learning approaches and techniques for student dropout prediction. Data Sci. J. **18**, 1–10 (2019). https://doi.org/10.5334/dsj-2019-014
3. Shahiri, A.M., Husain, W., Rashid, N.A.: A review on predicting Student's performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422. (2015). https://doi.org/10.1016/j.procs.2015.12.157
4. Rastrollo-Guerrero, J.L., Gómez-Pulido, J.A., Durán-Domínguez, A.: Analyzing and predicting Students' performance by means of machine learning: a review. Appl. Sci. **10**, 1042–1058 (2020). https://doi.org/10.3390/app10031042
5. Beaulac, C., Rosenthal, J.S.: Predicting university Students' academic success and major using random forests. Res. High. Educ. **60**, 1048–1064 (2019). https://doi.org/10.1007/s11162-019-09546-y
6. Hoffait, A.S., Schyns, M.: Early detection of university Students with potential difficulties. Decis. Support Syst. **101**, 1–11 (2017). https://doi.org/10.1016/j.dss.2017.05.003
7. Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A.: Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. **115**, 36–51 (2018). https://doi.org/10.1016/j.dss.2018.09.001
8. Thammasiri, D., Delen, D., Meesad, P., Kasap, N.: A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Expert Syst. Appl. **41**, 321–330 (2014). https://doi.org/10.1016/j.eswa.2013.07.046
9. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
10. Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem: a review. Int. J. Adv. Soft. Comput. Appl. **7**, 176–204 (2015)
11. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the International Joint Conference on Neural Networks, pp. 1322–1328 (2008). https://doi.org/10.1109/IJCNN.2008.4633969
12. Lema, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **40**, 1–5 (2015)
13. Hastie, T.J., Pregibon, D.: Generalized linear models. In: Statistical Models in S (2017)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995). https://doi.org/10.1023/A:1022627411411
15. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986). https://doi.org/10.1007/bf00116251
16. Pavlov, Y.L.: Random forests. Random Forests 1–122 (2019). https://doi.org/10.1201/9780367816377-11
17. Pedregosa, F., Gaël, V., Gramfort, A., Vincent, M., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **85**, 2825 (2011)
18. Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M.: Boosting methods for multi-class imbalanced data classification: an experimental review. J. Big Data. **7**, 70 (2020). https://doi.org/10.1186/s40537-020-00349-y

19. Friedman, J.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
20. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
21. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. In: NIPS 2018: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 6639–6649, December 2018. https://dl.acm.org/doi/10.5555/3327757.3327770
22. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Statist. **28**(2), 337–407 (2000). https://doi.org/10.1214/aos/1016218223