A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance

Pedro Strecht
INESC TEC
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
pstrecht@fe.up.pt

Luís Cruz
INESC TEC
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
luiscruz@fe.up.pt

Carlos Soares
INESC TEC
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
csoares@fe.up.pt

João Mendes-Moreira
INESC TEC
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
jmoreira@fe.up.pt

Rui Abreu
INESC TEC
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
rma@fe.up.pt

ABSTRACT

Predicting the success or failure of a student in a course or program is a problem that has recently been addressed using data mining techniques. In this paper we evaluate some of the most popular classification and regression algorithms on this problem. We address two problems: prediction of approval/failure and prediction of grade. The former is tackled as a classification task while the latter as a regression task. Separate models are trained for each course. The experiments were carried out using administrate data from the University of Porto, concerning approximately 700 courses. The algorithms with best results overall in classification were decision trees and SVM while in regression they were SVM, Random Forest, and AdaBoost.R2. However, in the classification setting, the algorithms are finding useful patterns, while, in regression, the models obtained are not able to beat a simple baseline.

Keywords

Regression, Classification, Academic Performance

1. INTRODUCTION

Recently, the University of Porto (UPorto) identified modelling of the success/failure of students in each course as one of its priorities. The goal is to use the models for two tasks: make predictions for the individual performance of students in courses and understand the factors associated with success and failure. These models are relevant to five levels of decision, namely: Course teacher, Program Director, Department Director, Faculty Director and University Rector. Course teachers and program directors can use the models to identify students at risk and devise strategies that can reduce the risk of failure. Also, program directors as well as department directors can find them useful in designing program syllabus. Finally, the top levels of university management can use these models to understand general trends and behaviours in student performance, which can lead to new or adapted pedagogical strategies.

The fact that models are needed for different levels of decision requires that these models have different granularities. In other words, course teachers and program directors are able to work with a few or a few dozen models, respectively. However, the other levels of management would have to deal with hundreds, maybe even thousands of models, which is not feasible. On the other hand, each course presents different particularities which makes the creation of a unique model to predict academic success for all the courses, an extremely hard task. Such a model would have to aggregate the different factors that influence success in very different courses. Therefore, we train a model separately for each course.

So far, the results obtained and the domain-specific constraints provide a satisfactory justification for the choice of decision trees. However, there is a need to understand the impact of this choice in the predictive accuracy of the algorithms, namely when compared with others. Additionally, although the problem of predicting if a student will pass or fail (classification task) is relevant for all levels of management of the university, the related problem of predicting the actual grade (regression task) may provide additional useful information. Therefore, this study also considers a comparative analysis of different regression algorithms. This comparison will also address the question of whether the features that are useful for classification are equally useful for regression.

The main contributions of this paper are: 1) to compare the predictive accuracy of different algorithms on the problems of predicting the performance of students in both classification (predicting success/failure) and regression (predicting the grade) tasks, particularly when comparing with decision trees, which have some other properties that deem

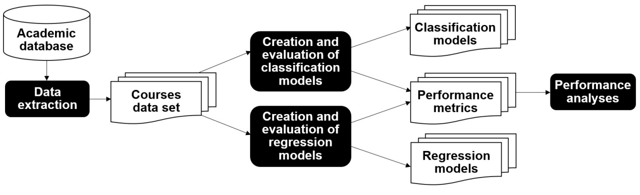

Figure 1: Experimental Setup

them suitable for this problem; 2) to assess whether the features which have obtained positive results in the classification task, and that represent essentially administrative information, are also useful to predict the grades.

The remainder of this paper is structured as follows. Section 2 presents related work. Section 3 describes the experimental set-up and methodology for both classification and regression models. Section 4 presents the results followed by section 5 with the conclusions and future work.

2. RELATED WORK

Predicting students' performance has been an issue studied previously in educational data mining research in the context of student attrition [24, 23]. Minaei-Bidgoli [13] used a combination of multiple classifiers to predict their final grade based on features extracted from logged data in an education webbased system.

Pittman [15] performed a study to explore the effectiveness of data mining methods to identify students who are at risk of leaving a particular institution. Romero et al. [16] focused on comparing different data mining methods and techniques for classifying students based on their Moodle (e-learning system) usage data and the final marks obtained in their respective programmes. The conclusion was that the most appropriate algorithm was decision trees for being accurate and comprehensible for instructors. Kabakchieva [10] also developed models for predicting student performance, based on their personal, pre-university and university performance characteristics. The highest accuracy is achieved with the neural network model, followed by the decision tree model and the kNN model.

Strecht, Mendes-Moreira and Soares [20] work predicted the failure of students in university courses using an approach to group and merge interpretable models in order to replace them with more general ones. The results show that merging models grouped by scientific areas yields an improvement in prediction quality.

3. METHODOLOGY

To carry out the experiments, a system with four processes was developed following the architecture presented in Figure 1. The first process creates the data sets (one for each course in the university) from the academic database, containing enrolment data. The courses data set were then used by two processes to create classification and regression models for each course using various algorithms. These models were evaluated using suitable performance metrics (different for classification and regression) that are collected to allow analyses and comparison by the final process.

3.1 Data Extraction

This process extracts data sets from the academic database of the university information system. The analysis done focuses on the academic year 2012/2013. A total of 5779 course data sets were extracted (from 391 programmes). The variables used were: age, sex, marital status, nationality, displaced (whether the student lived outside the Porto district), scholarship, special needs, type of admission, type of student (regular, mobility, extraordinary), status of student (ordinary, employed, athlete, ...), years of enrolment, delayed courses, type of dedication (full-time, part-time), and debt situation. The target variables are approval for classification and final grade for regression.

The final grade in these data sets is stored as a numerical value between 0 and 20. However, there are some special cases in which the grade is given as an acronym (e.g, RA means fail because of dropout), which is not feasible for regression. In such cases, in which a student failed, we converted the grade to 0.

3.2 Creation and evaluation of models

Two processes trained a set of models for classification and regression respectively for each course using different algorithms. For classification we have used k-Nearest Neighbors (kNN) [9], Random Forest (RF) [2], AdaBoost (AB) [7], Classification and Regression Trees (CART) [3], Support Vector Machines [21], Naïve Bayes (NB) [12] and for regression we used Ordinary Least Squares (OLS) [18], SVM, CART, kNN, Random Forest, and AdaBoost.R2 (AB.R2) [8].

This selection of algorithms was based on the most used algorithms for general data mining problems [22]. In this set of experiments a standard values of parameters was used. As baseline in classification we defined a model which always predicts failure. For regression, the baseline model predicts the average grade of the training set of a given course.

Models were evaluated using the k-fold cross-validation method [19] with stratified sampling [11]. The distribution of positive and negative instances is not balanced, thus it is necessary to ensure that the distribution of students in each fold respect these proportions. Failure is the positive class in this problem and we used F1 score for evaluation [5]. All regression models used 10-fold cross validation and the Root Mean Squared Error (RMSE) as evaluation measure [4].

Training and evaluation of models was replicated for each course. Courses with less then 100 students were skipped. This resulted in around 700 models for each algorithm in both classification and regression.

3.3 Performance Analyses

In both classification and regression, the algorithms were compared by placing box plots side by side relating to F1 and RMSE respectively. To get a better perspective of the distribution of results, violin plots are presented together with the box plots. The longest horizontal lines inside the boxes refer to the median while the shortest refer to the average. A few descriptive statistics were also collected and presented in tables.

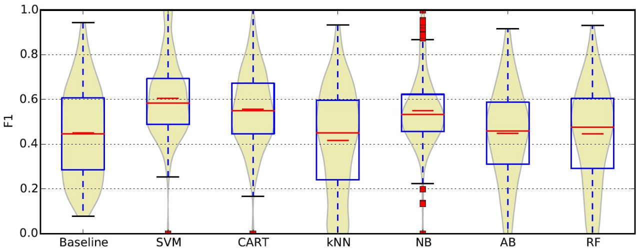

Figure 2: F1 score for each classification algorithm

In order to statistically validate the results obtained in the experiments we have used the Friedman test as suggested by Demšar to compare multiple classifiers [6]. We have used the typical value of 12 groups of models often referred as data sets in this context.

4. RESULTS

This section presents the results obtained by running experiments to train models for both classification and regression.

4.1 Classification

Figure 2 presents the F1 score distribution of models across algorithms. Table 1 presents some basic statistics about the results. Algorithms are ranked by descending order of values of the average and standard deviation of F1 scores.

The first fact that stands out from Figure 2 is that none of the algorithms present exceptional results. Albeit this, some of them seem to systematically outperform the baseline, namely SVM, CART and NB.

Table 1 confirms that SVM is the algorithm with the best performance, clearly outperforming the baseline. Not only it provides the highest average F1 score, 0.60±0.17, but sometimes it also achieves a maximum F1 score of 1.0, while the maximum score of the baseline is 0.94. Finally, although the minimum score is lower than the baseline's (0 vs. 0.08), the standard deviation is lower (0.17 vs. 0.20) which indicates that overall, it obtains more robust results.

Similar observations can be made for CART and NB. The performance of RF and AB is very similar to that of the baseline, while kNN is worse. The results of Random Forest, in particular, are surprising as this algorithm usually exhibits a very competitive performance [17].

In spite of the showing some systematic differences, the results are, overall, not very different. This is confirmed by the results of the Friedman test, $\chi^2(6) = 2.6071, p = 0.8563$, as the p-value is very high.

Table 1: Classification models results (F1)

Rank	Algorithm	Avg	Std Dev	Max	Min
1	SVM	0.60	0.17	1.00	0.00
2	CART	0.56	0.17	1.00	0.00
3	NB	0.55	0.16	1.00	0.00
4	RF	0.45	0.22	0.93	0.00
5	AB	0.45	0.21	0.92	0.00
6	Baseline	0.45	0.20	0.94	0.08
7	kNN	0.42	0.24	0.93	0.00

4.2 Regression

Figure 3 presents the distribution of RMSE values of models obtained by the algorithms. Table 2 presents some basic statistics about the results. The algorithms are ranked by ascending order of RMSE values.

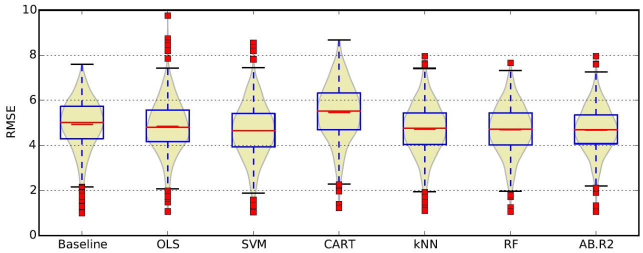

Figure 3: RMSE for each regression algorithm

As in classification, it is also quite straightforward that none of the algorithms present exceptional results. Also in this case, there is one algorithm which performs clearly worse than the baseline, CART (Table 2). Unlike classification, all violin plots show exactly the same shape, i.e., equally sized upper and lower tails. Therefore, differences are more related to overall performance (i.e. location). This shows that to compare models it is enough to consider the average and standard deviation.

The differences in performance are even smaller than in classification. However, Table 2 suggests that SVM was the best algorithm with an average of 4.65±1.19, but the standard deviation is quite large (1.19) taking into account the RMSE of the baseline (4.92). These observations are confirmed by the Friedman test ($\chi^2(6) = 3.3697, p = 0.7612$). In the case of regression, the value of the RMSE is interpretable, as it is in the same scale as the target variable. All algorithms obtain an error around 5, which is very high according to the scale (0 to 20).

In light of the results obtained in the classification setting, this is somewhat surprising, since the independent variables are the same and many of the algorithms used are based on the same principles.[1] Further analysis of the results is necessary to understand them and to identify possibilities to improve the results.

Table 2: Regression models results (RMSE)

Rank	Algorithm	Avg	Std Dev	Max	Min
1	SVM	4.65	1.19	8.54	1.03
2	RF	4.69	1.10	7.66	1.06
3	AB.R2	4.69	1.02	7.96	1.07
4	kNN	4.72	1.12	7.96	1.10
5	Baseline	4.92	1.11	7.59	1.00
6	OLS	4.84	1.19	9.75	1.06
7	CART	5.46	1.26	8.68	1.22

[1] Although this must be interpreted carefully as it is arguable to say that, for instance, SVM for classification and regression are the same algorithm.

5. CONCLUSIONS

Positive results were obtained on the classification approach where the goal is to predict whether a student will pass or fail a course. Surprisingly, however, the results on the regression approach, where the goal is to predict the grade of the student in a course, were bad. Additionally, we found no statistical evidence that the differences in performance between the algorithms are significant, although some trends are observed. Further analysis is necessary to better understand these results, which could lead to ideas for improvement. As a complement of the problems studied in this work, it should be interesting to predict an interval for a grade [1].

Some algorithms are more sensitive to parameter tuning than others. Thus it is not guaranteed that they ran with the best configuration. As future work, some optimisation could be made using an automate tuning methodology. In addition, feature selection and feature weighting can be carried out which has proven to yield good results in educational data [14].

Although the feature set used in the experiments provided some interesting results in classification, the same did not happen in regression. Thus, new features could be added. Features related to academic goals, personal interests, time management skills, sports activities, sleep habits, etc. are worthwhile investigating.

6. ACKNOWLEDGMENTS

This work is funded by projects "NORTE-07-0124-FEDER-000059" and "NORTE-07-0124-FEDER-000057", financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

7. REFERENCES

[1] R. Asif, A. Merceron, and M. Pathan. Predicting student academic performance at degree level: A case study. 2015.

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

[3] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

[4] T. Chai and R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? *Geoscientific Model Development Discussions*, 7:1525–1534, 2014.

[5] N. Chinchor. MUC-4 Evaluation Metrics. In *Proceedings of the 4th Message Understanding Conference (MUC4 '92)*, pages 22–29. Association for Computational Linguistics, 1992.

[6] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[7] T. G. Dietterich. Machine-learning research: four current directions. *AI magazine*, 18(4):97, 1997.

[8] H. Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997.

[9] E. Fix and J. L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, DTIC Document, 1951.

[10] D. Kabakchieva. Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1):61–72, 2013.

[11] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Conference on AI (IJCAI)*, pages 1137–1145, San Mateo, CA, 1995. Morgan Kaufmann.

[12] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, volume 33, pages 81–93, 1994.

[13] B. Minaei-Bidgoli. Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd ASEE/IEEE Frontiers in Education Conference*, pages 1–6, 2003.

[14] B. Minaei-Bidgoli and W. F. Punch. Using genetic algorithms for data mining optimization in an educational web-based system. In *Genetic and Evolutionary Computation—GECCO 2003*, pages 2252–2263. Springer, 2003.

[15] K. Pittman. *Comparison of data mining techniques used to predict student retention*. PhD thesis, Nova Southeastern University, 2008.

[16] C. Romero. Data mining algorithms to classify students. In *1st International Educational Data Mining Conference (EDM08)*, 2008.

[17] M. R. Segal. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, 2004.

[18] S. M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, pages 465–474, 1981.

[19] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–147, 1974.

[20] P. Strecht, J. Mendes-Moreira, and C. Soares. Merging Decision Trees: A Case Study in Predicting Student Performance. In X. Luo, J. Yu, and Z. Li, editors, *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 535–548. Springer International Publishing, 2014.

[21] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[22] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[23] A. Zafra and S. Ventura. Predicting student grades in learning management systems with multiple instance genetic programming. *International Working Group on Educational Data Mining*, 2009.

[24] J. Zimmermann, K. H. Brodersen, J.-P. Pellet, E. August, and J. M. Buhmann. Predicting graduate-level performance from undergraduate achievements. In *4th International Educational Data Mining Conference (EDM11)*, pages 357–358, 2011.