Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Use of Professor Comments in Predicting Student Success

Timothy H. BELL[a*], Christel DARTIGUES[b], Florent JAILLET[c] & Christophe GENOLINI[d]
[a, b, c]*Université Côte D'Azur, CNRS, I3S, France*
[d]*Zébrys, France*
*timothy.bell@univ-cotedazur.fr

Abstract: During their studies students receive written notes and comments from their professors assessing their grades, attitudes, qualities, and lacuna. These characteristics reflect a more subjective approach as opposed to the typical grading system. This paper, through topic modelling and word vectorization approaches, uses textual data to predict at-risk students in their first year of university studies with a Random Forest model. First, we introduce the used methods and analyze the corpus at hand. Then we vectorize the data (by Latent Dirichlet Allocation and other vectorizing methods) to categorize it and use it in the classifier. We then propose adding a dynamic element to the prediction through linear regression when using our data as a time series. Finally, we will review the prediction accuracy and feature importance to assert if these professor comments do indeed reflect the student's scholar capacities. After comparing with the raw numerical grade data, we have better or as-good-as results by using our augmented textual data.

Keywords: Student Success, Random Forest, Allocation, Text Prediction, Time Series

1. Introduction

In France 60% of students in their first year of higher education fail the school year as seen in the MESR-SIES report (2013). Recently more and more research attack this problematic (Alyahyan & Dustegor, 2020). During the high school years, students get graded, and receive comments and notes throughout their studies. The latter, these short texts, allow for a better understanding regarding the student's grades and behavior. We place ourselves in a context where the student is accepted in a university and the teachers want to avoid failure of said student. In the community of educational mining and when predicting student success or detecting at-risk students, most studies mostly use previous grades and input them in Machine Learning models. Numerous works have shown good results (Alyahyan & Dustegor, 2020) especially with Random Forests and more generally ensemble methods (Vijayalakshmi & Venkatachalapathy, 2019; Balaji et al., 2021; Unal, 2020). This paper shows the potential of using Random Forests with transformed textual data.

There are many ways to transform our textual data into numeric vectors, we'll go through some methods relevant to the data used in this work. For a better prediction model, some of the data will be derived to create new variables, by combinations, linear regressions, and frequency differences.

2. Related Works

The work related to student success prediction is very extensive and has always been a pillar of research in educational data mining. The common prediction method is to use the student grades from high school or sociodemographic indicators. Other articles also take in consideration the student's interests and hobbies. A recent extensive state-of-the-art can be seen in Albreiki et al. (2021). Some papers such as Li et al. (2020) and Bell et al. (2021) show that data augmentation can improve prediction. Augmenting the data can be done by rearranging the data or performing various calculations on the data to obtain new features that might be more relevant for the model.

Regarding the use of textual data for student success prediction, there is little research done. The only work, to our knowledge, can be seen in Fateen et al. (2021) and Jayaraman (2020), the first using data from a cram school and the latter using advisor comments for student dropout prediction. Other studies do not use non-student written notes. The aim of this paper is to add to this near-empty research scope in educational research.

3. Dataset

The samples for our data are taken from 2 French university departments (Marketing and Management), we have data from their last 2 years before entering their university curriculum. All the data used in this paper has been entirely anonymized and sensitive data has been omitted. Our target data is whether the students passed or failed their first semester of university. Additionally, we also have absenteeism reported during the first semester. Having binary target data, we are therefore solving a binary supervised classification problem. During high school, students receive grades and professor comments in each subject. Our dataset has a comment/grade pair for each trimester and each year, this work focuses on the last 5 trimesters of high school. The average number of professor notes per student is 50. The data is very short with an average text length of approximately 7 words, the texts are very straight-to-point and often follow the same structure. First commenting on the success or failure of the trimester and then sometimes adding the student's attitude and/or sometimes giving advice. Some examples of sentences are "Unsatisfactory results this trimester" or "okay overall, can still progress" (in this article all the sentences and words will be translated into English).
After removing stop words, we have 3200 unique words in the dataset. Table 1 displays the top words, with "trimester" the most frequent, this word alone doesn't help us (the reader) or the algorithm. So, we also create k-grams up to 3 as we find sentences such as "very good trimester". In Table 1, the top k-grams are also shown revealing terms such as: "very good trimester" and "fairly good trimester". Some k-grams might seem a bit odd such as "trimester good student", this is because stop-words and others are removed during the preprocessing operations.

Table 1. *Top k-grams in the Corpus.*

1-grams	2-grams	3-grams
trimester	good trimester	very good trimester
work	very good	fairly good trimester
results	individual work	trimester good student

Table 2. *Top 1-grams per Class.*

Class	Top 3 1-grams	Class Frequency Difference
Failed	trimester	0.0001
	work	0.00068
	result	0.00054
Passed	trimester	-0.0001
	good	0.00082
	serious	0.00041

If we divide into the passing and failing classes, the top 3 terms for each class are seen in Table 2 with the class frequency difference being simply $\Delta f_w = \frac{f_w}{|C_D|} - \frac{f_w}{|C_{D'}|}$, where f is the mathematical frequency, w is a word or k-gram, C is the corpus. D is a subset of samples, and D' is the subset of samples obtained from the set subtraction $C - C_D$, D corresponds to the current class, for instance the failing class may be the subset D and the succeeding class the subset D'. Due to the size of the corpus, there is variability in the vocabulary, it is therefore expected to have word frequencies close to 0.

4. Methods

The aim of this paper is to focus on the "thumbs up, thumbs down" aspect of these texts and make them as explicit as possible to feed our model. A simple way to pick through our data is by counting occurrences of certain words, words selected from the word rankings of each class. For instance, counting the number of times a sample student has the word "disruptive". So, a student that has the word "disruptive" appear many times will have more probability of belonging to a certain class if this word is decisive in any manner. But more complex methods exist, the next 2 Sections go through various operators that take transformed text features and create scalar values that describe the features' class.

1.1 Latent Dirichlet Allocation

LDA (Blei et al., 2003) is an automatic topic modelling algorithm, often used in NLP, that attributes a topic score for each word. In Figure 1, after preprocessing the raw data in (a) and obtaining vectors of k-grams in (b) for each document (a document here is one single comment or teacher note), we input all the data into an LDA model, and we obtain in (c) a topic distribution matrix containing the k-grams in their probable topic. LDA is found suitable in our case, as we have short sentences, limited vocabulary (teachers tend to reuse the same words), and words that will often appear in similar word contexts and orders of words (obtaining frequent k-grams). The typical way of classifying our documents is using the per-document-per-topic probability γ value of each document (also called Θ by Blei) and each topic.

We explore two other ways to use the values obtained from LDA. Firstly, where v_w is the topic probability vector obtained for each k-gram w in a document c, each value of the vector is the probability of the k-gram to be associated with a topic. A document can be represented as a set of vectors $c'' = v_1, .., v_n$ for n number of k-grams in the document. To reduce the dimensions of c'' to be able to input into our predictive model, we use $c''' = p(c'')$. With p the pointwise product of the k-gram vectors that return a vector the same dimension as v (of size n the inferred number of topics). The resulting vector c''' is split in t_n number of features, t_n is the number of inferred topics. This is illustrated simply in Figure 1 with (d) the set of vectors c'' (or matrix) of size (number of topics in LDA model) ∗ (number of k-grams in document). (e) are the newly obtained features.

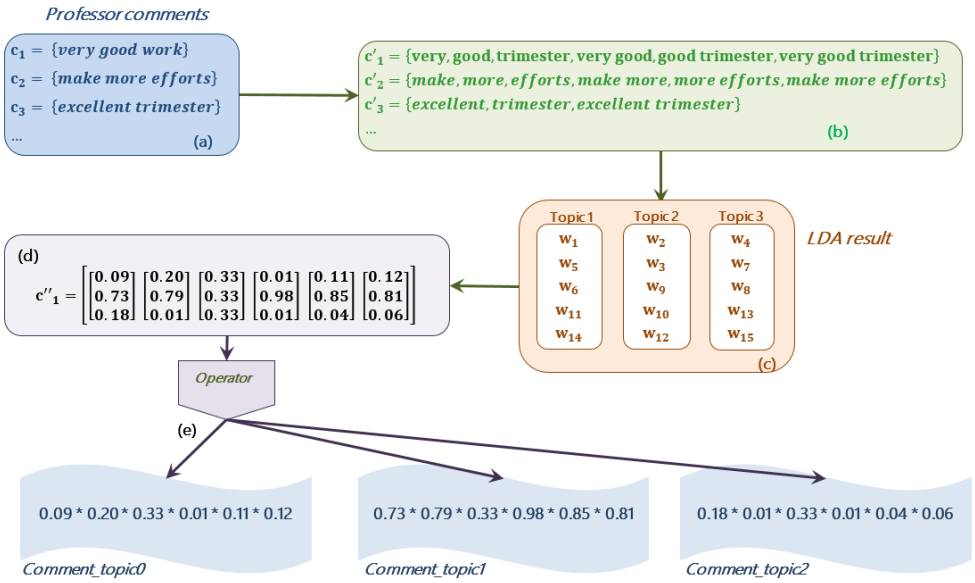

Figure 1. Summary of the RF Models Creation.

Additionally, to create a wider gap between the topics, so that our predictive model can better split depending on how strongly the vectors reflect the association of the k-grams with certain topics,

we perform a pointwise product on the vectors. The resulting scalar is used as a coefficient on indicator functions for each topic. Obtaining a set of features: $\{I(T(c) = t_1).c'''1, \dots, I(T(c) = t_n).c_n'''\}$ with n the number of inferred topics, c_n''' elements of vector c''', and the indicator function using the γ values for each topic to establish whether the class vector is the main document topic.

Another method of transforming our textual data is to take the results obtained in Table 2 and follow the same process as in Figure 1. Therefore, each sample feature would become a set of two features $A_+(c) = \prod_{w \in c} \Delta f_w\,,\ if\ \Delta f_w > 0$ and $A_-(c) = \prod_{w \in c} \Delta f_w\,,\ if\ \Delta f_w < 0$ where *c* is a document and *w* the k-grams of the document.

1.2 Dynamic Aspect

As found in Bell et al. (2021) and Li et al. (2020) we can find relevant features for an accurate prediction by taking account of the progress of a student over time. By doing linear regression on the values throughout the trimesters we obtain the regression coefficient β of the evolution of the student. As it may be more relevant to look at how the student improves at specific combinations of trimesters. For each subject (e.g., Mathematics), we would multiply the number of each feature by $\sum_{i=2}^{5} \binom{5}{i}$ extending our feature space by 26 for the yearly professor notes and (number of subjects) $*$ 26. In our dataset we have 9 different subjects that carry information through the 2 years. The data will therefore get 260 newly added features for each method of transforming the documents into numerical data. For the increase coefficient over the span of trimesters 1,4, and 5, we will note $\beta_{1,4,5}$.

1.3 Random Forest

We chose Random Forests (RFs) (Breiman, 2001) as it has proven its use in student success prediction (Zhang et al., 2021; Zeineddine et al., 2021). RFs are an ensemble method consisting of multiple decision trees. Each decision tree is trained on a random collection of samples as well as a random collection of features. As we have limited samples, we will cross-validate 10 RFs. We use accuracy and the F1-score to compare models. Each RF model is trained on detecting the passing class and the failing class. In our RF model we use the Gini index (Zhu et al., 2014) for the tree splits. Having this index gives us a built-in feature importance model. In our results section we will discuss the top features selected when training the model, helping us compare the advantages of adding certain sets of features.

We will create RFs for each feature group, and also one RF combining all feature groups, and then we'll compare the results with each other and with a prediction using only the numerical data. The compared groups are the k-gram occurrence, the γ value per-document-per-topic, the pointwise product of the LDA vectors of each document, the pointwise product as a coefficient, the class frequency product, the combination of all feature groups.

5. Results

After trialing, we decided that 3 topics should be informed to our LDA model. Dividing our word bank into a first category with k-grams like "Good Trimester" and "Good", another with words like "Difficult" and "Barely", and finally, "Suitable" and "Suitable Results" for the last category. With all the augmentation done, we obtain a total of approximately 6000 features, mostly features discussed in 4.2. Most of the features are irrelevant to our prediction and do not help in splitting the nodes in our RF. Ideally, work done in this paper should encourage feature selection to get rid of useless features.

1.4 Performance Review

LDA topic modelling works quite well with this data, probably since the texts are quite short and contain little noise. Also, the teacher wants to convey positive or negative assessments in a clear-cut fashion (e.g., "Serious student. Good grades."). Therefore, when looking at the topics as a matrix (illustrated in

(e) in Figure 1) we can see that the "bad" topic contains k-grams such as "need work", "bad", etc. But some k-grams may not be placed in the "good" topic, such as "serious overall" or "satisfying good work" which are put into the "neutral" topic.

In Table 3 we display accuracies for the augmented data groups, the group with the lesser performance is counting element occurrences, with an accuracy quasi-identical to the passing rate (56%). The Class Frequency Product doesn't perform much better either. The rest perform similarly and show sufficient results in our use case, additionally, when using all feature groups in a single predictive model we obtain good results.

Table 3. *Performance of feature groups.*

	Accuracy	Failing class f1-score	Passing class f1-score
Element occurrence	0.58	0.39	0.76
LDA γ value	0.65	0.36	0.82
LDA pointwise product	0.62	0.35	0.80
LDA coefficient pointwise product	0.65	0.41	0.78
Class frequency product	0.60	0.35	0.75
All combined	0.69	0.53	0.86

Predicting with only grades achieves an accuracy of 70% in the marketing department and 73% in the management department. The difference between this accuracy and what we get with our model (using the best accuracy) is 1% only for the marketing department and 6% for the other, both in favor of our model. This big difference may simply be since marketing in this specific establishment may be more demanding (only 56% pass).

1.5 Most Decisive Features

Ultimately, we are predicting at-risk students in their first year of university. Knowing what features describe this "failing" class of students is necessary. The best predictor for the marketing class is the β value of trimesters 3, 4, and 5 of the pointwise vector multiplication of the LDA vectors. As a note, in referral to the pointwise product illustrated in Figure 1, we tried different calculations (non-extensively) for step (e) and found that only the product can generate useful features. The k-gram occurrence counts only showed one feature present in the top 100 predictive features and it has a low importance score. Among the different subsets of data in our dataset, element occurrence counts are never in the top splitting. In the top 20 features for each department, 18% of features are LDA per-document-per-topic probabilities, 31% are coefficient based LDA pointwise products, 14% simple LDA pointwise products, the rest are mainly β values with combinations of trimesters. Less than 2% is of the class frequency product type.

A note on absenteeism. When using the model to predict the absenteeism obtain good results, for a proportion of 17% of reported absent students we have an 89% percent accuracy in prediction. When using the purely numerical data, this characteristic is unpredictable. The advantage of using teacher notes and comments for predicting things like absenteeism is quite clear. For this prediction, the occurrence of k-grams relating to "absent" or "disruptive" are features that appear in the top 20 most decisive features. The β value of these k-gram occurrences do not appear however to be decisive in any split of our RF model, this can be explained by the low presence of the said k-grams in some trimesters over the years, and therefore no trend can be extrapolated through linear regression.

When comparing with the raw grade prediction or other predictions from papers mentioned throughout this article, the accuracy of using only textual data shows that we obtain similar or sometimes better results. Attesting the not-so-subjective characteristics of theses short texts.

6. Conclusion

Over the last years, student success prediction has been a core field of study for the community. Typically, grades, and other meta data are used to predict student success, some ideas on possible

predictors can be found in Bell et al. (2021) and Zhao et al. (2020). In this paper we used comments and notes that describe the student's behavior among other things. These short texts carry information that cannot be conveyed through the previously mentioned features, they allow the teachers to gain more insight into the behavioral aspects of students.

Our goal in this study was to explore if written comments and notes from teachers can give out more information than using grades, as we believe that as well as discussing student grades and quality of work, these short texts carry information on other elements such as behavior or absenteeism. In this paper we found that using the textual data can carry more relevant information in the prediction model, as we see that augmenting the data by calculations or a dynamic approach, results in slight improvement in prediction. We also saw that predicting absenteeism has shown promising results, as information regarding this can only be extracted through written notes as in most datasets used for predicting student performance there are no metrics relating to absenteeism. Understanding information like this can be used as a tool for teachers to anticipate student needs as soon as the first semester of higher education.

In future works, we firstly hope to explore different methods of using the textual data in the hope of achieving better performance and then find ways to include numerical data to the model to get a more accurate prediction in the end. We would also like to—in the same way as absenteeism—predict "implication" of a student during his classes. Another way to further extend this study would be to perform sentiment analysis on the comments which may give us a different dimension to the text.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences, 11(9*).

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17(1)*, 3.

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences, 11(21)*.

Bell, T., Dartigues-Pallez, C., Jaillet, F., & Genolini, C. (2021). Data Augmentation for Enlarging Student Feature Space and Improving Random Forest Success Prediction. In *Artificial Intelligence in Education: 22nd International Conference*, Proceedings, Part II. 82–87. Springer-Verlag.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res., 3*, 993–1022.

Breiman, L. (2001). Random Forests. *Mach. Learn., 45(1)*, 5–32.

Fateen, M., & Ueno, K., & Mine, T. (2021). An Improved Model to Predict Student Performance using Teacher Observation Reports. In *Proceedings 29th International Conference on Computers in Education Conference*. 31-40.

Jayaraman, J. (2020). "Predicting student dropout by mining advisor notes," in *Proceedings of The 13th International Conference on Educational Data Mining (Ifrane),* 629–632.

Li, H., Ding, W., Yang, S., & Liu, Z. (2020). Identifying at-risk K-12 students in multimodal online environments: A machine learning approach. In *Proceedings of The 13th International Conference on Educational Data Mining*. 137–147.

MESR-SIES. Réussite et échec en premier cycle. Note d'Information Enseignement supérieur Recherche. (2013). Retrieved from https://www.enseignementsup- recherche.gouv.fr/sites/default/files/ imported_files/documents/NI_MESR_13_10_283447.pdf.

Unal, F. (2020). Data mining for student performance prediction in education. In *Data Mining, chapter 9*. IntechOpen, Rijeka.

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of Predicting Student's Performance using Machine Learning Algorithms. *International Journal of Intelligent Systems and Applications*, 11, 34-45.

Zeineddine, H., Braendle, U.C., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Comput. Electr. Eng.*, 89, 106903.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, 12.

Zhao, Y., Xu, Q., Chen, M., & Weiss, G. (2020). Predicting Student Performance in a Master of Data Science Program Using Admissions Data. In *Proceedings of The 13th International Conference on Educational Data Mining*. 325-333.

Zhu, W., Feng, J., & Lin, Y. (2014). Using Gini-Index for Feature Selection in Text Categorization. In *Proceedings of the 2014 International Conference on Information, Business and Education Technology*. 76-80. Atlantis Press.