Learning Analytics in Further and Higher Education

Jade Lyons[1], Kevin Curran[2]
Northwest College of Further and Higher Education[1]
School of computing and Intelligent Systems, Ulster University[2]
Northern Ireland

Abstract

Student retention is a problem faced by all education institutions. We demonstrate how we applied machine learning techniques to data about students, their courses, attendance records and whether they dropped out. We highlight how we used data analysis and visualization techniques as predictors for student drop out. We found that Random forest shows good results but specificity is low. More data would probably resolve this issue. Although neural networks has next greatest accuracy, k-NN's results over the different descriptors are reasonable. Specificity is quite low over all models. More work would need to be carried out to achieve higher accuracy. As the data was so imbalanced, it would be better to have a bigger dataset so oversampling would not be needed. This would help to achieve better results.

1. Introduction

As we have been living in the digital age for years now, large amounts of digital data has been accrued by schools and universities about students and the potential of this data is now beginning to be realized. Various government and education initiatives exist that are encouraging schools and colleges to make effective use of the data that is being held about their students, to enhance the student experience, identify 'at risk' students before they might leave and therefore improve retention. Applying data analysis techniques can help schools create strategies that will help them engage those students who may otherwise have dropped out, as well as making the education experience better for all students [1,2].

There are a great number of research papers available that have looked at how retention can be improved and if there are any specific predictors that make it more likely that students will drop out [3,4]. There are a few similar characteristics that seem to run through a lot of this research. John Bean, who has studied retention extensively since the 80's and has developed a psychological model of retention, suggests that factors in a student's background can affect whether they are likely to drop out or not [5]. Environmental factors and what the student's intentions are, also play an important role as indicators of a student's commitment to a course of study. Assuming that this is an accurate theory, then each individual college will be looking at their specific factors that have an impact on student retention. Crucial to guiding the development and implementation of measures to improve student retention at an institution is an understanding of the factors that influence retention at that institution [6,7].

The term Learning Analytics has been applied to the analysis of data held about students. Learning Analytics is "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [8].

They have published several documents about learning analytics and provided online learning resources to help schools who want to start making use of their student data to enhance their own teaching and learning [4]. In a wider context other than improving retention, the National Forum have stated that Learning Analytics can be used to enhance teaching and learning in a variety of ways:

- It can let teachers know which resources their students are using and how active they are.
- It can let students know how engaged they are with course material, relative to their peers.
- Real-time information can give both teachers and students the opportunity to take timely, informed action, as appropriate.
- It can inform the curriculum and programme design.
- It can identify patterns of activity that are most likely to engender deep learning and have a successful outcome for the student.
- It can identify at-risk students and empower them to change their academic trajectory before they suffer any negative consequences.
- It can identify and prescribe actions and resources that are most likely to yield a favourable outcome for students.
- It can be used to identify students with sudden changes in engagement that can be indicative of a wide range of non-academic issues. By identifying students that may be facing personal, emotional, medical, social or financial challenges, LA can help support staff to proactively intervene and provide relevant, targeted supports to students with the greatest need [9].

Learning Analytics can be divided into three levels: descriptive, predictive and prescriptive [10, 11, 12].

Purdue University used a predictive model and developed an early alert system that is an algorithm that processes the data about each student and uses a traffic light system to alert the teacher of 'at risk' students as well as alerting the student that they may experience difficulties completing that module. We look at the student data that is held by the North West Regional College over a three-year period.

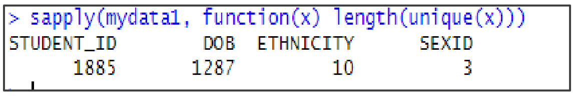
2. The Data

The North West Regional College has provided seven csv files with a variety of data relating to students and courses. This data has been harvested from a few different MIS systems within the college. The files, attributes and number of observations are as follows:

- Student.csv: STUDENT_ID, DOB, ETHNICITY, SEXID, LAST_NAME, TUTOR_STAFF_ID, MILESTOCOLLEGE (1890 observations)
- Course.csv: COURSE_ID, SUBJECT, TITLE, COURSE_AIM, INST_TIER1, INST_TIER2, INST_TIER3, TENANT_ID, AWARDING_BODY (68 observations)
- Course_instance.csv: COURSE_INSTANCE_ID, COURSE_ID, START_DATE, END_DATE, ACADEMIC_YEAR (170 observations)
- Staff.csv: STAFF_ID, FIRST_NAME, LAST_NAME, TITLE (182 observations)
- Staff_course_instance.csv: STAFF_ON_COURSE_INSTANCE_ID, STAFF_ID, COURSE_INSTANCE_ID (1039 observations)
- Student_course_membership.csv: STUDENT_COURSE_MEMBERSHIP_ID, STUDENT_ID, COURSE_INSTANCE_ID, WITHDRAWAL_REASON, COURSE_OUTCOME, COURSE_EXPECTED_END_DATE, COURSE_JOIN_DATE, COURSE_JOIN_DATE, COURSE_END_DATE (1890 observations)
- Attendance.csv: EVENT_ID, STUDENT_ID, EVENT_NAME, START_TIME, END_TIME, EVENT_ATTENDED, COURSE_INSTANCE_ID (799995 observations)

At first glance there appears to be relationships between each of the csv files as there appear to be primary and foreign key fields throughout the files. As there is a lot of information here, each file was looked at in turn to identify relationships and to decide which files will be used in the analysis. For the purposes of this analysis the following files were selected as the most use for determining student drop out: student, attendance and student_course_membership. These files can be joined on STUDENT_ID.

The student data frame was examined first, to check how many unique instances of each attribute were present. Interestingly, although the student data frame showed that it had 1890 observations which would suggest that there were 1890 individual students' records, there were only 1885 unique observations:



```
> sapply(mydata1, function(x) length(unique(x)))
STUDENT_ID        DOB  ETHNICITY      SEXID
      1885       1287         10          3
```

Figure 1. Unique observations in student data

There were five duplicates that were removed. Even if a student has completed more than one course they should only have one student record in this file. All the attributes of the student data frame should be useful in the analysis. STUDENT_ID will be used as the key to join the three selected files together. DOB is useful at this stage as it can provide a student's age. SEXID is useful as knowing whether more males or females drop out is interesting. MILESTOCOLLEGE will be a useful predictor as there may be a relationship between the distances that people have to travel and whether they are more likely to drop out or not. ETHNICITY may be useful if only to show if there is much diversity at the college. Figure 2 shows the number of each ethnicity of the students in the student file.

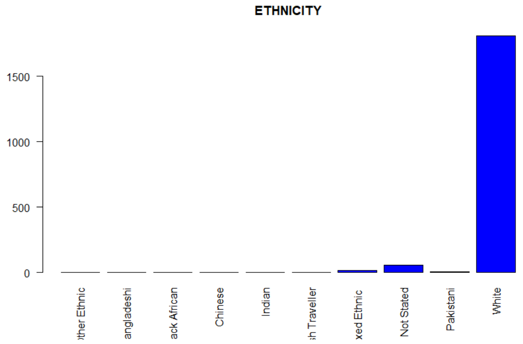


Figure 2. No of each ethnicity

From this plot it is obvious that students are mostly identifying as white. As there is such a clear difference in the number of white students compared to other ethnicities, this will probably not make a good predictor for the model.

The attendance file has a number of very useful attributes for this analysis. Every student's attendance at every class within a course has been recorded as either 'Y' or 'N'. Using this data, two new columns of data were created, ABSENCES and CLASSESATTENDED. These represent the total number of absences and classes attended for each student. These new attributes were then added to the student file, joined on STUDENT_ID. The data frame now has 3503 observations. Student_course_membership was examined in the same way and five duplicate student observations were removed which resulted in 1886 observations. Student and student_course_membership were linked via the STUDENT_ID, this resulted in 3503 observations and 16 attributes in the new data frame. There were 8 attributes that were removed from the data frame as they weren't useful as predictors. The COURSE_OUTCOME attribute reflected the WITHDRAWAL_REASON attribute with a different code that is used internally for management. This attribute was not needed for this analysis and was removed. There were other attributes that weren't useful for this analysis as they had no value as predictors, they were: COURSE_INSTANCE_ID, COURSE_JOINE_DATE (all the same each year), DOB (age observation exists), COURSE_EXPECTED_END_DATE, COURSE_END_DATE (these are the same for all courses) and STAFF_ON_COURSE_INSTANCE_ID as this is made up of COURSE and STAFF_ID.

The other files were discarded at this time for a number of reasons. They held duplicate data and a lot of the STUDENT_ID's didn't match the IDs in the student data frame. The COURSE_INSTANCE_ID which occurs in a few of the files did not have the same data format in each file so was of no use when trying to join the different files.

There is now one data frame made up of student, student course membership and attendance. The STUDENT_COURSE_MEMBERSHIP_ID attribute's data is made up of the student id and a course id:

FDFS3WELFR_ABB17107726

The course id (before underscore above) is useful as a predictor as students may drop out of certain types of courses more than others. The student id part of the data is no use and was removed, leaving each row of data as only the course id. This new attribute was called 'Course'. Figure 3 shows the attributes of the new merged data frame and the number of missing values (3503 observations):

Attribute	NA no
ETHNICITY	1618
SEXID	1618
Course	1618
WITHDRAWAL_REASON	1618
MILESTOCOLLEGE	1618
COURSE_JOIN_AGE	1618
CLASSESATTENDED	1038
ABSENCES	1038

Figure 3. Missing data in data set

There are a total of 3503 observations in the data frame and they account for nearly half of the data. Because there are so many, it was decided it wouldn't be appropriate to impute values because it could result in inaccurate results when building the models. For this reason the observations with missing values were removed leaving 847 complete observations which, although a lot smaller, will have a lot more integrity when building the models.

Figure 4 is a table plot that gives an overall view of the data that will be used in the analysis.

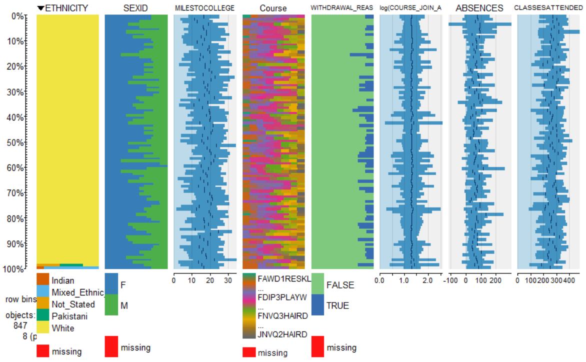


Figure 4. Table plot of dataset

From this plot it is obvious that the most popular ethnicity in this data frame is white. The plot also shows that there is an even spread of male and female students. It also doesn't look like anyone lives extremely far away from the college either. The withdrawal reason column shows that in this data frame, most of the students stayed until the end of the course. The WITHDRAWAL_REASON attribute's data may contain one of 37 codes that represent the reason why students leave, ranging from C01, 'college terminated because of attendance' to U01,'unknown reason'. The codes are only present if a student has left the college. Figure 5 shows which withdrawal code was used the most often for students leaving.

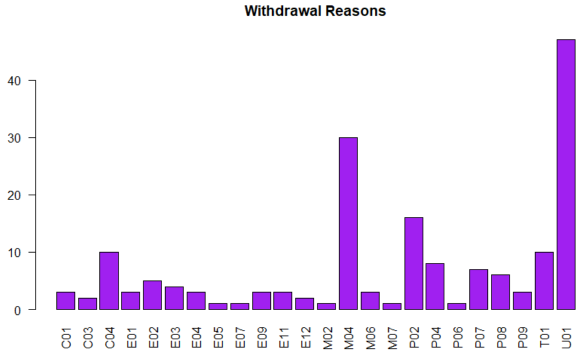


Figure 5. Total number of each withdrawal reason

The most common reason for students leaving was U01 which is 'unknown', followed by M04, 'gone into employment', PO2, 'family/personal issues' and C04, 'college terminated attendance unable to make contact'.

There are other plots that can be looked at to observe any relationships that may exist within the data. Figure 6 shows age and no of absences of the students.