Innovative Learning Analytics Research at a Data-Driven HEI

Azcona, David[a]; Corrigan, Owen[a]; Scanlon, Philip[a] and Smeaton, Alan F.[a]

[a]Insight Centre for Data Analytics, Dublin City University, Ireland.

Abstract

A university campus is comprised of Schools and Faculties attended by students whose primary intention is to learn and ultimately graduate with their desired qualification. From the moment students apply to a university and thereafter gain acceptance and attend the campus they create a unique digital footprint of themselves within the university IT systems. Students' digital footprints are a source of data that is of interest to groups including teachers, analysts, administrators and policy makers in the education, sociology, and pedagogy domains. Learning analytics can offer tools to mine such data producing actionable knowledge for purposes of improving student retention, curriculum enhancement, student progress and feedback, and administrative evolution. In this paper, we summarise three ongoing Learning Analytics projects from an Irish university, demonstrating the potential that exists to enhance Higher Education pedagogical approaches. First year students often struggle with making the transition into University as they adapt to life and study at a Higher Education Institution. The research projects in the area of Learning Analytics at our institution focus on: improving test performance using analytics from a general-purpose VLE like Moodle, identifying studying groups and the performance peer effect using on-campus geolocation data, and detecting lower-performing or at-risk students on programming modules.

Keywords: *Learning Analytics; Personalised Learning; VLE; Student Retention; Early Intervention; Data Mining.*

1. Introduction

From the moment a student applies to a University, each becomes a continuous source of data, especially relating to their activities on the campus. Each student can be defined by their own unique digital footprint which comprises of a number of components including:

- Demographic information
- Previous academic history
- Assignment and exam results
- Library attendance and book withdrawals
- Activity on University Virtual Learning Environments (VLEs)
- Use of ICT (Information Communication Technology) resources
- Access to WiFi or other networking systems
- Math Learning Centre and other drop-in centre records
- Class attendance, where measured

Learning Analytics in the context of our research projects can be defined as the collection and mining of a student's unique digital footprint to understand their interactions with each other and with the University systems. Siemens and Long (2011) advocate that Learning Analytics provides various degrees of actionable intelligence for learners, Faculties, course administrators and decision takers at departmental levels. We believe that the value goes further and can extend to students' external supports, including families.

The following three sections outline three individual and unique Learning Analytics projects, each using students' digital footprints in some way to positively benefit students.

2. Predictive Educational Analytics Using the University's VLE Logs

Research by Calvo-Flores (2006) has shown that student grades can be predicted with a good level of accuracy using features derived from the activity logs from a Learning Management System. The challenge that we address, is how to generate predictions of student grades in first year undergraduate classes, and how to use these predictions to help students achieve better results. Our solution does not require any input or interaction from Lecturers as instead, we provide direct feedback in terms of performance prediction and in recommending content to study, to the students themselves. Our solution was implemented at Dublin City University across a wide range of first year undergraduate modules over a two-year period. Analysis of the examination results for students from the first year of this study shows an increase in test scores for the majority of modules, that this was implemented for with some modules such as Mathematics for Business, achieving very significant improvement when comparing students who opted into receiving this feedback.

Overall, our work demonstrates that students respond well to automated interventions and that this does help to improve overall student performance and thus student retention.

We found that depending on the module type, and how far into the module content that we were (i.e how many weeks of semester), we could predict student exam outcome based on a cross-validated Area Under the Receiver Operating Characteristic (ROC AUC) as a performance metric for our binary functions. The ROC AUC score for modules we used in our work varied widely between modules. We found that the modules which were best suited to this sort of analysis had an ROC AUC score of 60% to 70% by the end of the semester. We also validated our predictions on a module, made while the module was still in progress as against the results achieved by the students. The F1 score, another evaluation metric which is the harmonic mean of precision and recall, for the prediction made in the final week was generally was between 70% and 90%, again depending on the module. This work was implemented across a cohort of 1,095 first year undergraduate students, of which 725 students chose to opt in to receiving this weekly feedback. We also observed a 3.5% improvement in actual exam mark between those who opted into the study and those who chose not to. (Corrigan (2015)). Once it is demonstrated that it is possible to predict students' examination grades with a reasonable level of accuracy, it is important to use this to help students in some way. Our overall goal is to use these predictions to assist students in improving examination performance, improve their learning experience and thus to improve student retention in University.

Our solution to the prediction challenge, is to build a model based on historical data. Test this solution using cross-validation to ensure that the predictions accuracy and then, using a regression model, we predict scores that the students are likely to obtain. We then repeat this process for every week of the semester. Once we predicts the grades that each student will likely receive, we developed a web application where students could check their own exam predictions weekly. This is presented as a chart showing how well the system believes each student is likely to perform relative to their peers as the prediction shown to students is their predicted place in class, in deciles. A sample of the personalised graph for just one module that students can visualise is shown in Figure 1. This graph illustrates that in week 12 of the semester (x-axis), this particular student in module SS103 was predicted to come only within the top 90% of the class (y-axis) and their whole semester shows a gradual decline in relative placing throughout the 12 weeks.

In addition to predicting place in class, we also sent tailored weekly alerts which included specific recommended resources taken directly from the LMS content for the course module that was deem to be helpful to the student. This is based on a case based recommender which uses data about online LMS 4 resources used by peers from the class who have a week-on-week prediction of a higher place in class, with a specific focus on LMS content that the student has not yet viewed.

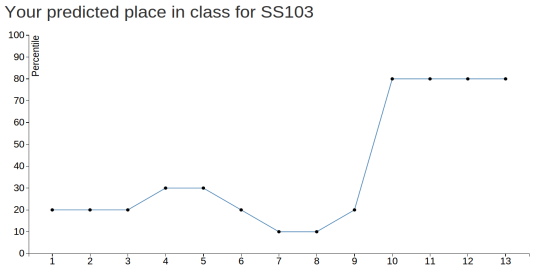

Figure 1: Week-by-week prediction for one student in one module. Source: Dublin City University (2016)

3. Wifi Geolocation Data, Student Groups and the Peer Effect

There has been a multitude of studies and research in the areas of the influence on an individual, of the heterogeneous social groups to which they become members. Within the educational domain the research emphasis is on the impact of peers on the academic performance of a student. Much of this research is reliant on data collection methods that are invasive, require participants' recall or are obvious to the subjects that they are part of a study. Recent developments in the area of formative educational methods have enabled other data collection options. Digitally collected data such as student digital footprints, has many advantages over that compiled manually through more traditional collection methods. Data-sets collected digitally are less susceptible to the inherent biases introduced through the intervention of human interpretations, furthermore it is often structured, complete and traceable. When a student registers with the university they are allocated unique credentials for use with the campus IT assets. As a student uses these IT assets they lay down a unique digital footprint.

This project examines an aspect of the digital footprint generated by students through the interaction of their wifi-enabled devices and the campus-wide Eduroam system. Using historical logs this project is carrying out a longitudinal analysis of the interactions among individual students and thus identifying group memberships. Once a group's members are identified, we correlate academic performance between group members and analyse how interaction with their peers affects their academic performance.

We identify which students interact on a continuous basis both within the framework of the scheduled lectures and labs and in areas where gatherings are more social, or are in shared studying locations. To ensure our research differentiates between the context of meeting places we subdivided the campus into academic (formal) and social (informal) locations.

Our hypothesis is that individuals become members of a number of emergent groups in the early stages of interaction before forming "friendships" within groups. We are further

exploring whether groups they become members of will influence academic performance. Figure 2 illustrates a wifi platform digital footprint captured between September and December 2014. This of second year students registered in one of the School of Computing's programs. It presents a pattern identifying Tuesday having a high level of wifi activity and Friday the lowest. This activity correlates with the programs academic timetable, i.e. there are no formal classes scheduled on Fridays. This figure also illustrates that as the semester progresses students spend more time on campus on Fridays, presumably in additional studies either individually or as part of a study group.

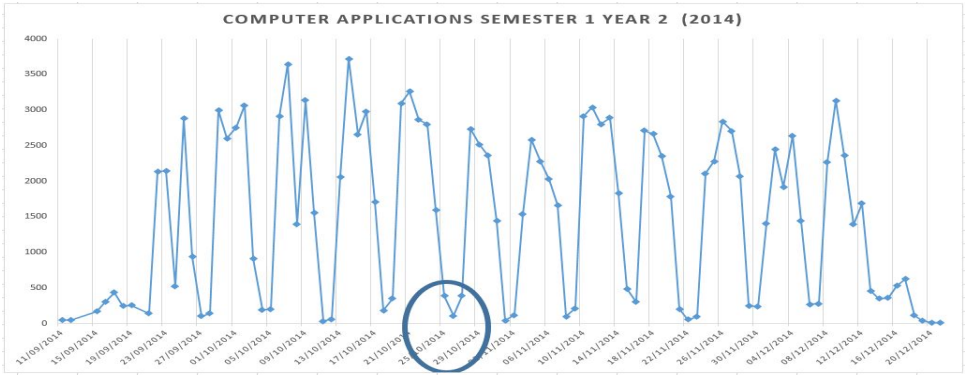

Figure 2.Wifi digital footprint. Source: SAP Hana and Predictive Analytics (2014).

Having identified each student's unique interaction with the wifi platform we can use a co-location algorithm to identify groups based on time, location and the number of their dyad interactions. Table 1 provides an example of the number of meetings between student pairs by location and category, i.e Academic or Social. It can be seen that the dyads with the smallest delta mark had both a large number of meetings in both Academic and Social locations. This can be construed as identifying students who spend a large portion of their time together in both Academic and Social setting, as friends and share similar academic abilities.

Student 1	Student 2	Academic Meetings	CLASS	LABS	LIBRARY	Social Meetings	HANG_OUT	TRANSIT	RESIDENCE	Total MEETINGS	DELTA
d399224566b1406f883ff88af289337a	fcd00bacc7327dc417e56d1ebdb67189	2588	1458	1130	0	355	222	127	0	2943	2.67
a878d00ec27f559e67e8071b4f4638ee	d399224566b1406f883ff88af289337a	2682	774	1908	0	217	121	94	2	2899	4.25
53f4fe4a9d759074c22be50786f9ff07	d399224566b1406f883ff88af289337a	2455	796	1659	0	104	72	24	8	2559	16.34
d399224566b1406f883ff88af289337a	f7aa5c504eff400c312b8664be16ea8d	1479	415	1012	52	471	360	87	22	1950	4.92
d399224566b1406f883ff88af289337a	f0381e3ba5f7f6e7c3cbf6ed0f3e3888	1824	476	1348	0	114	59	55	0	1938	9.92
f0381e3ba5f7f6e7c3cbf6ed0f3e3888	f68fda980005140df86bb6429ae0133e	1160	710	320	130	453	373	78	2	1613	0.58
d399224566b1406f883ff88af289337a	f7fb4a52b3bc08ada78be4813b7d9e85	1351	586	765	0	149	58	81	2	1500	15
6ccbcb030e54c7bf1dbf95f275e0570a	d399224566b1406f883ff88af289337a	1376	532	844	0	102	65	37	0	1478	16.84
9b1c28c65fa9cc7a2c877c9168b5ea80	a878d00ec27f559e67e8071b4f4638ee	1106	536	489	81	282	207	36	39	1388	13.75
757e5e772c06bace6801026a805be617	d399224566b1406f883ff88af289337a	1217	389	828	0	61	51	10	0	1278	10.42

Table 1.Student Dyad by location and exam delta.

An accepted methodology in the examination of the interaction of students within groups and communities, is Social Network Analysis (SNA). Wasserman et al. (1994) were one of

the earliest exponents of SNA being utilised to examine the interaction among students. We propose that SNA can be used to identify patterns and relationships amongst individuals, groups and communities within our study. Using SAP Hana and SAP Predictive Analytics Social Network analysis tools we have identified student friendship dyads and are carrying out network analysis between group membership and academic performance of the group(s) members. In the remainder of this project we will correlate group membership with academic performance and identify the optimum group size for shared exercises. We will also measure demographic features and academic performance and carry out feature extraction of students who disengage from the university. This behavioral analysis could allow early identification of students likely to opt out of study.

4. Computer Science and Personalised Programming Learning

We know that students' digital footprints commence prior to their arrival at the university as demographic and GPA (CAO points) are collected at the time of application. We analysed 950 first-year Computer Science (CS) entrants across a seven year period through the Leaving Certificate entry route. Early analysis showed a high correlation between the entry level GPA and first year final exams aggregate as shown in Figure 3.

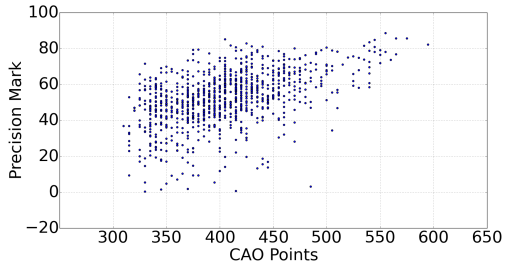

Figure 3. University entry points vs First-year precision mark. Source: Dublin City University (2010-2016).

At Dublin City University, 304 students dropped out over the last six years and 64% of these happened during their first year of studies in CS. We trained a machine learning algorithm to identify first-year CS students who might face difficulties before they even arrive on-campus to start their programme. The function is based on student characteristics including Age, Gender, Nationality, Domicile, Access route to campus, etc; and University Access exam scores: Irish Leaving Cert: GPA (CAO points) and Individual Exam level and results. Student Support and Services at Dublin City University are monitoring our analysis for the incoming cohort of 2017 students. Students who might be at risk based on their characteristics and entry points are being contacted and offered help. The learning algorithm we used is a Random Forest (which consists of multiple Decision

Trees). The training data is very imbalanced throughout the six years (21% of failures, which are 251 students), and this algorithm maximizes the metrics for the negative class.

Table 1. Random Forest metrics

Class	F-score	Precision	Recall
Fail Class	28.70%	30.93%	29.12%
Pass Class	83.36%	82.09%	83.44%

Source: Dublin City University (2010-2016).

This approach, based on static student information, doesn't take into account the student's effort or learning throughout the year and won't yield great insights. Thus, real-time analytics are being gathered from a custom learning environment (VLE) developed for the teaching of programming modules. Students can browse the module's material and submit their laboratory work online. Program submissions are verified in real-time on every upload. Every interaction with the material online and the programs developed along with the results are tracked, forming a digital footprint for each CS student who uses the platform. For each module leveraging the platform, a predictive model has been developed using historical data. The models, developed using machine learning techniques, predict student performance based on programming skills acquired and web interaction events. Students are assessed by taking laboratory computer-based programming exams and the passing likelihood is the target of our predictions. The empirical error minimization approach has been employed to pick up the learning algorithm with the fewest empirical errors from a bag of classifiers for each module analysed.

A learning algorithm was built for "Computer Programming I" in 2016-2017 using only one year of training data. Reports to lecturers was sent every week outlining real-time predictions about the students performance. Students struggling with the module could be identified at a glance plus the confidence of our classifier for each student. The classifier selected was Logistic Regression which also gives an associated probability.

Student	Fail / Pass Prediction	Fail Prediction Confidence	Pass Prediction Confidence
John Doe	Pass	26.95%	73.05%
Jane Roe	Fail	69.27%	30.73%
Johnny Smith	Pass	27.03%	72.97%

Figure 4. Prediction probabilities snapshot . Source: Dublin City University (2016).

At-risk students were targeted during laboratory sessions by lecturers and laboratory tutors. We cluster students in two cohorts depending whether they failed or passed their first

laboratory exam. The learning improvement between these two groups was 11.52% in 2015/2016 and 50.26% in 2016/2017 when assistance was offered to weak students. The differential learning improvement on average was then 4 times more on the academic year the predictions were run and reports sent, than on the previous one, and at-risk which means students learned 4 times more.

Personalised programming recommendations are coming as student feedback the following semester. We are expecting an impact on student engagement and learning on these CS modules.

5. Conclusions and Future work

Learning Analytics gathers student data about their learning progress and this raw data can be processed and turned into actionable knowledge which Faculty administrators and Lecturers can use to identify students having difficulties on courses or at university life in general. We have found that our VLE (Moodle) and custom in-module predictions reach a usable accuracy that increases towards the end of the semester. At our University, we are excited to provide tools that help and encourage student learning right across the campus. In this paper we have shown three examples where some part of students' digital footprints - their access to the VLE, their interaction with the campus WiFi and their interaction with an in-lab programming environment - can be usefully mined for such actionable knowledge.

We need to bear in mind when we are comparing our results between lower and higher-performing students that the former have more margin for improvement and if the latter maintains their academic performance, then this is already an accomplishment. In addition, we are in the process of aggregating different data sources about students, combining their characteristics with in-module analytics or geolocation data to better understand their digital footprint and make more informed decisions.

References

Calvo-Flores, M. D., Galindo, E. G., Jiménez, M. P., & Piñeiro, O. P. (2006). Predicting students' marks from Moodle logs using neural network models. *Current Developments in Technology-Assisted Education*, *1*, 586-590.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*(5), 30.

Corrigan, O., Smeaton, A.F., Glynn, M., & Smyth, S. (2015). Using Educational Analytics to Improve Test Performance. In *Design for Teaching and Learning in a Networked World* (pp. 42-55). Springer International Publishing.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.