Predicting "At Risk" Students from log data

Alan Smeaton, Sinead Smyth, Owen Corrigan,
John Brennan, Aly Egan

Centre for Data Analytics

Insight

Project Aims

Our target is to reliably predict whether a student will drop out. Students interact with many different systems in their time in university. They create a profile of activity usage. It is the aim of this project to use these profiles to predict whether a student will drop out. We then use these predictions to target interventions at the students most at risk.

Data Sources

The information Systems & Systems Department (ISS) in DCU have provided us with a wide variety of data about students. We have access to their demographic data such as location, age, gender, citizenship status, marital status and socio-economic status. We also have access to their leaving certificate results.
We are particularly interested in their Moodle activity for a few reasons. The first is that it is a very strong predictor of student success. From viewing the Moodle access logs on a week by week basis, we can see that for many modules there is a strong periodicity associated with the modules. The second is that we can make predictions for students mid-way through the semester. This allows us to alert students earlier that they are at risk of failing. Finally, it is perhaps more ethical to base our interventions on the actions of students rather than demographic information that they have no control over.

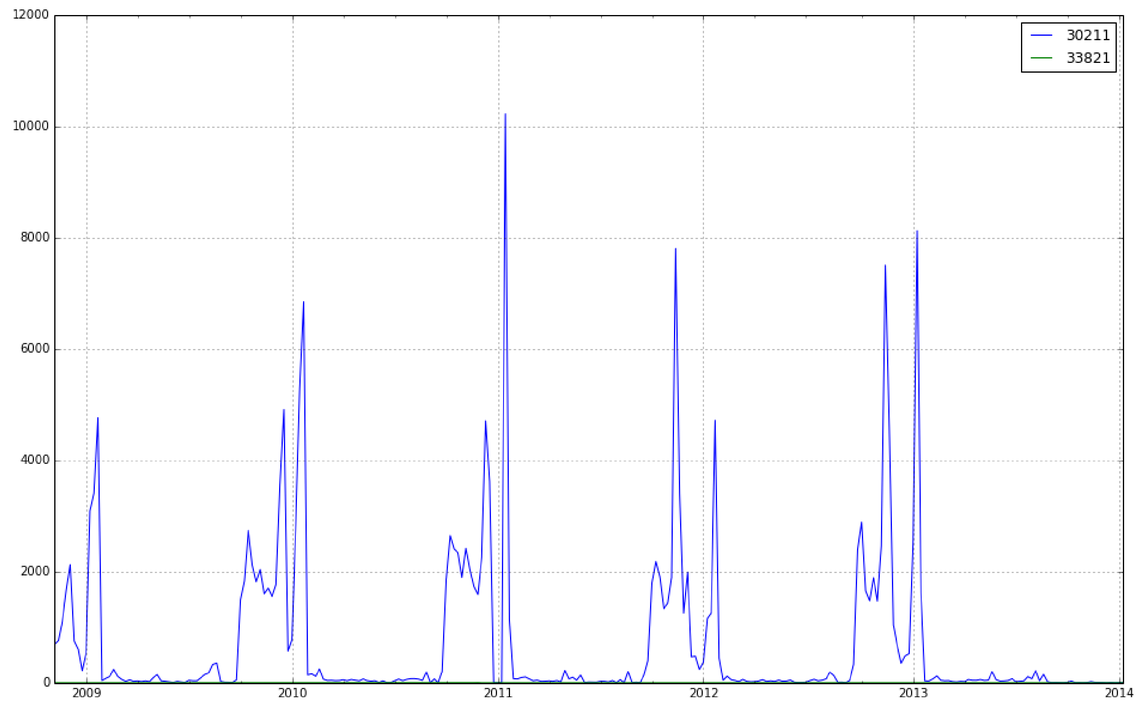

Figure: Sample Moodle Activity Logs

Feature Extraction

From their demographic data we extracted

- Age
- Gender
- Commute distance. This was calculated by looking up there home address using the Google Maps API
- Citezenship status. This was extracted using dummy coding

For each module, we divided the Moodle logs into chunks of weeks. We then extracted multiple features for each week

- A count of how many times they accessed moodle
- The ratio of on campus to off campus accesses using IP address
- The average time of day they used moodle
- How many times times they accessed moodle during the weekend

Classification

We used a Support Vector Machine (SVM) to classify students as a pass or a fail. We trained one SVM for each week of the semester up until the exams. Each training set contained all data available up until that week. For example the "week 7" SVM was trained with the demographic data, and all weekly Moodle log data up until week 7. This allows us to make predictions on new Moodle log data on a week by week basis.
The following process was used

- We selected only the first week of Moodle log data,
- We used 10-fold stratified cross validation
- We scored out model using Receiver Operating Charateristic Area Under Curve (ROC AUC) metric. This was used because it is less biased than accuracy when using imbalanced classes.
- We repeated this process for every week, including all of the Moodle data up to that week. We then graphed the performance of the classifier over each week.

This process gave us some intuition on whether the classifier was effective. If the performance of the classifier improves over time then the classifier was effective. We also noted that good classifiers typically achieve an ROC AUC score of above 0.6. We used these heuristics to select courses to target with intervention strategies for the coming semester.
When a SVM makes a prediction for a particular student it also produces a confidence score for that individual student. We can take advantage of this to rank the students according to their need of an intervention.

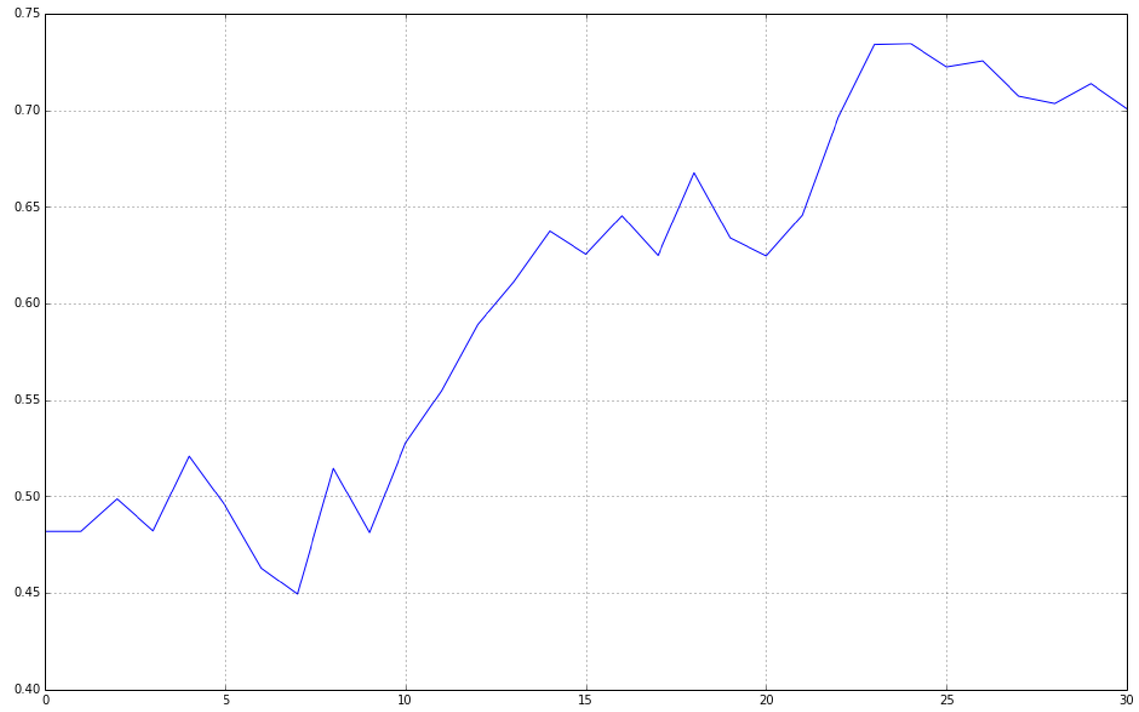

Figure: Sample classifier performance by week

PredictED

Some modules have been selected to run a trial based on the behaviourist idea of "Public Posting". Instead of lecturers intervening, students can compare themselves to others in their class in terms of predictions. Each week students receive an email with their predicted score and rank in the class. The are also given a link where they can compare themselves to everyone else in the class. Our hypothesis is that this will encourage those at risk of failing to positively change their behaviours. We will be running a trial of this project over the coming semester.

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN sfi Science Foundation Ireland