Old Dominion University

CS 773

Data Mining and Security

Course Project

Authors:
Grant Atkins
David Haslam

July 31, 2017

Contents

1 Executive Summary

The goal of this course project, originally stemmed from Open University Learning Analytics data set, is to identify "at-risk" students provided [2].

More needs to be done in this section...

2 Introduction

With analytics becoming an integral part of Learning Management Systems (LMS), Universities can analyze their data to intervene and aid "at-risk" students. This can be done by detecting failing or struggling students earlier on in their courses. These detections can later be used to predict "at-risk" students for future intervention. Analysis of demographic information for students may also prove useful to detect environments or character attributes that identify a struggling group.

The Open University (OU) in the United Kingdom, offered distance learning courses with its virtual learning environment (VLE) from 2013 to 2014. Students interacted with the VLE and OU then collects information such as: page names, clicks, and times connected to the website. Students are also required to register for these courses earlier on, which provides: registration date, unregistration date, class types and semester. Many students may end up failing or withdrawing from the VLE courses. The data sets that OU provided can be used to provide insight on 2013 to 2014 semester students.

3 Problem Statement

When analyzing students to observe for "at-risk" students it may seem easily done by looking at simply grade scores, however, there could be factors present in a student's personal history or personal accomplishments that correlate with course problems causing a student to be "at-risk." The goal of this paper is discuss and find data mining techniques that aid in the detection and prediction of "at-risk" students.

4 Solution Methodology

Finding a solution for this project was based on the data being observed. We first attempted to look at each data set, each csv provided, individually and then find attributes that would be the best indicators to determine the "final_result" of students. The tools used for this project included weka and R. Later we would merge the data sets together and then perform the following steps depending on the data set:

1. Find the best attributes for the data provided, often through information gain for feature selection or PCA for clustering
2. Test machine learning techniques such as:
   (a) Decision trees
   (b) Naive bayes
   (c) Bagging
   (d) K-means
3. Cross validation for measuring effectiveness

This paper first experiments upon the individual data sets and then the results of the merged data sets as explained in the next section.

5 Experimental setup and data used

There are many different ways to experiment on the OU learning analytics data set. There are many interactions between the different sets of data, such as demographic information can be linked to course assessments for each student as shown in Figure 1. The attribute "final_result" was the main attribute that was used in each data set because it determines whether students passed, failed, withdrawn or gained distinction. We found these attributes to be along the same lines, so we eventually paired them to Pass or Fail, where withdrawn became Fail and distinction became Pass.

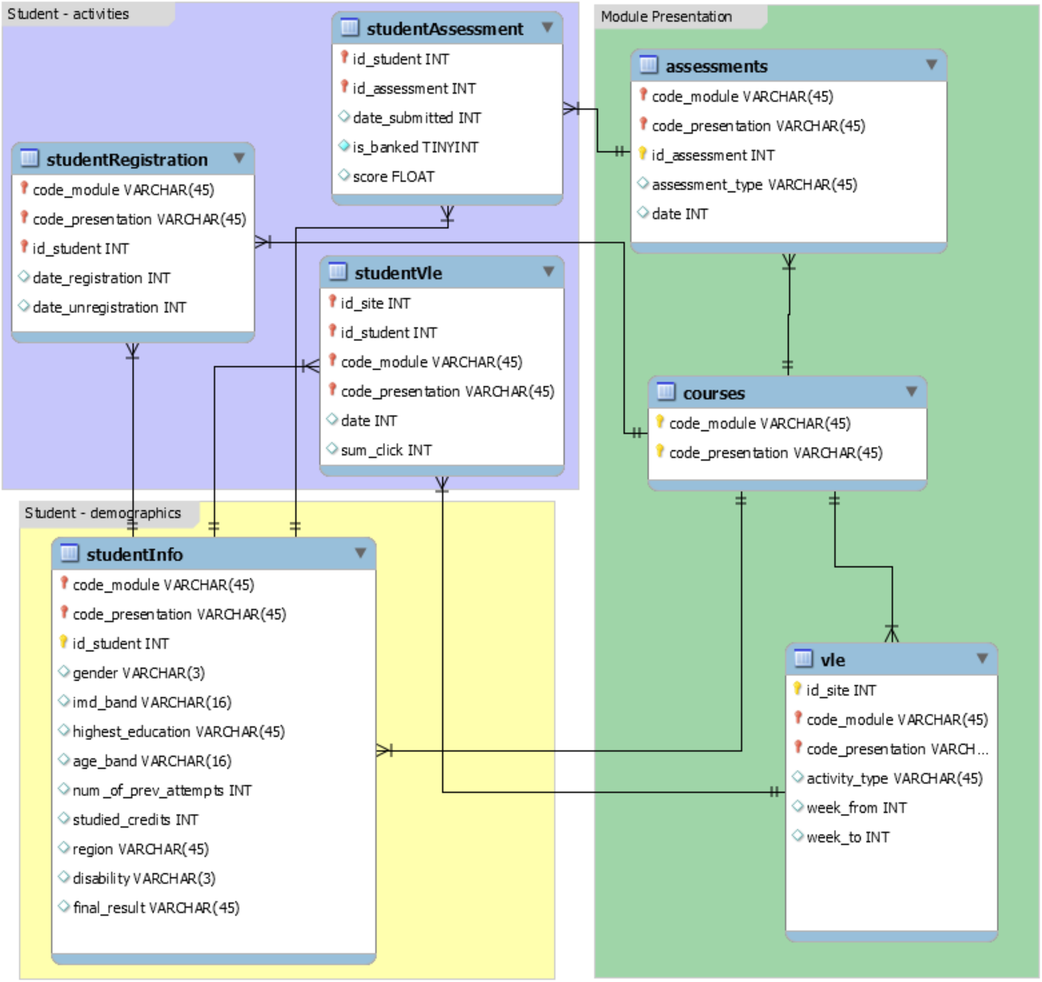

Figure 1: OU learning analytics data set relations [2]

While each file may contain valuable information about a student, many times important details can be lost among large amounts of useless data. In our attempt to isolate the most important attributes, we ran the files that were correlated with a student passing or failing to find information gain. These four files, ”studentAssessment”, ”studentInfo”, ”studentRegistration”, and ”studentVle” provided the most useful attributes to determine if a student passed or failed. The information gain for each of these files is displayed below in Figure xx. We found that the files ”course” and ”Vle” were almost completely useless.

When later testing this data, non-important attributes were purged. For example, student_id was removed because there is nothing to gain from specific individual students compared to other attributes such as score or code_module where there can be correlations.

After purging useless material from the input, we used various data mining techniques on each file separately to see if we could accurately predict whether a student would pass or fail. While certain files such as "studentVle" and "studentAssessment" did produce good results, it was not until we combined several of these files that we observed the best outcome.

In order to simplify the data, we only predict for two outcomes; pass and fail. The project states that we need to identify when a student needs counseling and are in danger of failing. Predicting for distinction would be useless because this is equivalent to passing. In addition, the main reasons for withdrawing is due to poor grades and lack of knowledge of the information. If a student was withdrawing, counseling would still be important and could possibly reverse the students decision.

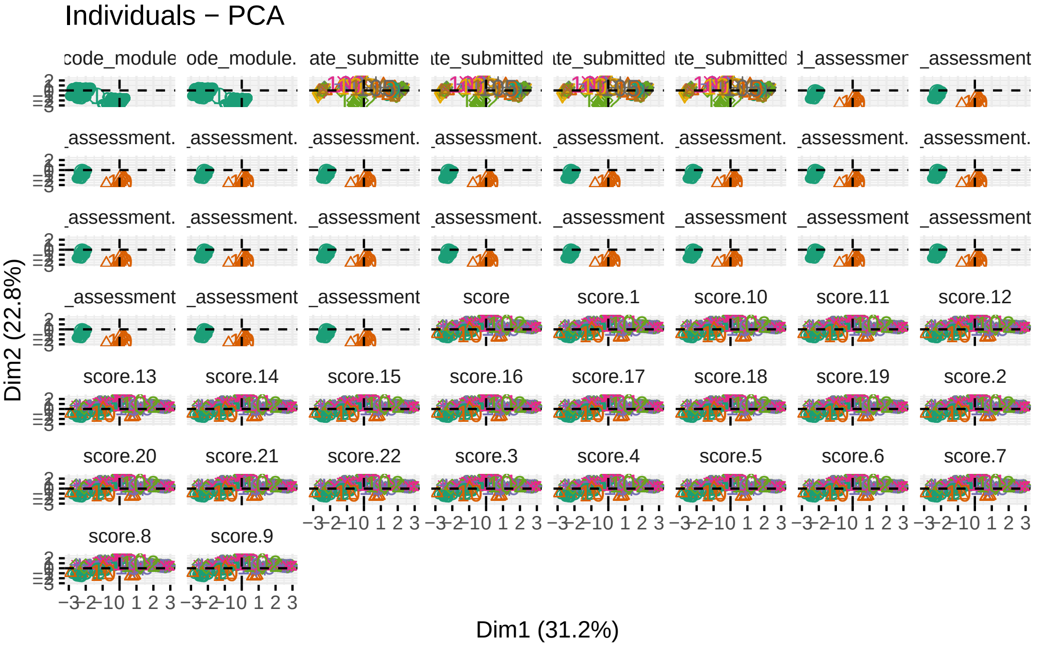

Figure 2: PCA of student assessment attributes

When first reviewing the data presented to us individually, student demographics was the first to be reviewed as this is information that is often received before a course has started. Approaching this set of data we first decided to use information gain to discover the best attributes to build a decision tree. This lead to interesting results as all of the attributes never had above a 5% information gain as shown in table 1. This deterred us away from using attributes from demographic information aside from final_result, which is the classifier for this data set.

Attribute	**Information Gain**
code_module	0.0333132948
studied_credits	0.0291538626
highest_education	0.0220755496
imd_band	0.0190484622
date_registration	0.0126263112
num_of_prev_attempts	0.0112197577
region	0.0099021067
code_presentation	0.0090550367
age_band	0.0047750479
disability	0.0030386879
gender	0.0003658152

Table 1: Information gain of student info and registration data merged

6 Results

Performing simple k-means using Weka's "densityBasedClusterer", showed very clear cluster when observing a students' click counts vs students' assessment scores as shown below in Figure 3. K-means chose number of clicks as the best attribute to perform clustering on.

If we view the clusters as final results, it becomes apparent that there is a correlation between click counts and scores for determining if a student passes or fails in the end. After approximately 14,500 clicks in the content, students generally all pass, even if a student receives a failing score after as shown in Figure 4.

Students who took the highest weighted assignment, where weight was 100, and scored below a 50 had a 87.88% chance to fail. This is fairly critical making students

Students who logged in to the VLE before the course started had a 67% chance to pass.

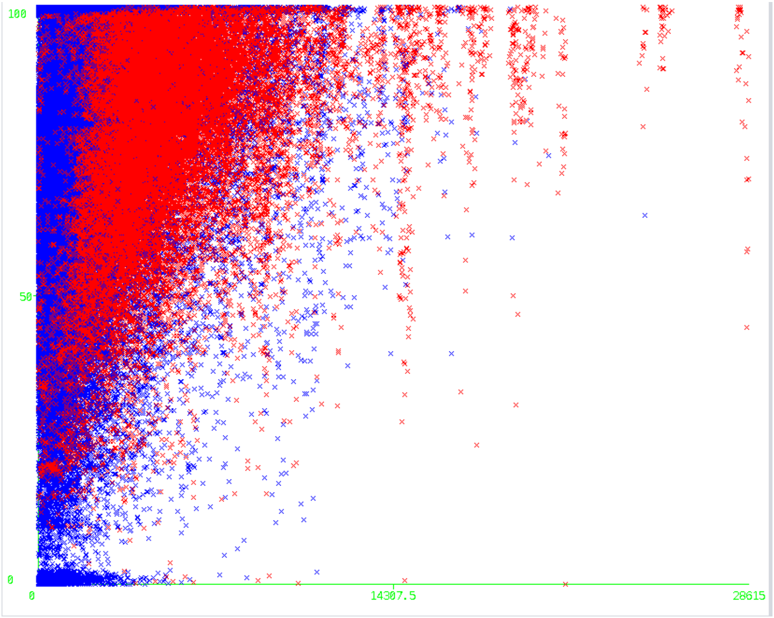

Figure 3: K-means clusters for click_counts vs scores

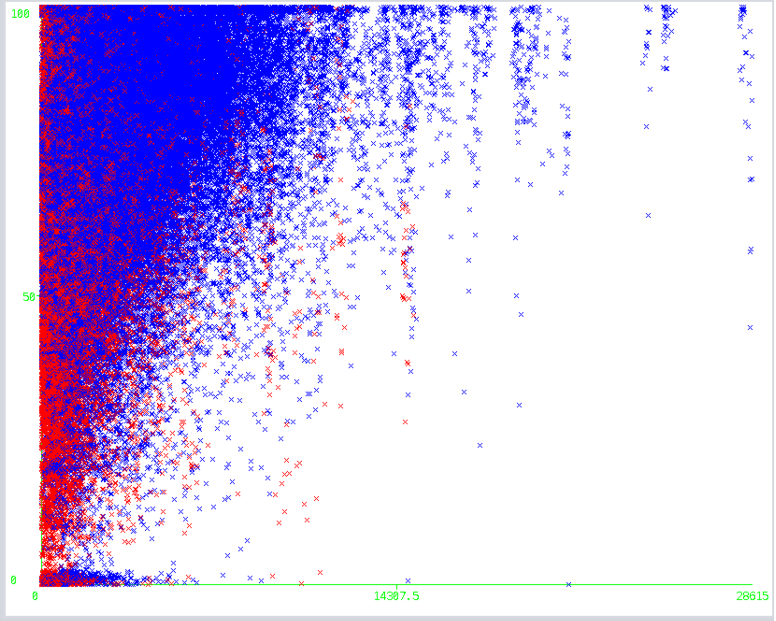

Figure 4: K-means results for click_counts vs scores with Blue being passing and red being failing

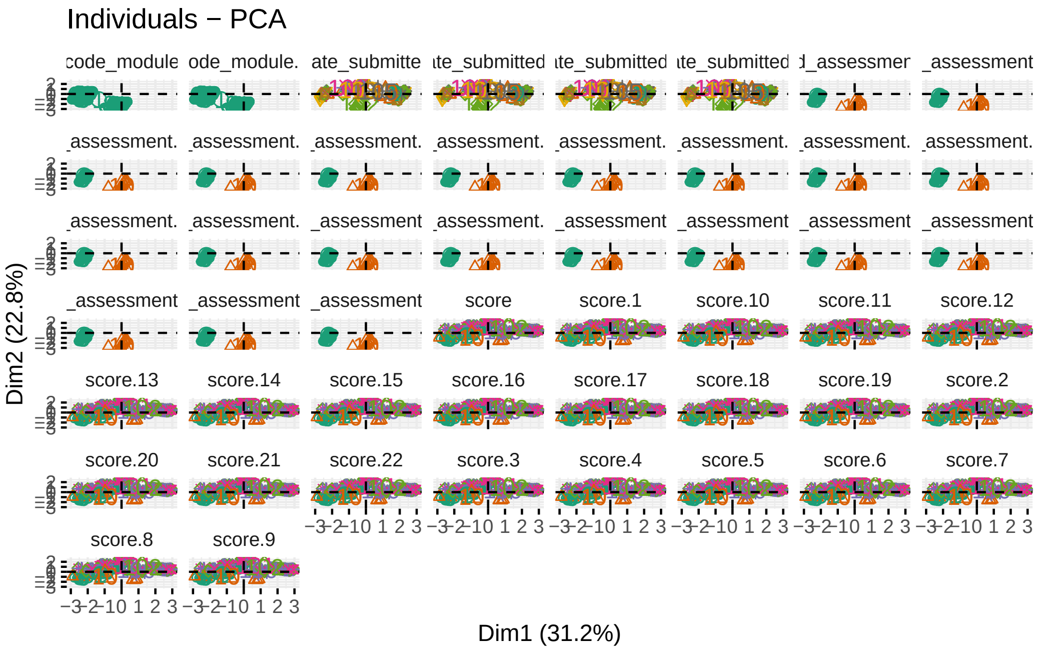

Figure 5: PCA of student assessment attributes

7 Conclusions

References

[1] Help on BibTeX entry types. http://nwalsh.com/tex/texhelp/bibtx-7.html. Accessed: 2015-03-12.

[2] Hlosta M. Herrmannova D. Zdrahal Z. Kuzilek, J. and A. Wolff. test. http://www.laceproject.eu/publications/analysing-at-risk-students-at-open-university.pdf, March 2015.