Factor Analysis and Outcome Prediction of QUT Higher Degree Research Students

Hugh Andersen
Kerrie Mengersen
Brendan van Rooyen

Introduction

Significant time, effort and resources are expended by students and universities to ensure student success and high university rankings. This has prompted research [1][2][3][4] using statistical models to analyse student profiles, engagement types and to predict student success or failure.

This research project uses corporate QUT demographic and engagement data on Higher Degree Research students from years 2015 and 2016 for the analysis. Basic data visualisation shows the make up of HDR students, this is followed by an exploratory factor and K means cluster analysis. The factor analysis results are used in a logistic regression model to determine which factors influence a successful student outcome.

Data Visualisation

The plots below illustrate some of the diversity of the data. QUT HDR students are predominantly Australian or Chinese, study Science and Engineering, Medicine or Creative Industries (Figure 1). Students use different forms of engagement, the highest is logged into QUT resources while on campus (Figure 2). Student course loading and engagement hours seems to affect outcome (Figure 3).

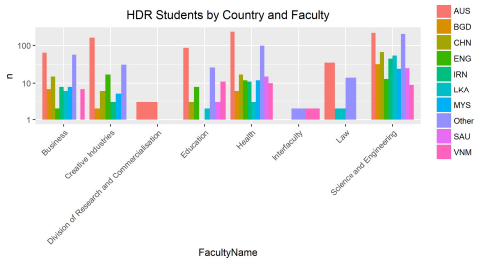


Figure 1: HDR Student make up

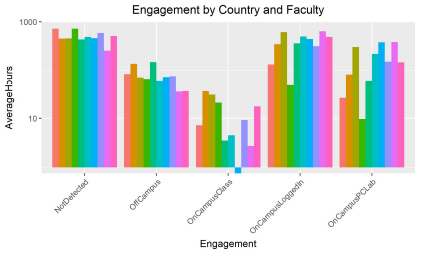


Figure 2: HDR Student engagement

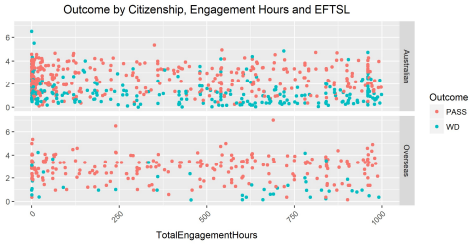
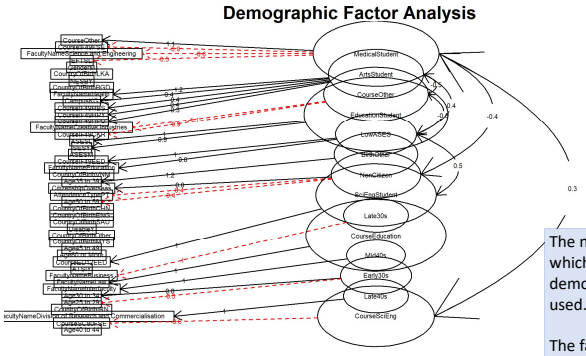


Figure 3: HDR Student outcome

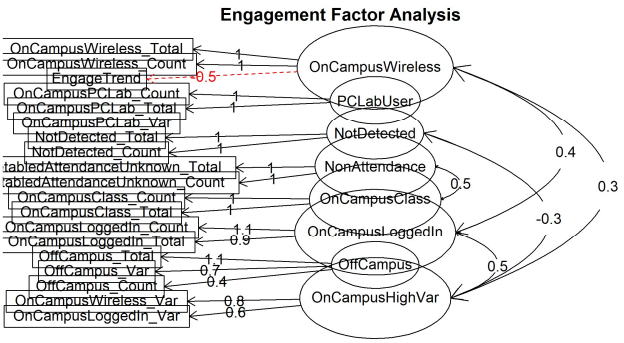
Factor Analysis

A factor analysis was conducted on the demographic and engagement data using the R Psych package [5] for variable reduction and to better understand student profiles.



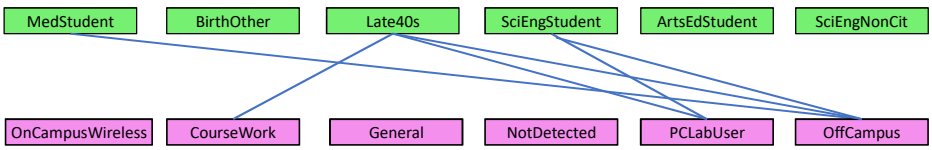
The number of factors was chosen using scree plots which showed that 14 and 8 factors for demographic and engagement profiles should be used.

The factor diagrams show correlation between factors, but it is high enough to be problematic. The factor analysis captured 0.43 and 0.80 of the cumulative variance.



K means Cluster Analysis

The factors were used to create clusters to further profile the students using the K means algorithm.
By counting number of students in each cluster a linkage diagram could be created.



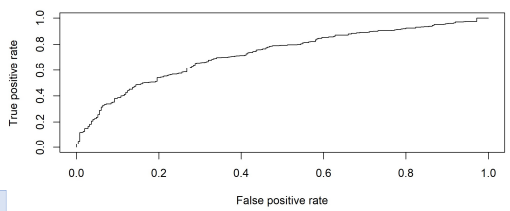
Only the main links between demographic clusters (top row) and engagement clusters are shown.
Medical, science and engineering and students in their 40s tend to work off campus and logon remotely.

HDR Outcome Prediction

A model to predict student outcome (pass or fail) was created using the student factors. A Logistic regression [6] model was chosen due to its interpretive ability. The model result is 68% accurate.

```
Confusion Matrix and Statistics

          Reference
Prediction  WD PASS
      WD   104   69
      PASS 146  363

               Accuracy : 0.6848
                 95% CI : (0.6484, 0.7195)
    No Information Rate : 0.6334
    P-Value [Acc > NIR] : 0.002824

                  Kappa : 0.2741
 Mcnemar's Test P-Value : 2.182e-07

            Sensitivity : 0.8403
            Specificity : 0.4160
         Pos Pred Value : 0.7132
         Neg Pred Value : 0.6012
             Prevalence : 0.6334
         Detection Rate : 0.5323
   Detection Prevalence : 0.7463
      Balanced Accuracy : 0.6281

       'Positive' Class : PASS
```

Logistic regression output shows:
- Accuracy = 0.68
- Kappa = 0.27
- Mcnemar's p value=2.18e-07



The ROC plot illustrates how well the model fits the test data.

Variable	PassPercent
(Intercept)	154.53
OnCampusWireless	14.18
PCLabUser	-(8.13)
NotDetected	-(16.20)
NonAttendance	-(26.77)
OnCampusClass	-(11.36)
OnCampusLoggedIn	-(15.94)
OffCampus	6.62
OnCampusHighVar	13.63
MedicalStudent	28.23
ArtsStudent	-(30.47)
CourseOther	-(9.44)
EducationStudent	-(3.85)
LowASES	-(7.71)
BirthOther	-(12.17)
NonCitizen	107.41
SciEngStudent	8.34
Late30s	3.67
CourseEducation	-(3.72)
Mid40s	22,446.68
Early30s	-(14.08)
Late40s	-(8.36)
CourseSciEng	37.53

This table shows the percent by how much each factor influences a student's probability of passing.

References and Acknowledgements

[1] Agrawal, H., & Mavani, H. (2015). *Student Performance Prediction using Machine Learning* (Vol. V4).
[2] Aragon, S. R., Johnson, S. D., & Shaik, N. (2002). The influence of learning style preferences on student success in online versus face-to-face environments. *The American Journal of Distance Education, 16*(4), 227-243.
[3] Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). *Data mining algorithms to classify students.* Paper presented at the Educational Data Mining 2008.
[4] Broecke, S. (2015). University rankings: do they matter in the UK? *Education Economics, 23*(2), 137-161. doi:10.1080/09645292.2012.729328
[5] The Personality Project, http://personality-project.org/r/psych/
[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. . (2015). *An Introduction to Statistical Learning* (Vol. 6): Springer Publishing.

My thanks to QUT Strategic Intelligence Unit for providing the data, and my supervisors.