Identifying Students at Risk of Dropping Out Their Second Year from Cape Peninsula University of Technology – A Case Study

Felipe Rodriguez

Department of Data Science, Bellevue University

Professor Andrew Hua

October 1, 2023

Introduction

This case study is reviewing a study from the South African Journal of Higher Education titled, “Applying Predictive Analytics in Identifying Students at Risk: A Case Study.” This study was performed by A. Lourens and D. Bleazard. The study aims to determine “likely student dropouts before or in the second year of study for a particular qualification at the Cape Peninsula University of Technology in Cape Town (CPUT)” (Lourens & Bleazard, 2016). By understanding the likelihood of dropout, recommendations can be made to further prevent the dropout rate of students entering their second year. As stated in the study by Lourens and Bleazard, “objective is to reduce the second-year dropout rate and to increase the university’s pass rates and throughput rates” (2016). Prior to this study, the University established two programs to aid in understanding student and first year retention. The project ‘Know your students’ was created to understand students at the institution and ‘First-year Experience’ was created with the aim of reducing dropouts between first and second year (Lourens & Bleazard, 2016).

This study was proposed to provide a predictive analytic component to the ‘First-year Experience’ program and to discover those who might be likely to dropout. It was determined to focus on the National Diploma in IT for this study. The data used for this analysis is gathered from CPUT’s institutional operational data for first time student enrolled from 2008 to 2014 to predict dropout by the second year ((Lourens & Bleazard, 2016). After the data was cleaned, there were a total of 1593 student records that were used in the analysis.

Method and Results

The data collected contained data from the years 2008 to 2014, however, the model used data from 2008 to 2013 and then used the 2014 data later as a validation set. A total of 1593 records were gathered and 452 of those student records were classified as second year dropouts. The case study used supervised learning models to create predictions of independent variables of the data. "KNIME was used to perform the predictive modelling with second-year dropouts as the dependent/target variable" (Lourens & Bleazard, 2016). Labels are needed for the supervised learning model, any student that did not return for the second year was labeled as 1 and those who did were labeled as 0. This labeling will go on to help the analysis and determining predictions. Another source of data was used for this study as well. This data included demographic information, that was gathered prior to enrollment, and academic performance data which was provided by the university. The models used within KNIME were Logistical Regression, Naïve Bayes, and Decision Tree. KNIME "is a low-code data science and data preparation platform that makes understanding data and designing analytic workflows accessible" (Emery, 2023).

The data from 2008 to 2013 was divided randomly into a 30-70 split, 30 percent being testing data and 70 percent being training data. Prior to creating the model, the authors also tested for collinearity. The records from the 2014 year were used to predict the outcome depending on the model. There were eight final prediction variables and the models selected were applied to these. The list of prediction variables was as follows: each was "binned first-year module marks in Technical Programming (TP), Information Systems (IS), Development Software (DS), Information Technology Skills (ITS), Systems Software (SS) plus Financial Aid (NSFAS) (Yes or No), Grade 12 Mathematics (Yes or No), and Type of accommodation (resident student or not)" (Lourens & Bleazard, 2016). The models analyzed some statistical components to establish adequacy. These were "a confusion matrix indicating the percentage correctly predicted, the sensitivity and the specificity, Cohen's kappa, the area under the receiver operating characteristic curve (AUC), and the validation data set" (Lourens & Bleazard, 2016).

Upon completing the models, the results showed that "the models performed well, and the Logistic Regression model had the highest percentage accuracy (88.6%) with a 1.3 percent error rate and was subsequently used on the validation data set" (Lourens & Bleazard, 2016). The table created in the case study demonstrates the accuracy and error of each model.

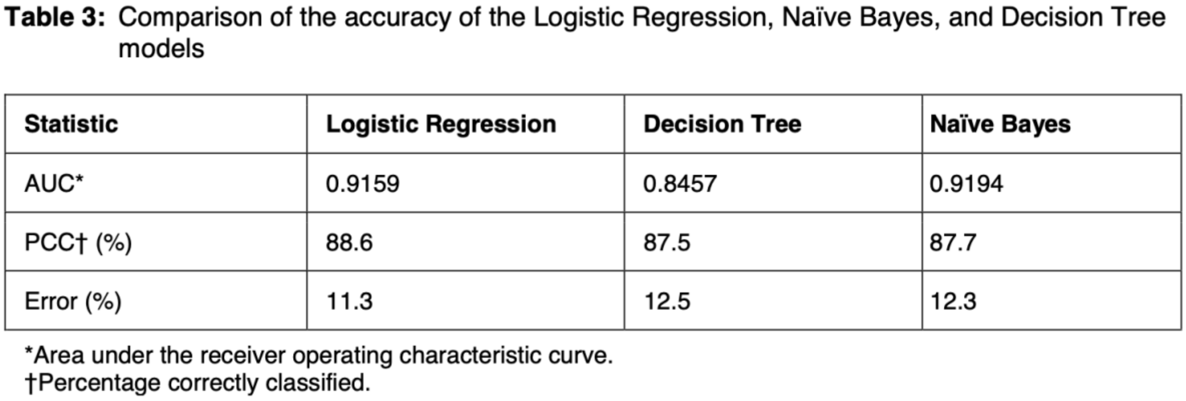

Table 3: Comparison of the accuracy of the Logistic Regression, Naïve Bayes, and Decision Tree models

Statistic	Logistic Regression	Decision Tree	Naïve Bayes
AUC*	0.9159	0.8457	0.9194
PCC† (%)	88.6	87.5	87.7
Error (%)	11.3	12.5	12.3

*Area under the receiver operating characteristic curve.
†Percentage correctly classified.

Using the model, names of students were determined to see which were more at risk of dropping out. One form of cross check conducted in the study was verifying six names of those produced by the models against enrollment records, it was found that 5 out of 6 had already cancelled their registration and the sixth was labeled as high risk. This gave further comfort in using the model for the department.

Conclusion

The model created by the team provided high accuracy and great predictions when selecting the 6 students to compare the model results versus their enrollment status. This study helped the department in determining if there are any factors that lead to students dropping out their second year. Based on the results of the model, there is some consistency and pattern of behavior that leads to the students dropping out their second year. Because of predictive analytics, the department can create opportunities to reach at-risk students to prevent dropouts within their second year. A benefit from these analytics is “it can assist higher education institutions to implement targeted intervention strategies in the first year of study in order to reduce the number of students leaving prematurely by their second year” (Lourens & Bleazard, 2016). The university can use this information to improve retention rate as well as student satisfaction with the university. Another beneficial factor of this study is that the income from tuition fees generated will no longer be lost if the students can be retained in the university (Lourens & Bleazard, 2016), resulting in more income for the university. An actionable item identified by the team is to incorporate some of the questionnaire data of incoming students into the model to create better results in the models.

In the future, the team should aim to incorporate more records of data to see if the model can determine more accuracy as well. Furthermore, the team can incorporate the use of predictive analytics to other departments and expand recommendation that will widely aid the institution. The factors identified can also be shared with sister schools to see if there is any pattern between the two. The model can also be applied additional years to see how the drop indicators change over the course of a course of a four-year traditional college, even though it is less likely to dropout as time goes on, there could be similar items that lead to dropping out later in their schooling.

Reference:

Emery, J. (2023, August 9). *What is Knime and tips for getting started*. phData. https://www.phdata.io/blog/getting-started-with-knime/

Lourens, A., & Bleazard, D. (2016). Applying predictive analytics in identifying students at risk: A case study. *South African Journal of Higher Education*, *30*(2). https://doi.org/10.20853/30-2-583