PREDICTING ACADEMIC SUCCESS USING LOGISTIC REGRESSION ON BIOGRAPHICAL DATA

Innocent Mamvura
Wits Business Intelligence
Siyaphumelela Conference 2017
27-29 June 2017

Acknowledgements

- Wits Business Intelligence Services
- Wits DVC Academic
- Wits Academic Advisors

Data is the Key to Success

"We are not going to run out of data anytime soon. Its maybe the only resource that grows exponentially. Maybe every 1.5 years we are seeing data double, and much of that data is social data, data about ourselves" Andreas Weigend

Abstract

- The University of Witwatersrand adopted the Seidman(2006) Retention model for student success to identify students who are most likely to fail first year based on the Admission point score, High School Marks. The university has two faculty advisors in each faculty providing interventions to students. This paper used a Binary Logistic technique to predict academic success using qualitative variables collected through a structured questionnaire by the university. IBM SPSS was used to analyze the data and Computer skills, library usage skills, First Generation, Financial Aid, Residence and Distance was found to be significant. The findings contribute greatly to the existing work in the field of students success and the university will now focus its resources to the most needy students.

Problem Statement

- Wits is collecting large amounts of data pertaining to student background, school information, financial background, level of competence in computers skills, library usage skills etc.
- The existing Student Early Warning System is based on only quantitative variables.
- Identifying at risk students using the existing system is resulting large numbers of first year students with less resources to assist these students.
- Greater shortage of resources.
- No Improvement in first year success rates.

Findings

- Distance to campus, Financial Aid, Residence, First Generation, Computer Skills, Library Usage is a predictor of academic success.
- Working part time is not a significant.
- The predictive accuracy of the Binary Logistic model is 81.2%.
- These findings will help with allocation alignment of resources where they are needed the most.

Independent and Dependent Variables

- Dependent Variable is Success with 1 and 0
- Independent Variables are First Generation, Distance, Residence, Financial Aid, Computer Skills, Library usage , Work part time

Definitions

- Binary Logistic Regression is a technique used when the outcome variable is a dichotomous variable (has two values). Logistic Regression uses Binomial Probability Theory in which there are only two outcome categories. The technique forms a function using the Maximum Likelihood Method, which maximizes the chances of grouping the observed data into the suitable category given the regression coefficients.

Assumptions of the Model

- Assumptions of Logistic Regression

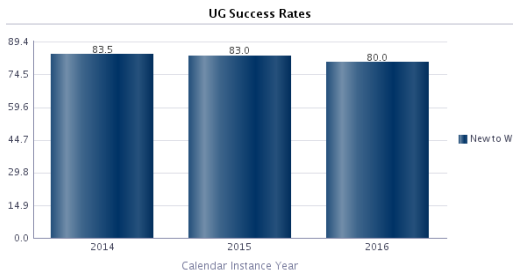
❑No Linear Relationship between the outcome and predictor variables is required;

❑The outcome variable must have two categories;

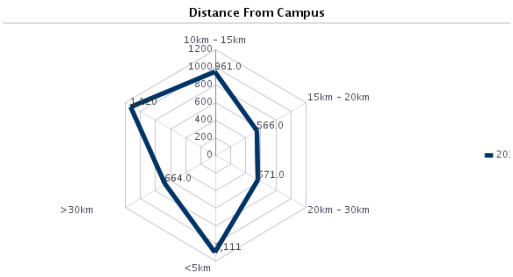
❑The predictor variables does not follow a normal distribution, or linear relationship;

❑ Maximum Likelihood coefficients are large sample estimates

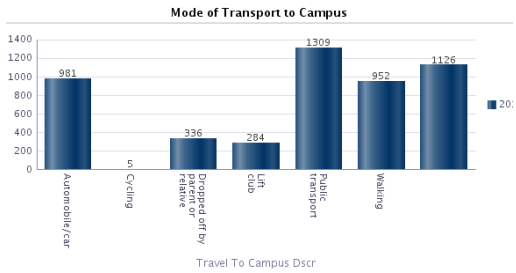
Wits UG Success Rates



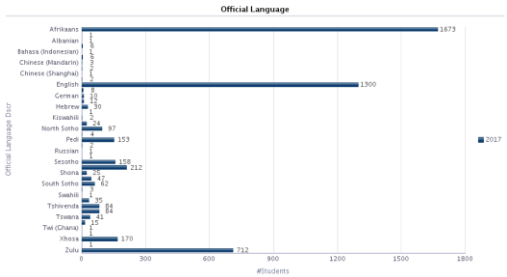
Data Visualization



Data Visualization



Data Visualization

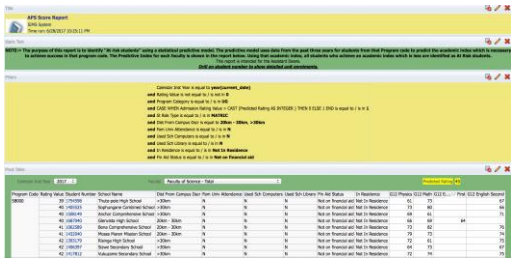


Binary Logistic Model

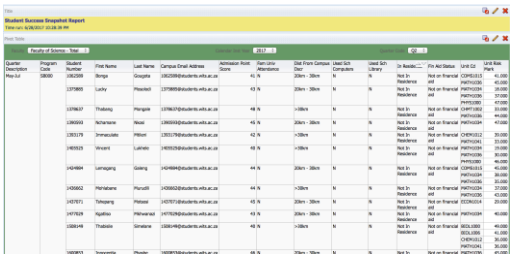
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Work(1)	.038	.091	.170	1	.680	1.038
	Distance			33.641	5	.000	
	Distance(1)	.118	.122	.939	1	.332	1.125
	Distance(2)	.523	.125	17.470	1	.000	1.688
	Distance(3)	.594	.149	15.786	1	.000	1.810
	Distance(4)	.532	.142	13.939	1	.000	1.702
	Distance(5)	.067	.202	.109	1	.742	1.069
	Residence(1)	.629	.196	10.267	1	.001	1.876
	FinancialAid(1)	.622	.085	53.876	1	.000	1.863
	FirstGeneration(1)	.478	.084	32.245	1	.000	1.613
	LibraryUsage(1)	.206	.090	5.220	1	.022	1.229
	ComputerSkills(1)	.326	.093	12.350	1	.000	1.385
	Constant	.245	.108	5.189	1	.023	1.278

Model Implementation & Dashboards



Model Implementation & Dashboards



Future Work

- Include more variables from the BQ data
- Use other machine learning techniques
- Explore deep learning algorithms
- Put the power of data into the hands of students
- Collaborate with other Universities on BQ

Conclusion & Recommendations

- Students travelling long distances to campus, not on financial aid, first generation, not in residence, no computer skills, no library usage skills are most likely to fail their first year.
- Working Part time is not significant.
- The university could provide accommodation to students staying >20km from campus and also provide access to computers and workshops to the students who had not been exposed to computers at high school. The university could also perhaps provide studying skills workshops and general use of library resources to these students. Academic mentors could also be arranged for first generation students as they require motivation and guidance.


Statistics
Predictive Analytics
Data Scientist
Tableau
Hadoop
Apache Spark
Interview
ython
Big Data
SAS
Data Mining
Scala
Machine Learning
Neural Networks
Text Mining
Data Science
Data Science
Business Analytics
Deep Learning
Modelling
Text Analytics
Analytics