Finding Patterns in Data: Creating a Model to Predict Student Success

Jose A. Muguira[1], Reinaldo Sanchez-Arias, Ph.D[2]
[1]Miami Dade College, Wolfson Campus, Miami FL
[2]St. Thomas University, School of Science, Miami Gardens, FL

Abstract

Student success is a major focus in the educational system, where a variety of predictors are used to estimate and measure how well students do in their different classes at the end of the academic year. Our research project aims towards proposing a model capable of demonstrating how student success can be predicted based on a series of indicators gathered from work submitted by the student throughout the semester.
We studied the student's performance in an online homework assignment system for a mathematics course, taking into account the final score in a given assignment, but also the number of times every problem was tried by the student before obtaining a correct answer.

Introduction

Some instructors at St. Thomas University use an open-source online homework system for mathematics courses called WeBWorK [1]. By accessing this resource, students have a system that lets them work at their own pace and helps reinforce the different topics covered in a given course.

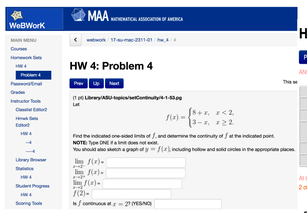


Fig 1 : Sample calculus assignment given to students using WeBWorK

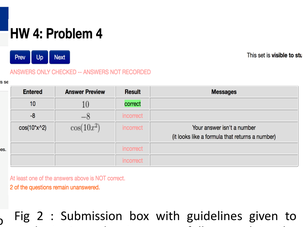


Fig 2 : Submission box with guidelines given to students in order to successfully complete the assignment

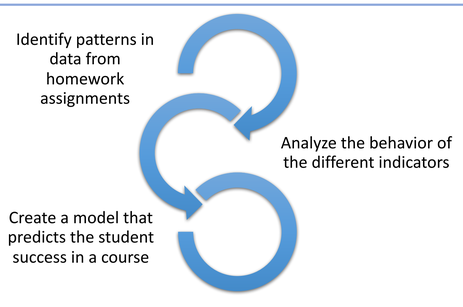
How can we identify if a student is struggling with a particular topic/problem/assignment?

WeBWorK stores information of the scores for each assignment, as well as a variable called "success index", an indicator of the number of attempts made by the student to complete the assignment. This information was used to create our data sets, run a clustering analysis and differentiate between students that are very efficient in the assignments and those who need more trials before completing a given homework set.

$$ind = 100\frac{1}{\text{average number of attempts per problem}}\left(\frac{\text{total correct problems}}{\text{total problems}}\right)^2$$

Fig 3 : Formula used by WeBWorK to compute the "Homework Index" for each student

Objectives



Data Analysis and Mathematical Modeling

What specific data was used in the research?

The main indicators used were labeled the "Homework Percent", which represents the score obtained by a student in a given assignment, and the "Homework Index", recorded by WeBWorK to measure the *success* indicator for the problem set.

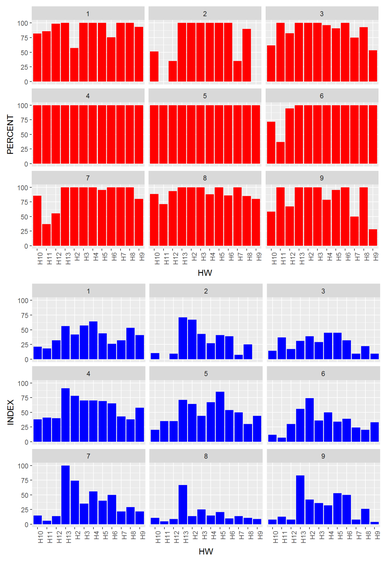


Fig 4. Behavior of the "Homework Percent" and the "Homework Index" of nine different students in a MAC2311 course

What differences a high-success student with a low-success student?

As part of our analysis, we tried to identify the difference between a student with a high grade in the class and a student with a low grade. Using the Homework Index variable for our analysis, we can appreciate how the student with lower grades has generally also a lower index than the student with higher grades.

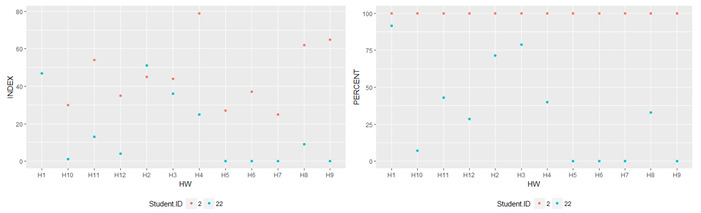


Fig 5 : Analysis between a high-grade student (Student #2) and a low-grade student (Student #22)

Based on this observations, we used all the records and created a graph to help us visualize the different groups in the students. As observed in the graph we were able to easy identify four groups, using a K-means algorithm in R. Students with high index and percent, students with average scores, students with low index but a high percent and students with low index and low percent.

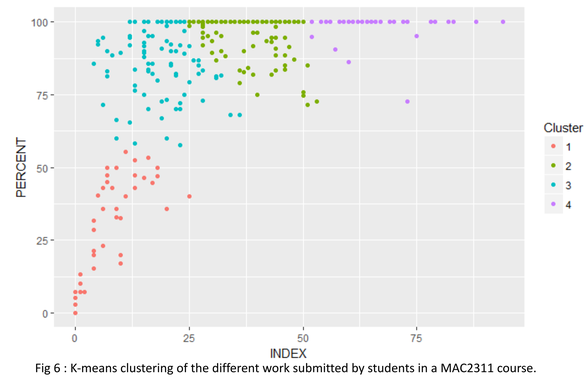


Fig 6 : K-means clustering of the different work submitted by students in a MAC2311 course.

Multinomial Logistic Regression Model

After all the data was successfully processed analyzed and having identified patterns, we created a mathematical model to interpret this analysis. In order to do that we grouped the grades into 4 categories: Superior, Merit, Satisfactory, and Poor. Then we divided the average of the index for every student into the categories of High, Middle, and Low. Using this information and with the help of the R statistical programming language [2], we created a *multinomial logistic regression model*, capable of calculating the probability of a student to succeed on the course based on the average homework percent and the average homework index. The 4 different categories created can be then translated into an estimated final grade in the course. The model studied in this project showed high accuracy in the predictions and allowed for an easy interpretation of the model outputs. Our framework was tested with data from one Pre-Calculus and two Calculus I classes at St. Thomas University.

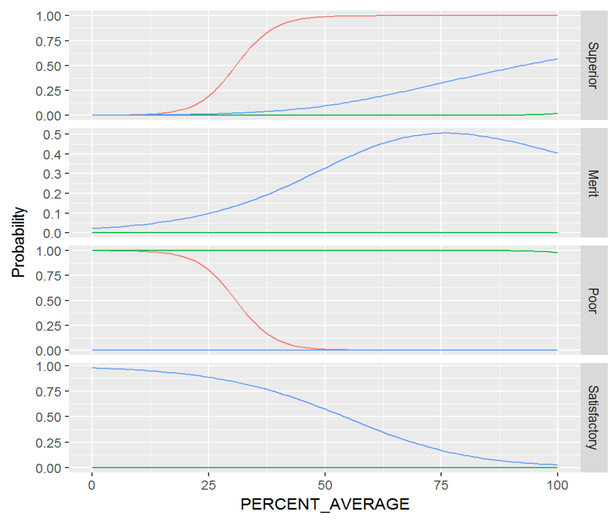


Fig 7: Multinomial logistic regression model of a MAC 2311 Calculus I course.

RStudio

Programming was essential in the development of this project. R is an open source language widely used in the data science community, with focus on statistical data analysis, data visualization and machine learning methods. During this project, tools from the "tidyverse" package [3] were used for data wrangling and data visualization with the help of RStudio [4] , an open source integrated development environment (IDE) for R.

```
library(tidyverse)
library(ggplot2)
  elim_HW_SID_p <- select(index_vs_percent_p,INDEX, PERCENT)

km.out_p = kmeans(elim_HW_SID_p, 3)
ggplot(index_vs_percent_p, aes(x= INDEX, y=PERCENT, color =as.character( km.out_p$cluster)))+
  geom_point()+
  labs(color="Cluster")
```

```
percent_average_sample_p <-read.csv("precalculusPAVERAGE.csv")
  percent_average_table_p <- select(percent_average_sample_p,Student.ID, PERCENT_AVERAGE, Qualification)

index_average_sample_p <- read.csv("precalculusIAVERAGE.csv")
  index_average_table_p <- select(index_average_sample_p, Student.ID, INDEX_AVERAGE)

index_vs_percent_average_p <- full_join(percent_average_table_p, index_average_table_p)
```

Fig 8 : Sample code used in the development of the different graphs and tables used for the project.

Conclusions

In this research project we designed a model capable of predicting a student's probability of success in a given course, based on student's work submitted throughout the semester in the form of online homework assignments. Data from one Pre-Calculus and two Calculus I courses at St Thomas University was used to create a multinomial logistic regression model. The model takes into account the student's scores in all assignments during a semester, as well as the student's "success index" per assignment, a fairly good indicator of how well the student is grasping the concepts evaluated in every assignment.
The model can be extended to include other predictors, and additional observations could be easily added to our framework for an even more robust model and prediction accuracy. The development of a web application (Shiny App) will be explored, taking advantage of the tools available within the R programming language. Such web application might be used for students to track their performance in a course and have an estimate of their final grade.

References

[1] WeBWorK, open-source online homework system for math and sciences courses. (Mathematical Association of America, MAA)
https://webwork.maa.org/wiki/Scoring_a_Problem_Set
[2] R: free software environment for statistical computing and graphics.
https://www.r-project.org/

[3] tidyverse: an opinionated collection of R packages designed for data science.
https://www.tidyverse.org/
[4] Rstudio: open-source integrated development environment (IDE) for R
https://www.rstudio.com/

Acknowledgements

I want to thank Dr. Reinaldo Sanchez- Arias for being such a great role model and guiding me successfully through the steps of analyzing and creating effective representation of data. I also want to highlight the support of my research partners for always being present when needed and the Faculty at the School of Science of St. Thomas University.
This project was supported, in part, by U.S. Department of Education grant award P03C1160161 (STEM SPACE), P031c160143 (STEM EngInE), P120A160036 (STEM ISLE), 1161177 (STEP Up), P120A140012 (SPARC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding agency.