WHAT ARE THE DRIVERS OF STUDENT DEMOGRAPHICS IN ANALYTICS AND DATA SCIENCE PROGRAMS? DOES ANALYTICS PROGRAMS HAVE ANY RELATIONSHIP WITH STUDENT'S DEMOGRAPHICS? WHAT IS THE CAUSE OF THE PROGRAM'S GREATER DIVERSITY?

Hsin Yu Pan (Candice), Oladimeji Adekoya (Dayo)
Pan351@purdue.edu; oadekoya@purdue.edu;

Abstract

This is a study that was conducted in order to identify the statistical components of students' demographics in the Analytics and Data Science Programs in the United States. Such as Age, percentage of Women to men, percentage of international Students and students work months etc. Also, by identifying the drivers of students' demographics in the Analytics and Data Science programs, universities will be able to re-strategize and re-design the program contents to admits better students.

Business Problem:

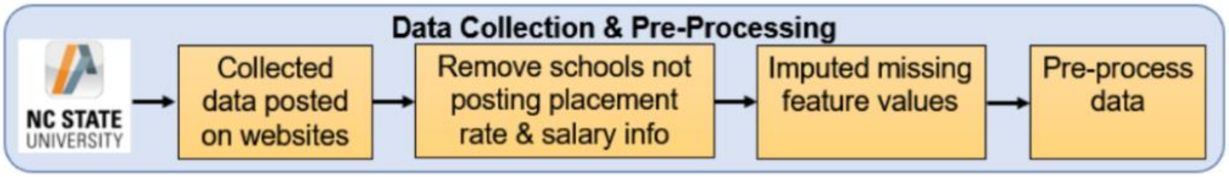
The idea behind this study is to use analytics (both descriptive and prescriptive) to provide a better insight to future students who may be unsure as to which school to choose, whether the return on investment in the Analytics and Data Science Program is worth the time and resources. The stakeholders in this study are; future Analytics Students, University recruiting bodies, University decision makers, and program coordinators. These stakeholders have been able to identify and come to an agreement that the drivers of students' demographics in this program, once identified, would play a vital role in the future of Analytics and Data Science programs in the United States.

Analytics Problem Definition:

The study considers insights through data collected from various Universities in the United States, in identifying the drivers of students Demographics in Analytics programs and describe relationships between students' demographics and analytics programs offered.

Data:

The data collected was first studied and understood by running the mean and hist for all the needed columns. That gave us a better understanding of our dataset and that helped in prioritizing and cleaning our dataset. Below is a figure showing the process of getting our data ready for DSS:


Data Collection & Pre-Processing
NC STATE UNIVERSITY
Collected data posted on websites
Remove schools not posting placement rate & salary info
Imputed missing feature values
Pre-process data

To understand and identify important relationships in the data, we used a descriptive visualization approach in our DSS by plotting the graphs.

Methodology Selection:
The analytical approach used to support this study are descriptive analytics and Predictive analytics. For descriptive, we used plot function to plot few histograms to show the relationships between the datasets and identify the key drivers in students' demographics. Linear regression was used in our predictive analytics using the *lm() function*. R is believed to be better analytical tool than most other statistical packages because of its very sophisticated graphics capabilities, and we have selected our analytical approach based on our audience's mental accounting.

Model Building:
Assumption: For R, python and SAS course. We replace all the missing value to 0. (1= yes, it provides course & 0 = no, it doesn't provide course)
From our analysis, we were able to identify a strong correlation between Duration of Analytics programs and Percentage of women in the program which indicated that Duration of the program is one of the drivers of students' demographics in the Analytics and Data Science Programs in the US. We also use predict() function to predict the percentage of women and student Age.

GUI & Functionality:
The Decision-Support System was designed to be user friendly and easily operated by our audience. The DSS shows a general information tab which allow users to choose from a dropdown list of Universities in United States offering Analytics and Data science program. Selecting a university from the list will populate some useful information about the Analytics and Data Science program. Another tab in our DSS, shows a statistical summary of Students Demographics in the university selected. The last two tabs shows plots of our cost Linear Regression and Work Months Linear Regression. It can interact function of excluding outliner. R-package, **ggplot**, was used in our descriptive analytical approach, R-package **Shiny** was used in designing our DSS while **Caret** package was used in our predictive analysis.

Conclusion:
We were able to identify Program Duration, and Students' Work Experience as key drivers in the analytics and data science program. We were able to identify few parameters that shows a linear relationship between analytics program and student demographics. We also found out that more data would be needed to be able to identify the cause of greater diversity in the program.

ShinyApp: https://pan351.shinyapps.io/masterprogram/
Github: https://github.com/candicepanpan/Shinyapp_project/tree/main
Video: https://youtu.be/gNM4gv9qhSQ

Reference:

Professor Lanham's lecture code
https://rstudio.github.io/shinythemes/
https://shiny.rstudio.com/gallery/plot-interaction-exclude.html