Debt, Degrees, and the Economy - SAS Advisement Analytics

Team1: HiEdu Consultants

1. Demi Constantine
2. Steven Schoiber
3. Braden Schrage
4. Charles Utt
5. Christopher Woods

SPRING 2019, MSDS SEC57, Capstone Project

SPS, Northwestern University, Evanston, IL

Advisor: Nethra Sambamoorthi, Ph. D

Team 1 - SAS Advisement Analytics

Index

Team 1 - SAS Advisement Analytics

1. Introduction

College and university students are taking longer to finish their degrees and are dropping out completely at a higher rate than seen previously. In 2012, the national average for full-time students at 4-year degree-granting institutions was 59 percent (Engelmyer, 2019).In 2019, over half of undergraduates have not completed their degree plans within six years (Selingo, 2018), and only 66% of graduate students had finished their degree plan at the end of four years (Council of Graduate Rates, 2019). Graduation rates and length of completion are heavily weighed factors when deciding on an institution to pursue a higher education, as college tuition rates increase.

Moreover, the percentage of students who graduate in six years is a federal government reporting requirement (Engelmyer, 2019). Several factors contribute to this, such as student loan debt amount, acceptance rates, local job market, relevancy of degrees offered, and the national economy. This consulting team will analyze academic, economic, and socioeconomic data, correlate possible root causes, and identify solutions for an institution to maintain high enrollment and matriculation rates.

2. Profile of the organization and background of the opportunity

Organization:

Student Advisory Services (SAS)

1234 School Road

Chicago, IL 60606

Overview:

This project will analyze national education and economic data for student higher education advisement. The project will enable SAS to offer student advisors the service to more effectively advise their students. The project will work to accept inputs and preferences from students to advise on the potential institutions to pursue. The service would be available for subscription to education institutions, their advisors, and third parties providing advisement.

3. Research Questions and Key Metrics

Student loan debt amount:

As of 2018, Americans have aggregated $1.56 billion in student loan debt over more than 44 million borrowers. More than 5 million (11.5%) of these student loans are 90 or more days delinquent. Do these rates influence a students inability to complete the program?

Acceptance rates:

Data indicates a thorough vetting process correlates with a higher completion rate in undergraduate programs and is more pronounced on graduate programs. How much can be tied to a school's acceptance rate? If a strong correlation exists, more focus is warranted

Labor market:

Lack of a robust local and national job market demand may discourage completion. What are the offerings in the local market that would encourage or discourage graduation? What is the overall outlook for areas of interest?

Relevancy of degrees offered:

As STEM degrees become more and more necessary in the workforce, are academic institutions keeping up?

National Economy:

The economy may be responsible for the majority of the surges and depressions of degree usefulness and student debt amount. How closely are the two events correlated? Should the outlook of economy support the decision processing of choosing the education opportunities available?

4. Data requirements and data strategies

The analysis discussed above will require education and economic data. The analysis will focus on the data below:

- Graduation Rates, Enrollment, Retention, Post-Graduation Employment
- Student Loan Statistics
- Economic data (Labor, Employment, etc.)

The data required will be sourced from multiple data sources. The strategies for the required analysis will depend on the state and type of the data acquired. The strategy will include one or more of the below strategies.

- Market Sentiment
- Bayesian Classifiers
- Factor Analysis
- PCA
- Clustering
- Regression
- Classification

5. Data Quality Overview

The dataset consists of 1,899 variables across 7,175 rows of varying sources and data types including continuous, discrete, and categorical data. Along with the data is the related data dictionary used for a reference of variables, descriptions, data types, and other information providing context. An summarized view of the data can be found below.

Data Type	Count
Float	1303
Integer	584
String	12

For variable reduction, we decided the best approach would be using a theoretical approach. Since the project is focused on creating a tool that recommends colleges, we are able to eliminate many post completion variables very quickly. For example, we are not concerned about the post completion variables in the categories of root, completion, earnings and repayment. This eliminated 1,383 variables. The variables we considered important to our analysis fall within categories of academics, admissions, cost, school and student attributes. We believe the strongest reasons for choosing a college are based on location, cost of tuition, program availability, admission availability, school size and the school's demographics.

An initial data profile was created to gain further insight into the condition and characteristics of the data. The full profile can be found within the supporting documentation of this report. A number of initial key variables have been provided below as an example.

Variable	Category	Definition	Data Type	Original/Derived	Expected direction of impact on the dependent variable
Institution Name	School	Name of school	String	Original	Response
Operating Institution	School	School is still operating	Binary	Original	+
Distance Education	School	Online school only	Binary	Original	+
State	School	State of school	String	Original	+
Size	Student	Size of student body	Categorical	Derived	+
Gender	Student	Gender of student body	Categorical	Derived	+
Ethnicity	Student	Ethnicity of student body	Categorical	Derived	+
SAT Score	Admissions	Average SAT score of student body	Integer	Original	+
Admission Rate	Admissions	Average admission rate of school	Float	Original	+
In-State	Cost	Average	Integer	Original	+

Tuition		in-state tuition cost of school			
Out of State Tuition	Cost	Average out of state tuition cost of school	Integer	Original	+
Associate Degree	Academics	Type of Associate degree being pursued	Categorical	Derived	+
Bachelor Degree	Academics	Type of Bachelor degree being pursued	Categorical	Derived	+
Certificate	Academics	Type of Certificate being pursued	Categorical	Derived	+
Share of students received Pell Grant	Student		Float	Derived	+
First-time, full-time student retention rate at four-year institutions	Student	Retention Rates	Float	Derived	+

First-time, full-time student retention rate at less-than-four-year institutions	Student	Retention Rates	Float	Derived	+
First-time, part-time student retention rate at four-year institutions	Student	Retention Rates	Float	Derived	+
First-time, part-time student retention rate at less-than-four-year institutions	Student	Retention Rates	Float	Derived	+
Share of Gender (2)	Student	Student Gender ratios	Float	Derived	+
Share of Enrollment for Ethnicity (14)	Student	Student ethnicity rations	Float	Derived	+

Of the dataset consisting of 1899 variables, it was decided we would keep 211 variables. However, many of these variables will be consolidated resulting in 36 total variables. Please note that the size variable is also being derived, but does not need to be consolidated with other variables as the ones below. The size variable will be transformed into a categorical variable (I.E. large, small, etc.) after further exploratory data analysis is conducted. Below is a table that depicts the amount of variables consolidated for the derived variables:

Variable	Category	Definition	Data Type	Original/Derived	Number of Concatenated Variables
Ethnicity	Student	Ethnicity of student body	Categorical	Derived	16 variables
Gender	Student	Gender of student	Categorical	Derived	2 variables
Associate Degree	Academics	Type of Associate degree being pursued	Categorical	Derived	38 variables
Bachelor Degree	Academics	Type of Bachelor degree being pursued	Categorical	Derived	38 variables
Certificate	Academics	Type of Certificate being pursued	Categorical	Derived	114 variable
Median	Earnings	Median	Categorical	Original	4 variables

Earnings		Earnings by years 6,8,9,10 post-graduation			
Earnings by sex	Earnings	Earnings Female/Male for years 6 and 10 post-graduation	Categorial	Original	4 variables

As far as data quality on the original data set, the only issue to note is null records. The data is sourced from existing government data sources with consistent data structures. The data consists of reported data from the registered education institutions. Majority of null values can be attributed to data privacy (suppression) and data sources which is noted in the source files and website. All populated data seems very clean and no transformations need to be performed. After further exploratory data analysis, we will decide how to handle the null records. Below is a chart showing the number of observations and number of null records to better understand the percentage of clean data.

Variable	Category	Definition	Number of Observations	Number of Null Records	Percentage of Clean Data
Institution Name	School	Name of school	7,175	0	100%
Operating Institution	School	School is still operating	7,175	0	100%
Distance Education	School	Online school only	7,175	438	94%

State	School	State of school	7,175	0	100%
Size	Student	Size of student body	7,175	728	90%
Gender	Student	Gender of student body	7,175	728	90%
Ethnicity	Student	Ethnicity of student body	7,175	728	90%
SAT Score	Admissions	Average SAT score of student body	7,175	5,749	20%
Admission Rate	Admissions	Average admission rate of school	7,175	4,950	31%
In-State Tuition	Cost	Average in-state tuition cost of school	7,175	3,111	57%
Out of State Tuition	Cost	Average out of state tuition cost of school	7,175	3,330	57%
Associate Degree	Academics	Type of Associate degree being	7,175	438	94%

		pursued			
Bachelor Degree	Academics	Type of Bachelor degree being pursued	7,175	438	94%
Certificate	Academics	Type of Certificate being pursued	7,175	438	94%

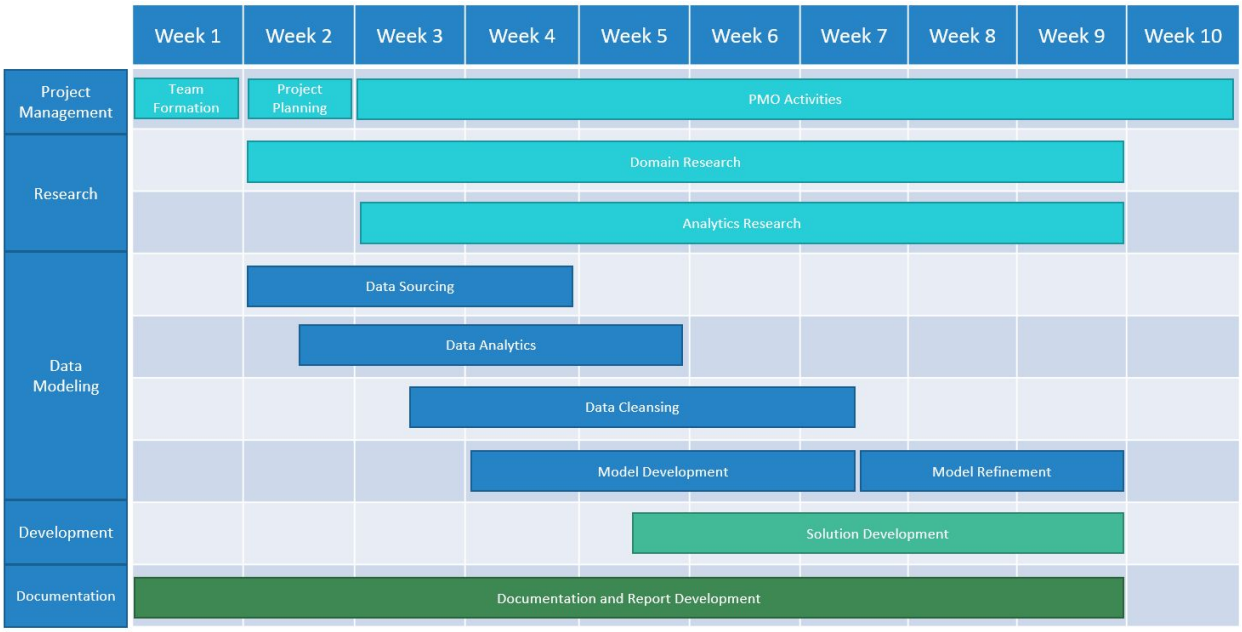
The variables to be consumed are expected to change with addition and reduction. Based on the analytic methods to be used, the data may require creation of columns derived from existing columns (e.g. binary flags). Further data analysis and data processing will lead to iterative variable analysis and reduction.

6. Project Management Plan

Project Management Approach

The project will leverage the Waterfall project management approach. The activities will begin consecutively with concurrent execution. Activities regarding data processing and analytics/model development will consist of iterative tasks. These tasks will be supported by project management, research and documentation tasks. The activities have been planned to the following timeline, cost, and resources.

Project Timeline



Cost

Summary Budget – List component project costs	
Project Component	**Component Cost**
• Personnel Resources	**$137,000**
Project Manager(~120k/yr@10weeks+profit@75%)	$21,000
Solution Architect(~150k/yr@10weeks@75%+profit)	$28,000
Higher Education Consultant(~100k/yr@75%+profit)	$17,000

Business Analysts(C~70k/yr; SC~80k/yr;@10weeks+profit) Developers(DE~70k/yr; DS~110l/yr;@10weeks+profit)	$34,000 $40,000
• Hardware (Hosting Fee)	$10,000
• Software/Licensing (Python Open Source & Existing Licenses Fee)	$5,000
• Travel Budget (18% of Total Fees)	$27,000
Total	**$179,000**

Deliverables

The following deliverables are to be completed as part of the SAS Advisement Analytics project. Any changes to these deliverables must be approved by the project sponsor.

- Project Report – The report will introduce the problem researched, identify key ideas and assumptions, present results, and suggested improvements and next steps.
- Prototype - A prototype representative of scenarios and outputs of the analytic models created.
- GitHub Repository - The repository will contain all documentation, data, and code related to the project.

Requirements

This project must meet the following list of requirements in order to achieve success.

- Analysis must comply with data privacy, discrimination, and other regulatory requirements
- Modeling methods must be researched, reviewed, and validated by the team
- Report of the research, analysis, and results must be delivered and presented
- Supporting documents must be included in final deliverables including data, identified sources, and functioning code

Additional requirements may be added as necessary, with project sponsor approval, as the project moves forward.

Constraints

The following constraints pertain to the SAS Advisement Analytics project.

- All hardware, software, and third party licenses must is limited to those readily available
- No hardware, software, or third party licenses may be purchased without the approval of project leadership
- All data, hardware, and software must be procured, configured, and maintained by project team
- No more than 5 full time project team members with appropriate task assignments
- Strict deadline of June 2nd, 2019

Assumptions

The following are a list of assumptions. Upon agreement and signature of this document, all parties acknowledge that these assumptions are true and correct:

- This project has the full support of the project sponsor, stakeholders, and all departments
- The adherence to regulatory and compliance requirements is valid only for those defined at during the execution of the project
- Any hardware, software, or third party licenses may be used if available and approved by SAS Advisory leadership
- Further subscription, development, and pricing of follow-on services is not in scope for this project.

Risks

The following risks for the SAS Advisement Analytics project have been identified. The project manager will determine and employ the necessary risk mitigation/avoidance strategies as appropriate to minimize the likelihood of these risks:

- Outside development of regulatory and compliance guidelines and requirements
- Complexity of models due to time frame and resources
- Accuracy of data due to open data availability

- Sensitivity to unexpected economic events

Resources

Project Sponsor

The project sponsor is the champion of the project and has authorized the project by signing the project charter. This person is responsible for the funding of the project and is ultimately responsible for its success. Since the Project Sponsor is at the executive level communications should be presented in summary format unless the Project Sponsor requests more detailed communications.

Stakeholders

Normally Stakeholders includes all individuals and organizations who are impacted by the project. These are the stakeholders with whom we need to communicate with and are not included in the other roles defined in this section. The Key Stakeholders includes executive management with an interest in the project and key users identified for participation in the project.

Change Control Board

The Change Control Board (CCB) is a designated group which is reviews technical specifications and authorizes changes within the organizations infrastructure. Technical design documents, user impact analysis and implementation strategies are typical of the types of communication this group requires. The CCB will consist of the Project Sponsor and a representative from each stakeholding party.

Project Manager

The Project Manager has overall responsibility for the execution of the project. The Project Manager manages day to day resources, provides project guidance and monitors and reports on the projects metrics as defined in the Project Management Plan. As the person responsible for the execution of the project, the Project Manager is the primary communicator for the project distributing information according to this Communications Management Plan.

Team

The Project Team is comprised of all persons who have a role performing work on the project. The project team needs to have a clear understanding of the work to be completed and the framework in which the project is to be executed. Since the Project Team is responsible for completing the work for the project they played a key role in creating the Project Plan including defining its schedule and work packages. The Project Team requires a detailed level of communications which is achieved through day to day interactions with the Project Manager and other team members along with weekly team meetings.

Steering Committee

The Steering Committee includes management representing the departments which make up the organization. The Steering Committee provides strategic oversight for changes which impact the overall organization. The purpose of the Steering Committee is to ensure that changes within the organization are affected in such a way that it benefits the organization as a whole. The Steering Committee requires communication on matters which will change the scope of the project and its deliverables.

Sources

Barshay, J. (2019, April 13). Federal data shows 3.9 million students dropped out of college with debt in 2015 and 2016. Retrieved from https://hechingerreport.org/federal-data-shows-3-9-million-students-dropped-college-debt-2015-2016/

Okahana, H. (n.d.). Master's Completion Project. Retrieved April 14, 2019, from https://cgsnet.org/masters-completion-project

Selingo, J. J. (2018, June 08). Why do so many students drop out of college? And what can be done about it? Retrieved April 14, 2019, from https://www.washingtonpost.com/news/grade-point/wp/2018/06/08/why-do-so-many-students-drop-out-of-college-and-what-can-be-done-about-it/?noredirect=on&utm_term=.632b047bd263