Enabling Advisement for Higher Education

PREDICT 498, SEC57, SP2019

1. Demi Constantine - Research, Team Communication, Document Prep
2. Steven Schoiber - Research, Survey Design
3. Braden Schrage - Research, Data Integration, Data Modeling
4. Charles Utt - Research, Data Integration, Data Modeling
5. Christopher Woods - Research, Document Prep

SPS, Northwestern University, Evanston, IL

Table of Contents

Roles and Responsibilities

Team Member	Role	Responsibility
Demi Constantine	Research, Team Communication, Documentation	QC, problem set, research, sync sessions, notes,
Steven Schoiber	Research, Survey Design	Survey methodology, integration
Braden Schrage	Research, Data Integration (R)	Data transformation, variable selection, EDA, k-means, problem set definition
Charles Utt	Research, Modeling (Python)	Steering committee, abstract, EDA, PCA, k-means
Christopher Woods	Research, Documentation	Research sources, methodology, problem statement, QC

Abstract

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).

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.

The issue tree below visualizes a number of issues related to the problem of decreasing student retention. Student retention is defined as ability to retain students in the programs or institutions they have been enrolled. This study works to identify a way to promote higher retention rates through student advisement that incorporates more focus on student preferences in pursuing higher education. By working to better advise students to attend colleges with offerings that best related to the student's needs and preferences, we can better prepare students and advisors for improving retention rates.

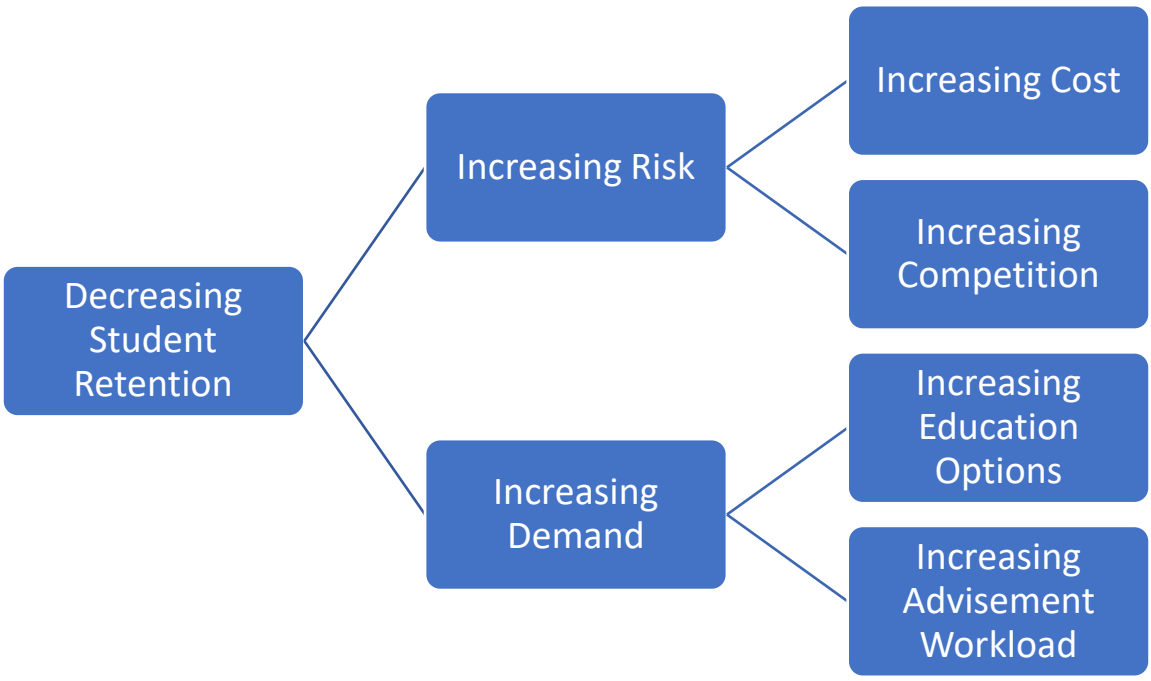


Figure 1: Higher Education Issue Tree

Problem Statement

Problem Description

A rise in high school students seeking to attend four-year higher education institutions has caused a greater demand for guidance counselor advice to select the right place based on a number of variables. With a declining advisor-to-student ratio of 1:367 in 2017, down from 1:282 in 2003, there is a greater need for efficiency in advisement portion part of the high school guidance counselor's role. HiEdu Consultants solves a student ratio challenge for high school guidance counsellors to improve their quality advice output for students by personalizing results for ideal higher education institutions based on a proprietary formula.

Problem Objective

HiEdu Consultants developed a tool which filters several variables to determine a meaningful solution optimized to student needs, preferences, and ideal outcomes. The tool assists guidance counselors, who already have a baseline of knowledge in higher education advising, with a goal of producing relevant and personalized results to accelerate the time to value of advice (i.e. reducing meetings).

After conducting a market scan to determine if other firms were seeking to solve this problem, we found there is no widely available solution. Based on market research assessing the competitive landscape of the education technology market, we found the market size to be $50B+ in 2017 according to research conducted by Navitas Ventures. Below is a market map which highlights the many segments.

HiEdu Consultants offers student advisory services for high school guidance counsellors in the Student Management sub-segment of the Manage segment in the above EdTech market map. As previously stated, there is no head to head competition, however, there are several focusing on various edges of college advice. Some examples include both College Interactive who focuses on bi-lateral communication and profile matching between students and universities, and UniSuccess, who focuses on peer-to-peer advice connecting students attending institutions of interest with prospective (high school) students.

Research

Literature Overview

In order to gain better insight about the problem for which we are solving, HiEdu Consultants reviewed reports, journals, and research to better understand a direction for developing a solution. Hanover Research, by far, had the most relevant and meaningful research as they are one of the market leaders on education consulting.

Hanover's literary work on *Best Practices in Higher Education Retention Strategies* (2017), provided insight into both key variables, such as financial aid, to consider in developing our tool as well as some retention strategies, which helped us level set the importance of student outcomes. In another literary work, *Academic Advising: Strategies for Improving Retention and Completion* (2012), though the focus was on community colleges, the underpinning principles inspired our approach to developing a tool that would be simple, effective, and leveraged by student counsellors to drive efficiency in student advising. In Hanover's report, *Best Practices in Engaging the Next Generation of Students* (2019), we were able to glean insight into the nuances of end-user personas to frame the potential future state of the tool. Lastly, in a piece published in the American Journal of Education titled *Institutional Factors Affecting Student Retention* (2003), we were able to reinforce key variables in the above mentioned 2017 research from Hanover. In sum, the above set the direction for the development of our tool and helped frame key inputs such as questions were drafted for our student surveys.

Additional resources reviewed were for technical development matters related algorithmic selection, such as KMeans, and modelling guidance as we sought to accurately filter relevant results.

Research Questions

The following research questions were defined to related to the higher education institution selection and student success.

Affordability/Cost:

Does cost factor into the decision to pursue and complete higher education?

Future Job Prospects/Labor Market Data:

Does the future outlook on future job prospects or labor market projections influence the pursuit and completion of higher education?

Satisfying major/relevancy of degrees offered:

How do degree offerings/programs influence the student interest in higher education institutions?

Location:

Does preference on the proximity to home regions play a significant role in student higher education decision making?

Student Interests:

Do personal preferences play impact the higher education decision making process?

Institutional Offerings:

Does the offerings of a school impact the higher education decision making process?

Research Metrics

The following metrics have been identified to relate to the success of students.

Retention Rate – the rate of which students enroll and remain at the institution

Completion Rate – the rate of which students complete the enrolled program at the institution

Graduation within 100% of time – the rate that students complete the enrolled program in 100% of the time required

Graduation within 150% of time – the rate that students complete the enrolled program in 150% of the time required

Methodology

Data Analysis

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 ranks colleges, we are able to eliminate many post completion variables very quickly. 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 EDA was run, and found the following information:

Variable	Category	Definition	Data Type	Original/Derived
Institution Name	School	Name of school	String	Original
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
In-State Tuition	Cost	Average in-state tuition cost of school	Integer	Original
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
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	Earnings	Median Earnings by years 10 enrollment	Categorical	Original	4 variables
Earnings by sex	Earnings	Earnings Female/Male for years 10 post enrollment	Categorial	Original	2 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%
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 pursued	7,175	438	94%
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.

Exploratory Data Analysis

The follow presents a selection of EDA conducted on the chosen datasets. Full reports can be found in the Appendix.

A stacked bar chart of type of higher education represented as private/public/other as well as the highest degree offered in each grouping. Private, for-profit schools have the highest representation of the schools. These include institutions like Devry or Trump University.

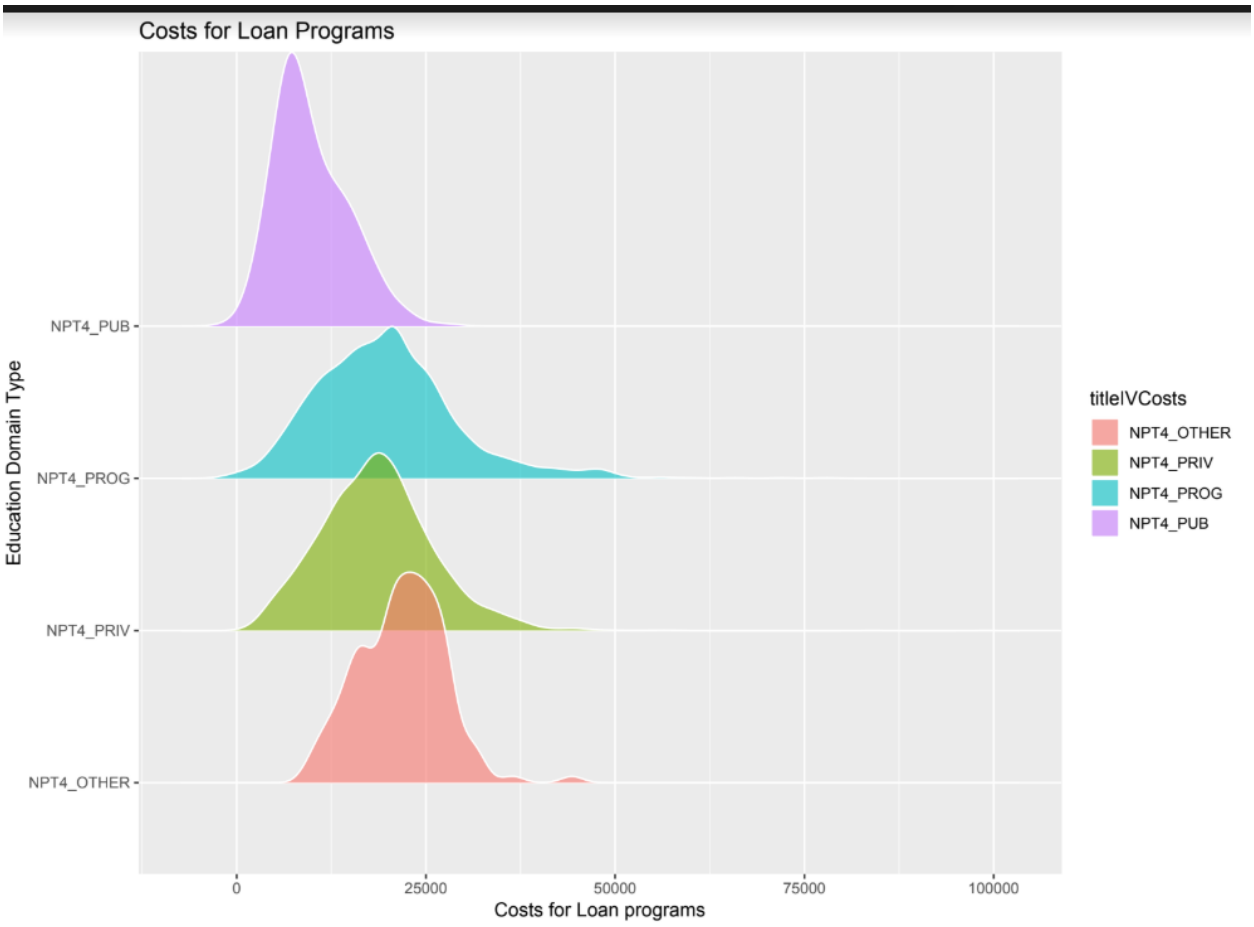


Figure X: Loan Cost by Education Domain

A cluster analysis by gathering the quartiles and variance measures for earnings of students post-graduation. The number of clusters was optimized at six and this is also shown in a supplemental scree-

plot. The cluster analysis shows dendrograms on the left and top of the chart which further shows the strength of relationship between institutions within each cluster. Cluster 3 and 4 is interesting given its small size and low earnings potential. Cluster 5 clearly has the highest earning potential and further investigation into correlation is warranted.

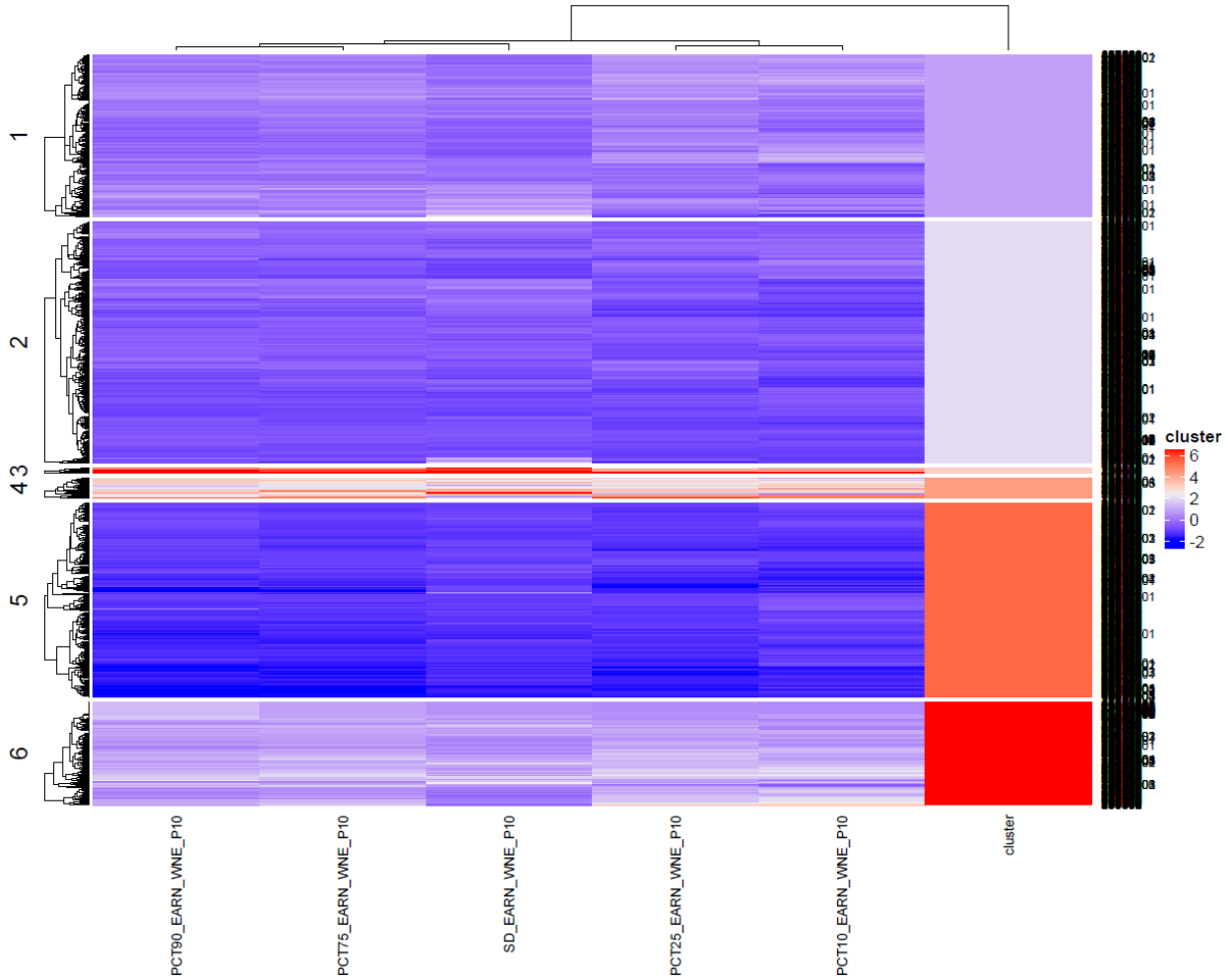


Figure X: Cluster Dendogram

A simple comparison through boxplots of tuition costs in-state and out-of-state. As expected, out-of-state is more expensive.

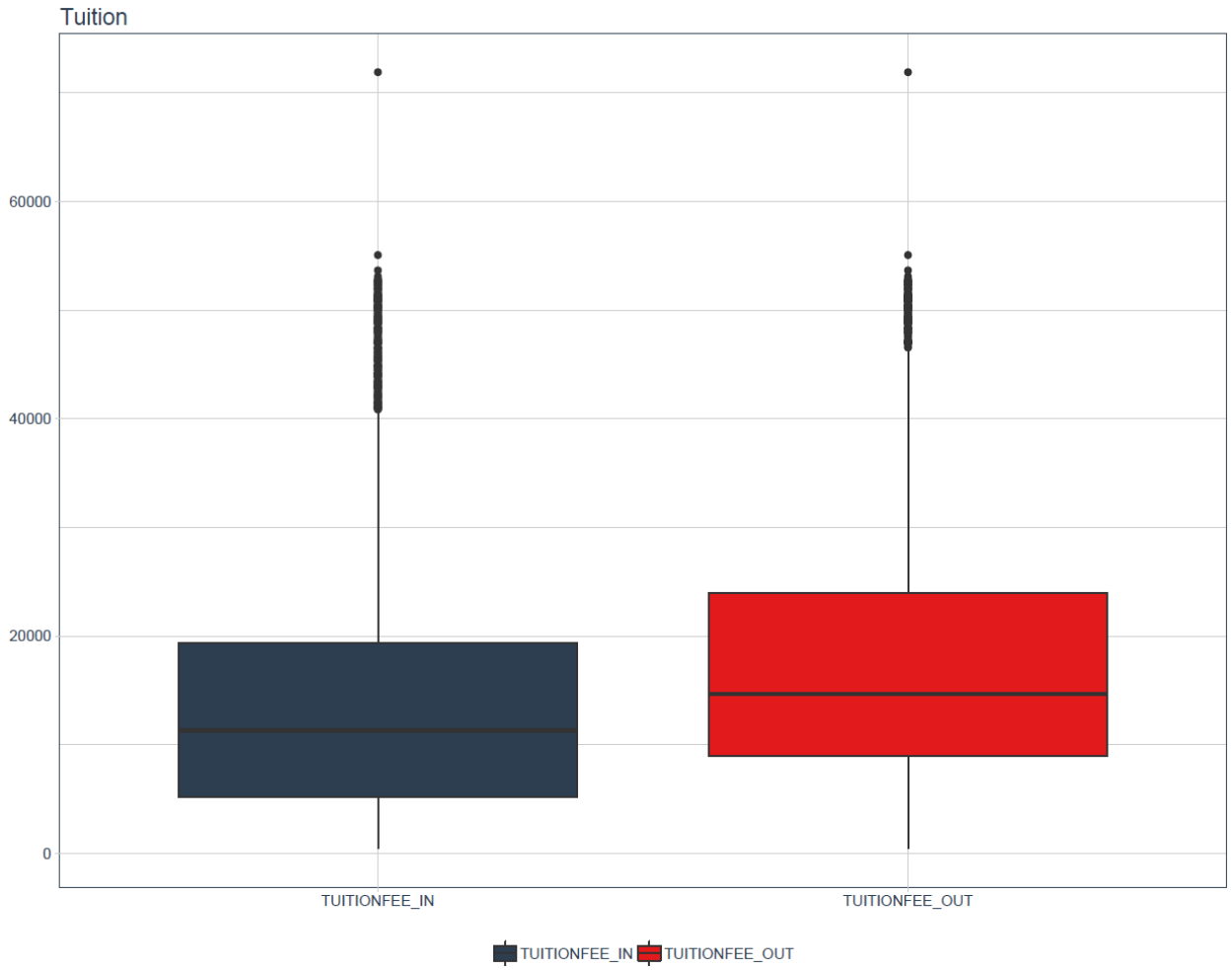


Figure X: Tuition Boxplot

A distribution of gender and earnings potential by private for-profit, public, private non-profit institutions.

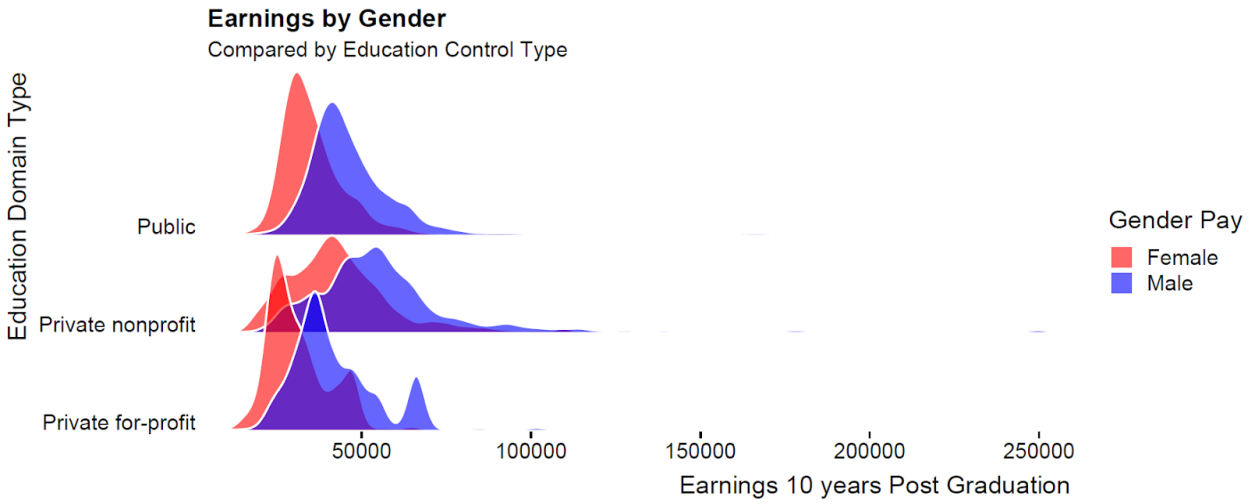


Figure X: Gender Earnings by Education Domain

Survey Design

The survey will focus on the topics on what drives a student's decision for college selection. At a high level, we focused the questions on topics such as school, student body, cost/aid, and academic preferences. The survey will have two types of questions on these subjects. The first type of questions will be ranking questions. The end-user will rate their preferences on a scale of 1-5 (1 being lowest and 5 being highest): For example, how important is it for you to attend a school in your home state? The second type of questions will be questions where students will select an answer from a list. For example, please select the following programs that you are interested in attending. Once the student finishes the survey, again we will feed their results into our model for school recommendation.

Data Integration

The openly available education data acquired must first be cleansed before any processing. The data quality review has identified the largest impediment as missing data. The process below outlines the steps required to gain value from the current data.

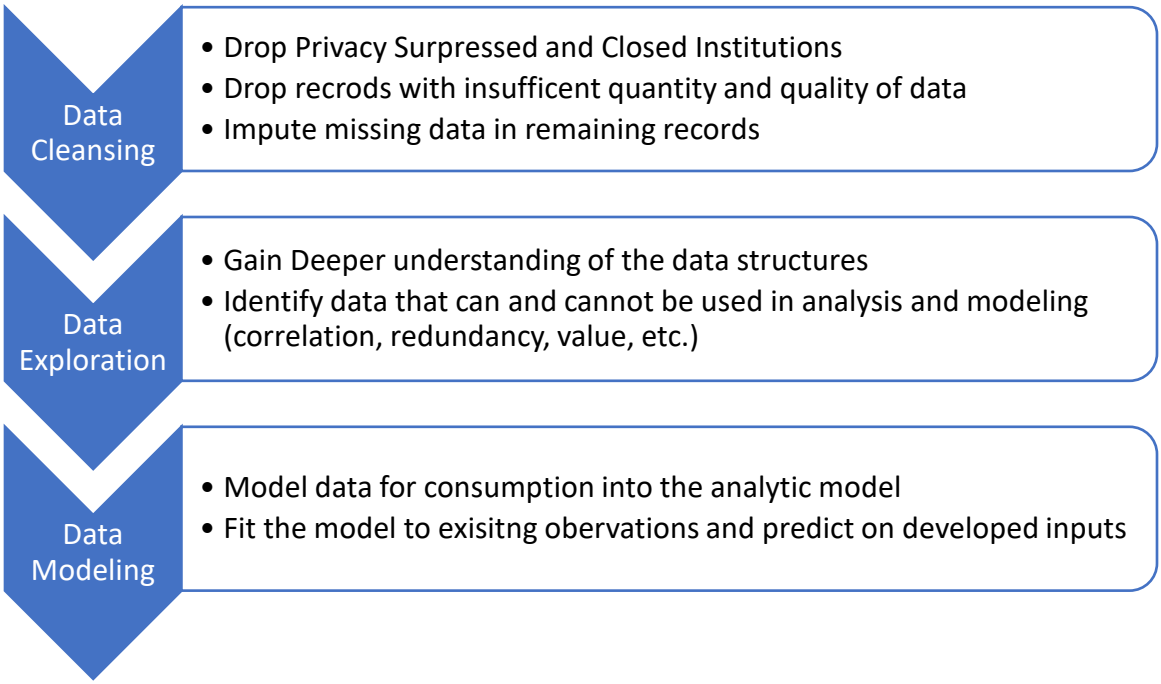


Figure X: Data Integration and Modeling Process

The process covers activities at a high level. It is expected that data cleansing and preparation will require the majority of effort. Cleansing will use techniques such as clustering to impute data. Exploration tasks will consist of defining categories and visualizing distributions among others. Modeling will include data preparation, such as derived columns, and consumption into the model. Outputs of these activities are included in the Appendix section below.

Model Description

The goal of the study is to develop an Unsupervised Learning model to provide the top Higher Education Institution recommendations based on survey inputs. We believe this type of model will best fit the data as the problem is more complex than straightforward classification prediction. The prediction will be based on the consumption of a single, or multiple, records into the model derived from answers to the recommended advisement questions. The model will work to identify the subset of institution groupings that are most similar to the input data. From the subset, further logic will be applied in filtering which will also be derived from the advisement survey (i.e. in-state, distance learning) to provide the recommendations that best fit the student's preferences.

The method used is K-Means Clustering. K-Means works to cluster similar data points around cluster centroids until the it has come to the minimum sum of squared errors (SSE). The modeling effort will involve the identification of the optimal, and most logical, number of clusters. The clusters will be validated using analysis of the SSE and the Silhouette Coefficient, two of the most common ways of cluster validation.

Solution Development

Solution Design

The goal of the model is to output a recommendation of the best cluster that fits the student's preferences. The analytic process will be designed to incorporate survey answers into the input records, execute the model, output a recommended cluster, and incorporate survey inputs to filter to the top recommendations. The process design can be seen below.

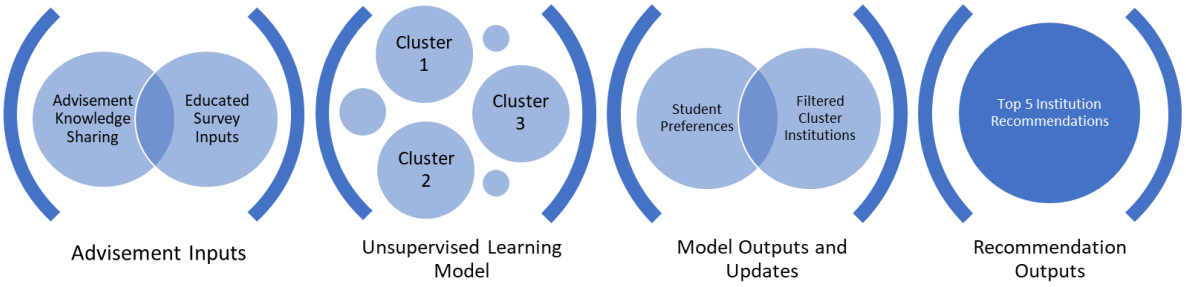


Figure X: Conceptual Design

The process assumes effectiveness in the Advisory process that precede the selection activities. It is assumed that advisors have sufficiently educated students and provided the requisite information to make an informed decision and identify key preferences. The solution will consume the quality inputs of

the preceding advising activities and provide the ability to analyze a much larger set of higher education institutions to reference in decision making.

Survey

Survey questions are used for both modeling and filtering. Questions that relate to the model work to assign a value to the related variables based on the data through association to quantile values of each variable. Questions that relate to filters will be applied to the cluster identified to be most similar to the student inputs. The questions and their use are noted in the table below.

Survey Question	Usage
Please select from one of the following for the state you currently reside in.	Filter - Location
On a scale of 1-5, how important is it to you to attend a school in your home state?	Model – Tuition Cost
Please select from one of the following for your gender.	Model - Gender
On a scale of 1-5, how important is it to you to attend a school that predominantly shares your same gender?	Model - Gender
Please select from one of the following for your ethnicity.	Model - Diversity
On a scale of 1-5, how important is it to you to attend a school that predominantly shares your same ethnicity?	Model - Diversity
Please select from one of the following for your age group.	Model - Diversity
On a scale of 1-5, how important is it to you to attend a school that predominantly shares your same age group?	Model - Diversity
On a scale of 1-5, how important is the cost of the school effect your decision to attend?	Model – Tuition Cost
On a scale of 1-5, how important is it to you to attend a school that offers a high rate of student endowments and or grants?	Model – Institution Investment
Please select from one of the following for the type of degree/certificate you want to complete.	Filter – Institution Classification
Please select from one of the following for the program type you want to complete.	Filter – Institution Classification

Data Integration

Data integration has been developed in Python. Integration includes the loadings and manipulation of data necessary to integrate data in terms of joining separate data sets and formatting for model consumption. Pandas, Numpy, Matplotlib, datetime, and pandas_profiling are the Python packages that are used in data integration tasks.

The first step of the process is to import the data. With data being sourced through CSVs, the Pandas package allows for derivation of data types. This allows for data types to be assumed for quick consumption of the data. Once data is loaded, the columns of relevant columns are grouped for their uses in the solution. The table below outlines the grouping of columns for their use in the solution.

Variable	Source	Usage
Unitid	Completion, Scorecard	All
Student Count	Completion	Modeling
Awards per value	Completion	Modeling
Aid value	Completion	Modeling
Endow value	Completion	Modeling
Grad 100 value	Completion	Modeling
Grad 150 value	Completion	Modeling
Pell value	Completion	Modeling
Retain value	Completion	Modeling
Ft fac value	Completion	Modeling
CCUGPROF	Scorecard	Filter
HBCU	Scorecard	Filter
MENONLY	Scorecard	Filter
WOMENONLY	Scorecard	Filter
TUITFTE	Scorecard	Modeling
INEXPFTE	Scorecard	Modeling
PFTFAC	Scorecard	Modeling
AGE_ENTRY	Scorecard	Modeling
FEMALE	Scorecard	Modeling
UGDS_MEN	Scorecard	Modeling
UGDS_WOMEN	Scorecard	Modeling
UGDS_WHITE	Scorecard	Modeling
UGDS_BLACK	Scorecard	Modeling
UGDS_HISP	Scorecard	Modeling
UGDS_ASIAN	Scorecard	Modeling
UGDS_AIAN	Scorecard	Modeling

UGDS_NHPI	Scorecard	Modeling
UGDS_2MOR	Scorecard	Modeling
UGDS_NRA	Scorecard	Modeling
UGDS_UNKN	Scorecard	Modeling
UGDS_API	Scorecard	Modeling
chronname	Completion	Descriptive, Filter
city	Completion	Descriptive, Filter
basic	Completion	Descriptive, Filter
site	Completion	Descriptive
long_x	Completion	Descriptive
lat_y	Completion	Descriptive
flagship	Completion	Filter

With the data sourced and its usage determined, the data must be cleansed in order for its respective consumption. Previous EDA provides detail on how data can be grouped together on a larger scale for accurate cleansing. EDA also lead to the identification of a more complete dataset with less cleansing requirements. The missing data is imputed with the average while Privacy Suppressed data is dropped completely. The cleansed data, identified usage groupings, and survey design leads to the development of the structure of the model and subsequent input observations.

Model Results

The Sklearn package is used to develop the KMeans clustering model. The initial observations are fed into the model with an initial iterations of 50 models iterating through each value of 1 to 50 clusters. This provides insight to the number of clusters that best fit the data. It is then tested on each value of 1 to 10 clusters while also plotting the SSE of each cluster. PCA is used to reduce the data dimensionality to 2 dimensions allowing the graphical analysis of clusters. Matplotlib is used to plot each number of clusters' SSE to assist in finding the optimal number of clusters using a Scree or Elbow Plot. This analysis determines that 2-4 clusters would be most optimal. The model consisting of 4 clusters is chosen despite it having a less optimal SSE among clusters. Further, the Silhouette Coefficient is used for further cluster validation where models aim for a score near 1. The 4 cluster model is determined to fit the problem

logically with an acceptable SSE and Silhouette Score among clusters in order provide more differentiation among institution clusters.

The following charts provide insight to the validation and final output of the model.

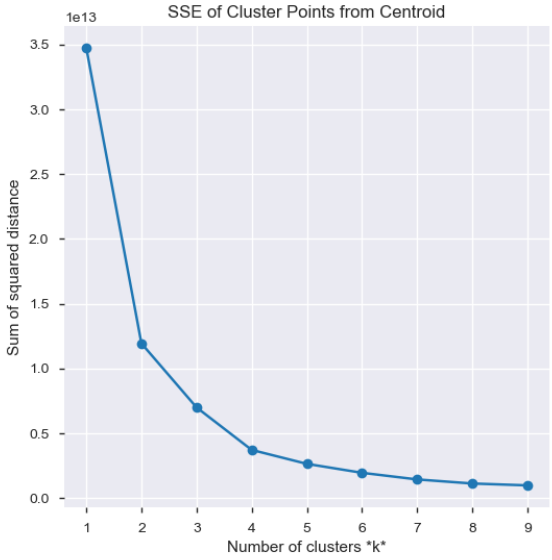


Figure X: Scree/Elbow Plot

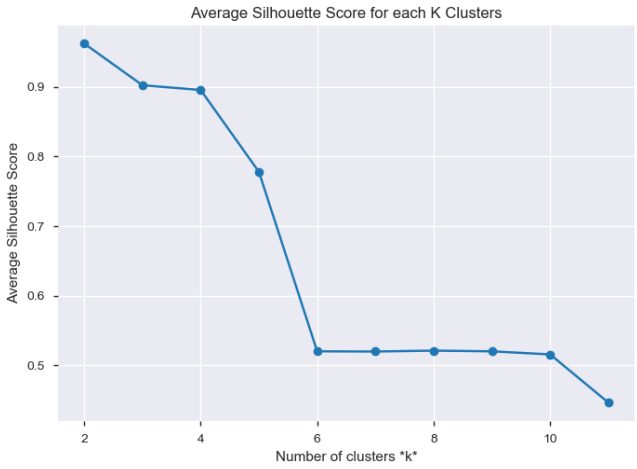


Figure X: Silhouette Plot

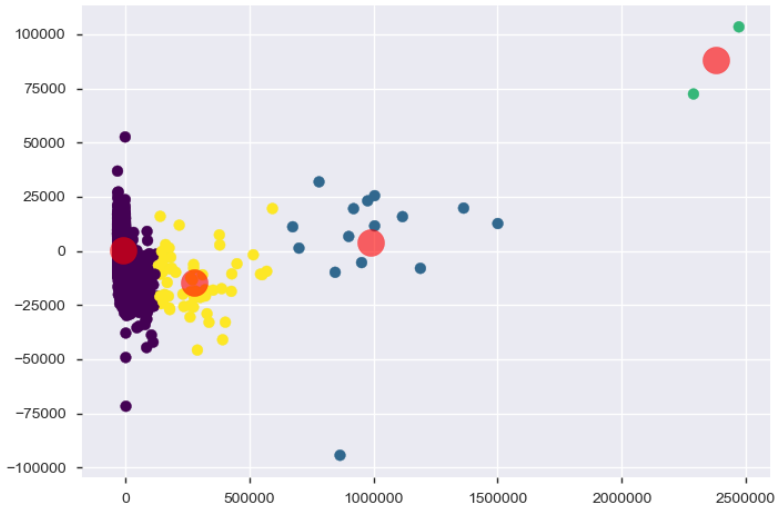


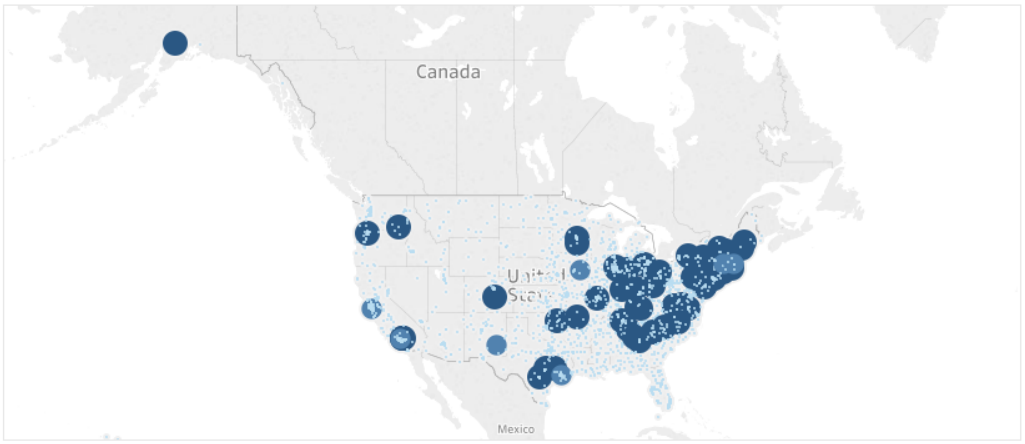
Figure X: KMeans Clusters

With a model developed and validated for accuracy and applicability to the problem. Cluster labels are then joined back to the initial observations. The inputs developed from survey answers are then fed into the model for the clusters that are most similar to the input. The clusters identified are then used obtain the initial cluster of institutions and are then further filtered for the top 5 recommendations ranked to the most important factors for each student.

Dashboard

These top recommendations are then exported and consumed into Tableau. Tableau is the used to develop visualizations of the location of each observation as well as the ranked list itself. The visualizations of a Map and Data Table are developed and combined into a single dashboard. This dashboard is designed to be interactive to provide the filtering of the Map through the ranked data table. This allows the advisor to focus in on each student and provide a single view of the top recommendations, their location, and specified details such as the Institution's name, location, and website. A preview of the dashboard for 3 example records is provided below.

HiEdu Selection Advisement Dashboard



School Ranked List

Student	Cluster	Rank	School	
Student1	0	1	Houghton College	
		2	Westminster College (Pa.)	
		3	Saint Vincent College	
		4	Alfred State College	
		5	Susquehanna University	
Student2	2	1	Stanford University	
		2	Pomona College	
		3	California Institute of Technology	
		4	New Mexico Military Institute	
Student3	0	1	Indiana University East	
		2	Missouri Southern State University	
		3	Purdue University North Central	
		4	Indiana University at Kokomo	
		5	Missouri Western State University	

Challenges

The main challenges faced were scope, resource constraints, and data. Scope refers to the topic area of Retention in Higher Education and the underlying issues. Resource Constraints refers to the constraints on the team such as availability and project duration. Data refers to the lack of depth and quality of data that is openly available through the Department of Education. Each challenged the team in different ways and ultimately limited development of the solution.

There are many potential areas of study as contributors to higher retention in higher education. Issues such as debt, student performance, and enrollment involve proprietary data that is not openly available and protected by a number of regulations. The nature of the issue at hand is complex as there are many different dimension of contributing factors. Many of these dimensions are subject to qualitative and societal, demographic, geographic, as well as environmental factors, and are therefore less measurable or are not currently or openly available for analysis. Further investment and research into the issue of retention will create more opportunity for this solution to be developed.

The project was limited to a duration of 10 weeks. The team was made up of 5 part time resources with varying experience in the topic area. A significant amount of time was spent sourcing data. More time and availability of resources would enable further and more in-depth development of the solution.

Data was a great challenge in terms of sourcing, EDA, variable selection, and overall value. Research had given great insight to the institution-based factors related to student retention. Unfortunately, the most relevant data, the student's individual behavior and ultimate actions, are not openly available to integrate into the model. Further data on student preference, ultimate choice, institution performance, and enrollment demographics, would provide great value to the model.

Though a number of challenges were faced, out team is confident that this solution provides value to advisors and students alike. We believe that the solution should be viewed as a continuous learning

model and solution that should be constantly iterated for improvement as further research and data becomes available. The solution is one that can continue to provide value as the higher education landscape, economy, and student demand changes.

Opportunities

One high value opportunity that is of great interest is the application of the solution to student level data. This would provide more insight to the students and what variables have more individual impacts on their higher education decision making process. Being able to model around more personality and performance variables would provide more effective clusters of student personas. Understanding the people that make up the student body of an institution would be much more valuable than the tracking of the generic concept of a student.

With more focus on personality traits, clusters could be identified among different types of students. This would provide deeper insight to the factors that are most important to students from different backgrounds and come to find the most common and different variables in their decision making. The current model is solely focused on the general concept of a student that wishes to pursue higher education with identified factors being those that are easily measurable without much detail into individual students.

Unfortunately, the data is not available to this study. Data would need to be collected or purchased in large quantities from a diverse population. This effort is out of scope of this project as a quality dataset would take many months or years to develop. With this opportunity presenting the most value in our opinion, the table below is used to identify further opportunities.

Opportunity	Scope	Impediment
Alternative Delivery Methods of Higher Education	Identifying emerging options for higher education in the form of trade schools, MOOCs, and online university platforms among others.	Data Availability and maturation of delivery methods.
Student Performance	Leveraging data of student performance across secondary and post-secondary education for performance based clustering.	Data availability and privacy of student level data.

Recommendations

The goal of the solution is to enable student success through completion of higher education programs by enhancing the advisement process for more informed decision making. The recommendation is the implementation and continuous study and iteration of the solution. With the understanding that data tracking the performance of the solution, measured over a number of years, and responsiveness to data depth, quality, and availability changes, the earlier the adoption of such an analytic solution will contribute to growing the success of higher education. The development of analytic capabilities through such solutions will better enable educators and advisors to promote further success of their students, and ultimately, graduation rates.

With the landscape of data across industries drastically changing, the solution is expected to deliver increasing value with the introduction of further research and data. The solution is a means to more efficiently leveraging the value of data and is able to provide the level of quality for which is fed into the model. Further research into student success factors and the acquisition of data will enable the solution to easily exceed initial investment.

Further, the continued advancement of technology can be leveraged to enable the solution to consume more data in terms of quantity and quality. With rapid developments of technology and supporting platforms for the execution of analytics, integration of the solution with such advancements provides opportunity for more efficient value delivery. Its current development enables future interactions with the use of open-source packages in development.

In summary, the continued development is dependent on the continued investment of our team in collaboration with our clients. Significant investment in educating institutions and stakeholders will be required to achieve the most value and efficiency. As the solution matures and advancements are leveraged, the value will continue to grow with increased success of students in higher education.

Conclusion

In summary, the tool created by HiEdu Consultants solves a problem that is not being directly addressed today by the EdTech market - relevancy and personalized results in four-year higher education pursuits for students and improved efficiency in student advising for high school guidance counsellors. The data used is some of the best available today and is collected and aggregated by the federal government from the institutions. As the institutional data continues to update and evolve, so too will the ability to track student-level data. We will be able to better measure Intended student outcomes (as time will go on) against real-world results (at the earliest in 4-5 years) in order to validate that recommendation made in for high school students were relevant. At a minimum, the value proposition of this tool, given current market conditions, accelerates time to value in both getting and giving advice for four-year higher education.

References

Hanover Research. *Best Practices in Higher Education Retention Strategies.* Hanover Research, 2017, pp. 1–49, *Best Practices in Higher Education Retention Strategies*.

Hanover Research. *Academic Advising: Strategies for Improving Retention and Completion*. Hanover Research, 2012, pp. 1–12, *Academic Advising: Strategies for Improving Retention and Completion*.

Hanover Research. *Best Practices in Engaging the Next Generation of Students*. Hanover Research, 2019, pp. 1–18, *Best Practices in Engaging the Next Generation of Students*.

Lau, Dr. Linda K. "Institutional Factors Affecting Student Retention." *American Journal of Education*, 2003, www.uccs.edu/Documents/retention/2003%20Institutional%20Factors%20Affecting%20Student%20Retention.pdf.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (n.d.). Data Cleaning: Overview and Emerging Challenges [Scholarly project]. Retrieved June 3, 2019, from https://www.cc.gatech.edu/~xchu33/SIGMOD2016Tutorial.pdf.

"k-means++: The advantages of careful seeding" Arthur, David, and Sergei Vassilvitskii, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007).

Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Appendix

Code

All code and related outputs can be found on the project Github under Deliverables, and Code.

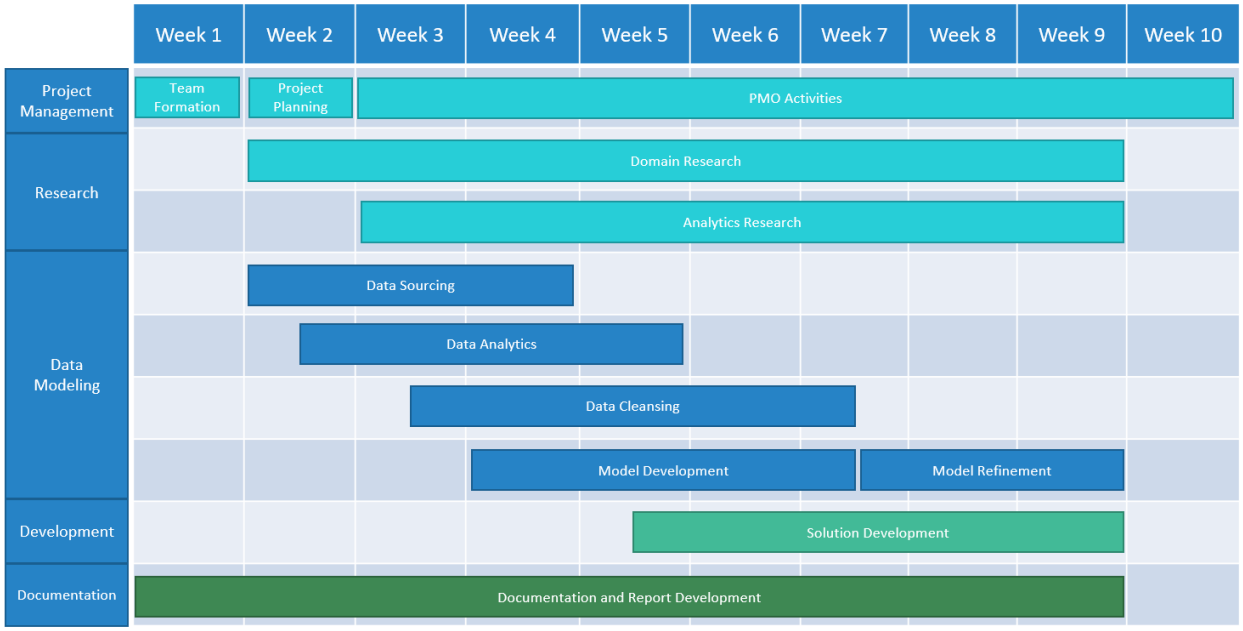
https://github.com/charlesautt/Northwestern-MSDS/tree/master/Capstone_Team1

Project 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) Business Analysts(C~70k/yr; SC~80k/yr;@10weeks+profit) Developers(DE~70k/yr; DS~110l/yr;@10weeks+profit)	$17,000 $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.