**A study of how student-level, major-level, and course-level characteristics are associated with whether a student graduates from FSB**

1. **Abstract**

Student success was measured by graduation within 6 years (12 terms) among Farmer School of Business students at Miami University of an unknown cohort. A logistic regression model was fitted to predict student success by student-level, course-level, and major-level attributes to suggest which attributes most likely affect student success. Attributes found to be strongly related to student graduation or not graduation were Greek life involvement, number of major changes, and first major. Additionally, student GPA scores from their first semester versus all semesters of studies were highly indicative of graduation within 6 years. Exploratory analysis of student success data provided by the student success advising team at Miami University’s Farmer School of Business as well as the fitting of a logistic regression revealed highly suggestive attributes that relate to graduation and provide recommendations for future advising practices.

1. **Background**

The Farmer School of Business (FSB) prides itself on providing a multitude of different resources to best help students succeed. However, the FSB advising office lacks the framework to define and understand the characteristics of a successful student. This research aims to create a standard definition of success, a model to identify at-risk students and characteristics of successful students, and an analytics workflow that the FSB advising team can use in the future.

The client is the FSB student success team whose scope is academic advising, study abroad opportunities, academic goal setting, major and career development, connecting students with campus resources and helping to keep students on track for graduation. Currently, the clients rely on anecdotal evidence to facilitate academic advising, lack a standardized process for collecting student data and have no process that allows identification of at-risk students.

A great deal of research has been conducted on using data mining techniques as a way to discern trends in what makes a student successful at higher education institutions. This is due to the predictive powers inherent in data mining techniques that use relevant information to predict how a student might perform. The framework for this analysis is largely dependent on the data available for analysis provided by the client. Previous research has cited prior academic achievement and performance, social life, family life, and psychological factors as having great importance on student success (Alyahyan & Düştegör, 2020).

This study creates a framework that will help define student success so that future academic advising processes can be driven by research of best practices. By analyzing the student data collected by the FSB’s advising office, a model was developed that could help identify at-risk students in terms of student success. This model can support the advising team in developing interventions for at-risk students that display the highest probability of not being successful. A reproducible workflow was created to provide the student success advising team with a way to iterate over the student success model for future uses in improving the quality of academic advising at the Farmer School of Business. The following research questions are posed for the purpose of this project:

* How can a metric for student success be defined at the Farmer School of Business?
* What major-level, course-level, and student-level attributes are most related to success criteria of graduating with an undergraduate degree within 6 years (12 semesters)?

1. **Data & Methods**

The raw data provided by Alan Clift, a Data and Planning Analyst at the Farmer School of Business, includes student demographic information for a specific cohort of students (unknown to us) that were projected to graduate from the Farmer School of Business. The data does not include details about when this cohort was actively enrolled at Miami University, nor does it identify student names to protect student privacy and eliminate bias in data analysis. The data collected for this cohort spans a timeframe of 6 years (12 semesters) and indicates graduation or no graduation with a degree within this time frame. The raw data consisted of six sets of csv files that described student attributes such as course enrollment, degree attainment status, student demographic information, extracurricular involvement, as well as unique identifiers for each student in the dataset.

To provide meaningful analysis and recommendations, the data were manipulated using RStudio (RStudio Team, 2022). All six datasets were read into an RProject folder that would allow reproducibility of data preprocessing steps across different hard drives. The pacman package (Rinker & Kurkiewicz, 2018) available in the RStudio environment was loaded into the data preprocessing notebook to improve reproducibility and workflow of the packages used. Packages used within the pacman function were DataExplorer (Boxuan Cui, 2020), tidyverse (Wickham et al., 2019), dplyr (Wickham et al., 2022), plotly (Sievert, 2020), zipcodeR (Rozzi, 2021), corrplot (Wei & Simko, 2021), and skimr (Waring et al., 2022). After all packages were loaded using the pacman function, the datasets were manipulated to produce records for each student identified by a PIDM value, or a unique identifier for a student in the given cohort. All six datasets were merged on the PIDM unique identifier to result in a singular dataset called “merged dataset” that would provide information about each unique student.

The resulting merged dataset contained 1180 rows, or records containing information for the number of students, as well as 48 variables that could be used for analysis. Inclusion of variables in the final dataset used for analysis and modeling that provide no value in defining success, or the inclusion of variables that overlap in information creates noise that could skew the data. The packages Scikit-learn (Pedregosa et al., 2011) and pandas (McKinney et al., 2010) were used in a Python notebook to balance the variance and bias (Singh, 2018) that would inevitably be calculated during the modeling phase of predicting student success. Variables that included unique values for each observation (PIDM), variables that were highly correlated with other variables (CT.Super.Class, Greel\_Involvement, Years.to.Degree.Average), variables that contained uniform values (CT.Class, CT.ACE, CT.Federal.Inclusion), and variables that were not informative to the aforementioned student success metric (CT.IPEDS.Race.Ethnicity, CT.Gender, CT.Age, CT.Zip.Code, CT.State, CT.Domestic.Minority, CT.Student.Type, CT.Housing.Status, CT.First.Generation.Student, and Last.Major) were dropped from the cleaned and merged dataset. Additionally, variables that were indicative of the response variable, graduation within 6 years, were dropped to avoid skewing model performance (Cumulative GPA, Degree Count).

In order to reduce the dimensionality of the data and enhance the predictive performance of the model created, variable collapse and dummy encoding was performed using a Random Forest Classifier (Ho, 1995) to predict feature importance of object variables. A benchmark value of 0.05 for all feature importance was established in the selection of dummy encoded object variables. The dummy encoded and feature selected variables included: First.Major, and course grades for core classes (ACC221, ACC222, BLS342, CSE148, ECO201, ECO202, FIN301). The final selection of variables to be used in modeling consisted of the following:

1. **Independent Variables:**

* *CT.First.Semester.GPA.Average:* The GPA obtained during the first term of studies by unique students from the cohort.
* *CT.BEST:* The highest reported ACT score each unique student of the cohort had prior to enrollment at the Farmer School of Business
* *Term.Count.Greek:* The number of terms each unique student participated in Greek Life organizations.
* *Major.Changes:* The number of times each unique student changed their primary major.
* *Distance:* The distance in miles unique students of the cohort lived in comparison to the location of Miami University.
* *Business\_Economics, Undeclared\_Bus, Undeclared, MusicPerformance, Chemistry, Mth\_Ldr, Information\_Systems:* Dummy-encoded first majors selected from feature importance analysis using a Random Forest Classifier compared against the feature importance benchmark of 0.05.
* *BLS342\_B:* A dummy-encoded core course selected from feature importance selection compared to a course grade of A in BLS342.
* *ACC222\_B:* A dummy-encoded core course selected from feature importance selection compared to a course grade of A in ACC222.
* *FIN301\_B:* A dummy-encoded core course selected from feature importance selection compared to a course grade of A in FIN301.
* *CSE148\_F:* A dummy-encoded core course selected from feature importance selection compared to a course grade of A in CSE148.
* *CT.Residency\_Resident:* A dummy encoded classification of U.S. residents as compared to non- residents of the U.S.
* *CT.Honors.Student\_Y:* A dummy encoded classification of students that are a part of the Honors Program as compared to students who are not.

1. **Dependent Variable**

* *Degree.Attained.Indicator:* A classification of 1 if the student obtained their Bachelor’s degree in the Farmer School of Business after 6 years (12 terms).

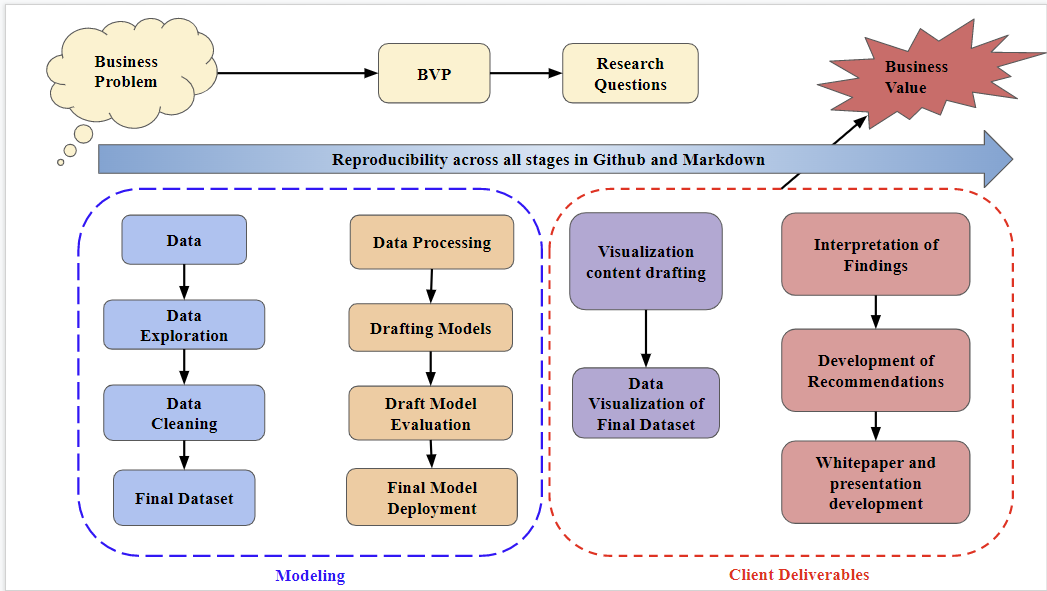
The final dataframe used for subsequent modeling consisted of 19 predictor variables (data type float64, int64, uint8), 1 response variable for the classification of graduation within6 years (12 terms), and 1047 observations of student data. The resulting data frame was split into train (70%) and test (30%) sets to improve model performance in future contexts of providing recommendations to the academic success advising team. Distribution reports about the final data frame revealed an unbalanced dataset with only one tenth of the 1047 observations not obtaining their degrees. The imbalance in the dataset was remedied by using SMOTE in the imblearn package (Chawla et al., 2002). Upsampling of the dataset consisted of randomly redistributing minority classes of data points to oversample the minority class (Soni, 2021).

* 1. **Modeling Framework**

To determine the attributes that are most related to student success at various levels, multiple supervised learning models were fit. The framework used to create such models is visualized in Figure 1.

**Figure 1**

*Modeling Framework*



*Note. A standardized framework for fitting models.*

* 1. **Modeling**

In order to establish predictive power in identifying student attributes associated with graduating in 6 years (12 terms), multiple supervised learning model packages were fit and accuracy, specificity, sensitivity, and misclassification metrics were reported for comparison. The models fit on the training data and evaluated for performance on the test data included Support Vector Classifier (Cortes & Vapnik, 1995), Logistic Regression (Cox, 1958), Random Forest Classifier (Ho, 1995), Naive Bayesian (Pedregosa et al., 2011), and XGBoost (Chen & Guestrin, 2016).

The final model was built on the Logistic Regression algorithm. Logistic Regression is a classification algorithm that can classify a response variable into multiple classes of interest. Logistic regression is most used and best suited for instances when the data in question has binary output. Other examples of instances where logistic regression can be helpful include predicting if an email is spam or not spam and diagnosing if a tumor is malignant or not (Kambria, 2019). Logistic regression allows access to the feature importance rankings, which rank the attributes used to predict graduation within 6 years in terms of how much they contribute to overall model performance. Feature importance rankings promote interpretability of the model’s ability to answer the research questions, giving insights into what factors are associated with student success. Additionally, cross-validation was used within the model to inform the bias and variance in the sampling of the training and test data split. GridSearch (LaValle et al., 2004) was also used to tune the Logistic Regression model to promote efficiency of model deployment as well as overall model performance in terms of predictive power.

1. **Results**
   1. **Model Evaluation**

When determining which model best predicts graduation within 6 years, a balance between specificity (the ability of the model to accurately identify all at-risk students) and accuracy (the rate of whether the model’s prediction is right or wrong for both at-risk and graduated students) is desired. The goal in creating a Logistic Regression model is to inform better identification of at-risk students without risking a high false negative rate, or mistaken identification of at-risk students in terms of success who are not in fact at risk. In other words, the model seeks to cast a big enough net to identify as many students as possible at risk of not graduating without falsely identifying too many at-risk students that could overwhelm the workload of the advising staff.

For comparison, accuracy, misclassification, sensitivity, and specificity metrics are described along with the established model.

**Figure 2**

*Model Performance Comparison*

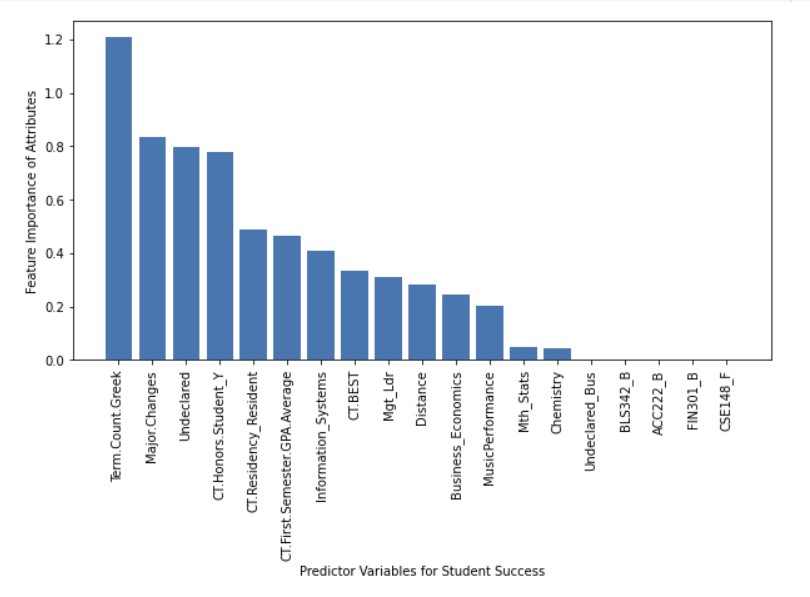
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Accuracy** | **Misclassification Rate** | **Sensitivity** | **Specificity** |
| **Logistic Regression** | 0.65 | 0.35 | 0.62 | 0.83 |
| **XGBoost** | 0.20 | 0.80 | 0.09 | 1.00 |
| **RandomForest Classifier** | 0.51 | 0.49 | 0.46 | 0.94 |
| **Support Vector Classifier** | 0.63 | 0.37 | 0.60 | 0.89 |

The Logistic Regression model has a relatively high specificity, which suggests that out of all the actual students that did not earn their degrees, the Logistic Regression correctly identifies 83% of these students. The model also has a low mis-classification rate of 0.35, which suggests that out of all predictions made, 35% of the predicted results were not correct when compared to actual results. Compared to some of the other models fitted, the Logistic Regression best predicted the students who actually graduated within the 6 year (12 term) timeframe (sensitivity), and best predicted the students that did not graduate (specificity). A model that performed well in terms of specificity, or the accuracy in predicting students who would not graduate from the Farmer School of Business, is critical in informing the clients as to which attributes are associated most to the students’ lack of success, or graduation within 6 years.

The figure below visualizes the attributes predicted to be indicative of graduating or not graduating within 6 years (12 terms) in descending order of feature importance, as predicted by the Logistic Regression model. The number of terms students participated in Greek life organizations, the number of times students change their majors, and majoring “undeclared” compared to the base level first major are predicted to be the most informative to student success. This information can be used to guide the academic advising staff to focus on specific attributes that have a high likelihood of predicting graduation within 6 years.

**Figure 3**

*Feature Importances of Student Success Attributes*



*Note.* Student success attribute importances are calculated from the Logistic Regression model (Cox, 1958).

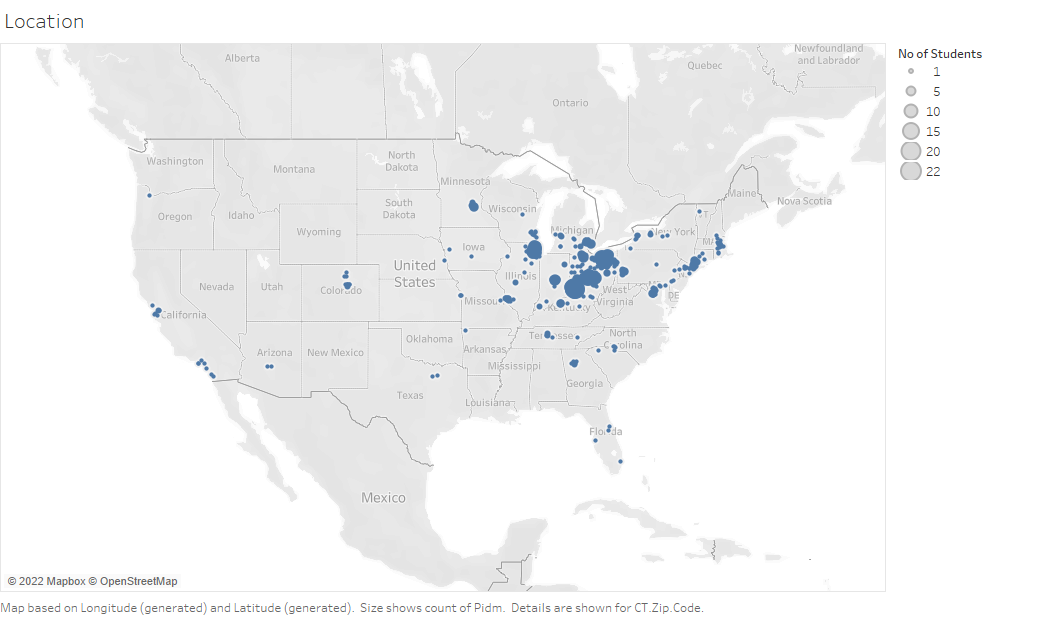
* 1. **Explanatory Dashboard**

The dashboard provides an exploratory data analysis to visualize demographic information about FSB students who are a part of the cohort of study. This demographic information includes gender, race distribution, the distance students’ home location is from Miami, the students’ Greek life involvement etc. This information is further broken down to characterize student-level, major-level, and course-level visualizations and analyses.

Gaining a better understanding of the student body in the Farmer School of Business is valuable in suggesting important factors affecting student success. For instance, analyzing the distribution of students based on their geographic locations reveals much of the student body resides in the Midwest region.

**Figure 4**

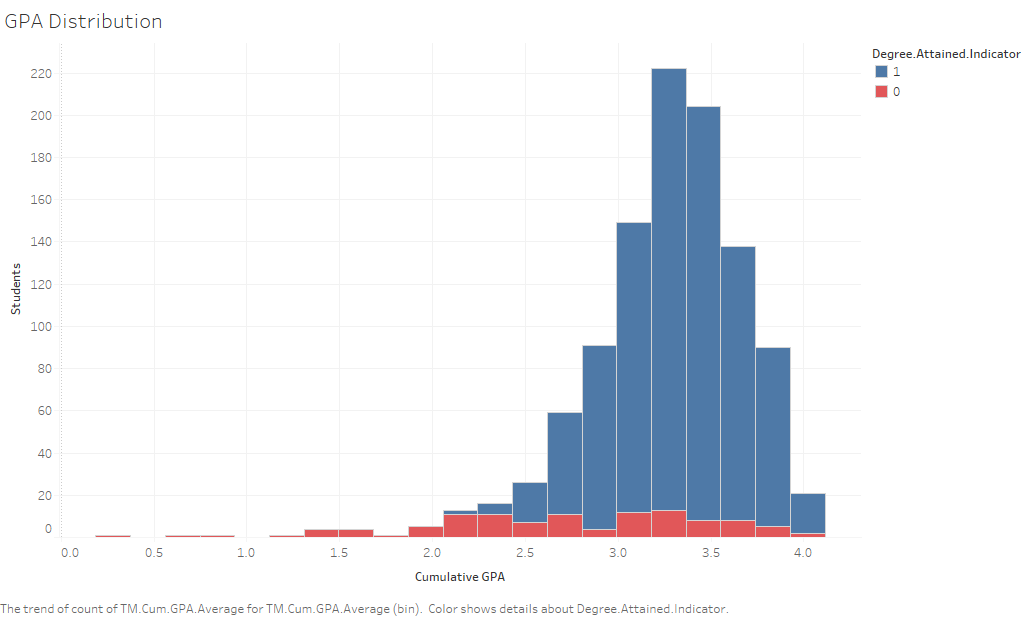
*Student Geographic Distribution by Population*

=

When looking at student success at a granularity of the student-level, the distribution of cumulative GPA can inform the chosen predictor variables for the Logistic Regression. Descriptive analysis suggests that a higher percentage of students who did not attain their degree were associated with having a lower cumulative GPA. The analysis was also honed in on students' GPA from their first semester and further compared to their GPA in later terms. These metrics were compared for students who did and did not graduate within 6 years. The findings suggest that students who had better first semester GPAs may have a higher likelihood of attaining their degree and have a higher overall college GPA, while students who obtained lower GPAs during their first semester were associated with a lower likelihood of graduating as well as obtaining lower cumulative GPAs.

**Figure 5**

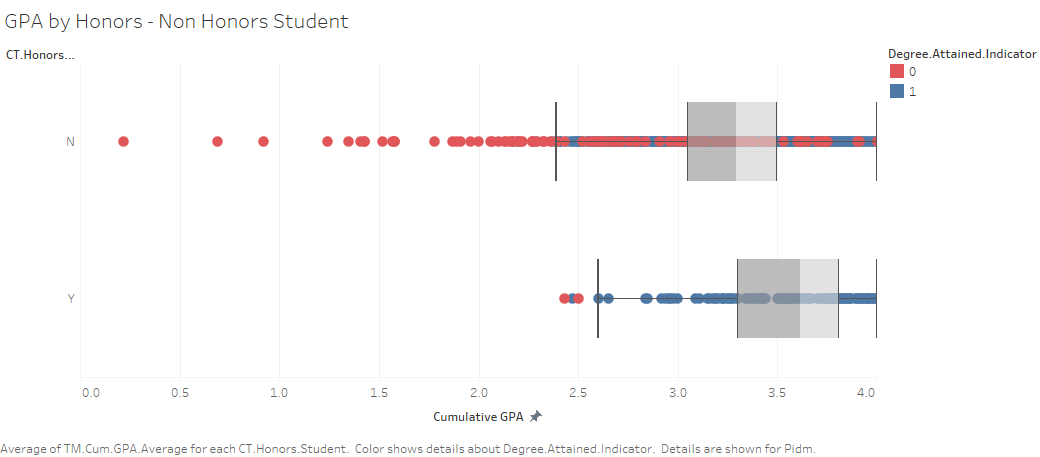
*Cumulative GPA distribution by frequency of students*



Further analysis explored honors program versus non honors program students. This analysis suggested that students who were in the honors program had a higher probability of achieving success as well as a higher likelihood of scoring a higher GPA than their counterparts. The median cumulative GPA for honors students was 3.615, while the median cumulative GPA for non-honors students was 3.29. Contextualizing the GPA metric among honors and non honors students revealed information about the distribution of this data as seen below.

**Figure 6**

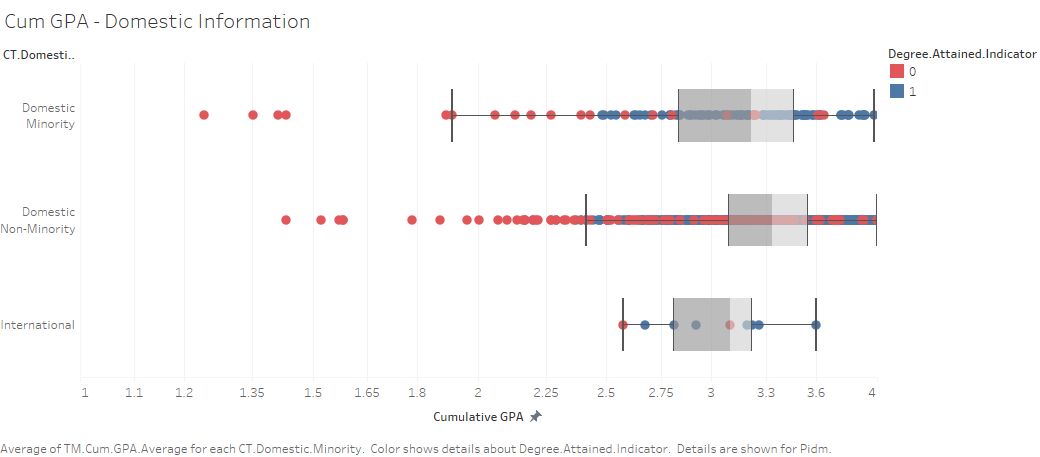
*GPA distribution among honors and non honors students*



Student ethnicity was explored by breaking down the cohort into three categories of Domestic Minority, Domestic Non-Minority and International students. In terms of cumulative GPA, domestic non-minority students performed better than domestic minority and international students with a median GPA of 3.33, while the other two groups had a median GPA of 3.215 and 3.1, respectively. However, this analysis lacked adequate sample size and distribution to appropriately represent the international students that are a part of the Farmer School of Business.

**Figure 7**

*GPA by Residency*



At the major-level granularity, most students were either Finance or Marketing majors. Every major had a small portion of students who did not attain their degree, but there was not one significant major where students struggled more when it came to attaining their degree. The distribution of length of time it took to obtain a degree was on average 4 years. For students that attained multiple degrees, four years was also the typical amount of time it took to achieve two degrees.

Box plots were created to visualize the spread of cumulative GPA based on each major, while comparing students who attained their degree with students who did not. In general, students who had a lower cumulative GPA were less likely to graduate with their degree. Also, there were some FSB major students with a high cumulative GPA that still ended up not attaining their degree. Majors Business Economics and Management and Leadership seemed to have the lowest quartile of GPA scores, with more students who did not attain their degree around those points. Each major had data points of students below the third quartile of cumulative GPA scores, most of whom did not attain their degree.

**Figure 8**

*Cumulative GPA by Major*



When looking at student success by course-level, the focus was on performance in specific preliminary, lower level courses that are required to be taken in order to graduate, sorted by student’s major. These courses are ECO 201, ECO 202, ACC 221, ACC 222, CSE 148, FIN 301, and BLS 342. Student’s performance was classified by letter grades A ,B, C, D, F, credit, no credit and W (Withdrawal from the course or university). This data gives a better estimation of whether students are at risk of not graduating based on their grades in these specific courses.

* 1. **Insights & Recommendations**

Defining a metric for student success provides a framework for the clients to use across the board in their future student support resources. Setting parameters on such a broad term as success must consider the unique and diverse experiences of all students at Miami. For this reason, the success metric was defined as graduation within 6 years (12 terms). Descriptive analysis and the deployment of a Logistic Regression supervised learning model are able to inform evidence-based decisions among the student success advising team right now and in future development of student resources. The descriptive analysis of student attributes suggests involvement in social organizations such as fraternities and sororities positively influence student success. Additionally, certain majors within the Farmer School of Business have a higher proportion of students that do not attain their degree within 6 years than others. Finance, marketing, and business economics majors comprise this larger proportion of students not attaining success and may require a more targeted advising approach. The Logistic Regression model suggested similar findings such that involvement in fraternities and sororities was related to a high probability of graduation within 6 years. A change in major and an undeclared major were also suggestive of a high probability of graduation within 6 years. The information gathered from the descriptive analysis and the supervised learning model suggest that the student success advising team should consider student social organization involvement as well as career exploration initiatives in their resource development.

1. **Conclusion**

The data provided by Alan Clift, a Data and Planning Analyst at the Farmer School of Business, includes student demographic information at levels of student, course, and major attributes. This data was used to develop a solution that could provide evidence-based recommendations to the FSB office of student advising on how to improve their advising practices to promote student success. Currently, the office of advising relies heavily on anecdotal evidence to guide their practices and development of resources. This project provides an iterative framework for improving academic advising at the Farmer School of Business for future cohorts by defining a success metric, deploying a model that can predict attributes of successful students, and creating an explanatory dashboard that can characterize successful and at-risk students. This project was conducted with a reproducible workflow using RMarkdown and Github softwares to allow this solution to be timeless. Not only will the office of advising be able to immediately apply the recommendations of promoting career exploration initiatives and social organization engagement, but they will also have a framework for deploying future analyses with different student attributes and different problems that the future of education may hold.

1. **Limitations**

The data provided contains variables that focus on students' academic performance at Miami University as well as internal factors, but lacks external factors such as student employment or information about student satisfaction that can be more difficult to quantify. Some factors that the data failed to fully encapsulate include academics in high school, prior to students enrollment in the university, social and family life, or any financial considerations. Some of these factors may have had a correlation to student success, but were not provided or able to be collected. Web scraping data that characterized these external factors could have been done to increase the number of meaningful variables. Additionally, many of the variables provided by the clients had to be dropped from the analysis because they were highly correlated with other variables, had uniform values, or had a large amount of null values. Such variables could have promoted model performance had they contained more variability.

**Works Cited**

Alyahyan, E., & Düştegör, D. (2020, February 10). *Predicting academic success in Higher Education: Literature Review and best practices - international journal of educational   
technology in higher education*. SpringerLink. Retrieved November 20, 2022, from   
https://link.springer.com/article/10.1186/s41239-020-0177-7

Boxuan, C. (2020). DataExplorer: Automate Data Exploration and Treatment. R package version 0.8.2. https://CRAN.R-project.org/package=DataExplorer

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Kambria. (2019). *Logistic regression for machine learning and classification*. Kambria. Retrieved November 14, 2022, from https://blog.kambria.io/logistic-regression-for-machine-learning/

Kuh, G. D. (2007). *"Piecing together the student success puzzle: Research, propositions, and   
 recommendations"* (5th ed., Vol. 32).

Kuh, G. D. K. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2006, July). *What   
 matters to student success: A review of the literature*. Retrieved November 20, 2022,   
 from https://nces.ed.gov/npec/pdf/Kuh\_Team\_Report.pdf

LaValle, S. M., Branicky, M. S., & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, *23*(7–8), 673–692.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & others (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825--2830.

Rinker, T.W., Kurkiewicz, D. (2018). *pacman: Package Management for R*. version 0.5.0. Buffalo, New York. http://github.com/trinker/pacman.

Rozzi, G.C. (2021), zipcodeR: Advancing the analysis of spatial data at the ZIP code level in R, Softw. Impacts. (2021) 100099.

RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

Sievert, C. (2020). Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida, 2020.

Soni, P. (2021, March 2). *Handling Imbalanced Datasets with imblearn Library*. Medium. Retrieved November 19, 2022, from https://medium.com/thecyphy/handling-imbalanced-datasets-with-imblearn-library-df5e58b968f4

Singh, S. (2018, May 21). *Understanding the Bias-Variance Tradeoff* . Retrieved November 19, 2022, from https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). “Welcome to the tidyverse.” \_Journal of Open Source Software\_, \*4\*(43), 1686. doi: 10.21105/joss.01686 (URL: https://doi.org/10.21105/joss.01686).

Wickham, H., François, R., Henry, R., & Müller, K. (2022). dplyr: A Grammar of Data Manipulation. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr

Waring, E., Quinn, M., McNamara, A., de la Rubia, E.A., Zhu, H., & Ellis, S. (2022). skimr: Compact and Flexible Summaries of Data. R package version 2.1.4. https://CRAN.R-project.org/package=skimr

Wei, T., Simko, V. (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot