PROJECT TITLE

AUTHOR

ABSTRACT

The higher education sector has vast amounts of unexplored data that can be used to drive decision making and better inform student success initiatives for undergraduate students. Student success can be measured objectively by using historical enrollment data to predict future-term undergraduate course outcomes for students enrolled in a degree seeking capacity at a higher education institution. The objective of this project is to predict whether a student will pass or fail an undergraduate course based on historical enrollment attributes, demographic data, and specific course grades. This information has the potential to facilitate the development of targeted student success programs to help students proactively by providing them with the support and resources they need to succeed.

INTRODUCTION

In recent years, there has been a slow but steady shift in the education sector towards a more data-driven decision-making process. A great example of this is that in 2014, the Florida Board of Governors (BOG) approved a Performance Based Funding Model that was built on ten metrics that would be used to evaluate all universities in the State University System (SUS). The metrics aligned with the SUS strategic goals and provided a data-based method of objectively evaluating each institution. This increased the need for educational data mining for universities in the system [1]. Educational data mining focuses on improving learning outcomes by collecting and performing analysis on education data. Much like other fields that use data mining to improve targeted metrics like increased revenue or reduced customer churn, educational data mining aims to improve student performance metrics such as retention, graduation, and course pass rates [2]. Due to the implementation of the BOG performance model, student success became an area of focus in the higher education space, especially for students in their undergraduate careers.

Student performance in undergraduate courses is important not only for the purpose of pushing the needle on a metric but also due to the impact that this first exposure to college has on the future success of a learner. Research shows that the drop-out rate for undergraduate college students in the United States is 40%. One of the reasons stated for dropping out of college is being unprepared for the academic workload that the student is required to take. These students are then discouraged and delayed in their academic careers by having to take remedial courses which often leads to financial impacts and late graduation [3]. It is for this reason that by focusing on student success initiatives will not only lead to a more prepared student overall but also to an increase in BOG metrics as an effect of good student performance.

MOTIVATION AND PROJECT APPLICATIONS

As previously mentioned, focusing on undergraduate student success is important to ensure that the learner is comfortable with college-level course work. This can be achieved by putting in place a pro-active approach to course enrollment as well providing those students who need support with access to the appropriate resources. However, the only way to do this is to be able to identify those students who are at risk of failing a course at the time of enrollment rather than waiting for end-of-term course results. By placing the point of intervention prior to the actual failing of a course, it allows the university to allocate the appropriate resources and personnel to help the student or provide an alternate option with a higher chance of success. This will not only aid students be better prepared but will also help the university boost metrics based on course- performance. This project aims to utilize historical enrollment data and course-specific grades to predict future-term student course performance by classifying them into a pass or fail category [4]. Although FIU has performed research studies on at-risk students [5], there is not at this time and to my knowledge any university wide predictive modeling available to identify at-risk students at the course level. This project will attempt to bridge that gap.

There are various potential applications for this type of predictive modeling, two of which are explored below:

1. **Course recommender system**: since this type of predictive modeling will use student historical enrollment data to predict future-term performance, it can be embedded as part of the platform students use to enroll in courses to create a personalized course-difficulty rating (on a low, medium, or high scale) without telling the student there is a potential. It would instead, require an appointment with their advisor if the difficulty is 'high'.
2. **Student at risk alert system**:
    a. **Advisors**: this model can be used as part of the student information system, allowing the advisors to be able to quickly get a sense of how challenging a course may be for a particular student. This would provide the advisor the opportunity to offer the student resources available to them, guide potential course selection discussions, and facilitate early intervention to avoid course failure.
    b. **Instructors**: this model could be used to signal to an instructor which students in his or her course are at risk of failing. A process could be put in place that would allow the instructor to support the student with the additional help or guidance on how to best success within the course. This could also inform the instructor on guide potential course modifications in that future that might help at-risk students perform better.

OBJECTIVES AND DELIVERABLES

There are two main objectives of this project, and these are detailed below.

TABLE 1: OBJECTIVES AND RESEARCH QUESTIONS

	Objective 1	Objective 2
	Build a classification model to identify students at risk of failing	Identify factors that are influential to a student's success
Question 1	Can a student's historical undergraduate course performance be used to predict future-term outcomes?	What are the most influential factors to student success in an undergraduate course?
Question 2	What is the best performing model for this classification task?	Is performance in gateway courses and university core curriculum courses influential in students' future-term course performance?

The main deliverable for this project is a completed classification model that can classify a student under one of two labels: pass or fail. This model could then be used in a variety of ways, two of which are described in more detail in the section titled **Motivation and Project Applications**.

TOOLS USED & DATA COLLECTED

TOOLS

All preprocessing, model building and model testing for this project were run using Python 3.0 on Jupyter Notebook on a Macbook Pro running on a 2.3 GHz 8-Core Intel i9 processor with 16 GB of RAM and 1 TB hard drive. Tableau Desktop version 2021.3 was used for preliminary visual exploratory data analysis as well as for creating the visualizations for the project results. The following Python libraries were utilized for preprocessing, the bulk of exploratory data analysis, and model building and testing:

- SciKit-learn
- Imblearn
- Pandas
- Numpy
- Uuid
- MatplotLib
- Seaborn

DATASET

The data used in this experiment comes from Florida International University's student information system, Panthersoft (a re-branded version of Oracle's Peoplesoft). The process to obtain this dataset had to be initiated through the university's Institutional Review Board, approval was granted by both the Chief of Academic Administration as well as the university's Registrar. This request was then forwarded to the department of Accountability and Information Management (AIM) who ran the required queries and delivered the data in a comma separated (.csv) format.

The raw dataset consisted of 637,327 unique student/course pairings with 48 attributes. This included 38,738 unique undergraduate students with 2,899 different courses across the span of 17 unique enrollments terms running from Fall 2014 to Fall 2021. For the purposes of this project, only three colleges are examined and only two departments within each of these colleges are part of the experiment. The colleges and departments included are detailed in **Table 2** below.

TABLE 2: COLLEGES AND DEPARTMENTS INCLUDED IN EXPERIMENT

College	Department
College of Arts, Science, and Education (CASE)	Psychology
	Teaching & Learning
School of International and Public Affairs (SIPA)	Management & International Business
	Finance
College of Business (COB)	Criminal Justice
	Politics & International Relations

Although the dataset might seem to be extensive, it is important to note that not all observations were used and there was some repetition in the data source columns which inflates the number of attributes actually available for the experiment. In the end, the size of the dataset was severely reduced once the necessary constraints on the data were put in place and cleanup was performed (more on this **Section 3.3 Preprocessing**).

The dataset consists of structured data that is enrollment-based, meaning every row is an enrollment instance consisting of a unique student, unique course offering (course prefix, course number and course section number), and unique term combination. This means that each student appears more than once, with the upper bound of instances per student being set by the total number of enrollments in the time frame being studied. The dataset contains a variety of student-course attributes, student-only attributes and course-only attributes with the majority being student-course attributes such as enrollment terms, course enrolled, grade obtained, student admit type and GPA. Although the data was provided in csv format, this initial file was converted to Excel's Strict Open XML Spreadsheet (.xlxs) format and then imported and analyzed using Python 3.0.

Due to the fact this data contains personally identifiable information (PII) it is only able to be shown in an aggregate fashion. For the purposes of this experiment, the data was anonymized prior to being used. This will be discussed in **Section 4. Implementation**. To provide a more in depth view without compromising student anonymity, the visualizations below offer a summarized look at the raw dataset prior to any processing. From **Figure 1** below, it's easy to see that the student distribution by gender is over 50% female, that the most common ethnicity is Hispanic and that most students are from Florida (in-state). These student demographics help to better understand the type of learner this dataset represents.

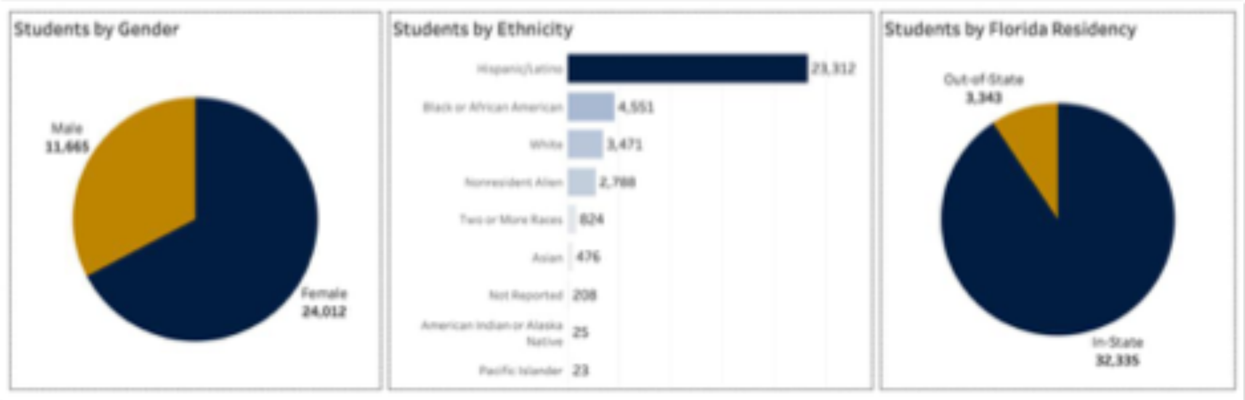


FIGURE 1: STUDENT DEMOGRAPHICS

Next, looking at the types of courses and modalities included in the dataset is also important. **Figure 2**. provides an overview the types of courses represented in the data as well as the modalities of the courses taught. It is of note, that this dataset did not include instructor-based attributes.

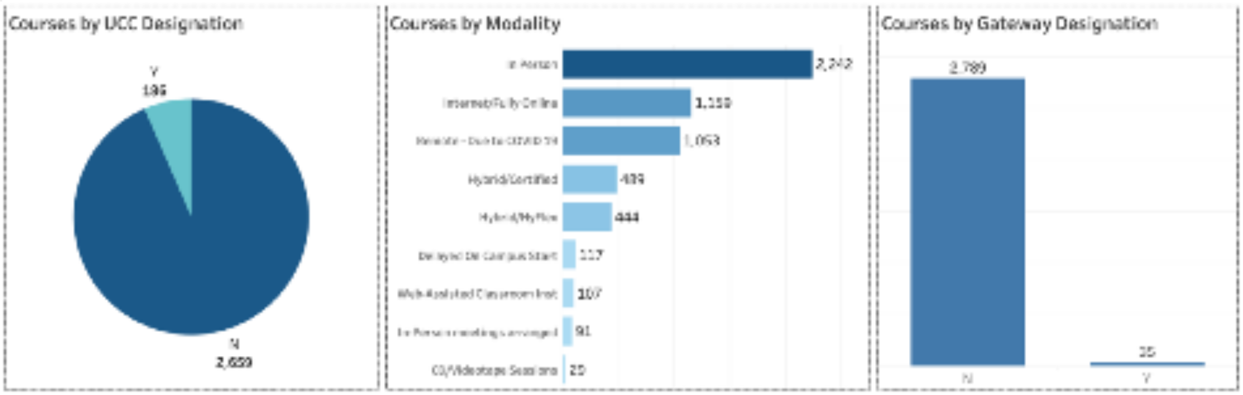
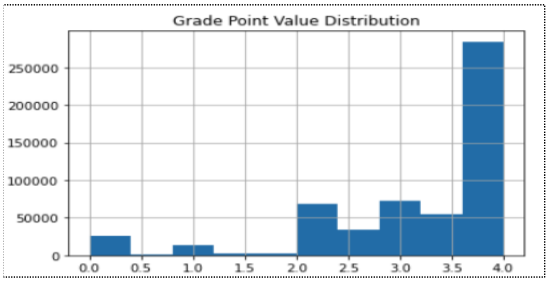


FIGURE 2: COURSE HIGHLIGHTS

Finally, **Figure 3** and **Figure 4** contain information on how the student population is distributed across colleges and what the grade distributions look like. It should be noted that student counts across college are not necessarily unique. As this dataset encompasses seven academic years, there is a possibility that students could have switched majors and therefore appear in counts for two different colleges.



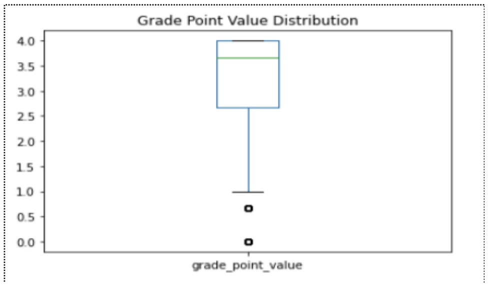


FIGURE 3: GPA AVERAGE DISTRIBUTION

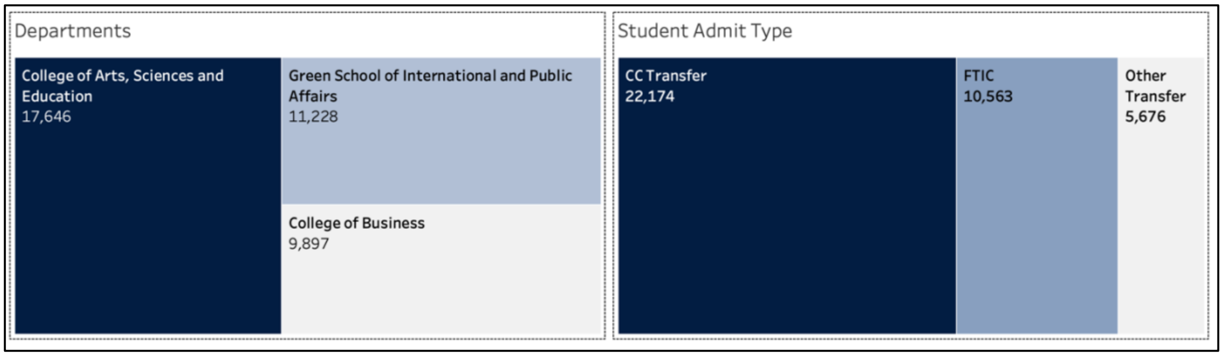


FIGURE 4: STUDENT DISTRIBUTION ACROSS COLLEGE AND ADMIT TYPE

IMPLEMENTATION

The experimental approach is performed in four stages as shown in the figure below. The stages will be explained in detail in the following sections of the report.

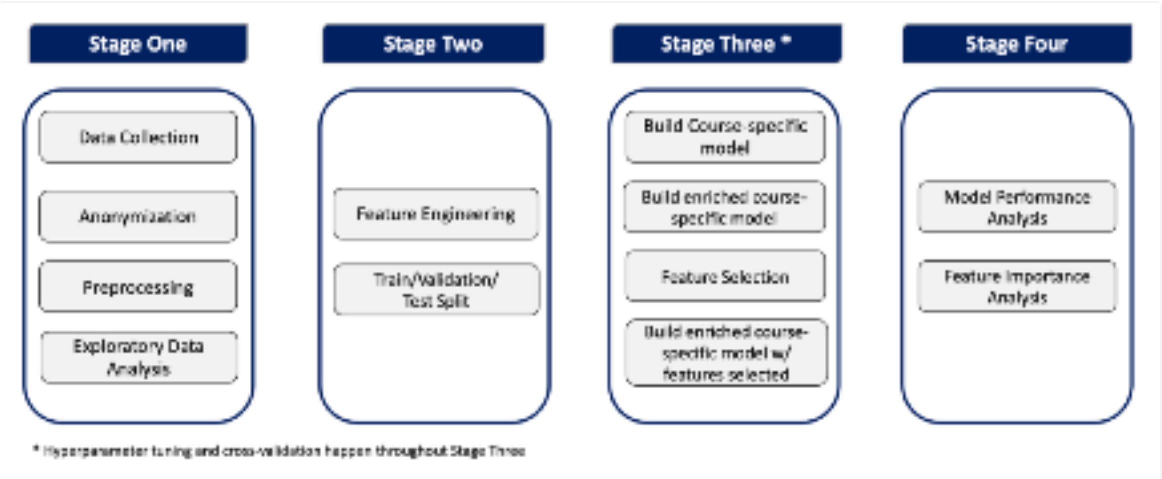

Stage One
Data Collection
Anonymization
Preprocessing
Exploratory Data Analysis
Stage Two
Feature Engineering
Train/Validation/ Test Split
Stage Three *
Build Course-specific model
Build enriched course-specific model
Feature Selection
Build enriched course-specific model w/ features selected
Stage Four
Model Performance Analysis
Feature Importance Analysis
* Hyperparameter tuning and cross-validation happen throughout Stage Three

FIGURE 5: EXPERIMENT IMPLEMENTATION IN STAGES

Stage one mainly consisted of data collection, anonymization, preprocessing, and exploratory data analysis. Stage two consisted of feature engineering and a manual train/validation/test split which will be discussed in detail **Creating Training/Validation/Test Splits**. Stage three entailed the evaluation of the data to pick the best attribute set to use for this experiment as well as the evaluation of three machine learning (ML) models with the objective of obtaining the best classification performance. Finally, stage four was the analysis of the results for the final selected model as well as the analysis of feature importance using the final selected model and the feature importance selection methodologies detailed in **Feature Importance**.

PREPROCESSING

As previously mentioned, this data contained personally identifiable information (PII) and to be able to use it for this project it first had to be anonymized. This was done using the UUID Python library which aids in creating unique identifiers by generating them using synchronization methods. It was important that the anonymization was reliable as well as consistent. This is because it was necessary to protect student privacy while still maintaining data integrity. Once the UUIDs were created, they were mapped to a unique Panther ID from the raw data and a separate excel file was created to store these mappings. This file was only kept as reference to avoid having a different UUID every time the script was run, but it was never used as part of the analysis or the experiment.

After the data was anonymized, clean-up had to be performed to ensure that the final dataset used for testing contained unduplicated, reliable data that was as concise and free of noise as possible. The steps to achieving that described below:

1. Removing duplicated attributes: the dataset contained multiple columns with the same information. For example, the enrollment term id appeared three times in three different formats. One column was a string and contained the textual description of the enrollment term (i.e. Fall 2020), the second column contained the numerical representation of the enrollment term (i.e. 1208), while yet a third column contained a different but equivalent numerical representation (i.e. 202008). This only served to add noise, so two out of the three columns were removed and the simple numerical representation of the term was kept.
2. Handling NULLs: although most attributes were under 20% NULL, there were some instances in which NULL were the majority. Some examples are the TOT_SAT and TOT_ACT scores; these attributes relating to standardize testing were removed. There was no imputation of NULL values in this particular experiment setup as it was important to capture the absence just as much as the presence of a value.
3. Handling NULLs: although most attributes were under 20% NULL, there were some instances in which NULL were the majority. Some examples are the TOT_SAT and TOT_ACT scores; these attributes relating to standardize testing were removed. There was no imputation of NULL values in this particular experiment setup as it was important to capture the absence just as much as the presence of a value.

After clean-up was complete, the next step in the process was performing preliminary data analysis to get some insight into what the dataset contained.

EXPLORATORY DATA ANALYSIS

In order to understand the dataset provided, it was necessary to dig deeper into the variables and their effects on the target for this experiment. **Figures 6** and **7** below quickly visualize some of the variables that were analyzed for impact on the target attribute. Figure 6 shows two attributes that have very little impact on the target variable when looking at the distribution between levels of each. **Figure 7** portrays Course Levels, which shows a more significant change in pass/fail distributions across levels. This is because even though there are a lot more observations for courses in levels 3 and 4, the higher fail rates are in course levels 1 and 2. This agrees with research that shows that most students struggle their first two years of college. Additionally, **Figure 7** includes features such as UCC and Gateway flags that also show a significant impact on the distribution of pass/fail outcomes.