Institutional Data Analytics Platform (IDAP)

Most higher education institutions have implemented student success initiatives with the goal of helping students succeed through their college career. These initiatives are often aided by data analysis. At Purdue, we are using data analytics and machine learning algorithms to better target initiatives to students in the way and time frame that they need assistance.

Predictive Analytics – Strategic Vision

Purdue's empirical tradition in relation to student success has been informed by many studies[1] focusing on subjects such as four year completion rates, transfer students, and at-risk groups. Technological solutions such as Hotseat, Passport, Pattern, and Course Signals have assisted students and faculty with classroom success. However, as big data technologies and practices are becoming a part of the higher education landscape, Purdue has expanded its efforts by utilizing new data sources that can provide real time markers of student behavior. These include data points such as wireless network usage and Purdue ID card swipes when accessing campus resources. Merging these new data sources with existing Student Information System (SIS) and Learning Management System (LMS) has become possible with the launch of Purdue's Institutional Data Analytics Platform (IDAP). The university is committed to utilizing more data sources beyond the traditional SIS and LMS to better understand factors that may be related to student engagement with campus resources and student academic progress. Leveraging this new space of big data and machine learning Purdue is now able to provide informative messaging on behavioral factors of successful students. Purdue has also provided a platform for increased collaboration with faculty experts on the technology.

Infrastructure

The infrastructure of IDAP includes Pivotal's Greenplum Database (GPDB) to store and compute structured data, and a Hadoop[2] platform to store unstructured data such as network logs. This infrastructure enables analysis on a Massively Parallel Processing (MPP-based analytics) environment allowing Purdue to process large amounts of data in a fast and efficient manner. The Pivotal suite of software includes a robust set of machine learning algorithms which are used to generate the predictive analytic results on big data.

The data sources currently in IDAP include traditional data sources such as SIS (Banner), LMS (Blackboard Learn), Degree Works as well as non-traditional data such as ID card swipe access data (dining transactions, co-recreational center access, residence hall door access), wireless network activity, and the Student Information Form (SIF). All data sources combined amounts to around 8 billion rows in GPDB and 80.9 TB raw text in Hadoop. Table 1 indicates several of the new or expanded models that have been examined in the new environment.

Table 1: Current IDAP projects
Predicting four-year graduation of individual students
Predicting course GPA for every undergraduate student in every unique undergraduate course
Transfer student retention
Predicting yield of individual students for enrollment management practices
Expanding the model predicting at-risk students at time of entry to Purdue
Curricula structure graphing of present programs of study[3]
Predicting 2nd year and 3rd year retention
Collaboration with faculty and academic areas
English Writing Lab assessment
Student messaging of predictive factors
Social network and trajectory analysis
Validation of new decision tree techniques

Results of Initial Models

Table 2 shows the results of some of the initial models that have been built in IDAP. The models are trained to capture the largest percentage of students in a risk classification (recall) as possible. The recall for the Course GPA and Pre-Entry First-Term-GPA models indicate that the models capture 73% and 72% of the risk populations respectively. To evaluate these recall numbers, compare them to what we would achieve when we select students at random as being at risk based purely on their prevalence in the overall population. In other words, if 10% of students typically fail, simply pick a random 10% of the student body and check recall. If we were to do this, the random selection recall for a student earning below a C in a course is approximately 10% and for earning less than a 2.5 first-term GPA is approximately 19%, far lower than those achieved by the models.

Table 2: Model Results			
Model	**Precision**	**Recall**	**F-Score**
Course GPA < 2.0	0.72	0.73	0.73
Pre-Entry First-Term GPA <2.5	0.41	0.72	0.52

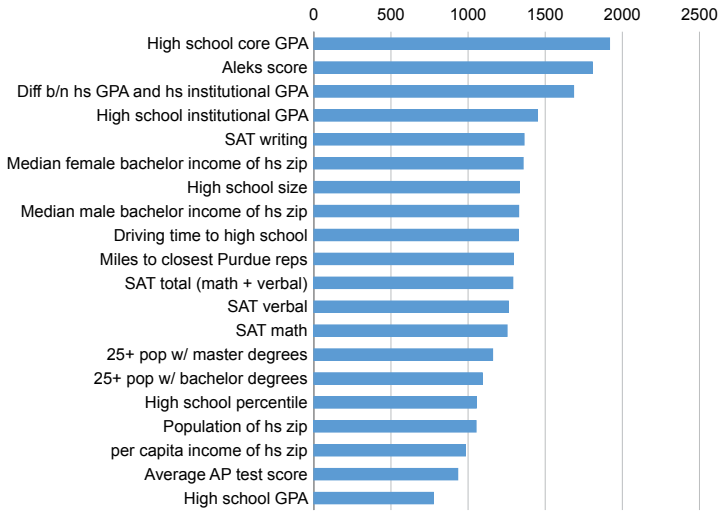
Pre-Entry First-Term GPA

Historically, Purdue has predicted first semester GPA for new admitted students prior to them arriving for their first semester at Purdue. The prediction helps identify students potentially at-risk of earning a first semester GPA of 2.5 or below. The list of at-risk students is shared with advisors and student success personnel so that appropriate outreach can be implemented. The model outputs an importance score for each predictor as well as a predicted probability of each student falling into the risk classification. A higher probability means that the student is more likely to have a first-term GPA at or below 2.5.

Institutional Data Analytics Platform (IDAP)

Figure 1 lists the top predictors of the at-risk model. The top predictor is the student's high school core GPA. The core GPA is a weighted GPA of core high school courses such as Math, Science and English courses. The feature named 'High school institutional GPA' is the historical average first year cumulative Purdue GPA of all students from a particular high school. This feature indicates the preparedness and success factor at Purdue for the set of students graduating from a particular high school. Other important predictors are related to socio-economic indicators related to the student's high school zip code.

Figure 1: Top Features of Pre-Entry First-Term GPA Model



Forecast

Purdue is utilizing the information gained from IDAP to create a student facing application named Forecast. Forecast displays to the student data about behaviors that are correlated with the academic success of past Purdue students. When the student logs in they are then able to see their individual data in relation to the overall historical trend, as well as receive information about services at Purdue to help them succeed. This will allow students to make a better informed decision about their own activities.

For instance, the application will display the negative relationship of term GPA to the number of days after the start of the semester the student waits to add a class (figure 2). If the student adds a course late then the application will provide them with advice from faculty on how to be successful (figure 3). The goal is to use this data to nudge students to be proactive in adding courses early at the beginning of the semester.

Figure 2: Public Data for Adding a Class Late Module

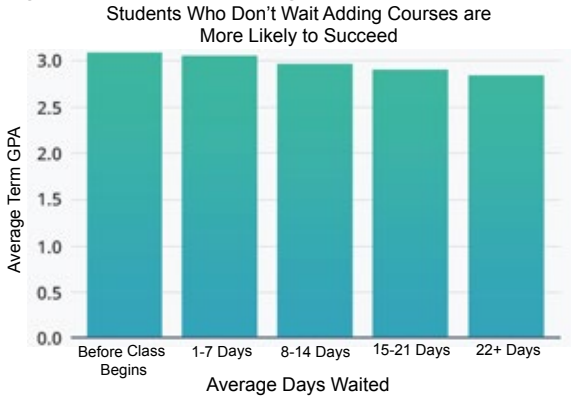
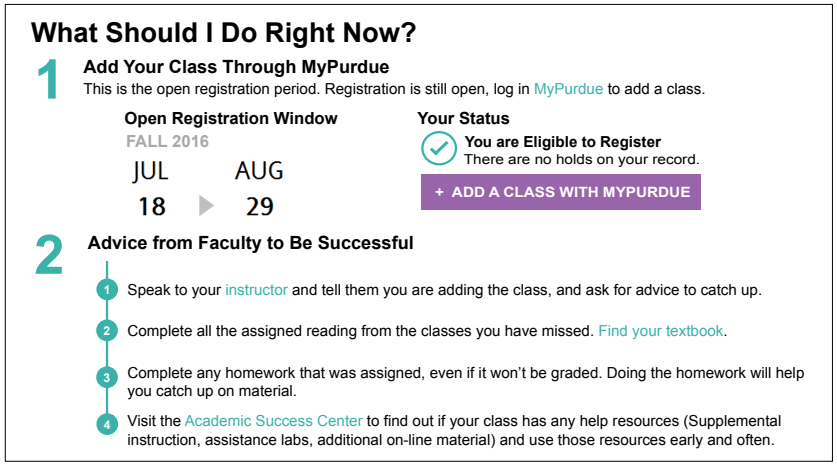
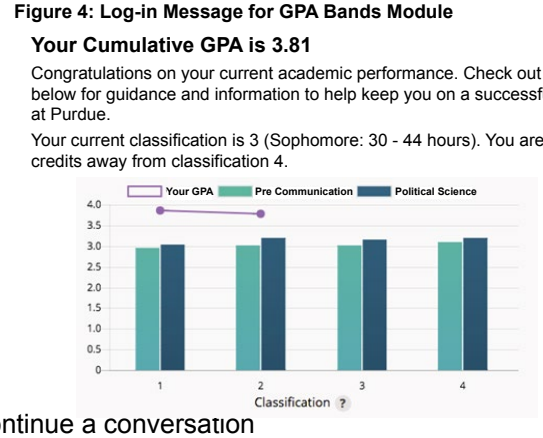


Figure 3: Log-in Message for Adding a Class Late Module



In addition, Forecast will allow students to view their performance relative to other Purdue students who were successful in their major. It displays the student's term GPA in relation to the historical term GPA by student classification (Freshman, Sophomore, etc) of students who have graduated from a particular major (figure 4). This is accompanied by links to services that can help the student during the semester such as tutoring and supplemental instructions.

Figure 4: Log-in Message for GPA Bands Module



The Future of IDAP

The infrastructure developed for IDAP will allow Purdue to continue to greatly expand its analytics efforts around student success and efficient use of university resources. As Purdue continues to increase the scope of data sources examined in IDAP it is important that we continue a conversation with our students, faculty, and administrative units about the potential use cases of the information.

[1] http://www.purdue.edu/enrollmentmanagement/aboutem/presentations.html
[2] http://pivotal.io/big-data/pivotal-hdb
[3] http://www.purdue.edu/oirae/documents/OIRAE_Briefings/Curricular_Efficiency_April_2016.pdf

Office of Institutional Research, Assessment & Effectiveness Ernest C. Young Hall 155 S. Grant Street West Lafayette, IN 47907-2114
Phone: (765) 494-7139 Email: oirae@purdue.edu www.purdue.edu/oirae