# CAN THE LEGANTO SOLUTION HELP PREDICT STUDENT SUCCESS?

## Report on Ex Libris-Curtin University Collaboration

Tomer Katz, Data Scientist
EX LIBRIS, A PROQUEST COMPANY

DECEMBER 2019

ExLibris
a ProQuest Company

# CONTENTS

# ABOUT THE PROJECT

In 2017, Ex Libris and Curtin University (curtin.edu.au/) began a collaborative effort to research the relationship between Curtin students' engagement with the Ex Libris Leganto® reading list solution and their academic success. Curtin provided encrypted, anonymized data about its students and their units of study. Ex Libris provided data regarding student engagement with Leganto and performed the machine-learning analysis described in this document. The project involved meetings between Curtin staff and the Ex Libris team and was presented at the May 2019 conference of The Higher Education Technology Agenda (Theta), the July 2019 meeting of the Ex Libris Users Group – Israel (MELI), and the August 2019 International Group of Ex Libris Users (IGeLU) meeting.

Curtin University was an excellent candidate for this project. In 2016, the university implemented the Leganto solution, and the rollout went very well. The usage of the system has been increasing steadily (Figure 1). As of September 2019, Leganto had handled more than 2,300 unique units, for which over 100,000 readings were made available to students.

## Campus Engagement Program Activities

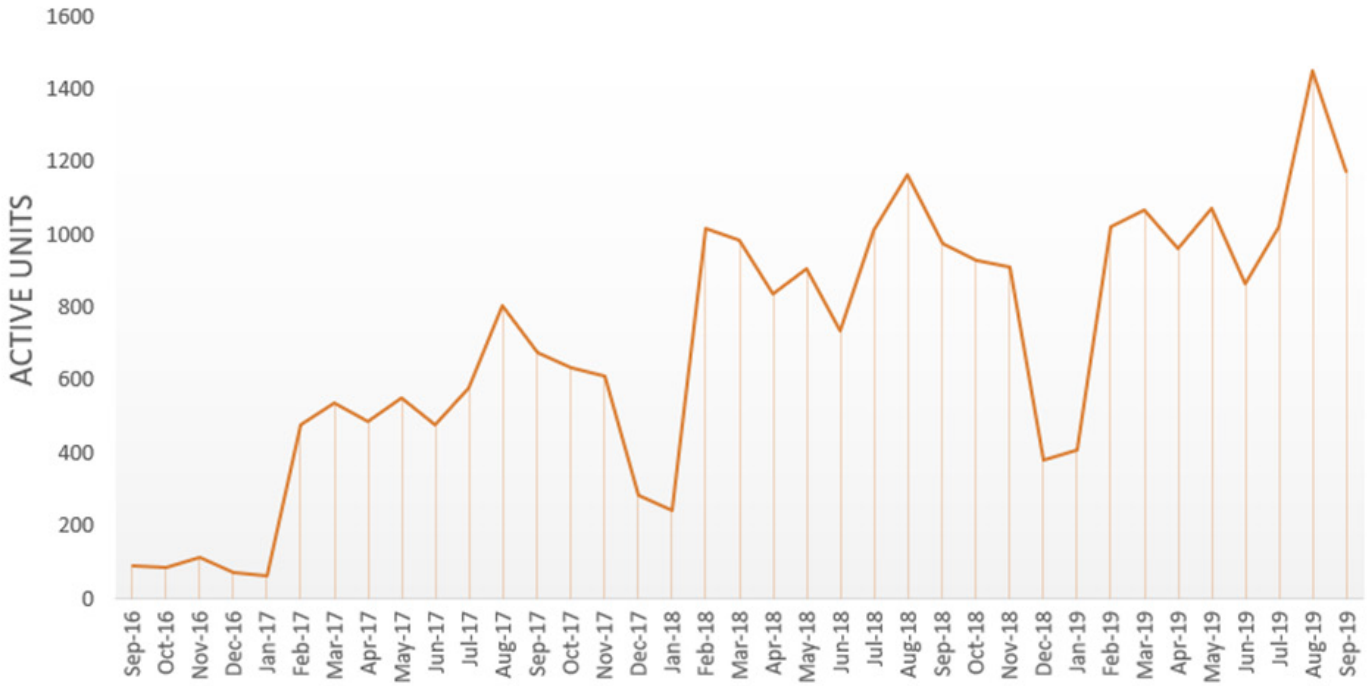


*Figure 1. Curtin's Leganto activity since the 2016 rollout*

## Terminology

This document uses the following terms:

**confusion matrix:** "A specific table layout that allows visualization of the performance of an algorithm… Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa)" (https://en.wikipedia.org/wiki/Confusion_matrix).

**feature:** "An input variable used in making predictions" (https://developers.google.com/machine-learning/glossary#f)

**feature engineering:** "The process of determining which features might be useful in training a model, and then converting raw data from log files and other sources into said features" (https://developers.google.com/machine-learning/glossary)

**ground truth:** "Information provided by direct observation (i.e. empirical evidence) as opposed to information provided by inference" (https://en.wikipedia.org/wiki/Ground_truth)

*machine learning:* "A program or system that builds (trains) a predictive model from input data (https://developers.google.com/machine-learning/glossary#m)

**precision:** A metric that "identifies the frequency with which a model was correct when predicting the positive class. That is:

$$Precision \quad \frac{True\ Positives}{True\ Positives + False\ Positives}$$

(https://developers.google.com/machine-learning/glossary).

**random forest:** "An ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees" (https://en.wikipedia.org/wiki/Random_forest)

**recall:** A metric for classification models that answers the following question: Out of all the possible positive labels, how many did the model correctly identify? That is:

$$Recall \quad \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

(https://developers.google.com/machine-learning/glossary#recall).

**success:** A student's satisfactory completion of a unit's requirements.

**unit:** "A discrete component of study within a subject area" (http://handbook.curtin.edu.au/definitions.html#U). The equivalent term in United States higher education is *course*.

**usage event:** A single student's interaction with the Leganto software (such as accessing a reading list, viewing a citation in full, or "liking" an item on a reading list)

## Data

The overall objective of our project was to identify students who are apt to struggle in their studies. The project was based on a machine-learning method called supervised learning, in which we used historical data that included end results (whether the student had passed or failed units in the past). Our aim was to train a machine-learning algorithm to correctly detect students who may have difficulty succeeding in a unit. Identifying such students as early in the semester as possible could enable the university to take action and assist the students as necessary.

The analysis focused on data from two semesters in 2018—February 26 to June 22 and July 30 to November 23—but also included data from previous semesters. Data was collected from 7287 students in all (undergraduates and graduate students), some of whom were enrolled in units in both semesters. In semester 1, the analysis incorporated data from 4188 students, and in semester 2, from 5617 students. Units in which a student had at least one Leganto usage event were included in the data set; the student's grades were pulled and included in the machine-learning project.

For the purposes of our study, we grouped all the data into three periods.

The data from the first period was historical, covering the second semester of 2016 up to (but not including) the first semester of 2018. The data consisted of two values:

- The student's average grade in units taken during that period
- The number of times that the student clicked in the Leganto software—a quantification of the student's engagement with Leganto—during the period

The data from the second period was collected at the six-week point in the semester under investigation and consisted of the student's Leganto engagement value (number of clicks).

The data from the third period was collected at the end of the semester in question and consisted of the student's final grade in the unit.

We established an "outcome indicator" reflecting student outcomes in their units of study, with three possible values: *pass*, *incomplete*, or *fail*.

Finally, we defined a variable, *target*, that represented the students whom we wanted to identify (those who may need extra help). For students whose outcome indicator was *pass*, we set the variable's value to *not at risk* (a numeric value of 0); for students with an indicator of *incomplete* or *fail*, the target value was set to ***at risk*** (a numeric value of 1). Our objective was to learn how to identify the students whose *target* value was *at risk*.

Each semester's data was collected separately according to the three periods. Then we merged the two semesters' data into one data set.

We believe that the way in which students interact with the Leganto software is related to their area of study. For example, humanities students are likely to be more engaged with Leganto than students in the sciences. We wanted our data set to consist of students who were likely to interact in similar ways with the Leganto software, so students in Curtin's Faculty of Health Sciences became the focus of our study.

Curtin University provided student-related data, unit-related data, and data about each student's success in the unit, as follows.

Data about each student:

- ID (encrypted to ensure privacy)
- Academic year
- Age
- Gender
- Degree sought
- Department

Data about each unit:

- ID
- Title
- Number of students enrolled

Data about student success in a unit:

- Each student's grade at the end of the semester
- Each student's outcome indicator (passed, incomplete, or failed) at the end of the unit

The Leganto student engagement data, provided by Ex Libris, complemented the Curtin data. The engagement data covered the following actions:

- Accessing the reading list
- Accessing the full view of a citation
- Clicking to view the full text of an item
- "Liking" an item
- Commenting
- Reading an item on the reading list
- Downloading a file on the reading list

## Feature Engineering

A variety of features were considered, where each sample (a row in the data set) represented a student enrolled in a unit. Each of the following features was evaluated for each student enrolled in a specific unit.

Features that were based on data provided by Curtin:

- RISK_AVG
  The student's risk of failure, based on the historical average of the student's grades in the first period (before the semester in question had started):

$$\frac{\textit{Student's number of units with an outcome of incomplete or failed}}{\textit{Student's number of units}}$$

  For example, a student took four units in the period beginning with the second semester of 2016 and continuing up to the first semester of 2018. The student passed one unit and failed three, yielding a feature score of 3/4 = 0.75.

- GRADE_AVG
  The average of the student's grades before the semester began: for the first semester, before February 26, 2018, and for the second semester, before July 30, 2018

- 1st_Year
  Whether the student is in his/her first year

- AGE
  The student's age

Features that were based on Leganto data:

- USAGE_BY_UNITS_PAST
  The average number of a student's Leganto usage events in the units that the student took in the first period. For example, a student who took three units before the project started and accumulated 6, 12, and 3 usage events, respectively, would have a feature score of (6+12+3)/3 = 7

- UNITS_COUNT_PAST
  The number of units in which the student used the Leganto solution during the first period

- TOTAL_USAGE_PAST
  The total number of a student's usage events in Leganto during the first period

- STU_AVERAGE_USAGE_PAST
  The average number of a student's usage events per unit compared to that of the student's classmates and based on units taken in the first period:

$$\frac{1}{M_i}\sum_{j}^{M_i}\frac{U_{ij}}{U_j}$$

$M_i$: Number of past units taken by student i
$U_{ij}$: Student i's usage in unit j
$\bar{U}_j$: Average unit usage

  The average unit usage is the number of clicks that took place in a unit divided by the number of students enrolled in the unit. This feature demonstrates whether a student is more engaged with the Leganto software than others in the class. Because the feature considers only the relationship between the student's usage events and those of the student's classmates (that is, the feature does not provide a pure usage-event count), it ignores whether engagement with the reading list differs from unit to unit.

  USAGE_FULL_TEXT
  The number of a student's usage events of the type "view full text" and "download a file" compared to the number of the student's usage events in the unit, based on the second time period (the first six weeks of the semester)

- TOTAL_FULL_TEXT
  The number of a student's usage events of the type "view full text" and "download a file," based on the second time period

- STUDENT_SEMESTER_USAGE
  The total number of a student's usage events in all the units in which the student was enrolled, in the second time period

- USAGE_STUDENT_UNIT
  The number of a student's usage events in a particular unit in comparison to the number of the student's usage events in all the units in which the student was enrolled in that semester, in the second time period. This feature indicates the units in which the student was most engaged in the Leganto software during the first six weeks of the semester

- STU_USAGE_IN_UNIT
  The number of the student's usage events in the unit during the second time period

- USAGE_RELATIVE_W1
  The number of the student's usage events in comparison to the other students' usage events in the unit, in the first week of the semester

## Data Analysis

We initially performed a data analysis to better understand the collected data and to validate its integrity. Interestingly, when we plotted the historical average risk of each student (that is, the number of times the student's outcome indicator was *incomplete* or *failed* divided by the number of units that the student took before the beginning of the semester in question) vs. the student's historical average grade (Figure 2), it became clear that most of the students earned a *passed* outcome indicator (with a value of 0 on the x axis) and their historical average grade were distributed from 60% to 90% (the distribution on the far right is close to the Gauss distribution). The graph shows the areas where most of the data points appeared, and on the lower right, it shows a small group of students who failed most of their historical units (with a value of 1 on the x axis) and whose historical average grade was low, with a value of about 60%.

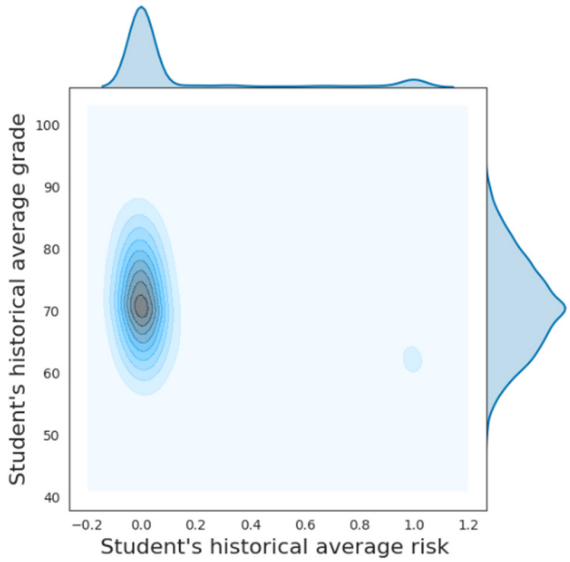


*Figure 2. Data portraying student's historical average grade vs. student's historical average risk*

By demonstrating that most students passed their units and that students who failed in the majority of their historical units have poor grades on average, this analysis suggests that the collected data makes sense.

The data analysis just described indicates a relationship between a student's historical average risk and historical average grade but does not specify whether the student passed or failed the unit. To analyze historical average risk and historical average grade in relation to the *target* variable (the student's outcome indicator for the unit), we split the data set into two groups— students whose *target* value is *at risk* and students whose *target* value is *not at risk*. This categorization was made possible because the relevant data was collected after the semester had ended (period 3) and included the information of whether the student had failed the unit. For a student who failed the unit or received a grade of incomplete, we set the *target* variable's value to 1; for a student who passed the unit, we set the *target* variable's value to 0.

We used a box plot to show both groups of students (those who passed and those who failed or received a grade of incomplete) with respect to their historical average risk (Figure 3, left). We noticed that most of the students whose *target* value was 0 had passed their units in the past (indicated by the wide stretch of the curve with the blue filling, which represents the density function). The group of students who failed or received an incomplete, with a *target* value of 1, had historical average risk values that are close to a uniform distribution. We can see that there is a major difference between the two distributions, which implies that academic success in historical units is significant in predicting future success. A similar graph, with historical average grade (Figure 3, right), shows that the difference between the groups is relatively small and, therefore, might not be significant in predicting future success. A graph with Leganto usage on the y axis (Figure 4) demonstrates a small difference between the two groups of students. We suspect that the number of Leganto usage events will have a small effect in predicting students' future success.

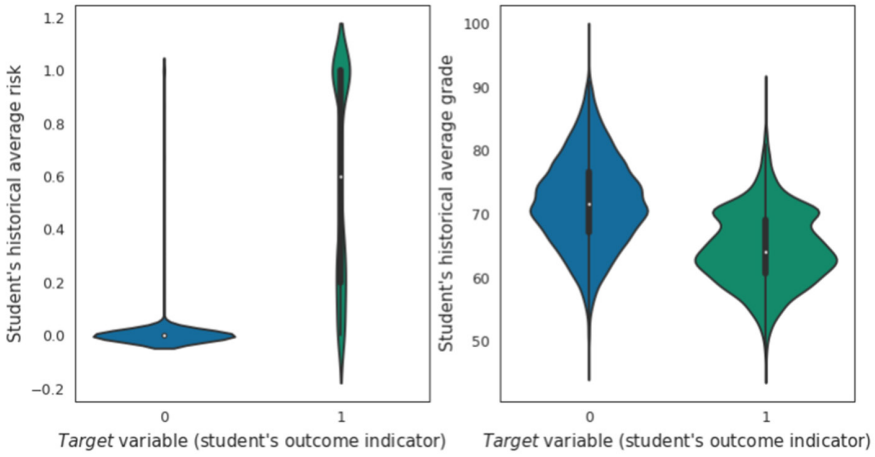


*Figure 3. Box plots of our data analysis. The left plot uses the student's historical average risk to distinguish between students who failed or received an incomplete and students who passed. The right plot uses the student's historical average grade to distinguish between the two groups of students.*

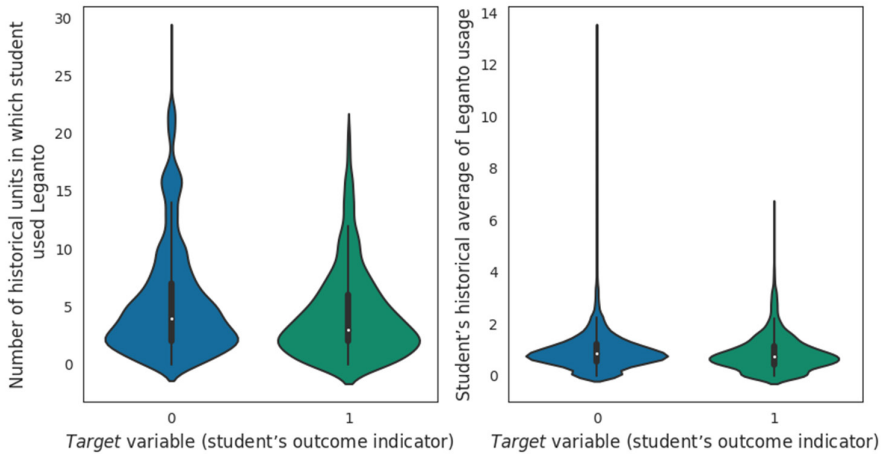


*Figure 4. Box plots of our data analysis showing the difference between the number of students who failed or received an incomplete and the number who passed relative to the number of units they took in the past (left) and the average number of times they used the Leganto software (right).*

## Model Creation and Results

The first item on our agenda was to create a simple prediction model without the involvement of machine learning, to serve as a baseline. We looked for a smart decision rule (without machine learning) to identify whether a student will be at risk of failure in the unit. During our data analysis stage, we found that the historical average risk feature would most likely be a good indicator, so we came up with the following decision rule: If a student failed at least one unit in the past, the student would be considered at risk of failure in a unit taken during the semester at hand. Next, we tested this decision rule on our data set. The test results of this rule would form the baseline of our machine-learning model, enabling us to predict the student's performance going forward.

Our data set has 16,740 rows, each representing a student who was enrolled in a unit during the semester under study. We know whether the student failed the unit or received a grade of incomplete in at least one unit in a previous semester (RISK_AVG > 0), and from the student's *target* value, we know whether the student passed the unit in the semester in question. Therefore, we can build a confusion matrix to illustrate the performance of the decision rule (Figure 5).

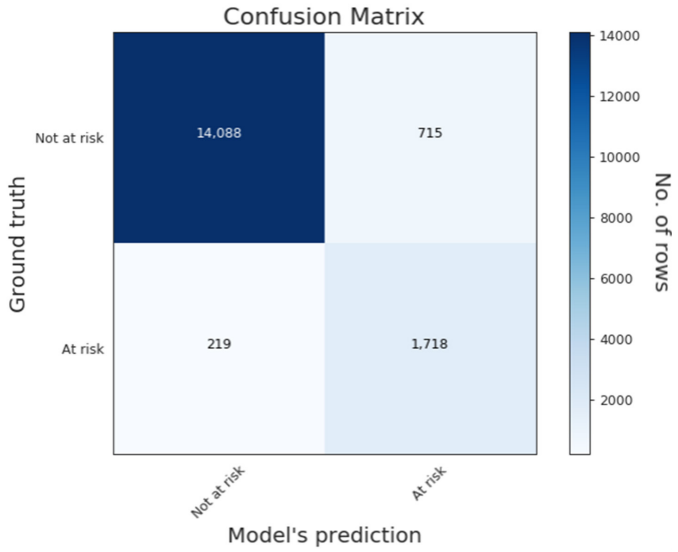


*Figure 5. Baseline confusion matrix, with a total accuracy of 94.4%. The numbers represent the rows in the data set classified by ground truth and the model's prediction*

A confusion matrix is a method used to evaluate the accuracy of a classification model. When reviewing a confusion matrix, we look at three main key performance indicators (KPIs): total accuracy (the model's accuracy in predicting any class—in our case, at risk or not at risk), precision ("the fraction of relevant instances among the retrieved instances" [https://en.wikipedia.org/wiki/Precision_and_recall]), and recall ("the fraction of the total amount of relevant instances that were actually retrieved" [https://en.wikipedia.org/wiki/Precision_and_recall]).

We can see that the performance of the baseline decision rule is good (Figure 5). Its total accuracy is 94.4% (which we calculate by adding the values of the diagonal and dividing them by the four total values). However, a more interesting KPI is precision, which expresses the accuracy of identifying a

student as at risk when this student is truly at risk. We believe that this is the most important KPI to consider. Because assistance to each at-risk student costs valuable staff time, it is important to make the staff's actions as effective as possible.

In our decision rule, the precision KPI is calculated as follows:

$$\frac{\textit{Number of students correctly identified as at risk}}{\textit{Number of students identified as at risk}} \qquad \frac{1718}{715+1718} = 0.71$$

This equation means that there is a 71% chance that a student who was identified by the decision rule as at risk will actually be at risk.

For the machine-learning algorithm, we chose the random forest classification algorithm, which was found to be most appropriate for the problem at hand. We used 80% of the data for the training set and 20% for the test set. The purpose of the training set was to enable the algorithm to learn which decision rule is best suited to identify at-risk students. We used that decision rule to examine the algorithm's performance on the test set. Then we constructed a confusion matrix and reviewed the model's KPIs (Figure 6).

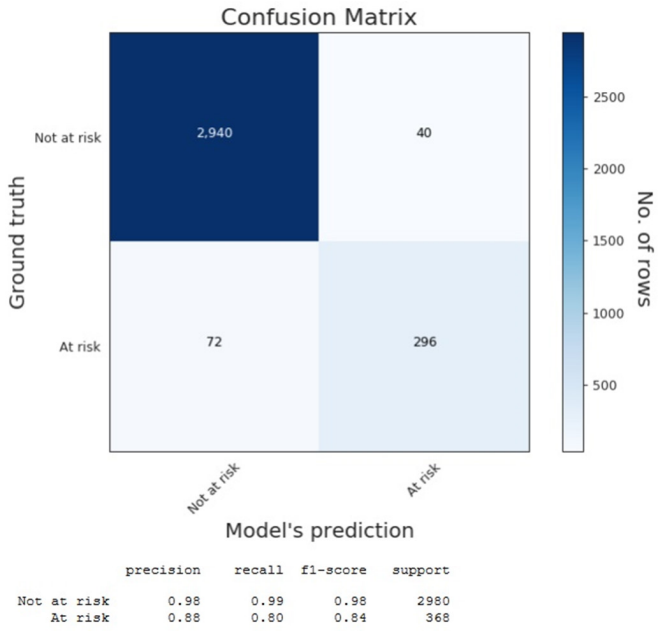


*Figure 6. Model's confusion matrix, with a total accuracy of 96.6%*

The total accuracy is 96.6%, which is slightly better than the baseline decision rule (94.4%). Furthermore, the precision KPI is 88%, which is 17% greater than the baseline. This means that using the machine-learning algorithm improves the probability of correctly identifying students who are at risk; as a result, any retention action taken by the university staff would be more effective by 17%.

To determine which features the algorithm used the most in creating the decision rule, we can look at the model's feature importance values (Table 1). The most important feature by far was RISK_AVG,

followed by GRADE_AVG (the average of the student's grades in the first period, before the semester in question had begun). In third place was the UNITS_COUNT_PAST feature, the number of units in which the student used the Leganto solution in the first period.

*Table 1. Model's feature importance values*

| FEATURE | IMPORTANCE |
|---|---|
| RISK_AVG | 0.8 |
| GRADE_AVG | 0.107 |
| UNITS_COUNT_PAST | 0.017 |
| 1st_Year | 0.017 |
| USAGE_BY_UNITS_PAST | 0.013 |
| TOTAL_USAGE_PAST | 0.012 |
| STU_AVERAGE_USAGE_PAST | 0.007 |
| AGE | 0.005 |
| USAGE_Full_Text | 0.005 |
| TOTAL_Full_Text | 0.004 |
| STUDENT_SEMESTER_USAGE | 0.004 |
| USAGE_STUDENT_UNIT | 0.003 |
| STU_USAGE_IN_UNIT | 0.003 |
| USAGE_RELATIVE_W1 | 0.003 |

Student engagement with Leganto had some effect in predicting student success, but the most predictive features were the students' past successes.

## Conclusions

This project set out to determine whether machine learning and engagement with the Leganto solution can help predict student success and indicate which students are more likely to have difficulty in their studies. On the basis of our results, we can draw two main conclusions:

- Using a machine-learning algorithm improves the probability of correctly identifying students who are at risk.
- Combining the results of a machine-learning algorithm with a metric expressing students' use of the Leganto reading list solution somewhat increases the accuracy of identifying at-risk students.

In light of the trial's results, Ex Libris plans to enable libraries that want to take part in institutional student success or retention initiatives to provide Leganto usage data to their institution's learning analytics platform.

## About Ex Libris

Ex Libris, a ProQuest company, is a leading global provider of cloud-based SaaS solutions that enable institutions and their individual users to create, manage, and share knowledge. In close collaboration with its customers and the broader community, Ex Libris develops creative solutions that increase library productivity, maximize the impact of research activities, enhance teaching and learning, and drive student mobile engagement. Ex Libris serves over 7,500 customers in 90 countries. For more information, see our website and join us on LinkedIn, YouTube, Facebook, and Twitter.