
# University of Wisconsin System Innovation Fund Final Report

**Project Title:**
**Retention and Enrollment Predictive Analytics Modeling**

**Project Leader:**
**Lynsey Schwabrow, Director, Institutional Research and Planning**

**Innovation Fund Grant Amount: $20,000**

## Executive Summary

The purpose of participation in the Innovation Fund opportunity was to acquire the means to create a data model that utilizes predictive analytical technology to forecast student population enrollment and retention trends. The project objectives included acquisition of Rapid Insight, Inc. software including QuickStart templates for enrollment and retention modeling, user licenses, and enterprise server application, as well as completion of documentation involving data definitions, model customizations, and results interpretation. Sufficient documentation from implementation will support the sustainability of the operation with existing staff moving forward. The development of a fully functional predictive model for both retention and enrollment projections required additional resources due to limited staffing, expertise, and time constraints. The Innovation Fund supported both the cost of software licenses, an enterprise server application to facilitate shared resources among project team members, and consultation services with the vendor.

## Purpose and Objective

The Office of Institutional Research and Planning at UW-Whitewater is an innovative, technology-driven unit committed to creating efficiencies to allow capacity for work that proactively exercises the potential that data analytics can have in impacting student success. Funding opportunities such as the Innovation Fund offered through UW System are valuable to such offices when initial acquisition of software can be costly and resources for training are limited. Funding from the Innovation Fund was sought to acquire predictive modeling templates, software licenses, and consultation with the vendor for the benefit of fast-tracking the implementation and use of the product on campus. Specifically, the objectives of the 120-day project were to:

- Purchase two predictive modeling QuickStart templates (Enrollment and Retention models) from Rapid Insight, Inc.
- Produce documentation to support the sustainability of the predictive modeling work, including data sources, data definitions, model customizations, and analytical interpretations
- Identify indicators of student success at UW-Whitewater
- Produce shareable visualizations for internal audience

The purpose of employing predictive analytics at UW-Whitewater is to improve enrollment and retention decision-making. Related challenges and opportunities for this analysis include goals of addressing the equity gap between minority and non-minority students, and assessing the impact of remedial coursework on retention and graduation success. These are just a couple of the challenges currently facing UW System that can be better researched through predictive modeling. In addition, a project such as this inherently improves knowledge of important data elements to the campus which can facilitate improvements to data quality.

## Organization and Approach

The project was organized into six phases to be completed within the 120-day duration. The phases and activities conducted within each are detailed below.

*Phase One: Acquire predictive modeling templates from Rapid Insight, Inc. Pre-plan for consultation sessions to conceptualize data sources and consider necessary data elements to track student success. (Duration: 1 week/7 days)*

The Innovation Fund provided the funding to acquire two QuickStart templates offered by Rapid Insight, Inc. which contained a Peoplesoft model for analyzing predictors of both enrollment and retention. Research on university predictive models and spending time watching webinars provided by the vendor assisted with an understanding of the end goal and the innovative ways in which universities are applying the modeling to operations.

The project described throughout the remainder of the report focuses on the retention model, as that is the only model researched and created during this timeframe. While this funding opportunity provided the necessary resources to purchase the enrollment model, it will be developed outside of the timeframe of this project due to significant time commitment for each of the models.

*Phase Two: Begin consultation sessions with Rapid Insight to set up templates. Identify useful campus data sources, including disparate sources which are currently not available for direct connections. Determine whether reliable data exists for required data elements. (Duration: 1 week/7 days)*

Consultation sessions with the vendor resulted in consideration of the data sources most appropriate for each variable in the model, and allowed for extended consideration of the data elements that may not be applicable or available to our campus.

*Phase Three: Plan for use of local data elements by documenting data definitions and sources. Consult with data custodians to troubleshoot and evaluate data quality. (Duration: 1 week/7 days)*

Documentation was created to confirm the data elements and definitions employed in the model. Much discussion was needed to evaluate several variables that have multiple interpretations. This phase is not complete as the campus can improve collaboration to determine the best source of particular data points. Decisions for the project were made at this point using the best information available to us at the time.

*Phase Four: Customize Rapid Insight QuickStart templates. Determine data element formats needing transformation. Discuss automation possibilities of data models and outputs. (Duration: 1 month/30 days)*

Multiple consultation sessions with the vendor were necessary to customize the retention model. Selecting the applicable data, transforming and recoding variables, applying formulas, and setting up baseline data within the tool is a significant undertaking for a small Institutional Research office. Documentation of the automated jobs that inform the model and testing comprised this phase of the project.

*Phase Five: Demonstration of models and outputs, and training of staff. Peer training and additional consultation from Rapid Insight as necessary. Gather end-user feedback. (Duration: 1 month/30 days)*

The retention model was demonstrated among internal staff, however is not yet refined enough to share as an end-product. Peer training and additional consultation will solidify the understanding of the model built and ensure confidence in the results. Gathering end user feedback will occur after that stage. Therefore, this phase of the project was not completed during the 120-day timeframe, however is currently in progress.

*Phase Six: Plan for data analysis, reports, and ongoing training allowing for sustainable use of the model following intern/staffing changes. Consider feedback obtained during Phase Five prior to closing out project. (Duration: 1 month/30 days)*

The project concluded with Phase Five in progress, therefore shareable reports have not yet been developed. Plans are in place, however, to gain an internship position within the office in which the incumbent will have the opportunity to develop the outputs necessary following the peer training in Phase Five.

# Analysis and Findings

### Student Population

The population of students targeted for these analyses included first time, first year undergraduate students enrolled at UW- Whitewater between Fall 2010 and Fall 2014. Of these, over 8,000 students were identified as having attended UW-Whitewater.

### Predicting First Year Retention

Using the QuickStart Template for retention, the project team developed a model predicting the likelihood of first time, first year students being retained to their second year. The model used admissions, financial aid, and demographic data as predictors (*see Figure 1*). As such, first year retention at UW – Whitewater was able to be predicted based on data available prior to the start of students first semester.

### Data Mining

To create the retention model the project team aggregated relevant data into a Multi-Year Model by developing four branch models: Financial Aid, Admissions, Retention, and Enrollment. The project team began by looking at which Financial Aid awards were most prevalent (*see Figure 2).*

**Figure 1. Retention Predictive Model**

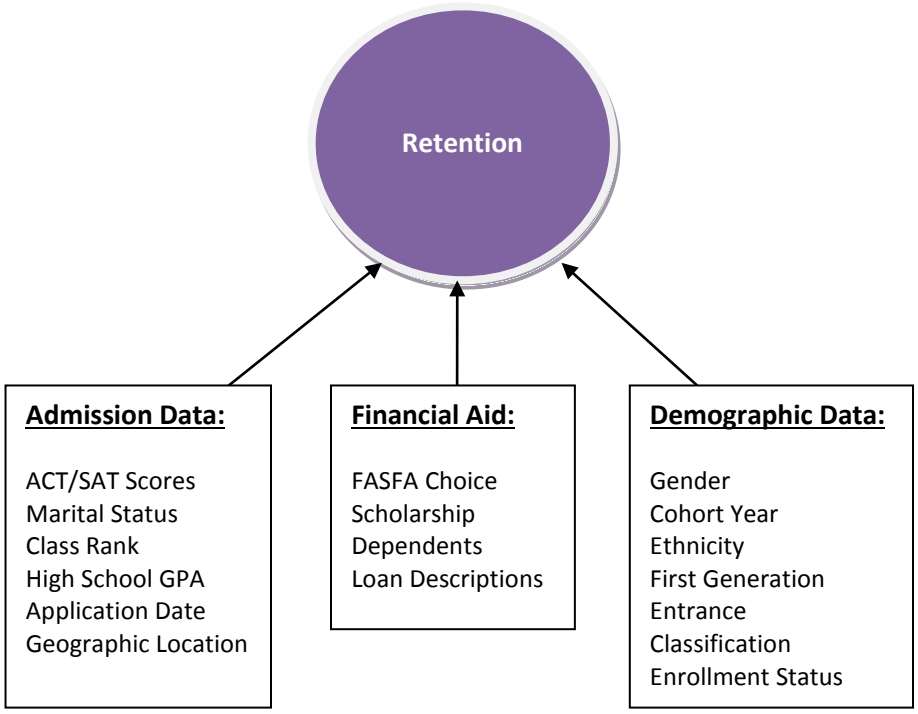


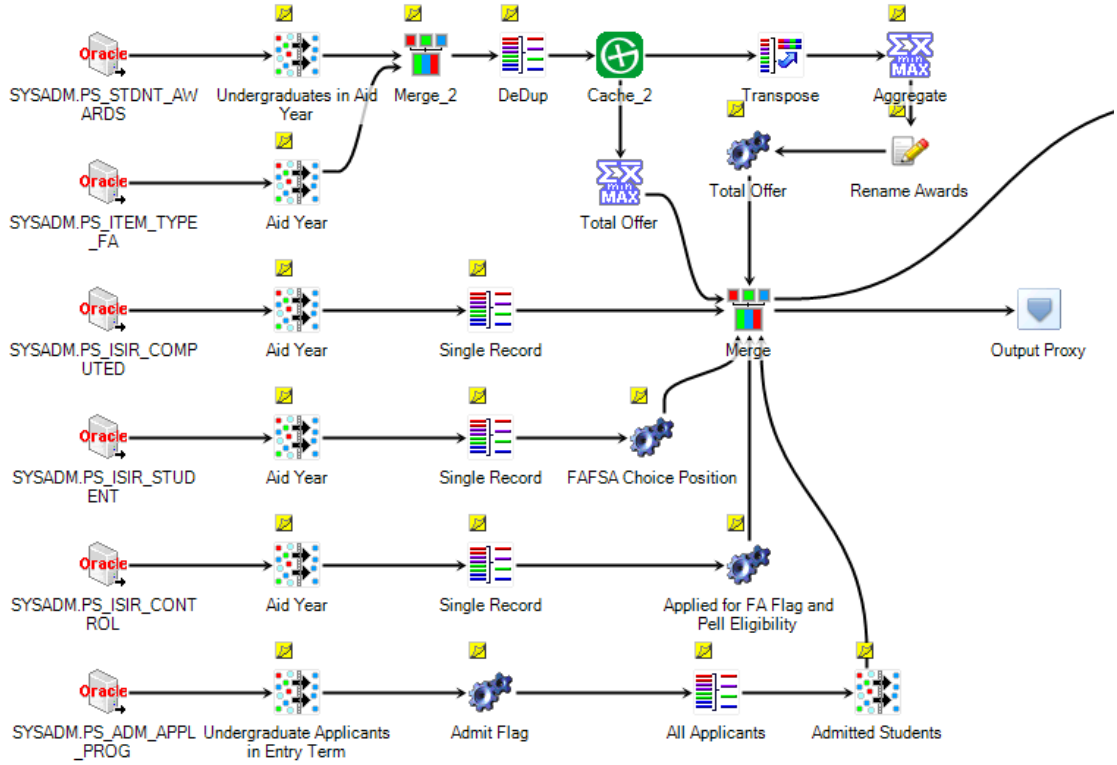

**Figure 2: Financial Aid Awards**

| Obs # | AID_YEAR | DESCR | EMPLID_Distinct | HIGHEST_OFFER_AMT_Mean | HIGHEST_OFFER_AMT_Sum |
|---|---|---|---|---|---|
| 1 | 2010 | Unsubsidized-Fed. Direct Loan | 8745 | 3066.01797489896 | 28826701 |
| 2 | 2010 | Subsidized-Federal Direct Loan | 6047 | 3351.86192337764 | 21435157 |
| 3 | 2010 | Maximum Parent Loan | 6041 | 86.6623075649727 | 523527 |
| 4 | 2010 | Pell Grant - Federal | 2662 | 3615.15965439519 | 9623555 |
| 5 | 2010 | WI Higher Ed. Grant | 1981 | 2265.86925795053 | 4488687 |
| 6 | 2010 | Work Study - Federal | 1501 | 590.572951365756 | 886450 |
| 7 | 2010 | Supplemental Grant - Federal | 1147 | 527.489102005231 | 605030 |
| 8 | 2010 | Parent Loan - Federal | 1127 | 6420.27166666667 | 7704326 |
| 9 | 2010 | Tuition Increase Grant | 1058 | 221.531190926276 | 234380 |
| 10 | 2010 | Perkins Loan - Federal | 819 | 726.433455433455 | 594949 |
| 11 | 2010 | Private Scholarship | 717 | 1691.18055788006 | 1212576.46 |
| 12 | 2010 | UWW Foundation Scholarship | 697 | 1060.76235294118 | 739351.36 |
| 13 | 2010 | ACG Grant Year 1 | 576 | 558.256944444444 | 321556 |
| 14 | 2010 | UWW Admissions Scholarship | 503 | 552.286282306163 | 277800 |
| 15 | 2010 | Unsubsidized-Fed. DL Summer | 497 | 1674.63276836158 | 889230 |
| 16 | 2010 | Pell Grant 2 - Federal Summer | 357 | 1135.52661064426 | 405383 |
| 17 | 2010 | Private Scholarship Notificatio | 353 | 511.459922512749 | 180545 |

The top ten highest value financial aid awards received by students were identified. In addition to tracking these ten variables the team recorded Pell Grant eligibility, FAFSA Choice position, and

Lawton Grant recipients. These variables were collected from local PeopleSoft data tables (*see Figure 3*) and then merged by an output proxy within the retention data model.

**Figure 3: Financial Aid Model**




In addition to financial aid data the project team utilized admissions and demographic data as well. The project team established 21 variables per student record including but not limited to; gender, birthdate, class rank, high school GPA, residency, ethnicity, and marital status. While examining certain variables, the team considered the best source for the data and in certain cases, the model was modified to accommodate data sources outside the student information system. The testing and verification of various data sources is an ongoing maintenance item for the data model. This data, similar to the financial aid data, was then merged within the retention data model through an output proxy.

Finally, the financial aid, admissions, retention, and enrollment outputs were merged to extract the data for each first time, first year freshman enrolled from Fall 2010 to Fall 2014 (8,385 student records).

## Retention Model Analysis

Results of the multi-year model were imported to Rapid Insight Analytics 3.0 for analysis. Prior to analyzing the data, the project team took time to investigate how each variable was categorized. There are five different data types; binary, categorical, continuous, text and dates (*see Figure 4*).

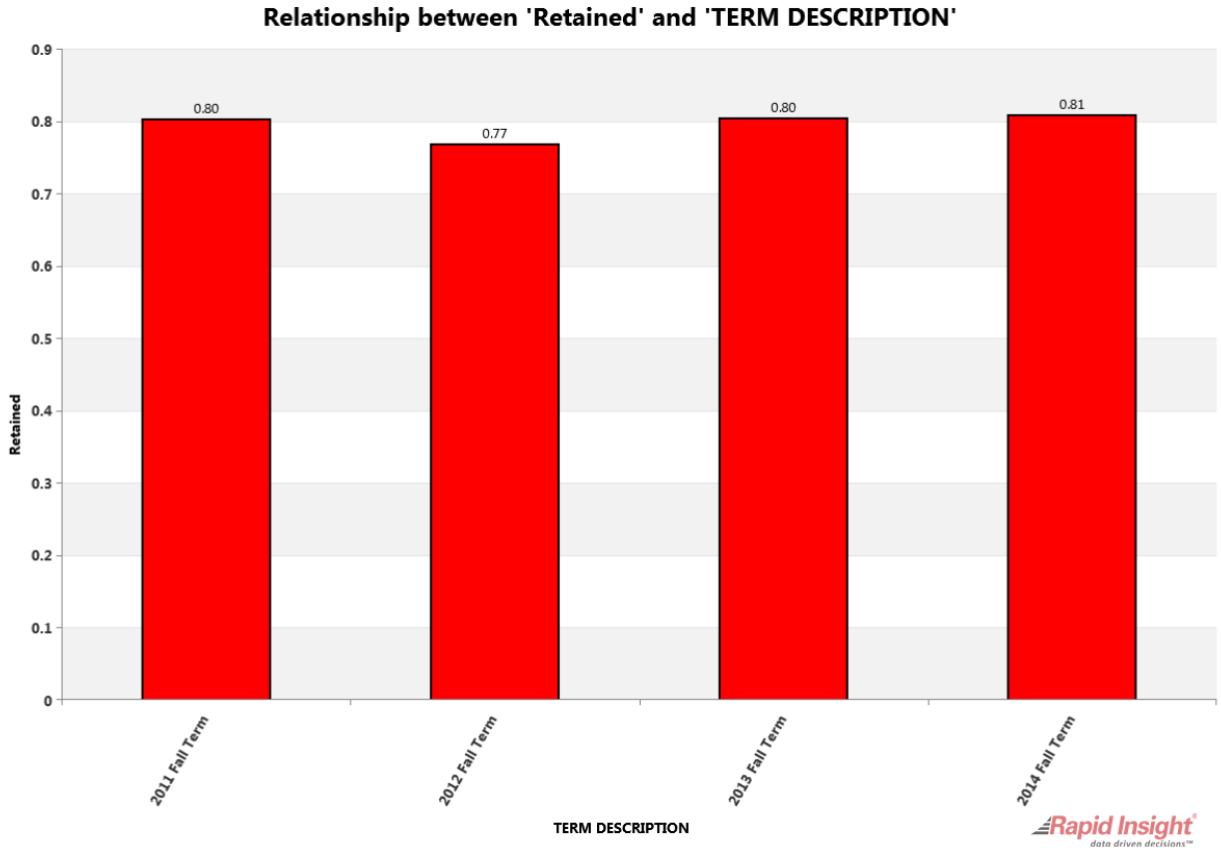
**Figure 4: Data Types**

| Variable | Type | Mean | Std Dev | Min | Max | Coeff Var | Missing | # Obs | Range | # Distinct | Notes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EMPLID | Text | n/a | n/a | n/a | n/a | n/a | 0 | 8385 | n/a | | |
| TERM DESCRIPTION | Categorical | n/a | n/a | n/a | n/a | n/a | 0 | 8385 | n/a | 4 | |
| ACAD_PROG | Categorical | n/a | n/a | n/a | n/a | n/a | 0 | 8385 | n/a | 7 | |
| ACAD_PLAN | Categorical | n/a | n/a | n/a | n/a | n/a | 120 | 8265 | n/a | 108 | |
| ACT_MATH | Continuous | 22.25 | 4.113 | 12.00 | 36.00 | 0.1848 | 120 | 8265 | 24.00 | 25 | |
| ACT_ENGL | Continuous | 21.82 | 4.335 | 7.000 | 36.00 | 0.1987 | 120 | 8265 | 29.00 | 30 | |
| ACT_COMP | Continuous | 22.30 | 3.486 | 11.00 | 34.00 | 0.1563 | 120 | 8265 | 23.00 | 24 | |
| ACT_READ | Continuous | 22.56 | 4.648 | 8.000 | 36.00 | 0.2060 | 121 | 8264 | 28.00 | 28 | |
| ACT_EW | Continuous | 19.46 | 7.249 | 0 | 34.00 | 0.3725 | 3517 | 4868 | 34.00 | 27 | |
| LAST_SCH_ATTEND | Categorical | n/a | n/a | n/a | n/a | n/a | 477 | 7908 | n/a | 816 | |
| REGION | Categorical | n/a | n/a | n/a | n/a | n/a | 8383 | 2 | n/a | 2 | |
| FIN_AID_INTEREST | Categorical | n/a | n/a | n/a | n/a | n/a | 473 | 7912 | n/a | 2 | |
| HOUSING_INTEREST | Categorical | n/a | n/a | n/a | n/a | n/a | 538 | 7847 | n/a | 2 | |
| FAFSA First Choice | Binary | 0.7179 | 0.4500 | 0 | 1.000 | 0.6268 | 0 | 8385 | 1.000 | 2 | |
| FAFSA Second Choice | Binary | 0.07168 | 0.2580 | 0 | 1.000 | 3.599 | 0 | 8385 | 1.000 | 2 | |
| FAFSA Third Choice | Binary | 0.03137 | 0.1743 | 0 | 1.000 | 5.557 | 0 | 8385 | 1.000 | 2 | |
| FAFSA First choice No second choice | Binary | 0.5348 | 0.4988 | 0 | 1.000 | 0.9327 | 0 | 8385 | 1.000 | 2 | |
| FAFSA Choice Position | Continuous | 1.309 | 0.9122 | 1.000 | 10.00 | 0.6966 | 1237 | 7148 | 9.000 | 10 | |
| DEPENDENTS | Categorical | n/a | n/a | n/a | n/a | n/a | 1237 | 7148 | n/a | 2 | |
| NUM_FAMILY_MEMBERS | Categorical | n/a | n/a | n/a | n/a | n/a | 0 | 8385 | n/a | 7 | |
| Applied for FA | Binary | 0.8525 | 0.3546 | 0 | 1.000 | 0.4160 | 0 | 8385 | 1.000 | 2 | |
| PELL Flag | Binary | 0.2951 | 0.4561 | 0 | 1.000 | 1.546 | 0 | 8385 | 1.000 | 2 | |
| TOTAL_INCOME | Continuous | 91,597.98 | 76,665.60 | -6,000.00 | 2,942,470 | 0.8370 | 1237 | 7148 | 2,948,470 | >5,000 | |
| Days Between App and Term Start | Continuous | 284.63 | 55.94 | 12.00 | 538.00 | 0.1965 | 473 | 7912 | 526.00 | 258 | |
| SAT Comp | Continuous | 1,038.02 | 163.20 | 410.00 | 1,460.00 | 0.1572 | 8289 | 96 | 1,050.00 | 49 | |
| SEX | Categorical | n/a | n/a | n/a | n/a | n/a | 156 | 8229 | n/a | 2 | |
| MAR_STATUS | Categorical | n/a | n/a | n/a | n/a | n/a | 156 | 8229 | n/a | 2 | |
| CLASS_RANK | Continuous | 81.89 | 87.93 | 0 | 812.00 | 1.074 | 479 | 7906 | 812.00 | 416 | |
| CLASS_SIZE | Continuous | 248.02 | 1,069.84 | 0 | 53,547 | 4.314 | 479 | 7906 | 53,547 | 575 | |
| EXT_GPA | Continuous | 3.563 | 5.209 | 0 | 98.09 | 1.462 | 479 | 7906 | 98.09 | 1818 | |
| CONVERT_GPA | Continuous | 3.219 | 0.5059 | 0 | 5.162 | 0.1572 | 479 | 7906 | 5.162 | 1726 | |
| PERCENTILE | Continuous | 48.93 | 31.20 | 0 | 100.00 | 0.6377 | 479 | 7906 | 100.00 | 99 | |
| Distance From Campus | Continuous | 58.78 | 101.39 | 0 | 4,191.88 | 1.725 | 290 | 8095 | 4,191.88 | 852 | |
| International Flag | Binary | 0.01133 | 0.1058 | 0 | 1.000 | 9.341 | 0 | 8385 | 1.000 | 2 | |
| In State | Binary | 0.8208 | 0.3836 | 0 | 1.000 | 0.4673 | 0 | 8385 | 1.000 | 2 | |
| Age | Continuous | 18.22 | 0.9899 | 16.00 | 53.00 | 0.05432 | 156 | 8229 | 37.00 | 20 | |
| COUNTRY | Categorical | n/a | n/a | n/a | n/a | n/a | 152 | 8233 | n/a | 44 | |

Validity of the data was tested by graphing retention rates, and the data set was found to match the actual retention rate in past years (*see Figure 5*).

The project team continues to explore results in this stage of analysis to determine whether results are as expected for each variable. The process of ensuring the most accurate data sources for the variables are being used will continue beyond the timeframe of this project.

**Figure 5. Determining Validity**




## Data Validation

In 2014 UW-Whitewater recorded an actual freshman to sophomore retention rate of 80.9% for first time, first year freshmen, while the logistic regression model predicted a retention rate of 79.6%. This example confirmed that our actual versus predicted retention rates were within 1% for the 2014 freshman cohort (*see Table 1*).

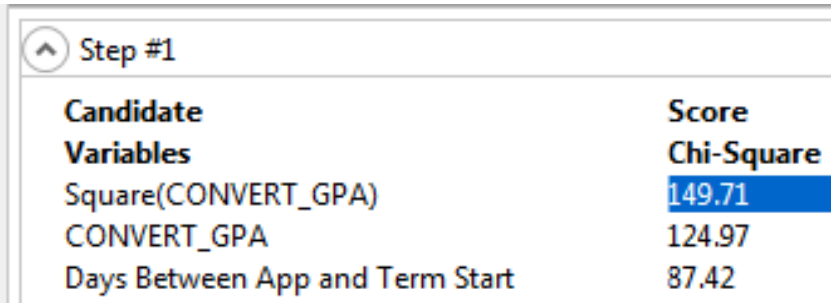
**Table 1. Actual versus Predicted Retention Rates**

| **Cohort Year** | Cohort | Actual Retention % | Predicted Retention % |
|---|---|---|---|
| **2012** | 2,155 | 76.9% | 79.7% |
| **2013** | 2,096 | 80.5% | 79.9% |
| **2014** | 2,141 | 80.9% | 79.6% |

## Results

Using Rapid Insight Analytics 3.0, 84 variables from the data model were assigned a chi-square score to determine whether a significant relationship with existed with retention. The process was repeated until the chi square score was no longer significant, or in this case received a score less than six (*see Figure 6).*

**Figure 6. Related Variables**




**Key Findings**

- *High School GPA:* Students' high school performance was found to be a significant predictor of student first year retention
- *Timing of application*: Variables associated with the application date were found to be significant predictors of student first year retention
- *UWW Foundation Scholarship*: Receipt of the UWW Foundation Scholarship was a significant predictor of retention for underrepresented minority students.

**Project Challenges**

While building the predictive retention model the project team encountered unforeseen challenges. One such challenge was the particular nature of the Quick Start templates. The templates were developed by Rapid Insight Inc. based on the PeopleSoft database. The project team chose to incorporate data elements from other sources which was outside the original scope of the project. Additional consultation with Rapid Insight was needed to customize the predictive retention model.

Another obstacle the project team encountered was a software update to Rapid Insight Analytics 3.0. The updated interface and functionality required the project team to seek additional training and change existing documentation to accommodate the updates.

The greatest challenge to this project was the 120-day timeframe which proved too short to complete the development of both the retention and enrollment predictive models. While this funding opportunity provided the necessary resources to purchase the Quick Start templates for both models, the development of the enrollment model was not completed due to the significant time commitment for each. The subsequent work to complete the enrollment model will be accomplished outside of this project timeframe.

# Conclusions and Recommendations

The Office of Institutional Research and Planning at UW-Whitewater identified a robust predictive analytics tool to examine first-year retention in first time, first year freshman. Employing predictive analytics can help mitigate challenges encountered by universities

including the achievement gap between minority and non-minority students. The Innovation Fund helped acquire the predictive modeling templates, software licenses, and consultation with the software vendor. Further collaboration with the software vendor and university data custodians is needed to refine the predictive model. Peer training and gathering end user feedback will occur after the model is further validated.

Expediency of producing predictive analytics is an issue throughout UW System due to limited staffing, expertise, and budget resources. The project team plans to facilitate discussion of the emerging practice through the CDR Liaisons across UW System. The CDR Liaisons regularly share information about tools being used to meet increasing demands for data and performance indicators.

# Appendix A: Proposed and Actual Budget

**Proposed Budget:**

| | Item Description *(person or item)* | "Hours and Rate" (if labor) or "Purchase Cost"(if non-labor) | Line Total |
|---|---|---|---|
| 1 | Enrollment Modeling and Student Retention QuickStart Templates by Rapid Insight: 2 templates x $4,000 | $8,000 | $8,000 |
| 2 | Rapid Insight Veera and Analytics licenses/Veera Enterprise Server: 1 Server cost x $5,000, 3 Single User Suite Licenses x $3,500/each = $10,500 | $15,500 | $15,500 |
| 3 | | | $ |
| 4 | | | $ |
| 5 | | | $ |
| 6 | | | $ |
| 7 | (add lines as necessary) | | $ |
| | | **Total Request:** | $20,000 |
| 1 | Matching Funds (Source: Institutional Research and Planning office will cover costs above the $20,000) | | $3,500 |
| | (add lines as necessary) | | $ |
| | | **Total Matching funds:** | $3,500 |

**Actual Budget**

| | Item Description *(person or item)* | "Hours and Rate" (if labor) or "Purchase Cost"(if non-labor) | Line Total |
|---|---|---|---|
| 1 | Rapid Insight Server | $5,000 | $5,000 |
| 2 | Rapid Insight Single User Suite Licenses: 3 Veera + Analytics = $3,000/ea | $9,000 | $9,000 |
| 3 | WI Software Contract Fee | $70.11 | $70.11 |
| 4 | Early renewal cost for Rapid Insight Server = $5,000, 3 Veera + 3 Analytics User Suite Licenses = $3,000/ea*<br>**Both a pro-rated Analytics license renewal ($675.00) and prepay renewal discount ($1,332.50) were subtracted from the cost.* | $11,992.50<br>*($5,929.89 paid for with Innovation Fund, with remaining balance of $6,062.61 paid for by Institutional Research and Planning* | $5,929.89 |
| | | **Total Innovation Fund expenditures:** | $20,000 |
| 1 | Matching Funds (Source: Institutional Research and Planning) | $6,062.61 | $6,062.61 |
| | | **Total Matching funds:** | $6,062.94 |

Notes: QuickStart templates in the Proposed Budget were unable to be funded due to the purchase date being prior to the current fiscal year. Therefore, funds were applied to licenses that will extend the use of predictive analytics through June 2017 for the Institutional Research and Planning office at UW-Whitewater.

# Appendix B: Project Team Members

**Project Leader**
Lynsey Schwabrow, Ph.D.
Director of Institutional Research and Planning
University of Wisconsin Whitewater
Office of Institutional Research and Planning

**Project Coordinator**
Benjamin Prather
Interim Policy & Planning Analyst
University of Wisconsin Whitewater
Office of Institutional Research and Planning

**Project Collaborator**
Rochelle Day
Business Intelligence Analyst
University of Wisconsin Whitewater
Office of Institutional Research and Planning