

2019

Institutional Effectiveness and Accreditation

University of Idaho

2/28/2019

Refinement of a Predictive Model for Progression Risk for Undergraduate Students

Contents

[Reexamination of Spring to Fall Progression 2](#_Toc2259601)

[Executive Summary 2](#_Toc2259602)

[Theoretical Framework 3](#_Toc2259603)

[Phase 1: Forensic Model: 4](#_Toc2259604)

[Background 4](#_Toc2259605)

[Phase 2: General Undergraduate Predictive Model: 5](#_Toc2259606)

[Background 5](#_Toc2259607)

[Treatment of High School data within General Undergraduate Model 5](#_Toc2259608)

[Treatment of Racial Self-Identification with General Undergraduate Model 6](#_Toc2259609)

[General Undergraduate Final Model: 6](#_Toc2259610)

[Consideration with Racial and gender elements 7](#_Toc2259611)

[Summary Findings (take-away): 8](#_Toc2259612)

[References 9](#_Toc2259613)

[Appendix A: Model Maximum Likelihood Analysis Summary 11](#_Toc2259614)

***Contact: Dale Pietrzak, Ed.D., Director, Institutional Effectiveness and Accreditation***

*dalepietrzak@uidaho.edu*

# Reexamination of Spring to Fall Progression

## Executive Summary

1. A final predictive model has now been revised to the point where it is within the functional range commonly seen for these models in the literature and reported at conferences. The model is consistent with data from the Cluster Analytic examination of the non-retained and initial data from the Student Experience Survey.
2. It can be used to provide a set of risk scores and prioritization indicators which can be used by advising or others in working with students. Meetings to determine how to provide the information to those desiring to use it (i.e., dashboards, lists, etc.) are being initiated. These discussions will necessarily include defining optimal intervention points regarding level of risk. These can be used to develop targeted student contact points and assist with general advising. Conversation with Advising and/or others who would have utility for it are being scheduled.
3. Using the relative importance estimates of predictors, along with the groupings provided, can provide a data-based means to focus on the common elements in many students’ situations. However, as “importance” is a personal element, the relative importance of something in the general population and that for any individual student can be quite varied. Professional judgement is always required.
4. As was determined with previous models, significant measures of student behavior and attitude, key to refining these models, are largely absent from the U of I system. Steps are underway to remedy this, but they are not operationalized yet. Some of these include planning to integrate the data that is collected by VandalStar. Conversations were had with Advising, and other potential end users and the decision was made to replace the Freshman Survey (CIRP) with the ACT Engage. This would be administered close to the student’s arrival on campus and should allow for the development of key data in this area. Lastly, a Student Experience Survey is being piloted (second semester) which may provide information to assist in the refinement, over time, of these models.
5. The findings here indicate that developing separate models for key audiences and along student characteristics may be of value. The factors of full-time/part-time, student standing (i.e. freshman, sophomore, junior and senior) and perhaps college are of possible utility. However, developing them would a) be dependent on a strategy for implementation and b) be made in context of what the end user needs. The substantial time required to develop these, if they are not going to be used, would be better targeted to other projects. It should be noted that it is also likely these models will differ in the classification accuracy and predictors. The final models may not provide additional utility over the General Undergraduate Model, or may prove to be of increased value.
6. Finally, a note of caution: this model, like similar models elsewhere, still has a higher than desired false positive rate. This is to some extent an artifact of the context of such models but given its limitations it does have sufficient utility.

## Theoretical Framework

As we have been able to increase and integrate more data within our system, and it has been some time since the last set of predictive models was developed, now seemed a good time to revisit the development of these models. In doing so we engaged in a set of training workshops to ensure that the best practices in the area were current.

At the time of the last model development the determination was made by IEA that while useful, the model was limited as there were no financial and non-academic data available. As the screen-shot of a workshop slide below indicates, it is important that the models consider various student experience elements in for the best prediction model. As U of I had a limited set of these data available, effort was previously focused on finding ways to develop this area. In working with various offices and constituent groups on campus, it was determined there was insufficient bandwidth and resources available to develop the area in most locations. Progress has been made in this area as the U of Idaho has now begun the implementation of VandalStar. Additionally, a Student Experience Survey is being piloted and developed within IEA. IEA, working with both the Student Success Office and Advising, has begun the reallocation of resources from the current Freshman Survey (CIRP) towards Engage (ACT). This will provide additional actionable information for offices of Student Support and Advising as well as aid in the development of future models.

A review of the literature is provided at the end of this paper. An understanding of these data is framed from the Bean and Eaton model of retention. The reference list at the end includes some of the material that was used in the literature review.



*From Herzog, (02-21-2019), Student Success Perditions at your fingertips: Developing online dashboards with Microsoft Power BI.*

## Phase 1: Forensic Model:

### Background

Phase one in the process of developing a predictive model is the development of a model that accounts for the historical pathways a student may have used to leave the university. This model will be called the Forensic Model. It is based on data from 2012-2017. Some of the points of exodus may be viewed as desirable (graduation) and others less positive (stop/drop out). Developing this understanding of the historical predictors of the student’s exodus helps to inform the development of the predictive model. The Forensic Model provides clues to key focus areas and locates possible missing data elements. It is also used to refine the model over time.

Based on the literature review and review of previous models an initial Forensic Model was developed. This model included all undergraduate students. Data were explored in various ways including recoding of various factors such as ethnicity, and class level as binary (0 or 1), to better understand the role of these attributes in the final model. The predicted outcome was “not registered” the next fall. The outcome variable was coded as a 1 if the student was registered at Census in the fall and 0 if the student was not registered. The initial base model saw a relative lack of fit (Pearson) and suggested there was a need to deal with outliers. The typical process was followed of developing a rough model and then outliers were removed. Further, as the high school variables were non-significant in the base model and they were high-percentage missing value items constituting a large loss of the sample, they were excluded from the trimmed model. Tuition base was also not significant and was removed from the model.

The trimmed Forensic model completed appropriately (found solution). The overall model has an area under the curve (ROC) of 0.94, and rescaled R-Square of 0.64. Pearson Goodness of fit indicated a reasonable fit of the model to the data. This is a relatively strong model that is consistent with information available from the literature and conferences.

The relative importance of the various elements in the model accounting for student exodus is shown in the chart below. We see the largest factor by far associated with leaving U of Idaho is graduation. The next group, or natural break, begins with Freshman Standing and is followed by Junior Standing. The next natural break appears to begin with CUM hours earned followed by End-of-term GPA. The next grouping begins with End-of-term Hours Earned and is followed by Academic Standing at Census and Institutional Credits. The next grouping appears to start with Sophomore Standing followed by Full-/Part-time, New Transfer, and then New Freshman from High School. The next grouping begins with Age at Start of Term followed by Gender and International Students. The last set of predictors starts with CUM GPA followed by Unknown race, Campus Housing, First Generation and then Hawaiian or Pacific Islander.

A summary list of the statistically significant predictors and their operation is provided below.

* Graduated
* Freshman
* Junior
* Cumulative Earned Hours
* End-of-Term GPA
* End-of-Term Hours Earned
* Academic Standing Census
* Institutional Credit
* Sophomore
* Full- or Part-time
* New Transfer
* New Freshman from HS
* Age at Start of term
* Gender
* International Student
* Tuition Aid Difference
* CUM GPA
* Unknow race
* Campus Housing
* First Generation
* Hawaiian or Pac. Isl.

In the trimmed Forensic model, we can see the interaction of non-retention and graduation happening with the mix of various elements. For example, seniors, who have higher cumulative credit hours earned, more institutional credit hours, and who tend to be older, exit through graduation. The end-of-term GPA tends to pick up early campus exodus as much as it does academics. It and lower cumulative GPA and new freshman from HS tend to exodus due to non-retention.

## Phase 2: General Undergraduate Predictive Model:

### Background

This model is also based on data from 2012-2017. While the Forensic model provides some understanding of where students have exited the university, it contains several elements that are not useful for predicting risk of non-retention. It also includes several elements not available prior to the start of the next term. Further, graduation is typically considered a positive outcome. As such, the predictive model needs to use data available, at least by census of the initial term, and focus on the prediction-of-risk students not returning. The formulation of these models removed those who graduated and targeted the risk of non-retention.

### Treatment of High School data within General Undergraduate Model

An initial exploratory base model included multiple elements over the years and has included High School elements and other factors. While High School data is marginally useful for a General Undergraduate Model, it is data with a high percentage of missing values. The high school elements of high school rank and class size were significant but the contribution to the model (estimates) were “0” when rounded to two decimal places and were 0.001 at three. The amount of missing data exceeds that which can be reasonably addressed with imputation within the general undergraduate student population. It will likely prove valuable for a model addressing new freshman from high school, or perhaps even freshman level more specifically, but is not viable for the general population of undergraduates. It is important to mention that the HS elements of school size (smaller schools’ greater risk), and class rank (lower-class rank slightly more at risk) were found and are consistent with the Cluster Analytic examination completed previously.

### Treatment of Racial Self-Identification within General Undergraduate Model

It is important to note is that the initial examination of racial identification used white as the reference group. This approach was not able to locate elements useful to the model (no significant elements). The approach used here racial coded identification relative to the combined general population and within the context of all variables in the model (for example, white compared to non-white or Native American compared to all non-natives, etc.).

### General Undergraduate Final Model

The final model was developed from reviews of the previous models completed previously and by refinement of a general base predictive model. All variables available that had in some way previously been linked empirically or were suggested in the literature were examined. The base model was further refined by removal of outliers. It should be noted that while resident status is not included in the model like many other variables, it was tested but failed to be included but failed to be of utility. Most of these are removed as they act detrimentally on classification rates.

The final predictive model targeted the probability of not being registered at the University of Idaho in the fall following spring semester at census date. The model found solutions and had a ROC of 0.88 with a rescaled R-Square of 0.34. Pearson Goodness of fit indicated a reasonable fit of the model to the data.

The relative importance of the significant predictors in the model are provided in the graph below. The color of the bars indicates suggested grouping of similarly important predictors. The predictors appear to intuitively fall into five groups.

The first group is composed of the cumulative GPA of the undergraduate student. This is the most pronounced element in the model. The second group begins with New Transfer Students and is followed by Junior Standing and then Full- or Part-Time Status. The third group begins with Institutional Credits and is followed by Freshman Standing, Age at Start of Term, Academic Standing and then International Students. The fourth group begins with Tuition Aid Difference followed by CUM Earned Hours, New Freshman from High School and then First Generation. The last group began with COE followed by CLASS, Black or African American and finally Gender.

### Consideration with racial and gender elements

It is important to note that gender and racial elements typically serve as proxies for social, engagement, and behavioral elements not included in a model. These typically signify areas where there is greater difficulty finding a sense of connection and adjustment to the university setting. This may be due to limited access to social support, the students being more introverted, or otherwise not locating ways to feel a sense of community.



A summary list of the statistically significant predictors and their model specific operation is provided below.

* CUM GPA (a lower GPA is at higher risk)
* New Transfer (new transfer is at an increased risk)
* Junior (junior status is lower risk)
* Full- or Part-Time (part-time is higher risk)
* Institutional Credits (fewer credits are higher risk)
* Freshman (freshman have increased risk)
* Age Start of Term (older students are at increased risk)
* Academic Standing (other than good standing at increased risk)
* International Students (decreased risk)
* Tuition Aid Difference (larger difference at greater risk)
* CUM Earned Hours (higher more risk)
* New Freshman from HS (increased risk)
* First Generation (increased risk)
* COE (decreased risk)
* CLASS (increased risk)
* Black or Af. American (increased risk)
* Gender (males increased risk)

## Summary Findings (take-away):

There are several implications of this examination of the final model.

1. The final predictive model has now been revised to the point where it is operationally within the functional range common for such models discussed in the literature and reported at conferences. The model is consistent with data from the Cluster Analytic examination of the non-retained and initial data from the Student Experience Survey.
2. It can be used to develop a set of risk scores, and prioritization, that can be used by Advising or others to the extent they desire. Meeting to determine how to provide the information to them (i.e., dashboards, lists, etc.), what levels of risk that are wised to be targeted, and what student level data is desired in that delivery needs to be determined. Conversation with Advising and/or others who would have utility for it are being scheduled.
3. The model can also serve as the basis for a decision tree-based decision-making process when conceptualizing how to think about students’ risk of exodus. The relative importance and grouping of predictors can be used to prioritize what areas of a student’s situation to consider first, and as most important. However, as “importance” is a personal element, the relative importance in the general population and that for the individual student can be quite varied.
4. As with previous models, the elements of student behavior and attitude are key missing components in the development of our model relative to those developed in other locations. Steps are underway to remedy this, but they are not yet operational. Data collected by VandalStar may enable such development here. Further, after conversations with Advising, and other potential end users, the Freshman Survey (CIRP) administered by IEA is scheduled to be replaced (pilot) with ACT Engage. This would be administered soon after a student’s arrival on campus and should allow for the development of key data in this area. Lastly, a Student Experience Survey is being piloted (second semester) and it may provide information to assist in the refinement of the model over time.
5. The role factors of full-time/part-time, Student Standing (i.e. freshman, sophomore, junior and senior) and perhaps College play in the model suggest that separate models may be of utility. However, the decision to develop them would need to be made within the context of their utility. There would be little point in developing these if they are not going to be used. If the additional models are developed it should be noted that it is likely these models would differ in the classification accuracy. If so, they may not provide additional utility over the General Undergraduate Model. It is also possible that these models could result in increased classification accuracy.
6. A note of caution. This model, like similar models elsewhere, still has a higher than desired false classification rate. This is to some extent an artifact of such models but has sufficient utility to be used to enter an advising discussion. This also means that this approach is currently not likely to lead to greater precision in predicting numbers for next-term enrollments than approaches.

## References

Allison, P. (2014). Measure of fit for logistic regression. Paper 1485-2014. Statistical Horizons and University of Pennsylvania. (SAS)

Bean, J. & Eaton, S. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention*, 3(1) 73-89.

Caison, A. L. (2006). Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education* 48(4): 435-451*.*

Collier, P. & Morgan, D. (2007). “Is that paper really due today?”: Differences in first-generation and traditional college students’ understanding of faculty expectations. *Higher Education*. DOI:10.1007/s10724-007-9065-5.

Demetriou, C. & Schmitz-Sciborski, A. (2012). Integration, motivation, strengths and optimism: retention theories past, present and future. In R. Hays (Ed.) Proceedings of the 7th National Symposium on Student Retention, 2011, Charleston (pp. 300-312). Norman, OK: The University of Oklahoma.

DesJardins, S. T. (2002). An Analytical Strategy to Assist Institutional Recruitment and Marketing Efforts. *Research in Higher Education* 43(5).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861-874.

Herzog, S. (2019), Student Success Perditions at your fingertips: Developing online dashboards with Microsoft Power BI. Percontor Workshop, 02-21-19.

Herzog, S. (2019). Exploring the power of predictive analytics: A step-by-step introduction to building a student-at-risk model. Percontor Workshop, 02-07-19.

Herzog, S. (2006). “Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression.” In J. Luan & C. Zhao (eds.), *Data Mining in Action: Case Studies of Enrollment Management. New Directions for Institutional Research,* no. 131. San Francisco, CA: Jossey-Bass*.*

Herzog, S. (2005). “Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen.” *Research in Higher Education,* 46(8): 883-928.

Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regression (Second Edition).* New York: John Wiley & Sons, Inc.

Luke, C. (2009), "An Examination of Psychological Factors That Predict College Student Success and Retention." Ph.D. dis., University of Tennessee. <https://trace.tennessee.edu/utk_graddiss/71>.

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research, Volume 2.* San Francisco, CA: Jossey-Bass

Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students’ academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138(2), 353-387. DOI: 10.1037/a0026838.

Rodgers, K. & Summers, J. (2008). African American Students at Predominantly White

Institutions: A Motivational and Self-Systems Approach to Understanding Retention. *Educ Psychol Rev*, 20, 171-190. DOI 10.1007/s10648-008-9072-9.

Thompson, D. (2009). Ranking Predictors in Logistic Regression. Paper D10‐2009. Assurant Health. (SAS)

## Appendix A: Model Maximum Likelihood Analysis Summary

**Note may variables are coded in the software by defaults as:**

|  |  |
| --- | --- |
| **0** | 1 |
| **1** | -1 |

