Student Success Regression Analysis Summary

Draft v3

2015-01-15

Craig Rudick, PhD – Associate Director for Institutional Research

Vince Kellen, PhD – Senior Vice Provost for Analytics and Technologies

Roger Sugarman, PhD – Director of Institutional Research

Adam Lindstrom – Data Scientist, UKAT

Andrew Johnson, PhD – Data Scientist, UKAT

Introduction

Over the course of the last 18 months the UKAT Institutional Research and Advanced Analytics teams have spearheaded a renewed effort to understand and quantify the factors that influence the success of students at UK. These analyses are based on student-level data from our HANA database. Our primary methodology has been to use multiple logistic regression, to model the impact of various independent variables on the success of students at UK. This document summarizes our methods and results from our research program, and highlights the variables we have identified as being particularly useful indicators of student success. However, given the scope of our work, this document is intended to highlight only the most salient features of the data and our findings, and certainly does not constitute a complete account of our student success research.

The Student Population and Student Success

The majority of students enter UK as first-time, full-time, degree-seeking undergraduates (i.e., traditional college “freshmen”). Given the size and homogeneity of this cohort, our analyses have largely focused on this important group. When running logistic regression models, our default student population includes the 13,289 GRS Cohort students who started at UK in Fall 2011, Fall 2012, or Fall 2013. All analyses in this document refer exclusively to this population unless otherwise specified.

The primary student success metric utilized in our analyses is 2nd Fall retention; i.e., whether or not a student is still enrolled at UK in their second Fall semester. Nearly half of all GRS Cohort students who fail to graduate from UK within 6-years drop out before the start of their second Fall semester, making this a particularly important point in time for measuring student success. Our research has shown that the significance and effect size of the independent variables is nearly identical when using 2nd Fall retention or other retention time points (e.g., 3rd Fall or 4th Fall retention) or even graduation (4-year, 6-year, etc.). 2nd Fall retention has the additional advantage of being measurable sooner than these other options, allowing for earlier validation on upcoming cohorts. Thus, unless otherwise indicated, our models focus on 2nd Fall retention as the primary student success metric and dependent variable.

Independent Variables

Our regression models utilize over a dozen independent variables to predict student success. These include both continuous and categorical variables. Many of our variables have some missing data. In a few cases this may be attributed to poor data quality, but more often it occurs for legitimate business reasons; for example, not all students starting at UK have taken the ACT test. In the models discussed in this document, we simply eliminate records containing missing data from the analyses. This is appropriate since our main goal here is to identify the factors which influence student success.

The most important factor which distinguishes between groups of independent variables is the point in time when we are able to gather information about a student. As a student applies, enrolls, and completes coursework, our knowledge of the student continually increases, allowing us to build increasingly robust models and better predictions about a student’s probability of success. There are three key time points in a student’s first year which provide additional data for our models: admission to the University, with data from the application; the start of the Fall semester, with data on students finances and their course load; and the end of the first semester, with a student’s course grades and other involvement data. The sections below provide detailed descriptions of the data and the pertinent results of our regression analyses.

Time Point 1: Admission

Data available at the time of admission is particularly interesting because this is the only data upon which we can base decisions on whom to admit to UK. However, this data is rather limited and does not allow us to build models with strong predictive power. There are essentially two types of data available at admissions: high school academic performance variables in the form of HS GPA and ACT/SAT scores, and categorical demographic variables.

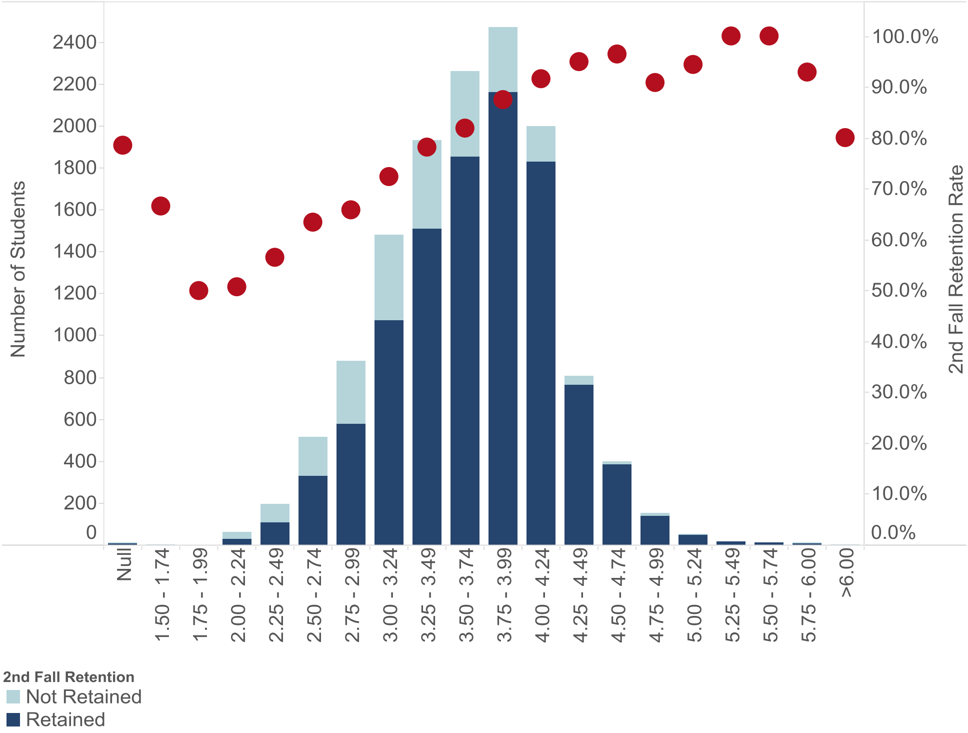
High School Academics

HS GPA

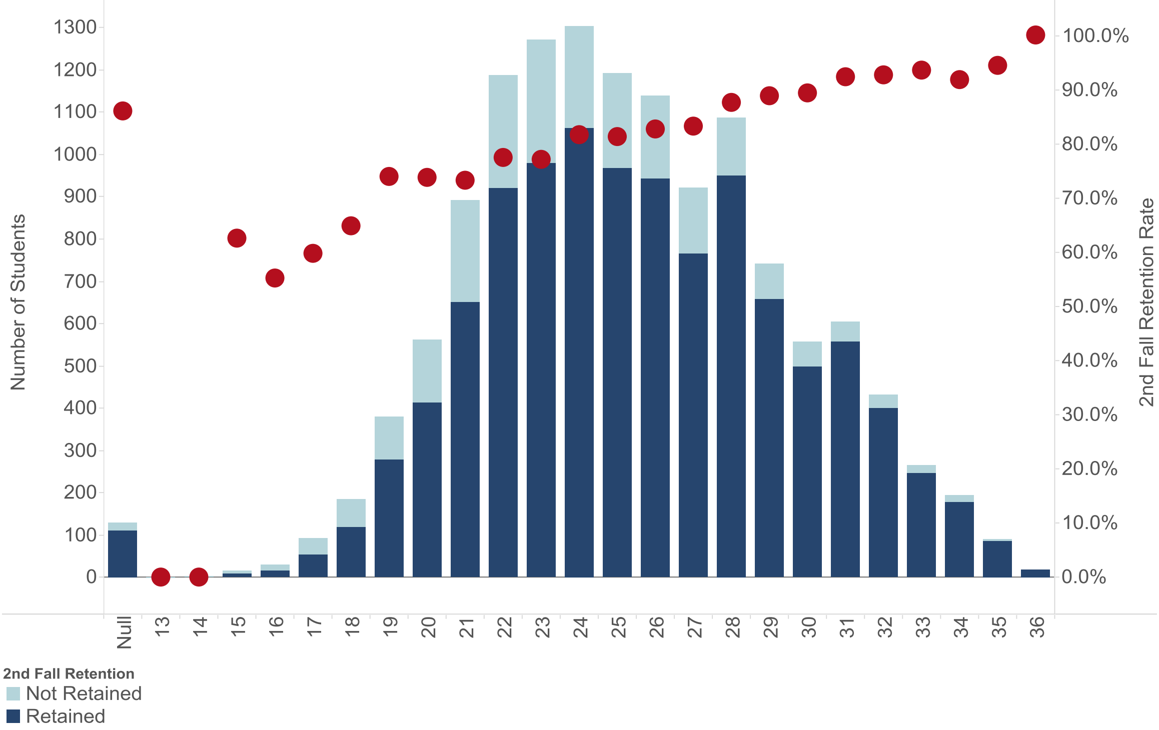
The single variable with greatest power to predict students' success before they arrive at UK is their high school grade point average, or HS GPA. A univariate logistic regression with HS GPA alone has a **p-R^2 = 0.070**. Adding a quadratic term very marginally increases the p-R^2 but we do not find this term to be significant in the full multiple model.

One potential drawback of this measure is inconsistency of the metrics value across high schools; i.e., it is not clear that achieving a certain GPA at a "good" high school is equivalent to achieving the same GPA at a "poor" one. Our preliminary analyses suggest that this is a relatively minor effect. We find a small number of high schools whose students clearly out-perform their peers with equivalent GPA's, but the vast majority of high schools do not show evidence of such metric bias.

Retention vs. HS GPA



Retention vs. ACT or Equivalent SAT



ACT/SAT Scores

The vast majority of students who attend UK take the ACT test (92.3%), although some take the SAT instead (6.8%). Our standard practice is to use the ACT score when available and to convert SAT scores into their equivalent ACT scores for students who have taken the SAT. Throughout this document, any references to ACT scores are assumed to include equivalent SAT scores, unless otherwise specified.

It is often assumed that a student's ACT score should be a more reliable metric of ability than HS GPA because it is standardized and not plagued by issues of varying quality, as are students' high schools. However, we find that a univariate logistic regression on ACT score gives a ***p*-R^2 = 0.032**, explaining less than half the variance of HS GPA. This is likely due to the fact that HS GPA is a composite score earned over four years of continuous work, while the ACT represents simply a few hours effort. High school grades also reflect the impact of motivational dispositions, such as conscientiousness, effort regulation and 'grit', that are relatively absent in standardized test scores.

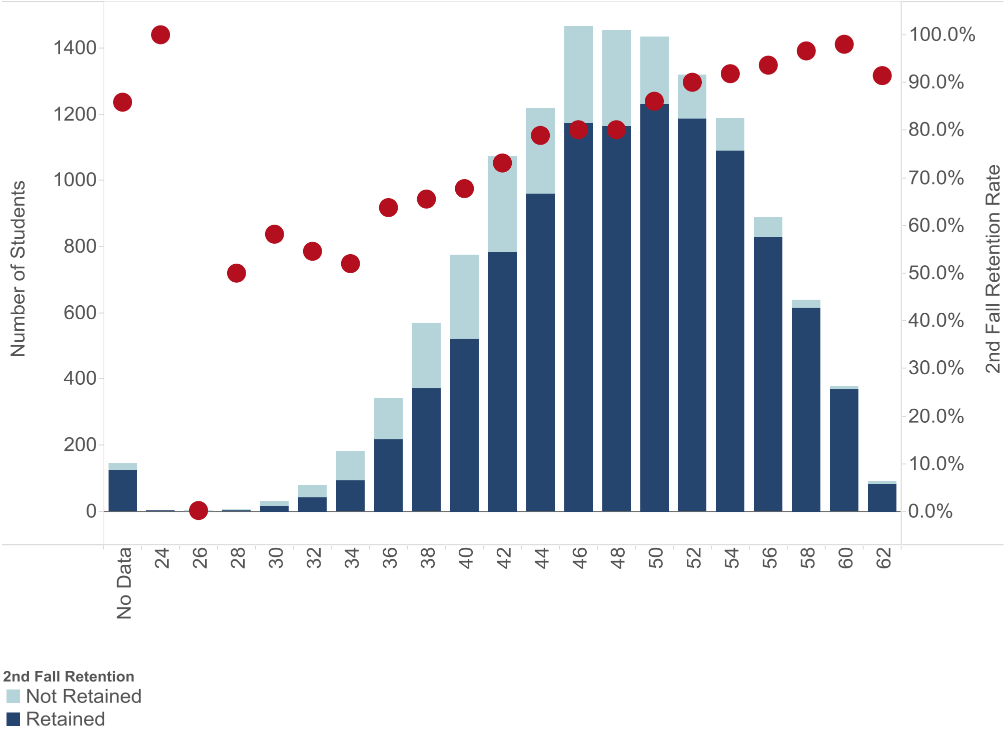
HS Readiness Index

Because HS GPA and ACT score are essentially our only pre-college indicators of academic ability, it is instructive to understand how the two interact with one another. Unsurprisingly, the two variables are significantly correlated, with a correlation coefficient of 0.53. A logistic regression model using these two variables yields a ***p*-R^2 = 0.073**, a small increase (+0.003) over using HS GPA alone.

In order to create a single variable which encapsulates student's pre-college readiness, we have used the results of the two-variable logistic regression to create the *HS Readiness Index*. This index uses an arbitrary scaling to convert the coefficients from the logistic regression fit to round, readily understandable numbers: **HS Readiness**

Index = 10 * HS GPA + ACT / 2. Effectively, this equation is primarily driven by the HS GPA, with a slight correction for ACT. Because of the linearity of the scaling, this single index provides the same level of predictive power as the two-variable model. The HS Readiness Index provides a useful ‘control’ variable, which allows us to account for students’ academic preparedness when exploring the effects of other variables, particularly categorical demographics.

Retention vs. HS Readiness Index



Demographics

The major demographic data we have on students at the time of admission are: gender, ethnicity, in-state vs. out-of-state residency, first generation status (did the student’s parents attend college), and athlete status. Including these categorical variables provides us with our full Admission time point model with a ***p-R*^{2} = 0.085** (an increase of 0.012 beyond the two-variable HS GPA and ACT model). A quantitative summary of this model appears in the table below. While there are a few formally significant and potentially interesting interaction terms which can modestly increase the *p-R*^{2} up to 0.089 (see the Appendix), we have chosen not to implement them in this model since the interpretation of the individual variables is much more straightforward when they are not included.

The most important demographic variable in this model is first generation status, with an odds ratio of 0.56; i.e. first generation students (19% of the student population) have nearly half the odds of being retained relative to their non-first generation peers. Athletes see an even stronger effect on retention, with an odds ratio of 2.2, but only 3% of students are athletes. We also find that out-of-state students are retained at lower rates than in-state students, and that Black, Hispanic, and Asian students each show statistically significant increases in retention odds over their White peers.

Variable	p-value	Odds Ratio	Notes
HS GPA	<0.001	1.84	Standardized
ACT	<0.001	1.18	Standardized
Residency – Out-of-State	0.005	0.86	Default is In-State
Gender – Male	0.738	0.98	Default is Female
First Generation – Yes	<0.001	0.56	Default is No
Athlete – Yes	<0.001	2.29	Default is No
Ethnicity - Black	0.017	1.21	Default is White
Ethnicity - Hispanic	0.039	1.30	Default is White
Ethnicity – Multi-Racial	0.364	0.90	Default is White
Ethnicity – Asian	0.011	1.67	Default is White
Ethnicity – Other	0.697	1.05	Default is White
intercept	<0.001	0.03	

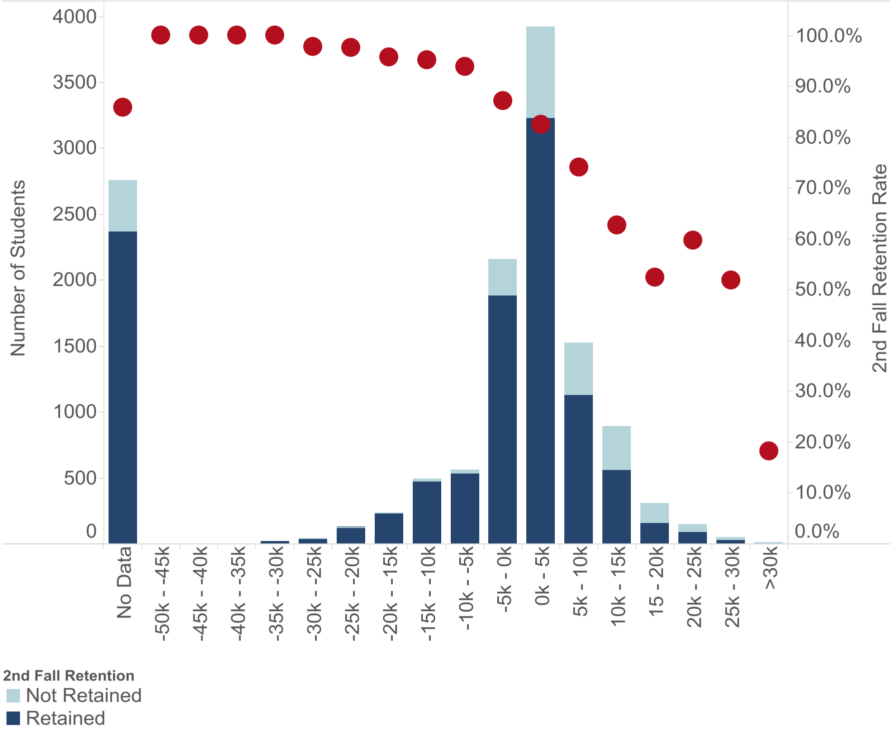
Time Point 2: Start of the Fall Semester

Between the time students apply for admission and the start of their first Fall semester, we gain several important pieces of information which can be used to predict their success at UK. These data fall into three primary categories: finances, housing, and course and program enrollments.

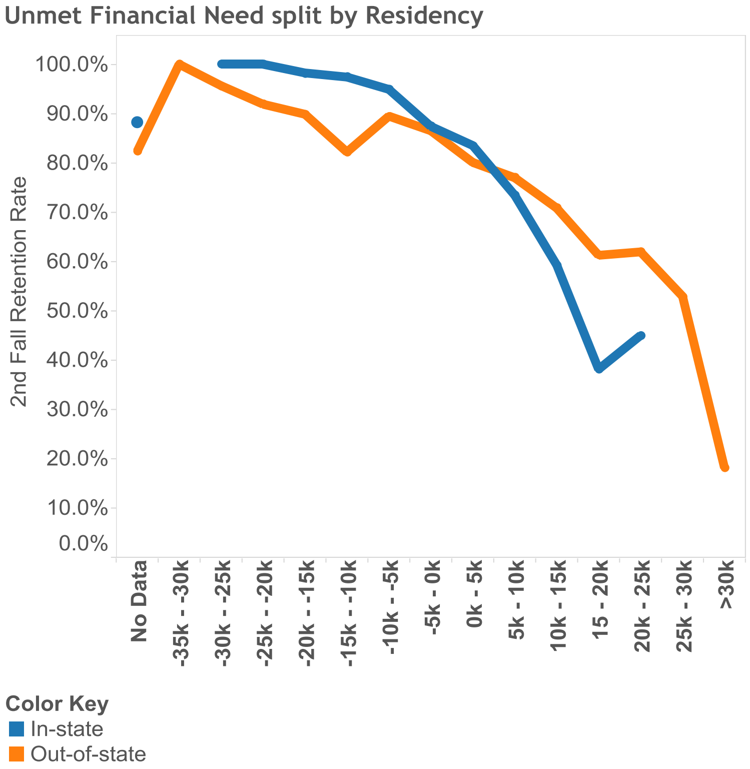
Finances – Unmet Need

Our most detailed financial data comes from the FAFSA, which some students fill out in order to receive financial aid. We have completed FAFSAs from 79% of our target population, which is a large enough proportion for this data to be potentially very useful. Additionally, it is likely that most students who do not complete the FAFSA choose not to do so because they do not need financial aid. Our analyses show that the single most predictive financial variable from the FAFSA is *unmet need*, which is a student's total financial aid package and expected family contribution subtracted from the total cost of attendance at UK.

Retention vs. Unmet Financial Need



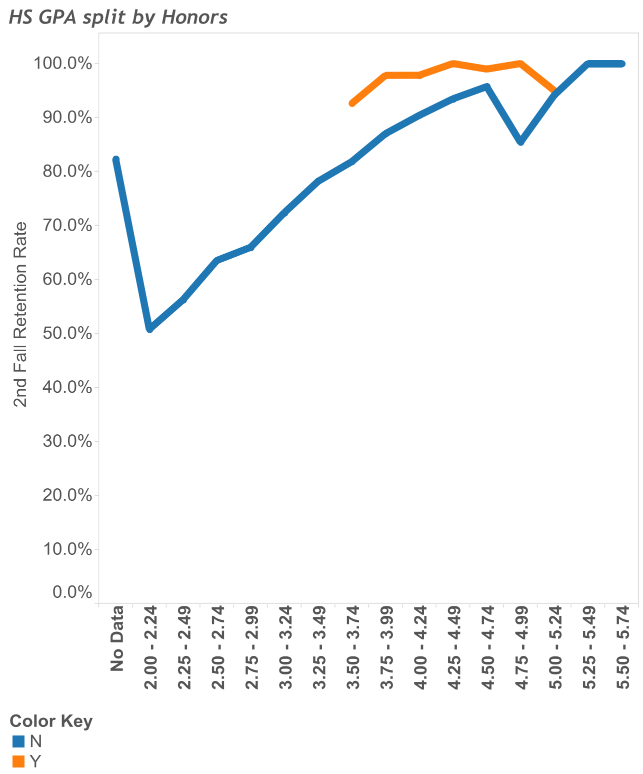
A univariate logistic model using unmet need alone has a ***p-R²*** **= 0.073**, equivalent to the result from the two-variable model using both HS GPA and ACT score. Furthermore, our analysis reveals a strong interaction term between unmet need and residency status. We find the retention rate of in-state students shows a much stronger dependency on unmet need than that of out-of-state students. Including residency and the interaction term in the regression model increase the explained variance by a full percentage point, with ***p-R²*** **= 0.083**. Adding HS GPA and ACT to this model brings the ***p-R²*** up to **0.121**, increasing the explained variance by 66% over the two-variable model with only HS GPA and ACT.

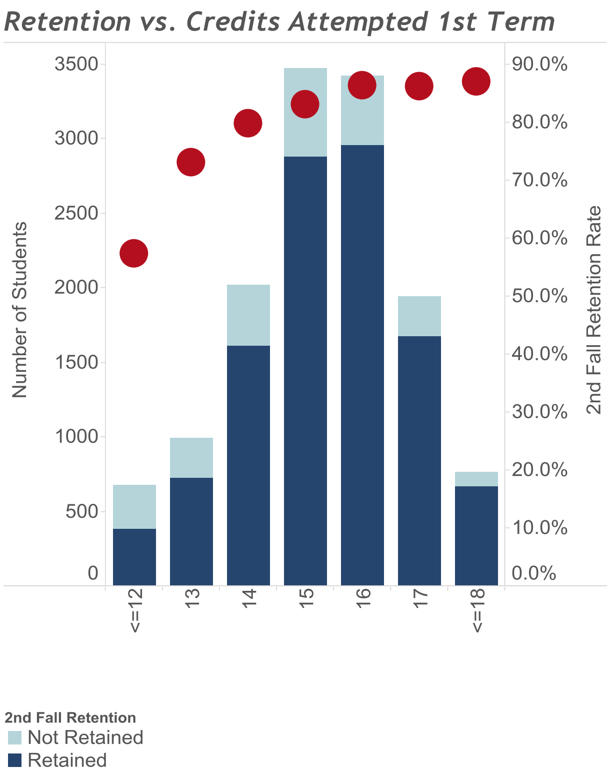


Academic Programs

As students arrive on campus, they enroll in classes, declare majors, and otherwise begin to participate in UK's academic programs. This data can be broken down in a variety of ways to potentially provide useful information on students' likelihood of success, as well as to retrospectively evaluate the effectiveness of student programs. Because of the complexity of this data, our investigations are on-going, and a complete account is far beyond the scope of this work.

Here, we identify two variables which show particularly interesting results: students' involvement in the Honors Program and the number of credit hours they attempt in their first term. The plots below clearly show that each of these variables is strongly associated with retention. However, because these variables have appreciable correlations with students' housing choices, we defer our quantitative regression model to the next section in which we discuss housing. In that multivariate model, we see that each of these variables is significantly associated with retention, with Honors program students having more than double the retention odds of their non-Honors peers, and students with a larger number of attempted credit hours having higher retention odds.

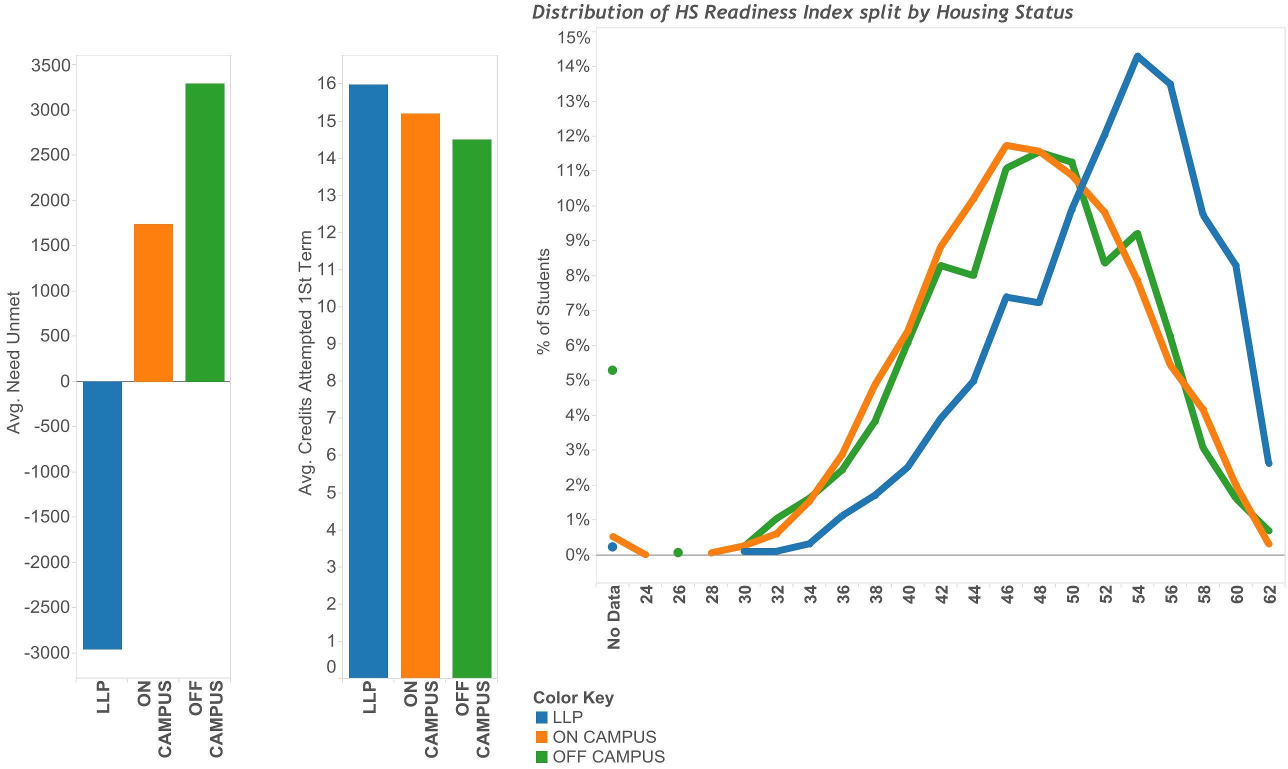




Housing

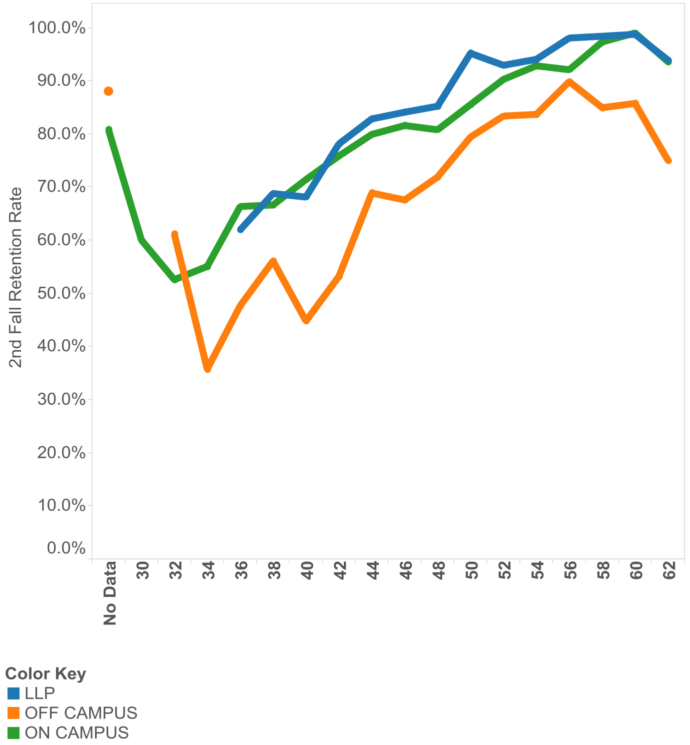
The major distinction between UK students' various housing options is whether the student lives on-campus in UK-provided housing, or off-campus. Within on-campus housing, one program of particular interest is the Living Learning Program (LLP) due to the heavy investment and significant expansion it has seen in recent years. Thus, our housing analyses have concentrated into splitting students on three housing categories: off-campus, standard on-campus, and LLP.

In order to fully appreciate the effects of housing on retention, it is particularly important to use multivariate methods which account for the inherently different student populations in the three housing categories. For instance, a naïve univariate analysis indicates that the retention odds of LLP students are more than double those of students in standard on-campus housing (retention rates of 91.1% and 82.1%, respectively). However, LLP students are significantly better academically prepared than students in standard on-campus or off-campus housing, and the level of unmet need for LLP students is less than for those in standard on-campus housing which is in turn lower than for off-campus students. Additionally, LLP students attempt more credit hours and are more likely to be involved in the Honors program (Honors has its own LLP community). Because we have previously shown that these other variables are associated with retention, it is critical that any analysis of the effects of housing on retention account for these population differences.



In the model detailed below, we have added this three-factor housing variable in our regression model, and included Honors program status and attempted credit hours as discussed in the previous section. The addition of only these variables increases the explained variance by about two percentage points, while adding in the demographics variables gives small further increase to $p\text{-}R^2$ = **0.149**. Looking at the regression coefficients and the raw data, it is clear that compared to standard on-campus housing, living off-campus has a much larger effect than residing in an LLP. Students living off-campus have nearly half the retention odds of their similarly prepared peers, while students in LLPs see a much more modest effect of increased retention odds by 26%.

Retention vs. HS Readiness Index split by Housing Status

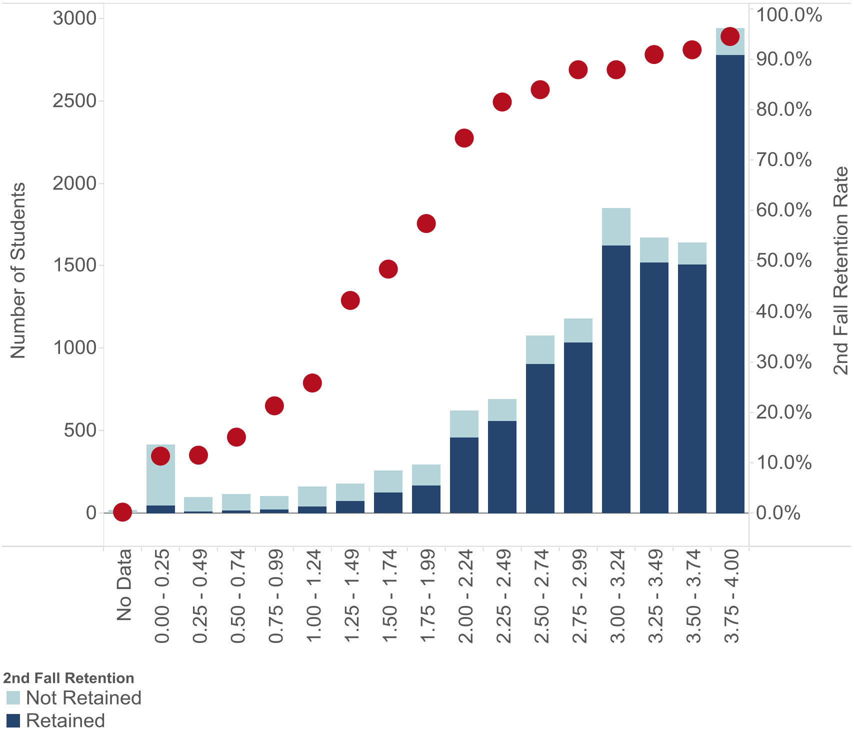


Variable	p-value	Odds Ratio	Notes
HS GPA	<0.001	1.63	Standardized
ACT	0.566	0.98	Standardized
Unmet Need	<0.001	0.92	per $1k
Residency	<0.001	0.72	Default is Out-of-State
Unmet Need, Residency interaction	<0.001	1.05	
Credits Attempted 1st Term	<0.001	1.15	
Honors Program	0.002	2.41	Default is No
Housing – LLP	0.019	1.26	Default is standard On-Campus
Housing – Off-Campus	<0.001	0.56	Default is standard On-Campus
Gender – Male	0.954	1.00	Default is Female
First Generation – Yes	<0.001	0.65	Default is No
Athlete – Yes	0.035	1.54	Default is No
Ethnicity – Black	0.556	0.95	Default is White
Ethnicity – Multi-Racial	0.184	0.84	Default is White
Ethnicity – Hispanic	0.091	1.28	Default is White
Ethnicity – Other	0.656	0.93	Default is White
Ethnicity – Asian	0.001	2.32	Default is White
intercept	<0.001	0.03	

Time Point 3: After the Fall Semester

Once students have completed a full semester at UK, we have a wealth of data on their behavior and academic performance which can be incorporated into our models. Similar to the situation discussed above for the data on students' enrollments at the start of the semester, the vastness and complexity of this data demands in depth targeted investigations that are beyond the scope of this work. However, one variable in particular stands out as providing tremendous information on students' likelihood of retention: their 1st semester UK GPA.

Retention vs. First Fall UK GPA



A univariate logistic regression using 1st Fall UK GPA yields a ***p*-*R*^{2}** = **0.234**. Adding this variable to those used at the start of the semester nearly doubles the explained variance, bringing ***p*-*R*^{2}** up to **0.277** (+0.128); while formally significant, adding a quadratic term in this variable provides only a modest increase in the explained variance (~0.3%). Interestingly, there are several variables which show substantial changes in their effect size in this model, as compared to the previous models which did not include 1st Fall UK GPA. For instance, HS GPA, which was previously among the most important variables, is no longer significant in this model. Another particularly interesting switch is gender, which was not significant in any previous model, but now shows that males have retention odds 26% higher at the same 1st Fall UK GPA than females; this suggests the possibility that males may achieve lower UK GPA's given the same academic preparedness, but also have lower standards for the own academic achievement. Many other important variables - such as unmet need, residency, first generation status, and housing - show very small changes in their effect sizes. The complex behaviors exhibited by these variables with the inclusion of 1st Fall UK GPA suggests it may be useful to compare the factors associated with high UK GPAs to those associated with persistence.

Variable	p-value	Odds Ratio	Notes
1st Fall UK GPA	<0.001	3.19	Standardized
HS GPA	0.347	1.04	Standardized
ACT	0.407	0.97	Standardized
Unmet Need (per $1k)	<0.001	0.93	per $1k
Residency (out-of-state)	<0.001	0.54	Default is Out-of-State
Unmet Need, Residency interaction	<0.001	1.03	
Credits Attempted 1st Term	0.217	1.02	
Honors Program (yes)	0.041	1.81	Default is No
Housing – LLP	0.084	1.21	Default is standard On-Campus
Housing – off-campus	<0.001	0.65	Default is standard On-Campus
Gender (male)	<0.001	1.26	Default is Female
First Generation (yes)	<0.001	0.73	Default is No
Athlete (yes)	0.412	1.19	Default is No
Ethnicity – Black	0.870	1.02	Default is White
Ethnicity – Multi-Racial	0.544	0.92	Default is White
Ethnicity – Hispanic	0.133	1.26	Default is White
Ethnicity – Other	0.236	0.82	Default is White
Ethnicity – Asian	0.001	2.56	Default is White
Intercept	<0.001	0.03	

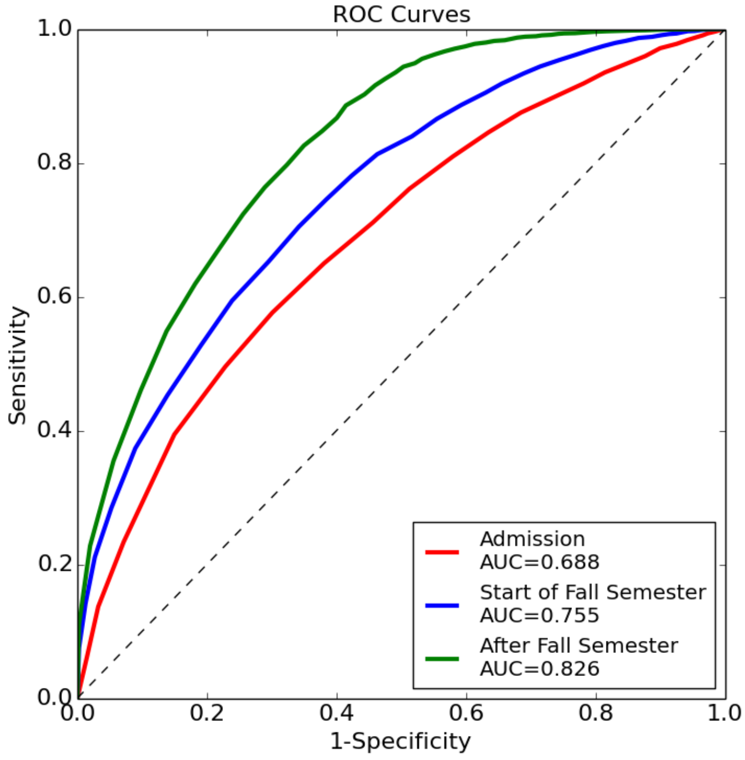
Summary

In this document we have presented the results of a set of multiple linear regression models of increasing complexity, corresponding to the increasing information we are able to gather on students as they apply, enroll and complete courses at UK. The table below summarizes the results of these regression models:

Model	$p\text{-}R^2$	AUC[*]
HS GPA (univariate)	0.070	
ACT (univariate)	0.032	
HS GPA + ACT	0.073	
Time Point 1: Admission (HS GPA + ACT + Demographics)	0.085	0.688
Unmet Need	0.073	
Unmet Need + interaction with Residency[**]	0.083	
Unmet Need + HS GPA + ACT	0.121	
Time Point 2: Start of the Fall Semester (Unmet Need + Academics + HS GPA + ACT + Demographics)	0.149	0.755
1st Fall UK GPA (univariate)	0.234	
Time Point 3: After the Fall Semester (1st Fall UK GPA + Unmet Need + Academics + HS GPA + ACT + Demographics)	0.277	0.826

[*] AUC is the integral of the ROC curves shown below, and is a direct measure of a metric's discriminatory power.

[**] For all models below this row in the table, the interaction term between Unmet Need and Residency is assumed whenever Unmet Need is used.



Future Research Goals

The logistic regression analyses outlined in this document constitute only a brief summary of our work on student success and persistence, and represent the tip-of-the-iceberg of our planned research in this area. In many ways, we feel that we are now completing the first stage of our research program by understanding the broad outlines of student success and retention at UK. In the coming months and years we intend to continue these investigations by adding new variables which explain student success, diving deeper into specific subjects, and expanding our statistical and computational methodologies. An outline of our planned future research program can be found on the Research Consortium On Student Success website: https://sites.google.com/a/g.uky.edu/research-consortium-on-student-success/home.

Appendix

Notes on Individual Variables

HS GPA

The HS GPA comes in two variants: standard (weighted) and unweighted. The primary distinction is that unweighted HS GPA is calculated on a pure 4-point scale, while standard HS GPA can exceed 4.0 for students who take Honors, AP, or other advanced classes. We use standard HS GPA as our default throughout this document, because we find that it has marginally greater predictive power than unweighed GPA, but all results presented here are extremely similar using either variant.

Additionally, there are two small data cleaning tasks performed on HS GPA before it is used in this analysis. First, we find that the predictive power of the variable ceases to increase at values above 4.5, and we thus use this as a ceiling value; i.e., all values above 4.5 are set to 4.5. This is particularly important when calculating the **HS Readiness Index**. Finally, the raw data contains some values of HS GPA of exactly 0.0; these are the result of data quality issues and we treat these as missing data.

ACT

The ACT score used throughout this document is the ACT composite score. The ACT, however, provides four subject sub-scores in Math, Science, English, and Reading. Running our regression models using these four sub-scores, we find that only the Math and English sub-scores are significant predictors, and that the predictive power of the model increases only slightly. We have chosen to use only the ACT composite score the models in this document for several reasons. Firstly, in our opinion the increase in model complexity caused increasing the number of variables is not worth the increase in predictive power, especially considering how little predictive power ACT has once other variables are considered. Secondly, the SAT only breaks down students' scores into three categories, making the conversion of SAT scores onto the ACT scale even more difficult for the sub-scores. Finally, we prefer to use the ACT composite score in HS Readiness Index, since this variable was specifically designed to capture as much predictive power as possible with a minimum of complexity; in fact, our goal for the HS Readiness Index is that students should be able to easily and quickly calculate their scores mentally.

Demographics Interactions

In the full Admissions time point model, we mention that there are a pair of formally significant interactions terms which we do not discuss in the main document. These include an interaction term between HS GPA and Residency whereby the persistence of in-state students shows a stronger dependence on HS GPA than for out-of state students and an interaction between score and Ethnicity whereby the success of African American students shows a weaker dependence on ACT score than for White students. Our search for significant interaction terms was not exhaustive, and it is possible that others exist. However, given our thorough explorations of the data, we find it unlikely that there exist any unknown interactions between our variables which have even modestly large effect sizes.

Unmet Need and other Financial Variables

In addition to Unmet Need, we have investigated the retention effects of several other financial variables, including: adjusted gross income, expected family contribution, gross financial need, and Pell Grant receipt. In this document we have focused solely on unmet need because we find that its predictive power for retention is vastly higher than for any of these other variables.