DEVELOPING A PREDICTIVE MODEL TO FORECAST AN INCOMING FIRST-YEAR CLASS

The landscape of higher education is undergoing rapid and significant changes, particularly in the area of student enrollment. Factors such as fluctuating birth rates, economic uncertainty, increasing competition between institutions, and shifts in student preferences have made it more difficult for universities to predict the size and composition of their incoming first-year class. Additionally, global events like the COVID-19 pandemic have had lasting impacts on both domestic and international student enrollment patterns. As a result, institutions are facing financial pressures, resource allocation challenges, and difficulties in long-term planning.

Accurately forecasting first-year student enrollment is critical for the operational well-being of a university. Enrollment numbers directly influence budgeting, staffing, academic offerings, housing, and student support services. Without reliable predictions, universities risk overcommitting resources or, conversely, being unprepared for an unexpectedly large or small cohort of students. These challenges make it essential for institutions to develop robust methods for predicting first-year enrollment, ensuring they can adapt to the evolving landscape of higher education.

Summit Ridge University is a large public university located in the American Heartland region. Like other institutions of its size, type and location, it has been experiencing some of the challenges mentioned above and would like to prepare better for the upcoming enrollment cycle. To address these challenges, this datathon invites participants to develop a predictive model that will forecast an incoming first-year class for Summit Ridge using a variety of factors. While traditional indicators such as past enrollment trends, academic performance, and geographic location are important, other variables—such as economic and behavioral indicators can be equally meaningful.

The aim of this datathon is to explore how demographic, socioeconomic, academic, and behavioral data can be used to develop a more precise and dynamic model of student enrollment. Which variables can be leveraged to create a model that will provide a more accurate forecast for future first-year student cohorts?

By developing this predictive model, participants will help the leadership of Summit Ridge University with the critical decision-making process, ensuring that it is better equipped to manage its resources and meet the needs of incoming students.

DATA CONSIDERATIONS

While this dataset is synthetic, it closely resembles the data used in similar projects within higher education, so, there are several important aspects to consider as you are preparing for the analysis:

You may need to transform some of the variables and create new ones to help improve results.

1. Geography: while we may not have state or city information, we have the residency status as well as miles from student's home to campus. We know that most students enroll within 100 miles from home, so, residents will have the higher rates of matriculation, following students from the bordering states. Students from non-contiguous states tend to enroll for reasons different from the other two groups: some pursue very specific programs of study, some want to get as far as possible from home, some follow their friends, etc. You may want to consider the different effect of each group on enrollment in your analysis.

2. Academics: the dataset provides you with a variety of academic variables: HS GPA, ACT and SAT scores and sub-scores, the number of high school core courses taken in math, English, foreign languages, social science and science, as well as the number of credits in specific subjects (e.g., algebra) taken at the high school or college level prior to applying and enrolling at the university. Additionally, there are variables that indicate whether a student is interested in pursuing a graduate degree, expressed the need in additional academic help or pursuing a major in STEM.

   While you can try each of these separately in the model, you may want to create additional indicators. For instance, high achieving students tend to enroll at the schools similar to Summit Ridge at a different rate from academically average students, because they have more options. Students who are planning to go to grad school will enroll at a different rate as well. Be mindful of potential behavioral differences across the academic ability groups. To understand the achievement threshold, consider the averages for academic indicators for students that apply and enroll at Summit Ridge.

3. Demographic: characteristics like sex, age or race/ethnicity may be important both on their own and in combination with geographic and/or academic characteristics. Consider the current format of these variables and feel free to experiment with alternative coding (e.g., dummy, effect, etc.)

4. Socio-Economic characteristics: while we don't have any direct indicators of student's financial origin, we have a variety of variables that can be used as proxies. For example, being a first generation student vs. being legacy, or paying your

application fee or acceptance deposit vs. having the fee waived or deferred. Additionally, academically average or below average students tend to receive less (if any) merit scholarships, which means paying full tuition. The expense is higher for non-resident students, which allows you to infer some of their families socio-economic standing and ability to pay.

5. Finally, consider time: time variables in this dataset will require additional transformation. At the moment, each time variable exists as a set of three separate columns: day, month and year. You will likely need to consolidate each set into a single date variable.

   Consider the behavioral markers these dates bring into your analysis, especially the application date, visit date and acceptance(deposit) date relative to the start of the semester (late August). By creating a tentative semester start date for each cohort year you can calculate the number of days between application and semester start and/or deposit and semester start to infer students' interest in the institution. You may also observe difference in these numbers between students of different academic standing.

NOTES ON STUDY DESIGN AND MODELING

Think about informational trade-offs that come along with the method of analysis you select and consider how you will supplement informational gaps. Certain techniques can help fine-tune the accuracy of the prediction, while others can provide insights into factors and characteristics that influence the outcomes and can be actionable (i.e., the admissions team of Summit Ridge can take steps to emphasize the positive factors and de-emphasize the negative ones).

Both, accuracy of the prediction and understanding of specific influential factors can be instrumental in the process of enrollment. If you are pursuing the insights into factors, provide visual support that makes your findings approachable to a non-technical audience.

VARIABLE DICTIONARY

student_id	
entry_year	Year when a student starts their first fall semester in college
credit_hrs	Number of credit hours student is registered for their first fall semester
app_dt	Day of application
app_mo	Month of application
app_yr	Year of application
app_fee	Application fee status: Paid, Deferred, Waived
acceptance_fee	Acceptance fee status: Paid, Deferred, Waived
hs_gpa	High School GPA student graduated with
act_cmpst	ACT Composite score
act_engl	ACT English sub-score
act_math	ACT Math sub-score
sat_comb	SAT Combined score
sat_vrbl	SAT Verbal sub-score
sat_math	SAT Math sub-score
deposit_dt	Day a student accepted their offer of admission
deposit_mo	Month a student accepted their offer of admission
deposit_yr	Year a student accepted their offer of admission
hs_math	Number of high school math credits taken in high school
hs_sci	Number of high school science credits taken in high school

hs_engl	Number of high school English credits taken in high school
hs_ss	Number of high school Social Science credits taken in high school
hs_flang	Number of high school foreign language credits taken in high school
residency	Student's residency relative to the university: Resident - lives in the state the university is located at; Contiguous - lives in the state bordering the university's state; Non-Contiguous - lived farther away, in a non-bordering state.
applied	Student applied to the university
app_completed	Student completed their application
admitted	Student was admitted to the university
deposited	Student paid their acceptance deposit
enrolled	Student enrolled at the university
undecided	Student is pursuing an Undecided/Undeclared major
hs_rank	Student's rank in high school
first_generation	Student is a first generation college student (neither parents have a Bachelor's degree)
legacy	Student has a family member previously graduated from this university
campus_visit_tot	Total number of campus visits
visit_dt	Day of the most recent visit
visit_mo	Month of the most recent visit
visit_yr	Year of the most recent visit
first_contact_dt	Day of the first contact with a student, meaning the first time a student engaged with the university in some way
first_contact_mo	Month of the first contact with a student

first_contact_yr	Year of the first contact with a student
stealth	Was a student a part of promotional communication flow or did they apply without prior contact with the university ("stealth" application=1)
grad_school_expct	A student expects to go to graduate school=1
ed_plan_asst	Student reports needing help with education planning
wrtng_asst	Student reports needing assistance with writing
read_asst	Student reports needing assistance with reading comprehension
study_skll_asst	Student reports needing assistance with study skills development
math_asst	Student reports needing assistance with math
hs_algbr_cr	Number of hs credits in the subject
hs_geom_cr	Number of hs credits in the subject
hs_trig_cr	Number of hs credits in the subject
hs_calc_cr	Number of hs credits in the subject
coll_algbr_cr	Number of college credits in the subject prior to enrolling at the university
coll_geom_cr	Number of college credits in the subject prior to enrolling at the university
coll_trig_cr	Number of college credits in the subject prior to enrolling at the university
coll_calc_cr	Number of college credits in the subject prior to enrolling at the university
coll_math	Number of college credits in the subject prior to enrolling at the university
coll_sci	Number of college credits in the subject prior to enrolling at the university
hs_only_math	Number of hs credits in the subject
hs_only_sci	Number of hs credits in the subject

rotc_interest	Interest in ROTC
miles_from_campus	Miles between student's home and campus
veteran	Is student a military veteran or a family member eligible for veteran benefits
age	Student's age
stem_major	Is a student majoring in a STEM field=1
urm	Student is a part of an underrepresented minority group
sex	Student's biological sex
college	Undergraduate college within the university