**ISA 491/591 Retention Predictive Modeling Project**

**Purpose of You Doing this Project:**

* Gain experience working with a real problem with data collected by a real system (Banner Web).
* Gain experience “data wrangling”.
* To walk through the predictive modeling process virtually unguided and present a result.

**Project Scope**:

The director of Enrollment Research and Analysis, has given us a task of **assigning a probability of leaving the university to incoming all freshman (probability of not being retained)**. The university definition of retention rate is the proportion of students who returned for a third semester. Thus the retention rate reported in US News and World Report is the percentage of students who start as typical freshman and returned for the first semester of their sophomore year. Note that “return” means that the student was enrolled in class on October 15 of that semester. If a student left Miami after that date they would be counted as being “retained”. This will give Miami information as to which students to give extra attention to upon admittance.

**ISA 491 Students:** You will have a partner for this project.

**ISA 591 Students:** This will be an individual project.

**Data**:

Anonymized application data stored on Banner Web from freshman entering in the fall semester of 2005 through freshman entering in the fall semester of 2017 has been combined “domestic\_data.csv”. Please pay careful attention to how years 2016 and 2017 retained variable is coded (its different and needs to be fixed).

The competition dataset: “domestic\_2018.csv” has no retained variable (yet). This is the data we will use to have the Data Mining World Championships on 11/30/2021. But you might want to take a look at how that data is coded (it changes by year) and what variables are included (those also change by year). Remember to use new data in a model, it must match the data the model was constructed on, exactly.

A word document titled “varsByYear.docx” contains a list of all of the variable contained in each year. You should note that there are different variables in different years. In particular note what variables are contained in 2018 data.

A somewhat useful data dictionary is contained in the file “Data Dictionary.xlxs”.

There is also a list of major codes in the file “Major codes and Descriptions.xlxs”.

**Retention Predictive Modeling Project Part 1**

In this first part you will be wrangling and cleaning the data you will use to train and validate your models. Part 2 of the project will be the modeling building part. *I do not expect you to build any models for Part 1*. At the Data Mining World Championships, I will give you a hold out sample which is the retention data taken on October 15, 2019 (which is the retention data for the freshman who started in August of 2018 that matches the domestic\_2018 data I have posted). The goal is to build the model that will predict this hold sample retention rate the best. There is extra credit involved having the best performing model at the Data Mining World Championships.

Note, whatever data preparation was done to the data to construct the model will have to be done to the hold sample data as well. Please note the variables contained in the 2018 data are listed in the “varsByYear.docx” file and the data is posted.

In this part of the project you will decide on the amount of data you will use for your modeling exercise. Will you use all of the available years or just a subset? Which subset?

You will quickly notice that even though this data set is titled domestic data there are some anomalous observations. Our goal is to construct a model for domestic students only.

**End Goal of Project Part 1:**

1. A properly formatted domestic student dataset as an .RDS file that is completely ready for modeling which includes:
   1. imputed variables as well as indicators when appropriate.
   2. properly coded response variable (1=left Miami, 0=still here).
   3. properly collapsed categorical variables.
   4. properly created dummy variables.
   5. only the variables that will be used in the modeling process (i.e. dates are not acceptable in a model, you will have to transform them).
   6. no redundant variables.
   7. proper dimension reduction techniques applied.
   8. any outside data that you choose to merge in. Please take into consideration the deployment of the model.

**Due: 11/4/2021 or any time before that**.

To turn in this part of your project you or your team will upload your data as an .RDS file to canvas as well as your code. If I give it back to you to fix, you lose 10 points.  If I give it back to you again, you lose 10 more point, etc. You must first complete this part of the project successfully before you can move to part 2.

**Important information from the director:**

* There were two “bad” years in terms of enrollment 2005 and 2009. This means that for whatever reason less students decided to come to Miami than is typical, making those classes small. This has huge budget implications and so the subsequent years 2006 and 2010 the class profile is lower (it was a little easier to get in) so that Miami was sure to fill the empty seats from the short classes. The goal is to admit 3600-3700 students per year.
* There is a considerable increase in average ACT between 2012 and 2013.
* Some teams last year chose to drop some years from the data, other choose to keep the entire set of years. The caveat being that if you look through the data files you will notice that there are a lot more inputs available in the later years than in the earlier years. You will have to decide which is better more variables or more time.
* The Academic Read Score (ACAD) is a measure of how challenging a student’s high school curriculum is. You will see some 0 values, meaning there is no challenge. However in recent years you will also see some negative values (I guess that means you just show up and get A’s). In other words, 0 and negative values are legitimate and should not be replaced.
* Scott considers international students to be a slightly different population than US residents. He typically models them separate.
* The Division has changed 00 used to be undecided now its lumped with Arts and Sciences. Also, recall, FSB has direct admits starting the in the Fall of 2012.
* Acceptance date is when the data was entered into the computer, not necessarily the date the acceptance was received. More recently the lag is just a day, however, in the past that could have been as much as a week.
* Conformation date and acceptance date are two identical fields, just coded differently.
* Scott is not interested in which Division a student is retained to. For example if you start in FSB and move to Arts and Sciences, he is happy. All the university wants is for students to stay at Miami.
* Decision date=OI, OE, OF, OM. OI=International, OE= early (these are the November students and they receive the highest aid as in scholarships). OF=regular decision and receive some financial aid in scholarships and OM=the march “the fillers” and receive the least amount of scholarships.
* ASU2 are students who did not get directly admitted into FSB but get into Arts and Sciences and get accepted later, might consider lumping these guys with the FSB students.

**Important things to consider:**

* Whatever transformations, replacement, imputation you perform on your original data file you will also have to do on the 2018 data (in class on 11/30/2021 at the “World Data Mining Championships”). Just be able to do the same manipulation fairly quickly, within one class, on the “holdout” data.
* The race variable is a total mess. You are going to need to reconcile the different coding. Use the definitions below.

1 American Indian or Alaska Native

2 Asian

3 Black or African American

4 Native Hawaiian or Other Pacific Islander

5 White

6 Hispanic/Latino

7 Multi Racial

8 Unknown

9 Non-Resident Alien

* The way that Banner codes data it uses missing values as place holders. For example, the variable called FirstGen (if the person was a first generation college student or not) is Y or missing. The missing simply means N. You will have to recode all of the variables that you wish to include in your model by filling the missing values. There are many variables like this in the data. You can figure it out via summary.
* **There are A LOT of redundant variables**. Some of them were created by previous modelers. For example there are columns that show a student’s “School Choice” but Scott as already created a column called “Miami Ranks” so you can disregard the school choice column.
  + On a side note, see what the modeler did in that above bullet? He took information and formatted it in a way that makes most sense for the problem at hand. Feel free to “create” variables that will be relevant. ☺
* Redundancy can come from several things (1) variables calculated from another variable and (2) different variables that capture the same information. For example, recently I was attempting to build a model with *Visa Status* and *Race*. What I found was the *Visa Status* kept coming up as important. I found that odd. When I actually *thought* about it, I realized that *Race* and *Visa Status* were really capturing the same information. So I just included Race in my model. Please do not take this bullet as specific instructions, it is an example. I was not modeling the data I gave you. In your retention data, the relationship might be different. You will have to put careful thought into what variables you want to include and what you choose to get rid of. I discovered/tested this using decision trees. But graphical methods and missing value patterns will be a great way to find this out.
* Hint: A variable with the same pattern of missing data as another variable is most likely redundant.
* The variable HSClust is the result of a cluster analysis to help reduce the dimension. The information that went into this analysis is the demographic information about the affluence of the high school or hometown area. It does not include ACAD.
* Consider outside sources of data, like free government data such as Census data.
* Remember, your brain cannot do model selection so cast a wide net.