**Comprehensive Student Risk Model Overview**

**Summary**

This document provides information about the predictive model for students’ risk of withdrawal in the next year. It defines the population of interest, details the variables in the model, describes the data sources and methodologies used, provides intuitions for how the machine learning algorithms work, and outlines the process by which the model predictions are classified.

The student risk model uses student data and supplemental socioeconomic data to predict the probability that a student will not persist into the next academic year. This is accomplished by first “training” the statistical model using historical data from prior year cohorts and then using that trained model to predict the outcomes for the current year cohort.

These predictions are represented as values between zero and one, which are the probability estimates that students will not be retained in the next year. It is important to note that probability values nearing one do not mean each of those students will not be retained. Likewise, probability values nearing zero do not mean each of those students will be retained. Instead, the probabilities reflect average expectations. For example, a probability value of .800 (or 80.0%) reflects the average expectation that eight out of ten students with factors leading to the same estimated risk of withdrawal will not be retained in the next year.

**Population**

Domestic full-time first-year, second-year, and transfer undergraduates.

**Outcome (w/ source) \***

*enrl\_ind (internal)*

An indicator variable for enrollment in the next year. For the predictive part of the model, the outcome calculated is the predicted probability of non-enrollment in the next year for current year students.

**Features (w/ source)**

*male (internal)*

An indicator variable for gender with a reference category of female.

*pop\_dens (supplemental)*

A continuous variable for the population density in the postal code where the student last attended school.

*educ\_rate (supplemental)*

A continuous variable for the education rate in the postal code where the student last attended school.

*underrep\_minority (internal)*

An indicator variable for underrepresented minority status with a reference category of non-underrepresented minority status. Underrepresented minorities include students identifying as Black, Hispanic, American Indian, Native Hawaiian and Pacific Islander, or two or more races.

*pct\_blk (supplemental)*

A continuous variable for the percentage of Black residents in the postal code where the student last attended school.

*pct\_hisp (supplemental)*

A continuous variable for the percentage of Hispanic residents in the postal code where the student last attended school.

*pct\_ai (supplemental)*

A continuous variable for the percentage of American Indian residents in the postal code where the student last attended school.

*pct\_hawi (supplemental)*

A continuous variable for the percentage of Native Hawaiian and Pacific Islander residents in the postal code where the student last attended school.

*pct\_two (supplemental)*

A continuous variable for the percentage of residents of two or more races in the postal code where the student last attended school.

*pell\_eligibility\_ind (internal)*

An indicator variable for Pell grant eligibility with a reference category of non-Pell grant eligibility.

*honors\_program\_ind (internal)*

*AD\_DTA (internal)*

An indicator variable for an associate of arts degree (DTA) with a reference category of non-degree holder.

*AD\_AST (internal)*

An indicator variable for an associate of science degree (AST) with a reference category of non-degree holder.

*AP (internal)*

An indicator variable for Advanced Placement (AP) exam credit holder with a reference category of non-credit holder.

*RS (internal)*

An indicator variable for Running Start (RS) credit holder with a reference category of non-credit holder.

*CHS (internal)*

An indicator variable for College in High School (CHS) credit holder with a reference category of non-credit holder.

*IB\_AICE (internal)*

An indicator variable for International Baccalaureate (IB) or Advanced International Certificate of Education (AICE) credit holder with a reference category of non-credit holder.

*first\_gen\_flag (internal)*

An indicator variable for first generation status with a reference category of non-first generation status.

*fall\_avg\_difficulty (internal)*

A continuous variable for the average “difficulty” of the courses in which the student is registered in the fall.

*fall\_avg\_pct\_CDF (internal)*

A continuous variable for the average percentage of C, D, and F letter grade to enrollments for the courses in which the student is registered in the fall.

*fall\_avg\_pct\_withdrawn (internal)*

A continuous variable for the average percentage of withdrawals to enrollments for the courses in which the student is registered in the fall.

*spring\_avg\_difficulty (internal)*

A continuous variable for the average “difficulty” of the courses in which the student is registered in the spring.

*spring\_avg\_pct\_CDF (internal)*

A continuous variable for the average percentage of C, D, and F letter grade to enrollments for the courses in which the student is registered in the spring.

*spring\_avg\_pct\_withdrawn (internal)*

A continuous variable for the average percentage of withdrawals to enrollments for the courses in which the student is registered in the spring.

*fall\_lec\_count (internal)*

A discrete variable for the number of lecture courses in which the student is registered in the fall.

*fall\_lab\_count (internal)*

A discrete variable for the number of lab courses in which the student is registered in the fall.

*fall\_stu\_count (internal)*

A discrete variable for the number of studio courses in which the student is registered in the fall.

*fall\_oth\_count (internal)*

A discrete variable for the number of courses other than lectures, labs, or studios in which the student is registered in the fall.

*spring\_lec\_count (internal)*

A discrete variable for the number of lecture courses in which the student is registered in the spring.

*spring\_lab\_count (internal)*

A discrete variable for the number of lab courses in which the student is registered in the spring.

*spring\_stu\_count (internal)*

A discrete variable for the number of studio courses in which the student is registered in the spring.

*spring\_oth\_count (internal)*

A discrete variable for the number of courses other than lectures, labs, or studios in which the student is registered in the spring.

*total\_fall\_contact\_hrs (internal)*

A discrete variable for the total number of contact hours for the courses in which the student is registered in the fall.

*total\_spring\_contact\_hrs (internal)*

A discrete variable for the total number of contact hours for the courses in which the student is registered in the spring.

*fall\_withdrawn\_hours (internal)*

A discrete variable for the number of credit hours a student has withdrawn from since census day of the fall term.

*spring\_withdrawn\_hours (internal)*

A discrete variable for the number of credit hours a student has withdrawn from since census day of the spring term.

*resident (internal)*

An indicator variable for residency with a reference category of non-residency.

*gini\_indx (supplemental)*

A continuous variable for the Gini index of income inequality for residents in the postal code where the student last attended school.

*median\_inc (supplemental)* \*\*

A continuous variable for the median income of residents in the postal code where the student last attended school.

*fall\_cum\_gpa (internal)*

A continuous variable for the high school grade point average of the student.

*fall\_midterm\_gpa\_avg (internal)*

A continuous variable for the averaged midterm GPA for the courses in which the student is registered in the fall.

*fall\_midterm\_gpa\_avg\_ind (internal)*

An indicator variable for the presence of a midterm GPA for the student in the fall with a reference category of non-graded.

*spring\_midterm\_gpa\_avg (internal)*

A continuous variable for the averaged midterm GPA for the courses in which the student is registered in the spring.

*spring\_midterm\_gpa\_avg\_ind (internal)*

An indicator variable for the presence of a midterm GPA for the student in the spring with a reference category of non-graded.

*remedial (internal)*

An indicator variable for remedial coursework with a reference category of non-remedial. Remedial coursework is defined as the student being registered in one or more courses designated as remedial.

*cum\_adj\_transfer\_hours (internal)*

A continuous variable for the cumulative number of credit hours transferred by a student.

*parent1\_highest\_educ\_lvl (internal)*

A categorical variable for the highest education level attained by the first parent/guardian of the student.

*parent2\_highest\_educ\_lvl (internal)*

A categorical variable for the highest education level attained by the second parent/guardian of the student.

*unmet\_need\_ofr (internal)*

A continuous variable for the unmet need of students relative to the amount of financial aid offered to them.

*unmet\_need\_acpt (internal)*

A continuous variable for the unmet need of students relative to the amount of financial aid accepted by them.

count\_week\_from\_term\_begin\_dt (internal)

A discrete variable for the week number in which the student applied during the application cycle.

*cahnrs\_anml (internal)* \*\*\*

An indicator variable for primary plan holder in Animal Sciences with a reference category of non-plan holder.

cahnrs\_envr (internal)

An indicator variable for primary plan holder in School of the Environment with a reference category of non-plan holder.

*cahnrs\_econ (internal)*

An indicator variable for primary plan holder in School of Economic Sciences with a reference category of non-plan holder.

*cahnrext (internal)*

An indicator variable for primary plan holder in CAHNRS other than those listed above with a reference category of non-plan holder.

*cas\_chem (internal)*

An indicator variable for primary plan holder in Chemistry with a reference category of non-plan holder.

*cas\_crim (internal)*

An indicator variable for primary plan holder in Criminal Justice with a reference category of non-plan holder.

*cas\_math (internal)*

An indicator variable for primary plan holder in Mathematics and Statistics with a reference category of non-plan holder.

*cas\_psyc (internal)*

An indicator variable for primary plan holder in Psychology with a reference category of non-plan holder.

*cas\_biol (internal)*

An indicator variable for primary plan holder in School Of Biological Sciences with a reference category of non-plan holder.

*cas\_engl (internal)*

An indicator variable for primary plan holder in English with a reference category of non-plan holder.

*cas\_phys (internal)*

An indicator variable for primary plan holder in Physics and Astronomy with a reference category of non-plan holder.

*cas (internal)*

An indicator variable for primary plan holder in CAS other than those listed above with a reference category of non-plan holder.

*comm (internal)*

An indicator variable for primary plan holder in College of Communication (including Strategic Communication and Journalism) with a reference category of non-plan holder.

*education (internal)*

An indicator variable for primary plan holder in College of Education (including Teaching and Learning and Kinesiology) with a reference category of non-plan holder.

*medicine (internal)*

An indicator variable for primary plan holder in College of Medical Sciences (including Speech and Hearing Sciences and Health Policy and Administration) or ESFCOM (including College of Medicine and Nutrition) with a reference category of non-plan holder.

*nursing (internal)*

An indicator variable for primary plan holder in College of Nursing with a reference category of non-plan holder.

*pharmacy (internal)*

An indicator variable for primary plan holder in College of Pharmacy and Pharmaceutical Sciences with a reference category of non-plan holder.

*vcea\_bioe (internal)*

An indicator variable for primary plan holder in Chemical Engineering and Bioengineering with a reference category of non-plan holder.

*vcea\_cive (internal)*

An indicator variable for primary plan holder in Civil and Environmental Engineering with a reference category of non-plan holder.

*vcea\_desn (internal)*

An indicator variable for primary plan holder in School of Design and Construction with a reference category of non-plan holder.

*vcea\_eecs (internal)*

An indicator variable for primary plan holder in EECS with a reference category of non-plan holder.

*vcea\_mech (internal)*

An indicator variable for primary plan holder in School of Mechanical and Materials Engineering with a reference category of non-plan holder.

*vcea (internal)*

An indicator variable for primary plan holder in VCEA other than those listed above with a reference category of non-plan holder.

*vet\_med (internal)*

An indicator variable for primary plan holder in College of Veterinary Medicine (including Neuroscience and Molecular Biosciences) with a reference category of non-plan holder.

Footnotes:

\* The outcome variable is used for model training based on prior years’ data. For the current year, the trained model tries to predict student enrollment outcomes next year.

\*\* These features have been adjusted for inflation in constant 2018 US dollars.

\*\*\* Excluded from this list are primary plan holders in Office of the Provost as including them yields a linear combination of variables.

**Data**

*Internal*

All institutional characteristics of students are drawn from internal snapshot or overnight data. These data adhere to the practices and procedures adopted by the Office of Institutional Research and the University.

*Supplemental*

The social and economic characteristics of the locales where students last attended school comes from the American Community Survey (ACS) TIGER/Line with Selected Demographic and Economic Data: <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-data.html>. Complete coverage of each of the geographies provided by the US Census Bureau occurs by aggregating single-year estimates over five-year periods. For example, the 2018 ACS 5-year Summary File spans from 2014 to 2018, while the 2017 ACS 5-year Summary File spans from 2013 to 2017.

Although widely used in research, academic literature varies in its approach to these multi-year data. Some scholars elect to have the aggregated data correspond to the ending year. In effect, the 2018 data would then be treated as if it were representative of the social and economic characteristics of the locales in 2018. Other scholars elect to have the aggregated data correspond to the middle year. In effect, the 2017 data would then be treated as if it were representative of the social and economic characteristics of the locales in 2015.

Due to data limitations, as well as the practical needs of the predictive model, the decision was made to follow the former logic with the addition of a lag period of two years. This was done for two reasons. First, the 2018 ACS 5-year Summary File is the latest data available from the US Census Bureau, with these releases typically occurring on a yearly basis on or around December 10. Second, given that new students entering in the current year likely attended high school in the prior year, a lag of a year or more was appropriate to better capture the time-ordered effects that the social and economic characteristics of the locales would have on students’ last attended schools. (An example SAS code file for preparing ACS data for the student risk model can be found here: [Z:\Nathan\Models\student\_risk\student\_risk\_acs\_prep.sas](file:///Z:\Nathan\Models\student_risk)).

Theoretically speaking, these accommodations can be justified in this way: as the primary interest is treating the variables derived from the ACS as social and economic characteristics that load on the locale where the students’ last attended schools, and not on the students themselves, the use of multi-year data in conjunction with a two-year lag provides the most relevant historical context. A student arriving in the 2020 academic year (Fall 2019 term) is using data from the 2018 ACS 5-year Summary File—spanning from 2014 to 2018—which covers the range of their high school career.

When necessary, the economic variables were adjusted for inflation according to the final year present in the released data. The methodology for this adjustment, as well as the annual average consumer price index (CPI) data needed to calculate it, are provided by the US Census Bureau: <https://www.census.gov/topics/income-poverty/income/guidance/current-vs-constant-dollars.html>.

**Models**

The modeling strategy used is that of an ensemble model, which can employ any number of machine learning algorithms commonly used for binary classification including logistic regression, support vector classification, stochastic gradient descent, multi-layer perceptron classification, and random forest classification. The intuition behind this strategy is that each algorithm has its own inherent strengths and weaknesses that can be balanced against one another to yield an aggregate model that is more robust to overfitting while also maintaining predictive accuracy.

*Logistic Regression*

Logistic regression with L2 regularization. Regularization is a method of penalizing complex models to prevent overfitting by including a penalty term along with the loss function being minimized. For this variant of L2 regularization, the penalty term is the sum of the squared coefficients scaled by the inverse of a hyperparameter, which is then added to the error term. A hyperparameter being an exogenous adjustable parameter. Lowering the value of the inverse hyperparameter promotes model underfitting, while raising the value of the inverse hyperparameter promotes model overfitting. The ideal value for the inverse hyperparameter produces a model that generalizes well to new data.

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

*Support Vector Classification*

Support vector classifier with a linear kernel. Support vector classification represents the data in an input space the dimension of the number of features. In the case of a binary classifier, the goal is to split the multi-dimensional input space into two partitions that divide the classes using a hyperplane. A hyperplane being an affine subspace that is of one dimension less than the input space. With a linear kernel, this partitioning is done linearly, which does not necessitate mapping the data to a higher-dimensional feature space as required by non-linear kernels. The general rule of thumb is parsimony. A linear kernel is preferred over non-linear kernels wherever suitable, especially when there are many features.

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

*Stochastic Gradient Descent*

Stochastic gradient descent with a modified Huber loss function. This classification model is a generalization of other stochastic gradient descent algorithms. Used with a modified Huber loss function, it is equivalent to a quadratically-smoothed support vector machine. Support vector machines with non-linear kernels represent the data in an input space the dimension of the number of features and map the data to a higher-dimensional feature space. In the case of a binary classifier, the goal is to split the multi-dimensional feature space into two partitions that divide the classes using a hyperplane. A hyperplane being an affine subspace that is of one dimension less than the feature space.

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html>

*Multi-Layer Perceptron*

Multi-layer perceptron neural network with a rectified linear unit (ReLU) activation function. While neural networks are mathematically more complex than other classification models, the intuition behind them is not quite so complicated. The objective is to find a set of mathematical functions (represented by the hidden layers) that map the features of the observations (represented by the input layer) to their target values (represented by the output layer) with minimal error. Neural networks allow for the hidden layers to be inherently non-linear through their combined use of weights, biases, and activation functions. An activation function is a non-linear monotonic function. Using the ReLU activation function, the output is of a given node is zero when its input is less than zero and it is equal to its input otherwise. One advantage of ReLU is its efficiency. More sophisticated activation functions can make it harder for the neural network to learn the weights and biases that produce the minimal amount of error.

<https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html>

*XGBoost Gradient Boosted Random Forest Classification*

Gradient boosted random forest classifier with a logistic objective function. This random forest classification aggregates several boosted decision tree classifiers that are fit to datasets sub-sampled from the original dataset using replacement. At a fundamental level, tree-based classification searches for split points among the data that partition the classes in such a way that the resulting loss function is minimized. The decision tree algorithm branches out iteratively by finding the next best feature in the sub-sampled dataset (and its associated value) that results in reducing the loss even further for the previously split nodes and so on until there is no improvement. The randomness introduced by combining the results of multiple boosted decision tree classifiers run on many bootstrapped variations of the original dataset produces an averaged classifier that is more robust to overfitting than standard gradient boosted decision tree classifiers.

<https://xgboost.readthedocs.io/en/stable/tutorials/model.html>

**Classification**

*Voting Classification*

Voting classifier with weighted average probabilities. Voting classification is an ensemble method that combines the predictions of the included estimators to improve their generalizability. The weighted average of the predicted probabilities for the base models is calculated for each observation. The chosen weights determine the relative contribution of each algorithm to the average outcome.

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html>

*Risk Thresholds*

The risk thresholds are set based on the predicted probabilities of non-enrollment in the next year produced by the voting classifier. Those students with predicted probabilities equal to or above the .66 level are considered high risk, those equal to or above the .33 level and below the .66 level are considered medium risk, and those below the .33 level are considered low risk. For model evaluation purposes a threshold of .50 is used. Those at and above the .50 level are evaluated as being predicted to withdraw while those below .50 level are evaluated as being predicted to be retained.