Machine Learning in Support of Student Success

Rachel Rucker, Vinh Dang (Ph.D.), Dipak Singh (Ph.D.) and Keith Hubbard (Ph.D.) | Stephen F. Austin State University

Our goal is to predict whether a student will finish the semester on academic probation by mid-term using university data.

Data

Most of the data about student activity for a given semester was scattered throughout multiple databases on campus. Our first step was combining all the data into one comprehensive dataset.

Features (Student Characteristics)

1. We began by grouping student meal plan swipe usage into weekly counts and as either breakfast, lunch or dinner.
2. Then, we merged these features with demographic and academic data including a student's gender, race, major, and the type of semester: Fall or Spring.
3. Because we would like the models we develop to be able to predict a student's semester outcome by mid-terms, we did not consider meal plan activity or any other data from beyond Week 8 of the semester.

Prediction

A student was labeled <u>on probation</u> if their semester GPA was less than 2.0.
There are very few students on probation compared to the large amount not on probation. This will cause some problems.

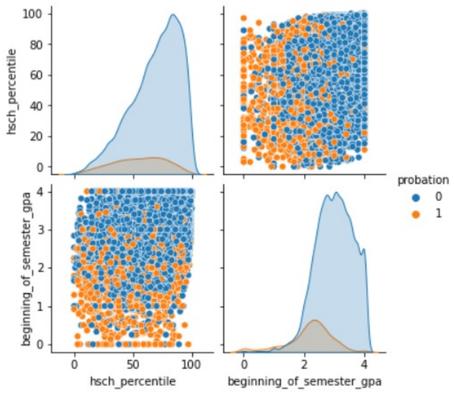


Fig 1: High School Percentile vs Incoming Term GPA

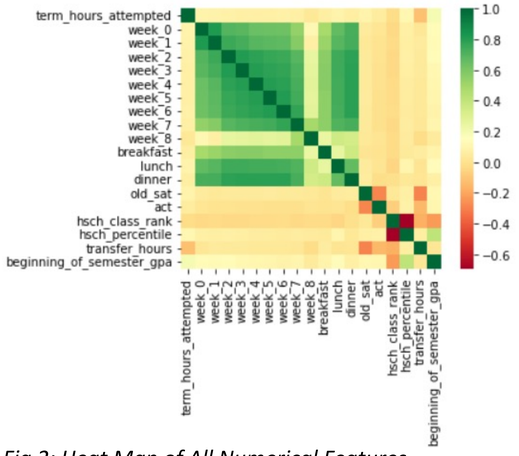


Fig 2: Heat Map of All Numerical Features

Models

Here, we tried to use the features that were gathered and joined in data processing from the first half of the term to predict student academic standing at the end of the 16-week term. These predictions, made halfway through the semester, will allow for intervention to help students succeed.

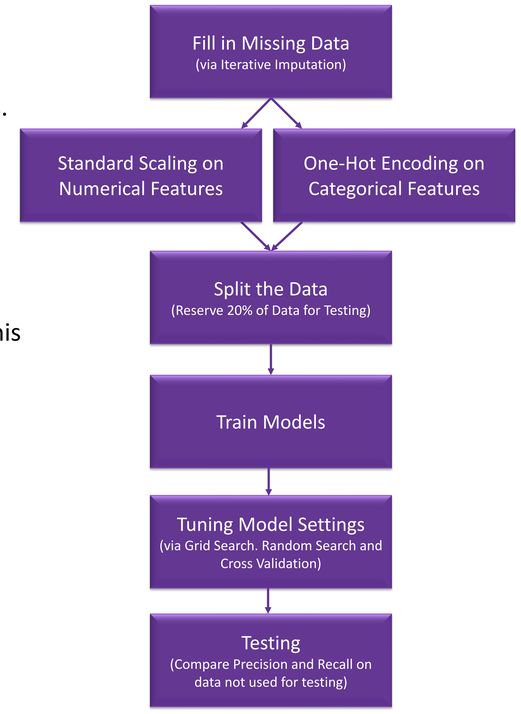
We used the pipeline to the right to train six individual models. These models are various methods by which the computer learns about the data before it predicts if each student will be on probation at the end of the semester.

Precision: Of all students that the model predicts will be on probation, this proportion truly are on-probation.
Recall: Of all students truly on probation, the model catches this proportion.

Table 1: Performance Metric Results

Model	Precision	Recall	F1-Score
Logistic Regression	0.711	0.113	0.195
LDA	0.581	0.225	0.325
QDA	0.420	0.289	0.342
KNN	0.770	0.067	0.124
Random Forest	0.708	0.108	0.188
SVC	0.661	0.133	0.222
Ensembled Model	0.775	0.065	0.120

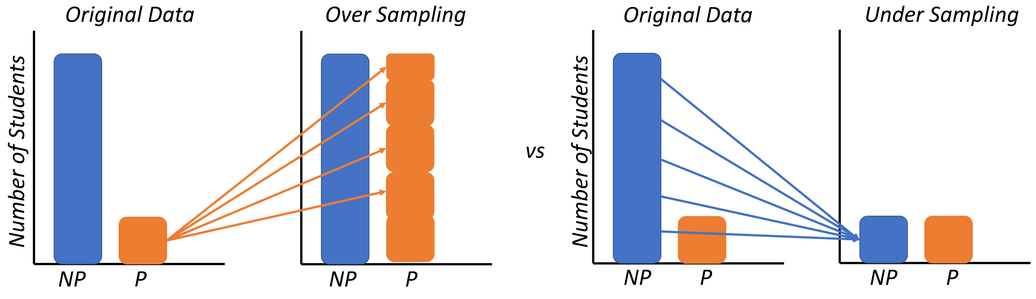
Pipeline



Improvements

Unsatisfied with the results above, we realized the imbalance between on probation and non-probation students was causing initial six models trouble. To fix this issue, we tried two different methods:

1. Over Sampling: We randomly choose students that were on probation and added duplicates of those records until we had an equal number of on-probation and non-probation students.
2. Under Sampling: We randomly choose only enough students that were not on probation to train the model with such that we would have equal numbers of each group.



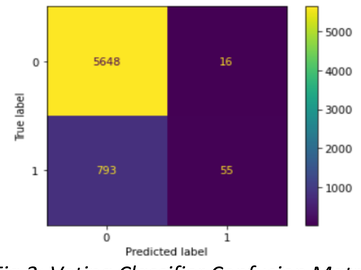


Fig 3: Voting Classifier Confusion Matrix

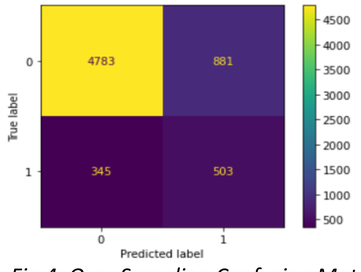


Fig 4: Over Sampling Confusion Matrix

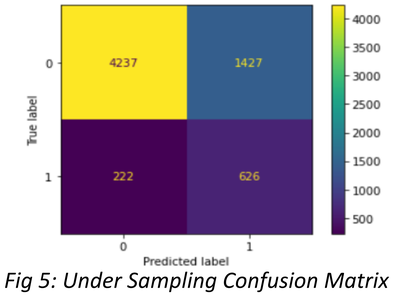


Fig 5: Under Sampling Confusion Matrix

Table 2: Performance Metric Results with Sampling

Sampling Approach	Precision	Recall	F1-Score
Over Sampling – Voting Classifier	0.363	0.593	0.451
Under Sampling – Voting Classifier	0.305	0.738	0.432
Voting Classifier on Imbalanced Data	0.775	0.065	0.120

This project was funded by SFA's President's Innovation Fund and completed as part of DS⁴ – Data Science Solutions for Student Success Research Group.