Delayed graduation rates, overpopulated classrooms, and money spent on additional resources to accompany these challenges have plagued universities since the boom of college attendance of the 21st century. One of the main causes of these problems are students retaking required classes, which causes bottlenecking. Bottlenecking is the scenario where many students retake a class, which causes that class to fill up very quickly, and thus results in new students who need to take the class the inability to successfully add into the class. This requires students to take that class in later semesters or if that class is a prerequisite for other classes, potentially forces them to drop that semester. This problem yields investigation into those classes where high-failure rates are present.

One of the classes at SDSU that has a high failure rate is Stat 119, an introductory business statistics class for non-stats majors that covers material from basic probability concepts to hypothesis testing and confidence intervals. Stat 119’s failure rate is around 24% give or take, which is really high considering this class is required for a multitude of majors. This class is usually taken by first-years, but there’s a wide demographic of students taking this class. Considering these attributes, this is the exact kind of class that should yield interesting information with in-depth data analysis.

The goal of this project is to investigate what type of student is likely to fail the class and if, with enough data, an early alert system can be implemented in future semesters that can intervene with students who are more likely to fail the class (perhaps an email be sent out to those students recommending supplemental instruction). Machine-learning models will be fit to the dataset, which originally contained 3123 observations and 131 variables. This dataset is the result of several merged datasets, as this data was collected from Fall 2017 through Fall 2018 via Blackboard and the school’s demographic libraries. The dataset contains demographic information such as prior academic history (HS GPA, SAT scores, units, etc.), descriptive attributes such as gender and ethnicity, and class information such as the grades received for various assignments and exams. The response variable is the grade variable, which was transformed to Pass/Fail, with C and above passing, and C- and below failing.

Grades on assignments are obviously highly correlated with the final grade. The purpose of this project is to uncover findings that are less intuitive and obvious in nature. As a result, grades predicting grades wouldn’t be very helpful in our analysis; clearly a student who does well on exams is likely to pass the class. A unique approach will thus be utilized in this analysis. Machine-learning models will be fit in two different phases, once before the semester begins (so, essentially only demographic information on the students taking the class), and then 4 weeks of class data will be added to those models to see if a better prediction rate can be obtained (and therefore this will narrow down students who are likely to fail the class).

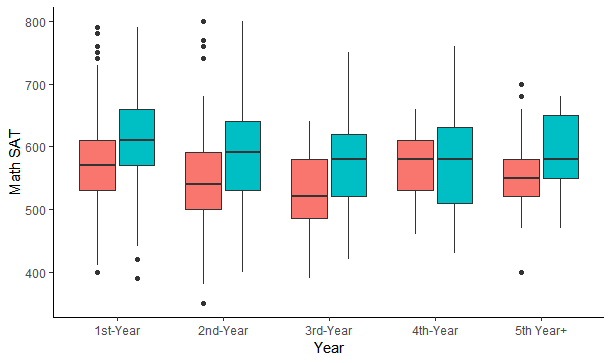
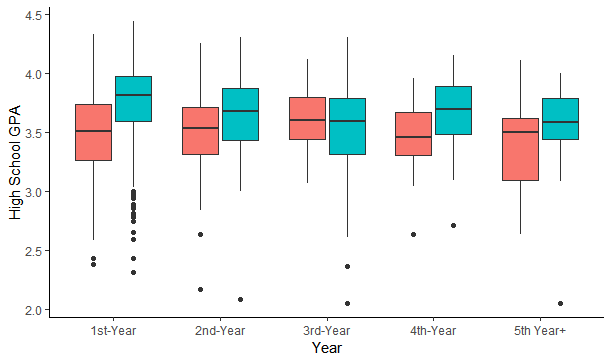
A support vector machine model and a random forest model will be implemented in this project. These are two classic machine-learning models (a support vector machine is more of a “computer scientist” approach to this problem whereas a random forest model is more of the “data scientist” approach). Before implementing these models, however, the data had to be preprocessed as it was very messy.

The appendix will detail specifics on how the data was cleaned and imputed, but in summary, columns and observations with large amounts of missing values were omitted as imputation would deviate too much from our analysis and was unnecessary. For a more granular analysis of our data, imputation of these observations would be something to consider, however. Important variables that were imputed include the SAT scores (which were imputed using a matching function with the ACT scores) and HS GPA. Demographic variables such as parent’s education levels or ethnicity were not imputed as these variables weren’t of high interest in our analysis, and coincidentally, students with those missing values also had missing values for many other demographics, so were dropped completely. The remaining continuous variables with missing values were imputed using the MICE package in R using the Random Forest method. After imputation and deletion of observations or irrelevant variables, there 2794 observations and 25 variables including the response variable. Please refer to the appendix for a list of all the predictor variable names.

Exploratory data analysis was based on intuition as far as what variables may be important in the analysis. First, some basic exploratory plots revealed a consistent failure rate between each period (Fall 2017, Spring 2018, Fall 2018), and a gender distribution relatively equal; of the 1432 women in the class, 23.7% failed and of the 1362 men in the class, 24.3% failed. In addition, the class was largely made of first years; of the 2135 freshmen, 21% failed, of the 423 second years, 38% failed, etc. Referring to the appendix will give detailed tables of this information.

Other variables also examined included the comparison of pass/fail rates with newly matriculated students and the distribution of colleges/majors. It’s well known that many new freshmen and transfers often struggle their first semester, but EDA proved otherwise as first-semester freshmen only had a 20% failure rate as compared to a 29.1% failure rate for other students. In addition, nearly half of the students taking the class were from the College of Business; this portion of the students had the lowest failure rate at around 20%. The other colleges were somewhat evenly distributed as far as the count goes and the failure rates ranged from 23-31%. It should be noted that for these variables the data was skewed towards newly matriculated freshmen and business majors; if these variables were normalized, we may see different results. As such, other variables are more likely to be important in the models. You may see further details listed in the appendix.

Since many of these students were freshmen, prior academic history such as HS GPA and SAT scores were intuitively worth examining. From the boxplots below, you can see that there’s a significant difference in HS GPA and Math SAT scores and the pass/fail rate:



Of those who passed the class, 1st years had the highest average high school GPA at around 3.7 and the highest average Math SAT at a little over 600. These averages decreased slightly for the other years, but the differences of the GPA/Math SAT were present throughout. In later years, the differences are less significant, but there were also a lot less 3rd+ years in these classes/transfers may have skewed the data slightly. It’s clear the Math SAT and the HS GPA are great indicators for being successful in the class. It should also be noted the SAT Composite scores were great indicators of success, but the variable is highly correlated with the SAT Math score, so it’s clear that relationship would be significant. Please refer to the plots in the appendix.

The Random Forest model was expected to run well. The Before Semester test misclassification rate was \_ and after adding the 4 weeks of homework assignments and quizzes, the test misclassification rate was \_. While the 4 weeks data model yields a better test misclassification rate, this method just reaffirms the belief that grades would simply predict grades; a variable importance chart in the appendix shows that the most important variables in the model were the various homeworks/quizzes.

The support vector machine model had a test misclassification rate of \_ before the semester and a then a test misclassification rate of \_ 4 weeks into the semester. The support vector machine ran better than expected as some of the variables were correlated, but as there are so many support vectors, this model may be overfitting.

|  |  |  |  |
| --- | --- | --- | --- |
| Random Forest | | Support Vector Machine | |
| Time Frame | Test Misclassification | Time Frame | Test Misclassification |
| Before Semester |  | Before Semester |  |
| ~4 Weeks In |  | ~4 Weeks In |  |

Both models ran mediocre test misclassification results, and in reality, although the 4 week models are more accurate, it’s simply because of the grade data (grades predict grades). The demographic dataset provides a more honest prediction of a student’s likelihood to pass/fail, especially when it comes to indicators like the Math SAT and HS GPA. The variable importance charts in the appendix reflect those sentiments as well.