US-China Education Review B, Mar.-Apr. 2023, Vol. 13, No. 2, 75-88
doi: 10.17265/2161-6248/2023.02.001

The Impact of Prior Courses and Grades on STEM FTIC Likelihood of Passing Calculus 1: A Naïve Bayes Classifier

Jay Sparks, Anthony Okafor, Josaphat Uvah
University of West Florida, Pensacola, U.S.A.

In this study, we investigate students' likelihood of passing Calculus 1, a keystone course, based on prior mathematics courses taken and their performance in those courses. With a sample of 918 First Time in College (FTIC) in Science, Technology, Engineering, and Mathematics (STEM) majors from 2015 to 2017 at a mid-sized public university in the United States, we used Naïve Bayes classification to analyze the data. We generated a predictive model with approximately 73% accuracy. Among the variables, receiving an A grade in a prior mathematics course was most likely to predict passing Calculus 1 at approximately 49%; it was also highly protective against failing. Our analyses showed that the type of prior course was only predictive up to 30% and was less useful than the prior course grade. Students' majors exhibited high variability and were not very predictive. Overall, courses that prepare students for Calculus 1 were less important than students' performance in those courses. In this regard, the higher the grade received in the prior course, the more likely the student was to pass Calculus 1. Thus, in terms of improving STEM retention and graduation rates, this suggests that improving students' performance would be more effective than revising curriculum in the prior courses.

Keywords: first time in college (FTIC), STEM and Calculus 1

Introduction

The United States government has prioritized STEM education at the tertiary level. Governing bodies in states have identified retention and graduation rates of students among the metrics on which institutions of higher education are evaluated, hence the need to create an incentive structure to improve the output of STEM graduates. Moreover, the marketplace rewards STEM graduates with higher median wages and job security. Thus, students are eager to pursue STEM majors. The poor retention and graduation rates of STEM majors in general points to at least one underlying problem in secondary education. Many students graduate from high school unprepared for the challenges of a STEM education in college. Almost uniformly, students in STEM majors are required to pass at least Calculus 1 before taking other higher-level core subjects. For this reason, students entering a STEM major should be "calculus ready", but many are not. It is theorized that this readiness gap contributes to the lower retention and graduation rates of STEM majors. Oftentimes, students who fail Calculus 1 either drop out of college or change away from STEM. Thus, Calculus 1 is a critical first step in completing a STEM education. Since early performance signals retention and timely graduation, it is important to develop models that can help to predict the probability that a student will pass or fail Calculus 1.

Jay Sparks, M.S. in Mathematical Sciences, Data Analyst, Predictive Analytics and Modeling (PAM) Lab, HMCSE, University of West Florida, Pensacola, FL 32514, U.S.A.
Anthony Okafor, Ph.D., Associate Professor of Mathematics & Statistics, University of West Florida, Pensacola, FL 32514, U.S.A.
Josaphat Uvah, Ph.D., Professor of Mathematics & Statistics, University of West Florida, Pensacola, FL 32514, U.S.A.

Statement of Problem

FTIC students enter STEM majors with varying levels of ability in mathematics as there are several paths that a student can take to Calculus 1. It is essential to investigate the effectiveness of these different paths by determining a student's probability to pass or fail Calculus 1 based on which path they took and how well they performed prior to Calculus 1. The primary goal of this study is to determine the likelihood that an FTIC student who takes a given path will pass or fail Calculus 1. Since performance levels at which a student can pass a course vary, we also determine the likelihood of a student's success in Calculus 1 given that student's grades in prior mathematics courses.

Relevance

It is common knowledge that a student's performance in the first year indicates that student's retention and graduation. Early higher performance improves student confidence and mastery of core subject matter. The more confident and prepared a student is, the more likely he or she is to complete the degree. The foundational role of Calculus 1 places it early in STEM degree programs. Thus, Calculus 1 is a keystone course for STEM majors. In order to improve students' learning, retention, and graduation rates, intervention strategies typically target students who are at the risk of failing keystone courses such as Calculus 1. It is in this regard that a good means of predicting students' likelihood of passing Calculus 1 becomes a tool for improving the effectiveness of intervention programs. More effective interventions for at-risk students can improve retention and in turn help fill the national gap in STEM graduates.

Literature Review

Research on retention and graduation rates in STEM is complex and often contradictory. Among other factors, attrition has been attributed to students' uncertainty about their goals and expectations, the difficulty in adjusting to college demands, and under-preparedness. Institutional research usually focuses on academic preparation as a predictor of retention and graduation because it is more easily inferred using available standardized data. However, the results from using measures of academic preparedness such as demographics, standardized test scores, and high school grade point averages (GPA) are mixed. We argue that students' performance in calculus courses is a good alternative to other predictive variables.

Chimka, Reed-Rhoads, and Barker (2007) performed survival analysis based on a Cox proportional hazard model that they used to examine the graduation rates of 429 engineering students at an Oklahoma university. They argued that higher standardized test scores in mathematics and science had a greater influence on graduation rates supposedly because they show better academic preparation. Yet when several variables were examined to predict retention and graduation rates for 8,573 first time in college (FTIC) students at a Florida university between 1984 and 1988, SAT and ACT scores had almost no relationship with those rates (Waugh, Micceri, & Takalkar, 1994). High school GPA was the best overall predictor. In another study, Mesa, Jaquette, and Finelli (2009) showed that advanced placement credits in high school might provide some improvement in college performance.

As Callahan and Belcheir (2017) and Simpson and Fernandez (2014) concluded in their respective studies, good performance early on in college improves retention and graduation. However, incoming high performers are not always retained. A group of 515 students at a major private university who were not retained from 2011 to 2012 was examined using latent profile analysis (Sulak, Massey, & Thomson, 2017). Among the classes of non-retained students, 40% were considered high achievers but only 32% left for academic reasons. Demographic factors have been suggested to predict the risk of non-retention because they infer social adjustment and socioeconomic status. Social factors and integration are known to improve retention and interventions for at-risk groups. See for example Dagley, Georgiopoulos, Reece, and Young (2016) and Scott, Thigpin, and Bentz (2017). Yet, other studies (Waugh et al., 1994; Xu, 2018) have concluded that demographics are a weak predictor of retention and graduation rates. For example, in one study (Waugh et al., 1994), African American students had the lowest retention rate because they had the largest proportion of under-prepared as deduced from their high school GPAs. However, when that proportion was controlled for, African American students had a rate of retention like other groups at the low end and higher rates of retention for those with higher high school GPAs. That is, well-prepared African American students were more likely to be retained than their white counterparts.

In general, STEM majors have lower retention rates than non-STEM majors. For instance, in the state of Tennessee, retention rates are significantly lower than the national average. A case in point: Xu (2018) invited 4,188 STEM majors from three public universities in Tennessee to complete a survey to identify factors influencing their intention to drop out or change majors. The results suggested that academic quality and support from faculty, cumulative college GPA, interest in the major, and ability to pay for education were most important. It is noteworthy that demographics had no effect.

Despite efforts to predict how students will do in STEM based on their characteristics or history, there is increasing evidence to suggest that training received in college is more important than demographics. See, for instance Scott et al. (2017) and Dagley et al. (2016). It has also been suggested that retention in STEM is self-selecting, meaning that less scientifically minded students are more likely to leave STEM. However, Jensen Neeley, Hatch, and Piorczynski (2017) in a study of 877 students from an elite, private midwestern university showed that scientific reasoning skills are developed through training in the first year of college. This indicates the importance of coursework in the scheme of things. Focusing on coursework as a predictor for retention highlights the need to take mathematics courses early and to do well. Predicting retention by coursework implies assumptions about the relationship between preparedness and performance. It is implied that students taking remedial coursework are less prepared for college, are less likely to persist; ergo, it follows that STEM students need to be "calculus ready" when entering college. However, this may not be the case.

Data

We obtained institutional data from a public university in the southeastern United States with an average enrollment of 12,940 students for the years of 2015, 2016, and 2017. The students were primarily nonresidential undergraduates. During those years, approximately 77% of students were undergraduates, 25% were STEM, and 5% were FTIC STEM. Table 1 below gives the breakdown of the data.

Table 1

Institutional Enrollment

	2015	2016	2017
Total Enrollment	12,808	12,979	13,033
Graduate Enrollment	2,606	2,915	3,079
(% of Total)	(20.35%)	(22.46%)	(23.62%)
Undergraduate Enrollment	10,202	10,064	9,954
(% of Total)	(79.65%)	(77.54%)	(76.38%)
Undergraduate STEM	2,654	2,711	2,906
(% of Undergraduate)	(20.72%)	(20.89%)	(22.30%)
Undergraduate FTIC	1,328	1,283	1,074
(% of Undergraduate)	(10.37%)	(9.89%)	(8.24%)
Undergraduate FTIC & STEM	440	490	443
(% of Undergraduate, % of FTIC)	(3.44%, 33.13%)	(3.78%, 38.19%)	(3.40%, 41.25%)

We obtained records of STEM majors from the college identified as FTIC students who took Calculus 1 after taking other mathematics courses. The total sample size was 948 across the three years.

Table 2

Cohort Frequency

	2015	2016	2017	2015-2017
Total	357	324	267	948
Pass	227	203	181	611
(% of Cohort)	(63.59%)	(62.65%)	(67.79%)	(64.45%)
Fail	130	121	86	337
(% of Cohort)	(36.41%)	(37.35%)	(32.21%)	(35.55%)

Description of Data

The FTIC records for STEM majors included numerous demographic variables along with course enrollment and grades. The only variables of interest in this study were a student's major, the mathematics courses taken prior to Calculus 1, respective grades earned in those courses, and their performance in Calculus 1. We excluded 30 students because their course grades prior to Calculus 1 were simply pass/fail or the students had changed out of STEM majors during that time. These majors include pre-licensure Bachelor of Science in Nursing, Information Technology, General Business, and Criminal Justice. Therefore, the sample size for analysis was reduced to 918 and their failure rate for Calculus 1 was around 32% across all three years.

There were five mathematics courses that students could enroll in before Calculus 1. These are College Algebra (MAC 1105), Intensive College Algebra with Lab (MAC 1105C), Trigonometry (MAC 1114), Pre-Calculus Algebra (MAC 1140), and Pre-Calculus with Trigonometry (MAC 1147). In terms of the contribution of each course to the cohort, Trigonometry had the highest enrollment (281, 31% of cohort) and number of passing students (187). Intensive College Algebra with Lab had the lowest enrollment (23, 3% of cohort) and passing students (16). The average enrollment was 184 students with 121 (65%) passing and 63 (35%) failing. However, for the ratios of students passing to failing Calculus 1 within individual classes, Trigonometry was close to average across the cohort while Intensive College Algebra with Lab was higher than average. Intensive College Algebra with Lab had the highest percentage of passing students (70%). College Algebra had the lowest percentage of passing students (57%). Figure 2 below shows the details.

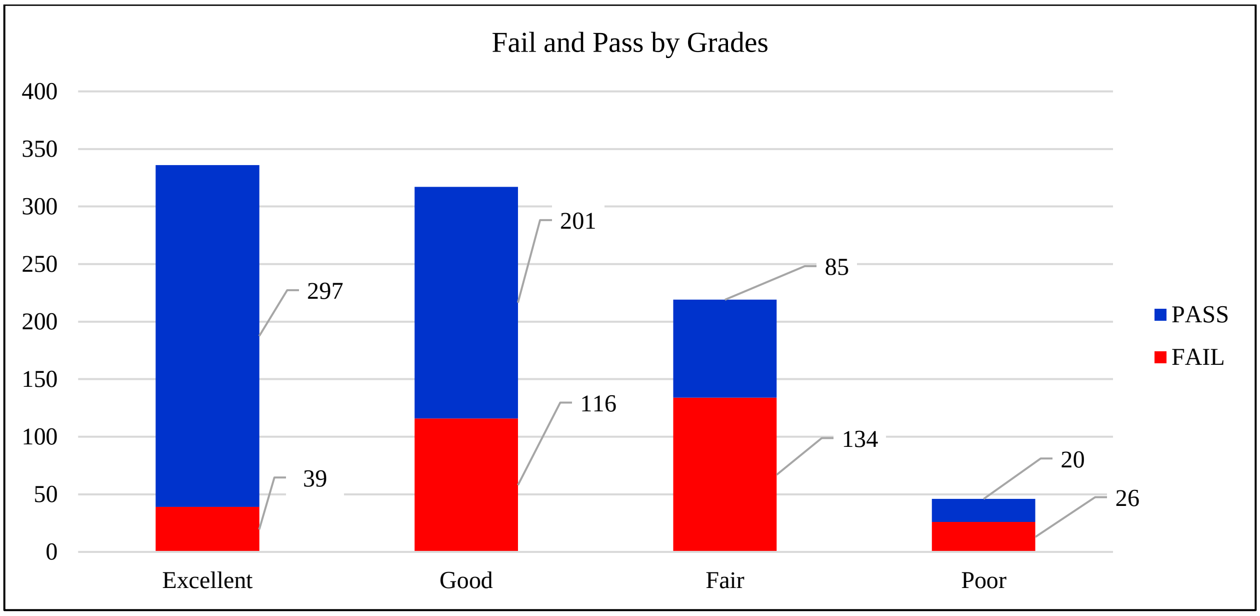

Figure 1. Frequency by grade rank.

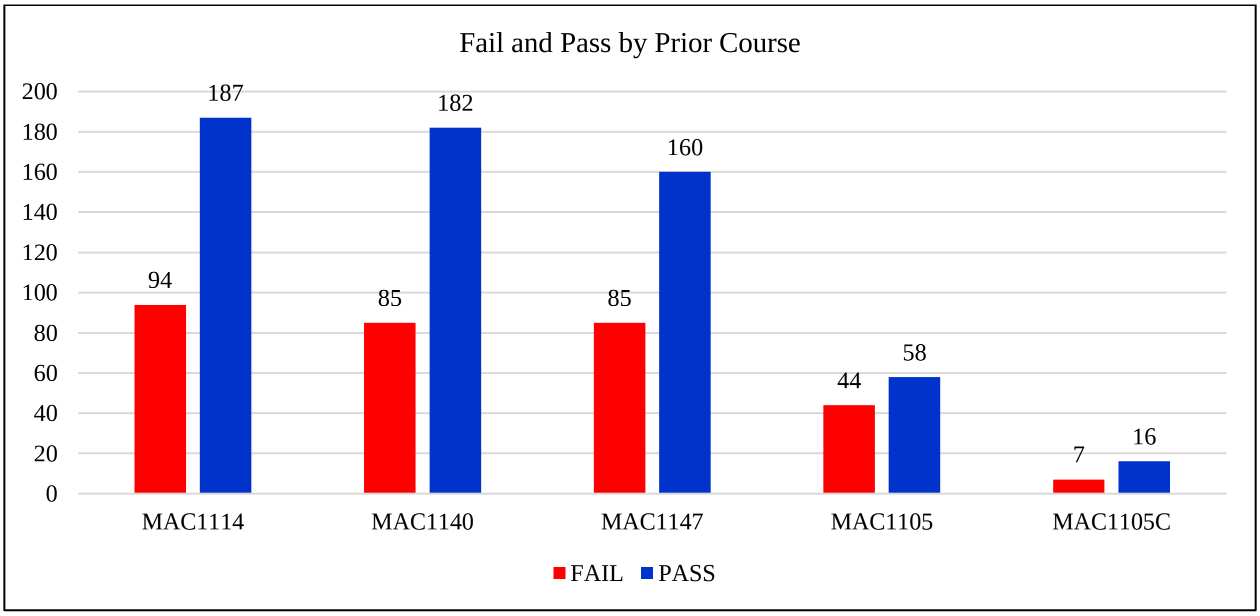

Figure 2. Frequency by course.

Among the 19 majors, the average enrollment was 48 students. The average number of passing students was 32. The top three most populous majors were Biomedical Sciences (127), Preprofessional Biology (121), and Marine Biology (101). These same three had the highest number of both passing students and failing students (Biomedical Sciences, 84, 43; Preprofessional Biology, 82, 39; and Marine Biology, 66, 35). In contrast, the least populous majors were Biochemistry (8), Engineering (9), and Pre-Computer Engineering (10).

In terms of ratios of passing to failing students among individual classes, the largest majors regressed closely to the cohort mean of 2.57. The situation among the smallest majors was much more variable. Biochemistry had the highest ratio of passing to failing students (7), while Engineering and Pre-Computer Engineering were among the lowest. This suggests that the ratio of passing to failing students is influenced by factors other than size of enrollment.

The top three majors with the highest pass rates were Biochemistry (88%), Physics (85%), and Mathematics (83%). Accordingly, these three had the lowest failing rates and best ratios of passing to failing students. Conversely, the top three with the highest fail rates were Natural Science (71%), Engineering (67%), and Cybersecurity (48%). Likewise, they had the lowest pass rates and the worst ratios of passing to failing students. The complete breakdown is shown below.

Table 3

Frequency by STEM Majors

STEM Majors	Fail (315)	Pass (603)	Total (918)
Biochemistry	1	7	8
Biology	26	40	66
Biomedical Sciences	43	84	127
Chemistry	6	19	25
Computer Engineering	14	18	32
Computer Science	24	49	73
Cybersecurity	28	30	58
Electrical Engineering	12	27	39
Engineering	6	3	9
Marine Biology	35	66	101
Mathematics	3	15	18
Mechanical Engineering	30	49	79
Natural Science	10	4	14
Physics	2	11	13
Pre-Computer Engineering	4	6	10
Pre-Electrical Engineering	19	42	61
Preprofessional Biology	39	82	121
Software Engineering	7	27	34
Undergraduate Non-Degree	6	24	30

There was performance variation within the cohort. The smallest majors had the highest overall performance while the largest majors were close to average. Enrollment size may suggest popularity, but less popular majors split students into both very high and very low performing groups. The smaller in size with higher performing majors may have been more selective or may have required greater mathematics preparation for their FTICs.

Method of Analysis

The features of this cohort were supervised in the sense that they were expressed at levels such as course names, grades that were awarded, and majors. Considering the limited amount of data (918) and the need for an efficient way to study the problem, the most useful linear classifier was Naïve Bayes Classifier (NBC). NBC is used because, as Hand and Yu (2001) have previously verified, its simplicity makes it highly effective.

Naïve Bayes Classification is a conditional probability model popular in machine learning that uses a finite set of prior outcomes to assign a probability to any future problem instance based on its features. In this study, the finite set of prior outcomes used to train the model was the FTIC STEM records from 2015 through 2017 as described by major, prior mathematics course, and prior mathematics course grade. The most important assumption for NBC is that the pre-courses, grades, and majors are independent (naïve) of each other.

Variable Coding

Each student record was coded based on 32 possible variables across three groups (Major, Prior Course, and Grades) as follows: (a) the 19 STEM majors; (b) the five possible Prior Courses (taken prior to Calculus 1); and (c) the four levels of performance. For this study, we classified grades of A and A- as Excellent, B+, B, and B- as Good, C+, C, and C- as Fair, and any other grade as Poor. We note that the Poor category comprised grades D+, D, F as well as I (incomplete) and W (withdrawal from course). As shown in Figure 2 above, grades that fall in the Fair, Good, and Excellent categories (i.e., A - C-) were regarded as passing grades while those in the Poor category were considered as failing grades. Table 4 below shows the data coding.

Table 4

Data Coding Table

Variables		Code	Description
Dependent	Calculus 1	Pass	A, A-, B+, B, B-, C+, C, and C-
		Fail	D+, D, F, I (Incomplete), W (Withdrawn)
Independent	Prior Course	MAC 1105	College Algebra
		MAC 1105C	Intensive College Algebra with Lab
		MAC 1114	Trigonometry
		MAC 1140	Pre-Calculus Algebra
		MAC 1147	Pre-Calculus with Trigonometry
	Grades	Excellent	Earned grade A or A-
		Good	Earned grade B+, B, or B-
		Fair	Earned grade C+, C, or C-
		Poor	Earned grade D+, D, F, I, or W
	Major	Biochemistry, Biology, Biomedical Science, Chemistry, Computer Engineering, Computer Science, Cybersecurity, Electrical Engineering, Engineering, Marine Biology, Mathematics, Mechanical Engineering, Natural Science, Physics, Pre-Computer Engineering, Pre-Electrical Engineering, Preprofessional Biology, Software Engineering, and Undergraduate Non-Degree	

Methods

The variation in passing or failing Calculus 1 was assumed to be affected by some combination of major, prior mathematics course, and prior mathematics course grade. Accordingly, this study proposes the following NBC probability model for predicting how likely a student is to pass or fail Calculus 1 based on these features. NBC is appropriate because its use of class labels and assumption of independence allow the formulation of a predictive model based on the already observed outcomes as discussed below.

Naïve Bayes Classification

Naïve Bayes classification applies class labels drawn from a finite set to problem instances. These labels are features or variables that are independent from one another. Given a problem representing some independent variable (or feature), the model assigns a probability to each possible outcome. The model uses a finite set of prior outcomes and features to determine the probability of future outcomes comprised of similar features. The calculation of the model uses Bayes' Theorem that can be stated as follows:

$$Pr(A|B) = \frac{Pr(B \cap \ A)}{Pr(B)}$$

We rewrite the equation in the form:

$$Pr(B \cap A) = Pr(A|B)P(B)$$

Consider the independency equation:

$$Pr(A \cap B) = Pr(B|A)P(A) \ \ since \ \ Pr(A \cap B) = Pr(B \cap A)$$

To produce Bayesian probability terminology:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

Considering Bayes' Theorem, a more explanatory form is given by:

$$Posterior \ \ Pr(A|B) = \frac{likelihood \ \ Pr(B|A) * prior \ \ Pr(A)}{evidnece \ \ Pr(B)}$$

Expanding the theorem to consider conditional independency assumptions:

$$PosteriorPr\left(A = \ a_k \middle| B_1 \cdots B_j\right) \text{ where,}$$
$$j = \ \{1, 2, \cdots \ \text{j}\} \text{ independent variables,}$$
$$k = \{1, 2, \cdots \ \text{k}\} \text{ possible outcomes,}$$
$$\text{prior } \ Pr(A) = Pr(A = a_k)$$

So that:

$$\text{likelihood } \ Pr\left(B_1, B_{2,} \cdots B_j \middle| A = a_k\right)$$
$$= Pr(B_1|A = a_k) \cdot Pr(B_2|A = a_k) \cdots \ Pr\left(B_\text{j} \middle| A = a_k\right) = \prod_j Pr\left(B_j \middle| A = \ a_k\right),$$

and:

$$\text{likelihood } \ P\left(B_j \middle| A = a_k\right) \cdot \text{prior } \ P(A = a_k) \ \ \propto P(A) \prod_{i=1}^{n} P\left(B_j \middle| A = \ a_k\right),$$

where $\propto$ denotes proportionality evidence:

$$\text{evidence } \ Pr(B) = \ \sum\nolimits_k Pr(A = a_k) \prod\nolimits_j Pr\left(B_j \middle| A = \ a_k\right).$$

Since the focus is the probability of B, notice that the *evidence Pr*(*B*) does not depend on a_k and the Na ïve Bayes classification rule is:

$$A \ \leftarrow \text{arg}\,\text{MAX}_{a_k} Pr(A = a_k) \prod\nolimits_j Pr\left(B_j \middle| A = a_k\right)$$

Assumptions of the Model

Given two events, A and B, the conditional probability (*P*) of the two in relationship to each other can be represented using Bayes' Theorem. Na ïve Bayes' has advantages in that each feature is assumed to be independent of any other feature. This decoupling means that the distribution of one feature can be independently estimated as a one-dimensional distribution. The model does not need to scale dimensionality with features. The model compensates for deficiencies in probability because it produces decisions based on the relative probability of a class in comparison to other possible classes. Thus, for a given feature, the model itself does not need to be accurate as long as its output has a higher probability than that of other possibilities.

Formulation of the Model

For this study, the *maximum a posterior* (or MAP) decision rule is as follows:

$$\Pr(A_{k=\{fail\ \text{or}\ pass)}) \leftarrow \arg MAX_{k=\{fail\ or\ pass\}} Pr\big(A = k_{(fail\ \text{or}\ pass)}\big)\prod_{j=1}^{n} Pr\big(B_j \big| A = a_k\big),$$

$$\text{where}, n = \text{prior courses, grade ranks, and majors}$$
$$k = \text{calculus class pass or fail}$$

For example, setting the outcome to be the failing probability of a biology major with an A grade in MAC1114 would be as follows:

$$Pr(fail\ |(MAC1114) \cdot (grade\ \ A = 1) \cdot (major = boiology)$$
$$= \arg\ \ MAX_{fail} Pr(fail) \prod_{i=1}^{\text{n}} Pr(MAC1114|fail) P(grade\ \ A = 1|fail) Pr(major$$
$$= biology|fail)$$

Analysis and Results

The major steps in Na ïve Bayes Analysis involve training the model using prior probabilities derived from the sample. That produces a model which then needs to be validated. The results are discussed below.

Descriptive Statistics

Of the cohort, Figure 3 (below) shows the prior portion in decreasing contributions by Prior Course. The greatest proportion comes from Trigonometry (31%), which contributed the most to pass (20%). The smallest comes from Intensive College Algebra with Lab (2%), which contributed the least to pass (3%). Note that the contributions to both pass and fail decrease with the proportion of each class.

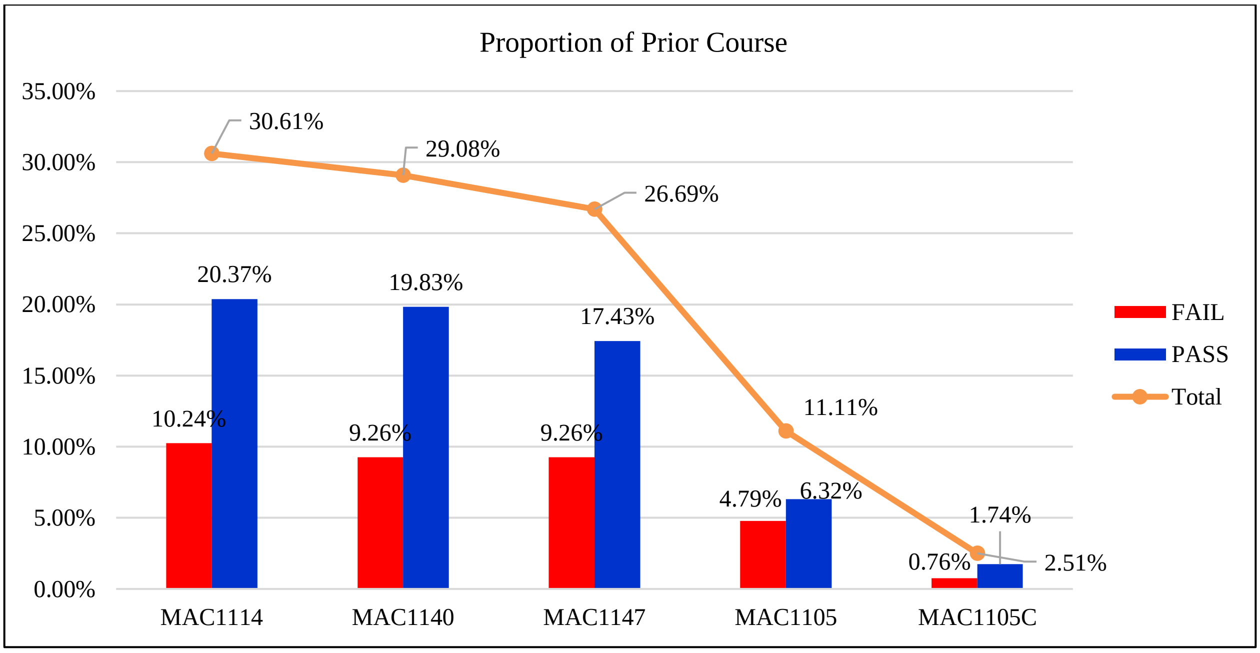

Figure 3. Proportion by course (each % is based on total sample size).

Of the cohort, Figure 4 below shows the portion in decreasing contributions by grades. As expected, the greatest proportion (31%) comes from students with excellent performance in their prior mathematics course (MAC1114), who contributed the most to pass (20%). The smallest contribution (2%) comes from students with MAC1105C, who contributed the least to pass (3%).

Interestingly, contributions to fail do not have an inverse relationship. They peak with students whose performance was in the C grade (including +/-) range. While students in the fair grade contribute more to pass than those in the poor grade (9% vs. 2%), they contribute much more to fail (15% vs. 3%). The fair grade group is the only one to contribute more to failing than passing. The contributions of students in the poor grade group to passing and failing are almost equal.

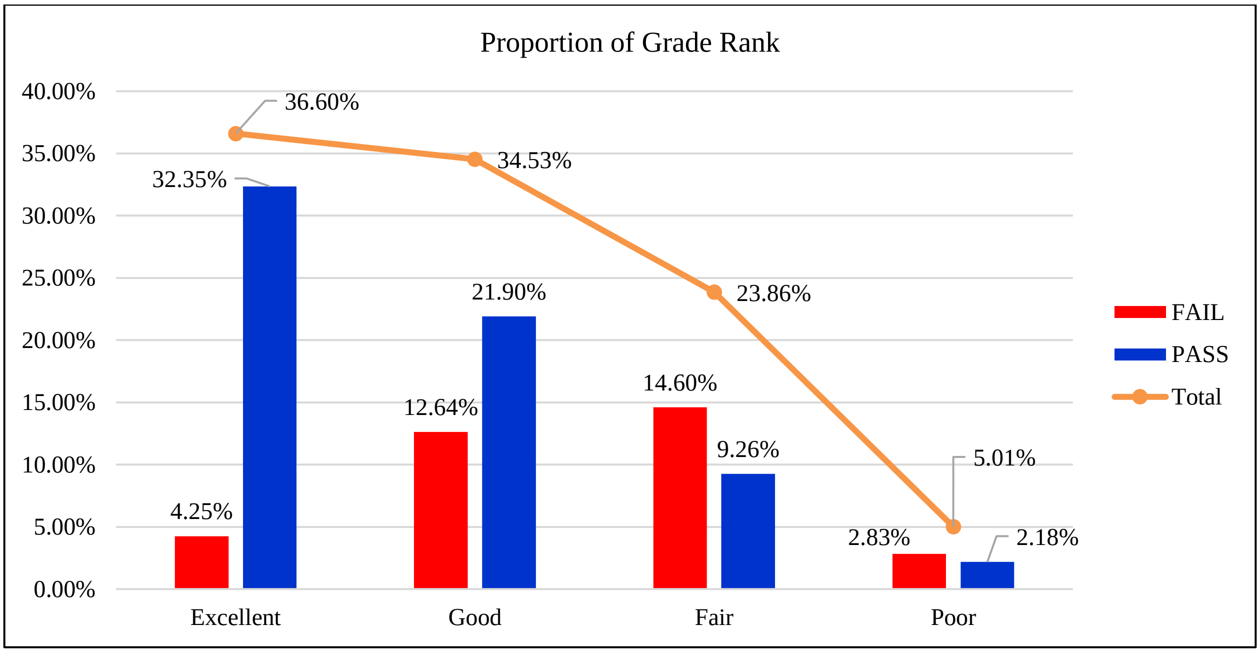

Figure 4. Proportion by grade rank (each % bases on total sample size).

Of the cohort, Figure 5 below shows the portion in decreasing contributions by major. This situation is more complex. The top three majors by contribution are Biomedical Sciences (14%), Preprofessional Biology (13%), and Marine Biology (11%). These majors had the greatest contributions to both pass and fail, with a ratio of almost 2 to 1. The bottom three majors by contribution were Biochemistry (1%), Engineering (1%), and Pre-computer Engineering (1%). These majors were dissimilar because Biochemistry contributed much more to pass than fail, at a ratio of almost 7 to 1. In contrast, Engineering contributed much more to fail than pass, at a ratio of almost 2 to 1.

The contributions of majors to pass and fail did not show a clear linear relationship. Grossly, majors with greater total contributions tended to contribute more to pass, but not always (e.g. Pre-Electrical Engineering vs. Biology, Undergraduate Non-Degree vs. Computer Engineering, Biochemistry vs. Engineering, etc.). Likewise, contributions to fail did not always follow total contribution (e.g. Cybersecurity vs. Biology, Computer Engineering vs. Electrical Engineering, Natural Science vs. Mathematics, etc.). These differences may be as a result of variability in requirements or selection admission criteria by major.

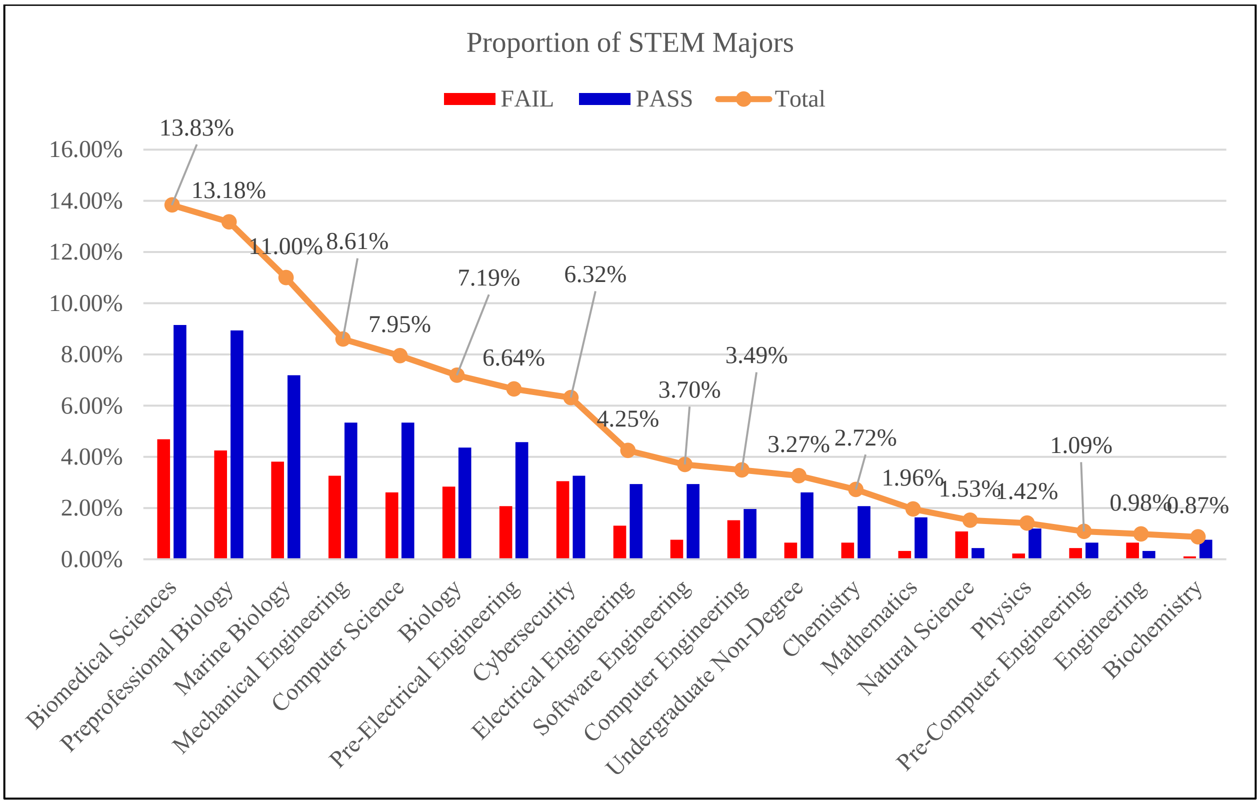

Figure 5. Proportion by major.

Software

The Na ïve Bayes Analysis was conducted using R (version 3.5.1). This included generating prior probabilities to train the model as well as model validation.

Na ïve Bayes Prior Probability

The descriptive statistics for the evidence variables received some discussion in this Section. Their prior probabilities as independent variables are shown below. The results of the Na ïve Bayes Classification Analysis are as follows. As shown in Table 5, the priori probability of the cohort to fail Calculus 1 was approximately 34%. The conditional probability of passing was 66%. The ratio of students likely to pass versus those likely to fail was close to 2 to 1.

Table 5

Priori Probability

	Calculus course	
Priori Probability	Fail	Pass
	34.31%	65.69%

The conditional probability for Calculus 1 performance is shown below. In terms of Prior Course, Trigonometry was most likely to pass (0.3101) and Intensive College Algebra with Lab was least likely (0.0265). Trigonometry was also most likely to fail (0.2984), and Intensive College Algebra with Lab was least likely to fail (0.0222). This tends to reflect the relative contributions of those two courses to the cohort as Trigonometry had the largest enrollment and Intensive College Algebra with Lab the smallest.

In terms of performance grades in prior courses, Excellent was the most likely to pass Calculus 1 (0.4925) and Poor (D, I, F, W) was the least likely to pass (0.0331). This tends to reflect the benefit of better prior performance as well as the peak failure for the fair group discussed above. At the high end of performance, the risk of failing decreases significantly when moving from the Good group to Excellent. Again, this illustrates the benefit of better prior performance.

In terms of majors, Biomedical Science was the most likely to pass (0.1393) and Engineering the least likely to pass (0.0049). This tends to reflect their relative total contributions. We note that Biomedical Science was also the most likely to fail (0.1365) and Biochemistry the least likely to fail (0.0031). Biomedical Science had almost an equal chance to pass or fail, while Biochemistry had a much better chance to pass than fail.

Table 6

Prior Probability by Category and Variable

Independent variables	Categorical	Priori probability to pass Calculus 1	
		Fail	Pass
Prior Course	MAC 1105C (Intensive College Algebra with Lab)	0.0222	0.0265
	MAC 1105 (College Algebra)	0.1397	0.0962
	MAC 1114 (Trigonometry)	0.2984	0.3101
	MAC 1140 (Pre-Calculus Algebra)	0.2698	0.3018
	MAC 1147 (Pre-Calculus with Trigonometry)	0.2698	0.2653
Grade Rank	Excellent (A or A-)	0.1238	0.4925
	Good (B+, B, or B-)	0.3682	0.3333
	Fair (C+, C, or C-)	0.4253	0.1409
	Poor (D+, D, F, I, or W)	0.0825	0.0331
Major	Biomedical Sciences	0.1365	0.1393
	Preprofessional Biology	0.1238	0.1359
	Marine Biology	0.1111	0.1094
	Mechanical Engineering	0.0952	0.0812
	Cybersecurity	0.0888	0.0497
	Biology	0.0825	0.0663
	Computer Science	0.0761	0.0812
	Pre-Electrical Engineering	0.0603	0.0696
	Computer Engineering	0.0444	0.0298
	Electrical Engineering	0.038	0.0447
	Natural Science	0.0317	0.0066
	Software Engineering	0.0222	0.0447
	Undergraduate Non-Degree	0.019	0.0398
	Chemistry	0.019	0.0315
	Engineering	0.019	0.0049
	Pre-Computer Engineering	0.0126	0.0099
	Mathematics	0.00952	0.0248
	Physics	0.0063	0.0182
	Biochemistry	0.0031	0.0116

Validation Results

In machine learning, data are partitioned for training a model and then validating it through testing. In this study, the resulting model above was generated using the training data set (0.9). The validation results below represent the remaining test set (0.1).

The model training data produced 448 true positives, 130 false positives, 97 false negatives, and 159 true negatives. The accuracy was approximately 73%. Validation data produced 52 true positives, 12 false positives, 6 false negatives, and 14 true negatives. The accuracy was approximately 79%. The results are shown in the combined confusion matrices below.

Table 7

Training (90%) and Test (10%) Confusion Matrices (True Predictions in Italics)

Predicted	Actual			
	Training data (Model)		Test data (Validation)	
	Pass	Fail	Pass	Fail
Pass	*448*	130	*52*	12
Fail	97	*159*	6	*14*
Accuracy	72.78%		78.57%	

A more detailed breakdown of the model based on validation is shown below. The results indicate relatively high accuracy at 0.7278. Since the accuracy rate is higher than the no information rate (0.6535), the model is reliable. Testing did not reveal any major inconsistencies in the model.

Table 8

Model Validation Results

	Train data (90%)	Test data (10%)
Prior probability	(0.3465, 0.6535)	(0.3095, 0.6904)
Accuracy	0.7278	0.7857
95% CI	(0.6962, 0.7578)	(0.6826, 0.8678)
No inform. rate	0.6535	0.6905
p-values [ACC > NIR]	2.57E-06	0.03518
Kappa	0.3824	0.4646
McNamara’s test *p*-values	0.0337	0.2385
Sensitivity	0.8220	0.8966
Specificity	0.5502	0.5385
Pos. pred. value	0.7751	0.8125
Neg. pred. value	0.6211	0.7000
Prevalence	0.6535	0.6905
Detection rate	0.5372	0.6190
Detection prevalence	0.6930	0.7619
Balanced accuracy	0.6861	0.7175

Conclusions

We examined data for FTIC STEM students by major, prior course, and grades. We used Naïve Bayes analysis to classify the results of FTIC STEM majors across three years of data. The resulting model revealed significant differences in the likelihood of passing calculus based on major, prior course, and grades. Table 6 above summarizes those results. As discussed above, the model produced reliable results.

The likelihood of passing calculus based on categories of variables (from most to least influential) was the performance grades of the prior mathematics course, then the prior mathematics course, and then major. The grades most likely to contribute to passing were in the Excellent category (A, A-). The prior course most likely to contribute to passing was Trigonometry. The major most likely to pass was Biomedical Science. Our analyses showed that a student's prior mathematics course is highly likely to affect their passing Calculus 1. Of greater importance are the grades earned in their prior course.

Limitations and Suggestions for Future Research

Calculus 1 was selected for this study because it appears early in the sequence of courses to be taken by FTIC STEM majors; it cuts across majors and is known to affect retention and graduation in STEM. However, other keystone courses exist within majors or groups of majors. In order to have a fuller picture, these other keystone courses should receive similar examinations to truly determine the effect of prior courses and grades for STEM majors.

The current study did not consider demographic variables. Previous literature points to confounding results in this regard. While some studies suggest that these are important, others indicate that they are important only under certain conditions. Perhaps incorporating demographic variables into a Naïve Bayes analysis to compare them with variables included here may yield good results.

More populous majors had a greater impact on passing Calculus 1 than majors with a smaller number of students. However, the odds ratios for the larger majors appeared to regress towards a mean. There was more variation in odds ratios among smaller majors but this much higher variation in passing Calculus 1 was not readily explainable from our analyses. Perhaps the selection criteria or requirements for admission into those majors contribute towards their students' performance.

References

Callahan, J., & Belcheir, M. (2017). Testing our assumptions: The role of first course grade and course level in mathematics and English. *Journal of College Student Retention: Research, Theory & Practice, 19*(2), 1610175.

Chimka, J. R., Reed-Rhoads, T., & Barker, K. (2007). Proportional hazards models of graduation. *Journal of College Student Retention: Research, Theory & Practice, 9*(2), 221-232.

Dagley, M., Georgiopoulos, M., Reece, A., & Young, C. (2016). Increasing retention and graduation rates through a STEM learning community. *Journal of College Student Retention: Research, Theory & Practice, 18*(2), 167-182.

Hand, D. J., & Yu, K. (2001). Idiot's bayes—Not so stupid after all? *International Statistical Review, 69*(3), 385-398.

Heiny, R. L., Heiny, E. L., & Raymond, K. (2017). Placement model for first-time freshmen in calculus I (mathematics 131): University of Northern Colorado. *Journal of College Student Retention: Research, Theory & Practice, 19*(3), 270-283.

Jensen, J. L., Neeley, S., Hatch, J. B., & Piorczynski, T. (2017). Learning scientific reasoning skills may be key to retention in science, technology, engineering, and mathematics. *Journal of College Student Retention: Research, Theory & Practice, 19*(2), 126-144.

Mesa, V., Jaquette, O., & Finelli, C. J. (2009). Measuring the impact of an individual course on students' success. *Journal of Engineering Education, 98*(4), 349-359.

Scott, T. P., Thigpin, S. S., & Bentz, A. O. (2017). Transfer learning community: Overcoming transfer shock and increasing retention of mathematics and science majors. *Journal of College Student Retention: Research, Theory & Practice, 19*(3), 300-316.

Simpson, J., & Fernandez, E. (2014). Student performance in first year, mathematics, and physics courses: Implications for success in the study of electrical and computer engineering. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings* (pp. 1-4). Madrid: IEEE.

Sulak, T. N., Massey, J., & Thomson, D. (2017). Using latent profile analysis to harness the heterogeneity of nonretained college students. *Journal of College Student Retention: Research, Theory & Practice, 18*(4), 457-473.

Waugh, G., Micceri, T., & Takalkar, P. (1994). Using ethnicity, SAT/ACT scores, and high school GPA to predict retention and graduation rates. Retrieved from https://files.eric.ed.gov/fulltext/ED453704.pdf

Xu, Y. J. (2018). The experience and persistence of college students in STEM majors. *Journal of College Student Retention: Research, Theory & Practice, 19*(4), 413-432.