Tech Know Learn (2017) 22:299–315
DOI 10.1007/s10758-017-9334-z

ORIGINAL RESEARCH

Predicting Student Success: A Naïve Bayesian Application to Community College Data

Fermin Ornelas[1] · Carlos Ordonez[2]

Published online: 17 August 2017

Abstract This research focuses on developing and implementing a continuous Naïve Bayesian classifier for GEAR courses at Rio Salado Community College. Previous implementation efforts of a discrete version did not predict as well, 70%, and had deployment issues. This predictive model has higher prediction, over 90%, accuracy for both at-risk and successful students while easing interpretation and implementation. Predictive results across eleven courses and cumulative gain charts show potential improvements to be made in students' academic success by focusing on high level risk students. Researchers at other colleges might find this empirical application relevant for implementation of early alert systems.

Keywords At risk students · Cumulative gains · Naïve Bayesian · Predictive model

1 Introduction

Academic institutions today face several challenges driven by cost concerns, increasing accountability, and diminishing resources. For instance, a recent report (Oliff et al. 2013) assessed state financial cuts to higher education for the fiscal years 2008–2013. Among its main findings were: all states except for North Dakota and Wyoming saw severe reductions in higher education funding; 11 states cut educational funding by more than a third; 36 states shrank funding by more than 20%; and Arizona and New Hampshire occupied first and second place in the list among those states, decreasing their funding to higher education by 50%. Meanwhile, graduation rates for young adults have been stagnant for at least twenty

✉ Fermin Ornelas
Fermin.Ornelas@riosalado.edu

[1] Rio Salado College, Tempe, AZ, USA

[2] University of Houston, Houston, TX, USA

years. The Lumina Foundation (2015) reported that in the 1970s the college graduation rate was 40% and recently reported it at 45.8%. However, according to a newspaper report (Rampell 2013) the trend has improved in the last five years. Furthermore, there is an ongoing federal effort to increase graduation rates while preventing college cost increases (President Obama 2013 State of the Union Address 2013). Therefore, to address the low graduation rates problem institutions have responded with additional online academic options to entice students to continue their education. For instance, Allen et al. (2016) reports that more than one in four students takes at least one online course. Subsequently, to monitor student progress new decision tools primarily used by the business community are now being tailored to the needs of academic institutions. Predictive modeling, profiling and segmentation, which are tools used for portfolio risk management and targeted marketing in the financial industry, are now utilized to monitor students' academic progress and to customize programs for student academic engagement (Bienkowski et al. 2012; Eduventures 2013; ECAR-ANALYTICS Working Group 2015). Online behavior such as, students—instructors interactions, student-to-student contacts, number of logins to class material, on time or lack of assignment submission and grades are being appended to demographic attributes to predict student academic success (Hung and Zhang 2008).

Rio Salado Community College (RSCC) offers most of its courses on line and to monitor its students' progress it has implemented a Naïve Bayesian (NB) classifier into its LMS, Rio Learn. Moreover, it has implemented a faculty driven support program intended to provide timely customized course feedback to meet students' needs, Guided Evaluation Assessment Response (GEAR). The program is a "technology-based, faculty-developed solution that contains an integrated set of teaching tools intended to increase feedback quality and consistency, as a fundamental component for providing guidance that promotes learning as part of assessment. The system provides students with enhanced feedback, consistent grading, and an improved learning experience (Rio Salado College Report of Student Learning 2013). The predictive model uses student attributes to generate warning indicator levels alerting instructors on how his/her students are performing in the course. The NB classifier though, has converted the predictor attributes to comparative measures. The predictors used by the current classifier implemented at RSCC are: number of logins, number of site engagement activities, total points earned, total points submitted, credit load and weighted versions of logins and site engagement (Smith et al. 2012). This approach in the pilot study has presented roll out issues to additional courses and results are less appealing for analysis and actual decision-making. Therefore, the objectives of this paper are: (1) to report empirical findings of a redesigned approach; (2) to demonstrate predictive efficiency gains derived from modeling with continuous attributes in the relational database; and (3) to compare Naïve Bayesian classifier to logistic regression results. The next section focuses on related online-learning research, data description and aggregation will follow, next we discuss the methodology followed by empirical results and finally the paper concludes with some recommendations arising from this empirical study.

2 Literature Research

To put this project research into context, we looked at several studies focusing on student success. Barber and Sharkey (2012) reported on two logistic regression models predicting student course success at the University of Phoenix through course week 4. One model had the following variables: <65% points in prior courses, >85% points in prior courses, credits earned at the university of Phoenix and cumulative points earned. They considered three risk tiers: high risk, low risk, and grey zone. Their findings were that the model predicted passing (low risk) or failing (high risk) accurately often 90% of the time. The second model added to those variables, non-current financial status, credits earned to credits attempted ratio, transfer credits higher than 18, days until first activity date, number of online posts, and point delta to prior courses. Credits earned to credits attempted ratio and non-current financial status both were strong indicators of students' difficulties. The investigators concluded that adding these new variables increased the predicted accuracy of the second model.

Smith et al. (2012) were the early developers of the currently used classifier at RSCC. The predictors used by the current classifier implemented at RSCC are: number of logins, number of site engagement activities, total points earned, total points submitted, credit load and weighted versions of logins and site engagement. To guide instructor support, these researchers created a three-warning risk level system: low, moderate, and high for student successful class completion with a C or better. The model was tested on a pilot class and found that it correctly classified 70% of students in the high-risk category but did not do as well identifying students in the remaining warning levels. This effort was an analytic improvement as it has served to layout the foundations for tracking and monitoring student performance and further development in predictive modeling at RSCC.

Liu et al. (2009) wanted to measure the effect of social presence on course retention and final grade for students taking online community college courses. Using survey data on social presence they estimated two logistic regression models. One model was developed to predict course retention using a dichotomous indicator for success and another ordinal model to predict final grade. The later model specification, had grades as a multilevel dependent variable. From survey data collected after the third week into the semester, they concluded that there was a positive relation between social presence and course retention. The odds of course retention were 1.015 more times for each unit increase in social presence score. Similarly, in the ordinal model predicting grade level as a function of social readiness, they concluded that the higher the social presence the higher the chances of a better grade. Their recommendation was to develop tools for early identification of at-risk students and create effective interventions intended to increase students' social presence.

Hung and Zhang (2008) analyzed patterns of online behaviors to make predictions on learning outcomes for 98 students enrolled in a business course in Taiwan. The variables included in the analysis were: final grade, total frequency of LMS logins, total frequency of accessing course material, last time accessed course materials, number of bulletin boards messages posted, number of synchronous discussions attended, hours spent reading bulletin board messages, and number of board bulletin messages read. Descriptive and predictive analysis was undertaken and a decision tree was applied to build a predictive model of online learning performance. Decision trees are rule driven algorithms where an outcome variable—root-relation to a set of attributes is divided into various segments—nodes-based on significant Chi square values at each level, (Buntine 1992). Among the empirical findings discussed by Hung and Zhang, were: frequency of accessing course material was the most important variable for performance prediction. Students accessing the course material more than 44.5 times had improvement in their grade to 89.62. If students read more than 66.5 messages the corresponding grade would improve from 72.57 to 88. Overall, Hung and Zhang (2008) found that when students were more actively engaged tended to perform academically better. Thus, accessing the course material and actively participating on online discussion were strong performance predictors.

Shelton, Hung and Baughman (2016) utilized time-series clustering analysis to predict course failure for graduate students in teacher education. Data from the spring semester was collected and divided into training and validation at 60 and 40%, respectively. Static demographic variables combined with performance dynamic variables were employed to estimate six models: decision tree, boosting, logistic regression and rule induction. Based on their respective misclassification rates in the validation data set, the researchers concluded that the decision tree was the best model in predicting at-risk students at the 10th week into a sixteen week course; it captured 78.6% of at risk students. However, according to Shelton et al. (2016), predicting at the 10th week presented a problem when designing timely successful interventions. A subsequent paper by Shelton, Hung, and Lowenthal (2017) studying the same population, intended to enhance both timing and increase in prediction rates by taking into account variances in learning patterns and course activity requirements. Utilizing the time difference in login data they successfully augmented the model prediction accuracy earlier at week six into the semester. This new effort led to higher prediction rates, 85.45%, while model accuracy was at 89.26%.

Ifenthaler and Widanapathirana (2014) undertook a major learning analytical project for Australian universities. They focused on two aspects of the analytical framework: student profile and learning profile. The former refers to static and dynamic parameters inherent to the individual (i.e. demographics, learning strategies, motivation, social media skills, etc.). The latter, relates to variables within the learning management system such as: time per session, time on task, time on assessment. Other parameters included were: login frequency, task completion rate, assessment activity, assessment outcome, learning material activity, discussion activity, support access, rating of learning material, assessment, support, effort, etc. Their estimation methods applied were multiple regression and support vector machines (SVM). Several model versions were built to analyze the data for each profile. For the student profile a sample of 146,001 students encompassing 1509 study units were gathered. They identified as more important variables associated with study unit outcome: historical grade, historical cumulative fails and highest level of prior education. The preferred model (6) accounted for 80% accuracy in predicting study unit outcome. For the learning profile, they focused on two units at one institution impacting 12,002 students. According to them assessment attempts, learning materials accessed, and self-assessment were the most important variables. Of the two SVM learning models built, model 1 predicted the study unit outcome with an accuracy of over 90%. Moreover, to assess the performance of the model they divided the study period in four equal intervals. The explained variance went from .528 in the first period to .878 in the last period suggesting that the model performance enhanced with the increased interaction. Furthermore, the predictive accuracy of the SVM model increased overtime over 90%. In both analyses, the authors stressed the importance of addressing students' needs early in the study unit to prevent attrition and increase success.

Macfadyen and Dawson (2010) conducted a pilot study to assess the usefulness of LMS tracking data to predict student success in an online undergraduate Biology course at the University of British Columbia in 2008. Data gathered at the student level included term counts for frequency usage of course material and tools supporting content delivery, engagement and discussion, assessment and administration/management. Moreover, total time spent on tool-based activities such as: assessments, assignments, and total time gave a measure of time-on-task by the student. They estimated two statistical models: (1) a multiple regression model to predict grade as a function of total number of discussion messages posted, number of completed assignments and number of messages sent; (2) a binary logistic model with the same set of predictors where the class event defined students at risk if final grade was <60, otherwise the student was successful in the course. The main empirical finding from model (1) was that more than 30% of the variation in student final grade was explained by the set of independent attributes. Likewise, model (2) correctly identified 70.3% of the students at risk of failure. Interestingly the most predictive attribute in the logistic model was the variable measuring total student contribution to course discussion forums. This empirical fact validated student peer engagement as part of the learning process for student success. Note that these studies focused on predicting success and retention for a single course using course behavior predictors. Our study applies the NB model to eleven high enrollment courses using similar learning predictors because they are dynamic and highly modifiable through targeted interventions by faculty and instructors.

The outcome of interest, success and non-success in a course can be seen as an event classification problem and as such it can be described by a Bernoulli probability distribution (Elkan 2014) suitable for estimation with logistic regression or NB. In this research the authors pursued a continuous Naïve Bayesian classifier because the predictors selected were continuous, easiness of interpretation and implementation. In the following section we discuss data set collection steps for the development of the NB classifier and some sample characteristics.

3 Data Collection Process

To build the continuous Naive Bayesian classifier we extracted the data from various tables residing in the SQL server, generated from Rio Learn the internal LMS system. The activity table records all student transaction interactions with the course material and the corresponding instructor. These activities were condensed to create four learning dynamic variables: number of logins, site engagement, weighted logins, and weighted site engagement. A separate database from the Maricopa Community College District capturing course modality, number of credits, grading and course enrollment provided student performance fields used to elaborate points earned, points submitted, and credits load.

The courses selected experience high enrollment and since they are geared towards either an associate degree or a university transfer course length is mostly 16 weeks. The courses included are: BIO100, CHEM130, CHEM130LL, CRE101, ECN212, ENG101, ENG102, FON241, FON241LL, GBS233, HIS103, HIS104, and PSY101. Rio Salado has forty eight weekly start dates with classes beginning every Monday and with different course lengths. Going forward an adjustment in the attributes calculation and segmentation is likely to require further research to reflect this dynamic process. For this stage of the research, the data has been aggregated by student ID, actual class event and course included in the GEAR program for students enrolled in the fall 2012 through spring 2013. The fall semester data was utilized to develop the predictive model while the spring data set was kept separate for conducting out of sample validation. Actual success frequencies for both training and validation showing total success for the courses selected are provided in Table 1 below.

Success is defined as achieving a C or higher in a course. Success rates are 56.8 and 60.1% for training and validation data sets, respectively; while non-success figures are 43.2 and 39.9%, accordingly. Moreover, Table 2 provides some demographic characteristics of students enrolled in the fall of 2012 by success indicator. Since 60% of the students in the

Table 1 Frequency distribution for training and validation samples

Success indicator	Training		Validation	
	Frequency	Percent	Frequency	Percent
0	2564	43.2	1085	39.9
1	3372	56.8	1637	60.1
Total	5396	100	2722	100

considered sample are enrolled in a single course the demographic analysis is focused on these students.

In Table 2 we present some demographic characteristics of the studied population. It is noticeable that both genders have roughly the same level of success, but women have about twice the enrollment of men. Regarding race, whites achieved the highest success rate at 70%, at the other extreme Blacks had the lowest success rate at 42%, while Hispanics were in around the middle at 59%, Asians had almost the same success as whites but their actual numbers were smaller, American Indians had a 50% success rate but also their actual participation is much lower. When considered by age, success rates remained somewhat constant at around 60–67%, interestingly age categories 21–24 and 25–30 constitute 45% of the total sample. When looking at student work activity two large groups exist students declaring not working at all and those working 31 or more hours; thus nearly half of the sample is composed of non-traditional mature students most likely working at least 31 hours. Furthermore, a significant number of these students enrolled at RSCC are first college generation. Finally at the bottom of Table 2 we provide success rates for students taking one course and more than one course. The success rate for those taking more than one course is lower at 48% compared to 63% for those taking one course. Students enrolled in one GEAR course represent 77% of the training sample, while 19% of the sample took two courses and the rest of students the enrolled in three classes or more. It appears that more than half of them are first time in college students and are more likely to be females. Close to half of those students appear to be working 31 or more hours per term. Thus statistical evidence in Table 2 shows that 37% of the students taking one GEAR course in the training sample are not being successful.

The overall non-success rate for this sample is slightly higher at 43% compared to the actual validation sample 39.9% (Table 1). Therefore, we have undertaken this project research so that struggling students and instructors can use Rio PACE indicators monitoring their academic progress to receive timely course feedback and targeted assistance to enhance their likelihood of success.

To assess actual success distribution across the GEAR courses in both samples two histograms are provided below. Figure 1 shows the graphical distribution for the training sample, while Fig. 2 presents the distribution for the validation data set. Despite the different number of observations in the training and validation samples, both data sets appear to have similar distributions. Interestingly, in the validation data despite the smaller sample size success counts by course follow the same behavior as in the training sample. In both histograms higher success is observed in ENG102 and PSY101, while low success rates are present in BIO101, CHEM130 and HIS103. The latter two courses, BIO101 and CHM130, differ in complexity with respect to the other courses and that could account for such low success rates.

It is recommended that both training and validation data sets share similar distributions, so that the estimated model obtains better prediction rates for meaningful out of sample

Table 2 Selected demographic attributes for GEAR enrollees, fall 2012

Demographics	Success indicator				
Gender	Non-success	Non-success (%)	Success	Success (%)	Total
Females	837	37	1420	63	2257
Males	455	37	782	63	1237
Unknown	15	38	25	63	40
Race					
Unknown	89	37	151	63	240
Hispanics	250	41	366	59	616
American Indian	26	50	26	50	52
Asian	28	31	61	69	89
Black	294	58	217	42	511
Hawaiian, Pacific Islander	4	40	6	60	10
White	606	30	1382	70	1988
Two or more races	10	36	18	64	28
Age					
LE 20	253	40	382	60	635
21–24	281	39	431	61	712
25–30	314	36	558	64	872
31–35	170	34	331	66	501
36–40	109	36	194	64	303
41–45	73	33	145	67	218
46–50	57	41	83	59	140
51–55	27	28	71	72	98
GE 56	23	42	32	58	55
Work hours					
Unknown	40	40	59	60	99
None	456	41	655	59	1111
1–10	42	31	94	69	136
11–15	39	38	63	62	102
16–20	98	41	140	59	238
21–30	117	35	216	65	333
31 or more	515	34	1000	66	1515
First in college					
N	531	35	1007	65	1538
Y	776	39	1220	61	1996
GEAR courses taken					
1 GEAR course	1307	37	2227	63	3534
2 or more GEAR courses	1257	52	1145	48	2402

validation. If that were not the case then one is likely to find that actual and predicted success rates might differ, thus leading to a possible rebuilding of the model. Brooks and Thompson (2017) advise building a model on data available from one year, then construct a testing set consisting of data from the following year, rather than dividing the data set from a single year into training and validation. The approach followed in this paper is

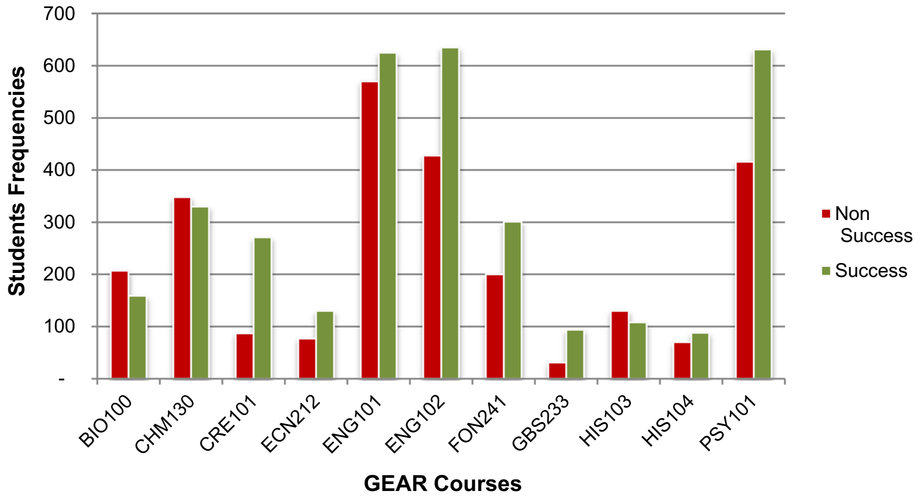

Fig. 1 PACE development sample, n = 5396

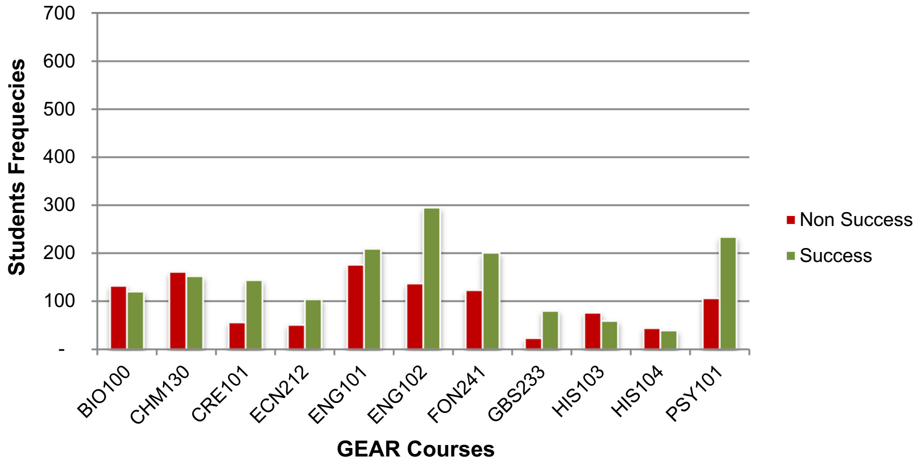

Fig. 2 PACE validation sample, n = 2722

consistent with their recommendation but applied to Fall and Spring terms. It is important to notice that success rates for the validation data set were not available at the time of the scoring of this sample. Final grades from the eleven courses were later appended to the validation data after being posted.

Next, we present two tables containing the set of sufficient statistics to be used by the scoring algorithm as specified in the methodology section. Tables 3 and 4 below provide these statistics, means and standard deviations, for the predictors included in the model. The class event, success/non-success is represented by a binary indicator for students achieving a grade of C or higher as successful or otherwise in the eleven courses listed in the table. Note that cell sizes by class event category and course are greater than thirty, thus adequate for model estimation and checking for normality of the predictors. These

Table 3 Mean statistics for GEAR courses by success indicator, fall 2012

Success indicator	Course	Ng	m_x1	m_x2	m_x3	m_x4	m_x5	m_x6
0	BIO100	207	23.86	19.68	8.13	6.61	1349.00	2812.80
0	CHM130	348	28.42	19.22	10.66	7.00	1013.22	1812.82
0	CRE101	87	15.89	12.23	3.99	2.80	137.14	427.64
0	ECN212	77	16.66	11.21	5.67	3.56	137.74	397.27
0	ENG101	570	25.38	19.52	8.78	6.37	233.78	676.09
0	ENG102	428	23.56	17.60	8.24	5.68	217.16	465.19
0	FON241	200	20.81	17.09	7.15	5.77	253.77	492.68
0	GBS233	31	21.39	15.26	7.64	5.26	173.68	535.81
0	HIS103	130	21.09	12.44	7.97	4.63	108.70	300.38
0	HIS104	70	15.39	8.81	5.16	2.91	67.77	257.14
0	PSY101	416	20.97	16.34	6.89	5.10	107.06	296.60
1	BIO100	159	54.77	46.75	26.13	21.98	4687.70	5838.36
1	CHM130	330	46.00	31.64	21.92	14.78	2415.44	2829.21
1	CRE101	271	49.15	42.94	23.25	20.18	899.31	969.24
1	ECN212	130	33.27	24.59	15.15	10.80	552.23	655.54
1	ENG101	625	55.18	46.66	27.04	22.88	889.42	997.46
1	ENG102	635	50.72	41.68	24.92	20.25	877.54	991.97
1	FON241	301	45.20	38.43	21.07	17.62	949.77	1081.06
1	GBS233	94	46.29	36.53	22.38	17.18	822.93	916.60
1	HIS103	108	41.98	28.27	20.65	13.61	422.81	499.54
1	HIS104	88	37.56	24.78	18.45	12.03	424.26	499.43
1	PSY101	631	47.58	38.79	22.58	18.09	425.45	495.73

statistical figures are derived from the algorithm in the estimation and used in the out of sample validation.

The corresponding variable definitions are given in Table 5 below. Given the model specification all the variables are defined as numeric easing their interpretation for analysis by administrators at the institution. Ideally as stated in the methodology section, we want these predictors to satisfy the normality assumption. Logins and site engagement with their weighted counterparts satisfied the mentioned condition, while points earned and points possible violated this assumption. That was somewhat expected as the latter two attributes are performance related, thus subject to more sample variability.

Also, the value ranges is wider for BIO100 and CHM130 undoubtedly related to subject matter complexity. The next sections focus on methodological aspects of the model, data interpretation and empirical findings.

4 Methodological Procedures

In developing the SQL algorithm for the continuous Naïve Bayesian model we followed research undertaken by Ordonez and Pitchaimalai (2010), and Pitchaimalai et al. (2010). The Naïve Bayesian model specification rests on the following assumptions: predicting attributes of success are independent and normally distributed. While the first condition,

Table 4 Standard deviation statistics for GEAR courses by success indicator, fall 2012

Success indicator	Course	Ng	s_x1	s_x2	s_x3	s_x4	s_x5	s_x6
0	BIO100	207	19.08	15.75	9.95	7.92	1454.56	1973.82
0	CHM130	348	17.15	12.24	9.29	6.51	685.36	917.78
0	CRE101	87	11.13	8.15	5.15	3.37	142.25	156.15
0	ECN212	77	11.45	7.86	6.30	4.10	144.04	163.14
0	ENG101	570	15.32	12.34	7.82	6.16	192.91	175.93
0	ENG102	428	15.22	11.78	7.68	5.62	180.37	215.10
0	FON241	200	16.92	13.99	8.82	7.13	238.56	280.82
0	GBS233	31	12.77	8.79	6.93	4.84	172.81	171.48
0	HIS103	130	15.09	9.18	7.91	4.71	122.19	139.84
0	HIS104	70	10.90	6.84	5.48	3.47	96.97	122.26
0	PSY101	416	14.30	11.01	6.98	5.22	102.90	107.04
1	BIO100	159	16.00	12.34	7.65	5.94	390.86	277.60
1	CHM130	330	14.64	10.48	7.29	5.20	259.12	207.42
1	CRE101	271	15.07	13.81	7.84	7.11	108.64	100.07
1	ECN212	130	11.71	8.24	6.04	3.96	47.31	17.40
1	ENG101	625	14.36	11.97	7.44	6.22	66.44	12.04
1	ENG102	635	14.78	12.22	7.60	6.17	77.60	32.48
1	FON241	301	15.00	12.27	7.56	6.26	248.10	224.27
1	GBS233	94	13.96	9.36	7.02	4.60	53.52	17.20
1	HIS103	108	11.63	8.38	6.14	4.19	34.20	4.81
1	HIS104	88	13.02	8.01	6.80	4.25	37.13	5.33
1	PSY101	631	14.00	11.63	7.15	5.80	34.81	11.56

Table 5 Variable definitions

Variable	Definition	Type
Success indicator	Achieving C or better	Binary
Course	Course Catalog Name	Descriptive
Ng	Count per Course and Class	Numeric
m_x1	Mean of Logins	Numeric
m_x2	Mean of Site Engagement	Numeric
m_x3	Mean Weighted Logins	Numeric
m_x4	Mean Weighted Site Eng.	Numeric
m_x5	Means Points Earned	Numeric
m_x6	Means Points Possible	Numeric
s_x1	Std. Deviations Logins	Numeric
s_x2	Std. Deviation Site Eng.	Numeric
s_x3	Std. Weighted Logins	Numeric
s_x4	Std. Weighted Site Eng.	Numeric
s_x5	Std. Weighted Points Earn.	Numeric
s_x6	Std. Deviations Points Poss.	Numeric

independence, is rarely satisfied Naïve Bayesian application results seems to be robust (Zhang 2004). Those assumptions facilitate the calculation of sufficient statistics necessary for model prediction and implementation at the course level. For alternative distribution specifications to predicting attributes see John and Langley (1995).

Let C_j represent an element belonging to the jth. class of the event of interest, i.e. success and non-success; h be the number of dimensions of a set of attributes given by X, i.e. number of site engagements; k be the number of GEAR courses students enrolled into; and n be the number of observations per each element of X, X_{ih}. Then for each class C_j, the continuous Naïve Bayesian basic statistics and probability density parameters require the following conditions:

$$L_{k\in j} = \sum_{x_i \in X_{k\in j}} x_i \tag{1}$$

Moreover, let $Q_{k\in j} = \sum_{x_i \in X_{k\in j}} x_i * x_i'$ be the cross product matrix. However, because of the independent assumption among the attribute elements in X we focus only on the diagonal elements of $Q_{k\in j}$.

These calculations apply to each element of attributes X per class $C_{k\in j}$. Furthermore, for each X_d belonging to the class event $C_{k\in j}$ corresponds a number of observations $N_{k\in j}$. Therefore one can obtain the Gaussian sample parameter estimates given as:

$$M_{k\in j} = \frac{L_{k\in j}}{N_{k\in j}}, \quad \text{and} \tag{2}$$

$$V_{k\in j} = \frac{Q_{k\in j}}{N_{k\in j}} - \frac{L_{k\in j}}{N^2_{k\in j}} * L'_{k\in j} \tag{3}$$

Both expressions for $M_{k\in j}$ and $V_{k\in j}$ are statistical representations of μ_{kh} and σ_{kh} per each dimensional class j for course k.

Once these statistics are computed subsequently for scoring Gaussian conditional probabilities and prior probabilities are derived at each data point in the data set X for each class event j. The set of prior probability values is given by $\pi(C_{k\in j}) = \frac{N_{k\in j}}{n}$ for each class event j in course k. Furthermore, the conditional Gaussian probabilities to compute final posterior probabilities can be expressed as:

$$P\left(X_{k\in i,h}\right) = \frac{1}{\sqrt{2\pi\sigma^2_{k\in j,h}}} * \exp\left\{-.5\left(X_{k\in i,h} - \mu_{k\in j,h}\right)^2 / \sigma^2_{k\in j,h}\right\} \tag{4}$$

The joint probability of each X element h is expressed as $\pi\left(X_{k\in i,h}|j\right) = \Pi_h P\left(X_{k\in i,h}|j\right)$, where $X_{k\in i,h}$ represents the h-dimensional value for X_i in each course k. To score both development and validation data sets, then optimum class C_j is determined by the following maximum probability expression:

$$P(k\in j|X_{k\in i}) = max_{k\in j}\pi_{k\in j}P(X_{k\in i,h}|k\in j). \tag{5}$$

Thus, these mathematical expressions were combined into a decision algorithm written in SQL in the Microsoft Server Management Studio 2012 for the selected GEAR courses. Both training and validation score codes were implemented using data for fall 2012 for the estimation of the model and for out of sample validation using data for spring 2013, respectively. In the next section, we will report on empirical results from applying this Naïve Bayesian specification to both training and validation data.

5 Empirical Results

In our research we wanted to predict student success for high enrollment courses participating in the GEAR program at the course level. The model was estimated using the following predictors: number of logins, weighted number of logins, site engagement activities, weighted site engagement, points earned, and points possible. Model results under the continuous Naïve Bayesian were encouraging as the classifier achieved higher non-success and successful identification compared with earlier work with a discrete version of this model specification focusing on one course (Smith et al. 2012).

Two algorithms were created one for the development of the model using the training data set, which generated the set of sufficient statistics in Tables 3 and 4. The second algorithm used those coefficients to predict probability of success for the validation sample. This step will provide a sense for how well the model is likely to perform in a production environment. Therefore, the discussion now centers on presenting our empirical findings for both training and validation samples.

The total sample for training encompassed 5936 students enrolled in GEAR courses during the fall of 2012. Nonsuccess was 43.2%, while success stood at 56.8%. The total validation sample amounted to 2722 students for the same courses during the spring of 2013. The corresponding nonsuccess/success rates were 39.86 and 60.14%, respectively. These figures were not known at the time the actual validation took place, but became available after the data warehouse was updated reflecting spring 2013 final grade results.

To evaluate the model fit, Table 6 below shows a cross tabulation of actual and estimated outcomes for students in the GEAR courses based on the training sample. Table 6 depicts how the NB model has performed comparing the actual event with the estimated outcome. For those students whose actual event was a success, 3290 were correctly classified while only 82 were identified as false negatives. That will represent 97.6 and 2.4% of students where the actual outcome is positive, respectively. For those failing the class, 2251 were correctly classified while 313 were classified as false positives. The latter represent 12.2% of the students whose actual outcome was non-success, while the former totaled 87.8% of the students failing a course. The overall classification rate of the NB was 93.3%. These results are expected since this is the training sample data set.

Since the intent of the model is to make predictions for the eleven GEAR courses we provide a graphical representation of actual versus estimated success rates for the training sample below at the course level (Fig. 3). As expected in this modeling stage estimated

Table 6 PACE GEAR courses confusion matrix, fall 2012

Success	Outcome		
	Fail	Pass	Total
0	2251	313	2564
	87.8%	12.2%	
1	82	3290	3372
	2.4%	97.6%	100%
Total	2333	3603	5936

rates are very close to actual rates in all courses except for BIO100, CHEM130, and CHEM130LL showing wider gaps in estimated rates.

Most courses show less than 10% difference between actual and predicted estimated success rates. Only those three courses show larger differences with CHM130 experiencing the most difference in excess of 10%. One possible reason the model is not performing as well for those three courses is that the points scale is very different from the other courses and the complexity of those courses. Moreover, both point variables did not meet the normality assumption.

Gains chart are frequently used in the financial industry to evaluate a model's ability to identify individuals more likely to respond to marketing offers. The larger the area between the two curves the better the model's classification. The 45 degree line usually represents random targeting of customers, while the curved line represents the additional customers to be gained if more selective targeting arising from a predictive model is initiated (Jaffery and Liu 2009).

In our research this gains chart, Fig. 4, identifies the unsuccessful students by sample tile that could be targeted for instructor-led interventions. We rank ordered students by probability of non-success and segmented them into 10 tiles for both training and validation results. Students in lower tile numbers are likely at higher risk of non-success than students in high number tiles. Figure 4 shows that at the 4th tile 89% of non-success students in the training sample are identified.

To assess how well the model would perform in a production environment the out of sample validation results for all eleven GEAR courses are shown in Table 7 below. Note that the success indicator was unavailable at the time predictions were made. It was appended to the data once final grades were uploaded into the data warehouse. This has allowed us to gauge the model's ability to predict success and non-success in the courses of interest.

The main findings for the validation sample were the following: 91.9% of the students were classified as true negative (997), i.e. unsuccessful, while 8.1% were predicted as false positives, successful but in fact not doing well (88); likewise, 1547 were identified as true positives, i.e. successful, while only 90 were misclassified as false negatives—successful students but misclassified as un-successful; percentage wise figures translate to 94.5 and 5.5%, accordingly. Thus, the overall prediction was 93.5%.

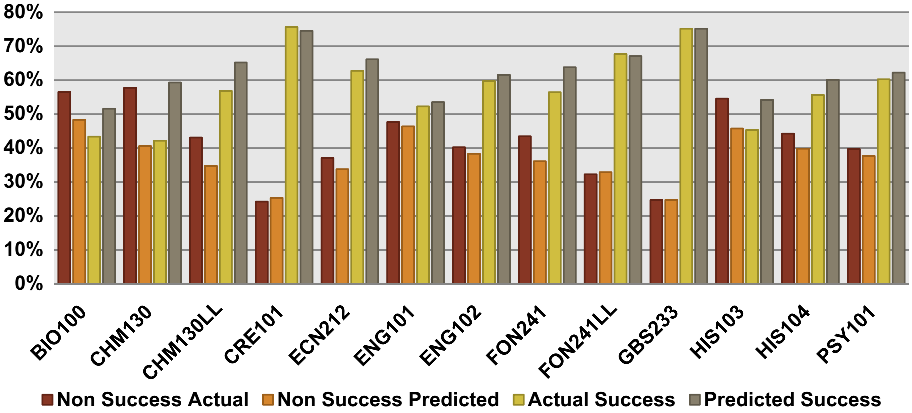

Fig. 3 PACE GEAR courses actual versus predicted success rates for training sample

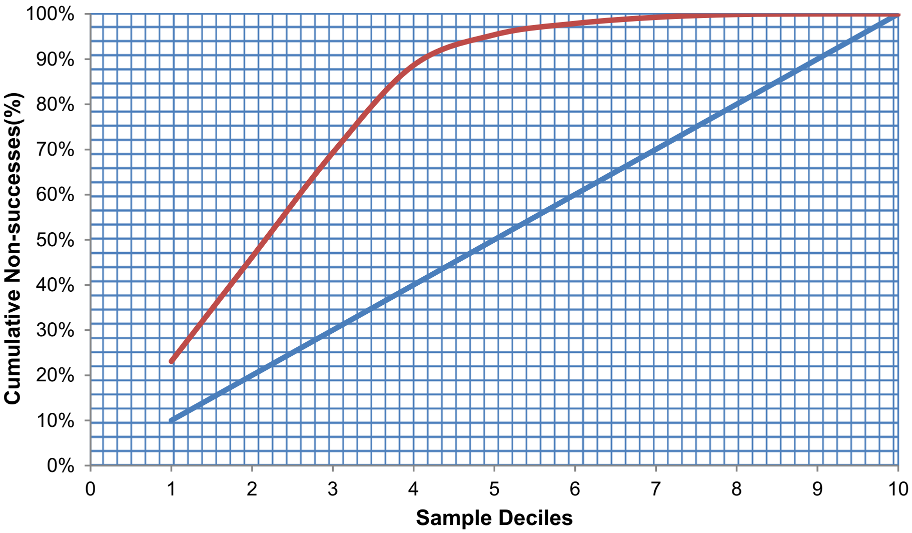

Fig. 4 PACE cumulative gains chart model training sample

Table 7 PACE GEAR courses confusion matrix, spring 2013

Success	Outcome		
	Fail	Pass	Total
0	997	88	1085
	91.9%	8.1%	
1	90	1547	1637
	5.5%	94.5%	100%
Total	1087	1635	2722

Next we present a graphical representation (Fig. 5) demonstrating how well the model predicted students' performance at the course level for the validation sample. As we can observe, the model predicts both non-success and success reasonably well across courses with slightly higher differences for CHM130, HIS103 for non-successes. Interestingly, for these courses the validation sample predictions are actually better than estimated results in the training sample.

The initial number distribution of the actual success/non-success charts provided in the data section is a useful tool to gauge both training and validation sample similarities, particularly when predictions need to be made (Brooks and Thompson 2017). This is an important feature that could ensure robust out of sample model performance. As seen in Fig. 6 below, the cumulative gains chart for the validation sample provided were derived based on the model using the predicting attributes from Table 3. Maximum lift and proper classification for true negatives occurs at the 4th tile where over 90% of the non-successful students are correctly classified.

The empirical interpretation arising from the chart is that proper identification of at-risk students can lead to better allocation of programmatic assistance resources such as: tutoring, advising and peer mentoring to high risk students. Thus, the institution can time

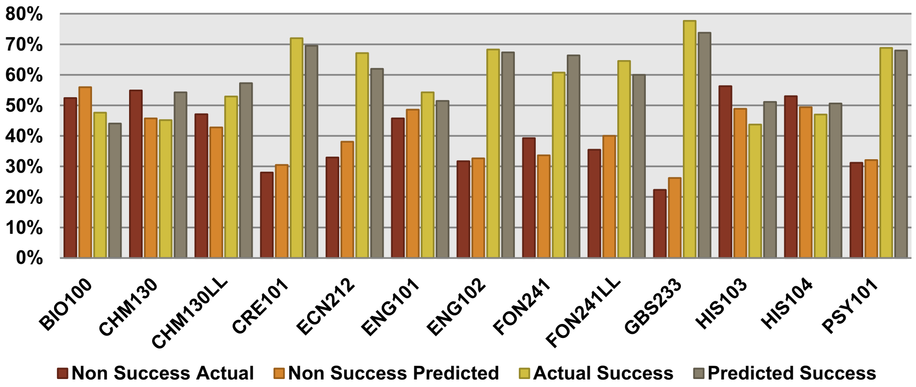

Fig. 5 PACE GEAR courses actual versus predicted success rates for validation sample

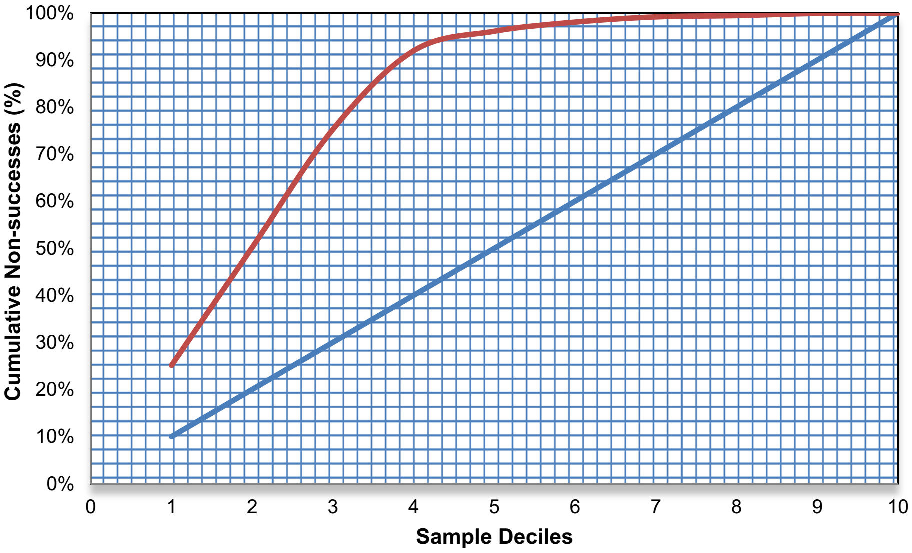

Fig. 6 PACE cumulative gains chart model validation sample

its resources more adequately to increase student success in the eleven courses in the GEAR program.

The overall performance of the model for the validation sample stands at 93.5%. Precision and accuracy rates are frequently used to determine the classification quality of binary classifiers (Vuk and Curk 2006); thus by these measures the NB classifier achieved accuracy and precision values of 93.3 and 91.3% in the training sample, while in the validation sample those figures were 93.5 and 94.6% respectively. It is worth mentioning that early estimation of the discrete model (Smith et al. 2012) predicted nonsuccess rates at 70% for one course, while our findings are in line with those of Barber and Sharkey (2012); however the latter analysis was applied to a different student population.

We also estimated a logistic regression model on the training data set. In this exploratory analysis, the dependent variable was an overall success indicator. Initial model diagnostics left us with three variables: weighted site engagement, points earned and points possible. The signs of these variables were as expected. For instance, the more engaged students were the more likely they were to be successful. Likewise, the higher the number of points earned the higher the chances students had of passing a course. Points possible had a negative sign suggesting that as points increase the success likelihood may decrease. Since this was a predictive model on the overall success indicator, we did not pursue further investigation on this because we were primarily interested in predicting success at the course level. To pursue this approach would require estimation of separate models for each course. However, the continuous Naïve Bayesian under the independence assumption facilitates estimation, prediction, and implementation across courses with a single model specification; thus we opted for this model over logistic regression.

6 Conclusions and Limitations of the Study

The intent of this empirical research was to develop and implement a continuous Naïve Bayesian model to predict student success for high enrollment courses under the GEAR program at Rio Salado Community College. This risk model is a key component of an alert system to enhance student success at the course level. Training and validation findings suggest that the model achieves high classification of non-successes and successes cases. Compared to the early version of the discrete NB classifier, the continuous model obtains higher rates of student classification and better prediction accuracy to a larger number of courses. This performance could be attributed to both better model specification and data measurement. The early model prediction non-success rate was 70% in a single course (Smith et al. 2012), while this new model predicts 91.8% on eleven courses in the GEAR program. The results of this research while not completely similar to those reported in the literature review are consistent with their findings.

The gains charts capturing at-risk students' distribution allow us to conclude that the model identifies success and non-success properly. Students belonging into 1–4 tiles are primary candidates for possible early intervention from instructors, advisors, and peer mentors. Targeting this subpopulation is likely to improve success rates possibly leading to higher persistence and completion rates.

For the training sample, the largest differences in predicting success were observed for BIO100, CHM130, and CHM130LL. One explanation for this is that the point scale values are different in these classes and material complexity. Surprisingly, for the validation sample, results were actually better. Further development will be required to rollout the model to other courses with shorter duration. Also early identification of at-risk students might require model modifications for predicting success within a shorter time window so that RSCC can program and target assistance resources accordingly. This is one limitation of the study as students in an online environment are more likely to drop early in the course. Predictions in the model are based on full term length but in exchange discussions on these empirical findings, faculty has expressed concerns that by then it could be too late. Recent empirical research (Shelton et al. 2016, 2017; Ifenthaler and Widanapathirana 2014) suggests that predicting early in the courses would result in timely interventions while addressing student retention. Thus, some modeling work is currently in progress to address these challenges. Finally, the results of this project could be relevant to other community colleges practitioners expanding into online learning and having the means to capture the needed data in their respective LMS to build similar early warning systems to assist at-risk students.

References

Allen, I. E., Seaman, J., Poullin, R., & Taylor, S. T. (2016). Tracking online education in the United States. *Online Consortium.* http://onlinelearningconsortium.org/read/online-report-card-tracking-online-education-united-states-2015/.

Barber, R., & Sharkey, M. (2012). Course correction: Using analytics to predict course success. In *LAK12: 2nd international conference on learning analytics & knowledge*, Vancouver, BC, Canada.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Department of Education.

Brooks, C., & Thompson, C. (2017). Predictive modelling in teaching and learning. In C. Lang, A. Wise, & D. Gasevic (Eds.), *Handbook of learning analytics* (pp. 61–68). doi:10.18608/hla17.005.

Buntine, W. (1992). Learning classification trees. *Statistics and Computing, 2,* 63–73.

ECAR-ANALYTICS Working Group. (2015). *The predictive learning analytics revolution: Leveraging learning data for student success.* ECAR working group paper. Louisville, CO: ECAR.

Eduventures. (2013). *Predictive analytics in higher education data-driven decision-making for the student life cycle.*

Elkan, C. (2014). *Maximum likelihood, logistic regression and stochastic gradient training.*

Hung, J. L., & Zhang, K. (2008). Revealing online learning behaviors and active patterns and making predictions with data mining techniques in online teaching. *MERLOT Journal of Online Learning and Teaching*, *4*(4), 426-437.

Ifenthaler, D., & Widanapathirana, C. (2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning., 19,* 221–240. doi:10.1007/s10578-014-9926-4.

Jaffery, T., & Liu, S. X. (2009). *Measuring campaign performance by using cumulative gains and lift charts.* Paper 196-2009, SAS Global Forum.

John, H. G., & Langley P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence*. San Mateo: Morgan Kauffman Publishers.

Liu, Y. S., Gomez, J., & Yen, C. (2009). Community college online course retention and final grade: Predictability of social presence. *Journal of Interactive Online Learning*, *8*(2), 165–182.

Lumina Foundation. (2015). *Strategic plan for 2017 to 2020.* http://www.luminafoundation.org/resources/lumina-foundation-strategic-plan-for-2017-to-2020.

Macfadyen, P. L., & Dawson, S. (2010). Mining data to develop an "Early Warning System" for educators: A proof of concept. *Computers and Education*, 54, 588–599. www.elsevier.com/locate/compedu.

Oliff, P., Palacios, V., Johnson, I., & Leachman, M. (2013). Recent deep state higher education cuts may harm students and the economy for years to come. *Center on Budget and Policy Priorities,* 1–21.

Ordonez, C., & Pitchaimalai, S. (2010). Bayesian classifiers programmed in SQL. *IEEE Transactions on Knowledge and Data Engineering (TKDE), 22*(1), 139–144.

Pitchaimalai, S. K., Ordonez, C., & Alvarado, C. G. (2010). *Comparing SQL and map reduce to compute Naïve Bayes in a single table scan.* doi:10.1145/1871929.1871932.

President Obama 2013 State of the Union Address. (2013).

Rampell, C. (2013). *Data reveal a rise in college degrees among Americans.* The New York Times.

Rio Salado College Assessment of Student Learning. (2013). *Annual report.*

Shelton, B. E., Hung, J., & Baughman, S. (2016). Online graduate teacher education: Establishing an EKG for student success intervention. *Technology, Knowledge and Learning, 21,* 21–32. doi:10.1007/s10758-015-9254-8.

Shelton, B. E., Hung, J., & Lowenthal, P. R. (2017). Predicting student success by modeling student interaction in asynchronous online courses. *Distance Education, 38*(1), 59–69. doi:10.1080/01587919.2017.1299562.

Smith, V. S., Lange A., & Huston, D. R. (2012). Predictive modeling to forecast student outcomes and effective interventions in online community college courses. *Journal of Asynchronous Learning Networks, 16*(3), 51-61.

Vuk, M., & Curk, T. (2006). ROC curve, lift chart and calibration plot. *Metodolozkisveski, 3*(1), 89–108.

Zhang, H. (2004). The optimality of Naïve Bayes. *American Association for Artificial Intelligence.* www.aaai.org.