RIO SALADO COLLEGE

Predicting Student Success

An Application to Community College Data

Fermin Ornelas, Ph.D., Carlos Ordonez Prof., Daniel Huston, M.Ed.

3/22/2014

This research focuses on developing and implementing the Naïve Bayesian Classifier to GEAR courses at Rio Salado Community College. It demonstrates that this predictive model has good prediction accuracy of at-risk students. Predictive results across courses and cumulative gain charts show potential improvements to be made in students' academic success by focusing at high level risk students.

Research paper submitted for presentation to Association for Institutional Research Conference in Orlando FL, May 2014, www.airweb.org.

Predicting Student Success: An Application to Community College Data
By
Fermin Ornelas, Carlos Ordonez, Daniel Huston

Introduction

Academic institutions today face several challenges driven by cost concerns, increasing accountability, and diminishing resources. For instance, a recent report by the Center on Budgets and Policy Priorities (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; eleven states cut educational funding by more than a third; thirty six 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 have been stagnant for at least twenty years; however according to a recent New York Times report (2013) that trend has improved in the last five years. Furthermore, there is an ongoing federal effort to increase graduation rates while preventing college cost increases. Therefore, to address the low graduation rates problem institutions have responded with additional online academic options to entice students to continue their education. 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. 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.

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. Based on student attributes the classifier generates 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 (V. Smith, et. al., 2012). This approach makes modeling and scoring more memory intensive 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 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.

Literature Research

To put this project research into context, the author 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 besides those variables, it also included 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. They concluded that adding these new variables increased the predicted accuracy of the second model.

V. Smith, Lange, and Huston (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, they 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, Gomez and Yen (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 fitted two logistic regression models. One model was developed to predict course retention success and another ordinal model to predict final grade. The survey data was collected after the third week into the semester. In the first model they concluded that there is a positive relation of 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 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. Some of the considered variables were: 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, 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. Among the empirical findings discussed in the paper 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 from72.57 to 88. Overall, Hung and Zhang found that when students were more actively engaged tended to perform academically better.

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 fitted 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 validates student peer engagement as part of the learning process for student success.

In the following section we discuss data set construction steps for the development of the continuous NB classifier and some sample characteristics.

Data Construction

To build the continuous naïve 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 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 for this NB classifier 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, a full description is given in the appendix. Rio Salado has 48 weekly start dates with classes beginning every Monday and with different course lengths. Going forward an adjustment in the field calculation and segmentation would require further research to reflect such dynamic process. Moreover because of course updates often responding to labor market needs, there is an additional field course master ID which would require additional data capturing for updated courses. Thus for a complete rollout of the NB classifier specification the data would also need to be adjusted.

For this phase 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 for model development while the spring data was kept for out of sample validation. Actual success frequencies for both development and validation are provided in table 1 below.

Table 1. Frequencies for Development and Validation Data Sets

	Development		**Validation**	
Success Indicator	**Frequency**	**Percent**	**Frequency**	**Percent**
0	2,564	43.19	1,085	39.86
1	3,372	56.81	1,637	60.14
Total	5,936	100	2,722	100

Success is defined as achieving a C or higher in a course. Success rates are 56.8% and 60.1% for development and validation, 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 sample considered enrolled in a single course the demographic analysis is focused on these students.

Both genders have about the same rate of success; however women have about doubled enrollment than 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 where 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 27% of the total sample. When looking at student work activity two large groups exist, those that are likely full time students and those working 31 or more hours. Furthermore a significant number of them students enrolled at RSCC are first college generation. Finally at the bottom of the table 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.

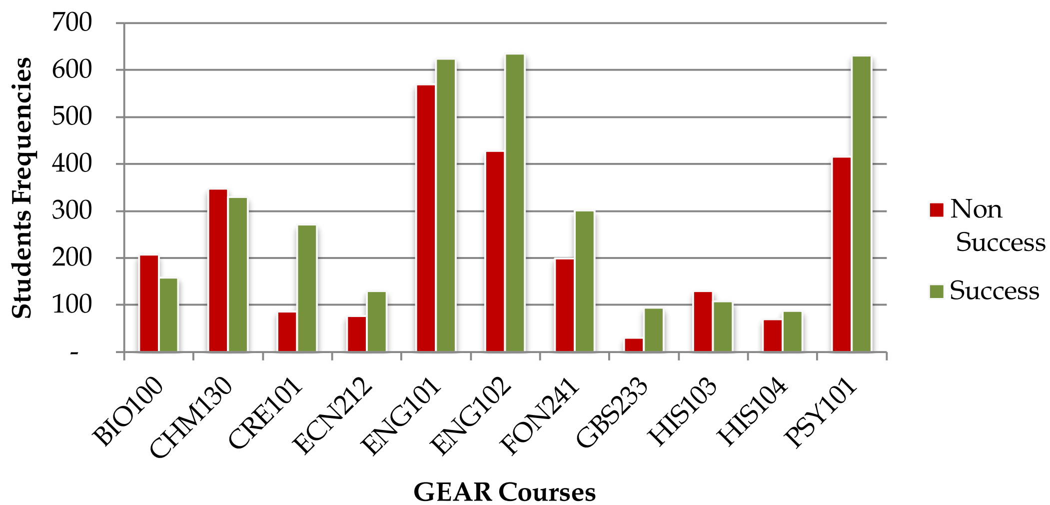
Table 2. Selected Attributes for GEAR Enrollees in One Course, Fall 2012

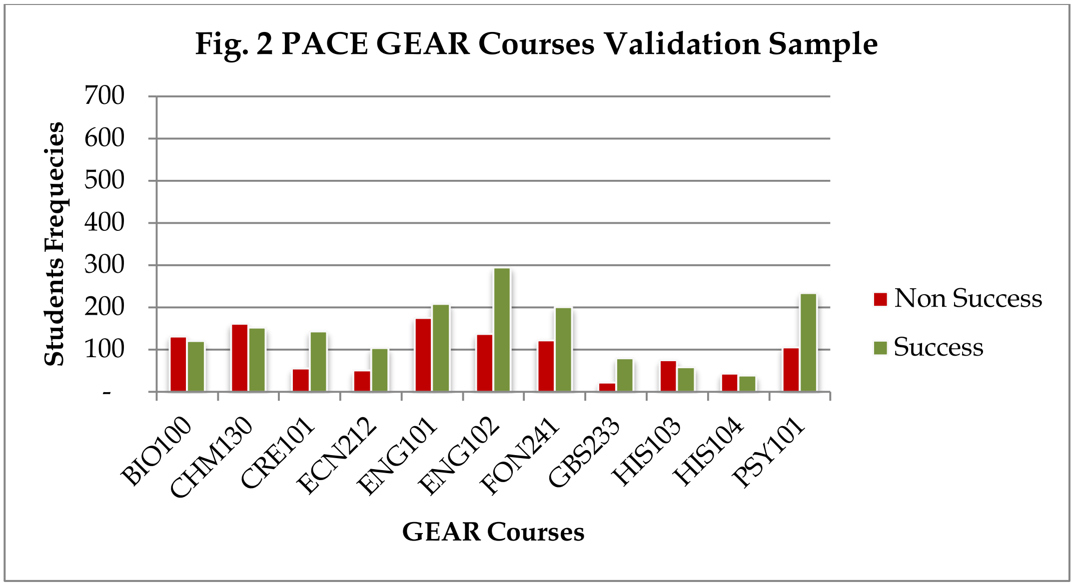
Demographics	Success Indicator				
Gender	Non-Success	Non-Success (%)	Success	Success (%)	Total
Females	837	37%	1420	63%	2257
Males	455	37%	782	63%	1237
Unknown	15	38%	25	63%	40
Race					
Unknown	89	37%	151	63%	240
Hispanics	250	41%	366	59%	616
American Indian	26	50%	26	50%	52
Asian	28	31%	61	69%	89
Black.	294	58%	217	42%	511
Hawaiian, Pacific I.	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
56 - +	23	42%	32	58%	55
Work Hrs.					
n.a.	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 +	515	34%	1000	66%	1515
First in College Generation					
N	531	35%	1007	65%	1538
Y	776	39%	1220	61%	1996
GEAR Courses					
Total 1 GEAR Course	1307	37%	2227	63%	3534
Total 2 or more Courses	1257	52%	1145	48%	2402

Students enrolled in one GEAR course represent 77% of the development sample, while 19% of the sample took two courses and the rest of students enrolled in three classes or more. It appears that more than half of them are first time in college students more likely to be females. Close to half of those students appear to be working 31 or more hours per term. Given these facts one should not be surprised that 37% of the students taking one GEAR course are not being successful. The overall non-success rate is slightly higher at 43% (table 1). This is the main reason we have undertaken this research project so that struggling students can get timely targeted assistance to address their poor performance and subsequently achieve a higher success rate for the classes considered.

To assess the actual success distribution across the GEAR courses two graphs are provided below. Interestingly, despite the smaller sample size in the validation, 2,722, the trend in success counts follow the same behavior as in the development sample. In both charts higher success is observed in ENG102 and PSY101, while low success rates are present in BIO101, CHEM01 and HIS103. These facts are relevant in the estimation as they facilitate better prediction results in the statistical model.

Fig. 1 PACE GEAR Courses Development Sample





In the appendix, table 1 provides basic statistics for the variables included in the model. These variables were checked for normality: logins and site engagement with their weighted counterparts satisfied the mentioned condition, while points earned and points possible did not. That was expected as the latter two attributes are performance related. Also, the value ranges is wider for BIO100 and CHM101. The next sections focus on methodological aspects of the model, data interpretation and empirical findings.

Methodological Procedure

In developing the SQL algorithm for the continuous Naïve Bayesian model we followed research undertaken by S. K. Pitchamalai, C. Ordonez, and C. Garcia-Alvarado (2010) as well as by C. Ordonez and S.K. Pitchamalai (2010). The model specification rests on the following assumptions: predicting attributes of success are independent and normally distributed. While the first condition, independence, is rarely satisfied naïve Bayesian application results seems to be robust, see H. Zhang (2004).

Let C_j represent an element belonging to the j^{th} class of the event of interest; h be the number of dimensions of a set of attributes given by X; 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:

$$L_{k\in j} = \sum_{x_i \in X_{k\in j}} x_i \qquad (1)$$

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:

$$M_{k\in j} = \frac{L_{k\in j}}{N_{k\in j}} \text{ , and} \qquad (2)$$

$$V_{k\in j} = \frac{Q_{k\in j}}{N_{k\in j}} - \frac{L_{k\in j}}{N_{k\in j}^2} * L_{k\in j}' \qquad (3)$$

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. Furthermore, the conditional Gaussian probabilities to compute final posterior probabilities can be expressed as:

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

The joint probability of each X element h is expressed as $\pi\left(X_{k\in i,h}\middle|j\right) = \Pi_h P\left(X_{k\in i,h}\middle|j\right)$, where $X_{k\epsilon i,h}$ 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

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

Thus, this methodology required a written SQL code in Microsoft Server Management Studio 2012 for the selected GEAR courses. Both development and validation score codes were implemented using data for fall 2012 for the development phase and for out of sample validation data for spring 2013. In the next section, we will report on empirical results from applying the naïve Bayesian to RSCC validation data.

Empirical Results

In our research we wanted to predict student success for high enrollment GEAR courses. Model estimation under the continuous naïve Bayesian results was encouraging as the classifier appeared to correctly identify non-success and successful students properly.

The total sample for development encompassed 5,936 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 2,722 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 estimation took place, but became available after the data warehouse was updated to reflect spring 2013 grade results.

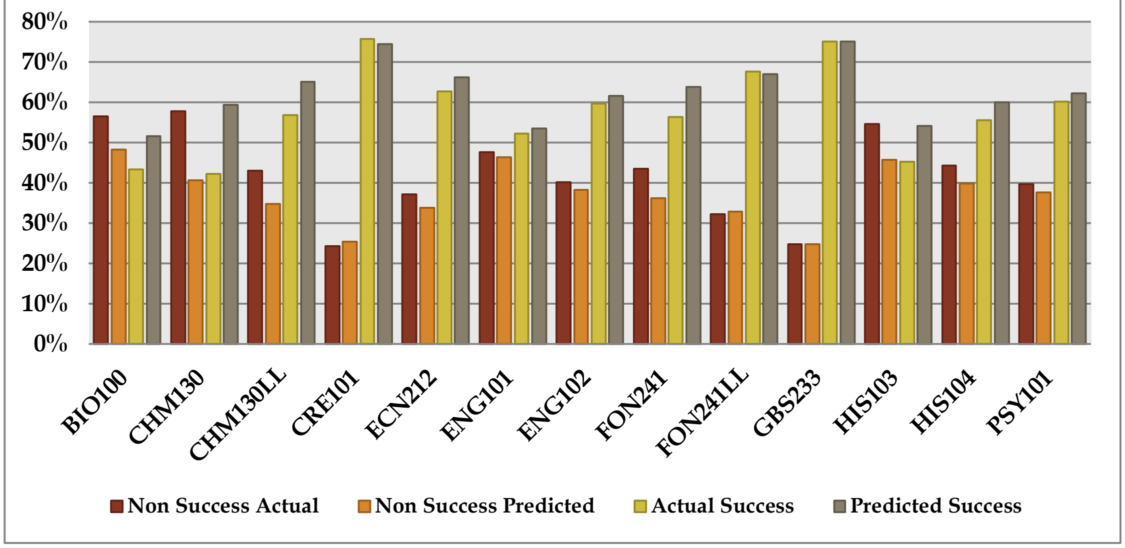
To evaluate the model fit, table 3 below shows a cross tabulation of actual and predicted outcomes for students in the GEAR courses based on the development sample. This table gives a good idea on how the NB model has performed compared to the actual event. For those students whose actual event was a success, 3,290 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 good, 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 prediction of the classifier was 93.3%.

Table 3. NB, GEAR Courses Confusion Matrix

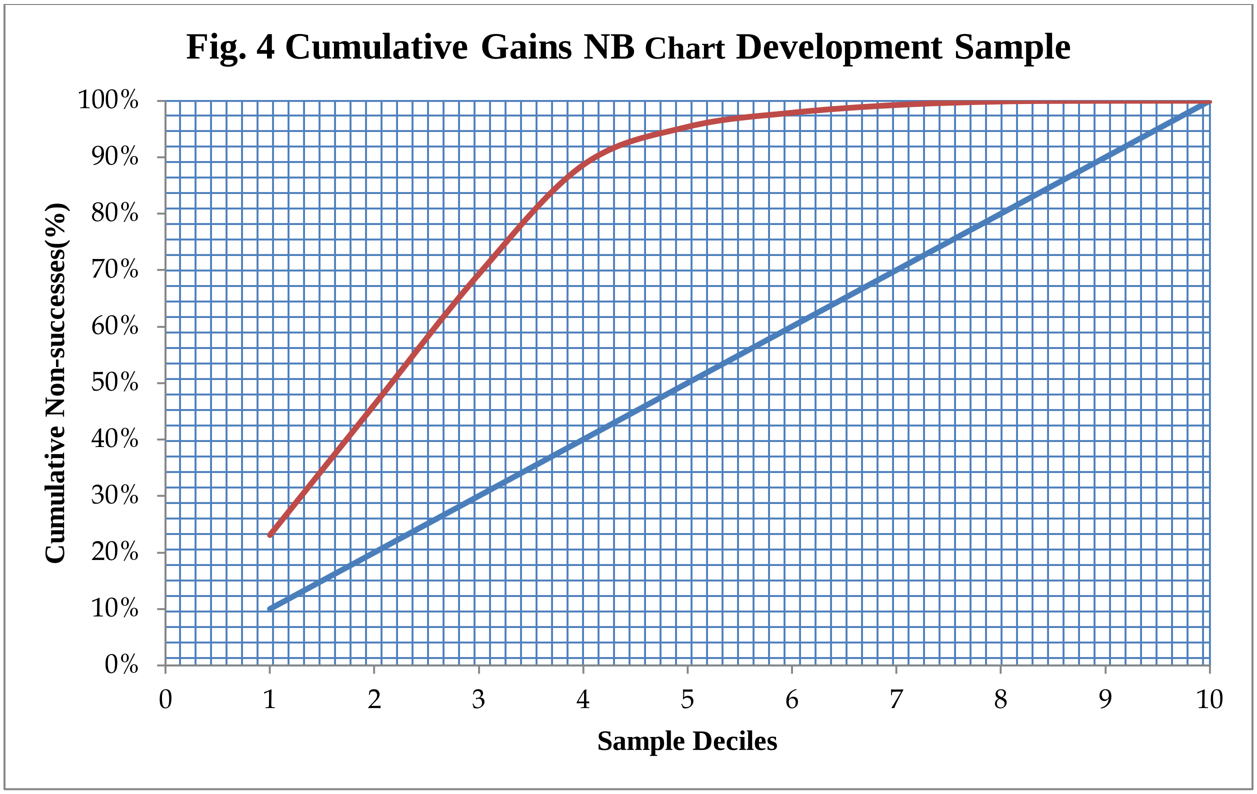
Table of Success by Outcome Development, Fall 2012			
Success	**Outcome**		
	Fail	**Pass**	**Total**
0	2251	313	2564
	87.8%	12.2%	
1	82	3290	3372
	2.4%	97.6%	
Total	2333	3603	5936

Since the intent of the model is to make predictions for the eleven GEAR courses we provide a graphical representation of actual versus predicted success rates for the development sample below. As expected in this modeling stage prediction rates are very good in all courses except for BIO100, CHEM130, and CHEM130LL showing less-stellar prediction rates. Most courses show less than 10% difference between actual and predicted 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 predicting as well for those three courses is that the points scale is very different from the other courses. Moreover, both point variables did not meet the normality assumption.

Fig. 3 PACE GEAR Courses Actual vs. Predicted Success Rates Development Sample



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 represent 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 (T. Jaffrey et. al., 2009). In our research this gains chart 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 development and validation prediction results. Students in lower tile number are likely at higher risk of non-success than students in high number tile. In the development sample maximum lift occurs at tile 4 where the model captures close to 90% of the unsuccessful students. Thus, selective intervention efforts focused on student in tiles 1 through 4 may render fruitful results by increasing student success rates in that section of the sample.



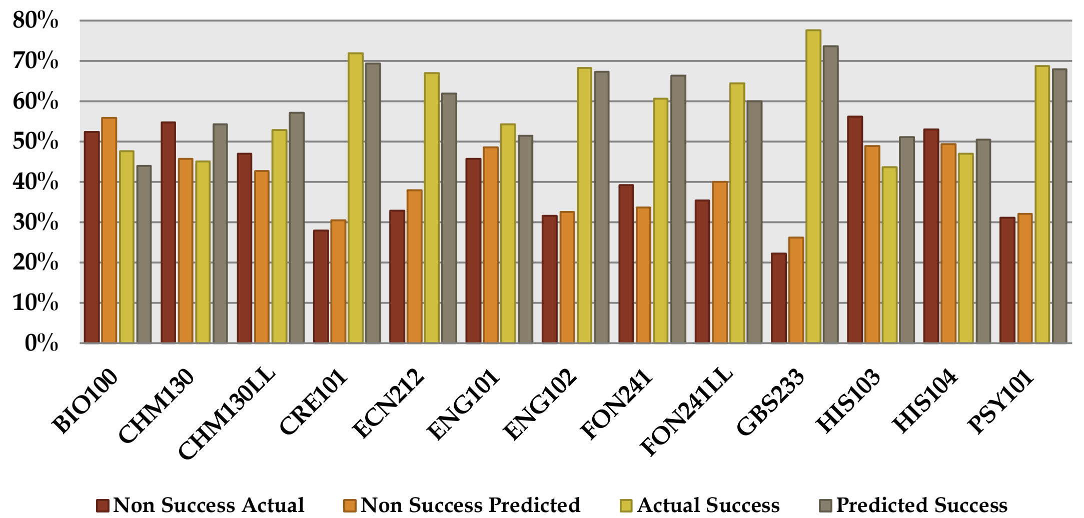
To assess how well the model would perform in a production environment the out of sample validation results for all the GEAR course are shown in table 4 below.

Table 4. NB, GEAR Courses Confusion Matrix

Table of Success by Outcome Validation, Spring 2013			
Success	**Outcome**		
	Fail	**Pass**	**Total**
0	997	88	1085
	91.9%	8.1%	
1	90	1547	1637
	5.5%	94.5%	
Total	1087	1635	2722

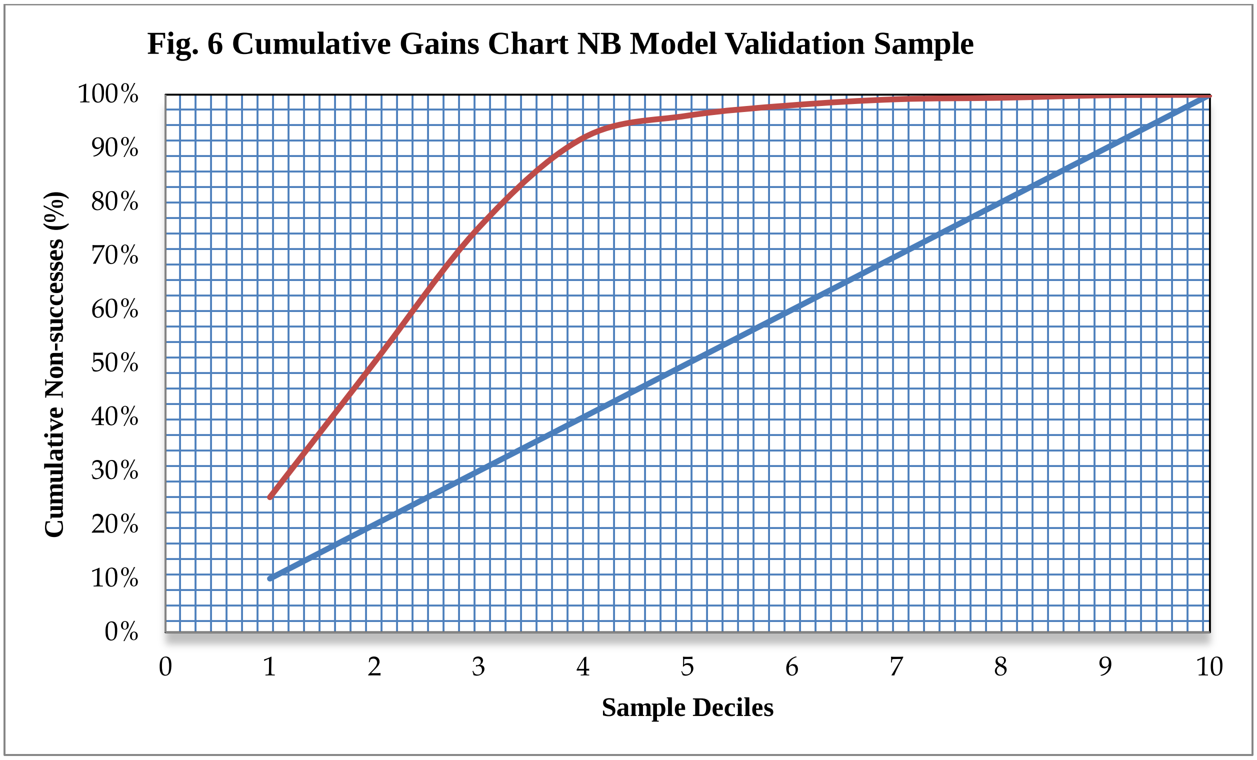
The main findings for the validation sample were the following: 91.9% of the students were classified as true negative (997), i.e. unsuccessful, while 8.1% were predicted as false positives (88); likewise, 1,547 were identified as true positives, i.e. successful, while only 90 were misclassified as false negatives, percentage wise figures translate to 94.5%, and 5.5%, accordingly. Thus, the overall prediction was 93.5%. Next we present a graph demonstrating how well the model predicted students' performance at the course level.

Fig. 5 PACE GEAR Courses Actual and Predicted Success Rates Validation Sample



Comparison of actual versus predicted rates became possible once grading results were available in the data warehouse. 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, further details are provided in the appendix. Interestingly, for these courses the validation sample predictions are actually better than in the development sample.

The initial number distribution of the actual success/non-success charts provided in the data section is a useful tool to gauge both development and validation sample similarity. This is an important feature that could ensure good out of sample model performance. This fact is supported by the cumulative gains chart in the validation sample provided in figure 6 below. Here the maximum lift also 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.

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 development sample, while in validation those figures were 93.5% and 94.6% respectively. It is worth mentioning that early estimation of the discrete model by Smith, et. al., (2012) predicted nonsuccess rates at 70%, 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 fitted a logistic regression model to the data set. This exploratory analysis was conducted on the development sample only. The dependent variable was the 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 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 positive 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.

Conclusions and Recommendations

The intent of this empirical research was to develop and implement a continuous Naïve Bayesian model to predict student success for selected GEAR classes at Rio Salado Community College. Development and validation results findings suggest that the model correctly classifies non-successes and successes cases. Compared to the early version of the discrete NB classifier, the continuous model achieves higher rates of student classification and better prediction accuracy. As reported by Smith et. al. (2012) the early model prediction non-success rate was 70%, while this new model predicts 91.8%. 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 early intervention from instructors, advisors, and peer mentors. Targeting this subpopulation is likely to improve success rates and ensure more students continue their college education.

For the development 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. Surprisingly, for the validation sample, results were actually better. We may need additional research on these courses to improve predictive accuracy.

Further development requires full rollout of the model to other courses with shorter duration. Also early identification of at-risk students might require model and modifications for predicting success within a shorter time window so that RSCC can program and target assistance resources accordingly. Some modeling work is currently on progress. Finally, the results of this project could be relevant to other community colleges expanding onto online learning and having the means to capture the needed data in their LMS to build similar early warning systems for at-risk students.

References

Rebecaa Barber and M. Sharkey. Course Correction: Using Analytics to Predict Course Success. LAK12: 2nd International Conference on Learning Analytics & Knowledge, 29 April – 2 May 2012, Vancouver, BC, Canada.

Jui-Long Hung and Ke Zhang. Revealing Online Learning Behaviors and Active Patterns and Making Predictions with Data Mining Techniques in Online Teaching. MERLOT Journal of Online Learning and Teaching, Vol. 4, No. 4, December 2008.

Simon Y. Liu, J. Gomez and C. Yen. Community College Online Course Retention and Final Grade: Predictability of Social Presence. Journal of Interactive Online Learning, Vol. 8, No. 2, Summer 2009.

Leah P. Macfadyen and S. Dawson. Mining Data to Develop an "Early Warning System" for Educators: A Proof of Concept. Computers and Education, 54, 588-599, 2010. www.elsevier.com/locate/compedu.

Phil Oliff, V. Palacios, I. Johnson, and M. Leachman. Recent Deep State Higher Education Cuts May Harm Students and the Economy for Years to Come. Center on Budget and Policy Priorities, March 2013.

C. Ordonez and S. Pitchamalai. Bayesian Classifiers Programmed in SQL. IEEE Transactions on Knowledge and Data Engineering (TKDE), 22(1):139-144, 2010.

C. Rampell. Data Reveal A Rise in College Degrees Among Americans. The New York Times, June 12, 2013.
Miha Vuk and T. Curk. ROC Curve, Lift Chart and Calibration Plot. Metodolozkisveski, Vol 3, No. 1, 2006, 89-108.

Tariq Jaffery and S. X. Liu. Measuring Campaign Performance by Using Cumulative Gains and Lift Charts. Paper 196-2009, SAS Global Forum 2009.

S. K. Pitchaimalai, C. Ordonez, and C.G. Alvarado. Comparing DQL and Map Reduce to Compute Naïve Bayes in A Single Table Scan.
http://doi.acm.org/10.1145/1871929.1871932

Vernon S. Smith, A. Lange, and D. R. Huston. Predictive Modeling to Forecast Student Outcomes and Effective Interventions in Online Community College Courses. Journal of Asynchronous Learning Networks. Vol. 16, Issue 3. June 2012.

Zhang Harry. The Optimality of Naïve Bayes. American Association for Artificial Intelligence (www.aaai.org), 2004.

Appendix

Table 1. Descriptive 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	s_x1	s_x2	s_x3	s_x4	s_x5	s_x6
0	BIO100	207	23.86	19.68	8.13	6.61	1349.00	2812.80	19.08	15.75	9.95	7.92	1454.56	1973.82
0	CHM130	348	28.42	19.22	10.66	7.00	1013.22	1812.82	17.15	12.24	9.29	6.51	685.36	917.78
0	CRE101	87	15.89	12.23	3.99	2.80	137.14	427.64	11.13	8.15	5.15	3.37	142.25	156.15
0	ECN212	77	16.66	11.21	5.67	3.56	137.74	397.27	11.45	7.86	6.30	4.10	144.04	163.14
0	ENG101	570	25.38	19.52	8.78	6.37	233.78	676.09	15.32	12.34	7.82	6.16	192.91	175.93
0	ENG102	428	23.56	17.60	8.24	5.68	217.16	465.19	15.22	11.78	7.68	5.62	180.37	215.10
0	FON241	200	20.81	17.09	7.15	5.77	253.77	492.68	16.92	13.99	8.82	7.13	238.56	280.82
0	GBS233	31	21.39	15.26	7.64	5.26	173.68	535.81	12.77	8.79	6.93	4.84	172.81	171.48
0	HIS103	130	21.09	12.44	7.97	4.63	108.70	300.38	15.09	9.18	7.91	4.71	122.19	139.84
0	HIS104	70	15.39	8.81	5.16	2.91	67.77	257.14	10.90	6.84	5.48	3.47	96.97	122.26
0	PSY101	416	20.97	16.34	6.89	5.10	107.06	296.60	14.30	11.01	6.98	5.22	102.90	107.04
1	BIO100	159	54.77	46.75	26.13	21.98	4687.70	5838.36	16.00	12.34	7.65	5.94	390.86	277.60
1	CHM130	330	46.00	31.64	21.92	14.78	2415.44	2829.21	14.64	10.48	7.29	5.20	259.12	207.42
1	CRE101	271	49.15	42.94	23.25	20.18	899.31	969.24	15.07	13.81	7.84	7.11	108.64	100.07
1	ECN212	130	33.27	24.59	15.15	10.80	552.23	655.54	11.71	8.24	6.04	3.96	47.31	17.40
1	ENG101	625	55.18	46.66	27.04	22.88	889.42	997.46	14.36	11.97	7.44	6.22	66.44	12.04
1	ENG102	635	50.72	41.68	24.92	20.25	877.54	991.97	14.78	12.22	7.60	6.17	77.60	32.48
1	FON241	301	45.20	38.43	21.07	17.62	949.77	1081.06	15.00	12.27	7.56	6.26	248.10	224.27
1	GBS233	94	46.29	36.53	22.38	17.18	822.93	916.60	13.96	9.36	7.02	4.60	53.52	17.20
1	HIS103	108	41.98	28.27	20.65	13.61	422.81	499.54	11.63	8.38	6.14	4.19	34.20	4.81
1	HIS104	88	37.56	24.78	18.45	12.03	424.26	499.43	13.02	8.01	6.80	4.25	37.13	5.33
1	PSY101	631	47.58	38.79	22.58	18.09	425.45	495.73	14.00	11.63	7.15	5.80	34.81	11.56

Table 2. Variable Definitions

Variable	Definition	Type
Success Indicator	Achieving a C or better	Binary
Course	Course Catalog Name	Descriptive
Ng	Cell 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. Deviation Logins	Numeric
s_x2	Std. Deviation Site Eng.	Numeric
s_x3	Std. Deviation W. Logins	Numeric
s_x4	Std. Deviation W. Site Eng.	Numeric
s_x5	Std. Deviation Points Earn.	Numeric
s_x6	Std. Deviation Points Poss.	Numeric

Table 3. GEAR Course Definitions

Course	Description	Credits
BIO100	Biology	4
CHM130	Chemistry	3
CHM130LL	Chemistry Lab.	1
CRE101	Critical Reading	3
ECN212	Economics	3
ENG101	English	3
ENG102	English	3
FON241	Food and Nutrition	3
FON241LL	Food and Nutrition Lab	1
GBS233	General Business Systems	3
HIS103	History	3
HIS104	History	3
PSY101	Psychology	3

Table 4. Development Sample Actual and Predicted Success

Courses	Non Success				Success				Total
	Actual	Actual (%)	Predicted	Predicted (%)	Actual	Actual (%)	Predicted	Predicted (%)	
BIO100	207	57%	177	48%	159	43%	189	51.6%	366
CHM130	219	58%	154	41%	160	42%	225	59.4%	379
CHM130LL	129	43%	104	35%	170	57%	195	65.2%	299
CRE101	87	24%	91	25%	271	76%	267	74.6%	358
ECN212	77	37%	70	34%	130	63%	137	66.2%	207
ENG101	570	48%	555	46%	625	52%	640	53.6%	1195
ENG102	428	40%	408	38%	635	60%	655	61.6%	1063
FON241	148	44%	123	36%	192	56%	217	63.8%	340
FON241LL	52	32%	53	33%	109	68%	108	67.1%	161
GBS233	31	25%	31	25%	94	75%	94	75.2%	125
HIS103	130	55%	109	46%	108	45%	129	54.2%	238
HIS104	70	44%	63	40%	88	56%	95	60.1%	158
PSY101	416	40%	395	38%	631	60%	652	62.3%	1047
Total	2,564	43%	2,333	39%	3,372	57%	3,603	60.7%	5936

Note: success is defined as achieving at least a C in a course.

Table 5. Validation Sample Actual and Predicted Success

Courses	Non Success				Success				
	Actual	**Actual (%)**	**Predicted**	**Predicted (%)**	**Actual**	**Actual (%)**	**Predicted**	**Predicted (%)**	Total
BIO100	132	52%	141	56%	120	48%	111	44%	252
CHM130	96	55%	80	46%	79	45%	95	54%	175
CHM130LL	65	47%	59	43%	73	53%	79	57%	138
CRE101	56	28%	61	31%	144	72%	139	70%	200
ECN212	51	33%	59	38%	104	67%	96	62%	155
ENG101	176	46%	187	49%	209	54%	198	51%	385
ENG102	137	32%	141	33%	295	68%	291	67%	432
FON241	84	39%	72	34%	130	61%	142	66%	214
FON241LL	39	35%	44	40%	71	65%	66	60%	110
GBS233	23	22%	27	26%	80	78%	76	74%	103
HIS103	76	56%	66	49%	59	44%	69	51%	135
HIS104	44	53%	41	49%	39	47%	42	51%	83
PSY101	106	31%	109	32%	234	69%	231	68%	340
Total	1,085	40%	1,087	40%	1,637	60%	1,635	60%	2,722

Note: success is defined as achieving at least a C in a course.