Modern Predictive Models for Modeling the College Graduation Rates

Emma A. Gunu
Office of Institutional Research
Central Michigan University, USA
gunu1ea@cmich.edu

Wilson K. Gyasi
Department of Mathematics,
Central Michigan University, USA
gyasi1wk@cmich.edu

Carl Lee
Department of Mathematics
Central Michigan University, USA
Carl.lee@cmich.edu

Robert M. Roe
Executive Director
Office of Institutional Research
Central Michigan University, USA
roe1rm@cmich.edu

Abstract:
Modern predictive modeling techniques are commonly used for modeling a target of interest based on a list of input variables. In general, these techniques are capable of identifying input variables associated with the target, but not for the purpose of identifying the causation relationship between target and inputs due to the fact that the data are observational data. Advanced technology has made data collection very easy and fast. As a result, when applying predictive modeling methods, the issue of data cleansing becomes critical. This article aims at comparing ten modern predictive modeling techniques for predicting college graduation rate within6years. The input variables include variables on 'pre-college' performance, 'first-year' college performance and various social-economic variables, as well as some variables related to university learning environment. The issue of data quality and modeling technique selection are discussed. Some pitfalls and cautions of applying predictive modeling techniques are discussed.

Keywords: Decision Tree, Ensemble, Gradient Boosting, LASSO, LARS, Logistic Regression, Neural Network, Random Forest, Support Vector Machine

I. INTRODUCTION

Modern predictive modeling techniques are commonly used for modeling a target of interest based on a list of input variables. The process of building a predictive model often involves steps starting from data collection, data integration, and data cleansing prior to variable selection and possible variable transformation to prepare data for modeling. The complexity of data preparation is becoming more complicated and difficult due to the fact that advanced technology has made data collection very easy and fast. Most of the data used for model building are observational data, and come from a variety of different sources. Each part of the data was originally collected for a specific purpose. Thus, the validation of data quality and integrity is critical for a successful predictive modeling project.

Data preparation often takes 60% to 80% of the project time because a quality data is essential for any modeling project. Once an appropriate data set is prepared, one can begin exploring the data to manipulate and fine-tune the data including variable selection, variable transformation, and missing data handling. In addition, using the insight and domain knowledge of the project is very important in order to identify the list of inputs for modeling.

Model building is an iterative process. Good knowledge about different modeling techniques, especially understanding the pros and cons of each is essential for selecting and finalizing the best model to predict the target. The process often involves building the optimal model based on a selected criterion for each modeling technique using training data for model building and validation data for optimizing the model. Once an optimal model is selected for each modeling technique, a testing data is applied to compare these 'best' models based on a selected criterion to choose the final 'best' model.

In general, predictive modeling techniques are capable of identifying input variables associated with the target, but not for the purpose of identifying the causation relationship between target and inputs due to the fact that the data are observational data. During the process of tackling the project, it is important to keep in mind the goal of the project throughout the process and take into consideration the context of the project in order to obtain a 'useful' model. The usefulness of a model relies upon solid data quality, valid modeling methodology, and a close connection with the purpose of the project.

The models obtained may be parametric, such as logistic regression and LASSO model; or maybe nonparametric, such as Decision Tree and Random Forest models. Therefore, interpretation of a model differs depending on if the model is parametric or non-parametric. For a parametric model, one needs to be careful to interpret the estimates in terms of association relationship with the target, not the causation relationship. For non-parametric model, since there are no parameter estimates, some techniques, such as Random Forest or Decision Tree, provide the degree of importance of inputs, which are useful for interpreting the association of inputs with the target. Some nonparametric models do not provide any information regarding the importance of the inputs. A common approach is to fit the predicted target using parametric techniques such as LARS, or a Decision Tree to identify the importance of input variables. Regardless of what techniques are applied, a good understanding of the context of the problem and solid understanding of the methodology is essential, in order to give a meaningful and appropriate interpretation of the relationship between target and selected inputs.

978-1-5090-5756-6/17/$31.00 ©2017 IEEE
SERA 2017, June 7-9, 2017, London, UK

II. THE CHALLENGE OF DATA QUALITY

Data quality could be an issue for any form of data, including quantitative data, text data and image data, as well as web data. The problem is becoming more complicated and difficult due to the easiness of data generation using modern technology. The issues of data quality may occur during data production and data manipulation. The process of data production may be further classified into data source identification, collection, extraction and integration. This process often takes over 75% of the time for a project [1]. Wang and Strong [12] defined data quality as "fitness for use". They identified four categories in fifteen data quality dimension. McKnight [2] summarized seven sources of poor data quality during this process, which includes (1) entry quality, (2) process quality, (3) identification quality, (4) integration quality, (5) usage quality, (6) aging quality and (7) organizational quality. Radhakrishna et al. [3] considered the components of data quality and provided a check list for checking eight components of data quality as follows (1) validity (2) reliability (3) objectivity (4) integrity (5) generalizability (6) completeness (7) relevance and (8) utility. The National Institute of Statistical Sciences (NISS) [13] defined three principles of data quality: (1) data are a product, with customers, to whom they have both cost and value; (2) data, as a product, have quality, resulting from the process by which data are generated; and (3) data quality depends on multiple factors, including (at least) the purpose for which the data are used, the user, the time, etc. Cai and Zhu [14] discussed the challenges of data quality in the big data era and proposed a five dimension of data quality criteria and an assessment metric for evaluating each criterion. The five dimensions are (1) availability (2) usability (3) reliability (4) relevance and (5) presentation quality.

Once data production is complete and data quality is evaluated, prior to model building, there is another process of data exploratory analysis and manipulation. The purpose of exploratory and graphical analysis is to identify possible erroneous cases and potential outliers, and investigate causes of missing data, and problems may occur due to different measurement units, as well as practically irrelevant inputs from domain knowledge point of view. At the data manipulation stage, solutions are decided and applied to handle error data, outliers, missing data and unify the measurement units, investigate relationship between target and each individual inputs and conduct a preliminary variable selection and transformation. The objective is to select inputs that are statistically and practically meaningful, and relevant to the target. The preliminary variable selection criteria often are set to be very conservative to ensure possible relevant inputs are not overlooked.

For situations where there are different data sources and different measurement units from a large amount of data, missing data is often a major concern when preparing the data. For most modeling techniques, missing one data value of an input means losing the entire case. If missing data are randomly occurring among variables and cases, even though the total number of cases may be large, if deleted, it is possible that there are only very few cases or worst, no valid case remains for model building. Most softwares implement default methods for imputing missing data. However, these default methods may be misleading or better methods can be implemented based on the context of the input variables. For example, missing of prices of a product should be handled differently from missing score of an exam. Some types of missing cannot be imputed, such as gender; while some missing should be treated as zero, such as missing a test. Some missing is due to the sensitivity of the question asked. Thus, individuals who chose not to answer such questions should be grouped and indexed so that one can investigate the insight of this group compared with others. Prices of a product, even though it is missing in a certain time period due to no sales does not mean there is no prices for the same product in the market. Thus, time series interpolation provides a good imputation for the missing prices.

III. CHOICE OF MODELING TECHNIQUES

Many different modeling techniques are available. Each technique has its strength and weakness. In this article, we apply the following ten modeling techniques to model the college graduation rates. In particular, we are interested in investigating inputs that are highly associated with the graduation at the 6^{th} year after entering a university: (1) Decision Tree (DT), (2) Logistic Regression(LR), (3) Least Angle Regression (LAR), (4) Least Absolute Shrinkage and Selection Operator (LASSO), (5) Gradient Boosting (GB), (6) Neural Network (NN), (7) Partial Least Square (PLS), (8) Random Forest (RF), (9) Support Vector Machine (SVM), and (10) Ensemble of DT, LR, LAR and NN. Some of these modeling techniques have been applied to many real world projects. For example, Mizuno et al. [27] studied fault-prone module prediction. Nakagawa et al. [28] applied SVM method to study the defect rates of electronic board.

A general strategy of predictive model building involves four steps: (1) data partitioning, (2) model building, (3) model optimization and (4) model comparison. The data partition step partitions the entire data into Training data set for building models, Validation data set for optimizing the final model, and Test data for model comparison. The model building step is aimed at selecting important variables for predicting the target based on a predefined criterion. For interval target, a common criterion is the average squared error (ASE). The model optimization step is to select the optimal model from the list of models built in the model building step by applying each model to the Validation data set based on a given assessment criterion such as AIC, BIC, and validation error. The model comparison step is to apply each "optimal" model obtained from each given modeling technique using a selected criterion to select the final "best" model.

A brief summary of each modeling technique is described in the following. For details about the modern predictive modeling, one may refer to the literatures [4-11].

- DT: This is a rule-based modeling technique. It is easy to apply and interpret, it allows missing data (that is, no imputation is needed), and automatically takes the interaction among inputs into account. The major weakness is that it discretizes interval target, thus, the prediction is no longer an interval scale. In addition, the DT is sensitive to the choices of inputs in order. It will give totally different rules when a different input is selected in

the early stage of the splitting step. Using the idea of variance and bias trade-off point of view, DT models often result in large variance of the predictions.

- LR: This is a parametric modeling technique, thus, easy to interpret. It is additive and hence, the parameter estimates can be interpreted as the pure contribution of the input to the target. One can perform inference on the parameter estimates. The logistic regression with two levels of target is $\ln((p(Y=1)/(1-p(Y=1))) = X\beta + \varepsilon.$ Some weaknesses are: (1) it is only good for categorical target, especially for binary target, (2) it is only good for capturing the general pattern of linear relationship between inputs and target, and (3) there is a tendency of obtaining many superficial significant inputs due to the nature of hypothesis testing when sample size is large. Validation data and model optimization step help to reduce the risk of overfitting.

- LAR: This method requires k steps to select k predictors and build the model. It is considered to be one of the most efficient methods for building models. It is parametric, thus easy to interpret and allows for inference. The weakness is that it is only good for capturing linear relationship between target and inputs.

- LASSO: The model coefficients $\beta = (\beta_1, \beta_2, \cdots \beta_p)$ are the solution to the constrained optimization problem: Minimize $\|Y - X\beta\|^2$ subject to $\sum_{i=1}^{p} |\beta_i| \le t$, where Y is the target, X are standardized. The method implicitly serves as a model selection process and the result is a regression model with estimated parameters that can be interpreted. However, it assumes that the full model, includes all input variables, as the 'true model' could be troublesome. The typical confidence interval and hypothesis tests do not apply. The final selected model may be overfitting.

- GB: This method is a boosting approach. The GB approach applies a series of decision trees by fitting the residual of the prediction from the earlier tree in the series, which are combined by a series of weights [5]. Unlike single DT, this method combines a series of DTs. The final tree is not as sensitive to the order of the input selected. For prediction purposes, it has lower ASE than a single DT. It requires the users to determine the number of trees to be added, the weight to be used to adjust the subsequent tree and how large each tree should be. Generally speaking, the weight for each tree is taken to be small (.05 or smaller), the smaller the weight, the more the trees should be built; while the size of each tree often is chosen to be of depth one or two. The weakness is that it requires decision on the number of trees, weight and tree depth, which often take many attempts to obtain the final model. In addition, it may suffer from the possibility of overfitting.

- NN: The target $Y_{i.}$ is modeled as the function of the linear combination of hidden layers $H_{i.}$ defined as $Y_{i.} = f(W' H_{i.}') + \varepsilon$, where f is the activation function connecting hidden layers to the target. A hidden layer H is a linear combination of the inputs: $H_{i.} = g(Z' X_{i.}'),\ i = 1,2,\ldots,N$, where g is the activation function and Z is the weight matrix of the inputs. The hidden layer activation function g is taken to be hyperbolic tangent function. The NN modeling technique often results in too many inputs and the estimated weights for selected inputs cannot be interpreted, known as 'Black Box'.

- PLS: This method applies the following modeling strategy:
  1) The input variables $\{X_1, X_2, \cdots X_m\}$ are first transformed and combined into a smaller number of orthogonal 'principal components'. Each principal component is a linear combination of the inputs, $H_i = \sum_{j=1}^{m} w_{ij} X_j$, i =1,2,$\cdots k$ with $k << m$, which explains a proportion of the variation of the target variable. The first component explains the highest proportion of the dispersion. Each principal component can be considered as a latent input variable, which may be interpreted empirically based on the estimated weights of each original input.
  2) Using the 'latent' variables as the inputs to build the best model for predicting the target and replacing H_i back to $X's$, we have the final model that involves all input variables, $y = \sum_{i=1}^{k} a_i H_i = \sum_{i=1}^{k} a_i \sum_{j=1}^{m} w_{ij} X_j$
  Since the number of 'latent variables' selected in the 'best' model is usually dependent on the decision of the amount of variance explained.

- RF: This is an improved bagging method. It runs a large number trees, K, based on the resampling data of size n. For each tree, a randomly chosen subset of inputs m out of p inputs are applied to build the tree. The final model is the average of the predicted target for interval target, and majority voting of predicted values (or average predicted probabilities) for categorical target. The strategy of random forest results in low bias and reduces variance of predictions.

- SVM: The SVM method can be used for categorical and interval target. The idea is described for the binary classification problem, which involves finding a parametric linear or nonlinear function that describes a hyperplane that separates two sets of points in R^m. The hyperplane can be linear or nonlinear, depending on the kernel specification.

- Ensemble of DT, LR, LAR and NN: Each predictive method has its own strength and weakness. LR captures general linear pattern. DT captures the interaction. LAR captures local effects. NN is a network system that often takes into account many inputs that are not selected by other method. Ensemble approach is a technique combing the results of different models. This method results in a better prediction, however, the inputs cannot be properly identified in a model form. Overfitting is a major concern.

The data for analytics projects are usually observational, and the models are for prediction purposes. Hence, one should be cautious since the relationship between target and inputs are associations.

IV. A CASE STUDY: WHAT FACTORS ARE ASSOCIATED WITH COLLEGE GRADUATION AT THE 6TH YEAR?

For the remainder of this article, the case study on investigating the inputs that are highly associated with the college graduation at the 6th year will be described and the process of data preparation and model building will be illustrated. The target is binary and defined as Graduated within 6th year since enrollment: Yes/ No. The input factors are classified into (1) pre-college academic performance, (2) social/economic status, (3) college living environment and (4) academic performance of the first year upon entering the university. Details on data collection, preparation and exploration is given in the next Section.

College graduation rate has been a concern for most universities in higher education in the USA. It is an important factor used for ranking universities in USA [15] and the world ranking universities [16]. Many studies have been published to investigate factors associated with graduation rates in higher education. The American College Testing report on college retention and graduation rates from 1991 to 2012 indicated that the percentage of student graduation for all institutions within five years decreased from 54.4% in 1991 to 51.9% in 2012. Findings of important factors associated with graduation and retention rates often include race, income, pre-college preparation and cumulative GPA [18, 19, 21]. The first-semester GPA was identified in [17] as an important factor for predicting graduation rates of underrepresented students. Rogulkin [20] found that student cumulative GPA in first and 2nd year is a significant predictor for graduation at the 6th year. High school GPA was also found to be an important predictor by some studies [22]. Yet, the financial factor was found to be a predictor for graduation within 6th year [24].

Research on reasons and framework for explaining retention at higher education can be traced back to early 1960's. A commonly cited theory is the Interactionist theory proposed by Tinto [23]. The theory indicated that students with higher level of academic and social integration were believed to persist in college and graduate. Tinto's model identified various important relationship between the pre-college characteristics, such as family background, formal schooling; and integration of social and academics among students and interaction with faculty. Up-to-today, Tinto's theory seems to explain reasons of retention and graduation rates adequately in general.

Among many literatures studying the retention and graduation rates, majority used a limited data sources using a specific modeling technique, e.g., Logistic regression is one of the most common techniques for modeling binary target, and multiple regression are most commonly applied to model interval targets. Recent development of modern data mining techniques provides broader and in-depth advanced modeling techniques for studying graduation and retention rates. These techniques have also been applied to model graduation rates and to address higher education related projects. The applications of data mining techniques for education projects can be traced back to early 1990's [25]. Since then, various applications of data mining techniques to study a variety of higher education performance metrics have been studied. Perhaps the most rigorous application of these modeling techniques was by Raju and Schumacker [26]. They studied students' graduation at the 6th year since enrollment in a large research university using the data from 1995 to 2005. The modeling techniques used in their study are LR, DT, and NN as described in Section III. The significant factors found by their study were High School GPA, Full-Part Time Status, Gender, First Semester GPA and Residency Status. It is interesting to note that our finding in this study suggests that some of these factors are also found to be important; while there are some differences in our findings.

V. DATA PREPARATION AND EXPLOROTORY ANALYSIS

The data used for the study are from different databases compiled by the Office of Institutional Research of a research II university. The records are from 2006 to 2010 for a total of 19288 cases and 60 variables. The university for this study is located in a campus town. Over 99.5% of students are full time. Any identifiable information about individual student are deleted and replaced by an ID number. Variables that are not associated with the graduation status are dropped. Various variables are combined to create new inputs. After preliminary screening, twenty-two inputs are selected for modeling purposes. These variables are classified into four categories as described in Table 1.

TABLE 1. THE CHARACTERISTICS OF INPUT

Factor type	Factor names (Factor Labels)
Pre-college Academic Performance metrics	High school GPA, ACT-Composite, ACT-Math, ACT-English, ACT-Reading, ACT-Science, ACT-Writing
Social/Economic Background	Gender, Age, Minority (Yes/No), Eligible for Pell Financial aids (Yes/No), 1st-Generation (Yes/NO)
Institution living environment	Years-after High School prior to University, Full-time (Yes/No), Residential (Yes/NO), Persist to 2nd term (Yes/No), Persist-to-2nd year (Yes/No)
First-year academic performance metrics	1st-term GPA, 1st-year GPA, 1st-year Cumulative Credits Attempted, 1st-year Cumulative Credits Passed, 1st-year Cumulative Credits received grades.

The target variable is a binary target defined by whether a student graduated within six years from their first semester enrolment (GRAD6). The levels are Yes or No. Among the total of 19288 students who enrolled as part or full-time, 11132 (57.71%) graduated within 6 years.

Some of the input variables have missing values. 2.4% of student took SAT, instead of ACT. The corresponding ACT scores are imputed using the decision tree imputation method by imputing each of the missing ACT scores using

the other input variables. For example, the missing ACT Composite is imputed using all other 21 inputs based on the decision tree imputation method. Table 2 summarizes the frequencies of categorical inputs and Table 3 summarizes the descriptive statistics of interval inputs for each group of target, GRAD6. The Chi-square test for testing the association between GRAD6 and each individual categorical input indicates that each individual input has a significant association with the target. Thus, these variables provide useful information for modeling. The importance of these inputs will be selected and determined by each modeling technique. The summary statistics of interval inputs in Table 3 do not include the imputed missing data values. The interval inputs are all students' academic performance related inputs, either from pre-college or during the 1st year at college.

TABLE 2: TWO-WAY CROSS TABULATION SUMMARY OF CATEGORICAL INPUTS VS. TARGET

Grad6	Persist to 2nd term		Persist to 2nd Year		Low Income Status		Minority		1st Generation Status		Gender	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	F	M
No	1705 (20.9)	6451 (79.1)	4253 (52.2)	3903 (47.9)	6678 (81.9)	1478 (18.1)	7221 (88.5)	935 (11.5)	5767 (70.7)	2389 (29.3)	4141 (50.8)	4015 (49.2)
Yes	74 (0.7)	11058 (99.3)	196 (1.8)	10936 (98.2)	9888 (88.8)	1244 (11.2)	10307 (92.6)	825 (7.4)	8568 (77.0)	2564 (23.0)	6703 (60.2)	4429 (39.8)
p-value*	< .0001		< 0.0001		< 0.0001		< 0.0001		< 0.0001		< 0.0001	

*: The p-value is from the Chi-Square test

TABLE 3: DESCRIPTIVE SUMMARY OF INTERNVAL INPUTS BY TARGET

Grad6		HSGPA	ACT Composite	ACT Reading	ACT Math	ACT English	ACT Science	1st Year GPA	1st Year Cum. Attempted Credits	1st Year Cum. Graded Credits	1st Year Cum. Passed Credits
No	n	8013	7924	7908	7907	7906	7909	8156	8156	8156	8156
	Mean	3.144	21.655	22.157	20.995	20.841	22.018	2.164	26.430	23.080	21.299
	s.d.	0.405	3.215	4.622	3.909	4.062	3.352	1.016	8.006	7.508	10.318
Yes	n	10980	10965	10952	10952	10951	10953	11132	11132	11132	11132
	Mean	3.398	22.614	23.023	22.073	22.051	22.675	3.092	32.145	28.324	30.969
	s.d.	0.406	3.468	4.805	4.140	4.379	3.479	0.569	6.822	3.598	7.454

Table 3 indicates some interval inputs such as High School GPA, Year1 GPA, Year 1 Credits Attempted, Year 1 Cumulative Credits received Grades and Year 1 Cumulative Credits Passed seem to be quite different between the Yes and No group of GRAD6. The ACT scores do not seem to differ between Yes/No groups. The standard deviations appear quite different for the 1st-year performance inputs. Students who did not graduated in six years had a much higher variation of these 1st-year performance inputs. For example, the s.d. of 1st Year GPA for the "No" group is 1.016 compared with 0.569 for the "Yes" group.

VI. METHODOLOGY FOR MODELING THE TARGET OF GRADUATION WITHIN SIX YEARS

Input variables for predictive modeling are often from various data sources, which are collected for different purposes. Consequently, there are a wide variety of different measuring units of the inputs. For consistency, all interval inputs are standardized and categorical variables are transformed into indicator variables. Ten modeling techniques are applied to build predictive models. Data partition technique is applied. The partition is set as 55/30/15 percent, respectively as training/validation/testing data, which corresponds to the number of cases 10609/5786/2893. A brief description of each method is given in section III. The best model from each technique is obtained and compared using Receiver Operating Characteristic (ROC) curve, Misclassification Rate(MISR) and Root Average Squared Error (RASE) computed from the test data. For details of the predictive modeling techniques, one may refer to the literatures [4-11]. Table 4 summaries the ROC, MISR and RASE for the best model obtained by each modeling technique. Based on ROC curve, the best four models are NN, LR, RF and GB (from 0.875 to 0.872). Based on the Misclassification rate, the best four models, in sequence, are RF, NN, LR and EM (from 0.1907 to 0.1935). Based on the Root Average Squared Error, the best four models are, in sequence, NN, LR, GB and DT (from .3691 to .3714). Based these comparisons and taking into the consideration of interpretability and applicability, it seems the logistic regression is the best model for our purpose of predicting GRAD6.

The selected important inputs for each model are summarized in Table 5. It clearly shows that the top three factors selected by every model are "Persistence to 2nd year", "1st Year Cumulative GPA" and "1st Year Cumulative Credits Passed". "High School GPA", "1st term GPA", and "Eligible for Pell Financial Aids" are the next three selected inputs. Among the top six inputs, the High School GPA is the pre-college academic performance. Eligible for Pell Financial Aids is the family income status. The other four inputs are related to the first year academic performances, especially the 1st Year GPA and 1st Year Cumulative Credits Passed are two most important factors associated with the graduation within six years. Other inputs that are chosen by six-seven models are 1st Year Cumulative Credit Graded, and ACT Science.

TABLE 4. MODEL COMPARISON BASED ON TEST DATA

	Test Data		
Models	ROC	MISR	RASE
Neural Network	0.875	0.1911	0.3691
Logistic Regression	0.874	0.1918	0.3700
Random Forest	0.873	0.1907	0.3720
Gradient Boosting	0.872	0.1963	0.3702
Least Angle Regression	0.870	0.1997	0.3742
LASSO	0.870	0.1997	0.3742
Partial Least Square	0.870	0.2160	0.3803
Decision Tree	0.867	0.1966	0.3714
Support Vector Machine	0.862	0.1945	0.4227
Ensemble Model	0.820	0.1935	0.4144

TABLE 5. IMPORTANT INPUTS SELECTED (*: the order selected by the model)

Selected Factors	Models										# of models selected
Label	NN	LR	GB	RF	PLS	LAR	LASSO	SVM	DT	EM	
Persist to 2nd Year	1*	1	1	1	1	1	1	1	1	1	10
1st Year Cumulative GPA	2	2	2	2	2	2	2	2	2	2	10
1st Year Cumulative Credits Passed	3	3	3	3	3	3	3	3	3	3	10
High School GPA	6	7	4	6		4	4	6	6	6	9
1st Term GPA	8		5	8	4	5	5	8	8	8	9
Low Income, Pell Eligible	7	4		7		6	6	7	7	7	8
ACT Science Score	5	5	6	5				5	5	5	7
1st Year Cumulative Credits Received Grades	4			4	5			4	4	4	6
1st Year Cumulative Credits Attempted		8	7		6						3
Persist to 2nd Term					7						1
First Generation Attending College		10									1
Resident Hall Status		6									1
Minority		11									1
ACT Composite Score		9									1

TABLE 6: PARAMETER AND ODDS RATIO ESTIMATES OF LOGISTIC REGRESSION MODEL

Inputs	Parameter estimates	Odds ratio estimates
Intercept	-5.5976	
Persist to 2nd year - No	-1.8589	0.024
1st Year GPA	1.0895	2.973
1st Year Cum Credits Pass	0.0722	1.075
Low Income – No	0.2083	1.517
ACT - Science	-0.0267	0.974
Resident Hall - No	-0.3301	0.517
High School GPA	0.4543	1.575
1st Year Cum. Credits Attempt	-0.0295	0.971
ACT - Composite	-0.0430	0.958
1st Year Generation - No	0.0857	1.187
Minority - No	0.1012	1.224

Random Forest, Logistic regression and Neural Network perform equally well. For interpretability and practical uses, we select logistic regression as the overall best model. This model selects eleven inputs. The model and the odds ratio estimates are summarized in Table 6. Some observations are summarized in the following.

Given the inputs chosen in the logistic regression model, the "pure" effects of some important inputs can be interpreted based on the odds ratios in Table 6 as follows.

- If a student cannot persist on to 2nd year, the odds of graduation in six years is reduced by 97.6% when compared with those who persisted on to 2nd year.
- One point higher in the 1st-Year Cumulative GPA is 1.97 times more likely to graduate in 6 years.
- One point higher in High School GPA has 57.5% higher chance to graduate in 6 years.
- Students did not stay in Resident Hall is estimated to have 48.3% lower chance to graduate in 6 years.
- Students who are not eligible for low income financial aids have 51.7% higher chance to graduate in 6 years.
- Students who are not a first generation have 18.7% higher chance to graduate in 6 years.

VII. DISCUSSION AND CONCLUSION

This purpose of this article is twofold. One is to briefly discuss the process of conducting an analytics project and introduce ten different predictive modeling methods. The other is to conduct a case study to illustrate the problem solving strategy using a predictive modeling approach in observational studies. The case study is to predict graduation in six years and identify important input variables that may be highly associated with the graduation in six years using data from four different sources. This article discusses the possible sources attributed to poor data quality and tasks that need to be addressed during the process of data production, data cleansing and manipulations. The data for the case study were student graduation and retention data from a Research-II university.

During the model building process, we address the tasks related to preliminary variable screening, missing data handling and others prior to building models. Ten modern

predictive modeling techniques are applied to build models and compared based on three criteria: ROC, Misclassification Rate and Root Average Squared Error. As summarized in Table 5, the best statistically and practically useful is the logistic regression model. Table 6 summarizes the factors chosen by the ten models. Three inputs are selected by all models: Persistence to 2nd Year, 1st Year Cumulative GPA and 1st Year Cumulative Credits Passed.

Due to the fact that the data are observational, one must be cautious that these relationships not be directly interpreted as causal relationships with the target.

REFERENCES

[1] M. Berry and G. Linoff, Data Mining techniques, 2nd Ed., Wiley Publishing, 2004.

[2] W. McKnight," Information Management: 7 sources of poor data quality", 2009. [Online]: www.information-management.com .

[3] R. Radhakrishna, D. Tobin, M. Brennan & J. Thomson, Ensuring data quality in extension Research and evaluation studies. *Journal of Extension*, Vol. 50(3), Article # 3TOT1. 2012. [Online]: http://www.joe.org/joe2012june/tt1p.shtml.

[4] H. Zou, "The adaptive Lasso and its oracle properties". *The Journal of American Statistical Association*, Vol. 101(467), pp. 1418-1429, 2006.

[5] J. H. Friedman, "Stochastic gradient boosting". *Computational Statistics & Data Analysis,* Vol. 38, pp. 367-378, 2002.

[6] R. A. Berk, Statistical Learning from a Regression Perspective. Springer Series in Statistics, 2010.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer Series in Statistics, 2009.

[8] J. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning. Springer Texts in Statistics, 2013.

[9] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society: Series B* 58: 267–288, 1996.

[10] E. Bradley, T. Hastie, R. Tibshirani, and Johnstone, "Least Angle Regression". *Annals of Statistics* 32 (2): 407–499, 2004.

[11] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd Ed. The Morgan Kaufmann Publisher, NY, 2011.

[12] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers". *Journal of Management Information Systems 12*(4), pp 5–33, 1996.

[13] F. K. Alan, A. P. Sanil, J. Sacks, et al., Workshop Report: Affiliates Workshop on Data Quality, North Carolina: NISS, 2001.

[14] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era". *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002, 2015.

[15] U.S. News & World Report [Online]: https://www.usnews.com/education/bestcolleges/articles/how-us-news-calculated-the-rankings , 2017.

[16] Times Higher Education World University Rankings [Online]:https://www.timeshighereducation.com/, 2017.

[17] S. Gershenfeld, H. W. Hood and M. Zhan, "The role of first-semester GPA in predicting graduation rates of underrepresented students". *Journal of College Student Retention: research, Theory & Practice*. Vol 17(4) 469-488, 2016.

[18] S. Mettler, "Redirecting and expanding support for college students". In T. Skocpol & L. Jacobs (Eds.), Reaching for a new deal: Ambitious governance, economic meltdown, and polarized politics in Obama's first two years. New York, NY: Russell Sage Foundation, 2011.

[19] E. Warburton, R. Bugarin, A. Nunez, and C. Carroll, Bridging the gap: Academic preparation and postsecondary success of first-generation students (NCES No. 2001-153). U.S. Dept. of Education, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2001/2001153.pdf , 2011.

[20] D. Rogulkin, Predicting 6-year graduation and high achieving and at-risk students. Fresno, CA. Fresno State Institutional Research, Assessment & Planning. Retrieved from: http://www.fresnostate.edu/academics/oie/documents/documents-research/2011/data%20mining%20report1.pdf , 2011.

[21] P. Attewell, S. Heil, and L. Reisel, "Competing explanations of undergraduate noncompletion". *American Educational Research Journal*, 48(3), 536–559, 2011.

[22] A. Seidman, College student retention: Formula for student success. Westport, CT: Praeger, 2005.

[23] V. Tinto, "Dropouts from higher education: A theoretical synthesis of the recent research". *A Review of Educational Research, 45*(1), 89-125, 1975.

[24] T. A. Walsh, T. A. (1997). Developing a postsecondary education taxonomy for interinstitutional graduation rate comparisons. Doctoral dissertation, University of New York at Buffalo. Buffalo, NY: Dissertation Abstracts International, 1997.

[25] M. J. Druzdzel and C. Glymour, Application of the TETRAD II program to the study of student retention in U.S. colleges. Working notes on AAAI-94 Workshop on knowledge discovery in databases (pp. 419-430). Seattle, WA, 1994.

[26] D. Raju and R. Schumacker, "Exploring student characteristics of retention that lead to graduation in higher education using data mining models". Journal of College Retention, Vol. 16(4), 562-591, 2015.

[27] O. Mizuno, N. Kawashima and K. Kawamoto, "Fault-Prone Module Prediction Approaches Using Identifiers in Source Code". *International Journal of Software Innovation*, 3(1), 36-49, 2015.

[28] T. Nakagawa, Y. Iwahori and M. K. Bhuyan, "Reduction of Defect Misclassification of Electronic Board Using Multiple SVM Classifiers". *International Journal of Software Innovation*, 2(1), 25-36, 2014.