Applying Data Science Methods for Early Prediction of Undergraduate Student Retention

Cen Li
Department of Computer Science
Middle Tenn State University
Murfreesboro, USA
Cen.Li@mtsu.edu

Michael Hains
Department of Psychology
Middle Tenn State University
Murfreesboro, USA
Michael.Hains@mtsu.edu

John Wallin
Department of Physics and Astronomy
Middle Tenn State University
Murfreesboro, USA
John.Wallin@mtsu.edu

Qiang Wu
Department of Mathematics
Middle Tenn State University
Murfreesboro, USA
Qiang.Wu@mtsu.edu

***Abstract*— This paper presents a case study of applying the data science methods to a large education data collected at a University over 7 years. The goal of the study is to understand the important features, and to derive models for predicting student retention. Issues dealing with real world data, for example variable definition, missing data handling, and data cleaning were discussed. A new recursive feature elimination based feature selection method was developed. This study derived features and models for four different student groups, the first-generation students, the African American students, the Hispanic students, and the disabled students. The features identified and the predictive models built were compared crossed the four groups.**

Keywords—Data cleaning, Feature selection, Predictive analytics, Classification, Retention.

I. Introduction

Students' academic success and retention has long been a concern of colleges and universities. In the US only 30% of students graduate from 2-year colleges in 3 years or less and approximately 50% graduate from 4-year colleges in 5 years or less [1]. Theoretical retention framework and predictive models to evaluate students' academic performance and improve retention based on empirical data has been established in the literature ([2], [3], [4]). These models identify the influencing factors such as precollege academic performance (such as high school GPA, SAT/CAT score), financial situations, geographic information, and social engagement, and use student's profile to predict whether or when a student may drop out. In these models, statistical methods such as descriptive statistics, ordinary linear regression, logistic regression, and hypothesis testing have been used. Longitude studies, in which individual students are followed by time, have often been used to identify risk factors, e.g, the simple cross tabulation [5], two samples comparison [6], Markov analysis [7], and survival analysis [8]. Success stories have been reported from many Universities after implemented predictive analysis schemes into student management systems, for example University of Central Florida was able to find 1,000 most at-risk students in a timely fashion, South Texas college was able to identify that students who register late for a course are more likely to fail or withdraw, Purdue University put in "signals" about students into their Blackboard system, and Tiffin University was able to improve the one-year student retention rate from 51% to 63% in five years [10].

The primary goal of this study is to improve minority student success through predictive data analysis using student data collected between 2007 and 2013 at Middle Tennessee State University (MTSU). The target minority groups include the following four groups: the first generation students, the African American students, the Hispanic students, and the disabled students. This data includes information about academic, social, and financial aspects of the students from before they enter MTSU, i.e., high school data, to the point when they graduate or leave MTSU. By determining the risk factors that have lead to students leaving college, we will provide the data and methods needed to design effective interventions to improve student retention.

The main questions to be answered by this study include:

- Which subset of variables is the most predictive of student retention outcome? Are the variables differ among the four target groups of students?
- Can data science methods be used to construct accurate predictive models for students' retention outcome? Which method is the most effective in building predictive models for this task?

The main steps used in this analysis include: (1) data pre-processing, (2) feature selection, and (3) predictive model construction. More details about each step are presented next.

II. Data Pre-processing

A. Compile the variable definition table

The data we initially obtained contained 157 variables. The names of the variables were in shorthand notations, mostly un-interpretable directly. Furthermore, the set of valid values as well as the meaning of these values for most of these variables were unclear. So, the first order of business was to establish a variable definition table that defines the meaning of each variable, the set of values that may occur for each variable, as well as the definition of each value.

A. Build a relational database for the data

Given the size of the data, which is around 200 megabytes, it was necessary to build an efficient database scheme to support the retrieval of appropriate subsets of data, corresponding to the target student groups, for analysis.

When we received the data, they were two giant .csv files, one contained demographic data for the students, and one contained grades of individual courses for the students. Subsequently, the data was imported into the MySQL database scheme. All features that were intensively used for queries were indexed and tables partitioned in order to achieve better data

This work was supported by the Tennessee Board of Regents Faculty Research Grant.

978-1-7281-5584-5/19/$31.00 ©2019 IEEE
DOI 10.1109/CSCI49370.2019.00250

retrieval performance. In addition, based on the nature of the data to be stored per variable, appropriate data types have been chosen to make the database more compact. All data that had values "N/A", "" and " " was treated as missing and was replaced with a value NULL.

C. Data cleaning

In the original data, missing data was present for many variables across many students. A feature value estimation approach was developed to fill in the missing values by estimating them based on the non-missing values in the data. In particular, the missing values were filled one variable/column at a time. *k*-nearest neighbor (*k*NN) method was used. For each variable with missing values, the distance to determine the nearest neighbors was computed as the standardized Euclidian distance of the top *d* highly correlated variables. Each missing value was filled by the mean value of same feature of the corresponding student's *k* nearest neighbors.

The approach can be summarized in the following steps:

1. Compute the correlation matrix for all the features. For each pair of features, the correlation is computed based on the observations that are available for both features;
2. For each feature that contains missing values, find the top *d* highly correlated features;
3. For each missing value, find features among the top *d* highly correlated features whose values are not missing for the same student. This is to guarantee the distance is computable;
4. For the features found in step 3, find all the students who do not have missing values in these features;
5. Standardize apart each feature. This is to prevent the feature with large variance to dominate the distance comparison;
6. Compute the Euclidean distance between the student with the missing value and the other students based on the standardized features;
7. Find the *k*-nearest neighbors. If there are more than *k* students having the same smallest distance, all students with this smallest distance will be selected;
8. Compute the mean value of the available features of the *k*-nearest neighbors. This will be the estimated value for the missing one;
9. Repeat step 2-8 until all missing values are filled.

For the parameters, we have chosen k=10 and d=10 in this study. In order to facilitate the computation of correlation matrix for the features, all the features were assumed to be numeric valued. Norminal feature valuse were mapped into numeric values. Binary valued features have values labeled as 0 for present and 1 for not present.

D. Characterization of Freshman retention data

To gain an understanding of the overall characteristics of the data regarding the target variable, i.e., student retention, we compared the percentage of students who stayed to finish the degree (Stayed), the students who transferred to other colleges (Transferred), and the students who dropped school (Dropped), for each of the four groups of students: African American students, disabled students, first generation students, and Hispanic students, as well as the rest of the students. The results are shown in Tables 1(a). For each group, for students who did not continue with their study at MTSU, a majority dropped out of school completely, only a small percentage of students transferred to other schools. An exception is the Disabled student group, having the highest percentage of students transferring to other schools. The percentage of students who dropped out of school is similar for African American students and for Hispanic students. Disabled students have the highest percentage of students dropping out of school among the four target groups.

Table 1. Percentage of students Dropped, Transferred, and Stayed (DTS) at MTSU after their freshman year

Group	Stayed	Transferred	Dropped
African American	0.78	0.07	0.14
Disabled	0.66	0.11	0.23
First Generation	0.87	0.05	0.08
Hispanic	0.78	0.06	0.16
Other	0.68	0.07	0.25

(a) Freshman having first year GPA >= 2.75

Group	Stayed	Transferred	Dropped
African American	0.46	0.14	0.40
Disabled	0.31	0.26	0.43
First Generation	0.49	0.11	0.40
Hispanic	0.52	0.08	0.41
Other	0.37	0.17	0.46

(b)Freshman having first year college GPA < 2.75

Group	Stayed	Transferred	Dropped
African American	0.61	0.11	0.28
Disabled	0.47	0.149	0.38
First Generation	0.68	0.08	0.24
Hispanic	0.65	0.06	0.29
Other	0.48	0.12	0.40

(c)All the students

From the results obtained with feature ranking and selection studies, it is observed that the retention results tie closely to student financial situation and student academic performance. One hypothesis is that students may be dropping out of school because their GPA has dropped below the minimum requirement (2.75) for them to retain the TN Hope scholarship. Tables 1(a) and 1(b) show the retention percentages for the 5 groups of students having GPA greater than and equal to 2.75, and those having GPA less than 2.75 respectively. The results correspond positively with our hypothesis. The retention percentage rates show quite significant changes when data is split along the students' college GPA value, especially for the groups of African American students, the first-generation students, and the Hispanic students.

III. FEATURE SELECTION FOR RETENTION PREDICTION

To rank the features/variables in terms of their importance in predicting student retention, a Recursive Feature Elimination (RFE) based method was developed.

A. Feature selection based on RFEs

We first ran the ridge regression (RR) to obtain an estimated model. The regression coefficient measures the importance of the corresponding variable. This gives the ranking of all the variables. When there are many variables to estimate and the sample size is limited, the estimated model may not be accurate and may lead to unreliable variable ranking. The idea of recursive feature elimination (RFE) technique is to remove the least important variable, run the ridge regression again, and re-rank the remaining variables. This process is repeated until there is no variable left for re-ranking. The variable ranking by RFE technique is much more reliable than the direct ranking by the single estimated model obtained from using all variables. We call this method RR-RFE.

The variable ranking by RFE technique may not be robust with respect to change of samples. It may be sensitive to outliers. Therefore, we combined the RFE and bootstrap method to refine the ranking. Two different methods have been used:

- Method 1: In this method, we randomly selected a subset of the whole data set to run RR-RFE in ranking the variables. This process was repeated 100 times, each time with a different randomly selected subset of data. The final variable ranking was then determined by the average ranking;
- Method 2: In this method, we randomly selected a subset of the whole data set to run RR-RFE in ranking the variables. Based on this variable ranking we cross-validated the number of variables and obtained an optimal regression model. We recorded the variables involved in this model. This process was repeated 100 times resulting in 100 optimal models. Then we counted the number of appearance of each variable in these 100 optimal models. The refined variables ranking is based on their frequency of appearance, the higher frequency a variable appears in these optimal models, the more important it is.

B. Feature selection results

The set of features identified to best predict retention outcome are listed below:

- The students' GPA is the best predictor for retention outcome for all four student groups.
- The students' financial problem is the next best predictor for all four student groups. Financial suspension and financial probation appeared in top 6 features across the board. In addition, other financial indicators, family total income and total aid received are very important indicators for African American student group and Hispanic student group, but not so much for the first generation group and the disabled student group.
- Student Age has some impact on retention for the African American group, the Hispanic group, and disabled student group, but not for the First Generation group.
- The percentage of courses a student attended being withdraw courses or DFW courses are important indicators for dropping school for African American group, disabled student group, and first generation group, but not for the Hispanic student group.
- The lack of math foundation knowledge is a good indicator for the African American students and the disabled student group, but not an indicator for the first generation group and the Hispanic student group.
- For the African American group and the disabled student group, the lack of English reading ability is a good indicator that a student may drop the school.
- Father and mother's education levels are good indicators for the disabled student group, are limited indicators for the Hispanic student group and the African American group, but has no indication for the first generation group.
- ACT scores are usually considered a good indicator for college academic performance. For the Hispanic students, the disabled students, and the first generation students, only ACT math scores appear to be an important indicator. For the African American students, only the ACT English score is identified as an indicator.
- Learning community is a good indicator only for the the African American student group.
- Requiring prescribed courses is a good indicator for the African American group and for the first generation group.
- Whether a student is required to take courses that help with student study skills is a good predictor for student retention outcome for the African American group and the disabled student group.

Features identified to be important predictors were used in the classification step.

IV. Classification Models for Predicting Retention

Classification problem has been well studied in the machine learning and data mining community ([10], [11]). Various classification systems based on naïve Bayes, decision tree, regression model, nearest neighbor, and neural networks ([12], [13], [14]) have been developed. To further increase the classification accuracy, sampling approaches and ensemble classification methods have been developed in conjunction with one of more types of the basic classification schemes[16]. Decision tree classification, regression classification, k-nearest neighbor, and Boosting ensemble approaches have been used in this study.

To evaluate the performance of the classification models, the following criteria have been used in this study: True Positive rate (TP rate), False Positive rate (FP rate), recall, and F measure. Since retention outcome is the target variable in this study, and there are 3 possible values of this variable: Stayed, Transferred, Dropped. The evaluation criteria values have been computed for each of these three classes. Ten fold cross validation was performed to obtain the projected prediction accuracies.

Table 2 gives the results for the first-generation student data. Models learned from this student group achieved the highest true positive and precision rate among the four student groups. With true positive rate for the "Stayed" group reaching 96.2% for regression classification models and 94.4% for the decision tree models. In addition, the precision for the "Dropped" class is also higher than those from the other three student groups, at 65.9% for regression model and 60.6% for decision tree models. Despite these, the false positive rate for "Stayed" hovers around 60% for the different classification models. Boosting ensemble classification using regression

classification is not able to further improve the classification results.

Table 2. Performance of the classification models for the first-generation student group

	TP	FP	Precision	Recall	F
Dropped	0.36	0.067	0.602	0.36	0.45
Stayed	0.944	0.606	0.787	0.944	0.858
Transferred	0.031	0.024	0.095	0.031	0.047

(a) Decision tree classification

	TP	FP	Precision	Recall	F
Dropped	0.222	0.112	0.359	0.222	0.275
Stayed	0.886	0.776	0.731	0.886	0.801
Transferred	0.031	0.009	0.61	0.031	0.054

(b) kNN (k=3)

	TP	FP	Precision	Recall	F
Dropped	0.439	0.064	0.659	0.439	0.527
Stayed	0.962	0.594	0.793	0.962	0.87
Transferred	0	0	0	0	0

(c) Regression classification

	TP	FP	Precision	Recall	F
Dropped	0.439	0.064	0.659	0.439	0.527
Stayed	0.962	0.594	0.793	0.962	0.87
Transferred	0	0	0	0	0

(d) Boosting Ensemble

For the African American student data, the model that generates the best estimation is the regression classification model with its true positive rate of 91.5% . The regression classification model also gives the highest precision rate for all three classes at 70.3%, 60.2% and 46.2% for "Stayed", "Dropped", and "Transferred" classes respectively. The performance of the classification models derived using the decision tree classification is quite close to those of the regression classification models. For the Hispanic student data, the regression classification models give the best performance with true positive rate for "Stayed" students reaching 92.5%. The false positive rate is also pretty high, at 77.3%. Many data with true class label "Transferred" and "Dropped" were misclassified as being "Stayed". This is an observation for all four student groups. For the disabled student data. The best true positive rate, precision and recall are from the "Stayed" group. Decision tree models received best classification performance in this student group. The best precision for "Dropped" group, 53.1%, as well as the highest precision for "Stayed" group, 70.15%, are both obtained using the decision tree approach. The boosting performance using regression classification actually shows a slight decrease of performance.

V. CONCLUSIONS

In this study we performed a study of the student retention data by developing and applying data science methods. Like dealing with many other real world data, in a data sceince project, data cleaning and preparation often requires extra effort. Through an overall student characterization study, it is discovered that there is a strong relationship between student retention outcome and the student academic performance, i.e., their college GPA, as well as its implication on the student's financial situation. Through the predictive feature ranking and selecting study, a group of features have been identified as strong indicators for student retention outcome for each of the four student groups. The feature lists obtained from the four student groups have been compared and contrasted among the student groups. The classification study shows that the classification models are good at capturing the characteristics of "Stayed" students, i.e., students who continued to complete their study at MTSU. The accuracy for classification for this class student reaches over 90%. Yet, the models are not as good at capturing the characteristics of the students who "Dropped" from their study at MTSU. The best classification accuracy obtained for this class of students is around 66%, for the first generation students.

One of the conclusions we derived from this study is that students that are performing academically at a level that would allow them to remain in school are dropping out of school when they have fallen below the grade criteria to retain their HOPE scholarship. Finding ways to assist this group academically and/or reducing the grade point average required to retain the HOPE scholarship are likely to have the biggest impact on improving student retention.

ACKNOWLEDGMENT

The authors would like to thank the Tennessee Board of Regents for provide the funding to support this research work.

REFERENCES

[1] Bill and Melinda Gates Foundation, (2010). Next generation learning Technical report, Bill & Melinda Gates Foundation, Seattle, USA.

[2] V. Tinto, Dropout from higher education: a theoretical synthesis of recent research. Review of educational research, 45:1, pp. 89-125, 1975.

[3] S. Herzog, Measuring determinants of students returns vs dropout/stopout vs transfer: a first to second year analysis of new freshmen. Research in higher education, 46:8, 2005.

[4] S. L. Ronco, J. Cahill, Does it Matter Who's in the Classroom? Effect of Instructor Type on Student Retention, Achievement and Satisfaction. AIR Professional File, 100. 2006.

[5] A. N. Avakian, A.C. MacKinney, and G.R. Allen, Race and sex differences in student retention at an urban university. College and University, 57, pp. 160-165, 1982.

[6] J. A. Naretto, Adult student retention: The influence of internal and external communities. NASPA Journal, 32, pp. 90-97, 1995.

[7] R. M. Heiberger, Predicting next year's enrollment: Survival analysis of university student enrollment histories. Proceedings of the American Statistical Association, Social Statistical Section, pp. 143-148, 1993.

[8] P.A. Murtaugh, L.D. Burns, and J. Schuster, Predicting The Retention of University Students, Research in Higher Education, 40:3, pp. 355-371, 1999.

[9] Higher Education Blog, How Data Mining Helped 11 Universities improve student retention strategies, 2012.

[10] T. M. Mitchell, Machine Learning. McGraw-Hill Companies, Inc. 1997.

[11] J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann Publishers, Inc. 1993.

[12] H. Kurt, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, Neural Networks, 2:5, pp. 359-366, 1989.

[13] A. F. Moller, A scaled conjugate gradient algorithm for fast supervised learning, Neural Networks, 6:4, pp. 525-533, 1993.

[14] M. Riedmiller and H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, Neural Networks, IEEE International Conference on , 1, pp.586-591, 1993.

[15] N. Cesa-Bianchi, C. Gentile and L. Zaniboni, Hierarchical Classification: Combining Bayes with SVM. Proceedings of the International Conference on Machine Learning, 2006.

[16] Y. Freund and R.E. Schapire, Experiments with a New Boosting Algorithm, Proceedings of the Thirteenth International Conference on Machine Learning, pp.148-156, 1996.