7th International Scientific Conference
Technics and Informatics in Education
Faculty of Technical Sciences, Čačak, Serbia, 25-27th May 2018

Session 1: Technics, Technology and Informatics in Education UDC: 37.018.43:519.2

Predictive Analytics for Students' Success

David D. Pokrajac[1*], Vladimir Mladenović[2]
[1] Delaware State University, Institutional Effectiveness, Dover, DE, USA
[2] University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia
[*] dpokrajac@desu.edu

Abstract: *We discuss needs and necessary prerequisites to efficiently utilize predictive analytics in the educational environment. Factors, such as data availability and quality and organizational commitment are considered in the predictive analytics framework. Various tasks, both academic and financial, where the predictive analytics can find its use at the institutions of higher education are identified, as well as the technology available to accomplish the analytics process.*

Keywords: *Student's success; predictive analytics; statistics; databases; machine learning.*

1. INTRODUCTION

Higher education in the 21th century is characterized by a number of features determining its development and future [1]. On one side, there are increased requirements for the number and quality of cadre—college graduates, that are the result of the educational process. On another side, the higher education encounters a number of problems and issues that may hinder its development [2]. First, the cost of education increases as determined by exogenous variables that an institution of higher education cannot easily control. Second, the support of governmental entities to higher education stagnates of decreases. Third, the quality of public education in K-12 system oscillates any may not be able to provide sufficient preparation of future university students. Fourth, the traditional public institutions have increasing competition from private for-profit and online programs that frequently compete based on the quality of customer service rather based on the quality of educational programs. These factors, if not addressed properly, may negatively affect financial future of traditional public institutions and severely limit their ability to provide meaningful, affordable and effective programs.

In order to cope with emerging issues, the institutions of higher education increasingly rely on data in order to better understand educational and financial aspects of their operations. The abundance of various information systems utilized provides an opportunity to collect large amounts of data. On the other hand, the maturity of predictive analytics technologies, starting from conventional descriptive statistics, to advanced machine learning techniques makes it possible to obtain powerful insights on complex relationships among the data. This paper discusses various challenges that an institution of higher education may encounter during the implementation of data-driven strategies to improve financial stability and students' success.

2. DATA

The data routinely collected by the universities differ in granularity (from summary data related to the whole university, to data pertaining to academic departments and units, to students' data) and temporal resolutions (from annual measures such as retention and graduation to students class participation and access logs to a learning management system, as typical extremes). Also, the quality of data generally varies, from high quality data (students' grades and class roster) to data related to admission and financial aid, that are typically provided by students and may have high noise and percentage of incomplete entries. Data cleansing remains the first step prior to utilization of university data in predictive analytics.

The applications and software systems that collect university data range from students' information system [3] to learning management systems (LMS) [4] to assessment software [5]. A number of these systems are designed as comprehensive, using a typical waterfall paradigm, and in addition to a steep learning curve, have a high maintenance cost and limited ability for customization. Albeit the majority of the systems uses the relational or object-relational technology, due to legacy reasons, the database designs do not always adhere to principles of relational design. Specifically, normalization is not always systematically implemented, which leads to potential data inconsistencies and problems with updates. The use of data warehousing reduces these problems, at the expense of additional maintenance overhead, which may not be

affordable to small and medium-size institutions with limited budgets. In spite of numerous attempts, the development of an affordable, adaptable and customizable data warehousing software, especially suitable for institutions of higher education remains a daunting task. An ideal system should support seamless data integration from various sources, SQL-free natural language querying and dashboard capabilities, that allow high-level non-technical university management to access, analyze and visualize data at various levels.

3. METHODS AND ALGORITHMS

Classic statistics provides a number of techniques for analysis of higher educational data [6, 7]. Descriptive statistics, such as counts and estimates of sample means and standard deviations and percentiles, remain an important tool to assess overall aspects of educational activities. Measures such as graduation rates (per university, department, program, cohort, etc.), retention rates (percentage of freshmen continuing into the sophomore year), persistence (percentage of upper classmen continuing education in the following semester), passing rates (per faculty, course, program), average students' debt, grade point average (GPA), are and will remain the cornerstone of basic reports at universities. Multivariate statistics, ranging from linear regression to principal component analysis provide an ability to determine and quantify relationship among various driving attributes that may influence students' success. Time series analysis can potentially determine long-term relationships in time-variable processes such as admission processes and cash flow management.

Machine learning techniques provide algorithmic engine for data mining and predictive analytics. Classification techniques are an excellent example of power that machine leaning can add to statistical techniques. The applications of multi-layer perceptron based classifier, than can be envisioned as generalization of standard logistic regression, provide useful models for predicting discrete outcomes based on a variety of numerical, categorical and ordinal variables. [8]. Support vector machines [9] are theoretically well-founded practical tools that provide high classification accuracy. A drawback of these models, however, is that they provide limited explanatory power; their complex structure is difficult to explain to a non-technical person which may limit their use and understanding. A solution is to utilize decision trees and decision rules [10] that result in graphical or natural language models and are typically appreciated by university administrators. Bayesian models [11] provide probabilistic dependence of outputs vs. input variables and can a be tool of choice for what-if analyses [12]. Further improvement of classification accuracy is possible through usage of classification ensembles [13], bagging and boosting [14, 15] but such models to a non-technical user also appear as a black-box and thus can predominantly be used embedded in other software. Deep-learning networks [16], demonstrated successful for a variety of artificial intelligence tasks, still await their application in educational domain.

Unsupervised learning technique, such as clustering, association rules and outlier detection analysis, can help identifying structure in unlabeled data. Clustering [17] groups data points based on feature similarity. Starting from relatively simple methods, such as k-means to agglomerative and density-based clustering, it helps to determine subsets of data that may help detect or better understand underlying processes leading to an underlying data structure. Association rules [18] are generalization of standard implications, and are related to fuzzy logic [19]. They introduce important concepts of significance (how frequently a particular rule applies in a data set) and confidence (how accurate is the rule) and can be inferred from large collections of data. Outlier detection, on the other hand, provides an opportunity to detect unusual data, that may require future attention and explanation. They are especially useful in multidimensional data sets, frequent in educational domains, where manual detection may not be practical.

4. TASKS

There is a number of processes within an institution of higher education where the utilization of predictive modeling can lead to measurable and substantial improvements. During the admission process, a university's goal is to recruit, admit and enroll high school graduates or other qualified persons according to pre-specified target enrollment numbers, in order to achieve university's vision and mission. Universities with specific missions (including but not limited to minority institutions, universities affiliated with specific religious denominations, etc.) may have additional admission goals. Market segmentation, that can be accomplished using statistical tools and clustering, is essential for efficient admissions. In addition, predictive models that can estimate the probability that an admitted student will enroll at the university help concentrate admission efforts on particular segments of students.

In the process of academic advising, a student is provided information about available majors (programs) at the university. It is demonstrated that appropriate major choice can lead to substantial financial savings to a student, and improve retention and graduation rates. Hence,

development of predictive models that can assist students in providing professional orientation contributes to students' success and readiness for future careers.

One of the most critical periods for students' academic future and accomplishment of their academic goals is the freshmen year. The student, encountered with issues such as monetary, academic in narrow sense, social and psychological, is at serious risk of leaving the university. Utilization of data ranging from demographic, to financial, to academic success data available through a learning management system, provide an opportunity to generate and test an early warning system that can be utilized by academic advisors or students themselves.

Data driven models can improve class scheduling and provide students an interactive tool to assess their progress, perform cases analyses and better manage their time and plan their coursework. At the same time, the models can improve resource management at the universities (time, space, personnel) and contribute to program prioritization and strategic planning. Techniques such as association rules can find application in these tasks.

Analyses of passing rates and average grades (by class, program, and faculty) contribute to better quality control of educational process. It can identify variance due to instructor and help determining the necessary corrective actions and personnel development. Further, it is possible to analyze effects of changing educational practice (placement tests, introduction of lecturers, mandatory usage of clickers or supplemental instruction) on students' success. A potentially interesting application is to investigate correlation between students' success in different classes, which can assist in redesigning curricula, adding or modifying prerequisites, etc.

The ultimate measure of students' success is their placement after the graduation. Analysis of meaningful employment data and their correlation with students' majors and other academic variables help tailoring universities' strategic planning.

Analysis of university financial data is an important segment that substantially contributes to financial stability and future of the institution. One of the key factors is to identify various cost drivers and their influence on fixed and variable cost of education. This is especially important for private (both non-profit and for-profit) institutions, due to their substantial dependence on tuition revenue. Better understanding of variable cost can help making strategic decision about tuition models, discounts through various forms of scholarships and incentives (special pricing for early completers and academically gifted students). Analysis of administrative processes (maintenance schedule, accounts payable, housing assignment and planning), especially when combined with six-sigma framework [20] improves efficiency and can contribute to reduction and better utilization of personnel and financial resources.

5. ORGANIZATIONAL ISSUES

There are several prerequisites that an institution of higher education needs to satisfy in order to fully utilize power of predictive analytics. Data awareness is here defined as understanding of the institution, starting from its management to administrators and educators, that data, when adequately analyzed and utilized, can substantially contribute to achieving the institutional goals. Our society can be defined as data driven and data abundant, resulting in requests to pursue data-supported decision making at higher education.

However, data awareness is only the *first*, albeit necessary condition. An institution of higher education, in order to truly utilize data-driven approach and predictive analytics, needs to make substantial investments in software and human resources. At the current stage of the technology, a team of data scientists, including expertise in databases, machine learning, statistics, data warehouses and visualization is essential to establish necessary data structures and launch predictive analytics systems. This makes usage of predictive analytics technology prohibitively expensive, especially for small to medium-size organizations why would, otherwise, most benefit from launching of the data-driven paradigm. Further efforts are needed to make data analysis less technical and more available to casual non-technical users. An alternative is to establish consortia of institutions of higher education which would share data analytics capabilities, and thus reduce the cost.

The true benefit of launching the data driven approach can be achieved only if the consumers of provided information have basic data literacy, and are capable of making conclusions based on provided information, reports and dashboard. Data literacy of all university constituents need be continuously increased through a series of professional development activities, workshops and hands-on exercises.

6. CONCLUSIONS

This paper represents an attempt to provide an insight on various issues and characteristics of application of predictive analytics and data driven modeling paradigm in higher education. We discussed needs, technical environment, potential tasks where the technology can help improving students' success and efficiency, and the organizational prerogatives. The paper is not intended to provide any definitive guidance in this,

still highly emerging field, but to identify potential venues for further development and investigation. While the paper reflects the experience of its authors, it does not necessarily represent the official standpoint of their institutions.

REFERENCES

[1] UNESCO (1998). *WORLD DECLARATION ON HIGHER EDUCATION FOR THE TWENTY-FIRST CENTURY: VISION AND ACTION*

[2] Bok, D. (2013). *Higher education in America, revised edition, Princeton University Press.*

[3] Felson, J. (2015). The low-hanging fruit of technology in Academia. *Academe, 101*(5), 35-37.

[4] *Sanga, M. W. (2016).* AN ANALYSIS OF TECHNOLOGICAL ISSUES EMANATING FROM FACULTY TRANSITION TO A NEW LEARNING MANAGEMENT SYSTEM, *Quarterly Review of Distance Education, 17 (1), 11-21, 56.*

[5] Bingham, R. P. et al, eds (2015) Leading Assessment for Student Success, Stylus Publishing.

[6] Coughlin, M.A., ed (2005) Intermediate/Advanced Statistics in Institutional Research, *The Association for Institutional Research.*

[7] Coughlin, M. S., Pagano, M. (1997) Case study applications of statistics in institutional research, *The Association for Institutional Research.*

[8] *Pokrajac, D. D. et al (2016) Prediction of retention at historically black college/university using artificial neural networksm 13th Int'l Symposium on Neural Networks and Applications NEUREL 2016, DOI: 10.1109/NEUREL.2016.7800124*

[9] *Kecman, V. (2001) Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models, MIT Press.*

[10] Hastie, T. et al (2009) The elements of statistical learning:Data Mining, Inference and Prediction, 2nd edn. Springer.

[11] Probabilistic Graphical Models*: Principles and Techniques by. Daphne Koller and Nir Friedman, MIT Press (2009).*

[12] Pokrajac, D*. et al (2017) Modeling Dormitory Occupancy Using Markov Chains, Proceedings of the 10th International Conference on Educational Data Mining, pp. 346-7.*

[13] Breiman, L (2001) Random Forests, *Machine Learning, 45(1) pp. 5-32.*

[14] Breiman, L. (1996) Bagging Predictor, *Machine Learning, 24(2), pp. 123-140.*

[15] Schapire, R.E., Freund, R. (2014) Boosting—Foundations and Algorithms, MIT Press.

[16] Goodfellow, I et al (2016) Deep learning, MIT Press.

[17] Aggrawal, C.C., Reddy, C.K. (2014) Data clustering, Algorithms and Applications, CRC Press.

[18] Adamo, J.-M. (2001) Data mining for association rules and sequential patterns, Springer.

[19] Ross, T.J. (2010) Fuzzy Logic with engineering applications, Wiley.

[20] Truscott, W. (2011) Six-sigma, Continual Improvement for Business, a practical guide, Routledge.