Utilizing Big Data Analytics to Improve Education

Annapoorna Manohar[1], Pooja Gupta[2], Veena Priyanka[3], Muhammad Fahim Uddin[4]
Department of Technology Management, School of Engineering
University of Bridgeport, Bridgeport, CT USA

Analytics can be defined as the process of determining, assessing, and interpreting meaning from volumes of data. It has been categorized in three different categories - descriptive, predictive and prescriptive. Predictive analysis can serve many segments of society as it can reveal hidden relationship which may not be apparent with descriptive modeling. Analytics advancement plays an important role in higher education planning. It answers several questions such as -which students will enroll in particular course, what courses are on trending or obsolete, what is the level of student satisfaction in the current education system, effectiveness of online study environment, how to design a better curriculum, likelihood of students transfer, drop out or failure to complete the course. Not only, data analytics helps in analyzing above points but also can be helpful in predictive modeling for faculty, administrative and students groups who are looking out for genuine results about the university rankings, based on which they make their decisions. Using the dataset "Academic Ranking of World Universities, 2003-2014", we studied and analyzed to forecast how university's management and faculty could adapt to changes to improve their education and thereby the ranking of their universities in the upcoming years. Microsoft SQL Server Data Mining Add-ins Excel 2008 was employed as a software mining tool for predicting the trending university ranking. This research paper concentrates upon predictive analysis of university ranking using forecasting based on data mining technique.
Corresponding Author:
Annapoorna Manohar, amanohar@my.bridgeport.edu
Pooja Gupta, pgupta@my.bridgeport.edu
Veena Priyanka, vkaluvak@my.bridgeport.edu
Muhammad Fahim Uddin, muddin@bridgeport.edu

Introduction/Background

As an emerging field within education, a number of scholars have contended that Big Data framework is well positioned to address some of the key challenges currently facing higher education [1]. Global trending is affecting education, additionally, there has been pressure from political and social changes for institutions of higher education to respond to these rapid changes effectively and on time. In the context of the strategic planning of higher education, Big Data Analytics is relevant nowadays since both regular and distance education bring about new data useful to support the making of decisions [2]. The plethora of useful data generated makes decision making tough, however, if higher educational institutes trace data they can adapt better [3].

Knowledge discovery and data mining approaches have been utilized to make sense of the unstructured data. There are several techniques or algorithms that are helpful in extracting the characteristics of the data and building a pattern. Big data has found its place in education and is predicted to be extensively implemented in institutions of higher education.

Big data and Data Mining

Big Data describes data that is fundamentally too big and moves too fast, thus exceeding the processing capacity of conventional database systems [4]. Big data has some key properties among them are: Volume, Velocity, Veracity, Variety, Volume etc. In addition to these properties, the stages required to unlock the value of data are – data collection, data analysis, visualization and application. Some of them are classification, clustering, regression etc.

Tools such as Moodle (Open-source learning platform), Blackboard Analytics (Packaged self-

service analytics applications), GISMO (Interactive tracking system built for Moodle that displays data through a graphical interface) and Meerkat-ED (Web tool that analyzes participants and their interactions in discussion forums) are some of the examples of types of Learning Analytics Resources[5].

According to [6], various challenges have been identified in achieving successful analytics such as affordability and resources which incorporates training and tools, misuse or inaccurate data, culture and individuals privacy rights, lack of expertise, insufficient ROI and data standardization.

For higher education to develop its analytics capacity, institutions will need to view funding for analytics as an investment in future outcomes, increase the amount of funding for analytics and invest in hiring an appropriate number of analysts to institute and develop an analytics program.

The Data Science is supported by some departments that has traditional focus on the various categories of the Big Data technologies. Consequently, the Data Science program implemented by different departments has the following specialization: (1) Data Models and Data Management; (2) Machine Learning, Data Analytics; (3) Big Data Infrastructure (cloud based), Computer Platforms, Big Data tools [7].

Although there is an increase in big data technology, it doesn´t mean that classical BI tools like Cognos, QlikView, SPSS have come to an end. Instead, trend is that BI tools can work with new Big Data technologies. Here is a list of some applications, working with traditional data and Big Data such as Apache Hadoop, IBM InfoSphere, BigInsights, Cloudera, IBM Cognos, IBM SPSS, SAS, Quest [8]. But, in Web-based Education Systems (WBES) that own several options for content, sequencing, and evaluation material, our student modeling offers a predictive support for student-centered education [9]. However, the Microsoft SQL Server Data Mining Add-ins Excel 2008 was employed as a software mining tool for predicting the trending university ranking. This tool was considered as it's an open source and will avoid the challenge of affordability.

Data mining is a technique that extracts hidden patterns from already processed data by using tasks such as classification, artificial neural network, decision trees support vector machine, clustering, regression association rules, forecasting technique. By utilizing data mining technique we discovered various models, summaries, and derived values from a given set of data [10]. Even though various models were discovered mining process was initiated with data cleaning, data integration, data selection, data transformation processes.

Pre-processed data undergoes a refining process as stated, for making better business decision, data analysis, pattern analysis, archaeology and dredging of data.

Related Study

A literature review of academic research associated with data analytics and descriptive modeling in the Educational sector reveals the following facts:

- Competition for Admissions: The advent of ranking systems has given students and society more data to evaluate the quality of an educational institute. Unlike the olden days when people had less knowledge about the quality of education being imparted in an educational institute, thanks to the extensive amount of data available in this age of information, many organizations that engage in ranking universities have come into existence and help college-goers choose the best institute that fits their set of requirements [11]. However, there has been little evidence that high competition has had positive effects on what students learn.

- Student Performance - Predictive Analysis: Research papers also pointed out towards a few factors that pre-empted the probability of success of a student [12]. These were:

    - Past Performance: If a student has a past record of scoring good grades, it became a strong

indicator of the future performance of the student.

- Demographic Outlook: Multiple research articles and surveys also proved that students who are married performed better at studies than single students. It was also mentioned in the research papers that the older the student is, the higher the changes of a better GPA are.
- Subject Choice: It has come to the fore through various researches that those students who chose math and honors in high school were deemed to succeed in undergraduate and graduate studies than those students that chose other subjects.
- Other Factors: There were some other factors noted in the research that proved to be strong indicators of students' success. These included the performance of a student in online classes and the ratio of attempted to that of credits completed.

- Academics & Business Intelligence: In all the researches that were undertaken, it was discovered that business intelligence was hardly used in the educational sector [13]. However, it has tremendous potential and can be used by educational institutes in increasing the enrollment numbers as well as sifting through student applications.

- Machine Learning: Another angle to data analytics in educational institutes that was explored in all the research literature was to do with machine learning algorithms. The C4.5 algorithm which is essentially a decision tree algorithm can be used to effectively design predictive models from the student data that has been accumulated over the years [14, 15].

Data Analysis Models

This paper focuses on the data mining add in of Microsoft SQL Server Data Mining Add-ins Excel 2008. A sample data set "Academic Ranking of World Universities, 2003-2014" extracted to undergo the lifecycle of a data mining process, which includes formulating/refining data, evaluating and analyzing mining models, thereby predicting results with the use of spreadsheet. For this process, user must have installed Microsoft Excel 2008 for the Table Analysis and Data Mining Client add-ins. Since the approach was based on Table Analysis Tools, we had to convert our raw data into table format that was supported by Excel.

The steps involved during the process were: Data Preparation, Data Modeling, Accuracy and Validation, and Model Usage.

In the Data Preparation process, picking the correct attributes from the source (exploring data), removing the outliers (cleaning data), splitting the data set into samples (partitioning data) were the common preparation needs.

Several Data Models are supported by the add-in, such as: Clustering, Decision Tree, Time-Series, Pie Chart, Neural Networks, Sequencing Clusters, and Histogram etc.

Accuracy and Validation generate estimation models that evaluate against the test data. Classification matrix, Accuracy Chart and Profit Chart are few of the parameter evaluators. In the Model Usage, there are two phase wherein in the browse part we explore the patterns from the output. In second phase, we query the model to predict from the new data.

Our dataset, "Academic Ranking of World Universities, 2003-2014" had various factors on which the descriptive and predictive modelling was done. Some of the factors were-

a) Alumni which had around 10% of total- It refers to the number of the alumni who wins Nobel Prizes and other Medals.
b) Award which had around 20% of total- The total number of the staff winning Nobel Prizes.
c) Highly Cited (HiCi) with total 20% - referring to of Highly Cited Researchers in twenty one different subject categories.
d) Publication PUB (20%) - Count of papers indexed in Science Citation Index and Social Science Citation Index in 2012.
e) Per Capita Performance PCP (10%) – weighted scores of above stated five values

divided by the count of full-time equivalent academic staff.

Based on different parameters, the ranking of university changes. For example in the below figure, Figure1, on the basis of PCP in 2014, the ranking for the university is high (good) for higher score. In 2014, California Institute of Technology had the lowest score of PCP, so its ranking was the best (6). This analysis is helpful for universities who can focus on improving their PCP score which is dependent on above stated indicators. More publication, more HiCI, more awards can help them get a better ranking.

United States

		Units	20
National Rank	Arizona State University	National Rank	4
	Boston University	National Rank	4
	California Institute of Technology	National Rank	
	Carnegie Mellon University	National Rank	3
	Cornell University	National Rank	1
Score on PCP	Arizona State University	Score on PCP	1
	Boston University	Score on PCP	2
	California Institute of Technology	Score on PCP	10
	Carnegie Mellon University	Score on PCP	3
	Cornell University	Score on PCP	3

Fig 1. Ranking of universities in USA based on PCP score.

Another analysis was based on number of awards, the ranking can always change. For Boston University, in 2005, there were no staff winning Nobel Prizes or Field Awards, as a result its ranking got dropped. While for California Institute of Technology had the highest awards won by the staff, so it's ranking was the best.

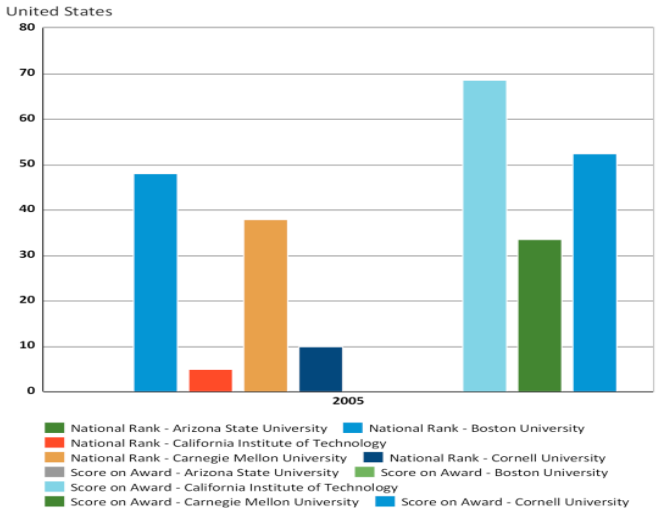

Fig 2. Bar Chart representation of ranking of universities in USA based on Awards

Similar models were generated using data mining add-in to analyze the factors and their influence on improving or deteriorating the ranking of universities.

In addition, from our research we examined that criteria like cultural, economic and historical stature cannot be the basis on which universities can be ranked. These ranking barriers may mislead students in deciding the university for their bright future.

Challenges

Some of the issues faced while implementing the data mining process for analyzing the trend in the university ranking were:

a. Data Fog situation, accuracy, multiple truths and extraction of data.
b. Finding the correct and related data set for the research.
c. Cleaning and refining the data set according to the requirements of the software.
d. Lack of data governance.
e. Understanding the algorithms provided by the data-mining add-in.

Conclusions

Microsoft SQL Server Data Mining Add-ins Excel 2008, the tool used could provide meaningful predictions upon which universities

can take corrective measures to enhance the quality of education system, improve their faculty contribution towards society. Further, the descriptive modeling can help evaluate the teaching staff and their excellence in imparting the education. This study provided vital information on which universities need to formulate new policies. They can design strategies according to the parameters they are falling behind on. However, for universities to incorporate the data-mining technique into their current systems will not be an easy endeavor. Bringing in changes to the already existing setup would require enormous transformation in terms of cost, resources and tools.

References

1. Siemens, G., *How data and analytics can improve education, July 2011*. Retrieved on August, 2011. **8**.
2. Amorim, J.A., et al. *Big Data Analytics in the Public Sector: Improving the Strategic Planning in World Class Universities*. in *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*. 2013.
3. Daniel, B., *Big Data and analytics in higher education: Opportunities and challenges.* British Journal of Educational Technology, 2015. **46**(5): p. 904-920.
4. Manyika, J., et al., *Big data: The next frontier for innovation, competition, and productivity*. 2011.
5. Reyes, J., *The skinny on big data in education: Learning analytics simplified.* TechTrends: Linking Research & Practice to Improve Learning, 2015. **59**(2): p. 75-80.
6. Bichsel, J., *Analytics in higher education: Benefits, barriers, progress*, 2012, and Recommendations.
7. Demchenko, Y., E. Gruengard, and S. Klous. *Instructional Model for Building Effective Big Data Curricula for Online and Campus Education*. in *Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on*. 2014.
8. Michalik, P., J. Stofa, and I. Zolotova. *Concept definition for Big Data architecture in the education system*. in *Applied Machine Intelligence and Informatics (SAMI), 2014 IEEE 12th International Symposium on*. 2014.
9. Lias, T.E. and T. Elias, *Learning Analytics: The Definitions, the Processes, and the Potential.* 2011.
10. Kantardzic, M., *Data mining: concepts, models, methods, and algorithms*2011: John Wiley & Sons.
11. M'Hammed, A., H. Wu, and Y. Cherng-Jyh, *Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade.* Journal of Educational Technology & Society, 2012. **15**(3): p. 77-88.
12. Ramesh, V., P. Parkavi, and K. Ramar, *Predicting student performance: a statistical and data mining approach.* International Journal of Computer Applications, 2013. **63**(8): p. 35-39.
13. mar Pal, A.K. and S. Pal, *Analysis and Mining of Educational Data for Predicting the Performance of Students.* 2013.
14. Bound, J., B. Hershbein, and B.T. Long, *Playing the Admissions Game: Student Reactions to Increasing College Competition.* The Journal of Economic Perspectives, 2009. **23**(4): p. 119-146.
15. Guster, D. and C. Brown, *The application of business intelligence to higher education: Technical and managerial perspectives.* J. of Information Technology Management, 2012. **23**(2).