SOLVING EDUCATIONAL PITFALLS WITH BIG DATA

Zeehasham Rasheed
Principal Software Engineer Data Science
Verizon
Adjunct Professor - George Mason University
zeehasham@gmail.com
# Table of Contents

Introduction ........................................................................................................................................... 4  
Role of Big Data in Education Data Analytics...................................................................................... 4 
Types of Educational Data ................................................................................................................. 4 
Investigating Potential Questions from Educational Data .................................................................. 5 
  Admissions ........................................................................................................................................ 5  
  Enrollment and Course Registration Pattern .................................................................................... 6  
  Retention and Graduation .................................................................................................................. 6  
  Costs ................................................................................................................................................ 6  
  Productivity ................................................................................................................................... 6  
  Learning Activity ............................................................................................................................... 6 
Proposed Education Data Analytics Framework ................................................................................. 7 
Data Analytics Models .......................................................................................................................... 9 
  Enrollment, Admission, Tuition Fee and Cost Model …................................................................. 9  
  Regression Model ............................................................................................................................. 9  
  Association Rule Model ................................................................................................................... 9  
  Decision Tree Model .......................................................................................................................... 10  
  Clustering and Segmentation Model ............................................................................................... 11  
  Sentiment Analysis Model .............................................................................................................. 11  
  SQL Query Engine ............................................................................................................................ 11  
Proposed Technology Specification .................................................................................................... 11 
Proof of Concept with Demo Application .......................................................................................... 12  
  Analysis Methodology and Outcome .............................................................................................. 14  
Conclusion ......................................................................................................................................... 16  
References .......................................................................................................................................... 17
Table of Figures

Figure 1: Proposed Framework for Big Data Analysis in Education .................................................. 8
Figure 2: Proposed Framework for Big Data Analysis in Education .................................................. 8
Figure 3: Technology Stack for Data Analytics and Reporting Solution System ................................ 12
Figure 4: Demo Application Main Screen .......................................................................................... 13
Figure 5: Demo Application Data Explorer View ................................................................................ 13
Figure 6: Demo Application Exploratory Analysis Dashboard ............................................................. 14
Figure 7: Exploratory Data Analysis on Students Homework Activity Logs ........................................ 15
Figure 8: Student Segmentation with Correlation Analysis using RFM Model .................................... 16

Disclaimer: The views, processes or methodologies published in this article are those of the author. They
do not necessarily reflect Dell EMC’s views, processes or methodologies.
Introduction

Data-driven decision making in schools and universities is a global yet overlooked trend in the education sector. The use of educational data to generate insights for accountability, self-improvement and support day-to-day operations discovers several critical issues in current initiatives for data analytics. However, initial results in teaching and learning analytics research using emerging data analytics technologies especially in the area of Big Data are promising. A vital aspect of these results can be considered as Data Literacy, which empowers institutes to make use of data in their decision-making processes and to meet both external requirements of regulatory standards as well as internal requirements of continuous self-evaluation and improvement needs [1].

Role of Big Data in Education Data Analytics

As the requirement of processing big data in terms of increasing volume, velocity and variety is surpassing the capability of conventional systems [2], more innovative tools and techniques are required to manage bigger data sets. The exploration of a massive amount of data is not novel, however, big data provides more challenges which are intended for investigating correctness of data (veracity) as well as different types and formats of data (variety) [3].

With such advancement in technology, the analysis of big data provides useful information and discovery of meaningful patterns, which is commonly defined as big data analytics. The impact is more evident when it unleashes its competency to provide effective data-driven decision making in teaching and learning analytics. Teaching and learning analytics facilitates teachers to reflect on their teaching methodology using evidence from the delivery to the students. It also provides insights about learners and their contexts to better understand and improve the learning environment.

In the same context, big data concepts is useful for a range of instructional and administrative applications in higher education such as monitoring and checking student performance, improving teachers capacities, and reviewing course evaluations [4]. This article concentrates on types of educational datasets, different questions related to those datasets and different methods for leveraging analytical capabilities to answers those questions.

Types of Educational Data

Schools and universities have a broad range of educational data generated by various sources, both internal and external to the school. For example student portfolios, grades, credit hours, and work schedules pile up over the years and educational institutes continue to uncover new ways to convert that data into actionable insights. Course Management Systems (CMS) and Learning Management Systems (LMS) such as Blackboard, Moodle, etc. have the capability to capture those extensive and time-sensitive transactional data points. Most such systems afford continual observing of learner activities, access to reading material, reactions, posts on a discussion board, solving of a test or quizzes, and final evaluation scores. Recording and investigation of these transactions in real-time can be used as input data for learning analytics application. Furthermore, additional sources of data can include any combination of
location, previous learning activities, health concerns (physical, emotional and mental), attendance, socio-economic data (parental income), and parental status that will add value to the overall analysis. Most universities store and aggregate this data under the umbrella of institutional statistics and can be categorized as

1. Student data, such as demographics and prior academic performance
2. Teacher data, such as competencies and professional experience
3. Activity data that is generated during teaching, learning, and assessment processes, both within and beyond the physical classroom premises, such as lesson plans, methods of assessments, classroom management, online activity monitoring and progress.
4. Human Resources, Infrastructure, Enrollment and Financial Plans, including educational and non-educational personnel and operational expenditures.
5. Student reviews and recommendations, social and emotional development such as support, respect to diversity and special needs.

With this richness in educational data, there is a need for the design and implementation of an analytics framework to increase institutional knowledge as well as improving responsiveness and real-time management. This, in turn, will make education institutes achieve dramatic improvements in both operational performance and the attainment of strategic objectives.

**Investigating Potential Questions from Educational Data**

The core objective of data analytics in education is to provide solutions that will operate cooperatively with the educational environment and build upon business rules, processes and practices. It should also bridge the gap between enrollment, student success and financial data and provide the ability to apply predictive and prescriptive modeling to answer a multitude of questions across the administration, teacher and student life cycles. Several questions that can be answered by leveraging educational data through data analytics. Some of them are listed below.

**Admissions**

1. Who is more likely to apply to the institute after initial prospecting?
2. Who will enroll in which discipline/program of study?
3. How does increasing the average SAT/GMAT/GRE scores affect incoming enrollment?
4. What will be the expected number of applications in the upcoming semesters?
5. What is the average GPA of students for each program?
6. What strategies in terms of marketing or financial aid would increase enrollment?
7. Who are the potential students for financial aid offering to maximize revenue?
8. How many students who applied for financial aid that meet all the requirements?
9. What is the impact of additional aid by program for financial aid leveraging?
10. How can we measure the performance of students based on the level of scholarship provided?
11. What combination of financial aid would increase the enrollment rate?
12. What is the impact of marketing efforts on out-of-state or in-state enrollment?
13. What is the net revenue gained by the upcoming class compared to previous years?
Enrollment and Course Registration Pattern
1. How does change in tuition fees affect the percentage of total enrollment?
2. Who is more likely to enroll in a particular course?
3. What is the frequent student course-taking patterns?
4. Which classes and courses are likely to fill up quickly or need additional sections?
5. Which courses have high dropout rates?
6. Who is likely to change a major discipline and how does it affect graduation rates?
7. Which courses are in high demand and why?
8. What is the average course load per student by program?
9. What is the number of students by status in each program at any given time?
10. How to leverage data for enrollment modeling by year and status along with tracking retention so that we compare with the approved budget.

Retention and Graduation
1. Who is likely to return (first, second, third year)?
2. Who is likely to graduate in four or six years?
3. Does lowering admissions requirements negatively affect overall retention?
4. How much financial aid will be vacated by students who do not return to the university?
5. What are the differences in terms of graduation rates for first time freshman and transfer students with respect to each academic program?

Costs
1. What is the instructional costs per student credit hour by program, department and level?
2. What is the average amount spent to enroll each student by program?
3. What is the total instructional cost per student by level, course, degree, program, department, school and college?
4. What are the revenues generated by each student enrolled by course, degree, program, department, school and college?

Productivity
1. Which degree programs are in demand and growing in enrollment along with increasing degrees awarded and at the same time contributing to revenue?
2. What is our current cash by program?
3. At any given point, what is our cash collection compared to calculated budget?
4. What impact would there be if we had different rate structures by school, college or by program?

Learning Activity
1. How can we predict learning sequences, student knowledge behavior and final grades?
2. How do we determine significant relations and patterns between student progress on consuming knowledge, activity time and usage on online learning systems and students grades?
3. How to discover the relationship between usability of the course materials and student learning performance?
4. How can we assess students’ performance based on their schedules, grades, disciplinary records and attending information?
5. How can we construct students’ metric based on engagement, page views, click rate, frequency of posting, number of logins, etc. using online activity logs from the learning systems?

**Proposed Education Data Analytics Framework**

The general process of extracting information from big data can be broken down into five stages [5]

1. Data acquisition and management
2. Extraction, cleaning and annotation
3. Integration and aggregation
4. Modeling and Analysis
5. Interpretation and presentation

With respect to the role of big data and learning analytics in educational institutes, a general architectural framework is proposed in Figure 1 that supports multi-structured data sets processing, analysis and presentation. The roles and abilities of two groups (i) Students, Teachers, Researchers and (ii) Data Analyst, Data Scientists, Data Engineers and Administrators are clearly defined and proposed within this framework and explained in the section below.

Data collection module maps, aggregates and cleans data from different sources. Structured data that constitutes about 5% of all the available data [6] such as Learning Management Systems (LMS), Administration and Enrollment data refers to the tabular data found in spreadsheets or relational databases. On the other hand, text, images, audio, video, and data from web logs are examples of unstructured data that sometimes lack the structural organization required by analytical methods.

A group of data-savvy people (Data Analyst, Data Scientists, Data Engineers) are responsible for preparing ETL (Extractions, Transformations, and Loading) layer. The ETL module includes functions such as data integration to relevant tables, data transformation, and loading of data specifically for advance analysis. Data scientists use this data to create complex predictive and business intelligence models that produce useful information and actionable insights. The output is then transferred to the presentation layer.

The presentation layer provides a user-friendly graphical interface where students, teachers, and researchers can easily retrieve information without the need for the in-depth data analysis knowledge, programming skills or database schema in the background.

We made this proposal, according to the most important aspects of the implementation of big data and analytics in the field of education that covers integration of data from different sources, embedded ETL and statistical computational layer and the possibility of high quality presentation layer to the end users (students, teachers, and researchers). The presented framework can be useful for providing an exceptional high performance and flexible data processing framework for ongoing study in educational research.
Figure 1: Proposed Framework for Big Data Analysis in Education
Data Analytics Models

Following are some data analytics models and methods to answer potential questions related to educational datasets

**Enrollment, Admission, Tuition Fee and Cost Model**

The primary task is to develop a model that can project future student enrollment. Simple methods such as moving averages with exponential smoothing can be used to predict the enrollment. Other complex time series models for example Box-Jenkins method, which uses Autoregressive Moving Average, would be another candidate model if required. Data preparation and model building requires complex calculation of counts, running averages and percentages.

Tuition fee model helps forecast net tuition revenue with respect to course, degree, program, department and school. It can also solve the use-cases regarding what/if scenarios under different tuition fee model and financial aid packages. Data preparation and model building requires complex calculation of counts, retention, aggregation, factor multiplication and percentages. Examples of calculations in tuition fee model include Attrition Rate, Percent change in tuition fee, Per Credit Hour Rate, Per Credit Hour Increase, and so on.

Accuracy of these models can be evaluated by applying a model on historical data. Actual and predicted values along with difference and percentage of error will determine the effectiveness and accuracy.

**Regression Model**

Regression model uses multivariate analysis to investigate the relationship between input variables with output variables. It is also used to analyze how the change of one variable can affect the outcome. For example, how a change in tuition fee can affect total enrollment can easily be investigated using regression model. If we consider tuition fee and enrollment use-case, we can define input and output variables as

1. **Input variables**
   a. In-state tuition and fees
   b. Out-of-state tuition and fees
   c. Total scholarships, fellowships and grants
   d. Other input variables: i.e. number of graduates, state population, unemployment rate, etc.

2. **Output variables**
   a. Enrollment headcount
   b. Total credit hours

Using regression, we can estimate the effect of tuition on total enrollment and total credit hours. The coefficients given by regression model for each input will determine their effect on output variables such that if their relationship is directly proportional or inversely proportional to each other.

**Association Rule Model**

Association Rules is an algorithm that helps uncover relationships and patterns between seemingly unrelated or scattered data in a relational database or other information repositories. An example of an
Association rule would be if a customer buys product X, he is more likely to buy product Y. We can also apply the same logic to discover course-taking patterns among undergraduate/graduate students.

For the input of Association Rule model, we require student information such as courses taken by each student in each semester throughout the academic career. Once the input is prepared, Association rules and patterns are generated by analyzing data for frequent patterns and using the criteria Support Confidence and Lift to identify the most important relationships.

Support is an indication of how frequently the course appears in the data. Confidence indicates the number of times a combination of two or more courses appear in the data. Lift interprets the importance of any rule. It is the ratio of the confidence and expected support values of the rule. A lift value greater than 1 indicates that the pattern appears more often together than expected and is significant.

**Decision Tree Model**

Decision Trees (DTs) are supervised learning method used for classification and regression. The goal is to create a model that predicts the outcome of a target variable by learning simple decision rules inferred from input data. DTs are simple to understand and interpret and rules can easily be visualized using tree or flow chart structure. We can start describing DT methodology by solving two use-cases of predicting student retention or student GPA. We can categorize the input and output variables of DT model in the following groups:

<table>
<thead>
<tr>
<th>Achievement</th>
<th>Demographics</th>
<th>Financial</th>
<th>Social/Situational</th>
</tr>
</thead>
<tbody>
<tr>
<td>College</td>
<td>Age</td>
<td>Scholarship Aid</td>
<td>BlackBoard/LMS Login per Course</td>
</tr>
<tr>
<td>Degree</td>
<td>Gender</td>
<td>Athletic Aid</td>
<td>BlackBoard/LMS Login per Week</td>
</tr>
<tr>
<td>Major</td>
<td>Ethnicity</td>
<td>Loan Aid</td>
<td>Transfer</td>
</tr>
<tr>
<td>Concentration</td>
<td>Area of residence</td>
<td>Household Income</td>
<td>Readmitted</td>
</tr>
<tr>
<td>hours registered</td>
<td>Marital Status</td>
<td></td>
<td>High School Graduation Year and Month</td>
</tr>
<tr>
<td>earned hours</td>
<td></td>
<td></td>
<td>Fitness Class Attendance</td>
</tr>
<tr>
<td>GPA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cumulative GPA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT Scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stem Courses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Stem Courses</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Decision Tree model analyzes each variable in a data set, determines which are most important variables using Information Gain, Gini Index and Entropy metrics, and then creates a tree of decisions which best partitions the data to predict the outcome. The tree is created by splitting data up by variables at each level by placing the most significant variable at the top/root and then the next significant variable on the second level and so on. The leaves (lowest level) of the tree contain the final output/outcome of the model.
that tells us the predicted retention status or GPA of the student along with the confidence score. The proposed DT can be customized and used with several use-cases and predict the outcome.

**Clustering and Segmentation Model**

Clustering and segmentation are unsupervised learning algorithms that are used for exploratory data analysis. Clustering algorithms such as K-Means and DBScan can be used to find different groups of students in the data and each group is further investigated based on their unique features as compared to other groups. This analysis also tells us how many different types of students exist within data. Similarly, segmentation model in general is widely used in the industry to segment customers such as loyal, new, likely to quit, lost, etc.) and help companies monitor churn rates. RFM (Recency, Frequency and Monetization) algorithm is a good candidate to segment customers based on their behavior. We can apply the same methodology on our student data and segment them. Using the outcome of RFM model, we can infer which group of students are eligible for Financial Aid or which students are having difficulty with course work and enable maximizing their efficiency.

**Sentiment Analysis Model**

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer’s attitude towards a particular topic, course, etc. is positive, negative, or neutral. We can evaluate course standing as well as instructor performance by processing student course evaluation comments and applying sentiment analysis on it. There are several open source packages available that include built-in functions to tokenize text, count frequency of words, find probabilities of each word as a positive, negative or neutral, and use classification algorithms to produce the output.

**SQL Query Engine**

This engine is the core of data analytics and reporting solution as several questions can be answered by Structured Query Language (SQL) queries. For example, which degree programs are growing and contributing to revenue, how many students by status at any given time, and so on. Also extracting and preparing the data for all the analytical models above require SQL query engine to join different data sources, filtering out unnecessary data points and feed them into the model.

The proposed SQL engine will support all kinds of simple to complex SQL operations such as SUM, COUNT, AVERAGE, RANK, PARTITION, GROUP BY, ORDER BY, HAVING, RANK, JOIN, etc. A custom SQL model can also be added in presentation layer where the end user can write customized query at runtime and get the output using graphical user interface (GUI).

**Proposed Technology Specification**

Figure 2 shows the proposed tiered technology stack that drives data analytics application and illustrates the relationship among three tiers and tools for each tier. We describe each component of technology stack in detail below.

Tier 1 is related to system infrastructure that contains tools for data ingestion (extracting and importing data). Tier 2 contains all the processing components including Query Engine, Application Processing
Interfaces (APIs) to interact with Tier 1 and Tier 3 and all the data mining and machine learning algorithms used for data analytics. Tier 3 provides an elegant and powerful framework for building web applications and provides an easy way to access actionable insights in the form of plots, tables, charts, or other kind of visualization.

![Technology Stack for Data Analytics and Reporting Solution System](image)

**Data Engines:** SQL Server, Oracle, DB2, Hive (Hadoop), Redshift (Postgres) etc.

**Enterprise Resource Planning (ERP) and Business Intelligence (BI):** PeopleSoft, SAP, Oracle, Banner, SAS, Microstrategy etc.

**Proof of Concept with Demo Application**

For this article, a novel application is developed which uses statistical analysis and Recency Frequency Monetization (RFM) model to analyze homework assignment activity logs and improve student performance by identifying different segments of students based on risk level. This application provides real-time insights and helps instructors perform evaluation at homework assignment, student and overall class levels while they are teaching a course. Below are some screen shots for this application.
Welcome to ZR and HR Analytics

ZR and HR Analytics is a web based application for Data Modelling and Data Analytics. This portal provides new insights and recommendations by applying cutting edge data mining and machine learning algorithms on students moodle, blackboard or any other online activity data.

Click Details in the side panel to get familiar with data exploration techniques and algorithm
Click Data Explorer in the side panel to view and download the data
Click Exploratory Analysis in the side panel to crunch your data

Figure 4: Demo Application Main Screen

Figure 5: Demo Application Data Explorer View
Analysis Methodology and Outcome

For demonstration purposes, student online activity logs are analyzed over some period of time during classes. Based on activity data over four homework assignments, relationships between number of online portal access, submission counts and students final scores are identified. This shows that student behavior and patterns of online activity during course work can be used to improve the course content. In addition, student segmentation is also performed to identify groups of students who show similar progress over time.

We use basic statistical methods such as min, max, mean, variance, standard deviation along with visualization techniques to explore whether online participation as reflected in the activity logs, is related to overall performance in the class. Particularly, we are interested in finding if (1) students engaging with the assignment submission and competition and (2) accessing the homework submission portal is associated with higher performance. Figure 6 shows statistical analysis on one homework submitted by students. In Figure 6 part (a) we see that students who get the maximum score (x-axis) have very few submission count (y-axis) for a particular homework, whereas students who got an average score have higher submission counts. In Figure 6 part (c) we see that larger number of accesses (y-axis) to online system does not guarantee good scores on homework assignments (x-axis). Furthermore, Figure 6 part (d) shows that the number of accesses increase as students proceed towards the end of course work.
Figure 7 shows the segmentation analysis using RFM model on student homework activity log data. The idea is to group/segment similar students based on their online activity and predict their grades based on segment information. In Figure 7 (b) we define six custom segments for students where (i) H refers to High Performers, (ii) X means Failing, (iii) L refers to Lost, (iv) R means Repeated or Recurrent, (v) N means New, and (vi) O means One Timers. In this model, Recency is considered as interval of last login or access by a student, Frequency as number of submissions by a student and Monetization as the score or grade earned by each student during that interval. In Figure 7(d), a correlation plot is generated between Predicted Score and Actual Score, which shows good correlation and validates the segmentation results.
Conclusion

With regards to the acceptance of data analytics benefits, it is an exceptionally practical time to explore analytics in educational institutes. Complex challenges that schools and universities encounter due to lack of actionable insights can be solved through analytics applications. Even though data analytics is has yet to be widely adopted in educational institutes, several administrators and faculties have confirmed that analytics brings significant improvement with respect to organizational resource provision, administration, finance, teacher capabilities and student achievements [7][8]. Furthermore, big data provide scientists and researchers with the opportunity to comprehend the meanings of these data and how they can be analyzed in a significant, effective, and consistent manner to provide actionable insights. For all these reasons, all education organizations need to have a strategy for how they can take advantage of big data analytics.

This article presents big data concept, possibilities, methodologies and tools for expanding the capabilities of educational systems. To summarize the overall objective, the importance of big data and analytics in education is twofold: effective administration and decision making in educational institutes, and helping faculty in improving their teaching and learning activities. The implications derived from current research will advance the discussion about building effective learning analytics systems for students, instructors and administrators based on big data. Ultimate success of this concept should not be evaluated based on the notion that analytics is just working for us, but whether this concept is used successfully elsewhere.
References

1. Demetrios G. Sampson, “Educational Data Analytics Technologies For Data-Driven Decision Making In Schools”, ELearning Industry, October 20, 2016
2. J. Manyika et al., “Big data: The next frontier for innovation, competition and productivity”, 2011

Dell EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” DELL EMC MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying and distribution of any Dell EMC software described in this publication requires an applicable software license.

Dell, EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries.