Mining Login Data For Actionable Student Insight

Lalitha Agnihotri
McGraw-Hill Education
lalitha.agnihotri
@mheducation.com

Ani Aghababyan
McGraw-Hill Education
ani.aghababyan
@mheducation.com

Shirin Mojarad
McGraw-Hill Education
shirin.mojarad
@mheducation.com

Mark Riedesel
McGraw-Hill Education

Alfred Essa
McGraw-Hill Education

ABSTRACT

Student login data is a key resource for gaining insight into their learning experience. However, the scale and the complexity of this data necessitate a thorough exploration to identify potential actionable insights, thus rendering it less valuable compared to student achievement data. To compensate for the underestimation of login data importance, in this paper we performed an exploratory data analysis of a large educational dataset consisting of 100 million instances of login data from 1.5 million unique students who attempted 783 thousand assignments. The data were from a McGraw-Hill Education web-based assessment platforms called *Connect*. Different data mining methods were employed to answers our initial questions regarding students' login behavior. Most of the findings were consistent with the intuitive expectations of student login patterns such as a considerable decline of activity on Saturdays, a visible peak on Sunday evenings, a high activity in September and February, and an increased activity toward later hours of the day. However, we also discovered an unexpected result while investigating the effects of the login activity, the performance scores, and the attempts. Surprisingly, this analysis showed a high positive correlation between login activity and performance scores, only up to a certain threshold. This provided us a new hypothesis on student groupings, which we explored through a cluster analysis. As a result of our exploratory efforts, a significant amount of patterns emerged that not only confirmed previously set forth expectations but also provided us new hypotheses, which can be leveraged to improve student outcomes.

Keywords

Exploratory data mining, assessment platform, clustering, log data, pattern & trend mining

1. INTRODUCTION

An increasing number of higher education institutions are incorporating online course management platforms, which creates a tremendous opportunity for monitoring learners' academic activity. These web-based learning environments capture immense amounts of login data that could be used for student monitoring and profiling ([11]). Educational literature suggests that monitoring students' academic activity is a key to a more effective and higher quality education ([2], [3], [7], [8]). Furthermore, research shows that college students would benefit from opportunities of introspection and cognitive monitoring of their progress in order to engage in careful academic planning ([1]). Hence, given its scale, these login data are a promising resource for shedding light onto students' academic behavior.

In this paper we explore login data from a McGraw-Hill Education's (henceforth MHE) web-based assessment platform. These data can serve as a basis for instructors' personalized intervention programs and feedback for student efforts toward self-regulated learning. While interest in login data analysis has been continuously increasing, there is no standardized way of analyzing this type of data ([9]) due to diversity of the data and uniqueness of research questions. Hence, we conducted exploratory data analysis without setting a priori limitations or hypotheses on our data. In Sections 2 through 4, we discuss our methods with detailed descriptions and their findings. Section 5 contains discussions about our results and conclusions along with future work.

2. METHODS

2.1 Participants and Materials

Our research data is collected via one of the MHE assessment platforms called Connect (`http://connect.mheducation.com`). Connect is a higher education web-based assessment and assignment platform, which provides students an online environment to do their coursework and logs user activity in order to provide feedback and support to its user needs.

In this paper we explored 100 million instances of user login data obtained from Connect between June of 2013 and June of 2014. For this analysis, we used data such as students' login dates, total number of logins, number of attempts on an assignment and assignment score. Depending on the analysis, some of these data were aggregated based on time or grouped by the unique students.

2.2 Procedures and Methods

To extract the necessary data for our analyses, we used Oracle's procedural language extension for SQL (i.e., PL/SQL) [4] and Python programming language [13], along with the necessary Python libraries to query, wrangle, clean, plot, and explore our login data. Our data contains the following attributes: student related data (e.g., student ID, student logins) and assessment related data (e.g., number of attempts, assessment score, number of attempts).

3. LOGIN BEHAVIOR ANALYTICS

3.1 Login Behavior

In this section we investigated the trends related to student logins. Figure 1 visualizes the overall pattern of student logins over the days of the week. The red line shows the average number of logins for any given day. This analysis validates the expected pattern of decreasing activity on Saturdays and increasing activity on Sunday evenings. This shows students' tendency to stay away from their homework assignments on the weekend until late Sunday when they attempt to prepare for the week. This finding is not surprising, in fact, it confirms the intuitive expectation of student academic activity on weekends vs. weekdays. If investigated further (i.e., A/B testing), this information could provide a basis for notifying students with customized and timely recommendations via Connect.

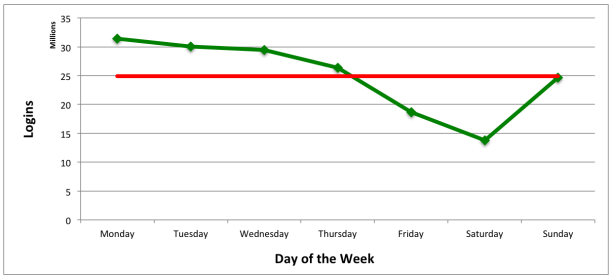

Figure 1: Logins by the day of the week. X-axis = Day of the week from Monday to Sunday; y-axis = Logins (in millions).

Next, in Figure 2 we investigated the number of logins per day. While the overall pattern of logins increasing in Fall through Spring and decreasing in Summer seemed very reasonable, the significant spike in Spring of 2014 seemed out of ordinary. To understand this unusual pattern, we requested more information from the Connect marketing team who explained that the spike in the Spring of 2014 is congruent with the new marketing effort making Connect assignments mandatory portion of students' coursework. This finding provided a data grounded confirmation of Connect team's marketing efforts.

4. PATTERN MINING & STUDENT PROFILING

For the analyses in this section, we used the average number of logins per assignment (henceforth logins), the average score per student (henceforth score), and the average attempt per assignment (henceforth attempt). In this section, we present our analysis of comparing the student login data with students' scores on assignments.

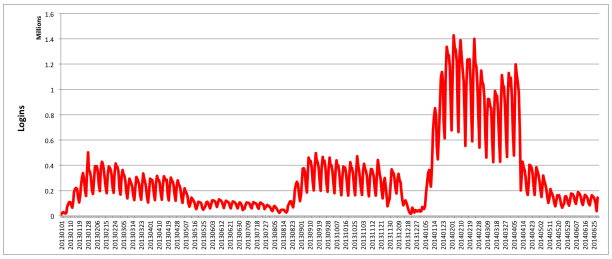

Figure 2: Logins by the month. X-axis = Days in months from 01/01/2013 to 06/25/2014; y-axis = Logins (in millions).

4.1 Login vs. Score Trends

To continue our data explorations, we decided to further investigate the potential patterns in the student login and student assignment score data.

4.1.1 Data Preparation

For this analysis, we looked at a total of 1.5 million users' assignments scored between June 2013 and June 2014. For each user, score, login and total number of attempts were normalized against users' total number of activities. Further, we eliminated some of the outliers by excluding the users with 1 or no attempts and eliminated users with more than average 50 logins which removed 100,000 users' data. On average, students login 5.5 times, have 1.03 attempts and have a score of 53% per activity.

4.1.2 Data Analysis

We plotted student logins per assignment vs. student's median score (see the green line in Figure 3). In this plot, we used the median score instead of the mean of the scores in order to account for the high variability of the distribution of scores. This figure shows that student median score grows as the number of logins increases. However, after a certain threshold, the score tends to decrease as the number of logins per assignment increases, thus showing the counterproductivity of the login activity. This contradicts to the intuitive assumption that more logins result in a better academic performance.

To further explore the relationship between login and scores, we performed a piecewise linear regression to identify possible segments in the data. Fitting a single regression line, the standard error (SE) of estimate with one regression line was σ_{est}=18. The SE for a model with two regression lines resulted in σ_{est}=12.5. We also tried fitting three regression lines through, which resulted in a higher SE of σ_{est}=16.8. Therefore, we used a model with two regression lines (see Figure 3). This resulted in a break at i=4 (i.e., Segment 1 = 0:4 and Segment 2 = 5:50). This suggests two distinct segments in the data. In the first segment, as the number of logins increase, the performance improves (slope = 6.48; correlation = 0.99). However, after a certain threshold, 4 logins, the scores plateaus, and gradually decrease as the logins increase (slope = -0.45; correlation = -0.93). This hypothesis is further explored in the next section through cluster analysis.

4.2 Student profiling

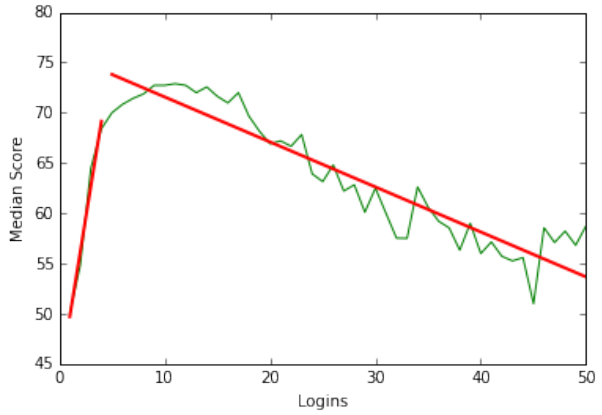

Figure 3: Piecewise linear model. X-axis = Number of logins per assignment; y-axis = Median score.

4.2.1 *K-Means Clustering Method*

Following the hypthesis formed in the previous section, we explored student login patterns through k-means clustering. In k-means clustering, data is partitioned into k clusters where each observation is assigned to the cluster with the nearest mean ([6]). The clustering process starts by choosing k random observations as initial cluster centroids. Thereafter, each observation is assigned to the nearest centroid and the new centroids are recalculated using the average of the data points in each cluster. We selected Euclidean distance as the distance metric in k-means clustering ([5]) where within-cluster sum of squares (hereafter, WCSS) is the cost function. Representing the data as a set of N observations $\{x_1, x_2, , x_n\}$, where each observation is a D-dimensional vector of D attributes, k-means clustering partitions N observations into k clusters $\{c_1, c_2, , c_k\}$ where WCSS is minimized as:

$$argmin \sum_{k=1}^{K} \sum_{X \in c_k} \parallel X - \mu_k \parallel^2$$

where μk is the mean of points in c_k. To accommodate the scale of our dataset, we have selected k-means clustering method due to its computational speed and efficiency compared to hierarchical clustering. In addition, k-means clustering is a robust approach, which results in non-overlapping clusters that are very easy to interpret. We have used the Elbow method ([12]) to identify the optimal number of clusters. In this method, average WCSS is measured as the number of clusters increase. Having more clusters results in smaller distances from centroids and hence a smaller average WCSS. However, the amount of drop is not constant as the number of clusters increase and the decrease in average WCSS flattens at a certain k value. This value, called the elbow metric, creates a break in the elbow graph and is a good measure for identifying optimal number of clusters.

4.2.2 *Clustering Results*

In this analysis, we used the same data aggregations for students' login, score and attempts as described in the beginning of this section to explore student groupings according to their login behavior. The elbow method is used to decide an optimum number of clusters. Figure 4 shows the average WCSS value as the number of clusters increases from 1 to 9. The graph nearly flattens after k equals to three, thus suggesting 3 as the optimal number of clusters.

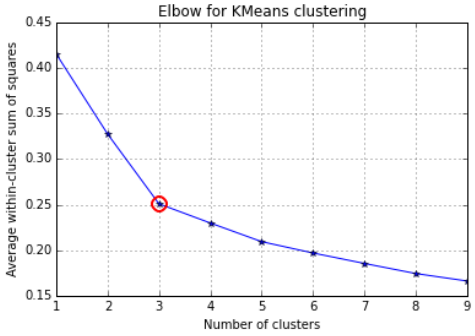

Figure 4: Elbow metric. k=3; x-axis = Number of clusters; y-axis = Average WCSS.

We used Scikit-learn python library ([10]) to implement k-means clustering. Figure 5 shows a 3D scatter plot of the three attributes used to cluster the data where the data points are colored by the cluster labels. Figure 5 shows

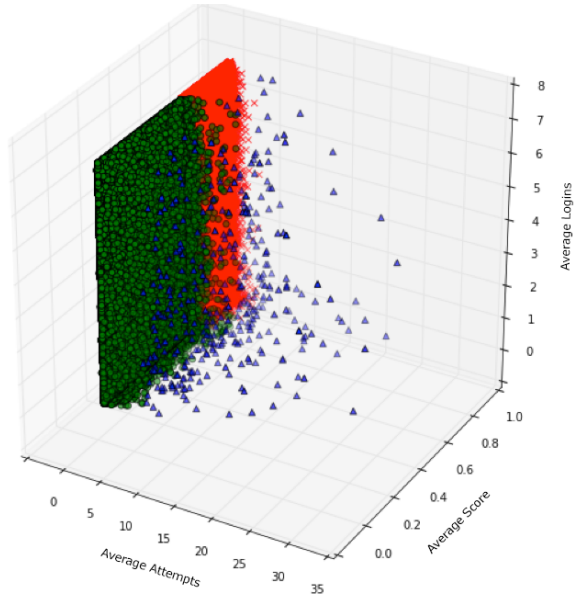

Figure 5: 3D scatter plot. Cluster 1 (red) = High Achievers; Cluster 2 (green) = Low Achievers; Cluster 3 (blue) = Persistent Students; Attempts = x axis; Logins = y axis; Score = z axis.

three sets of distinct student login profiles. The Cluster 1 (red), whom we label as *High Achievers*, represent a group of students with a low number of attempts, a medium number of logins, and a high score. The Cluster 2 (green), whom we label as *Low Achievers*, is the group with a medium number of attempts, and low number of both logins and score. Finally, the Cluster 3 (blue), whom we label as *Persistent Students*, is the most distinct group with a high number of both attempts and logins, and a medium score. To quantify

this information, in Table 1 we have tabulated the count, the mean and the standard deviation of these three attributes across each of the three clusters. In addition, we have simplified this content in Table 2.

Table 1: Cluster Statistics. Total = number of observations. SD = standard deviation.

		Average Attempts	Average Score	Average Login
Cluster 1 (High Achievers)	Count	1097675	1097675	1097675
	Mean	1.03	0.84	8.51
	SD	0.19	0.37	58.07
Cluster 2 (Low Achievers)	Count	780405	780405	780405
	Mean	1.05	0.24	6.03
	SD	0.25	0.17	21.33
Cluster 3 (Persistent Students)	Count	1220	1220	1220
	Mean	9.82	0.56	35.65
	SD	6.83	0.32	62.13

Table 2: Student groups based on cluster statistics.

	Login	Attempt	Score
High Achiever	Medium	Low	High
Low Achiever	Low	Medium	Low
Persistent Student	High	High	Medium

Table 2 shows that Cluster 1 (high achievers) includes students with the highest score among the three clusters. Low achievers, Cluster 2, stand out with a very low score and a low number of logins. This shows a relationship between the low logins and the low performance scores in students with very high or very low scores. However, students with medium score have very high average logins and high average attempts per activity. This fluctuation between average score and login indicates a non-linear and non-trivial relationship between student behavior (number of logins and attempts) and performance.

5. CONCLUSION & DISCUSSION

In this paper we explored student login data collected from MHE's Connect higher education platform. The investigation of student login activity reveals a non-linear relationship between student activity and performance. Piecewise linear regression revealed that students who do better on their assignments tend to login more. However, if a student logs in 5 or more times per assignment, their performance tends to plateau and then deteriorate. Thus, it would be beneficial for the instructor to intervene at this point as it might indicate that the student has not grasped the concepts required for the assignment. Finally, investigating student login behavior led to identifying three distinct groups of students: high achievers who login just optimum number of times to get high score, low achievers, who login very rarely and tend not to do well, and persistent students who show grit in their efforts to succeed by logging in and attempting the most but still perform less than high achievers. The educational value of such finding is in identifying and encouraging certain activity behaviors that are correlated with good performance.

Future work will be concentrating on factors such as the variability in the students' scores based on the due date of the assignments, time spent on assignments, potential recommendations or instructors actions and effectiveness of these recommendations via A/B testing. Finally, we will be attempting to join students academic performance gathered from Connect to their performance or other institutional or demographic data in order to predict student academic success.

6. ACKNOWLEDGEMENTS

This paper is based on work supported by McGraw-Hill Education. We would like to extend our appreciation for all the informational support provided by the Connect Team, the research support provided by the MHE CDO Stephen Laster, and the Analytics team at DGP. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect positions or policies of the company.

References

[1] N. E. Commander and B. D. Smith. Learning logs: A tool for cognitive monitoring. *Journal of Adolescent & Adult Literacy*, 39(6):446–453, 1996.

[2] K. Cotton. Classroom questioning. *School improvement research series*, 3, 2001.

[3] R. DuFour. Professional learning communities. 1998.

[4] S. Feuerstein and B. Pribyl. *Oracle pl/sql Programming.* " O'Reilly Media, Inc.", 2005.

[5] J. C. Gower. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.

[6] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[7] K. Leithwood, K. Seashore Louis, S. Anderson, K. Wahlstrom, et al. Review of research: How leadership influences student learning. 2004.

[8] R. Mazza and V. Dimitrova. Visualising student tracking data to support instructors in web-based distance education. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 154–161. ACM, 2004.

[9] M. Muehlenbrock. Automatic action analysis in an interactive learning environment. In *The 12th international conference on artificial intelligence in education, AIED*, pages 73–80, 2005.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[12] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[13] G. Van Rossum and F. L. Drake. *The python language reference manual.* Network Theory Ltd., 2011.