Study on Learnta Customer Success

Data Challenge B

Consultants: Wei Zhang, Yi Wang, Tianxing Ma, Shaocong Peng, Zhu Wang

Clients: Will Ma, Feifang Hu

May 6th, 2019

Abstract

OBJECTIVE The project aims to analyze relationship between engagement in Learnta system and learning outcomes which will be based on to provide statistical methods to measure the customer health and active degree based on the given data. The last but not least, making Learnta school member more active will eventually help students have success in study.

METHODS We calculate the usage frequency and using time, then combine these two measures for each organization as health score to quantify the degree of activity. To define the ability of Learnta maintain the customers, we combine the customer maintenance ability and the size of each organization as vital signs. To quantify the progress of students, we define the difficulty of each question as correction rate among all students who have solved this question.

RESULTS We find that 40% of Learnta member schools health score exceed fifty and most of member schools vital sign are around fifty. That means most Learnta schools will be willing to renew their membership and we could see their active degree is not bad.

CONCLUSION Learnta has spared no efforts to provide a good system and make the member school active, we could see that lots of members keep active and would like to renew their membership in the next quarter. Meanwhile, we also give some actionable advice to Learnta to improve the system so that can help students keep learning in the system.

1 Introduction

Customer Success is essential for Software as a service (SaaS) company, especially for the company like Learnta, who want to expand their software services into the field of education. Our customer success project aims to analyze the behaviors of Learnta organizations and provides statistical methods to measure the customer health and active degree based on the given data.

We divided our project into four parts: i) Firstly, we did a descriptive statistical analysis to get a glimpse idea of the 20 Learnta organizations; ii) Secondly, we come up with the ideas of how we calculate the health scores and active degrees for each organizations; iii) Moreover, based on the analysis of learning progress and time spent on using Learnta system for each organization, we evaluated the learning outcome of using Learnta system; iv) To conclude our results, we built a user interface of customer dashboard to track the behaviors of organizations.

2 Background and Dataset

Learnta was established in New York and Shanghai in 2016. It is a company of learning technology and analytics, focusing on "artificial intelligence + education". Based on the Chinese primary and secondary school curriculum system, the Learnta teams design and develop artificial intelligence algorithm engines, learning systems, course content and teaching modes to provide integrated intelligent learning solutions for primary and secondary institutions. In our study, we have a sample of 20 Learnta organizations (See Appendix 2.1).

The datasets of this study were provided by Learnta. The first dataset has 7 variables, mainly provides the information of login and logout time for individual users. It contains the id number, organization id, user's id, classroom id, status (1=enter classroom; 2=exit classroom; 3=log in; 4=log out), user's type(1=student; 2=teacher) and time. The second dataset is a supplement data with 22 variables. Each row of this dataset represent the information of student's behavior when doing questions with the system. It contains the information of the questions, and whether the student did the question correctly. For the first dataset, we have a total of 270337 observations. And for the second dataset, we have 186674 observations in total.

3 Methodology

3.1 Health Score and Active Degree

In order to quantify the degree of activity in using Learnta system, we calculate the usage frequency and using time, then combine these two measures for each organization as health score (figure 3.1). We define usage frequency as the average percentage of weeks for each user to use

Learnta system, that is, how much percentage of weeks among these 16 weeks, one user has login activities in the system on average. For each organization, we calculate the usage frequency by dividing the average number of weekly active users by the number of total users. For each organization, we use the average weekly spent in Learnta system per person as the using time. We scale both the using time and usage frequency so that these two set of values have the same mean and the same variance. We calculate the average of the two sets of scales values and then scale the average to (0,100) as the health scores.

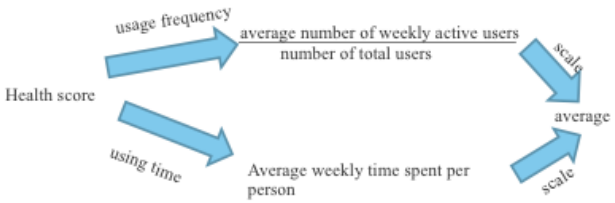

Figure 3.1 Workflow of health score calculation

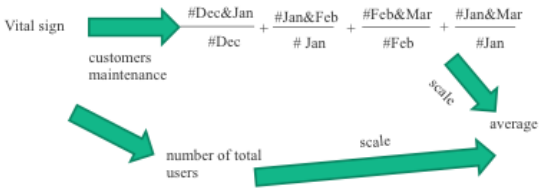

Figure 3.2 Workflow of vital sign calculation

To evaluate the ability of Learnta system to maintain active users in each organization and to predict their future usage, we calculate the vital score for this evaluation. We calculate the customer maintenance ability and the size of each organization and combine these two measures as vital sign (figure 5-3). We define customer maintenance ability as the average percentage of active users in each month that keep using Learnta system in the next one or two months. For example, we first calculated the number of active users of December and then calculated how many of them still use Learnta in January. We then calculated the total number of users of each organization. We use the same methods as the previous section to scale and combine the two sets of values to calculate vital signs.

3.2 Student Progress Analysis

To quantify the progress of students, we define the difficulty of each question as correction rate among all students who have solved this question. We define student's score in each question as the correctness (1:correct; 0: incorrect) subtract the difficulty of question. We then measure the ability of student at the beginning using the average score of the first 20% questions, and the ability of student at the end using the average score of the last 20% questions. The improvement of each student is therefore measured as the ability at the beginning minus the ability at the end. We conducted two sample t-test for the equality of average ability of students of each organization.

4 Data Analysis and Results

4.1 Overall Study

Active User Number

We calculated the frequency with duplication since users might login several times in different month. There is only one day in November 2018 in the data, so the number of users in this month is relatively low. From the Table 4.1 we can see, 2018.12 and 2019.1 have nearly the same number of active users, but there is decreasing in 2019.2.

Table 4.1 Monthly Active Users

Month	2018.11	2018.12	2019.01	2019.02
Number	6	527	524	300

From Figure 4.1 the 6th week and 14th week are at a relatively low level (14th week is not a complete week). Other week maintain a relatively stationary level. From Figure 4.2, the day with the largest number of users is 2018-12-09; The days with the minimal number of users is 2019-02-04~2019-02-08. What's more, data fluctuates greatly during different time periods and seems like exist a periodic.

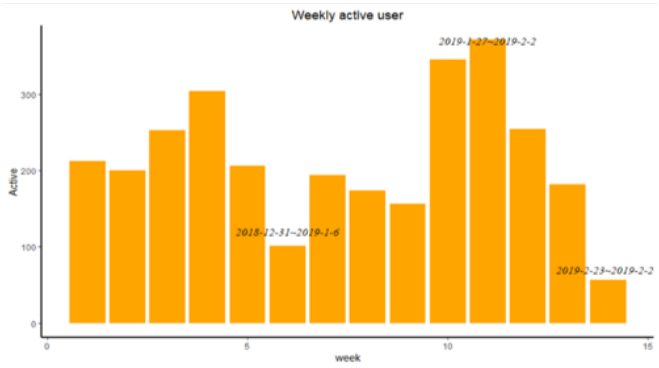

Figure 4.1 Weekly Active Users

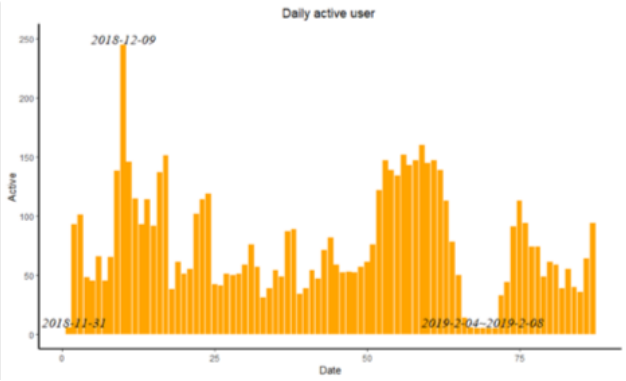

Figure 4.2 Daily Active Users

From Figure 4.3, the number of active students and teacher have similar trend, but the number of students is always higher than teachers'.

Days from open to active

To find out how much time needed for each organization to have frequent users' activities, we calculated the days before activation for each organization. Since we only have the data after November 2018, we excluded the organization that opens before November 2018. We

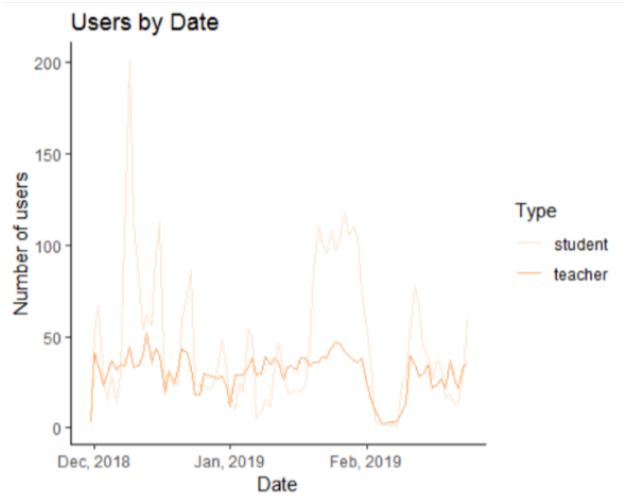

Figure 4.3 Monthly Active Users

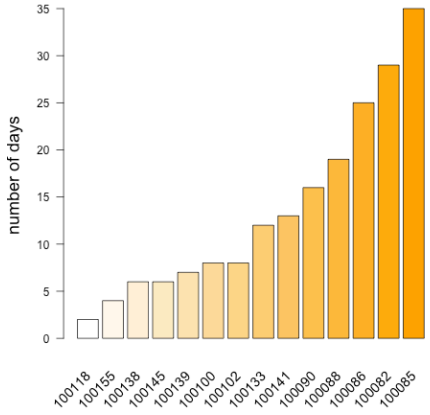

Figure 4.4 Days from opening to activation

also excluded the organization showing login activities before the days of opening. We calculate the days between opening and obtaining more than 10 daily login activities (figure 4.4). We find that the days before activation vary from 2 to 35 among these organizations. We find organization 100118 takes the shortest time to activate, while 100085 takes the longest time to activate.

The Trend of Correction Rate

The correction rate for every organization can represent the progress in some degree. From the Figure 4.5 we can see: 1) the trends for Math and English are always inverted; 2) there are no significant trend both in Math and English, in other word, when we use correction rate to determine the progress for organization, Learnta system may not have an obvious effect.

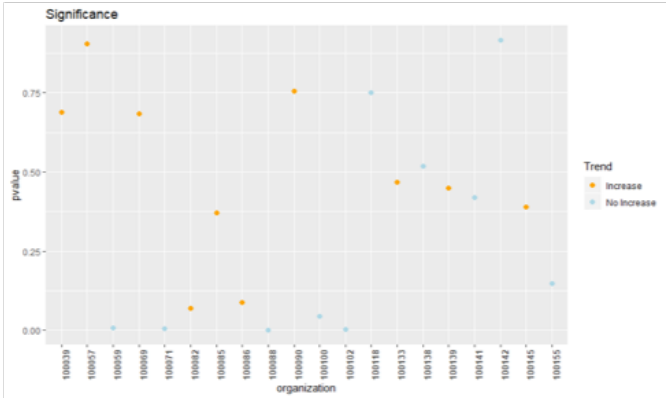

Figure 4.5 Correction Rate Trend

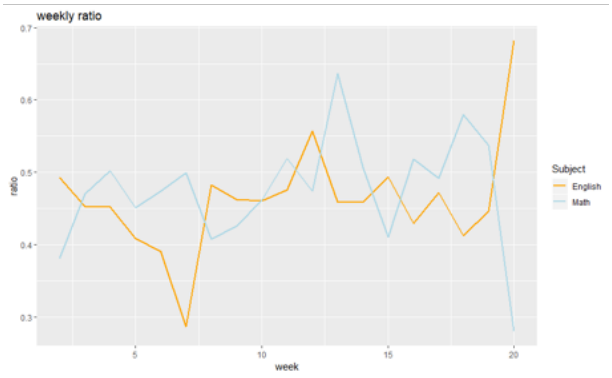

Figure 4.6 Trend Test

In Figure 4.6 orange point represent the organization which have an increase trend and the light blue point represent the organization which have a decreasing trend or have no trend.

From the plot we can see that half of the organizations have an increasing trend. Therefore, in other word, half of the organizations have some degree of progress after using Learnta system.

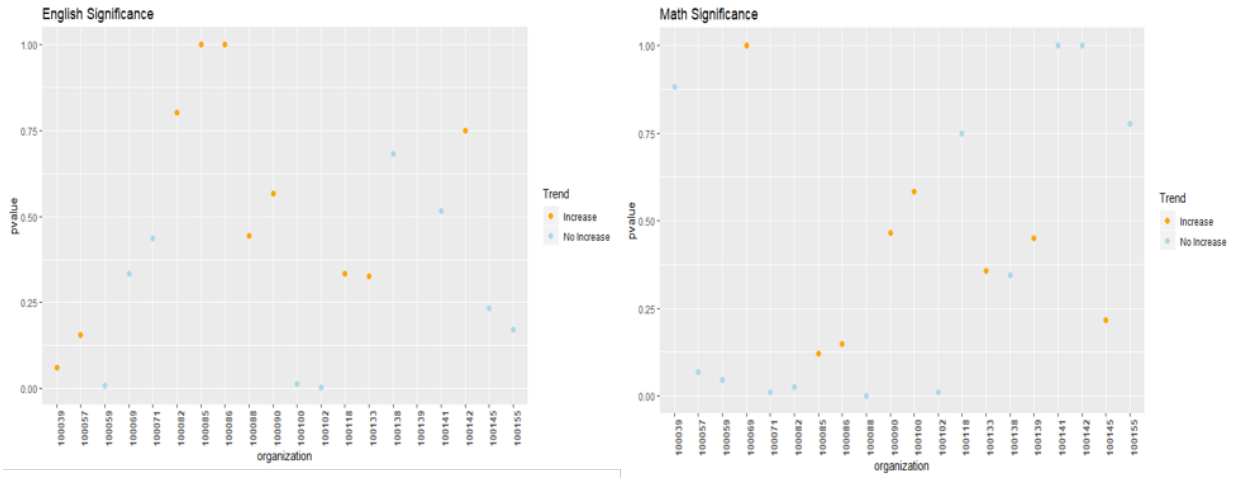

Figure 4.6 Trend for Different Subject

Then we analyzed the relationship between correction rate and time spending on Learnta system. From Figure 4.7, we can find that the spending time for each student are different. And the correlation between correction rate and spending time are significant positive (correlation=0.0989, p-value<0.001). In other word, we can assume, for a number of students, if they spend more time on Learnta system, they will have a higher correction rate which in some degree represent a progress for students.

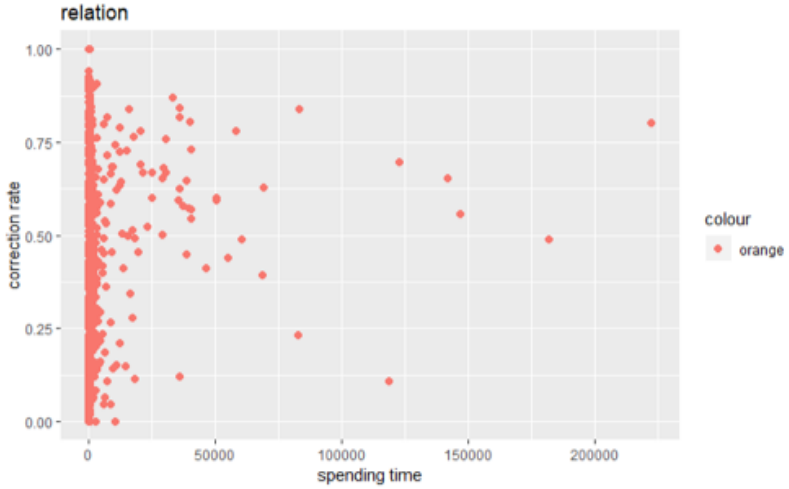

Figure 4.7 Correction Rate with Time Spent

4.2 Case Study: A Deeper Analysis for Organization 100082

This organization is located in Shandong province. It contains 66 students and 26 teachers. The ratio of students and teachers for this organization is 2.54, which is a common ratio for students and teachers. Moreover, this organization used the Learnta system in the whole period of the data. Therefore, we choose this organization to do the deeper analysis.

Active User Numbers

According to the Figure 4.8, there was a decreasing trend of using Learnta system from second month to the fifth month; and December 2018 have the most active users, February 2019 have the least active users. What's more, when we split the teacher and students, we find there were more active students for this organization which is normal in all the school. However, from Figure 4.9, we can find that the number of teachers' log in times is almost the same as the number of students'. That means, averagely, teachers log in more than students.

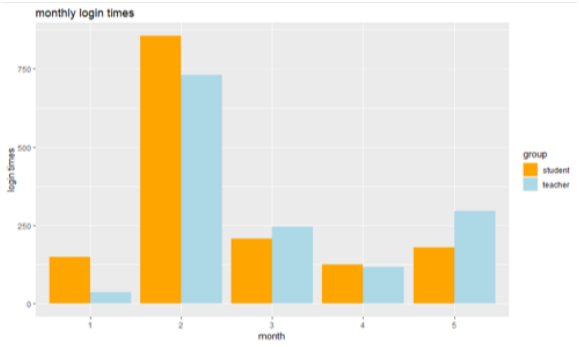

Figure 4.8 Monthly Active Users

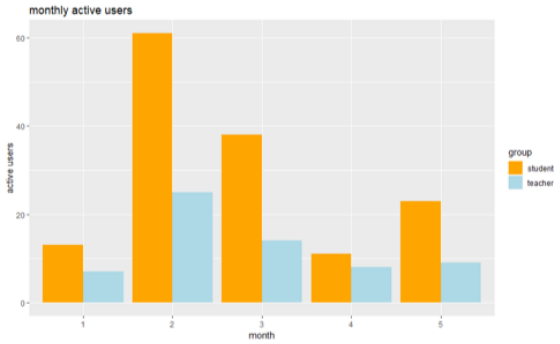

Figure 4.9 Monthly login Times

From Figure 4.10, the week started from 2018 12.9 have the most active users, and except this week the login times for every was almost stationary. The eleventh, twelfth, thirteenth and fourteenth weeks, or in other word, the entire February have the least active users because of the Spring Festival, but there is a small increase around the week started at the beginning of February. And from Figure 4.11, we can see there is an obvious periodicity time series trend. January and February have a lot of days with no log in which may due to the winter break. And we also can assume that 100082 have a decreasing trend of using Learnta system, because even after the Spring Festival, the number of active users did not return to the prior level.

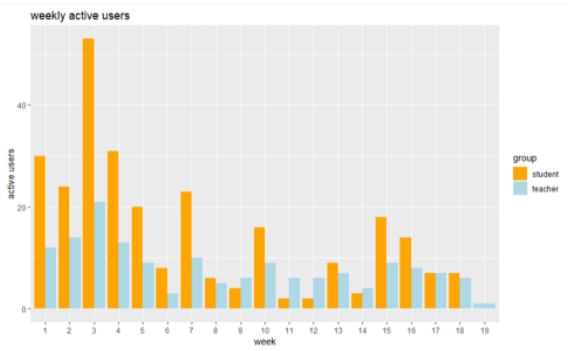

Figure 4.10 Weekly Active Users

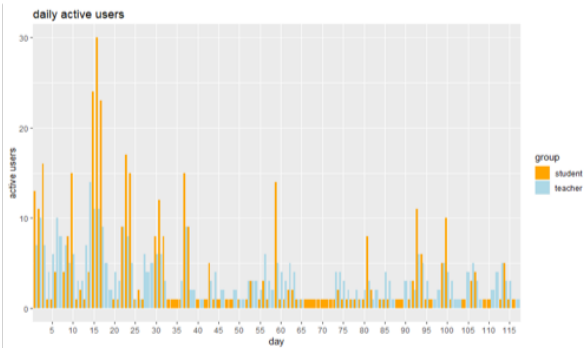

Figure 4.11 Daily Active Users

Frequency of Using Learnta System

To find out the daily usage activities of organization 100082, we calculated the daily login numbers of teachers and students in this organization. Figure 2 summarized the daily login number from the open date of this organization to the end of February. From figure 4.12, we find that the login frequencies in the beginning month are higher than the login frequencies in the flowing months. In addition, there are some outliers in the first month. Students' using frequency are generally higher than the using frequency of teachers.

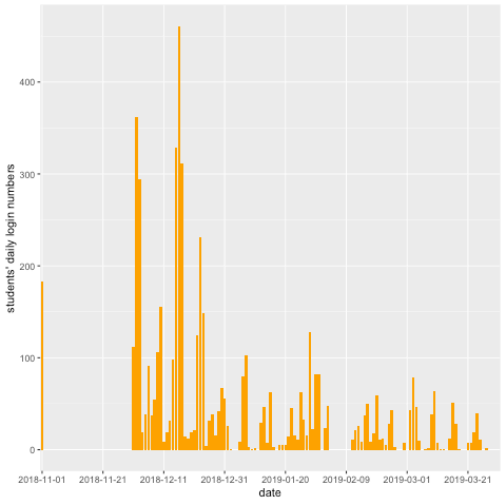

Figure 4.12 Daily login numbers of 100082

Take a closer look at the hourly frequency, we found out that users were more frequently using Learnta system at night. As figure 4.13, users frequently logged in the system during 19 pm to 3 am in December 15. Similarly situation also happened in other days. This situation might tells us that students or teachers were more likely to use the system as homework or supplements of study after class.

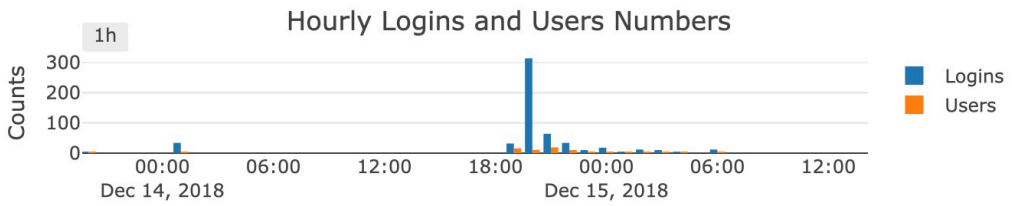

Figure 4.13 Hourly login and users of 100082

Time Spent on Learnta System

To see how users spent time on Learnta system, we plotted daily and weekly time spent plot. From the figure 4.14, we can see that except the beginning of February, teachers and students all spent some time on using Learnta system. Teachers spent more time than students on using the system. After February, we can clearly see that the daily average time spent on the system were decreased for both teachers and students. In the weekly average time spent plot, the trend of time spent on the system is clearly decreasing.

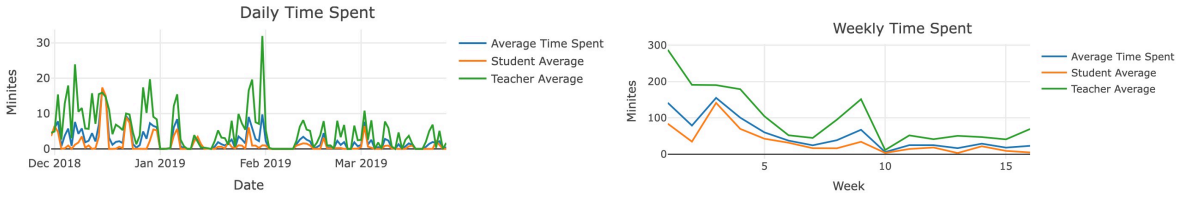

Figure 4.14 Time spent on Learnta system of 100082

When splitting the in class and after class time spent on Learnta system, we can see that for the Organization 100082, the in class and after class learning time is almost equal.

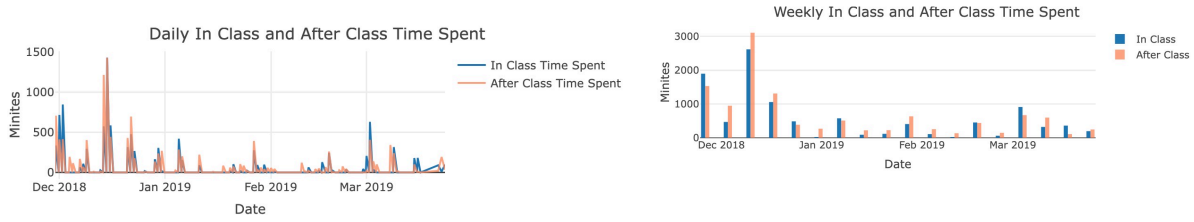

Figure 4.15 In class & after class time spent on Learnta system of 100082

Relationship Between Correction Rate and User Time Spent on Learnta

In this part we focus on one of the organizations and figure out the relationship between the correction rates and the spending time. We also divided the spending time into two part: in

class and out class. And we tested the relationship between spending time and correction rate for all the students in this organization. Since the sample size for each type of questions is generally small, we used the correction rate directly to calculate the improvement. There is no significant relationship between the entire spending time and correction rate. And there is also no relationship between correction rate and spending time in class. However, correction rate has a significant positive relationship with spending time out class (Figure 4.16). Which means, when the students in this organization spending more time after class, they will have a higher correction rate. And the scatter point plot is shown as below.

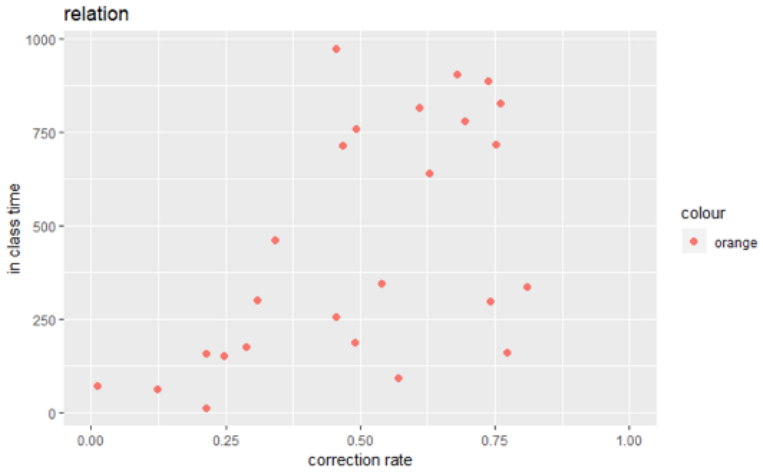

Figure 4.16 Correction Rate with Out Class Time Spent

5 Conclusion and Suggestion

5.1 Conclusion

Health Score and Vital Sign

The health score of each organization is shown in figure 4.17. We conclude that organizations 100039,100102,100085 have the highest health score, and organizations 100155,100133,100069 have the lowest health score. The vital sign of each organization is shown in figure 4.18. We find that organizations 100059,100057,100039 show the highest vital sign, and organizations 100142, 100118, 100139 have the lowest vital sign.

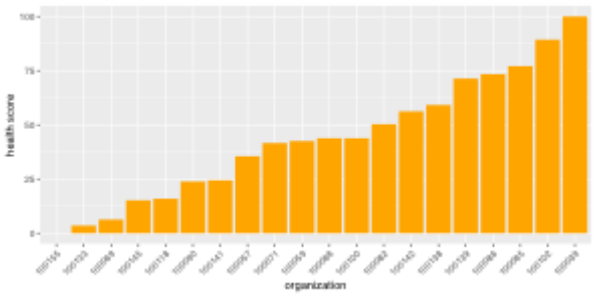

Figure 4.17 Health scores of each organization

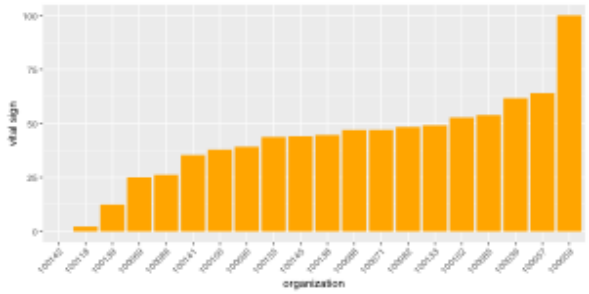

Figure 4.18 Vital signs of each organization

Overall Correlation Between Learning Outcome and Question Numbers

We conducted two sample t test for the equality of average students' ability of each organization. We found that there is not significant improvement in most organizations (table S2). Through calculating and testing the correlation between the improvement and number of questions solved, we find that such relationship is not significant for most organizations (table S3).

5.2 Suggestion

From previous research, we find that the active users vary from weeks to weeks, some students or teachers may not use the system for several days. To solve this issue, we can set a threshold time that not using the system continuously, if the users do not login the system for a long time(exceed the threshold time) , the students will receive a reminder through the application, and the teacher will receive the texts through their phones.

The second suggestion is that Learnta can update the questions in low valley period. We can see that during the final exam period, there are more active users than usual. Besides, there is a peak for every four weeks. We do not need to concern about the weeks with high active degree, while the low valley period needs more attention. So we recommend that developers could update some new questions or set some contests in the low valley period to encourage students to use the system.

The last but not least, we have already calculated the health score and active score for each organization, so we can combine the scores as "Effort Scores" to present in the summary report every week as references to teachers. Based on the health score and active score, we can provide the same evaluation for every student. If the student has better active degree than last week, the system can show them that they have a good performance this week. At the same time, the system can automatically send texts to their parents to reward their child.

Appendix

Table S1 Information of Organizations

orgId	Teacher	Student	Total	Adjusted Total	Accesstime	Condition	Location
100039	10	38	48	48	2018/6/15	existing	Jiangsu
100057	41	43	84	83	2018/8/13	new	Jiangsu
100059	14	289	303	300	2018/8/24	existing	Hunan
100069	10	1	11	11	2018/9/29	existing	Jiangsu
100071	17	26	43	43	2018/10/8	new	Henan
100082	26	63	89	89	2018/11/3	existing	Shandong
100085	4	9	13	13	2018/11/5	new	Shandong
100086	8	25	33	33	2018/11/7	existing	Shanghai
100088	17	27	44	43	2018/11/14	new	Yunnan
100090	12	33	45	43	2018/11/16	existing	Hunan
100100	6	12	18	17	2018/11/27	existing	Henan
100102	6	31	37	36	2018/11/30	existing	Shandong
100118	7	6	13	13	2018/12/7	existing	Zhejiang
100133	12	15	27	24	2018/12/20	new	Henan
100138	6	18	24	24	2018/12/26	existing	Hebei
100139	2	1	3	3	2018/12/26	new	Hunan
100141	3	0	3	3	2018/12/27	existing	Shanghai
100142	5	8	13	13	2018/12/29	existing	Sichuan
100145	4	0	4	4	2019/1/2	existing	Anhui
100155	10	6	16	12	2019/1/17	existing	Shandong

Table S2 Change of students' abilities

Organization	begin	end	p value
100086	0.007	-0.039	0.21
100039	-0.007	-0.025	0.37
100059	0.029	-0.053	0.00
100057	-0.049	-0.015	0.11
100069	-0.045	-0.001	0.29
100071	-0.027	0.019	0.26
100082	0.010	-0.019	0.21
100090	0.006	-0.016	0.45
100085	0.263	0.165	0.11
100088	0.008	-0.012	0.67
100102	-0.022	0.001	0.06
100100	-0.002	-0.004	0.97
100118	-0.078	0.006	0.11
100139	0.060	-0.117	0.53

100142	-0.379	-0.109	0.02
100138	0.031	0.058	0.52
100133	0.042	0.049	0.83
100145	-0.018	-0.015	0.89
100155	0.003	-0.043	0.13
100141	0.175	0.122	0.48

Table S3 Correlation of learning outcome and question numbers

organization	correlation	p value
100086	-0.09	0.60
100039	0.04	0.70
100059	-0.01	0.91
100057	-0.11	0.15
100069	0.69	0.31
100071	-0.26	0.06
100082	-0.07	0.43
100090	-0.01	0.90
100085	-0.31	0.31
100088	-0.09	0.58
100102	-0.08	0.07
100100	0.04	0.75
100118	0.21	0.22
100139	0.26	0.83
100142	0.36	0.64
100138	-0.56	0.01
100133	-0.03	0.77
100145	-0.08	0.51
100155	-0.22	0.02
100141	0.05	0.83