Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220095

The Relational Model Between College Students' Internet Behaviors and Academic Performance

Hanyun LUO[a], Yong SONG[b,1], Jiadong DONG[c], Xiaoling ZHANG[a] and Guoping NIE[a]
[a] *School of computer and information, Anqing Normal University, China*
[b] *School of Foreign Languages, Anqing Normal University, China*
[c] *School of electronic engineering and intelligent manufacturing, Anqing Normal University, China*

Abstract. Different internet behavioral habits have different impacts on the studies and life of college students. Using scholarship winners and students who were retaking courses after failing to pass exams as our research samples, we obtained a massive amount of data for these research samples from the internet surfing authentication system and internet behavior audit system within seven months. After data-masking, we used a binary classification logistic equation to establish a regression model of the relationship between college students' internet behaviors and academic performance. We applied SPSS to establish a mathematical model to calculate the duration of internet surfing and browsing content to analyze the internal relationship between college students' internet behaviors and academic performance. We designed an academic performance early warning mechanism for institutions of higher education to enhance evidence-based and targeted decision-making.

Keywords. Internet behaviors, academic performance, logistic regression model.

1. Introduction

Smart campuses supported by the Internet of Things, cloud computing, and virtualization characterized by high perception, strong collaboration, and strong service capabilities have grown significantly. With this enthusiastic promotion of information teaching methods, including smart classrooms and Rain Classroom, and the further popularization of mobile terminal devices, such as mobile phones, tablets, and laptops, among college students, internet-based learning has become an important part of college education and teaching. These methods significantly enrich and influence the learning of college students. Different internet behavioral habits among college students, however, affect the different ways in which these students learn. On the one hand, internet-based learning enriches the learning content of college students and broadens the learning approaches of college students, which helps college students acquire and accumulate knowledge and develop and cultivate personality. On the other hand, college students who have been using the virtual internet world for a long time are becoming disconnected from the real world and are affected by illegal as well as harmful information. These college students

[1] Corresponding Author, Yong SONG, School of Foreign Languages, Anqing Normal University, China; E-mail: songyong@aqnu.edu.cn.

are prone to developing online behavior anomie, which can seriously affect the ability of college students to learn [1–8]. We used two primary methods to study the behaviors of internet users: questionnaire surveys and an investigation conducted with a computer-aided telephone interview (CATI) system.

We used data that originated from the log files of the university's internet surfing authentication system and internet behavior audit system. The data used in this empirical study on the relationship between college students' internet behaviors and academic performance truthfully recorded college students' internet behaviors without these students being affected by psychological factors. Such data are different from data that are obtained from telephone interviews and questionnaire surveys. After data-masking, we used a binary classification logistic equation to establish a regression model of the relationship between college students' internet behaviors and academic performance. We tested the goodness of fit of this model with the sample data to explore the correlattion between college students' internet behaviors and academic performance from the perspective of big data.

2. Logistic Regression Model

In this study, we performed a comparative analysis on two groups of subjects: (1) National Scholarship and National Encouragement Scholarship winners, and (2) students retaking exams. In this way, the dependent variable Y could be regarded as a binary classification. If $\mathrm{Y} = 1$, the student had won a scholarship; and if $\mathrm{Y} = 0$, the student had retaken exams. The independent variables, $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3 \ldots \mathrm{x}_n)$, denoted college students' duration of internet surfing and internet behaviors, such as browsing websites, accessing forums and microblogs, watching movies, and chatting with people. We transformed the actual problem to be studied into an exploration of the probability p of college students to win scholarships as well as the relationship between p and internet behaviors. A binary classification logistic linear regression model [9–16] was used, as follows:

$$\mathrm{Logit}(p) = \ln\frac{p}{1-p} = \mathrm{b}_0 + \mathrm{b}_1\mathrm{x}_1 + \cdots + \mathrm{b}_n\mathrm{x}_n \,, \quad (1)$$

when $p \to 0, \mathrm{Logit}(p) = \ln\frac{p}{1-p} \to -\infty,$

when $p \to 0.5, \mathrm{Logit}(p) = \ln\frac{p}{1-p} \to 0,$

and

when $p \to 1, \mathrm{Logit}(p) = \ln\frac{p}{1-p} \to +\infty.$

Through this logistic regression model, we converted the value range of the dependent variable Y to [0, 1], and the value range of $\mathrm{Logit}(p)$ was the entire real number field with 0 as the symmetry point. There was a probability p corresponding to each value of the independent variable. Equation (1) could be transformed to obtain Equations (2) and (3), as follows:

$$p_1 = \frac{\exp\,(\mathrm{b}_0 + \mathrm{b}_1\mathrm{x}_1 + \cdots + \mathrm{b}_n\mathrm{x}_n)}{1 + \exp\,(\mathrm{b}_0 + \mathrm{b}_1\mathrm{x}_1 + \cdots + \mathrm{b}_n\mathrm{x}_n)}, \tag{2}$$

$$p_2 = 1 - p_1 = \frac{1}{1 + \exp\,(\mathrm{b}_0 + \mathrm{b}_1\mathrm{x}_1 + \cdots + \mathrm{b}_n\mathrm{x}_n)}, \tag{3}$$

where p_1 is the probability of a student winning a scholarship and p_2 refers to the probability of a student retaking an exam. By establishing a mathematical model based on the internet behavior data of the samples, we determined the regression coefficient of the equation and then calculated the probability using this equation. The closer p_1 was to 1, the closer p_2 was to 0, which meant the student was in good learning conditions and was very likely to win a scholarship. On the contrary, the closer p_1 was to 0, the closer p_2 was to 1, which indicated the student was in poor learning conditions and probably would fail to pass the exams.

3. Research Subjects and Sample Data Collection

3.1. Subjects

More than 20,000 students are in AnQing Normal University. Every year, a certain number of students win the National Scholarship and National Encouragement Scholarship. Only less than 4% of college students can receive a National Scholarship and National Encouragement Scholarship after several rounds of selections. All the scholarship winners stand out among the top students. Studying diligently and working hard to make progress, these students develop comprehensively in morality, intelligence, sports, aesthetics, and labor education and develop enviable academic performances. In contrast, many students have to retake exams. Failure in passing the exams is a direct reflection of poor academic performance. We asked whether the internet behaviors of the two types of students with significantly different academic performance might be significantly different as well. We performed a comparative analysis sample data on these two groups: (1) National Scholarship and National Encouragement Scholarship winners and (2) students who had to retake exams.

3.2. Sample Data Collection

We successively collected data of students enrolled in college in 2015, 2016, and 2017, including National Scholarship winners, National Encouragement Scholarship winners, and students who had to retake exams. Among them, 1,313 students had received National Scholarship and National Encouragement Scholarship, and 7,663 students had retaken exams. The specific data are shown in table 1.

Table 1. Sample information

Class	Number of National Scholarship winners	Number of National Encouragement Scholarship winners	Number of students retaking exams
2015	/	/	3086
2016	/	699	2408
2017	29	585	2169
Total	29	1284	7663

3.3. Collection of Internet Behavior Data of Samples

The TOPSEC Behavior Audit System is based on the apache + mysql + php technology architecture. In this study, we deployed a virtual machine to export the internet behavior data of the samples from the audit system by running the import.php script, forming a data file in the format of. csv. We then used bat.txt to merge the daily. csv files of each sample.

We selected internet behavior data during the seven months (214 days) from June 1 to December 31, 2017. We collected valid data from a total of 1340 samples, among whom 350 samples were National Scholarship and National Encouragement Scholarship winners, and 990 students had to retake exams. The total number of internet behaviors collected was 104,154,805. Table 2 is the statistical table of the number of internet behaviors.

Table 2. Number of internet behaviors*

Class	**Effective number of National Scholarship winners**	**Number of internet behaviors**	**Effective number of National Encouragement Scholarship winners**	**Number of internet behaviors**	**Effective number of students retaking exams**	**Number of internet behaviors**
2015	/	/	/	/	156	10,435,936
2016	/	/	138	14,520,023	239	18,204,140
2017	9	923,461	203	13,432,259	595	47,268,986
Total	9	923,461	341	27,952,282	990	75,909,062

* There is a big gap between the effective data and the sample size. The reason is that the student dormitories are covered by three internet services: the campus internet, China Telecom internet, and China Mobile internet. The students can freely choose the internet they want to use. This statistical table is based only on the data pertaining to student use of the campus internet.

4. Research Design

4.1. Processing of Internet Behavior Data

We exported internet behavior data from the TOPSEC Behavior Audit System, and then used Excel and Matlab to synchronously process the data and remove any abnormal data. The data structure after data-masking is shown in table 3.

Table 3. Internet behavior data

Browsing websites	Accessing forums and microblogs	Sending files	Sending information	…	Web search	Government sectors	Weekly magazines and media
10	0	0	7,958	…	0	0	0
28,508	11	0	40,696	…	15	0	4
193,812	8	0	202,544	…	15	0	0
16,696	18	0	24,801	…	37	0	0
……	……	……	……	…	……	……	……
8,197	3	0	6,497	…	0	100	15
6,052	1	0	7,455	…	9	0	0
0	0	0	2,584	…	0	0	0
106,861	67	0	97,333	…	20	0	0
5,154	0	0	32,708	…	3	0	29

This table contains 85 internet behavior fields, including browsing websites, accessing forums and microblogs, sending files, sending information, conducting web searches, receiving and sending emails, logging into accounts, completing IT-related behaviors, accessing Baidu Wenku, browsing Baidu news, accessing blogs, watching movies, and chatting with people. The tabular data denote the frequency. We used SPSS to classify and summarize internet behaviors and established a mathematical model based on mathematical equations to present the relationship between college students' internet behaviors and academic performances.

4.2. Relationship Between College Students' Internet Behaviors and Academic Performance

To objectively reflect the fitting effect of the logistic regression model, we applied SPSS to randomly select 210 scholarship winners and 210 students who were retaking exams. The internet behavior data contained 85 internet behavior fields, including the examples given previously. We classified and summarized data and then combined data in various ways to obtain several models, among which the internet behavior ratio model had the best effect. The data structure is shown in table 4. The predicted results are shown in the classification table of internet behavior data (table 5).

Table 4. Ratio of internet behaviors

Browsing websites	Accessing forums and microblogs	Sending files	Sending information	Web search	Receiving and sending emails	Account logging in	…
0.3301	0.0000	0.0000	0.6483	0.0011	0.0000	0.0109	…
0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	…
0.2857	0.0000	0.0000	0.7143	0.0000	0.0000	0.0000	…
0.8077	0.0000	0.0000	0.1799	0.0001	0.0000	0.0035	…
0.1685	0.0000	0.0000	0.6602	0.0009	0.0000	0.0010	…
0.5912	0.0001	0.0000	0.3951	0.0024	0.0000	0.0057	…
0.3997	0.0000	0.0000	0.5909	0.0008	0.0000	0.0051	…
0.6849	0.0005	0.0000	0.3036	0.0040	0.0000	0.0049	…
0.0024	0.0000	0.0000	0.9976	0.0000	0.0000	0.0000	…
0.2104	0.0000	0.0000	0.5586	0.0015	0.0000	0.0093	…
…	…	…	…	…	…	…	…

Table 5. Classification of internet behavior data [a]

Classification table [a]

Observed value			Predicted value		
			Exam results		Correctly predicted percentage
			0	1	
Step 1	Exam results	0	208	2	99.0
		1	75	135	64.3
	Overall percentage				81.7
Step 2	Exam results	0	203	7	96.7
		1	61	149	71.0
	Overall percentage				83.8
Step 3	Exam results	0	202	8	96.2
		1	52	158	75.2
	Overall percentage				85.7
Step 4	Exam results	0	205	5	97.6
		1	51	159	75.7
	Overall percentage				86.7
Step 5	Exam results	0	205	5	97.6
		1	51	159	75.7
	Overall percentage				86.7

a. Segmentation threshold: .500

As shown in table 5, the overall percentage of the model to accurately predicting the existing data reached 86.7%. Table 6 shows the specific expression of the model according to variations in parameters in the internet behavior model equation.

Table 6. Variation of parameters in the internet behavior model equation

		B	S.E.	Wald	df	Significance	Exp(B)
Step 1a	Other websites	4017.214	1253.846	10.265	1	.001	.
	Constant	−1.039	.136	58.723	1	.000	.354
Step 2b	Web search	1074.855	267.159	16.187	1	.000	.
	Other websites	4596.598	1344.785	11.683	1	.001	.
	Constant	−1.359	.159	73.374	1	.000	.257
Step 3c	Web search	747.382	254.135	8.649	1	.003	.
	Account logging in	191.334	54.597	12.281	1	.000	1.245E+83
	Other websites	5211.575	1454.037	12.847	1	.000	.
	Constant	−1.700	.192	78.669	1	.000	.183
Step 4d	Web search	755.616	266.600	8.033	1	.005	.
	Account logging in	55.342	66.025	.703	1	.402	1083315821127086100000000.000
	Instant messaging (QQ)	572.645	212.404	7.269	1	.007	4.971E+248
	Other websites	5351.212	1475.460	13.154	1	.000	.
	Constant	−1.773	.197	81.111	1	.000	.170
Step 5d	Web search	825.758	262.982	9.859	1	.002	.
	Instant messaging (QQ)	677.947	175.289	14.958	1	.000	2.684E+294
	Other websites	5286.957	1462.916	13.061	1	.000	.
	Constant	−1.738	.191	82.463	1	.000	.176

a. Variable input in Step 1: [%1:, 1:
b. Variable input in Step 2: [%1:, 2:
c. Variable input in Step 3: [%1:, 3:
d. Variable input in Step 4: [%1:, 4:

Following Step 15 in table 6, Equation (4) was obtained, as follows:

$$\mathrm{Logit}(p) = \ln\frac{p}{1-p}$$

$$= -1.738 + 825.758 \times \text{web search} + 677.947 \times \text{Instant messaging (QQ)} + 5286.957 \times \text{Other websites}. \quad (4)$$

On the basis of the analysis and interpretation of Equation (4), we found that 14 parameters—including accessing websites, sending information, online shopping, playing computer games, and constant—had a significance of $p < 0.05$. This finding indicated that their correlation with academic performances was statistically significant. The probability p calculated based on Equation (4) well revealed the academic performance of college students, with an overall accuracy reaching 86.7%.

5. Conclusions

In this study, we used the binary classification logistic regression model to study the relationship between college students' internet behaviors and academic performances, which was an effective exploration. We selected scholarship winners and students who were retaking exams as the research samples. The binary classification of the samples was the premise and assumption of applying the logistic regression model, which was the theoretical basis of this study.

We collected a total of 1,259,185 pieces of data about 8,976 college students' duration of time internet surfing from the internet surfing authentication system. We extracted 104,154,805 data points about these students' internet behaviors from the internet behavior audit system. We used multiple data-processing tools, such as script, batch processing, Excel, and SPSS.

We selected the internet behavior data of 210 scholarship winners and 210 students who were retaking exams, and used SPSS to conduct the logistic regression analysis. We established a corresponding mathematical model and determined the model parameters. With a prediction accuracy of 86.7%, the model achieved a great overall effect. In other words, the mathematical model determined by parameters, including web search, instant messaging (QQ), and other websites, relatively accurately revealed the relationship between college students' internet behaviors and academic performance. From the perspective of data, this study was innovative for establishing an academic performance early warning mechanism for institutions of higher education and provided a basis for decision-making related to the ideological and political education of college students.

The data originated from the log files of the internet surfing authentication system and internet behavior audit system. This study used these data to truthfully record college students' duration of internet surfing and other various internet behaviors. Unlike data obtained through telephone interviews and questionnaire surveys, the data used in this study were objective and truthful without being affected by psychological factors such as fear, shyness, or conformity. Therefore, the research results gained in this study were closer to the objective reality, which was an innovative aspect of this study.

6. Suggestions for Improvements

In terms of the methods, further study remains to be carried out using SPSS. The 85 internet behavior fields could be further explored. For instance, the fields could be combined, aggregated and classified; and additional fields could be included for a detailed analysis. For example, the internet behaviors of male and female college students as well as the relationship between their respective internet behaviors and academic performances could be studied, and the differences between male college students' and female college students' internet behaviors could be investigated.

Acknowledgement

This paper is supported by the CERNET Innovation Project under Grant No.NGII20170511, Key Program of humanities and social sciences in universities of Anhui Province under Grant No.Sk2015a380, 2020 Teaching Quality and Teaching Reform project in Anhui Province NO.2020qkl23.

References

[1] Deng Yankui, Kuang Xiaoxia. Analysis and intervention of college students' internet behavior anomie in an all-media environment [J]. Leading Journal of Ideological & Theoretical Education, 2016.9:151-155.
[2] Lan J, Ai D. The anomie behavior of information dissemination in WeChat public platform and its governance path [J]. Journal of Hunan University (Social Sciences), 2018, 32(3): 154-160.
[3] Wang Hua, Yang Lingling. Youth internet moral anomie and corresponding guidance [J]. People's Tribune, 2018(11):116-117.
[4] Zhu Lin. Types, Causes and countermeasures of college students' internet behavior anomie [J]. Journal of East China Normal University (Education Sciences), 2016.2:88-95.
[5] Feng Tianmin, Zhang Rujing. A review of research on online learning behaviors over the past five years in China [J]. New Media Research, 2019(05):75-79.
[6] Wang Yang. Research on the development of learners' online learning behaviors from the perspective of Big Data [J]. China Information Technology Education, 2019 (03): 102-105.
[7] Kohn, Karen. Using Logistic Regression to Examine Multiple Factors Related to E-book Use; Library Resources & Technical Services; Chicago Vol. 62, Iss. 2, (Apr 2018): 54-65.
[8] Anand, Nitin; Jain, Praveen; Prabhu, Santosh; Thomas, Christofer; Bhat, Aneesh; Internet use patterns, internet addiction, and psychological distress among engineering university students: A study from India. Indian Journal of Psychological Medicine; Kottayam Vol. 40, Iss. 5, (Sep/Oct 2018): 458-467.
[9] The Ministry of Education of China. 2016 Rolling Survey on College Students' Ideological and Political Status [EB/OL].
[10] The 44th China Statistical Report on Internet Development [EB/OL].
[11] Zhang C, Gao K, et al. An analogue comparison of discriminant analysis and logistic regression [J], 2010(1):19-25.
[12] Li Y, Zhang J. A simple linear regression model based on the compositional data. Journal of Applied Statistics and Management [J].2019(3): 442-449.
[13] Song J, Ge Y. The application of logistic regression in binary task pricing model. Journal of Nanjing Normal University (Natural Science Edition) [J],2018(4):33-38.
[14] Sun Yifan, Pan Kunfeng et al. Ideas and methods of predicting college graduates' career development—an attempt based on machine learning algorithms [J]. Education Research Monthly, 2019(1):25-35
[15] Bao W, Li B. Who is unemployed, employed or admitted to graduate school; an investigation of the employment situation of college graduates in China between 2003 and 2009[J]. Chinese Education&Society, 2015, 47(6):36-58.
[16] Lan Y, Liu Y, et al. The internet public opinion hot-degree dynamic prediction model oriented to big data. Journal of Intelligence [J], 2017(6):105-110.