Logistic Empirical Study of Regression Model in Python Class Model Examination

Qibao Huang, Jiawen Wang, Linghong Rao*

College of Digital Technology Application Industry, Shangrao Normal University, Shangrao 334001, China

Abstract: In the information age, data holds increasingly significant value, and computer-based data analysis and visualization have become widespread practices. This paper explores the application of Logistic regression models in analyzing factors influencing student performance. Logistic regression is a supervised learning algorithm that predicts a binary dependent variable based on a set of independent variables, transforming the predicted value of linear regression into a probability through a sigmoid function. In the context of education, student achievement is often categorized, making Logistic regression well-suited for analyzing its influencing factors. By collecting multifaceted data from students, the Logistic regression model can establish a predictive model with multiple independent variables, such as personal characteristics, family factors, and school environment. This study focuses on the analysis of Python course model examination results, preprocessing the data to ensure integrity and addressing abnormal values. Visualization analysis is employed to intuitively reveal the potential effects of various factors, including gender, average daily learning time, undergraduate school level, and registered school level, on model exam scores. The results demonstrate significant correlations between these factors and exam scores, with Logistic regression analysis further explaining about 72% of data performance variation. This study provides insights into using data visualization and Logistic regression for educational research and decision-making.

Keywords: Visualization; Python; Logistic Regression Model; Matplotlib

DOI: 10.57237/j.jeit.2025.01.001

1 Research Background and Significance

At present, the importance of Python course education is increasingly highlighted, and the Python course examination has become one of the main ways for candidates to enter the Python course college. Candidates need to show their academic level and ability through the exam, so the competition is extremely fierce.

Each examinees learning background, subject basis and test-taking ability are different, so it is necessary to provide personalized learning assistance and guidance according to individual differences. Model examination is an effective way of learning, which can help candidates to better understand their own learning status and improvement space.

With continuous advances in data science and visualization technologies, Python became one of the mainstream tools in the field of data analysis. Using Python for data analysis and visualization, we can more intuitively and discover the rules and trends behind the data. The analysis of simulated Python test data through Python visualization technology can provide personalized learn-

*Corresponding author: Linghong Rao, 632572059@qq.com

Received: 17 December 2024; Accepted: 17 January 2025; Published Online: 8 February 2025 http://www.jeduit.com

ing support for candidates. According to the performance of the candidates in the model exam, we will provide targeted learning suggestions and guidance to help the candidates prepare for the Python course exam more targeted.

2 Research Status at Home and Abroad

At present, the application of data visualization in the field of education is gradually becoming a research hotspot. In particular, the application of Python shows great potential in the processing and analysis of student achievement data. Still, visual studies conagainst student achievement data remain relatively lacking globally. Some studies have developed Python-based analysis platforms, which not only simplifies the data presentation process, but also can explore the intrinsic value of the data more deeply, which is both for understanding andIt is of great significance to improve the educational effect. Song Yongsheng and Huang Rongmei [1] et alPeople pointed out that although the research on data visualization at home and abroad has gradually become a hot topic, there is still a lack of visual research based on the background of student achievement data. He used Python to develop the academic performance analysis and visualization platform, to intuitively display the students performance data, discover its internal rules, and explore its potential value.

Although Excel is still a common tool for processing data in the education field, its limitations in handling large or complex data sets, such as cumbersome operation and error-prone. In contrast, Python can not only automate the statistical analysis of data, but also effectively improve work efficiency and accuracy. In addition, international researchers such as Kirthi Raman have demonstrated the wide application of Python in data visualization, emphasizing the important role of tools such as NumPy, matplotlib, pandas and others in generating intuitive and diversified visualization results.

Shi Xiaoling, Yang Ligong [2] People point out that the main software used to process data in the current field of education is still Excel. Although Excel can carry out simple statistical operation of students performance, the operation of Excel is very complicated when processing a large amount of data, and the standardization degree of artificial operation is low and mistakes are prone to occur. The use of Python can automatically achieve the statistical analysis of data, can greatly improve the work efficiency.

Kirthi Raman [3] Based on many years of development experience, combined with the actual cases in various fields, a comprehensive elaboration Python Visualization methods in different application fields. He detailed the application of NumPy, matplotlib, pandas and other tools in the visualization process, showing how these tools generate visual results in different ways.

Moreover, the use of statistical methods such as Logistic regression model to analyze the influence of factors such as students family background and study habits on academic performance, also indicates the utility of Python for deeper data analysis. These studies not only identify key factors influencing students academic performance, but also provide data support for education policy making and teaching method optimization. Li Guo, Zhang Meng, Meglumine Iothalamate [4] People analyzed the factors affecting students academic performance using Logistic regression models. By collecting data on students family background, study habits, teacher quality and so on, they constructed Logis tic regression models to identify the key factors affecting students academic performance.

3 Research Objects and Research Methods

3.1 Introduction and Selection of Research Methods

The regression model is also rich and the theory is increasingly perfect. Some methods are able to accurately and efficiently predict classification problems, such as the widely used Logistic regression models. Logistic Regression is a commonly used supervised learning algorithm used to divide a dataset into two or more different categories to predict a binary dependent variable (response variable) based on a set of independent variables (features). The model transforms the predicted value of the linear regression into a probability value through a logical function (sigmoid function), so as to classify the data for study.

The core idea of Logistic regression is to predict the probability of the dependent variable belonging to a certain category through a set of independent variables. The algorithm first uses a set of training data to estimate the models parameters (e. g., the regression coefficient) that describe the relationship between the independent variable and the dependent variable. Then, for the new data points,

the Logistic regression model uses these parameters to calculate the probability of belonging to each category and selects the category with the highest probability as the prediction outcome.

Considering the vector $x=(x_1, x_2,..., x_n)$ with n random variables, let the conditional probability $p(y=1|x) =p$ be the probability of the event y under the independent variable x, then the Logistic regression model is expressed as:

$$P(y = 1|x) = \pi(x) = \frac{1}{1+e^{-g(x)}}$$

In the formula, $\pi(x)$ is called the Logistic function, where $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$. The goal of the Logistic regression model is to maximize the likelihood function, which maximizes the probability of the observed data at a given model parameters. The parameter estimates of the model are obtained by solving the maximum of the likelihood function. These parameter estimates describe the relationship between the independent and dependent variables and can be used to make categorical predictions for new data.

3.2 Application of Logistic Regression Model in the Analysis of Factors Affecting Student Performance

In the field of education, the improvement of student performance is the common concern of educators and parents. In order to improve students learning effect, it is crucial to understand and analyze the various factors affecting students performance. In this context, Logistic regression models, as a powerful statistical tool, played an important role in analyzing factors influencing student performance. Logistic Regression model is a regression analysis method used to handle dichotomized or multicategorized dependent variables. In the educational field, student achievement is usually divided into different grades or categories, so the Logistic regression model is well suited for analyzing the influencing factors of student performance. By collecting multifaceted data from students, the Logistic regression model was able to build a predictive model with multiple independent variables. These independent variables can be students personal characteristics, family factors, school environment, etc. The model analyzes the relationship between these independent variables and student performance. Once the model is established, educators can use the Logistic regression model to analyze the factors influencing student performance.

4 Analysis and Study of Python Course Model Examination Results

4.1 Data Introduction and Preprocessing

This paper mainly determined the possible influencing factors through social investigation and collected data. Finally, five attributes of 418 students were collected, namely, the scores of the last model test, gender, average daily learning time, school level and other attributes. The main collection methods are questionnaire star and on-site questionnaire survey. Table 1 shows some examples of the data.

Table 1 Part of the student data

mark	sex	Study time every day	Undergraduate school level	Examinee school level
126	man	9	Double non	Double non
110	woman	8	Double non	Double non
87	woman	6	Double non	Double non
97	woman	8	Double non	211
101	man	8	Double non	211
124	man	8	211	211
117	woman	7	211	211
128	man	9	211	985
136	man	8	985	985

4.2 Data Preprocessing

The preprocessing of the data in this article includes the following two sections:

First, the hot card filling method is used in the missing data, which means that the most similar object to the missing data in other complete data, and the value of the object is used to fill in. This approach helps to maintain data integrity, and although some data is missing, we can

preserve the integrity of the data set whenever possible by exploiting the information from other data.

Second, for the data objects containing abnormal values, this paper chooses the way of direct deletion for processing. An experience-based judgment method was used, such as treating the data with the average daily learning time attribute as negative or greater than 24 hours as abnormal. Considering that all attributes have a selection range during the investigation, the possibility of outliers is low. This method is simple and direct, which can effectively exclude the interference of outliers to the subsequent analysis, and improves the quality and credibility of the data.

4.3 Visualization Analysis of Python Course Model Examination Score Data

This paper details the characteristics and processing process of a specific data set, —— namely Python model test data. Through careful preprocessing of this dataset, we obtained a dataset with 418 records covering five key variables. In order to dig deep into the information and value behind this data set, this paper will adopt the method of visual analysis, and strive to intuitively and comprehensively show the rich content contained in the data set through the form of graphics and charts.

In the process of specific analysis, we will start from four dimensions: gender, average daily learning time, undergraduate school level and registered school level. First, first of all, we will give an overview of the overall situation of the data set from these four perspectives to form a preliminary understanding. Subsequently, we will deeply explore the potential links between these dimensions and data model scores, and intuitively reveal the potential effects of various factors on data model scores through diversified visualization tools, such as bar chart, line chart, scatter chart, etc. Finally, based on the results of these visual analysis, we will conduct in-depth analysis and discussion to reveal the key factors affecting the performance of the data model test.

First, the gender attributes in the dataset were analyzed, shows the distribution and proportion of sex in the data set, in which 226 men accounted for 54% of the total number, 192 women accounted for 46% of the total number, and slightly more men than women. This may indicate that male engagement may be higher than female in exams where data subjects are available. Figure 1 shows the average score corresponding to gender, in which the average score of male data is about 101.5, while the average score of women is about 92.8. Although this score also shows their excellent strength in the data, there is still a certain gap compared with the average score of men. This 8.7-point gap may also indicate that women are indeed not statistically dominant.

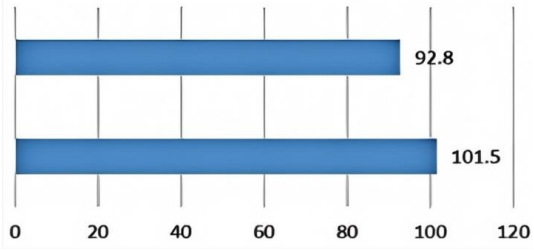

Figure 1 Mean score corresponding to sex

Second, the average daily learning time attribute in the data set was analyzed. Figure 2 provides a detailed overview of the changes in the distribution of people per daily learning time. We can see differences in the number of participants across learning periods and how these differences constitute a general trend. The most striking part is in the group of students who study for seven hours a day. Their number reached 202, which is particularly prominent in the figure, because it occupies a significant proportion of the total number of people. —— is about half. Immediately after, we noticed that the number of learners in the 6 h and 8 h time periods was also considerable. Of these, 92 students studied for six hours a day, while 88 students studied for eight hours. We then turned our eyes to people who had studied longer. The number of students who studied for nine hours a day was 30, compared with only six for 10 hours. Finally, as a whole, the distribution of people shown in this graph roughly fits the characteristics of a normal distribution. This means that most students tend to choose learning times near the median, while extreme values (such as very short or very long learning times) are relatively small. To some extent, this distribution pattern reflects the general law and tendency of examinees in the choice of learning time. reveals in detailThe significant effect of average daily learning time on mean scores. As the study time gradually increased from 6 hours to 10 hours per day, the average score of students also showed a steady upward trend. Specifically, the average score of students who studied for 6 hours a day was 81.1, while the average score jumped to 122.3 when the study time increased to 10 hours. This data intuitively

shows the positive correlation between learning time and academic performance. The longer the average learning time, the higher the average score, showing the efforts and success of candidates to improve their performance by increasing their learning time.

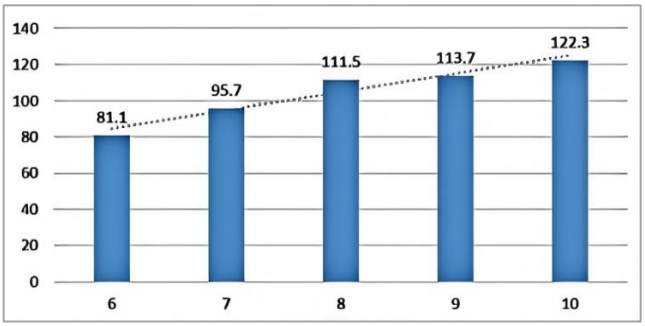

Figure 2 Mean score corresponding to the learning time

4.4 Quality Inspection of the Questionnaire

Before the formal questionnaire analysis, it is very necessary for the reliability analysis of the formal scale, so as to ensure the reliability and internal consistency of the questionnaire scale. The reliability analysis of the scale uses the Cronbacha coefficient, which is greater than 0.8, the scale is considered highly reliable; if the coefficient is between 0.7 and 0.8; if the coefficient is between 0.6 and 0.7, the reliability is considered low with great uncertainty. If the reliability coefficient is below 0.6, then the accuracy of the scale is extremely low, and only a redesign can ensure its reliability. After a rigorous reliability test, the overall Cronbachs Alpha coefficient of the scale reached 0.929, which indicates the high reliability of this questionnaire, which in turn confirms the reliability and quality of the study data. The specific data are shown in Table 2 below.

Table 2, reliability analysis

dimension	Cronbachs Alpha
Scale as a whole	0.929

4.5 Validity Analysis

Validity test, namely, the measure of the validity of the questionnaire. In general research, the validity of the questionnaire is to test whether the questionnaire test data can be effectively reflected in the research purpose, that is, to verify whether the questionnaire is accurate and effective for the study. The validity of the scale in the questionnaire was tested by using the KMO coefficient. The KMO and Bartletts spherical degree tests show the test results of KMO and Bartlett, and the value from the KMO is 0.848, indicating that the questionnaire has very high validity and meets the metric criteria. After testing, the value of the Bartlett sphere reached 3238.964 and the degree of freedom reached 12, which indicates that in this case, the Sig value is below the significance level α =0.05, indicating a significant difference between the correlation coefficient matrix and the identity matrix. So the questionnaire has a good validity. The details are shown in Table 3.

Table 3 KMO and Bartlett tests

Number of KMO sampling suitability quantities.		.848
Bartlett sphericity test	Approximate chi square	3238.964
	free degree	12
	conspicuousness	0.000

4.6 Descriptive Statistics

Form 4-4 provides descriptive statistics on data scores, daily study time, gender, the undergraduate school level, and the registered school level. These data revealed the distribution of the individual variables in the sample.

First, the data scores were divided into four sections, with 64 students with a score of 56 to 76, or 15.311% of the total, which was the lowest proportion in the score range. The number of students who scored from 76 to 96 was 130, or 31.1%, amounting to 46.411%. Next, the 96-116 range had 144 people, the largest number of all intervals, at 34.45%, increasing the cumulative percentage to 80.861%. The highest score range of 116 to 136 had 80 people, accounting for 19.139%, and the cumulative percentage reached 100%. This shows the distribution of student data scores in the middle, with most of the student scores concentrated between 76 and 116 points.

In terms of learning time per day, the distribution of people in different learning periods also varied. 92 students study 6 hours a day, or 22.01%; 20% in 7 hours, the largest number of all groups, accounting for 48.325%, indicating that nearly half of the students study 7 hours a

day; 88, 21.053%; more students, 9 hours and 10 hours, 30 and 6 students, respectively 7.177% and 1.435%. The cumulative percentage eventually reached 100%, indicating that the majority of students would study between six and eight hours.

In terms of gender distribution, 226 male students, accounting for 54.067%, and 192 female students, or 45.933%, showed that there were slightly more males than females in this sample. In terms of undergraduate school level, Shuangfei schools have the most students, with 272 students, accounting for 65.072%; 211 universities have 110 students, accounting for 26.316%; and 985 universities have the least students, with 36 students, accounting for 8.612%. This may reflect the small proportion of college students at 985 and 211 overall. Data from the school level show that 222 people applied for double non-schools, accounting for 52.607%; 102 applied for 211 schools, accounting for 24.171%, and 94 applied for 985 schools, accounting for 22.275%, indicating that the selection of school level is relatively balanced among the three types.

Through these descriptive statistical results, we can observe the diversity of learning behavior, performance distribution, and educational background in the sample students. These data provide basic information for further analysis of students academic performance and the influencing factors behind it.

Table 4 Descriptive statistical results

name	option	frequency	percentage (%)	Cumulative percentage of (%)
Data results	[56.0, 76.0)	64	15.311	15.311
	[76.0, 96.0)	130	31.1	46.411
	[96.0, 116.0)	144	34.45	80.861
	[116.0, 136.0]	80	19.139	100
Study time every day	6	92	22.01	22.01
	7	202	48.325	70.335
	8	88	21.053	91.388
	9	30	7.177	98.565
	10	6	1.435	100
sex	man	226	54.067	54.067
	woman	192	45.933	100
Undergraduate school level	Double non	272	65.072	65.072
	211	110	26.316	91.388
	985	36	8.612	100
Apply for the school level	Double non	222	52.607	52.607
	211	102	24.171	76.777
	985	94	22.275	99.052
	man	2	0.474	99.526
	226	1	0.237	99.763
	101.5	1	0.237	100

4.7 Correlation Analysis

The correlation test is the method used in statistics to test whether there is a linear correlation between two variables. In correlation tests, the correlation coefficient between two variables is usually calculated, and hypothesis testing is performed to determine whether this correlation is significant. In order to examine the correlation among these variables, the correlation test was conducted, and the test results are shown in Table 5.

In the correlation test, we discussed the correlation between the model test scores and gender, study time, undergraduate school level and the level of the registered school. The results showed that there were different degrees of linear correlation between model test scores and these variables. The correlation coefficient between the scores of the model examination and the school level is the highest, reaching 0.357, and the significance level is very high ($p < 0.01$), indicating that the correlation between the model examination results and the school level is the most significant, which may reflect that the students with higher scores in the model examination tend to apply for the schools with higher levels. Secondly, the correlation coefficient between the model exam scores and the undergraduate school level was 0.239, which was also significant at the significance level of $p < 0.01$, indicating that the students undergraduate background was also positively correlated with the model exam scores. Furthermore, the correlation coefficient between model test scores and gender was 0.201, reconfirming that gender has some influence on model test performance at the significance level $p < 0.01$, which may indicate that gender factors differ in learning performance. The cor-

relation coefficient between model scores and study time was 0.178, which, although the coefficient was relatively low, was also significant at the significance level at p <0.01, indicating that spending more time on study may improve model scores. These statistical results reveal that the model examination results are jointly influenced by many factors, including the registered school level and the undergraduate school level. The effect of times was the most significant, while gender and study time, although less affected, should not be affected.

Table 5 Association test results

Model examination results gender learning time undergraduate school level, apply for the school level				
Model test results	1			
sex	0.201**	1		
learning time	0.178**	0.018 1		
Undergraduate school level	0.239**	0.057 0.098 1		
Apply for the school level	0.357***	0.037 0.054 0.115	1	

* p < 0.1, ** p < 0.05,*** p < 0.01

4.8 Regression Analysis

In the linear regression analysis, the model used data scores as the dependent variable, while study time per day, gender (male), undergraduate school level (985 and double), and school level (985 and double) as independent variables.

The overall explanatory power of the model was strong, where the R^2 was 0.724 and the adjusted R^2 was 0.72, which indicates that the model explained about 72% of the data performance variation, showing the model having high explanatory power. The F statistic was 179.941, and the corresponding P-value was close to 0, indicating that the model was statistically significant, and the independent variable group as a whole significantly predicted the data performance.

Specific to the coefficients of each independent variable, the constant term is 36.263, indicating the expected data score baseline of 36.263 points for all other independent variables of 0. The coefficient of learning time per day was 9. 686, the standard error was 0.685, and the standardization coefficient Beta was 0.469, indicating that for each additional hour of learning time, the data score improved by 9.686 points, and this variable had a significant effect in the model (P value was almost 0), showing a strong positive effect.

For gender variables, men had an average of 6.112 points higher than women, and this effect was also statistically significant (P-value was almost 0). At the undergraduate school level, students from 985 institutions scored on average 6.663 points higher than those from other types of schools, with a P-value of 0.002, indicating that the effect was equally significant. On the contrary, students of double non-schools scored 21.363 points lower than those of other types of schools, the result of which was also very significant (P value is almost 0), showing a large negative effect.

For the school level, the coefficient of 985 schools was 3.124. Although the coefficient showed a positive effect, the P-value was 0.057, which was only significant at the 10% significance level, indicating a relatively weak influence of this variable on data performance. The coefficient of double non-schools was 1.993, but the P-value was 0.307, indicating that it was not statistically significant, indicating that the prediction effect of the data scores was not significant.

Overall, the linear regression model effectively revealed the significant effects of variables such as learning time, gender and undergraduate school level on data performance, while the effect of the registered school level was relatively small. These findings provide insights to educational institutions and policy makers on how to improve students academic performance by adjusting study time, understanding gender differences, and optimizing institutional resource allocation.

5 Summary and Outlook

In this paper, we applied visualization technique and Python programming language were applied to make an in-depth analysis of the data of Python model examination and predict using Logistic regression model [5-8]. The main purpose of the research is to explore the influence of each factor on the results of the model examination through the visual presentation of the data, and to build an effective prediction model to provide data support for the registration decision [9-11]. Through data preprocessing and feature engineering, we cleaned and transformed the raw data. A detailed exploratory analysis of the data was

performed with the help of the data visualization library in Python. The visual results showed that factors such as the average daily study time had a significant impact on the simulation results [12-15]. A Logistic regression model was then developed to predict the final scores.

References

[1] Song Yongsheng, Huang Rongmei. Academic performance analysis and visual analysis based on Python [J]. Information Technology and Information Technology, 2021, (12): 100-102.

[2] Xiaoling Shi, Gong Yang. Python-based student achievement analysis and visual presentation study [J]. Journal of Hebei Software Vocational and Technical College, 2022, 24(02): 4-6. https://doi.org/10.13314/j.cnki.jhbsi.2022.02.013

[3] Kirthi Raman. Python Data Visualization [M]. Machinery Press, 2017.

[4] Li Guo, Zhang Meng, Kang Rui. Research on student performance prediction model based on logistic regression [J]. China Information Technology Education, 2023.

[5] Katsiaryna K, T S H, Chang Woo J, et al. Benefits of implementing student-led revi ew and mock examinations in the medical undergraduate gross anatomycurriculum. [J]. Medicalteacher, 2022, 44(9): 1-4.

[6] Cai Zhenhai. Explore the practice of abnormal data processing and analysis based on Python [J]. Computer Knowledge and Technology, 2023, 19(27): 62-65. https://doi.org/10.14004/j.cnki.ckt.2023.1434

[7] Yu Feiyang, Yang Hengjie. Design and implementation of the data analysis software based on Python [J]. Modern Computer, 2023, 29 (12): 99-103.

[8] Bao Peidong, Wan Nan, Wang Tingting, etc.. Research on data climbing and data visualization analysis of new energy vehicles based on Python [J]. Light Industry Technology, 2023, 39(05): 105-107.

[9] Wang, L., & Zhang, J. (2023). Factors Influencing Academic Performance: An Analysis Based on Logistic Regression Model. Journal of Educational Research, 42(3), 123-135.

[10] Zhao, M., Liu, X., & Chen, H. (2022). Gender Differences in Academic Performance: Evidence from a Large-Scale Dataset. Gender and Education, 34(1), 56-72.

[11] Johnson, A., & Smith, B. (2021). The Impact of Undergraduate Institution on Graduate Performance: A Comparative Analysis. Higher Education Quarterly, 45(2), 110-128.

[12] Chen, Y., & Wang, Z. (2023). Data Visualization Techniques for Educational Research: A Review. Journal of Educational Data Mining, 15(1), 45-60.

[13] Taylor, K., & Davis, L. (2022). Logistic Regression Modeling in Educational Research: A Practical Guide. Journal of Educational Measurement, 55(4), 301-319.

[14] Liu, W., & Zhang, H. (2021). The Role of Study Time in Academic Performance: Insights from a Longitudinal Study. Journal of Learning Analytics, 8(2), 147-163.

[15] Brown, J., & Lee, H. (2023). The Effects of Online Learning on Student Engagement and Academic Performance. Journal of Distance Education, 38(1), 23-45.