Prediction of University Comprehensive Score Based on Regression Analysis

Yutong Li*

Computer Science, University of Bristol, Bristol BS8 1TH UK
**Corresponding author. Email: tt20694@bristol.ac.uk*

ABSTRACT
Machine learning [1-5] is regularly accustomed to various fields to resolve troubles that are not able to easily solve in founded on computer methods. The most straightforward and frequent algorithms in machine learning is linear regression. Linear regression is a method for performing predictive analysis based on mathematics. University ranking is a very important but also very challenging and controversial issue. The comprehensive strength of a university involves scientific research, teachers, students and other aspects. But what factors weigh heavily in rankings, and how a university's composite score is determined based on factors such as education quality, alumni employment, research output and citations. The linear regression is used in this study to forecast the strength of universities based on the rankings of teachers and scientific research provided by CWUR, and explores which factor has the greatest impact on university strength. This dataset [6] comes from Kaggle and contains various metrics and rankings of thousands of universities. After conducting a large number of linear programming predictions and simulations, the data show that Harvard University has the highest overall strength, and the quality of faculty is the evaluation indicator that has the greatest impact on the overall score.

Keywords: *Machine learning; linear regression; university ranking; prediction*

1. INTRODUCTION

Linear regression [7-9] is a mathematical test applied with estimating and appraising the connection with the reasonable variables. Fisher test Chi-square and t-testing and variance analysis (ANOVA) be widely used in considering to the outcomes of the other researchers in the analysis is not considerable. Therefore, Partial correlation and regression are therefore tests that enable scientists to link two variables to assess the effects of confounding.[4,10,11]. Linear regression [12] is regularly applied with mathematical research approaches. Measure the predicted effects and model them against multiple variables is effective for Linear regression. Linear regression can establish a linear correlation between dependent and independent variables. Therefore, Linear regression is helpful in evaluating and modelling data.

In current years, the impact of the World University Rankings has grown. Obviously, university rankings have become a significant standard for evaluating "first-class", which not only affects the evaluation of universities, but also influence students' options of universities; on the other hand, rankings have a greater influence on universities planning, and more and more of universities began to notice rankings as their own development goals.

The author hopes to build a machine learning model with a linear regression method so that people can easily and intuitively understand the overall competency of a university, predict the overall score of a university, and judge various advantages and shortcomings of the university.

2. THEORY

2.1. Regression

Regression [13] is a technique accustomed to two approaches. Primarily, regression examination are normally used for prophecies, because their application has great overlaps with the fields of machine learning. Then, regression analysis can be accustomed to many examples to decide causal relations between the independent and dependent variables. Significantly, regressions alone present only relations between a

dependent variable and a fixed dataset collection of various variables.

2.2. Least square method

The greatest approaches to show the best fit curve of line for one set of data point is the least square method(LSM)[14,15]. LSM reduces the amount of the squares of the offsets (residual part) of the points of the curve.

The LSM in the linear regression model use to find b0 and b1 predictions such that the cumulative squared distance from the real y_i response $\hat{y} = \beta_0 + \beta_1 x_i$ approaches the minimum of all possible regression coefficients β_0 and $\beta_1$option

$$(b_0, b_1) = arb\min \sum_{i=1}^{n} [yi - (\beta_0 + \beta_1 X_i)]^2 \quad (1)$$

The motivation behind the least squares approach has been to find the estimates of the parameters using the least squares that is the nearest line to all points (x_i, y_i). The least-squares results of the basic linear regression listed in this paper are computed by solving this system.

$$\frac{\partial}{\delta\beta_0} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \quad (2)$$

$$\frac{\partial}{\delta\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = 0 \quad (3)$$

Considering that b_0 and b_1 are the solutions to the above system,we can describe the relationship between x and y with the regression line $\hat{y} = b_0 + b_1 x$ indicated by convention.It is easier to solve for b_0 and b_1 by using a centralized linear model:

$$y_i = \beta_0^* + \beta_1 (x_i - \overline{x}) + \varepsilon_i \quad (4)$$

Where $\beta_0 = \beta_0^* - \beta_1 x^-$ .We need to solve for

$$\frac{\partial}{\delta\beta_0} \sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 \overline{x}_i)]^2 = 0 \quad (5)$$

$$\frac{\partial}{\delta\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 \overline{x}_i)]^2 = 0 \quad (6)$$

Taking the partial derivatives with respect to β_0 and β_1 we have

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \overline{x}))] = 0 \quad (7)$$

$$\sum_{i=1}^{n} [y_i - (\beta_0^* + \beta_1 (x_i - \overline{x}))](x_i - \overline{x}) = 0 \quad (8)$$

Note that

$$\sum_{i=1}^{n} y_i = n\beta_0 + \sum_{i=1}^{n} \beta_1 (x_i - \overline{x}) = n\beta_0^* \quad (9)$$

Therefore,we have $\beta_0^* = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$

Substituting β_0^* by $\overline{y}$ we obtain

$$\sum_{i=1}^{n} [y_i - (\overline{y} + \beta_1 (x_i - \overline{x}))](x_i - \overline{x}) = 0 \quad (10)$$

Denote b_0 and b_1 be the solutions. Now it is easy to see

$$b1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{s_{xy}}{s_{xx}} \quad (11)$$

And

$$b_0 = b_0^* - b_1 \overline{x} = \overline{y} - b_1 \overline{x} \quad (12)$$

The explanation behind the LSM is to determine parameter estimates by taking the “closest” row to all data points(x_i, y_i)[16]. Residual calculation have an impact on regression analysis. It must be that residual linear regression can be determined for the examination y_i and the fitted values $\hat{y}_i$ ’s,residuals can be shown. Remembered that the ε_i term is not found in the regression model. Therefore,regression error not found and the residual regression is detected[17]. Naturally the predicted value,the mean of the whole population,is not detected[18,19].

3. LITERATURE REVIEW

Hyun-il Lim. [11] A framework has been developed that uses linear regression when using software instruction-based code vector evaluation software characteristic definition applications. Experimental results suggest that linear regression can be an effective method for classifying the software in software analysis. However, experiments have already been proposed. In other words, a well-designed machine learning model can be easily used for software analysis. Using machine

learning for information analysis is also expected to increase the understanding of software functions.

Wang Xingang. [21] He proposed a weighted Bayes algorithm based on multiple regression (MLWNBC) by calculating weights using the MLR algorithm and removing duplication between properties. In addition, each property determines the influence size of each property according to its weight. MLWNBC makes WNBC more rational (weighted and small deviation classification algorithm). A classification study on 10 datasets in the UCI database showed that this algorithm has excellent performance, which can increase accuracy and reduce consumption time. The dataset is All properties are estimated. Some properties do not affect the result.

4. DATA OVERVIEW

[4]:

	0	1	2
world_rank	1	2	3
institution	Harvard University	Massachusetts Institute of Technology	Stanford University
region	USA	USA	USA
national_rank	1	2	3
quality_of_education	7	9	17
alumni_employment	9	17	11
quality_of_faculty	1	3	5
publications	1	12	4
influence	1	4	2
citations	1	4	2
broad_impact	NaN	NaN	NaN
patents	5	1	15
score	100.0	91.67	89.5
year	2012	2012	2012

Figure 1 The data of top 3 universities

Process the data, get some data, and get the indicators and comprehensive scores of the top 3 universities from the data. The quality of education, alumni employment, quality of faculty, publications, influence, citations, patents are included in the indicators. These metrics affect the university's composite score.

First, observe the average scores of the top ten schools in the world. To do this, one needs to average the scores of the same school in different years. Group schools, combine records from the same school and take an average score. Then sorting by the average score in descending order, and taking the data of the top ten schools for observation and analysis.

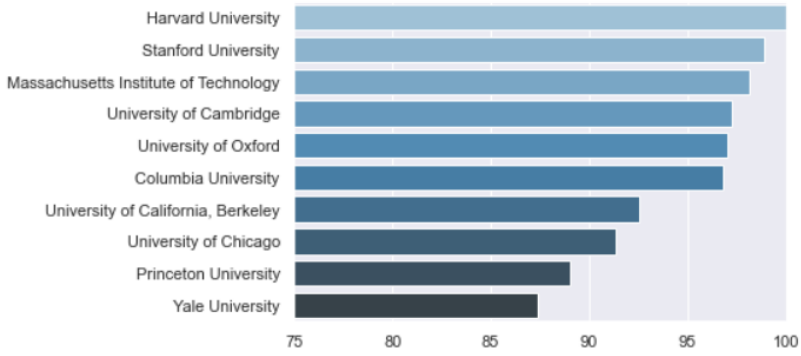

Figure 2 The top 10 university score

It can be seen from the figure2 that among the top 10universities, Harvard University has the highest average score, followed by Stanford University, followed by MIT. Yale University has the lowest average score in the figure.

4.1 Score prediction

The root-mean-square error (RMSE) is a normally accustomed to evaluate of the differences between values (population or sample values) forecast by a model or an estimator and the values detected. RMSD represents the square root of the second sample moment, which is the difference between the predicted and observed values, or the quadratic mean of these differences. When calculating data samples for estimation, these deviations are called residuals, and when calculating out of samples, they are called errors (or prediction errors).[22]

The RMSE of the test set is 3.999, which is an acceptable result under the prediction target of the percentile system. From the evaluation indicators, it seems that we can predict the comprehensive score with better ranking according to various aspects.

4.2 Indicator comparison

Select Tsinghua University, Peking University and Harvard University to compare the data, and observe the gap between Tsinghua, Peking and Top-1 Harvard in various indicators. The radar chart can be observed intuitively, which meets the needs of the experiment, and can visually display the differences between different categories in various indicators.

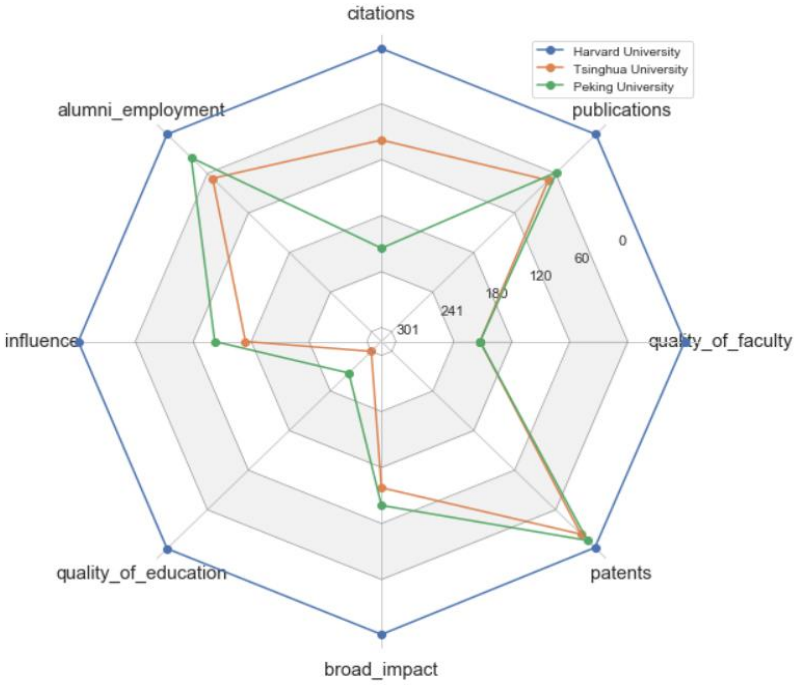

Figure 3 Radar chart comparison

It can be seen that Harvard University ranks among the best in various indicators. Although Tsinghua and Peking ranks higher in terms of publications, patents and employment rates, it has a large gap in the quality of teachers and education, and is at a medium level in terms of influence. From the internal perspective of Tsinghua and Peking, Tsinghua is significantly better than Peking University in terms of citations, but slightly weaker than Peking University in other aspects.

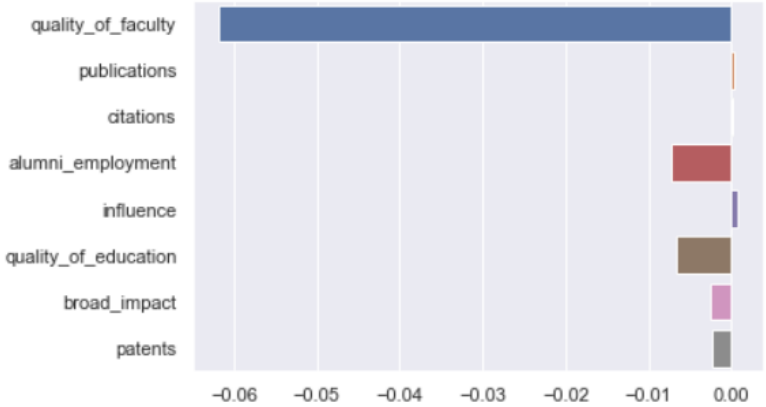

Figure 4 The impact of each indicator on the score

The figure 4 show,quality of faculty ratio is -0.06,alumni employment ration is -0.006,The quality of education ratio is -0.05. By analyzing the proportion of each index, we can know that quality of faculty is the evaluation index that has the greatest impact on the comprehensive score. Both alumni employment and quality of education will have some influence on the comprehensive score. The influence of other indicators accounts for than small.

In order to observe the relationship between the leading factor of "faculty quality",

"alumni employment","quality of education" and the comprehensive score, drawing the scatter plot distribution of quality of faculty and score, a scatter plot of alumni employment, quality of education and score, to find that their correlation with score is weaker than that of quality of faculty.

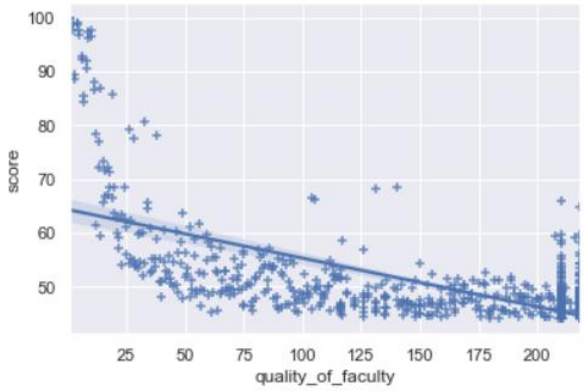

Figure 5 The scatter plot of quality of faculty for score

It can be seen from the figure5 that the distribution of the scatter plot is uneven. When the abscissa is between 0 and 25, the score is in a state of rapid decline and is distributed between 100 and 55. When the abscissa gradually increases, the score stabilizes between 60 and 50, with scattered points between 60 and 70.Linear fitting of quality of faculty and score shows that there is indeed a certain correlation between quality of faculty and score, but it is obviously not linear.

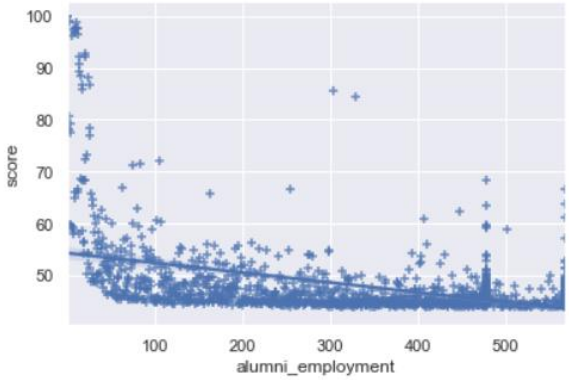

Figure 6 The scatter plot if alumni employment for score

It can be observed from the alumni employment graph that when the abscissa is between 0-100, the score is distributed between 50 and 100, the abscissa continues to increase, the score is stable between 35-55, and the scatter points are fitted, It is found that alumni employment and score have no obvious linear relationship.

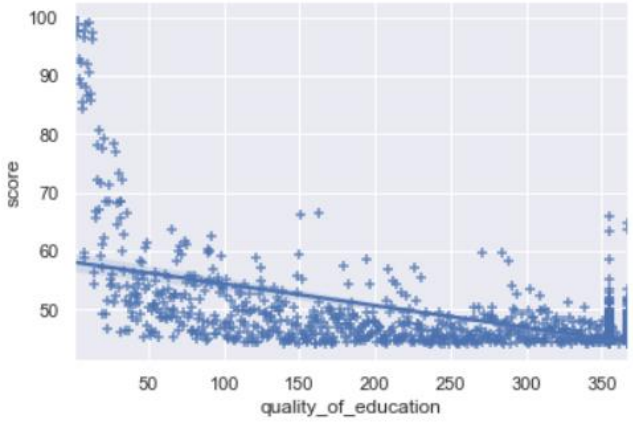

Figure 7 The scatter plot if quality of education for score

From the quality of education graph, it can be observed that when the abscissa is between 0-50, the score drops rapidly from 100 to 40, the abscissa continues to increase, and the score decreases steadily. When the abscissa increases to 250, the score stabilizes at around 40. Fit the scatter points and find that there is no obvious linear relationship between quality of education and score

5. CONCLUSION

Regression modeling is a mathematical method normally applied in research, especially for detected studies. The correct selection of a regression model, the selection and presence of model variables, are key behaviors that must be established and properly controlled to obtain valid statistics, and results may be inaccurate if an appropriate regression model is unavailable or incorrectly applied.

University rankings are related to many factors. Using linear regression to evaluate various indicators, it is found that using linear regression can better predict the comprehensive score of universities according to various aspects. However, there is no direct linear relationship between various indicators on the comprehensive score of the university, but it does have a great impact, among which quality of faculty has the greatest impact.

The linear regression method predicts the overall score of a university and the effect of each factor on the overall score. It enables you to know the ranking of universities objectively and helps people choose a school that is right for them as much as they want. At the same time, it makes the school realize its own shortcomings, which is convenient for improving one's own abilities.

A larger dataset cannot be obtained under limited conditions. As it is limited to the method used in the laboratory, there is a slight error that cannot match the reality in the effect of the overall score by the ratio of the calculation factors. Larger dataset of more accurate results can be obtained by selecting and improving research methods.

REFERENCES

[1] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms: Cambridge university press, 2014.

[2] K. P. Murphy, Machine learning: a probabilistic perspective: MIT press, 2012.

[3] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, pp. 78-87, 2012.

[4] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and Region Growing for Breast Cancer Segmentation," in 2019 International Conference on Advanced Science and Engineering (ICOASE), 2019, pp. 88-93.

[5] Bargarai, F., Abdulazeez, A., Tiryaki, V., & Zeebaree, D. (2020). Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio.

[6] Kaggle(https://www.kaggle.com/mylesoneill/world-university-rankings?select=cwurData.csv)

[7] B. Akgün and Ş. G. Öğüdücü, "Streaming linear regression on Spark MLlib and MOA," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 1244-1247.

[8] M. H. Dehghan, F. Hamidi, and M. Salajegheh, "Study of linear regression based on least squares and fuzzy least absolutes deviations and its application in geography," in 2015 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2015, pp. 1-6.

[9] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine Learning Supervised Algorithms of Gene Selection: A Review," Machine Learning, vol. 62, 2020.

[10] Zebari, D. A., Zeebaree, D. Q., Abdulazeez, A. M., Haron, H., & Hamed, H. N. A. (2020). Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097- 203116.

[11] Abdulazeez, A, M. A. Sulaiman, and D. Q. Zeebaree "Evaluating Data Mining Classification Methods Performance in Internet of Things Applications," Journal of Soft Computing and Data Mining, vol. 1, pp. 11-25, 2020.

[12] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943

[13] J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142

[14] X. Yan and X. Su, Linear regression analysis: theory and computing: World Scientific, 2009.

[15] Y. Fujita, S. Ikuno, T. Itoh, and H. Nakamura, "Modified Improved Interpolating Moving Least Squares Method for Meshless Approaches," IEEE Transactions on Magnetics, vol. 55, pp. 1-4, 2019.

[16] J. Wolberg, Data analysis using the method of least

squares: extracting the most information from experiments: Springer Science & Business Media, 2006.

[17] J.-H. Xue and D. M. Titterington, "$ t $-Tests, $ F $-Tests and Otsu's Methods for Image Thresholding," IEEE Transactions on Image Processing, vol. 20, pp. 2392-2396, 2011.

[18] R. Zhang and J. Tian, "Multi-parameter ocean surface wind speed retrieval based on least square method," in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2016, pp. 5835-5837.

[19] H. Chi, "A Discussions on the Least-Square Method in the Course of Error Theory and Data Processing," in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 486-489.

[20] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943.

[21] X. Wang and X. Sun, "An improved weighted naive bayesian classification algorithm based on multivariable linear regression model," in 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 2016, pp. 219-222.

[22] Hyndman, Rob J.; Koehler, Anne B. (2006). "Another look at measures of forecast accuracy". International Journal of Forecasting. 22 (4): 679–688.