University Ranking Prediction System

Vaibhav Singh[1] Akash Rawat[2] Dr. Pankaj Kumar[3]
[1,2,3]Department of Computer Science and Engineering
[1,2,3]Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, India.

Abstract— In this research paper, we are representing a system of developing university ranking prediction system by analysing global university performance indicators. Here, we've used a uniform dataset of Times education world university rankings. Firstly, we perform pre-processing on the info to see whether the info is suitable to use or not and if not then we continue on further process to correct the dataset. to create the accurate and efficient prediction model, we split the ranking dataset into training and test data. Then, supported score or results of previous years we generate predicted score for every influential feature using our proposed outlier detection and rectilinear regression Algorithm. After it, all the schools are ranked globally supported the anticipated total score predicted by the model. Then, we evaluate the prediction system accuracy based the difference between the anticipated output vs actual output. eventually they are often stratified or guaranteed that our proposed university ranking prediction system is compatible to assess the upcoming global university ranking.

Keywords: Ranking Prediction System, Global University Ranking, Data Analysis, Supervised Machine Learning, Computational Artificial Intelligence

I. INTRODUCTION

Ranking universities may be a basic fundamental issue not just for the scholars and academics but also for the university authorities, industry and even government. Global university ranking purely depends on several key parameters or factors like teaching, research, citations, international outlook, industry income etc. Usually, global university ranking is published on a year-on-year basis or just one occasion during a year. Before publishing the present or present year's ranking, it's essential to assess the upcoming rank of certain university for several reasons. Prospective students require this for applying to specific universities supported upcoming ranks. University authorities should assess the varsity's upcoming rank to enhance certain fields and upgrade their standards comparing with others. Industry and also as even government organizations should also attempt to predict and display the ranking of universities for providing permissions to the foremost eligible and appropriate institutions for subsequent year. Concepts and techniques or technology of knowledge, analysis and supervised machine learning are considerably useful to elucidate the past and predict the longer term by analysing and exploring the info.

There are various aspects or types of national and international university ranking prediction systems has been introduced during which different methodologies are adopted. the days education Global University Ranking system, founded within the UK (England) in 2010, is considered one among the foremost influential, popular and widely used or observed university ranking system. the days education World University Ranking system provides Ranks with the worldwide performance tables which justify research-intensive universities across all their prime objectives which are teaching, research, knowledge transfer and international outlook etc.

The worldwide ranking system is using currently total 13 calibrated performance parameters or indicators to supply the foremost accurate and balanced comparisons trusted by students, teachers, university authorities, industry and government also. In this research paper, we develop a worldwide university ranking prediction system by analysing all the university performance indicators of Times education World University Rankings dataset. First of all, we've made country wise university ranking data analysis so as to form difference within the actual effect of performance indicators and identify the highest most influential factors. we've splitted the entire or whole ranking dataset into training and test dataset for the aim of evaluating our proposed prediction system/model also. We've evaluated the prediction system accuracy supported the difference between the particular and predicted result or output. Therefor we've founded out that our proposed global university ranking prediction system is compatible for estimating upcoming global university ranking.

II. MOTIVATION

University ranking is extremely beneficial for college kids, teachers, university board trustees and government. Different ranking systems published varsity ranking yearly or biannually. People can see or check the present rank of a university from those ranking providers. But ranking may be a transitional and everchanging process, it changes time to time consistent with the parameters. If upcoming rank of any university are often predicted, it might certainly help or assist the potential graduate students to settle on desired institution beforehand as they might have longer.

Moreover the university authorities can observe their present situation and check out to enhance each of the influential performance indicators before publishing their upcoming rank. Prediction of world ranking of global universities are going to be considerably very helpful to the donors also as for government so as to form the choice of continuation or approval of grant for a selected university.

So, the sooner as we stated earlier that within the prior assessment or checking of performance indicators as they were for instance features teaching, research, citations, international outlook isn't only necessary to predict a university rank but also to supply future insight of admission, provision of grants, university development indicators. that is often, why we are motivated to form an analysis of the many and important performance indicators of a well-reputed university ranking system and build a university ranking prediction system which may easily ready to predict the rank of worldwide universities with more precisely and effectively with high accuracy.

III. Related work

Ranking may be a very crucial and important research topic in Information Retrieval process. Feature selection is one among the elemental issue or a drag within the ranking model. Previously, another study examined whether a university is to be ranked or not by analysing the impact of worldwide rankings in education. Recently there are many increasing interests in university rankings also because the comparison among the ranking systems. A recent study has shown, also compared different world university rankings employing a set of similarity measures parameters.

The comparisons during a research paper revealed that ranking results or outcome can vary, sometimes dramatically, thanks to methodologies and emphases of varied criteria. Considering of those relatable comparative analysis as, a unifying framework that's being used for ranking predictions has been proposed from training data called Boosted Ranking Model.

There is another study done that shows the likelihood to predict usability of a university website from university ranking systems. Of those research based studies encouraged us to analyse the influential performance indicators or parameters of a well reputed university ranking system and develop a university ranking prediction system which can predict the rank of worldwide universities efficiently and accurately.

IV. Methodologies

A. Data Analysis

For developing university ranking prediction system, here we exploit publicly available dataset of worldwide university rankings. At first, we analyse the dataset of worldwide university rankings to hunt out the influential performance indicators shown within the table present in feature selection. There are several different parameters or attributes or features within the dataset which are university name, country of the university, score of teaching, research, citations, income, international, total score, number of students, ratio of between student and staffs, number of international students, ratio of female and male including teachers and students, year of rank etc. described within the Table given in feature selection.

B. Features Selection

The dataset of Times education global university rankings consists of 13 features mentioned as performance indicators of a university. we've analysed country wise influence of all the performance indicators in last two years and located that the variation of scores in teaching, research, citations, international outlook mostly influence the ranking of universities as depicted in Fig below and Fig. then. So, we've tried to developed an efficient, effective and accurate global university ranking prediction system by utilizing these features of the dataset. as an example, if we glance at Caltech University data, we see, the other indicators e.g., student staff ratio, number of students, international student acceptance rate and female to male ratio remain constant no matter ranking.

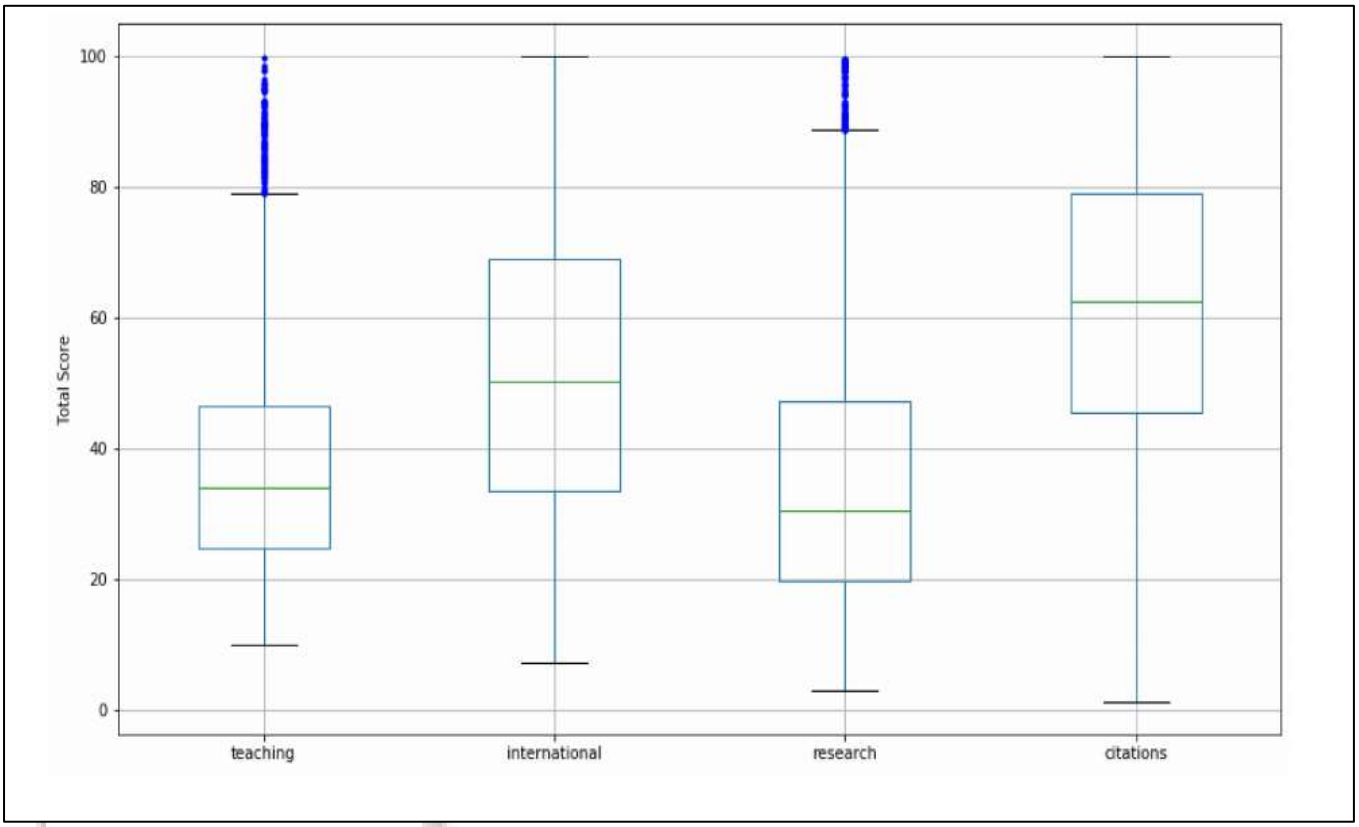
Analysis of Dataset for University Rank Prediction

university_name	Different universities all around the world
country	Different Countries around the world
teaching	Score indicating the learning environment
international	Score combining staff, students and research outlook
research	Score based on volume, income and reputation in research
citations	Score representing research influence
income	Score indicating industry income by knowledge transfer
total_score	total score combining weighted scores of other performance indicators
num_students	number of full time equivalent students at the University
student_staff_ratio	ratio of full time equivalent students to the number of academic staff those involved in teaching or research
international_students	percentage of students originating from outside the country of the University
female_male_ratio	ratio of female to male students at the University
year	Years from 2011-2016

Therefore, we eliminate these features from the performance indicators initially. For the rest five features (3,4,5,6,7 of Table above) we consider different weights to calculate total score and our predicted or calculated total score matches relatively more with the times total score once we consider negligible weight for income.

C. Outlier Detection

We have acknowledged that a feature of a selected university doesn't follow a selected trend or pattern throughout the years. as an example, analysing the data we see that the points aren't following any similar trend depicted in Fig below.



D. Z-scores for Removal of Outliers

Z-scores are can be used to quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the amount of ordinary deviations above and below the mean that each value falls. as an example, a Z-score of two indicates that an observation is 2 standard deviations above the standard while a Z-score of -2 signifies it's two standard deviations below the mean. A Z-score of zero represents a worth that is equal to the mean.

To calculate or compute the Z-score of the dataset for an observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is that the following:

$$Z = \frac{X - \mu}{\sigma}$$

The further away an observation's Z-score is from zero, the weird it's. a typical cut-off value for locating outliers are Z-scores of +3/-3 or beyond zero. The probability distribution below displays the distribution of Z-scores during a typical normal distribution. Z-scores beyond +3/- 3 are so extreme you'll barely see the shading under the curve.

E. Model Creation and Evaluation

Linear Regression could even be a machine learning algorithm supported supervised learning which performs a regression task on the data or dataset. Regression models calculate a prediction value supported independent variables. it's mostly used for locating out the connection between variables and forecasting.

See the below Fig. to know how Model is formed.

```
In [122]: # Import library for Linear Regression
          from sklearn.linear_model import LinearRegression

In [123]: LR = LinearRegression()

          # Train the model using the training sets
          LR.fit(X_train, y_train)

Out[123]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [124]: #value of y intercept
          LR.intercept_

Out[124]: -2609.0759991655
```


Model Evaluation could even be a crucial, necessary, and integral a neighborhood of the model development process. It helps find the only model that represents our data and thus the way well that chosen model will add the top of the day. Evaluating or measuring of the model performance with the help of training dataset isn't acceptable in machine learning because it can easily generate overoptimistic and overfitted models. There are two methods or process for evaluating models in data science named Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to Guage model performance.
See the Fig. below for Model Evaluation.

Model evaluation

```
In [126]: # Model prediction on train data
          y_pred = LR.predict(X_train)
          y_pred

Out[126]: array([46.00272073, 50.77053041, 65.53981316, ..., 61.44199404,
                 47.05003465, 70.02478732])

In [127]: # Model Evaluation
          print('R^2:',metrics.r2_score(y_train, y_pred))
          print('Adjusted R^2:',1 - (1-metrics.r2_score(y_train, y_pred))*(len(y_train)-1),
          print('MAE:',metrics.mean_absolute_error(y_train, y_pred))
          print('MSE:',metrics.mean_squared_error(y_train, y_pred))
          print('RMSE:',np.sqrt(metrics.mean_squared_error(y_train, y_pred)))

          R^2: 0.7896914796760776
          Adjusted R^2: 0.7885903879466329
          MAE: 3.323118095097534
          MSE: 16.997582221540107
          RMSE: 4.122812416487089
```

F. University Rank Prediction System

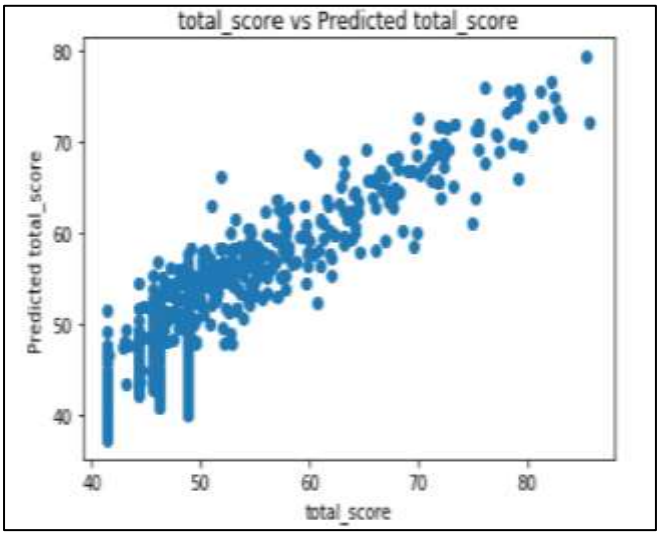
Firstly, we've considered or understood the university ranking dataset of Times education. Then we've selected the foremost influential performance indicators by analysing year wise variation of scores in various universities. to make also on evaluate the prediction model, we split the dataset as training data for year 2011 to 2015 and left the data of year 2016 for test purpose. Deploying the training dataset, we've detected outliers for each performance indicators using our proposed Algorithm named linear regression. Then, we've calculated the anticipated rank score of teaching, research, citations, international outlook using the linear regression Algorithm. then, we've generated total predicted rank score supported certain weight of each performance indicators. Finally, we've ranked universities globally using the anticipated total rank score. the entire process is illustrated within the flow chart of Fig.

V. CONCLUSIONS

In this research paper, we are trying to develop and present a system, how of developing a worldwide university ranking prediction system by analysing all the important university performance indicators. Here, we've used the dataset of Times education World University Rankings which comprises of worldwide performance tables that judge universities supported several performance criteria and parameters. we've splitted the ranking dataset into training and test dataset for the aim of building and evaluating our proposed prediction system. At first, we've made country wise university ranking data analysis to observe the variation of performance indicators and determine the foremost influential factors. Then, supported previous years score we've generated predicted score for those influential attributes using our proposed outlier detection method. After that, supported predicted or calculated result we've ranked all the faculties.

Finally, we've evaluated the prediction system accuracy. Thus, we've found out that our proposed method or model for university rank prediction system is suitable to estimate upcoming global university ranking.
Fig. Shown Below is that the difference between the anticipated and actual price.

```
In [131]: # Visualizing the differences between actual prices and predicted values
          plt.scatter(y_test, y_test_pred)
          plt.xlabel("total_score")
          plt.ylabel("Predicted total_score")
          plt.title("total_score vs Predicted total_score")
          plt.show()
```

ACKNOWLEDGMENT

We would like to gratefully acknowledge support of our guide and mentor for their generous assistance to get this project completed. Without their support and guidance it would not be possible to make this project.

We would also like to thank the Times Higher Education for providing us the Dataset required for the project of various world universities.

REFERENCES

[1] https://en.wikipedia.org/wiki/Linear_regression#:~:text=In%20statistics%2C%20linear%20regression%20is,as%20dependent%20and%20independent%20variables).&text=Such%20models%20are%20called%20linear%20models.
[2] C. C. Aggarwal, Data mining: the textbook. Springer, 2015
[3] https://www.researchgate.net/publication/315853600_University_ranking_prediction_system_by_analyzing_influential_global_performance_indicators
[4] "Dataset of world university rankings," https://www.kaggle.com/mylesoneill/world-university-rankings, accessed: 2016-10-10.
[5] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, pp. 407–414.
[6] M. N. Volkovs and R. S. Zemel, "Boltzrank: learning to maximize expected ranking gain," in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 1089–1096
[7] I. F. Aguillo, J. Bar-Ilan, M. Levene, and J. L. Ortega, "Comparing university rankings," Scientometrics, vol. 85, no. 1, pp. 243–256, 2010.
[8]] L. Hasan and E. Abuelrub, "Is it possible to predict usability of a university website from university ranking systems?" in Proceedings of the World Congress on Engineering, vol. 2, 2013, pp. 3–5.
[9] V. M. Moskovkin, E. V. Pupynina, N. P. Zaitseva, and R. V. Lesovik, "Methodology for comparative analysis of university rankings, with the mediterranean and black sea region countries taken as an example," Methodology, 2013
[10] https://pandas.pydata.org/
[11] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[12] https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7
[13] https://www.w3schools.com/python/python_ml_train_test.asp
[14] https://en.wikipedia.org/wiki/Training,_validation,_and_test_sets
[15] https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/#:~:text=One%2DClass%20SVM-,Outlier%20Detection%20and%20Removal,t%20fit%20in%20some%20way.&text=In%20this%20case%2C%20simple%20statistical,deviations%20or%20the%20interquartile%20range.
[16] http://scikit-learn.org/stable/modules/outlier_detection.html
[17] https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

