Predicting Global Ranking of Universities Across the World Using Machine Learning Regression Technique

Dr. Prakash Kumar Udupi [1]
[1]*Middle East College, Muscat, Sultanate of Oman*
prakash@mec.edu.om
www.mec.edu.om

Dr. Vishal Dattana [2]
[2]*Middle East College, Muscat, Sultanate of Oman*
vishal@mec.edu.om
www.mec.edu.om

Dr. Netravathi P.S.[3]
[3]Srinivasa University
Managalore, India
nethrakumar.ccis@srinivasuniversity.edu.in
www.srinivasuniversity.edu.in

Jitendra Pandey [4]
[4]*Middle East College, Muscat, Sultanate of Oman*
jitendra@mec.edu.om
www.mec.edu.om

Abstract

Digital transformation in the field of education plays a significant role especially when used for analysis of various teaching and learning parameters to predict global ranking index of the universities across the world. Machine learning is a subset of computer science facilitates machine to learn the data using various algorithms and predict the results. This research explores the Quacquarelli Symonds approach for evaluating global university rankings and develop machine learning models for predicting global rankings. The research uses exploratory data analysis for analysing the dataset and then evaluate machine learning algorithms using regression techniques for predicting the global rankings. The research also addresses the future scope towards evaluating machine learning algorithms for predicting outcomes using classification and clustering techniques.

Keywords: Digital Transformation, Teaching and Learning, Machine Learning, Regression, Classification, Clustering.

1. Introduction

Digital transformation facilitates organisations to understand their data and information, analyse these data for improve operational performances and competitive advantages using digital technologies (Peter et al., 2021). Data science as a part of digital transformation enables the organisations to generate new business models, develop strategy, create roadmap and build competitive advantage based on the understanding of information and patterns contained within the data. In order to establish new benchmark in the higher education domain, it is also important that the higher educational institutes needs to understand the best practices from the global context, indicators used for measuring performances and evaluation criteria (Vitenko et al., 2021).

Machine learning is a subset of artificial intelligence, which facilitates machine to learn from data using algorithms and predict the results without explicitly programming (Awad & Khanna, 2015). Machine learning helps the organisation to study the business operations, customer behaviours, analyse the patterns derived from the data, develop predictions and prepare relevant strategies. Machine learning problems are broadly classified as classification, clustering and regression problems (Sarkar,2021). Classification and regression are categorised under supervised learning and clustering is categorised under unsupervised learning (Alloghani et al., 2020). If the output variable is discrete, then classification or clustering can be applied, whereas if the output variable is continuous, then regression technique can be applied. Hence, machine learning technique can be used to study and learn historic global university ranking data, identify the patterns and predict the university rankings (Estrada & Cantu, 2022). Further, global ranking of top universities are continuous and regression technique is most suitable.

2. Study of machine learning regression framework

Machine learning framework begins with data gathering, and data pre-processing as shown in the below figure 1.

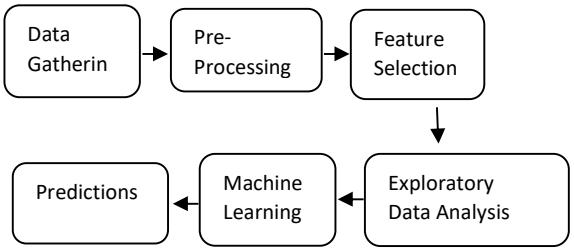

Figure 1. Machine learning regression technique framework

After the pre- processing, feature selection will be carried out. Exploratory data analysis will be performed after feature selection to find out various statistical parameters pertain to the selected dataset. Exploratory data analysis helps to understand behaviour of dataset, behaviour of variables, understanding of data attributes and characteristics of dataset. Once the exploratory analysis

completed, Machine learning algorithms are applied to predict the outcomes. The dataset will be split as training and testing dataset before input to the machine. Machine learns the dataset using various algorithms and predicts the outcomes.

3. Study of existing world university ranking framework

World university ranking is an annual declaration of top university ranking popularly conducted and published by Quacquarelli Symonds organisation. Relevant indicators and corresponding weightings are showcased in table 1.

Indicator	Weighting	Elaboration
Academic peer review	40%	Based on an internal global academic survey
Faculty/student ratio	20%	A measurement of teaching commitment
Citations per faculty	20%	A measurement of research impact
Employer reputation	10%	Based on a survey on graduate employers
International student ratio	5%	A measurement of the diversity of the student community
International staff ratio	5%	A measurement of the diversity of the academic staff

Table 1. Indicators and weightages

QS system rankings are comprising of three parts named as global overall ranking, subject wise global ranking and region wise rankings. QS world rankings are based on performance evaluation of key aspects such as teaching, research, employability, university mission and internationalisation.

4. Data collection, pre-processing and exploratory analysis

Top universities world ranking dataset for the year 2022 are used as a basis for predicting the outcomes using machine learning techniques. 21 column variables in the form of indicators are used as features and one column variable named global rankings is used as target variable. Total 1300 rows of data were used for pre-processing. After the removal of missing values, total 1225 rows of data were used for analysis. In order to understand characteristics of various data and relationship between variables, correlation test is applied into the dataset (Kumar & Chong, 2018).

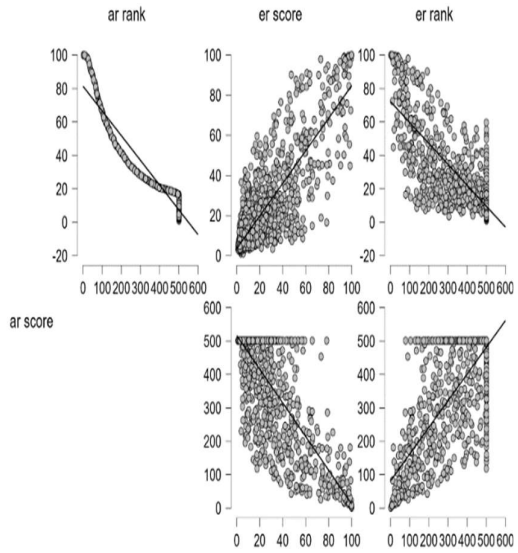

Figure 2. Scatterplot of feature variables

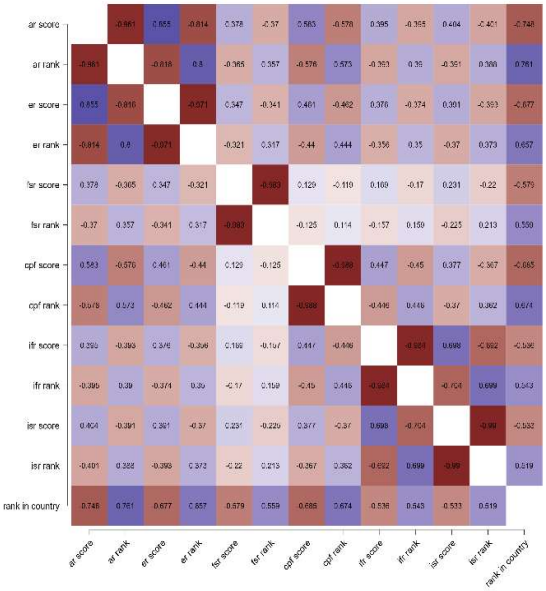

Figure 3. Heatmap of feature variables

Scatter plot of feature variables are illustrated in Figure 3 and relevant heatmap derived from correlation test is illustrated in Figure 3. Total 784 rows of random dataset was used for training, 196 rows of random dataset used for validation and 245 rows of random dataset data were used for testing purposes.

5. Result analysis

Application of regression algorithm into global university ranking dataset derives 66 trees and to achieve better accuracy, boosting regression is applied (Velthoen et al., 2021). To get increased predictive outcomes, Gaussian

loss function is used within the boosting regression (Sigrist,2021). The resultant outcome of boosting regression is demonstrated in Table 2. Test results are evaluated using mean square error(MSE) values. The lower mean square error value depicts better test result.

Boosting Regression							
Trees	Shrinkage	Loss function	N (Train)	N (Validation)	N (Test)	Validation MSE	Test MSE
66	0.100	Gaussian	784	196	245	0.170	0.201

Table 2. Boosting regression result

In regression problems, to check the performance of algorithms or models, evaluation metrics are used (Alexei,2018). In evaluation metrics, test results are compared using various values such as root mean square error (RMSE) value, mean absolute deviation (MAD), mean absolute error (MAE), mean absolute percentage error (MAPE) and R^2 (R-Squared or coefficient of determination). Table 2 demonstrates evaluation metrics.

Evaluation Metrics	Value
MSE	0.201
RMSE	0.448
MAE / MAD	0.382
MAPE	154.19%
R^2	0.841

Table 3. Evaluation Metrics

Relative influence helps to communicate the importance and percentage contributions of each variable for the results. Table 4 demonstrates the variable details and relevant percentage contributions for deriving the outputs. From the evaluation matrices, we observe that the R^2 value is.0841. Hence goodness of fit is 84.1%, which showcase that the data can be fitted well within the regression model and good prediction of global ranking possible using the selected variables in terms of information.

Relative Influence	Relative Influence
score scaled	76.584
ar score	7.649
fsr score	4.261
cpf score	4.090
country code	3.345
er score	1.997
country	0.689
fsr rank	0.586
ifr score	0.527
isr score	0.271
size	0.000
focus	0.000
research	0.000

Relative Influence	Relative Influence
age band	0.000
status	0.000
ar rank	0.000
er rank	0.000
cpf rank	0.000
ifr rank	0.000
isr rank	0.000

Table 4. Relative influence of variables

Relative influence plot provides visual illustrations of variable contributions for the overall results in terms of a plot. Figure 4 illustrates the relative influence plot. From the plot, we observe that scaled score contributes the most for predicting the ranks.

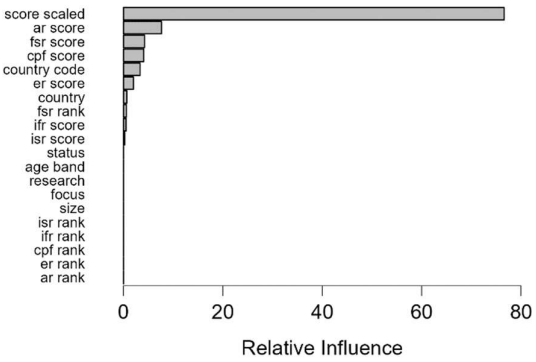

Figure 4. Relative influence plot

Out of bag improvement plot provides distribution of training data with reference to out of bag changes in Gaussian deviance versus number of trees. The out of bags improvement plot helps to estimate the prediction accuracies of boosting regression. Figure 5 illustrates OOB changes versus number of trees for the training dataset. We observe that as the tree approaches to 66, accuracies established and deviances reduced.

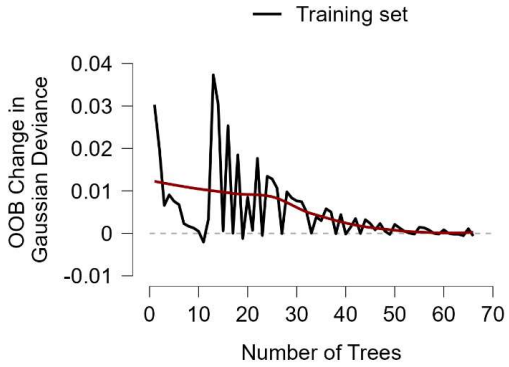

Figure 5. OOB improvement plot

Predictive performance plot provides the visual representation of predicted value versus observed value. Figure 6 demonstrates there is a linear relationship exists between of observed test values versus predicted test values.

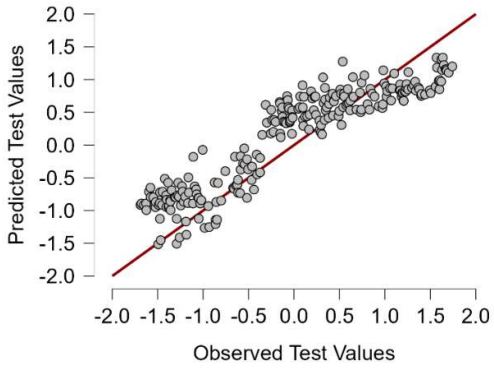

Figure 6. Predictive performance plot

Deviance plot provides the graphical representation of Gaussian deviance with reference number of tree formation. Figure 7 illustrates as the number of tree increases; deviance reduces and more prediction accuracy can be achieved.

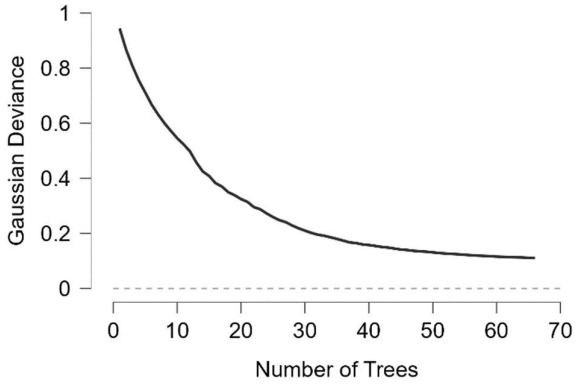

Figure 7. Predictive performance plot

6. Conclusion

Knowing global ranking among the top universities of the world is a complex process because various parameters of each universities has to be evaluated before allocating the rankings. Quacquarelli Symonds dataset used for predicting global ranking consists of 21 different performance parameters based on 6 different indicators with different weightages. Hence robust dataset with complex data structure was used for predicting global rankings. When applied regression technique as a part of machine learning algorithm, it has been observed that good predictions of global ranking possible. For the individual institutions, knowing its global ranking helps to understand its capabilities and competitiveness in turn helps to develop strategies to perform better. The research can be further extended and classification and clustering techniques also can be applied, Application of clustering and classification techniques helps the universities or academic institutions to understand on which group they fall or under which category they perform with in the groups.

References

[17]. Verhoef, C.P., Broekhuizen, T.,Bart, Y.,Bhattacharya, A.,John.,Fabian, N.,Haenlein, M.(2021).Journal of Business Research. https://doi.org/10.1016/j.jbusres.2019.09.022
[18]. Vitenko, T., Shanaida, V., Droździel, P., Madlenak, R.(2018). Assessment of higher education in global environmnet.INTED2018 Proceedings, pp. 4040-4045
[19]. Awad, M., Khanna, R. (2015). Machine Learning. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_1
[20]. Sarker, I.H.(2021).Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160. https://doi.org/10.1007/s42979-021-00592-x
[21]. Alloghani, M., Obe, D., Mustafina, J., Hussain, A., Aljaaf, A.(2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Supervised and Unsupervised Learning in Datascience. Springer. https://doi.org/10.1007/978-3-030-22475-2_1
[22]. Estrada-Real, A.C., Cantu-Ortiz, F.J.(2022). A data analytics approach for university competitiveness: the QS world university rankings. Int J Interact Des Manuf 16, 871–891. https://doi.org/10.1007/s12008-022-00966-2
[23]. Kumar, S., Chong, I. (2018). Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States. Int J Environ Res Public Health.15(12):2907. doi: 10.3390/ijerph15122907. PMID: 30572595; PMCID: PMC6313491.
[24]. Velthoen, V., Dombry, C., Cai, J., Engelke, S., (2021). Gradient boosting for extreme quantile regression.arXiv.https://doi.org/10.48550/arXiv.2103.00808
[25]. Sigrist, F.(2021). Gaussian Process Boosting.arXiv.https://doi.org/10.48550/arXiv.2004.02653
[26]. Alexei, B.(2018). Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio.http://dx.doi.org/10.2139/ssrn.3177507