Analysis of different university ranking systems around the world and predicting rankings of the universities for future years

Project report for IITB DS203 Programming for Data Science 2021

Link to GitHub repository

Rohan Kalbag
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
20d170033@iitb.ac.in

Asif Shaikh
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
20d070017@iitb.ac.in

Kalp Vyas
Dept. of Electrical Engineering
Indian Institute of Technology Bombay
kalp.vyas@iitb.ac.in

Abstract—**Quality education is essential for socio-economic growth, better standards of living and advances in technology. There are numerous ranking systems globally that rank university on the quality of overall education imparted by them. Our work involves Exploratory Data Analysis of the datasets of three of these ranking systems. The inferences from this shall be used to identify trends/biases in these ranking systems. Relation of the rank with various parameters and scores used by the ranking system shall also be identified. Using the aggregate score over past years a predictive model shall be trained to predict future rankings using the principles of regression.**

I. Introduction

Education is essential for every person. It is the acquisition of knowledge, skills that make a person more rational, scientific and employable. Education is not imparted uniformly everywhere. A pressing question currently is the identification of colleges that impart better education than others. As a student or researcher, which institution one should choose is a matter of confusion even today. Because education is a rare purchase and an increasingly important as well as expensive decision in one's life, students and researcher seek information that allow them make informed choices in the selection of a university and/or an academic program. Demand for consumer information on academic quality has led to the development of university rankings in many countries of the world.

University rankings are often heavily criticized because of their statistical inaccuracy, because of the measures chosen to represent academic quality, or because of their expected negative impact on the overall performance of universities. But research suggests that well designed organizational report cards can sometimes serve as effective instruments for public accountability. The provision of relevant information about universities to students and researchers allow them to be paired with an institution of their interest as well as be confident about the quality of education imparted there.

In this project, we will be looking at 3 such ranking systems: the QS (Quacquarelli Symonds) ranking system, ARWU (Academic Ranking of World Universities) ranking system also known as the Shanghai Rankings for global universities and the NIRF (National Institutional Ranking Framework) ranking system for Indian universities. This project will find out trends and biases in the different ranking systems as well as try to use different ML frameworks to predict the scores and the ranks given by the ranking systems for the future years after comparing how well it does on a test year.

II. Background and Prior Work

QS World University Rankings are published annually by Quacquarelli Symonds (QS). Between 2004 and 2009 QS published its university rankings in partnership with Times Higher Education (THE) magazine. Since 2010 THE and QS publish their own university rankings using separate methodologies. QS still uses the methodology used prior to 2010. The QS University Rankings along with Academic Ranking of World Universities (ARWU) and Times Higher Education (THE) World University Rankings is considered as one of the three most-widely read university rankings in the world. The QS rankings use six different indicators, namely

- Academic Reputation (40%)
- Employer Reputation (10%)
- Faculty/Student Ratio (20%)
- Citations per faculty (20%)
- International Faculty Ratio (5%)
- International Student Ratio (5%)

The Academic Ranking of World Universities (ARWU), also known as the Shanghai Rankings is published annually starting from 2003 which makes it the oldest global university ranking system. It was originally published by Shanghai Jiao Tong University and now by the Shanghai Ranking Consultancy. The Universities are ranked using six academic or research performance indicators listed below.

- Alumni (10%) - Alumni of an institution winning Nobel Prizes and Fields Medals

- Award (20%) - Staff of an institution winning Nobel Prizes and Fields Medals
- HiCi (20%) - Highly Cited Researchers
- N&S (20%) - Papers published in Nature and Science
- PUB (20%) - Papers indexed in Science Citation Index-Expanded and Social Science Citation Index
- PCP (10%) - Per capita academic performance of an institution

The National Institutional Ranking Framework (NIRF) is a methodology established by the Ministry of Education (MOE), formerly known as the Minister of Human Resource Development (MHRD), to rank higher educational institutes in India. The rankings are published annually since 2015 by the Ministry of Education. The framework uses several parameters for ranking institutes. These parameters are grouped into five metrics and each of the five metrics have been assigned a weight depending upon the type of the institute. The approved set of parameter groups are,

- TLR (30%) - Teaching, learning and resources
- RPC (30%) - Research, professional practice and collaborative performance
- GO (20%) - Graduation outcome
- OI (10%) - Outreach and inclusivity
- PR (10%) - Perception

III. Data and Methodology

The datasets were sourced from the NIRF, QS and ARWU websites. The datasets contain the ranks, parameter scores and geographical categorical data of all the top-ranking universities for the respective ranking systems.

The data for the following years were obtained from the sources mentioned above.

Ranking	Datasets Obtained	Number of years
NIRF	2017 - 2021	5
QS	2018 - 2020	3
ARWU	2005 - 2018	14

The data was cleaned and processed to be analyzed. Various plots were made in order to interpret the statistical properties of the data. Histograms were made for each of the parameters in order to study the nature of the distribution. For yearly and parameter-wise comparison of the distribution, and to identify outliers, Box and Whiskers plots were made. The mean value of parameters in uniform batches of 10 were plotted against the rank to study the dependency of the rank on various parameters. Annual heatmaps of the correlation matrix were plotted to study the correlation of various parameters with each other. The annual variation of the parameters of the top 10 ranked universities were also plotted. Plots were made to infer about the geographical distribution of these universities.

IV. Experiments and Results

Statistical analysis for geographical distribution:

Fig 1 shows a histogram of the countries the universities ranked in the QS rankings are present in. The United States, United Kingdoms, Australia and Canada have the highest number of universities in that order with the US and the UK having a distinctively large fraction of the total universities. These are followed by oriental countries such as China, Japan and South Korea

For the ARWU ranking, in Fig 2 we plot a histogram of no of top ranking universities present in a particular country. We notice that a majority of the top-ranked colleges are situated in the USA and UK, followed by Germany and then the oriental countries. Figs 3 and 4) show the distribution of the total number of the universities present in the continents for years 2005and 2018. These show that there is a shifting of educational quality from the west to the east over the years, indicating that the education is become in a sense ”global” and not local to the western countries alone. The number of education institutes in Asia and Australia is improving and worsening in North America.

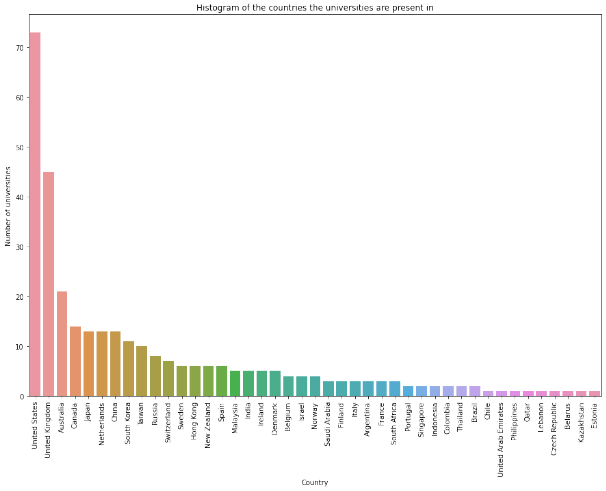

Fig. 1. Country Wise Distribution for QS

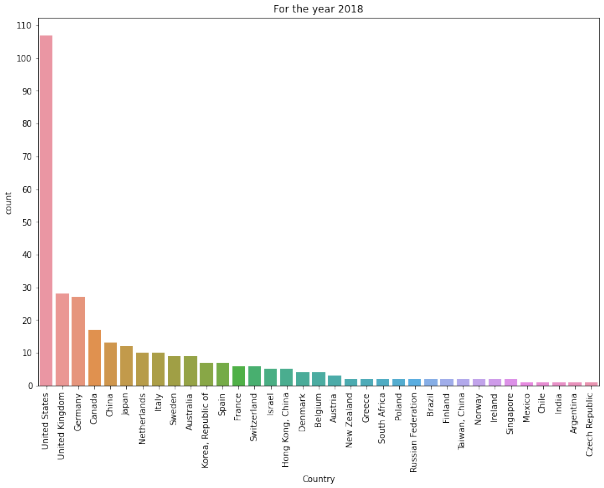

Fig. 2. Country Wise Distribution for ARWU

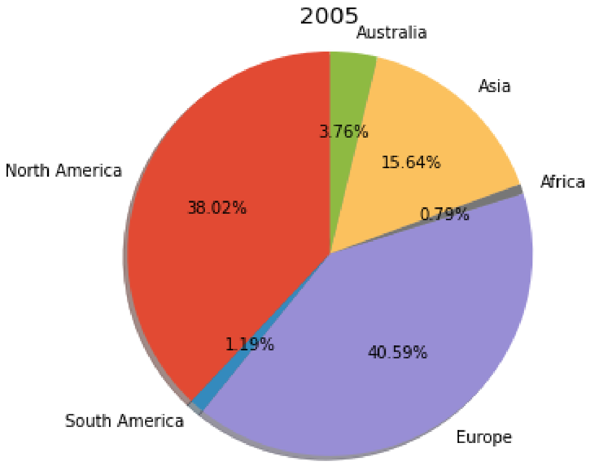

Fig. 3. Continent Wise Distribution for ARWU in 2005

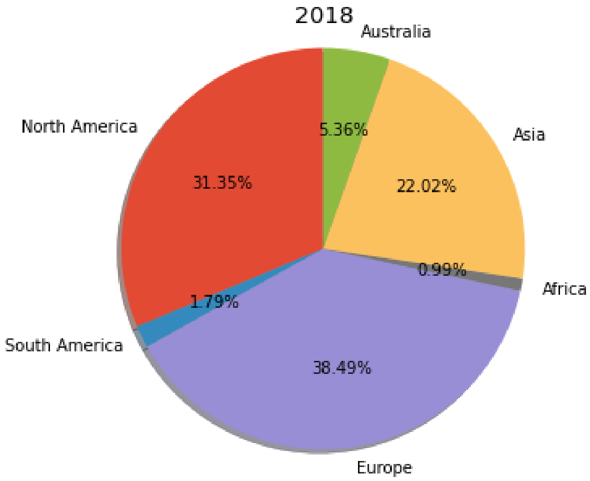

Fig. 4. Continent Wise Distribution for ARWU in 2018

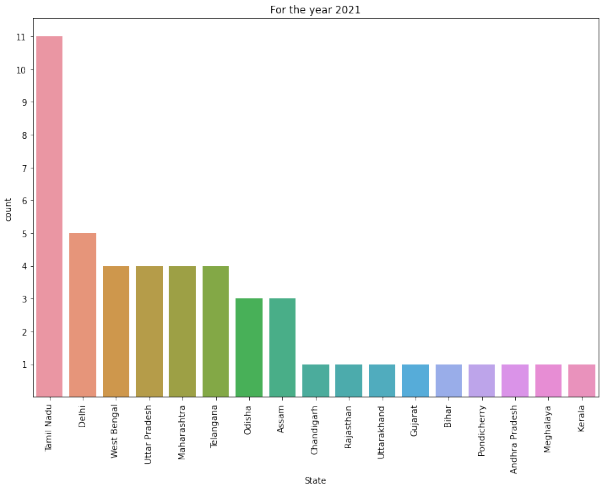

Fig. 5. State Wise Distribution for NIRF

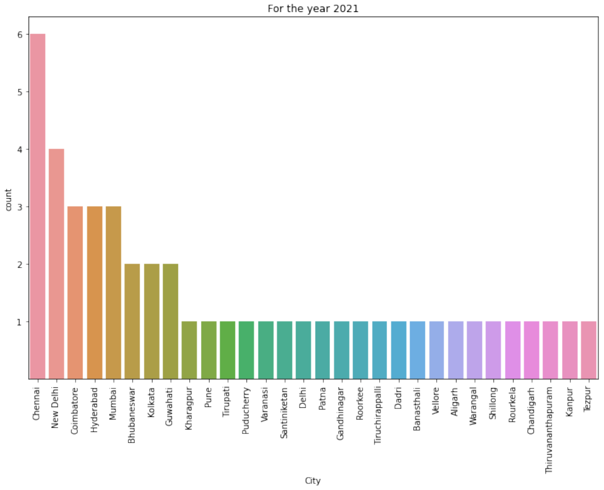

Fig. 6. City/Town Wise Distribution for NIRF

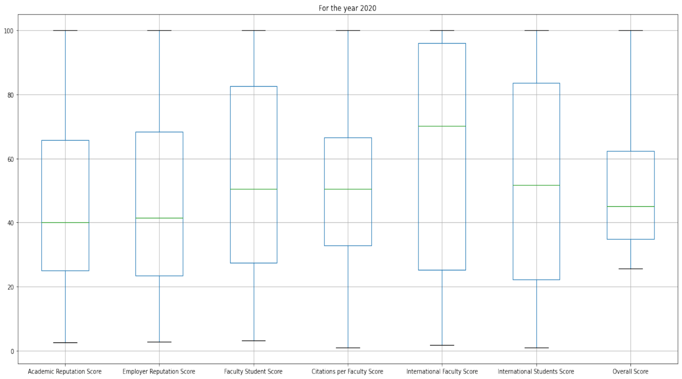

Fig. 7. Box and Whiskers Plot for QS

For the NIRF ranking, it was observed that the higher-ranked colleges are localised in a few states such as Tamil Nadu, Maharashtra, Delhi, Uttar Pradesh and West Bengal in Fig 5 we plot a histogram of no of top ranking universities present in a particular state. Tamil Nadu has a very large no of higher-ranked universities compared to the other states. Also, it was observed in Fig 6 we plot a histogram of no of top ranking universities present in a particular city. that a majority of the higher-ranked universities are located in metropolitan cities and state capitals rather than in relatively smaller towns and villages. This indicates that higher quality of education is localised and to only large cities, this could be because of better infrastructure and opportunities in cities and social life that students and faculty enjoy in cities.

Statistical analysis of ranking parameters:

For the QS World Rankings

In Fig 13 we plot the time average over all years of the average value of all the parameters for batches of 10 ranks each. It is noticed that having a lower/better QS Rank corresponds to a higher Academic Reputation Score, Employer Reputation Score, Faculty Student Score, Citations

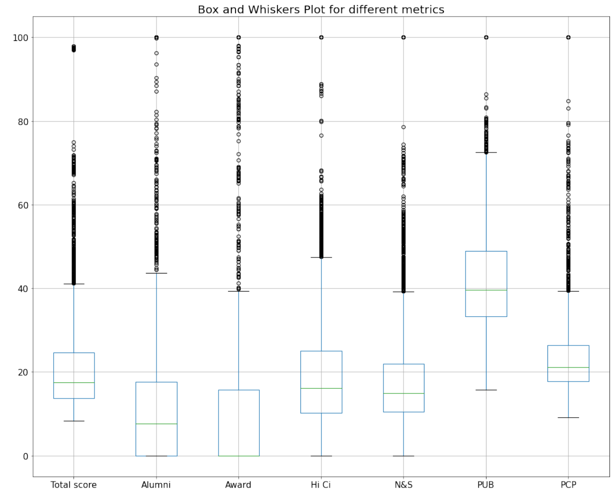

Fig. 8. Box and Whiskers Plot for ARWU

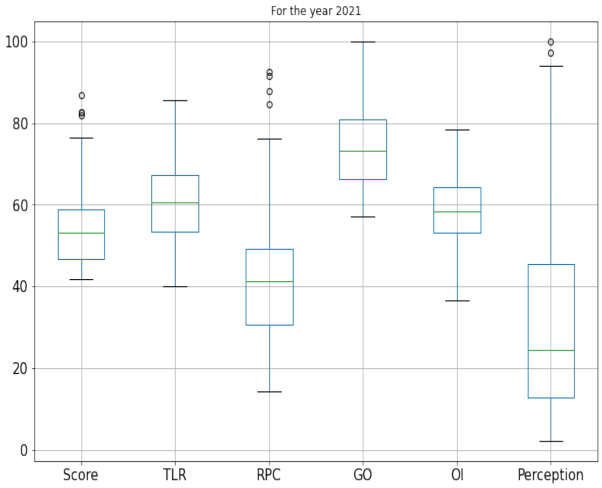

Fig. 9. Box and Whiskers Plot for NIRF

Per Faculty Score and Overall Score. However, we observe that there is as such no strictly increasing/decreasing relation of Rank with International Student and Faculty Scores as these show fluctuating behavior with an increase in Rank.

In Fig 10 the heatmap can also be used to study the correlation of various parameters with each other. Rank is highly negatively correlated with Academic Reputation Score, Employer Reputation Score and Overall Score and also has a small negatively correlation with Faculty-Student, Citations per Faculty. This means the parameters decrease rapidly and decrease as the ranks increase respectively. However we see that International Faculty and International Student score have nearly 0 to minimal correlation with rank as we had concluded from the previous paragraph.

Another observation that can be made is that Score has very high positive correlation with Academic Reputation Score, Employer Reputation Score and Overall Score

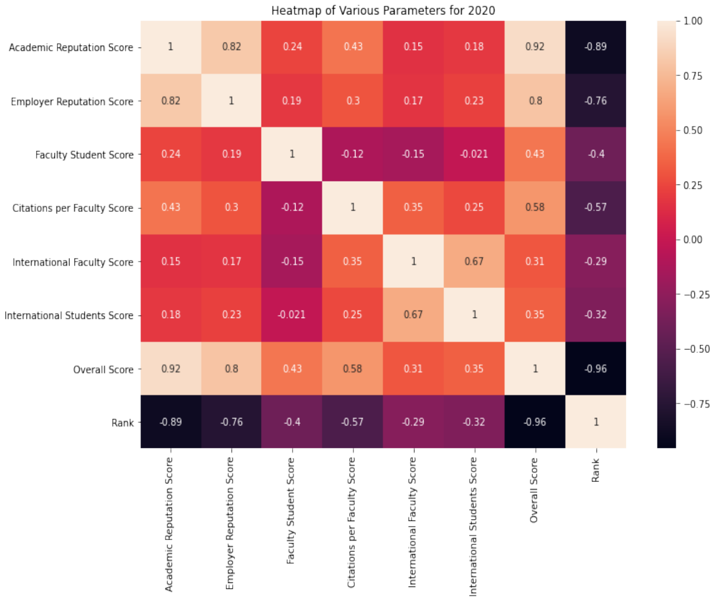

Fig. 10. Heatmap of correlation matrix of various parameters for QS

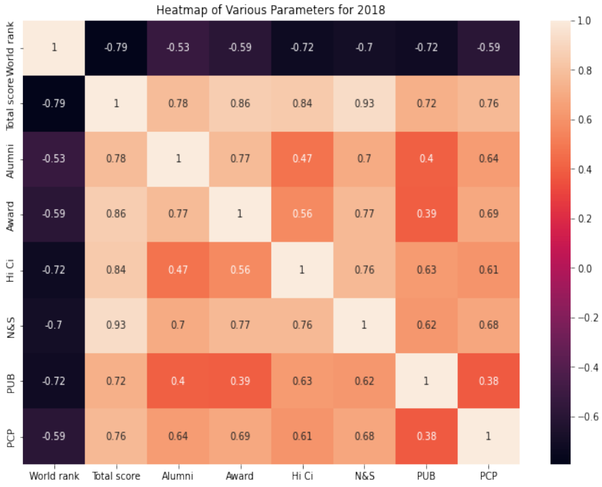

Fig. 11. Heatmap of correlation matrix of various parameters for ARWU

For the ARWU Rankings

In Fig 14 we plot the time average overall years of the average value of all the parameters for batches of 10 ranks each. From the plots we can see that the relative importance of of different indicators for the high ranking institutes is Award $>$ Alumni $\approx$ PUB $>$ HiFi $\approx$ N&S $>$ PCP. For the intermediate and the low ranking institutes order is PUB $>$ PCP $>$ N&S $\approx$ HiCi $>$ Alumni $\approx$ Award. The PUB scores have high relative importance for all ranks. Apart from the PUB scores the relative importance for the high ranking and the intermediate/ low ranking institutes is exactly opposite.

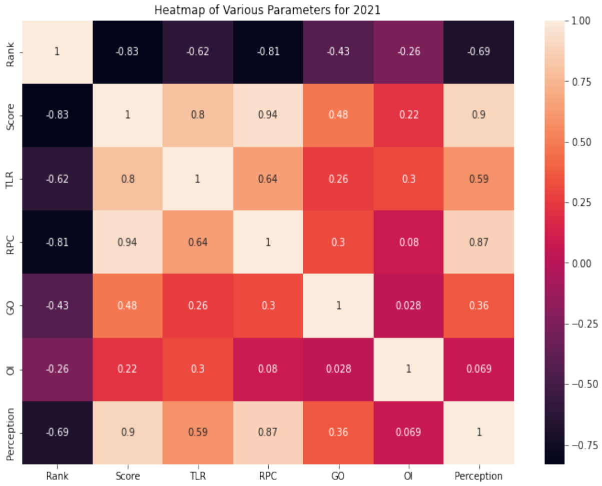

Fig. 12. Heatmap of correlation matrix of various parameters for NIRF

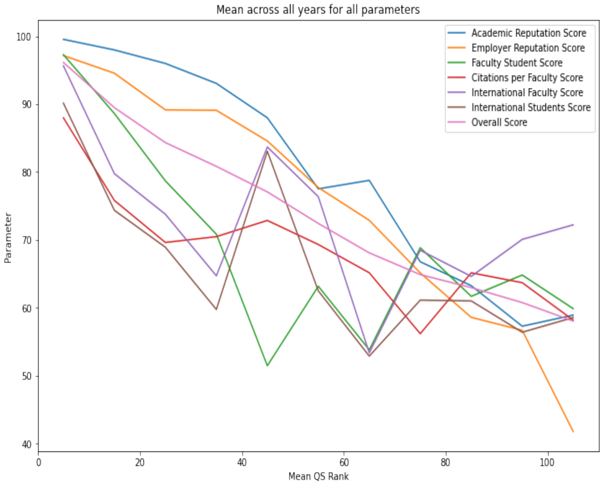

Fig. 13. Variation of the parameters vs rank across all years for QS

The Alumni and Award scores for low ranking institutes are nearly zero. This means that the majority of Nobel Prize and Fields Medal recipients are concentrated in the fewer higher ranking institutes.

From the heatmap in Fig 11, we notice here that the World Rank (AWRU Rank) has very high negative correlation with all the parameters other parameters if the parameters increase, then the AWRU rank decreases. Also since most of the parameters have a high positive correlation some of them could be redundant.

The box and whiskers plot for different indicators and the total score are shown in Fig 8. The Award indicator has nearly zero median value and the lower quartile values of Alumni, HiCi and N&S are also nearly zero. This means that the Alumni scores for 50% of the institutes are nearly 0 and the Alumni, HiCi and N&S for 25% of the institutes are nearly 0. The PUB scores are the closest to having a uniform distribution. Nearly all the indicators are rightward skewed.

For the NIRF Rankings

From the box and whiskers plots for the parameters in Fig 9 it is noticed that some parameters like Score, Perception and RPC have a lot of outliers. This could be due to these being relatively very high for few premier institutions like the IITs and IISc compared to any other educational institution. Also it was observed mean value of Scores, Perception is increasing over the years hemce the overall quality of education in India is improving over the years. Also the mean value of the parameter RPC is increasing which indicates that educational institutions are participating in more research. The number of outliers for the parameter Score have increased over the years, this corresponds to some colleges performing very well relative to the others for Score and RPC. This could be because of the Covid Pandemic.

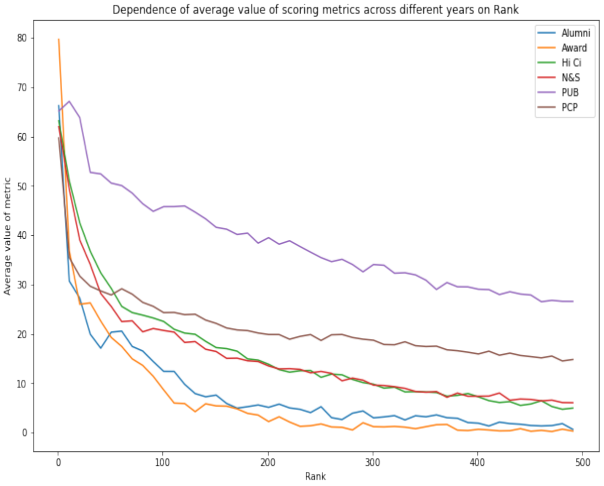

Fig. 14. Variation of the parameters vs rank across all years for ARWU

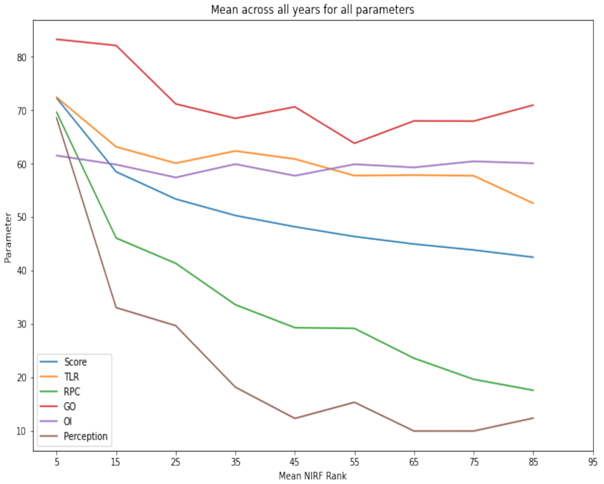

Fig. 15. Variation of the parameters vs rank across all years for NIRF

In Fig 15 we plot the time average over all years of the average value of all the parameters for batches of 10 ranks each. It is noticed here having a better NIRF rank corresponds to a higher values of Perception and RPC. The perception of the college with respect to the public, other colleges and employers is more provided the college has a better NIRF rank. The GO scores, As students prefer to join higher ranked colleges and employers prefer to hire from higher ranked colleges. Indicate institutions having a better NIRF ranking involve in more Research and Professional Practice. This could be because of a more talented student group and better employers. However there is no such strict increasing/decreasing trend for the parameters OI, GO and TLR. There is a slight decrease/ nearly constant behaviour in the other parameters as the rank increases.

From the heatmap in Fig 12 we can study the correlation of various parameters with each other Rank is highly negatively correlated with Perception, Score and RPC and also slightly negatively correlated with GO and TLR. This means the parameters decrease rapidly with increase in Perception, Score and RPC relative to GO and TLR.

Performance of IIT Bombay in World Rankings:

From the EDA it was concluded that IIT Bombay performs well in the academic reputation and employer reputation criteria, it does poorly in criteria like international student and faculty and this cause the overall ranking to drop, If ranking was done only on basis of academic parameters, IIT Bombay will perform much better than many top US universities too, but it needs to improve on the international student and faculty ratios in order to perform better overall in the world universities ranking. It can be further observed that almost all the indian universities in these rankings follow similar trends.

Predictive Results:

Predictive models were made to predict scores of the universities in the different ranking systems. Models were made for NIRF and Shanghai but not for QS, since for QS, only 3 years data was available and that wasn't enough to make a proper predictive model. Similar procedure was done for both Shanghai and the NIRF data, where an array of models was made for each university since each university had different properties. Six ML frameworks were used for performing regression with one feature as input which is the year and the score of the university as output. The six frameworks used were Linear Regression, Lasso Regression, Ridge Regression, SVM (Support Vector Machine), Decision Tree Regression and Random Forest Regression. The accuracy of the results were measured by the R^2 score on the validation/test data.

Before starting with the model, the data was filtered once and only the universities which were present were retained and the others were removed since it is pointless to have universities present for very few years. Since there was only one feature which is the year, which is categorical data, there is no need for normalization here. There was no hyper parameter tuning done for any of the mentioned ML frameworks due to nature of simplicity of the model. All the default parameters were used.

For the NIRF data, 4 year data (2017-2020) was used as train data and data for the year 2021 was used as test data. The results on the test data is as follows:

ML Framework	R^2 Score on test data
Linear Regression	0.878
Ridge Regression	0.927
Lasso Regression	0.960
SVM	0.969
Decision Tree	0.986
Random Forest	0.987

Clearly, Random forest gives the best R^2 Score among all the frameworks. This was further used to predict the rankings for the year 2022 which is next year where we got a quite similar pattern as the previous years. Here is the prediction of the top 6 universities as per the NIRF ranking system for the year 2022:

Rank	Institute Name	Score
1	Indian Institute of Technology Madras	83.88
2	Indian Institute of Science	82.67
3	Indian Institute of Technology Bombay	79.20
4	Indian Institute of Technology Delhi	78.69
5	Indian Institute of Technology Kharagpur	74.31
6	Indian Institute of Technology Kanpur	69.07

For the Shanghai data, the data from years 2005-2016 was used as train data, 2017 data was used as validation data and 2018 was used as test data. The results on the validation data is as follows:

ML Framework	R^2 Score on validation data
Linear Regression	0.9931
Ridge Regression	0.9931
Lasso Regression	0.9928
SVM	0.9887
Decision Tree	0.9959
Random Forest	0.9939

We can observe that here the R^2 scores were pretty high for almost all the frameworks which were tested. There are 2 reasons for this, firstly, a lot more train data was used here compared to that in NIRF, 12 years of training was used for only 1 year of validation and secondly, The shanghai data doesn't fluctuate very much and gives very similar scores every year i.e the correlation between year and the scores are fairly low and this gives very high efficiencies while making regression models. From the frameworks tested, the Decision Tree Regression gives the most accurate result on the train data and so it was used on the test data for the year 2018 and it gave an R^2 score of 0.9961510762613965, which is even higher than the validation accuracy! Following is comparison of the actual scores and the predicted scores for the top 5 universities for the year 2018:

Rank	Institute Name	Predicted	Actual
1	Harvard University	97.92	97.95
2	Stanford University	73.15	74.91
3	University of Cambridge	68.16	69.47
4	MIT	67.79	68.93
5	Princeton University	60.65	59.87

V. Learning, Conclusions, and Future Work

For the international rankings, we observed that Indian universities often don't perform well due to criterion like International Students and Faculty. This isn't a very fair criterion because US and UK have a much higher proportion of international students in comparison to India due to being more advanced and developed, if these criteria weren't included in the QS ranking, Indian universities will perform a lot better. This also leads to a lot of bias in the QS and ARWU rankings towards UK and US universities. It was also observed that there is a shifting of educational quality from the west to the east over the years, indicating that the education is become in a sense ”global” and not local to the western countries alone. A majority of Nobel Prize and Fields Medal recipients are concentrated in the fewer higher ranking institutes.

For the NIRF rankings, majority of the higher ranked universities are located in metropolitan cities and state capitals rather than in relatively smaller towns and villages indicating that higher quality of education is localised and to only large cities, this could be because of better infrastructure and opportunities in cities and social life that students and faculty enjoy in cities. Additionally, there was a lot of bias observed towards Tamil Nadu colleges. We can see that many Indian universities that performed well in the international ranking systems don't perform well in the NIRF ranking whereas many Tamil Nadu colleges which perform very well in the NIRF rankings don't perform well in the International rankings and a few don't even appear in the International rankings, indicating a clear bias for Tamil Nadu colleges in the NIRF ranking.

For the predictive analysis, we can conclude that most regression frameworks worked well here giving decent to high accuracy, in particular Random Forest and Decision Tree Regressor gave very high accuracies and are very usable models.

For future work, we can make more useful predictive models which can predict each individual criterion of the university that the ranking system desires which can help in further comparison between the predicted university rankings. We can maybe try a few more ML frameworks like using neural networks and it's various types like CNNs and RNNs to make more powerful models.

Acknowledgment

We would like to thank Prof. Amit Sethi, Prof. Manjesh Hanawal, Prof. Sunita Sarawagi and Prof. S. Sudarshan, our instructors for the course DS203 for giving us the opportunity to work on this project. We would also like to thank all the TA's involved with this course, who helped us a lot in doing the various assignments, and completing this course.

References

[1] National Institute Ranking Framework (NIRF) website
[2] Academic Ranking of World Universities website
[3] Quacquarelli Symonds (QS) World Rankings website
[4] Link to ARWU dataset
[5] Link to QS dataset
[6] Link to NIRF dataset