Analysis of features influencing World University Rankings and Prediction

Shriya Hukkeri[1], Akhil Maddipatla[2]
Northeastern University[1,2]
hukkeri.s@northeastern.edu[1], maddipatla.a@northeastern.edu[2]

Abstract

In this project, we take into consideration the three major datasets which are nothing but the world ranking datasets and test the performance of various regression algorithms by using them on the cleaned and standardized original datasets. We test the performance of these regression algorithms by taking their predicted value and comparing them against the baselines which are our original ranking datasets. The datasets are CWUR dataset, Times Ranking dataset and Shanghai Ranking dataset. We find the difference between the original rankings and the predicted rankings along with the accuracy of the algorithms. The decision tree algorithm along with Gradient Boost Regressor and Bagging Regressor gave the most accurate prediction with accuracies of 0.9949, 0.9973 and 0.9968. Also, we analyse how certain features like quality of education, alumni, awards and other factors influence the final raking outcome using various graphs and visualizations.

Introduction

Assigning a rank to a university is a very tedious task. Not only do students have to struggle with this but the universities heads, the faculty, industries, and the government too!

University rankings play an important part in the lives of many. It helps student decide the best university they could choose for a bright future, helps faculty decide the best place they could work at, helps the university authorities to know their position and develop skills accordingly, so on and so forth.

There are several global ranking systems that aid in ranking universities at national and international levels.

The most influential systems that we chose are, Times Higher Education (THE) that provides a list of best ranked universities with emphasis on the research mission, Centre of World University Ranking (CWUR) that uses 7 indicators grouped into 4 areas namely “Quality of Education”, “Alumni Employment”, “Quality of Faculty” and “Research Performance” and lastly the Shanghai ranking system (Previously known as Academic Ranking of World Universities) that ranks universities based on 4 major criteria namely “Quality of Education”, “Quality of Faculty”, “Research Output” and “Per Capita Performance” . All these three ranking systems have different methodologies for ranking a particular university.

A machine learning algorithm that could predict ranks accurately would be of great help to help make this process easier. In our project the main strategy of evaluating the performance of the algorithms is to see how accurately the algorithm could predict the ranks by taking into consideration the difference in the “predicted ranks” and the “original world rankings”. The smaller the difference the better. We also generate the accuracy scores (R2 scores) for every algorithm for their performance evaluation.

We would also analyse the importance that each feature has and how it contributes towards the rank of a university. We try to derive the most dominant features

from the rest by visualizing the datasets with the help of graphs like Bar charts, Heat maps, Radar Charts and Word Clouds.

Background

For loading data, applying pre-processing techniques we used in-built functions from Python Data Analysis (pandas) and scikit-learn python libraries (sklearn.preprocessing). For the application of mathematical operations, we used NumPy. For visualizing the data by plotting graphs we used Matplotlib. And for the support of regular expressions, we used the module "re". Since Rank Prediction is primarily a regression problem, we have chosen Regression algorithms and Ensemble methods. The regression algorithms used are Lasso Regression, Ridge Regression and Decision Trees. The ensemble methods used are Bagging Regressor and Gradient Boosting Regressor (GBR). All these algorithms were run on the built-in scikit-learn implementations.

A particularly interesting challenge was put in front of us when we discovered that the world rank feature in Times Higher Education and the Shanghai Datasets, that was of the primary importance for a lot of the analysis that we were supposed to do, was not presented in a suitable format. The world rank feature which is supposed to be an integer value assigned to a university was in fact represented as a rank range for most universities. For example: "University of Innsbruck" is ranked as 201-225. Now this would cause a problem since the data format would no longer be in integer format as required. Also, a rank range would not help us analyse features appropriately. Therefore, we found a way, only fair to assign ranks to the universities that had ranges as their ranks. We introduced a method called "rank_midpoint" which would take the rank range as the input and then get its midpoint and assign that as a rank to the university. In this was we were able to get the world rank column into a format that would give us accurate results. For the Radar chart visualization for TIMES and Shanghai datasets we have a function called "radar_chart", that is a common function that is called by both the datasets.

Related Work

There are projects which are primarily concerned with world university rankings but most of them are based on visualizations and investigate about how the ranking criteria such as alumni, quality of education, job employment and others weigh up for the final ranking outcome in terms of total score. There is a work from Anika Tabassum, Mahamudul Hasan, Shibbir Ahmed, Rahnuma Tasmin and Tasmin Musharatt from Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology which deals with analysis of the world university rankings by considering the university performance indicators. They have built an algorithm which calculates the performance indicators and outlier detection as well as a rank score calculator. They have used Java programming language and used only times university ranking data set. We are trying something different like taking the major algorithms used for regression and ranking and comparing them against each other by taking the difference between the original ranking and the predicted values respectively. In this way, the smaller the difference the greater is the performance of that algorithm. We have primarily worked with pandas and other python related libraries. We have also used three different ranking systems to test the accuracy of prediction of these algorithms. Also, we have analysed certain universities through spider graphs and plotted box plots for the returned values.

Project Description

Datasets

We make use of three datasets in our project as mentioned above. The first one is the CWUR dataset which is a Saudi ranking system. This dataset publishes rankings by assessing the quality of education, quality of faculty, alumni employment and research performance. It does not base its assessment on surveys and the data submitted by the university. It uses 7 indicators grouped into 4 major areas. They are: "Quality of Education" that is measured by the number of alumni who won major academic distinctions, "Alumni Employment" that is measured by the number of alumni who have worked on the top executive positions at the largest most reputed companies of the world, "Quality of Faculty" that is measured by the number of faculty at the university who have won major academic distinctions and lastly "Research Performance" that is divided into sub categories called "Research Output" that is measured by the number of

research papers published, “Publications” that is measured by the number of research papers that are present in some of the top journals, “Influence” that is measured by the number of research papers that are present in some of the highly influential journals and “Citations” that is measured by the number of highly cited research papers.

world_rank	institution	country	national_rank	quality_of_education	alumni_employment	quality_of_faculty	publications	influence	citations	broad_impact	patents	score	year
1	Harvard University	USA	1	7	9	1	1	1	1		5	100	2012
2	Massachusetts Institute of Technology	USA	2	9	17	3	12	4	4		1	91.67	2012
3	Stanford University	USA	3	17	11	5	4	2	2		15	89.5	2012
4	University of Cambridge	United Kingdom	1	10	24	4	16	16	11		50	86.17	2012
5	California Institute of Technology	USA	4	2	29	7	37	22	22		18	85.21	2012

CWUR dataset before data cleaning and application of data pre-processing techniques

The second dataset is the Times Higher Education (THE). This dataset gives a list of the world’s best universities giving major emphasis on the research work. There are 13 major categories that it takes into consideration like “Teaching” which describes the learning environment, “Research” that takes into consideration the volume, income and the reputation, “Citations” that measures the research influence, “Industry Income” that measures the knowledge transfer and “International Outlook” that comprises of staff, students and research.

world_rank	university_name	country	teaching	international	research	citations	income	total_score	num_students	student_staff_ratio	international_students	female_male_ratio	year
1	Harvard University	United States of America	99.7	72.4	98.7	98.8	34.5	96.1	20,152	8.9	25%		2011
2	California Institute of Technology	United States of America	97.7	54.6	98	99.9	83.7	96	2,243	6.9	27%	33 : 67	2011
3	Massachusetts Institute of Technology	United States of America	97.8	82.3	91.4	99.9	87.5	95.6	11,074	9	33%	37 : 63	2011
4	Stanford University	United States of America	98.3	29.5	98.1	99.2	64.3	94.3	15,596	7.8	22%	42:58:00	2011
201-225	Autonomous University of Barcelona	Spain	33.7	45.9	27.9	57.9	37	-	30,538	12.3	10%	59:41:00	2012

TIMES dataset before data cleaning and application of data pre-processing techniques

The third dataset is the Shanghai Rankings. This dataset takes into consideration all the universities that have any Nobel Prize winners, Highly Cited Researchers, Field Medallists, or papers that are published in Nature or Science. This dataset also considers 4 major areas as follows: “Quality of Education” that describes the alumni who won Nobel Prizes and Field Medals, “Quality of Faculty” that is divided into 2 sub categories called “award” that measures the staff winning Nobel Prizes and Field Medals and “HiCi” that describes the highly Cited Researchers, “Research Output” that is again divided into 2 sub categories called “NS” that describes papers published in Nature and Science and “PUB” that describes papers mentioned in Science Citation and lastly “Per Capita Performance”.

world_rank	university_name	national_rank	total_score	alumni	award	hici	ns	pub	pcp	year
1	Harvard University	1	100	100	100	100	100	100	72.4	2005
2	University of Cambridge	1	73.6	99.8	93.4	53.3	56.6	70.9	66.9	2005
3	Stanford University	2	73.4	41.1	72.2	88.5	70.9	72.3	65	2005
4	University of California, Berkeley	3	72.8	71.8	76	69.4	73.9	72.2	52.7	2005
101-152	Aarhus University	2		15.4	19.3	7.9	22.3	41.6	22.4	2005

SHANGHAI dataset before data cleaning and application of data pre-processing techniques

Data Transformations and Feature Extraction

The data at hand was not in the perfect format to be fed to the machine learning algorithms and hence they had to be converted into a suitable format. Many of the columns had NaN values that had to be replaced. In most case we replaced them with zeros. Certain columns also had missing values, to impute these missing values we replaced them with the mean value of the original column values. We used the fillna() to do the same. Also, there was a requirement that certain columns present their values in the integer format, like “number of students”, but the value was expressed as follows 20,153 for example. This makes it a string data type. So, we had to convert all such values to integer type. We also executed a method to convert the rank ranges like 200-300 for example as a single rank for that university. That method received a rank range as an input and found the midpoint of the range and assigned that as the rank for that university. We applied this method to the “world_rank” and “national_rank” columns. Certain columns that were only an additional information about the university and did not play an important part in the ranking of the university were dropped, so that a cleaner dataset could be fed to the machine learning algorithms. We used the StandardScalar method to standardize our dataset. And we use the LabelEncoder to handle categorical values such as “country” names or “university names”.

Empirical Results

Rank Prediction using the five different algorithms (CWUR)

First, we have the bagging regressor which is also a meta-estimator which fits the regressors-base. Before making an ensemble, it introduces randomization. Along with that we take the difference and the predicted values for the remaining algorithms similarly along with their accuracy. We display only 5 sample ranks here from each testing because of the space constraints. These algorithms run successfully when the datasets are cleaned and standardized. We set the random

The bagging regressor score is: 0.9977335431423

Serial No	world_rank	Predicted_BR_cwur	Difference_BR_cwur
1124	925	925.3	0.3
235	36	37.0	1.0
585	386	390.3	4.3
1469	270	289.7	19.7
2066	867	862.3	4.7

The Gradient Boost Regressor Score: 0.997369096

Serial No	world_rank	Predicted_GBR_cwur	Difference_GBR_cwur
1124	925	922.34	2.65
235	36	63.38	27.38
585	386	390.125	4.125
1469	270	270.99	0.9989
2066	867	862.3	6.75

The Lasso Regression Score is: 0.9420800320

Serial No	world_rank	Predicted_LR_cwur	Difference_LR_cwur
1124	925	852.127	72.87
235	36	205.766	169.76
585	386	385.11	0.886
1469	270	361.83	91.83
2066	867	836.164	30.835

The Ridge Regression score is: 0.9424210988

Serial No	world_rank	Predicted_RR_cwur	Difference_RR_cwur
1124	925	856.78	68.21
235	36	193.35	157.35
585	386	383.17	2.82
1469	270	368.64	98.64
2066	867	837.03	29.969

The Decision Tree Regressor Score: 0.99494132

Serial No	world_rank	Predicted_DTR_cwur	Difference_DTR_cwur
1124	925	925.3	0.3
235	36	37.0	1.0
585	386	390.3	4.3
1469	270	289.7	19.7

2066	867	862.3	4.7

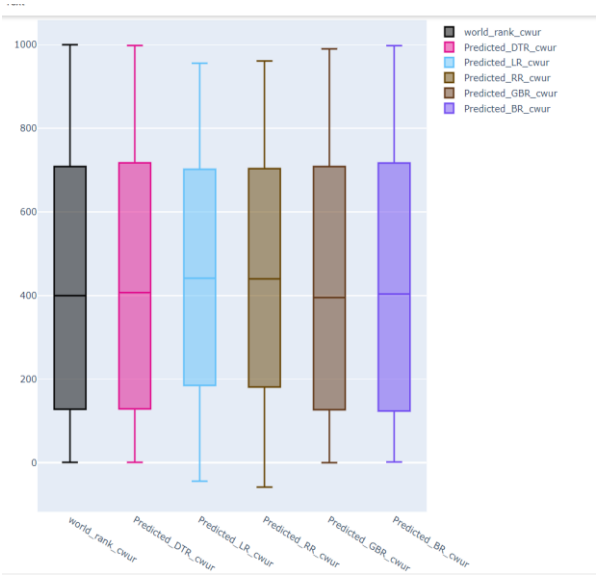
As we have taken the predicted values for CWUR ranking dataset against for each algorithm, now we compare it against our baseline world rank. Now we take and compare only the differences so that we can get the most accurate detail of which algorithms perform against each other.

As we can see below, the three primary algorithms which perform well are the Decision Tree regressor, The Gradient Boost Regressor and Bagging regressor because they hold the least differences with the original world rank.

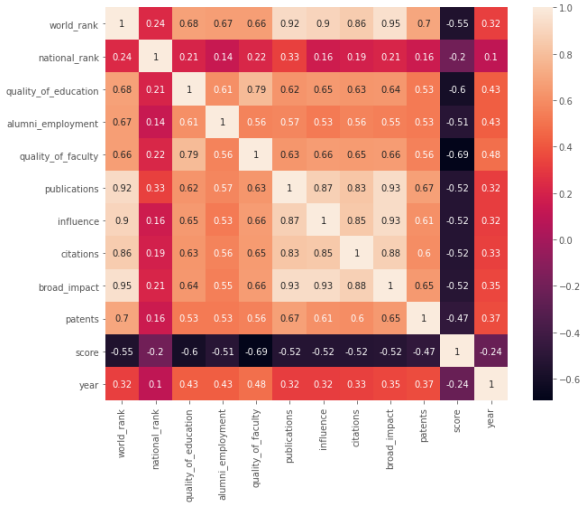
World_ rank	Difference_ DTR_cwur	Difference_ RR_cwur	Difference_ LR_cwur
925.0	4.0	68.21	72.87
36.0	0.0	157.35	169.76
386.0	2.0	2.82	0.88
270.0	1.0	98.64	91.83
867.0	2.0	29.96	30.83

World_ rank	Difference_ GBR_cwur	Difference_ BR_cwur
925.0	2.658257	0.4
36.0	27.3816	38.9
386.0	4.1258	1.7
270.0	0.998974	4.4
867.0	6.759993	5.6

Also, as a part of our results we have the box plot of the predicted algorithms against the world rank which is our baseline. The Box plot takes the 5 values which are min, quartile 1, median, quartile 3 and the max values along with the upper and lower fence values. The different algorithms are coloured differently so that they are distinguishable.



Box Plot for world rank against predicted values (CWUR)



Heat map for CWUR

The heat map for CWUR data set further extends our analysis by showing which factors are primarily correlated while taking correlation matrix and plotting them against each other. We can see for CWUR that the world rank is positively correlated with quality of education, publications, influence, citations and broad

impact. Hence, these features primarily determine the total score of the university.

Times data

We follow the same steps and procedure as done previously but now on our Times data set.

The bagging regressor score is: 0.9120442044

Serial No	world_rank	Predicted_BR_times	Difference_BR_times
1494	93.0	95.70	2.70
1492	91.0	103.40	12.40
1828	26.0	20.40	5.60
4	5.0	3.00	2.00
2188	375.5	313.00	62.50

The Gradient Boost Regressor Score: 0.921364490

Serial No	world_rank	Predicted_GBR_times	Difference_GBR_times
1494	93.0	107.459	14.45
1492	91.0	107.197	16.197
1828	26.0	21.685	4.314
4	5.0	10.59	5.59
2188	375.5	300.33	75.16

The Lasso Regression Score is: 0.79672

Serial No	world_rank	Predicted_LR_times	Difference_LR_times
1494	93.0	139.74	46.74
1492	91.0	143.64	52.64
1828	26.0	-47.86	73.86
4	5.0	-169.14	174.14
2188	375.5	409.61	34.116

The Ridge Regression score is: 0.79674530

Serial No	world_rank	Predicted_RR_times	Difference_RR_times
1494	93.0	139.35	46.35
1492	91.0	146.47	55.47
1828	26.0	-49.89	75.89
4	5.0	-172.82	177.82
2188	375.5	408.78	33.28

The Decision Tree Regressor Score: 0.85224038

Serial No	world_rank	Predicted_DTR_times	Difference_DTR_times
1494	93.0	80.0	13.0
1492	91.0	93.0	2.0
1828	26.0	18.0	8.0
4	5.0	3.0	2.0
2188	375.5	325.5	50.0

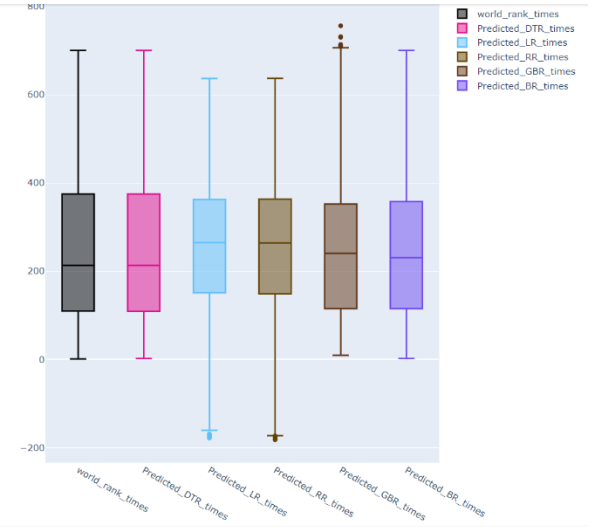
Now we take the differences of the predicted values against the times ranking data set.

World_Rank	Difference_DTR_times	Difference_RR_times	Difference_LR_times

93.0	6.0	46.35	46.74
91.0	2.0	55.47	52.64
26.0	7.0	75.89	73.86
5.0	1.0	177.82	174.14
375.5	50.0	33.28	34.116

World_ rank	Difference_ GBR_times	Difference_ BR_times
93.0	14.45	10.70
91.0	16.197	18.20
26.0	4.31	5.80
5.0	5.59	1.90
375.5	75.160	71.25

As we can see above even for times data set, the three primary algorithms that are performing well are Bagging regressor, Decision Tree and Gradient Boost Regressor.



Box Plot for world rank against predicted values (Times)

The next image is our heat map of times data set. This shows our correlation between the factors influencing times ranking. We can observe that total score is influenced very positively by teaching research and income. In this way these factors weigh more than the remaining ones to be deterministic about the final ranking outcome. Also, 'international' along with the remaining factors make very less impact on the final ranking outcome.

Shanghai Data set

Finally, we follow the same procedure for the shanghai data set as well by predicting the ranking outcome through using our five algorithms as mentioned previously.

The bagging regressor score is: 0.97400418979

Serial No	world_rank	Predicted_ BR_shan	Difference_BR_ shan
4575	175.5	175.65	0.15
167	177.5	176.50	1.00
1889	352.0	351.70	0.30
958	450.5	453.20	2.70
3010	450.5	451.70	1.20

The Gradient Boost Regressor Score: 0.96731823

Serial No	world_rank	Predicted_GBR_shan	Difference_GBR_shan
4575	175.5	162.75	12.74
167	177.5	206.183	28.68
1889	352.0	364.311	12.311
958	450.0	460.97	10.479
3010	450.5	461.52	11.024

The Decision Tree Regressor Score: 0.9545316796

Serial No	world_rank	Predicted_DTR_shan	Difference_DTR_shan
4575	175.5	125.5	50.0
167	177.5	176.5	1.0
1889	352.0	352.0	0.0
958	450.5	450.5	0.0
3010	450.5	451.5	1.0

The Lasso Regression Score is: 0.8387122250

Serial No	world_rank	Predicted_LR_shan	Difference_LR_shan
4575	175.5	269.26	93.76
167	177.5	280.00	102.50
1889	352.0	341.59	10.403
958	450.5	388.27	62.22
3010	450.5	373.767	76.73

The Ridge Regression score is: 0.839443531

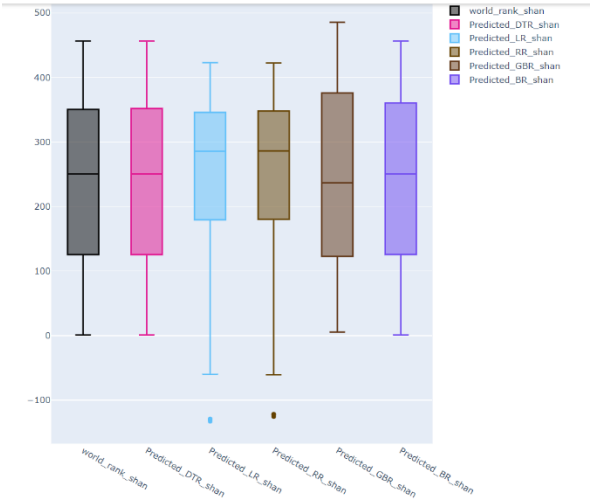
Serial No	world_rank	Predicted_RR_shan	Difference_RR_shan
4575	175.5	269.46	93.96
167	177.5	279.37	101.87
1889	352.0	344.08	7.917
958	450.5	392.15	58.34
3010	450.5	375.53	74.96

Now we take the differences of the predicted values against the shanghai ranking data set.

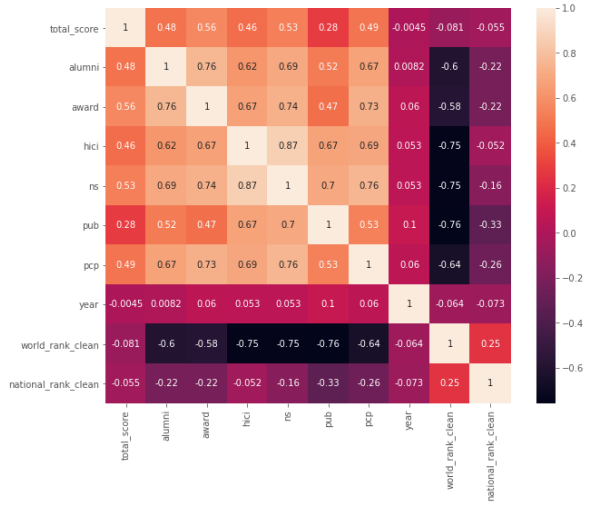
World_Rank_clean	Difference_DTR_shan	Difference_RR_shan	Difference_LR_shan
175.5	50.0	93.96	93.76
177.5	1.0	101.87	102.50
352.0	0.0	7.91	10.40
450.5	0.0	58.34	62.22
450.5	1.0	74.96	76.73

World_Rank	Difference_GBR_shan	Difference_BR_shan
175.5	12.74	2.85
177.5	28.68	1.40
352.0	12.311	0.30
450.5	10.47	4.80
450.5	11.0248	0.70

The three primary algorithms with the least difference to the original world ranking are the bagging regressor, Gradient Boost Regressor and the Decision Tree regressor.



Box Plot for world rank against predicted values (Shanghai).



Heat map of (Shanghai)

We can observe here that world rank is highly correlated with alumni, award, hici, ns, pub and pcp. These factors are nothing but Quality of faculty, Research Output and per capita performance.

Conclusion

In our project which deals with the World University Ranking, we develop a unique way of predicting the rankings again using various algorithms and thereby taking the difference between the original rankings to the predicted values. This gives us a comprehensive overview of the performance of these algorithms when tested against various datasets. The Gradient Boost Regression, The Decision Tree regression and Bagging regression have the most accurate results of prediction. We prefer Decision tree Algorithm since it not only closely matches with the original rankings but also considers all the possible outcomes of a decision and traces a path to conclusion. A comprehensive analysis of each branch is specified along with the decision nodes. Decision Tree regressor deals with less cleaning and missing values especially for world University rankings where a lot of data is either missing or is in an improper format. Also, when there is no factual information available, probability conditions are utilized to keep every choice under perspective. Furthermore, by using the correlated heat maps, we understand how many individual features of each individual ranking systems weigh to determine the final score which in turn is demonstrated on the world ranking system. In the future is there can be a possibility of combining the world ranking data sets and estimating the bias of each ranking system, thereby noting the partiality of the ranking system. This will provide the transparent system to many students across the world. Finally, my team would advise the future students who would like to work with our project to explore more about the datasets and work with many of the latest algorithms and various other ranking systems.

References

https://www.timeshighereducation.com/world-university-rankings/ -United Kingdom Ranking

http://www.shanghairanking.com/ -Shanghai Ranking

http://cwur.org/ - Saudi Arabia Ranking

https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2021-methodology

https://people.cs.vt.edu/anikat1/publications/university-ranking-prediction.pdf

https://www.shanghairanking.com/rankings/arwu/2020

https://en.wikipedia.org/wiki/College_and_university_rankings

Marklein, Mary Beth. "Rankings create 'perverse incentives' – Hazelkorn". *University World News. University World News. Retrieved 14 September 2016*

Philip G. Altbach (11 November 2010). *"The State of the Rankings"*. *Inside Higher Ed*. *Retrieved 11 June 2017*.

Source Code for our Project:

https://colab.research.google.com/drive/1_f4SsZUah5IRguTRMHZjNCkmjVPLH7_6?usp=sharing