UNIVERSITY RANK PREDICTOR

Vivekanand Sridharan, Karthik Sajeev, Sandesh George Oommen, Anesh Krishna J N, Marat Rinatovich Amaltdinov

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

sridha16@purdue.edu; ksajeev@purdue.edu; sgeorgeo@purdue.edu; ajayasud@purdue.edu; mamaltdi@purdue.edu

ABSTRACT

As graduate students, the struggle on finalizing colleges for admission was even more taxing than the exams and projects we undertake. This was the initial spark which got us working on designing a DSS Shiny app which would predict University rankings based on weightage given to the most important determinants like Quality of Teaching, Quality of Research, extent of Industrial collaboration and various other factors. We also realized in the process that this app would be even more useful to Universities because the percentage of revenue from just applications is enormous. Moreover, the incoming quality of students would also be more refined as the ranking gets better. We, as a team, decided on using Regression techniques for predictive analytics as Regression is a simple but efficient tool for elaborate data analysis. On further fine tuning, we used Multiple Linear Regression in the final model for prediction. The Final shiny app used six predictors for estimating the overall score or rank of the University. There is also an option of predicting the rank for the next successive year by altering the rank details of the university. By this design, Universities can work on the more important rank predictors and boost their standing overall. Students could also be aware of the most important factors to look out for when shortlisting universities.

BUSINESS PROBLEM

Our rank predictor app would help a University Management to understand the critical factors that determine their overall score and hence their rankings. The goal is to improve their existing ranking by tweaking some/all of these most important factors. As there is a set of predictor variables, which in turn can be used for prediction, this problem can be translated easily in to an analytics problem as well. This would mean their University appears on the shortlist of more students than before. University management, the major stakeholders here, would be benefitted by increase in revenue generated from the number of applications they would be receiving.

ANALYTICS PROBLEM

As formulated, we apply and compare different Regression techniques so as to use the best method to predict new rankings of Universities from the given predictor variables. The metrics and KPI's are measured by the percentage increase in Total Score and the percentage increase in Number of applications which are in turn driven by factors like Research Rating, Teaching Rating, Industry income rating etc. Although, advanced machine learning algorithms can be used for better prediction, we stick to regression techniques for prediction. University Management would be the outright benefactors if they invest in our project.

Although advanced machine learning algorithms are being used for prediction, we ignore such constraints and complex approaches and use simple regression techniques. There are four key assumptions related to any regression model which is being used to estimate the overall score:

1. Linear relationship
2. Multivariate Normality
3. No or little multicollinearity
4. Homoscedasticity

DATA

The data comes from the annual Times magazine rankings of universities across the world (https://www.statcrunch.com/5.0/shareddata.php?keywords=rankings). We used two different data sets with identical variables and numbers of observations. First set includes data for the 6 years period from 2011 to 2016. The second set includes data for the most recent year. Both sets have 200 observations and 13 variables. We cleaned the original set of data to eliminate the missing values. We predicted the Total Score using six factors: Teaching Rating, Research Rating, Citations Rating, Industry Income Rating, Placement Rating, and Percentage of International students. All of the predictor variables have a scale of 0 to 100. There are also other variables in the data sets but they were not used in constructing our predictive model, as they have no significance in predicting the Total Score.

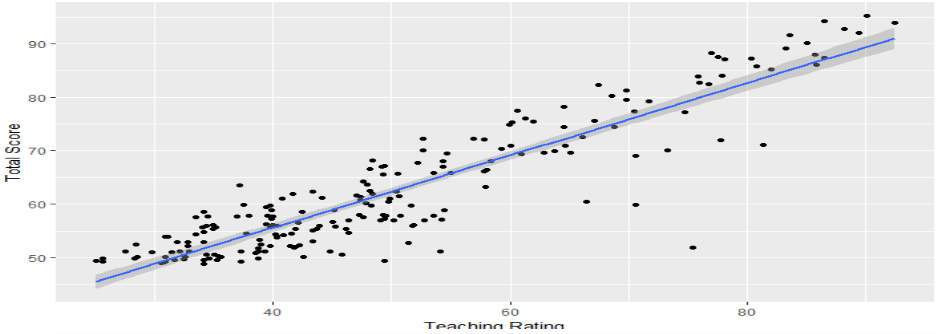
Description of all variables can be found in the following table:

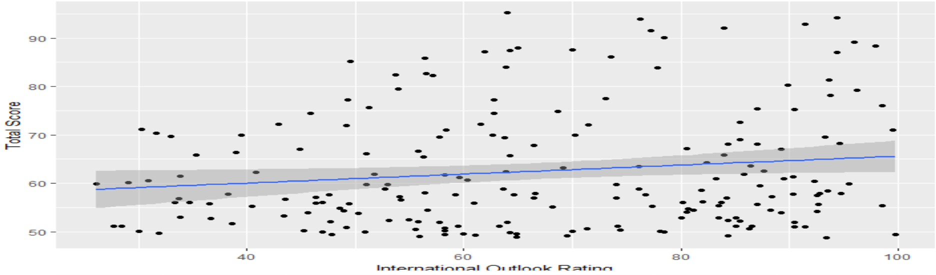
Column	**Description**
World_Rank	University rank for a given year
University_Name	The name of the university
Country	Location of university
Teaching_Rating	Rating from a 0-100 scale of the quality of teaching at the university. This rating is based on the institution’s reputation for teaching, it’s student/staff ratio, it’s PhD’s/ undergraduate degrees awarded, and it’s institutional income/ academic staff ratio.
Inter_Outlook_Rating	Rating from a 0-100 scale of the international makeup of a university. This rating is based the international staff percentage, and the percentage of research papers from the university that include at least one international author, number of students in student exchange programs
Research_Rating	Rating from a 0-100 scale of quality of research at the university. This rating is based on the university’s reputation, it’s research income/ academic staff ratio, and it’s production of scholarly papers.
Citations_Rating	Rating from a 0-100 scale of based on the normalized average of citations by other papers per paper from the university (how often the research from the university is cited by other papers).
Industry_Income_Rating	Rating from a 0-100 scale grading how much companies are willing to invest in the universities research. The rating is calculated based on the research income from businesses per academic staff member.
Total_Score	The final score used to determine the university ranking based on Teaching_Rating, International_Outlook_Rating, Research_Rating, Citations_Rating, and Industrial_Income_Rating.
Num_Students	Total number of students in a given year
%_Inter_Students	Percentage of student body who come from a foreign county
Placement Rating	Rating from a 0-100 scale of based on number of students getting placed in well established companies through career fair
Year	Academic year that the ranking was released. For example, 2016 denotes the 2015-2016 academic year

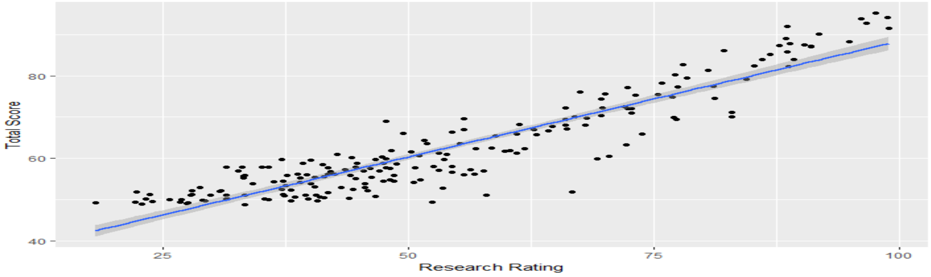
METHODOLGY

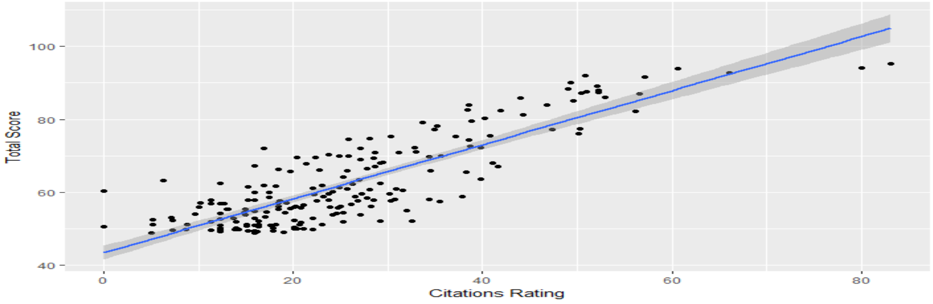
Scatterplots

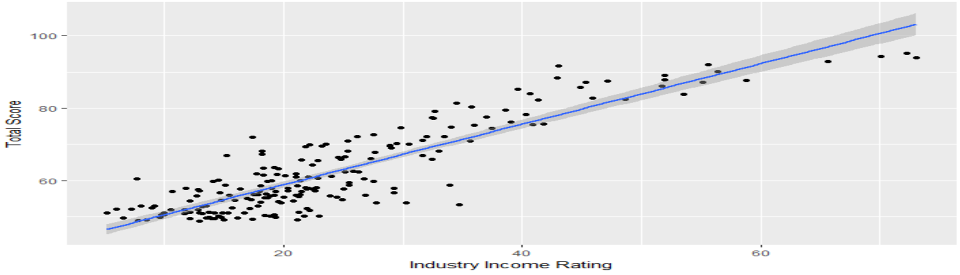
In order to assess the effect each of the eight factors has on the total score, the scatterplots are made and checked to see if linear model fits the data. It also serves the purpose of checking the linearity assumption, which is a must before any regression technique can be applied. The ggplot function is used to achieve this objective, keeping y as the Total score and assigning x as one of the eight factors each time. The plots obtained are shown below:

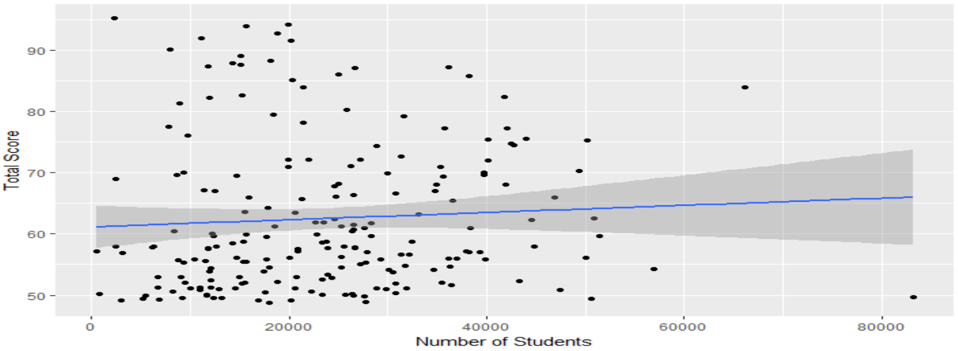


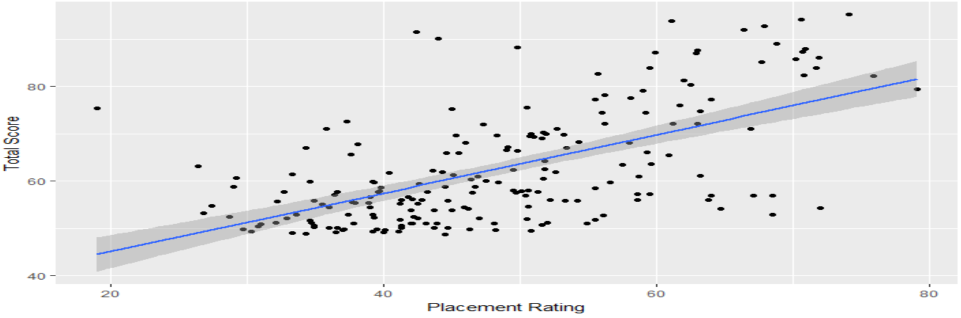


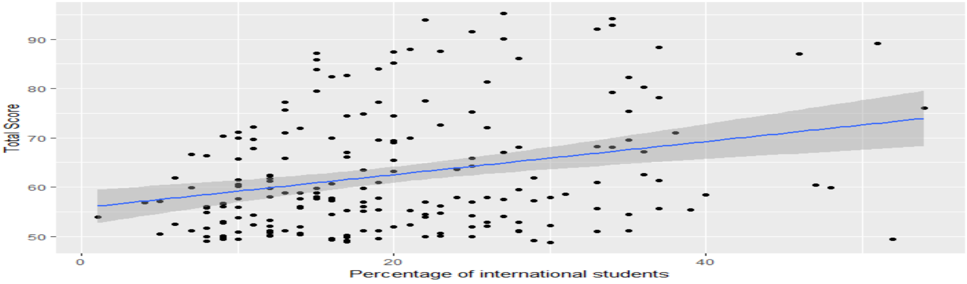












As seen from the plots, most of the factors have a good linear relationship with the Total score, which is evidenced by an increase in the value of Total score for an increase in the value of the factor. In a couple of plots, specifically, International outlook rating vs. Total score and Number of students vs. Total score, the relationship is not as strong: the slope is visibly smaller than that of the other plots and data points are not necessarily close to the line, but are spread out on either side. Despite these minor factors, since the assumption of linearity is not violated in any of these cases, all eight factors qualify to be considered for the next level of testing.

Simple Linear Regression

In order to compare and assess the strength of the influence each factor has on the Total score, regression analysis is carried out. In other words, the p- and f-values are computed and compared. The p-value for each factor tests the null hypothesis that the corresponding coefficient equals zero. A low value of p indicates that the null hypothesis can be rejected and the existence of a linear relationship can be confirmed. On the contrary, a relatively high value of p indicates the absence of linear relationship between the factor and the predicted variable. Through this method of analysis, it would thus be possible to extract those factors that do have a significant contribution towards the total score, and to eliminate those that do not.

In R, the lm function is used to perform simple linear regression and the summary function is used to obtain the F-statistic and the p-value. The following table shows the obtained results:

Testing Scenario	F-value	p-value
Total_Score ~ Teaching_Rating	946.1	2*e^-16
Total_Score ~ Research_Rating	1143	2*e^-16
Total_Score ~ Citations_Rating	509.8	2*e^-16
Total_Score ~ Industry_Income_Rating	748.5	2*e^-16
Total_Score ~ Placement_Rating	112.6	2*e^-16
Total_Score ~ X._Inter_Students	17.55	4.2*e^-5
Total_Score ~ Num_Students	4.627	0.03269
Total_Score ~ Inter_Outlook_Rating	0.8132	0.3683

As observed from the table, the factors ‘Number of students’ and ‘International outlook rating’ have much higher p-values when compared to those of the other factors. This indicates that these two factors do not influence the total score as much as the others do, and hence can be eliminated before performing subsequent steps.

Moving forward, the six influential factors that will be considered with regard to the different models are:

- Teacher Rating
- Research Rating
- Citations Rating
- Industry Income Ratings
- Percentage of International Students
- Placement Ratings

Comparison of models

Different models need to be compared and their performance with regard to the given dataset must be evaluated in order to choose the one that best describes the data. Here, the models being compared are: Multiple linear regression (MLR), Ridge regression, and Lasso regression.

Multiple linear regression - used to model the relationship between one response variable and multiple explanatory variables.

Ridge regression- used in datasets having multicollinearity

Lasso regression (Least absolute shrinkage and selection operator)- used to perform variable selection

By linking the dataset, we would be able to compare the performance of the models and choose the one that fits best.

The coefficient and intercept values are obtained as shown:

	MLR	Lasso	Ridge
Intercept	28.13138	32.36366380	28.47494220
Teaching_Rating	0.20446	0.18565066	0.20292440
Research_Rating	0.21088	0.20912987	0.21074994
Citations_Rating	0.11710	0.11107056	0.11660727
Industry Income Rating	0.21751	0.21525930	0.21735984
Placement Rating	0.04825	0.01187475	0.04528483
Percentage of International Students	0.10634	0.04468217	0.10132584

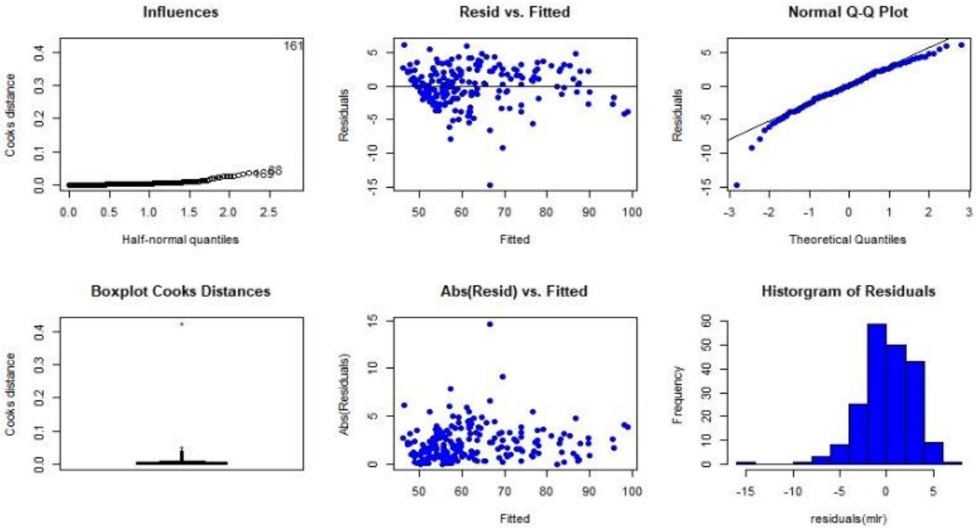
Next, the R^2 values of the three models are compared. A higher R^2 value indicates that there is a good match between the data and the fitted regression length.

	MLR	Lasso	Ridge
R square value	0.994	0.9442	0.943

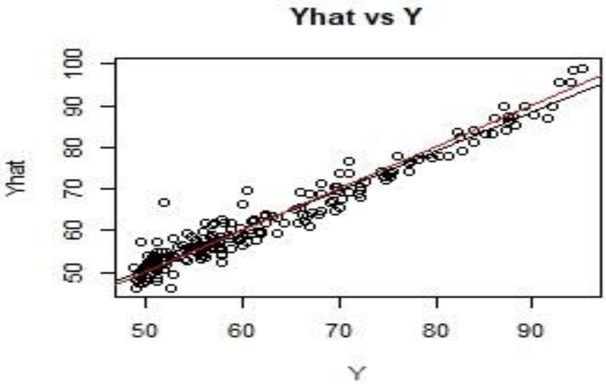
As seen from the table, R^2 values for MLR, Lasso and Ridge regression are very close to each other. As all three models are equally efficient in predicting the Total score, MLR is arbitrarily chosen as the model to carry out analysis.

MODEL BUILDING

- After we finalized on using Multiple Linear regression, we started off checking the basic assumptions for carrying out regression. The Normality and constant variance assumptions were reinforced through the qq plots and the residual plots respectively. The box plot and histogram (which was left skewed) showed possible outliers which might exist in the data.

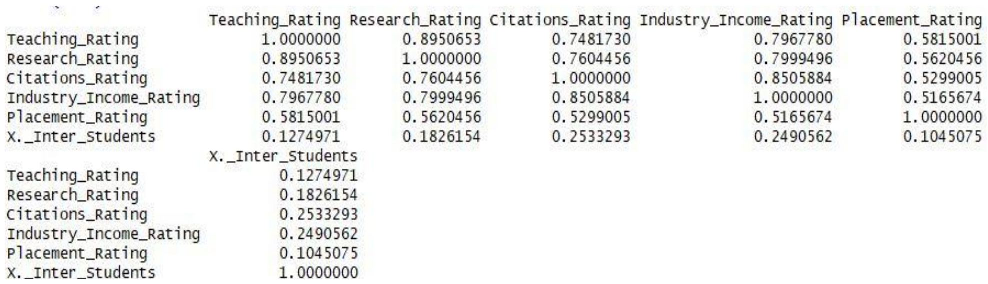


- We then ran the Y-hat vs Y plot which showed that the predicted value line was in line with the 45 degree actual value line. This was a starting point which ensured that we were on the right track using Multiple Linear Regression.



- There were 4 major characteristics which had to be validated for using the Regression model. They were:
  - Correlation and Multicollinearity between predictor variables
  - Influential observations
  - Outliers
  - Variable Selection

- We ran the cor command which is the Correlation command in R.



```
                       Teaching_Rating Research_Rating Citations_Rating Industry_Income_Rating Placement_Rating
Teaching_Rating              1.0000000       0.8950653        0.7481730              0.7967780        0.5815001
Research_Rating              0.8950653       1.0000000        0.7604456              0.7999496        0.5620456
Citations_Rating             0.7481730       0.7604456        1.0000000              0.8505884        0.5299005
Industry_Income_Rating       0.7967780       0.7999496        0.8505884              1.0000000        0.5165674
Placement_Rating             0.5815001       0.5620456        0.5299005              0.5165674        1.0000000
X._Inter_Students            0.1274971       0.1826154        0.2533293              0.2490562        0.1045075
                       X._Inter_Students
Teaching_Rating                0.1274971
Research_Rating                0.1826154
Citations_Rating               0.2533293
Industry_Income_Rating         0.2490562
Placement_Rating               0.1045075
X._Inter_Students              1.0000000
```

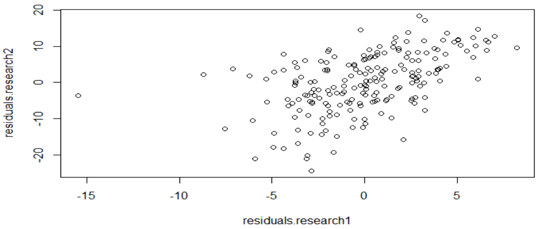
- The correlation between Teaching Rating and Research Rating was 0.89, which is not considered to be an alarmingly high value. We still decided to investigate this value because a significant correlation meant that both the ratings together were fruitful predictors in the model, but individually they might not be significant.

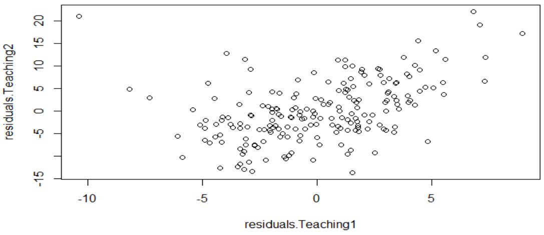
- We used the General Linear test approach (GLT) which is based on Hypothesis Testing. There were three cases in the GLT, which signaled if the 2 predictors were needed in the model. The Teaching rating coefficient is zero, or the Research Rating coefficient is zero or both are zero

P-values for 1st case	**< 2.2*e^(-16)**
P-value for 2nd case	<5.743*e^(*-15)
P-value for 3rd case	<2.219*e^(-10)

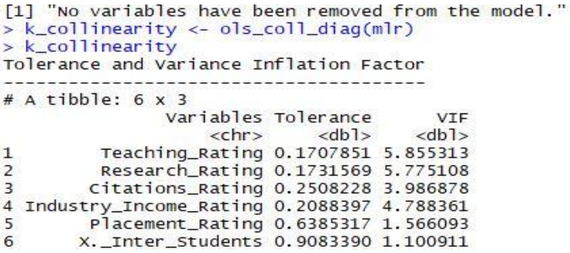
- Based on the result of the GLT we obtained, we obtained significant p values. Hence, we rejected the null hypothesis and chose the alternative which said that both the variables were needed together in the model.

- We confirmed the above assumption of non-correlation by running the partial residual plots for the 2 variables. Partial residual plots were plotted between the residuals obtained from predicting Y from the remaining predictors and Predicting X (either teaching or Research) from the remaining predictors. The plots showed a linear relationship which meant that both the predictors were needed in the model. i.e. The correlation value will not hinder the performance of the predictors



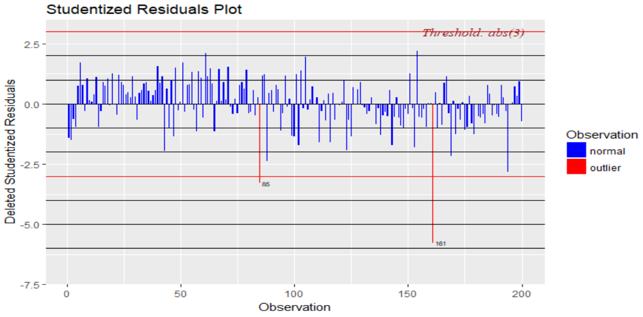


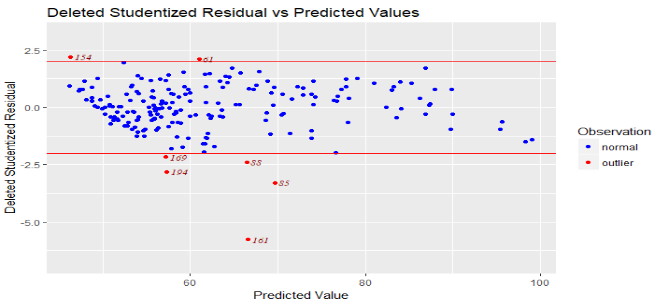
- For Multicollinearity, we decided to use Tolerance and VIF (Variance Inflation Factor) as a threshold. A tolerance value of <0.1 and a VIF value >10 were undesirable. The results obtained were within the specifications



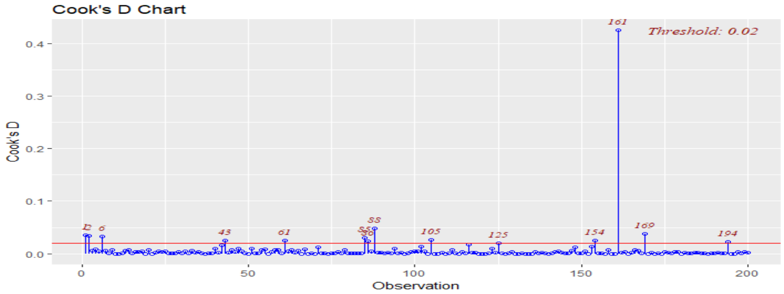
```
[1] "No variables have been removed from the model."
> k_collinearity <- ols_coll_diag(mlr)
> k_collinearity
Tolerance and Variance Inflation Factor
---------------------------------------
# A tibble: 6 x 3
              Variables Tolerance      VIF
                  <chr>     <dbl>    <dbl>
1       Teaching_Rating 0.1707851 5.855313
2       Research_Rating 0.1731569 5.775108
3      Citations_Rating 0.2508228 3.986878
4 Industry_Income_Rating 0.2088397 4.788361
5      Placement_Rating 0.6385317 1.566093
6     X._Inter_Students 0.9083390 1.100911
```

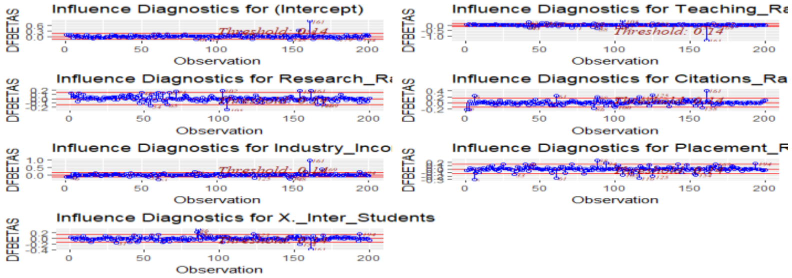
- For Outliers, we decided to use the Studentized Residual and the Studentized Deleted Residual as cut off values. The plots for the two also helped in visually observing the major outliers and removing them from the dataset. Studentized residuals are obtained by dividing the residuals by the standard error. Studentized deleted residuals are obtained by deleting a case and finding the predicted value, and then calculating the difference in predicted values with and without the observation. Cut-off values for the two are 3 and $t_{n-p-1}\left(1-\frac{alpha}{2}\right) = t_{193}(0.975)$. Based on the cut off values and the obtained plots, we noticed one major outlier (161) which had to be removed.

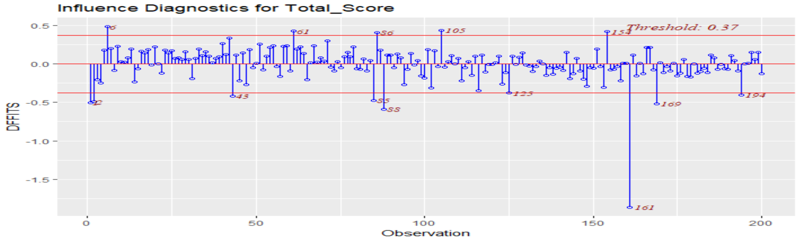




- For the influential observations, we used Cook's D (which measures the influence of case i on all the other observations), DFFITS (which measures the influence of a case i on its own fitted value) and DFBETAS (which measures the influence of case i on all the regression co-efficients). For Cook's D, the cut off value was $F_{p,n-p} = 0.2$, For DFFITS, the cut off value was 2*sqrt(p/n) = 0.37 and For DFBETAS the cut off was 2.

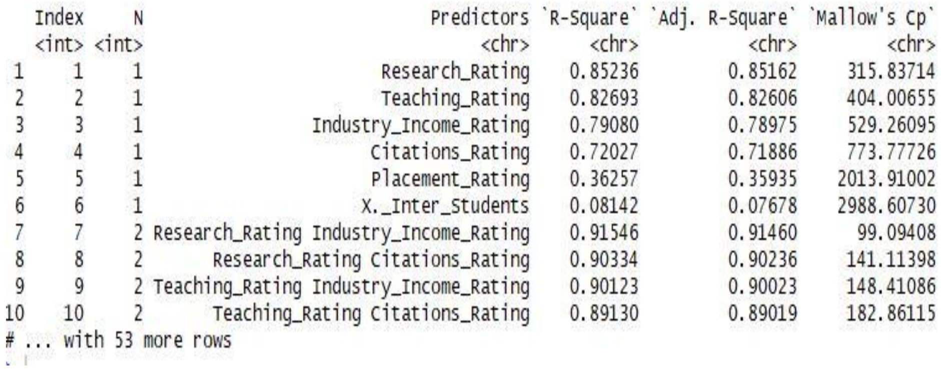






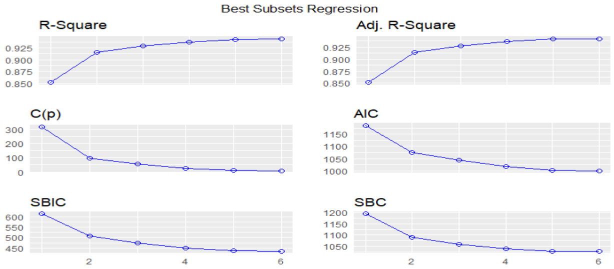
Based on the above graphs, we concluded that the observation at 161, which was also an outlier, turned out to be influential observation.

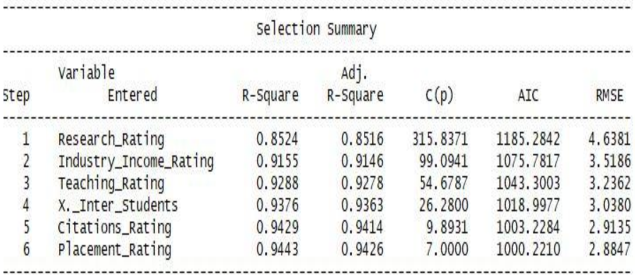
After performing diagnostics and extracting only the useful observations from the dataset, we had to use Variable Selection techniques to ensure that we were using the best set of predictors. We decided to use Stepwise Regression, which would return all possible 2^k combinations of models for k variables. For example, if the model has 3 predictor variables, the number of subsets returned would be 2^3 = 8. Hence, for Stepwise basic command, we had 2^6 models listed based on adjusted R-square and Cp.



```
   Index     N                          Predictors `R-Square` `Adj. R-Square` `Mallow's Cp`
   <int> <int>                               <chr>      <chr>           <chr>         <chr>
1      1     1                     Research_Rating    0.85236         0.85162     315.83714
2      2     1                     Teaching_Rating    0.82693         0.82606     404.00655
3      3     1              Industry_Income_Rating    0.79080         0.78975     529.26095
4      4     1                    Citations_Rating    0.72027         0.71886     773.77726
5      5     1                    Placement_Rating    0.36257         0.35935    2013.91002
6      6     1                    X._Inter_Students    0.08142         0.07678    2988.60730
7      7     2 Research_Rating Industry_Income_Rating    0.91546         0.91460      99.09408
8      8     2     Research_Rating Citations_Rating    0.90334         0.90236     141.11398
9      9     2 Teaching_Rating Industry_Income_Rating    0.90123         0.90023     148.41086
10    10     2     Teaching_Rating Citations_Rating    0.89130         0.89019     182.86115
# ... with 53 more rows
```

We then ran Stepwise (Forward) regression which took in variables one-by-one and rejected them if the p-values weren't significant, Stepwise (Backward) regression which took in the full model initially and kept throwing out variables if the p-value wasn't significant and Stepwise (Both ways) regression which performed all possible permutations. Based on the results obtained, the full model with all the predictors had the highest possible R-square value and most significant p-value.





```
--------------------------------------------------------------------------------
                              Selection Summary
--------------------------------------------------------------------------------
        Variable                     Adj.
Step         Entered       R-Square    R-Square     C(p)        AIC        RMSE
--------------------------------------------------------------------------------
   1    Research_Rating       0.8524      0.8516    315.8371    1185.2842    4.6381
   2    Industry_Income_Rating 0.9155     0.9146     99.0941    1075.7817    3.5186
   3    Teaching_Rating       0.9288      0.9278     54.6787    1043.3003    3.2362
   4    X._Inter_Students     0.9376      0.9363     26.2800    1018.9977    3.0380
   5    Citations_Rating      0.9429      0.9414      9.8931    1003.2284    2.9135
   6    Placement_Rating      0.9443      0.9426      7.0000    1000.2210    2.8847
--------------------------------------------------------------------------------
```

- Based on all the above considerations, we finalized on using all the six predictors and removing very few outliers.

FUNCTIONALITY

The main packages that are used in Model Building and developing Shiny App are:

- ISLR :
  This package is used for statistical analysis, mainly for simple regression and multiple linear regression
- ggplot2
  This package is used to plot various graphs used for evaluating all the variables taken into consideration
- glmnet
  This package is used for performing lasso and ridge regression and to get the regression coefficients
- caret
  This package is used to streamline model building and also in evaluation process
- olsrr
  This package is used for diagnostics and variable selection
- shiny theme
  This package is used to add themes to Shiny App to make it look professional and visually appealing. The theme used for ‘University Rank Predictor’ is ‘cyborg’.

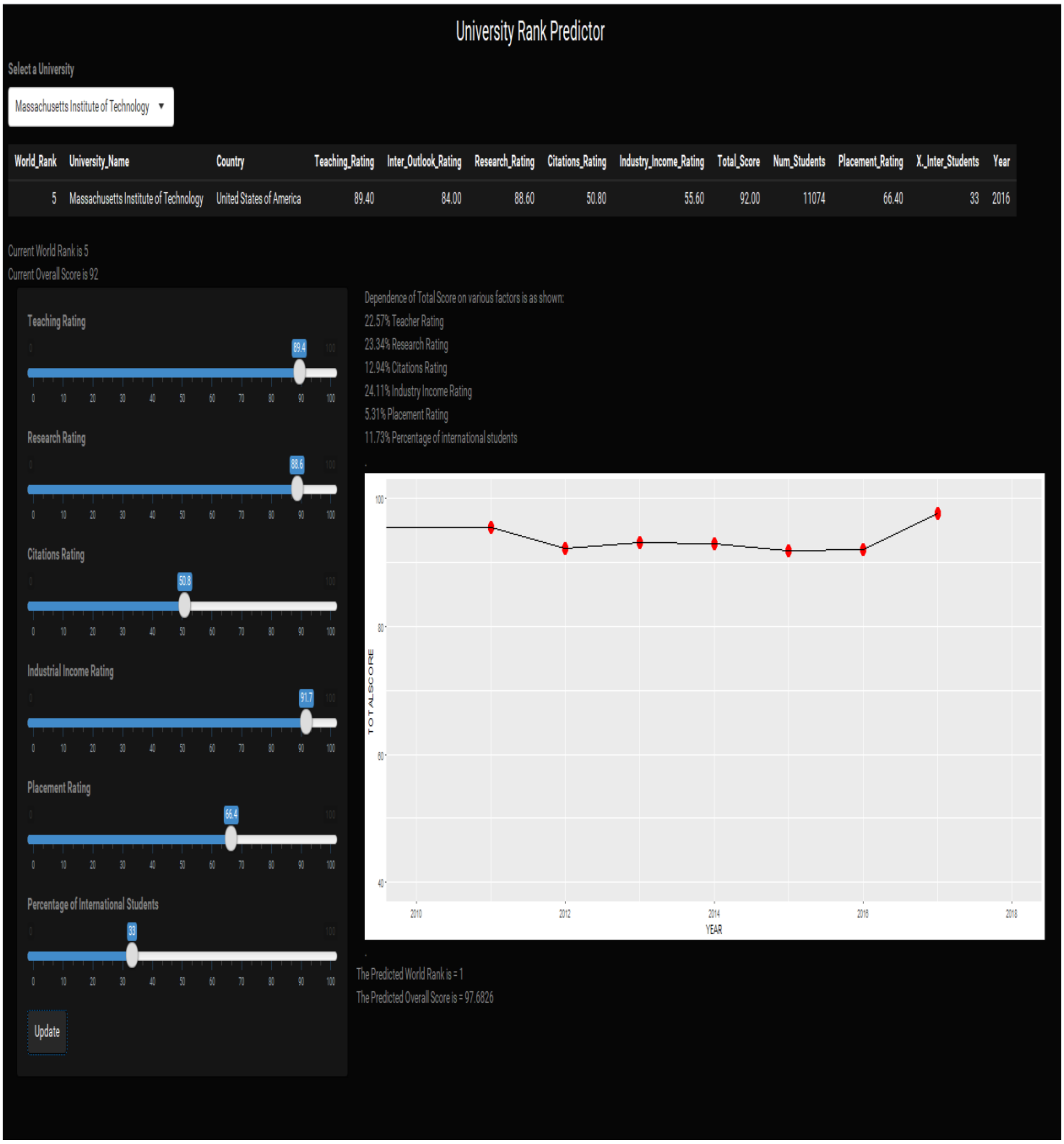
How the App works

The input parameter in our DSS Shiny App is University Name. Initially a university (university for which the rank is to be improved) is selected from a list of colleges in the drop-down menu. On selecting the university, its current ranking and overall score gets displayed at the top of the screen along with a table displaying all the ratings along with all the information related to the university directly from the dataset.

There are also 6 sliders available in the app in which the corresponding individual rating of all the factors influencing overall score gets appeared. The sliders display the corresponding teaching rating, research rating, citations rating, industrial income rating, placement rating and percentage of international students of the university selected (Figure 1).

On clicking the ‘Update’ button at the bottom of the window, the predicted world rank and overall score for the next year get displayed on the right side of the screen. A plot is also generated which shows the overall score of the university for the past 6 years and also the predicted score for next year(Figure 2).

The app is designed in such a way that the sliders act initially as the output and later on acts the input parameter. The university management can decide upon what factors to improve and check how the overall score increases for corresponding change in individual parameters being displayed in the sliders. On changing the ratings in sliders, the percentage increase in overall score differs for different parameters. For example, the percentage increase in overall score is maximum when industry income rating increases (Figure 3), and minimum when placement rating increases (Figure 4). The percentage dependence of overall score on each of the factors is also displayed in the window. Thus, the university management can invest on improving the significant parameters which affect the overall score and get an idea about how much they have to improve upon individual factors to improve their overall score to the desired extent.



University Rank Predictor

Select a University

Massachusetts Institute of Technology ▾

World_Rank	University_Name	Country	Teaching_Rating	Inter_Outlook_Rating	Research_Rating	Citations_Rating	Industry_Income_Rating	Total_Score	Num_Students	Placement_Rating	X_Inter_Students	Year
5	Massachusetts Institute of Technology	United States of America	89.40	84.00	88.60	50.80	55.60	92.00	11074	66.40	33	2016

Current World Rank is 5

Current Overall Score is 92

Teaching Rating: 89.4

Research Rating: 88.6

Citations Rating: 50.8

Industrial Income Rating: 91.7

Placement Rating: 66.4

Percentage of International Students: 33

Dependence of Total Score on various factors is as shown:

22.57% Teacher Rating

23.34% Research Rating

12.94% Citations Rating

24.11% Industry Income Rating

5.31% Placement Rating

11.73% Percentage of international students

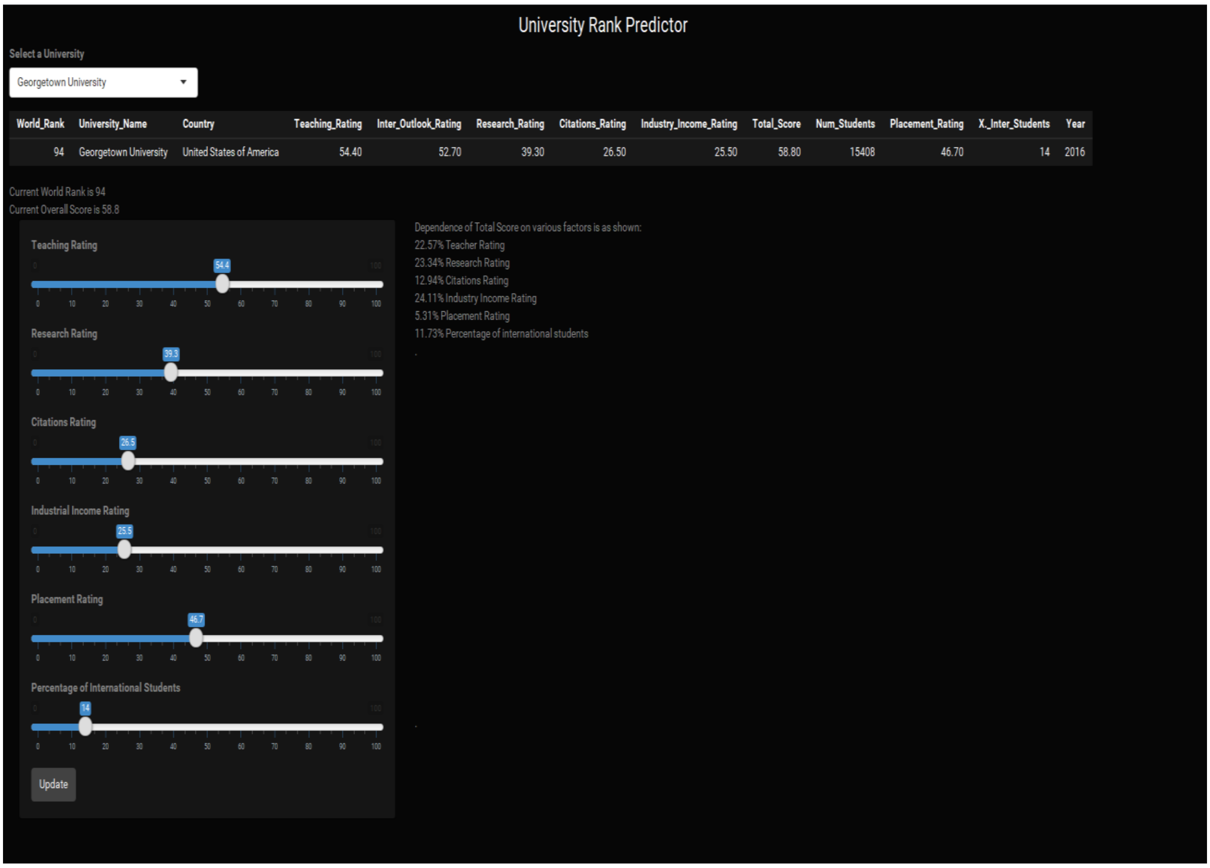
The Predicted World Rank is = 1

The Predicted Overall Score is = 97.6826

Enhancements

In our Shiny App, we have considered the overall ranking of a particular university. In fact, rankings vary for different departments in the university. Suppose if a dataset containing the ratings for all the departments in universities are available, it is possible to use the same techniques we have used to predict the individual rankings of all the departments of a particular university. The factors affecting the ranking could then be modified in such a way that the department rankings could also be improved.

GUI DESIGN AND QUALITY



University Rank Predictor

Select a University: Georgetown University

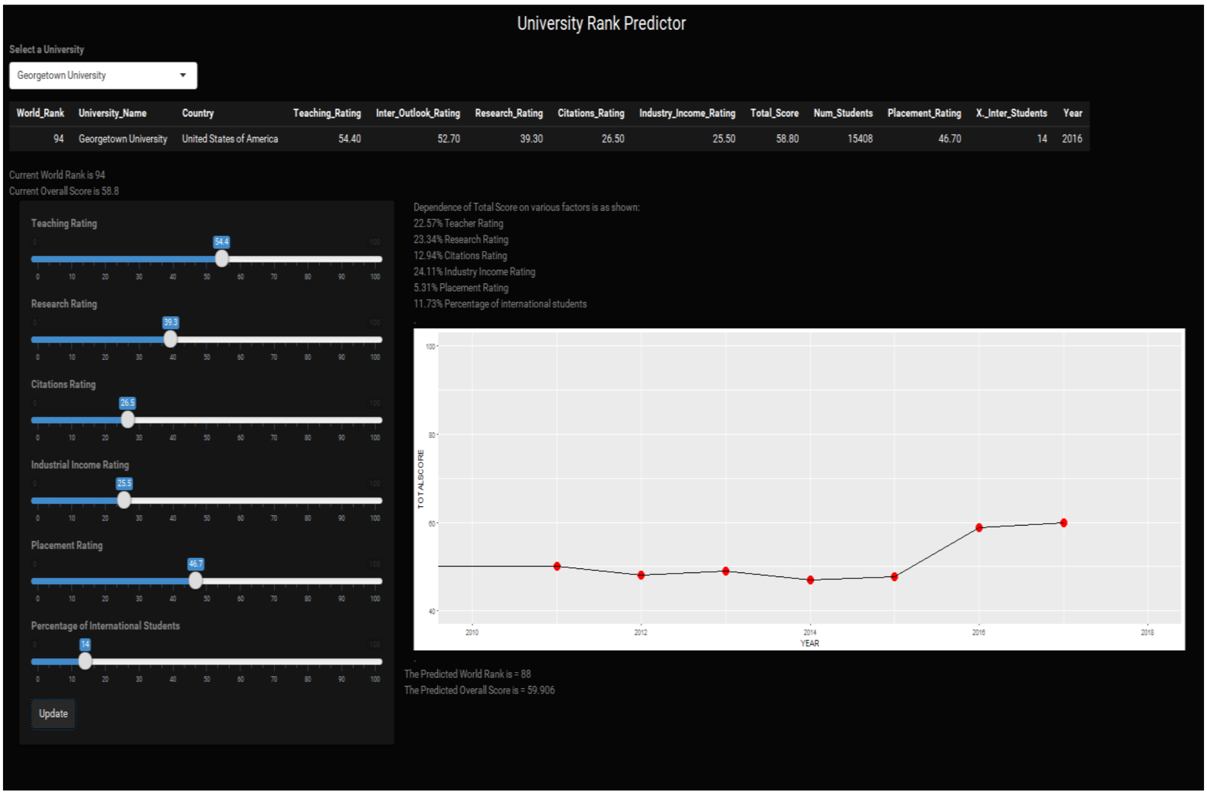
World_Rank	University_Name	Country	Teaching_Rating	Inter_Outlook_Rating	Research_Rating	Citations_Rating	Industry_Income_Rating	Total_Score	Num_Students	Placement_Rating	X._Inter_Students	Year
94	Georgetown University	United States of America	54.40	52.70	39.30	26.50	25.50	58.80	15408	46.70	14	2016

Current World Rank is 94
Current Overall Score is 58.8

Teaching Rating: 54.4
Research Rating: 39.3
Citations Rating: 26.5
Industrial Income Rating: 25.5
Placement Rating: 46.7
Percentage of International Students: 14

Dependence of Total Score on various factors is as shown:
22.57% Teacher Rating
23.34% Research Rating
12.94% Citations Rating
24.11% Industry Income Rating
5.31% Placement Rating
11.73% Percentage of international students

Figure 1: Initial setup



University Rank Predictor

Select a University: Georgetown University

World_Rank	University_Name	Country	Teaching_Rating	Inter_Outlook_Rating	Research_Rating	Citations_Rating	Industry_Income_Rating	Total_Score	Num_Students	Placement_Rating	X._Inter_Students	Year
94	Georgetown University	United States of America	54.40	52.70	39.30	26.50	25.50	58.80	15408	46.70	14	2016

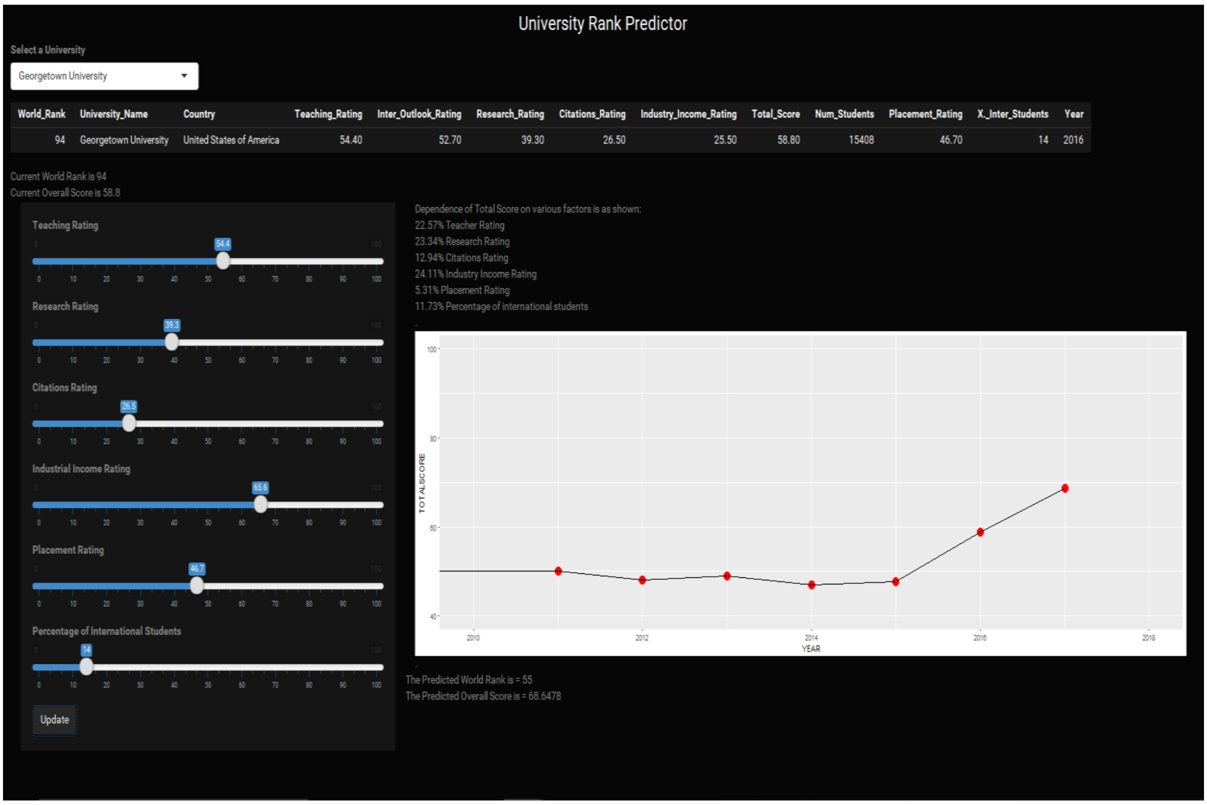
Current World Rank is 94
Current Overall Score is 58.8

Teaching Rating: 54.4
Research Rating: 39.3
Citations Rating: 26.5
Industrial Income Rating: 25.5
Placement Rating: 46.7
Percentage of International Students: 14

Dependence of Total Score on various factors is as shown:
22.57% Teacher Rating
23.34% Research Rating
12.94% Citations Rating
24.11% Industry Income Rating
5.31% Placement Rating
11.73% Percentage of international students

The Predicted World Rank is = 88
The Predicted Overall Score is = 59.906

Figure 2: After clicking ‘Update button’



University Rank Predictor

Select a University

Georgetown University

World_Rank	University_Name	Country	Teaching_Rating	Inter_Outlook_Rating	Research_Rating	Citations_Rating	Industry_Income_Rating	Total_Score	Num_Students	Placement_Rating	X._Inter_Students	Year
94	Georgetown University	United States of America	54.40	52.70	39.30	26.50	25.50	58.80	15408	46.70	14	2016

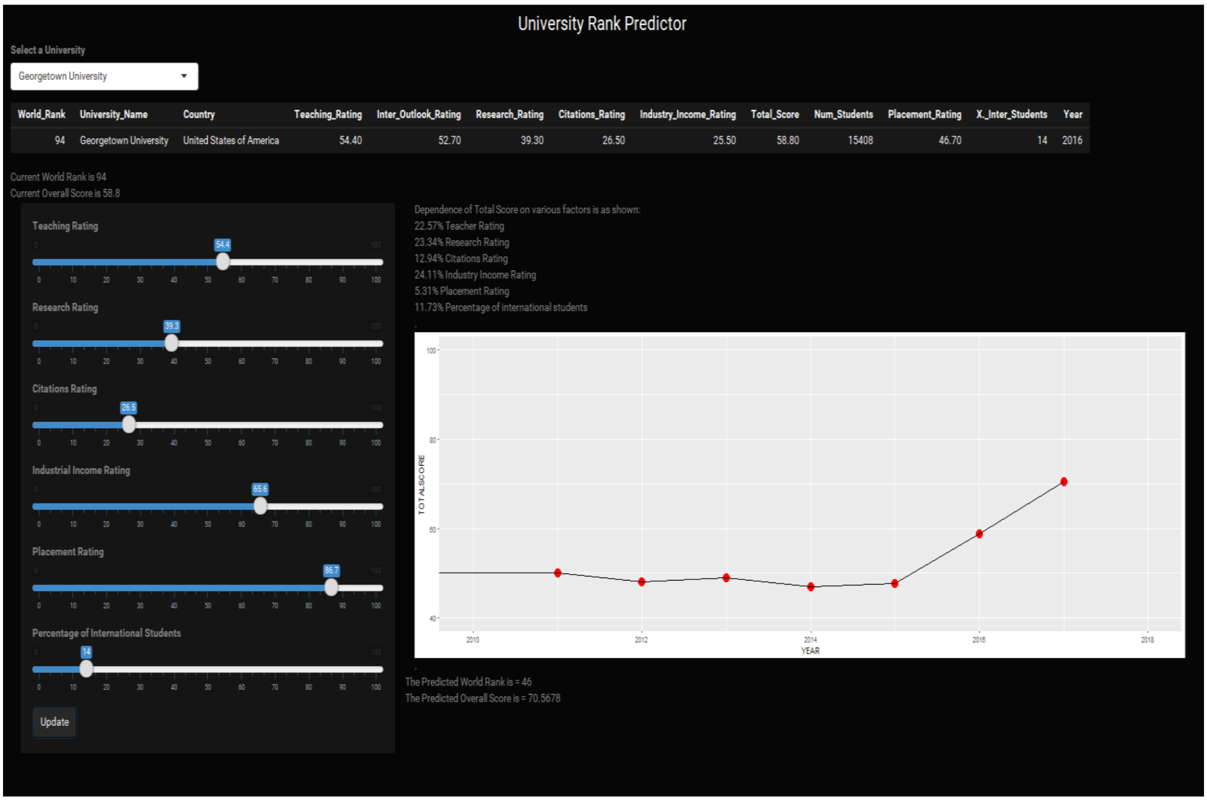
Current World Rank is 94
Current Overall Score is 58.8

Teaching Rating: 54.4
Research Rating: 39.3
Citations Rating: 26.5
Industrial Income Rating: 65.6
Placement Rating: 46.7
Percentage of International Students: 14

Dependence of Total Score on various factors is as shown:
22.57% Teacher Rating
23.34% Research Rating
12.94% Citations Rating
24.11% Industry Income Rating
5.31% Placement Rating
11.73% Percentage of international students

The Predicted World Rank is = 55
The Predicted Overall Score is = 68.6478

Figure 3: After increasing industrial income rating (for example, by 40%) the predicted score increases to a much higher value



University Rank Predictor

Select a University

Georgetown University

World_Rank	University_Name	Country	Teaching_Rating	Inter_Outlook_Rating	Research_Rating	Citations_Rating	Industry_Income_Rating	Total_Score	Num_Students	Placement_Rating	X._Inter_Students	Year
94	Georgetown University	United States of America	54.40	52.70	39.30	26.50	25.50	58.80	15408	46.70	14	2016

Current World Rank is 94
Current Overall Score is 58.8

Teaching Rating: 54.4
Research Rating: 39.3
Citations Rating: 26.5
Industrial Income Rating: 65.6
Placement Rating: 86.7
Percentage of International Students: 14

Dependence of Total Score on various factors is as shown:
22.57% Teacher Rating
23.34% Research Rating
12.94% Citations Rating
24.11% Industry Income Rating
5.31% Placement Rating
11.73% Percentage of international students

The Predicted World Rank is = 46
The Predicted Overall Score is = 70.5678

Figure 4: After increasing placement rating (for example, by 40%), the overall score changes only marginally

CONCLUSION

Based on the results from the DSS, we concluded that Industry Income Rating, Teaching Rating and Research Rating were the most significant predictors. Investing more time, money and research on these factors would increase the benefits multifold. We were also able to predict the rank for the next year using the DSS app. This app would be an ideal stepping stone for prediction and the best part is its flexibility for further upgradation using advanced data analytics techniques.

REFERENCE

1. https://www.statcrunch.com/5.0/shareddata.php?keywords=rankings