# Uni Rankings (CYO) Report

Riya Michael

24/06/2021

## I. Introduction

A significant method of Data Analytics is Machine Learning which facilitates model building. Originating from artificial intelligence, machine learning has been based on the concept of training a system to perform certain tasks by feeding in data. The system in turn, tries to identify patterns and finally optimizes models and arrives at conclusions while mitigating the need for human intervention.

Predictive Models in machine learning have been regarded as a highly beneficial tool which enables businesses and other sectors to make informed decisions. This is done with the help of historical data which is then trained and tested to provide highly accurate results. This enables the organisations to identify and track trends, gain insights on emerging markets, identify customer behavioral patterns and so on.

Some of the most common machine learning algorithms may include linear regression, logistic regression, decision trees, random forests, k - nearest neighbors (kNN) and Naive Bayes.

Successful implementation of Machine Learning Algorithms can be seen in the form of speech recognition technology such as Apple's Siri, self-driving cars such as Waymo, online education systems such as Duolingo and recommendation systems such as Netflix, to name a few.

The aim of this project is to develop a prediction model for any given university around the world, based on various variables such as national rank, influence, quality of faculty and so on. For this purpose, Generalized Linear Models (GLM Method) and Random Forests Algorithms have been executed.

***Data*** The historical data for this project has been collected from Kaggle. The data set is titled 'World University Rankings' and comprises of three files of global university rankings namely, The Times Higher Education University Ranking, The Academic Ranking of World Universities and The Center for World University Rankings. With each data set having its own pros and cons, after great consideration, the data published by The Center for World University Rankings has been used for this project. The details of the data have been explained below.

The World University Rankings data consists of 14 rows and 2200 columns, i.e. 2200 observations across 14 variables. It comprises of the data for various universities across 4 years , i.e., 2012- 2015.

```
dim(data)
```

```
## [1] 2200   14
```

```
summary(data)
```

```
##    world_rank     institution          country          national_rank
##  Min.   :   1.0   Length:2200        Length:2200        Min.   :  1.00
##  1st Qu.: 175.8   Class :character   Class :character   1st Qu.:  6.00
```

```
##  Median : 450.5   Mode  :character   Mode  :character   Median : 21.00
##  Mean   : 459.6                                        Mean   : 40.28
##  3rd Qu.: 725.2                                        3rd Qu.: 49.00
##  Max.   :1000.0                                        Max.   :229.00
##
##  quality_of_education alumni_employment quality_of_faculty  publications
##  Min.   :  1.0        Min.   :  1.0     Min.   :  1.0      Min.   :   1.0
##  1st Qu.:175.8        1st Qu.:175.8     1st Qu.:175.8      1st Qu.: 175.8
##  Median :355.0        Median :450.5     Median :210.0      Median : 450.5
##  Mean   :275.1        Mean   :357.1     Mean   :178.9      Mean   : 459.9
##  3rd Qu.:367.0        3rd Qu.:478.0     3rd Qu.:218.0      3rd Qu.: 725.0
##  Max.   :367.0        Max.   :567.0     Max.   :218.0      Max.   :1000.0
##
##    influence       citations      broad_impact       patents
##  Min.   :  1.0   Min.   :  1.0   Min.   :   1.0   Min.   :  1.0
##  1st Qu.:175.8   1st Qu.:161.0   1st Qu.: 250.5   1st Qu.:170.8
##  Median :450.5   Median :406.0   Median : 496.0   Median :426.0
##  Mean   :459.8   Mean   :413.4   Mean   : 496.7   Mean   :433.3
##  3rd Qu.:725.2   3rd Qu.:645.0   3rd Qu.: 741.0   3rd Qu.:714.2
##  Max.   :991.0   Max.   :812.0   Max.   :1000.0   Max.   :871.0
##                                  NA's   :200
##      score             year
##  Min.   : 43.36   Min.   :2012
##  1st Qu.: 44.46   1st Qu.:2014
##  Median : 45.10   Median :2014
##  Mean   : 47.80   Mean   :2014
##  3rd Qu.: 47.55   3rd Qu.:2015
##  Max.   :100.00   Max.   :2015
##
```

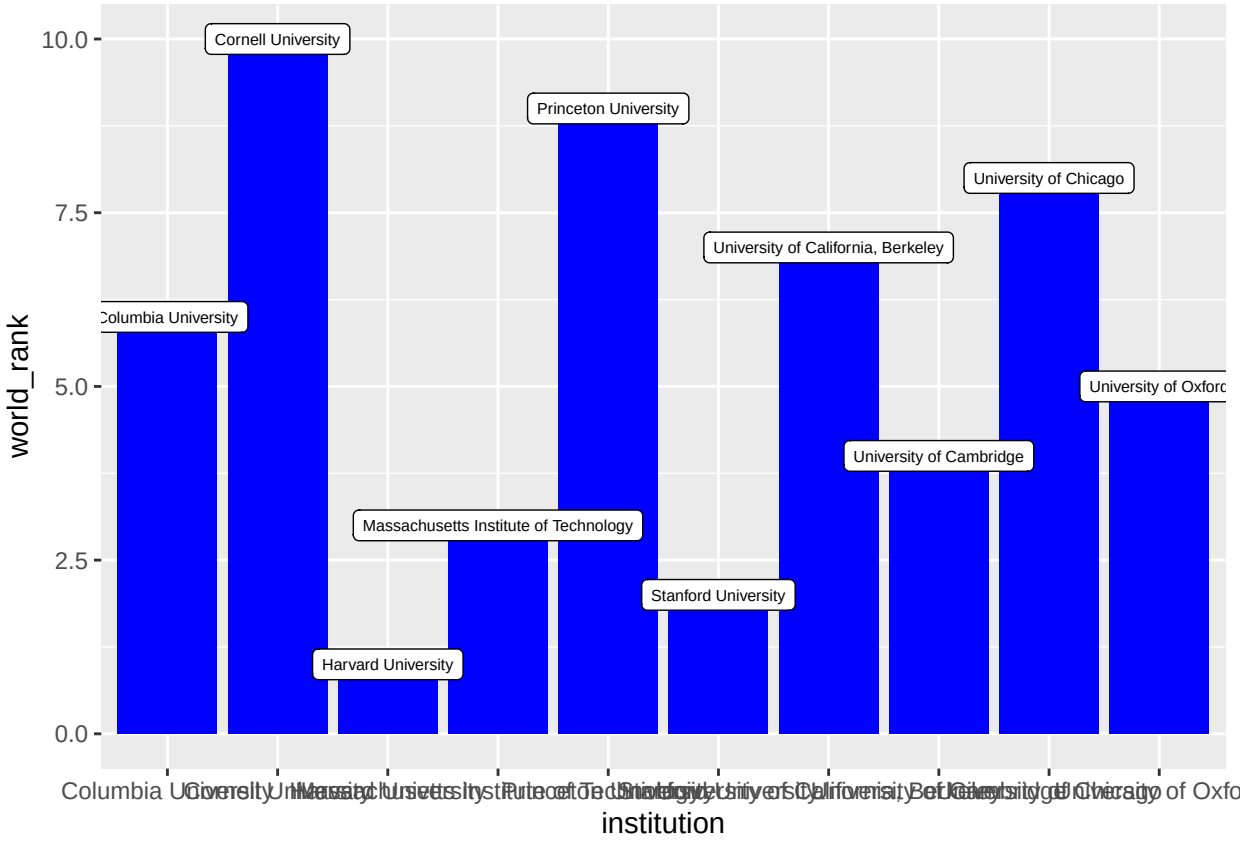
In the table below, the variables names along with their descriptions has been explained.

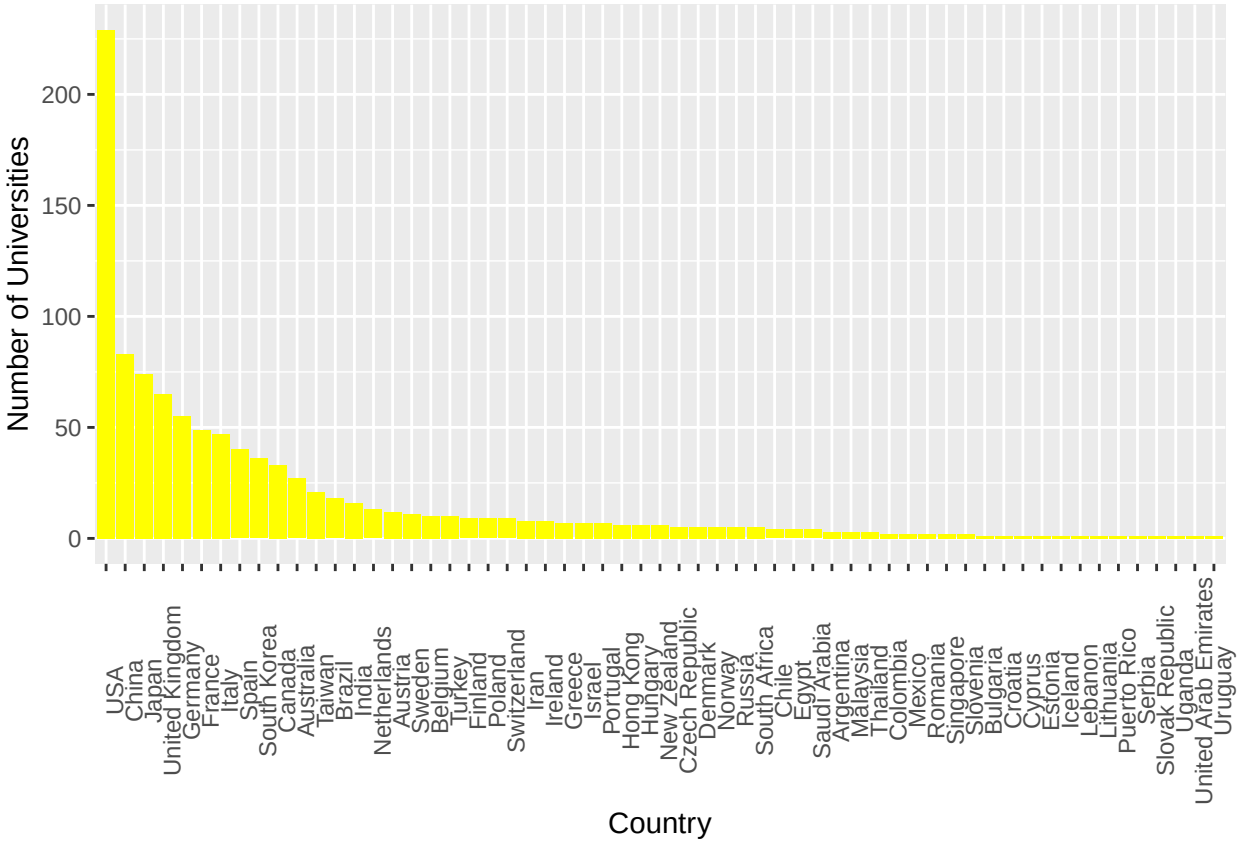
| Sr.No. | Variable Names | Variable Description |
|---|---|---|
| 1 | World Rank | The ranking of the university from a global perspective |
| 2 | Institution | The name of the university |
| 3 | Country | The country the university is situated in |
| 4 | National Rank | The ranking of the university from a national perspective |
| 5 | Quality of Education | The ranking of the university's quality of education |
| 6 | Alumni Employment | The ranking of the university's alumni employment |
| 7 | Quality of Faculty | The ranking of the university's quality of faculty |
| 8 | Publications | The ranking of the university's publications |
| 9 | Influence | The ranking of the university's influence |
| 10 | Citations | The number of students at the university |
| 11 | Broad Impact | The ranking of the university's broad impact (only available for 2014 and 2015) |
| 12 | Patents | The ranking of the university's patents |
| 13 | Score | The total score calculated by considering the previous factors and in turn used to determine the world rank |
| 14 | Year | Year of ranking (2012 - 2015) |

*Since the data provided under the variable "Broad Impact" contains entries only for 2014 and 2015, for

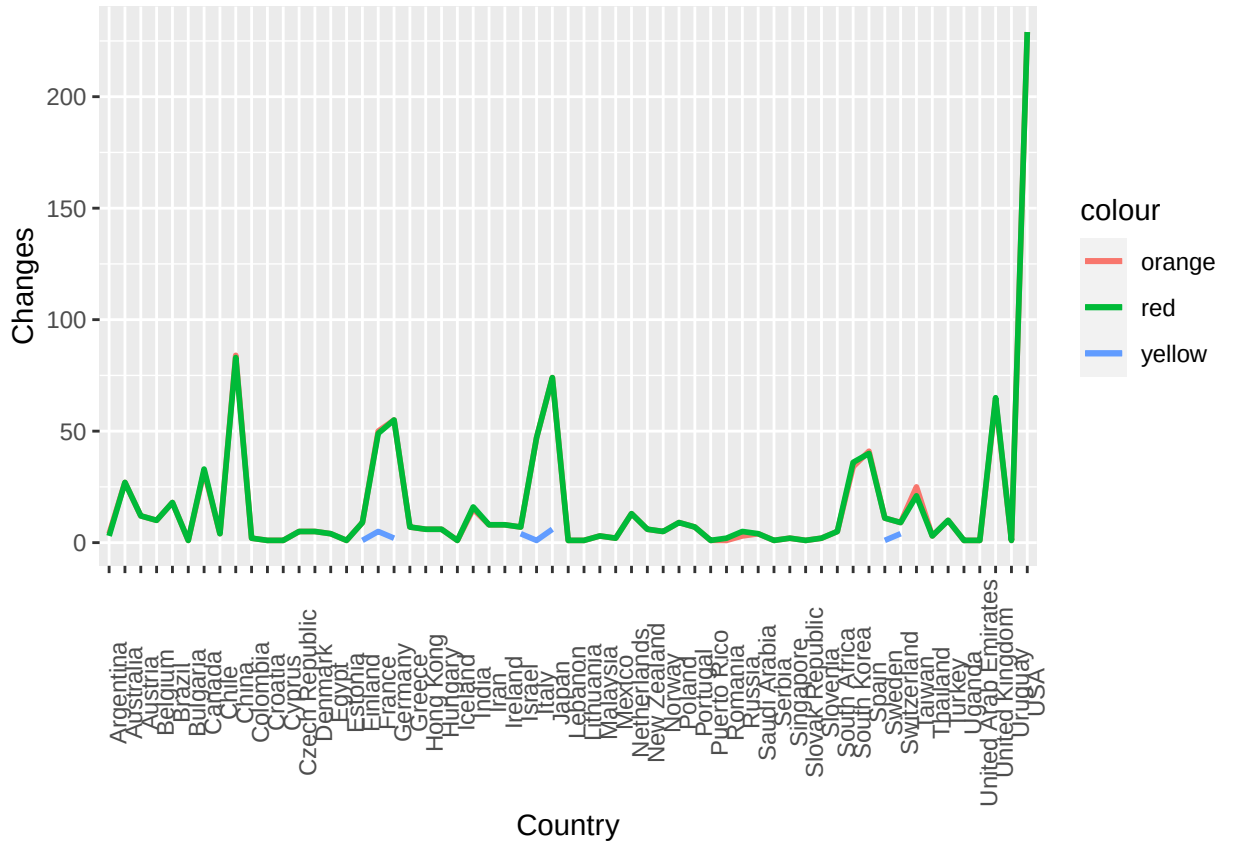
the purpose of uniformity, this particular variable has been omitted from the analysis. Hence, all other parameters, namely, national rank, quality of education, alumni employment, quality of faculty, publications, influence, citations and patents along with score as our main dependent variable have been used for this project.

We can analyse the universities which have higher ranks i.e. the top universities by observing the data for the most recent year which is 2015 as shown below.






From the above graph, we can observe that the data for some universities changes every year i.e. the number of universities whose data has been compiled has been varying over the years (for certain countries). The same can be ascertained from the following table.




**II. Analysis**

For the purpose of building a predictive model to determine the scores of various universities around the globe, two methods have been used to build the algorithm.

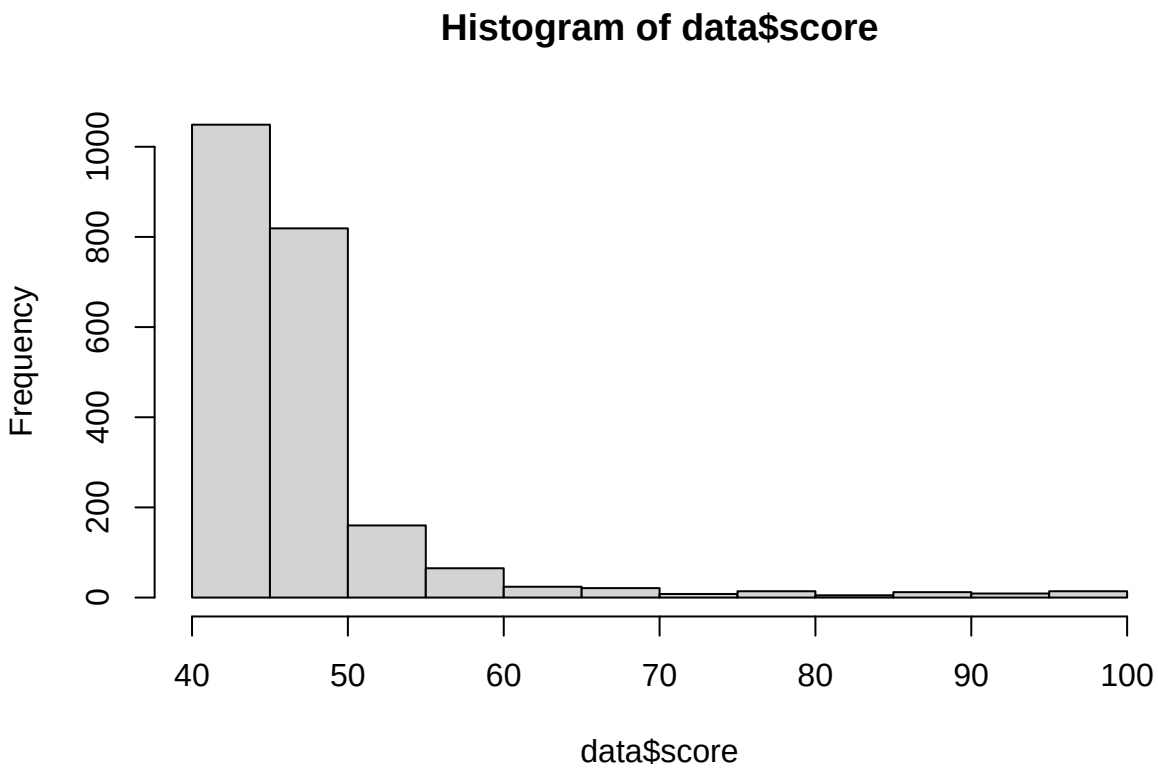
1. GLM (Generalized Linear Models)

2. Random Forests

Like any other algorithm in machine learning, here too, the data has been divided into two sets for training and testing. The train set will consist of 70% of the data and the test set will consist of the remaining 30%. The purpose of carrying out these steps is to use our train set to build the algorithm. The built algorithm will then be applied on the test set. By doing this, we can verify whether the fitted algorithm produces similar, if not same results by comparing it with the actual data of the test set. The proportion of 70:30 has been considered as the optimal ratio to split the data as a lower portion of the the train set (or higher portion of the test set) will be not be helpful in building the algorithm. Moreover, if we consider a lower portion of the test set (or higher portion of the train set), though it will aid in building the algorithm, the available data in the test set will not be sufficient to fit the algorithm i.e. to verify whether our predicted scores and actual scores match. Hence, 70:30 is the optimal ratio for splitting the data in this project. The same has be ascertained by calculating the RMSEs after accounting for 50:50, 60:40, 70:30 and 80:20 ratios. The RMSE for the 70:30 ratio turned out to be lower that the others, thereby proving that there are lesser chances of errors if we stick to this ratio.

In this project, RMSE or Root of Mean Squared Error has been used as our success indicator i.e. the figure that will help us analyse the accuracy of our prediction results. We can visually verify the accuracy of our algorithm by comparing the predicted scores and actual scores of the test set.

***Building the algorithm with the help of the train set***




**1. GLM**

As observed from the histogram above, the scores are not normally distributed. One of the biggest advantages of using the GLM method is that it does not only consider data sets having normal distributions, but also the data sets that are not normally distributed. Moreover, as observed from the output (later on), one can easily interpret the results and develop a clear understanding of the influence of every predictor on the outcome.

```
#Splitting the data into Train and Test Sets
set.seed(1)
train_index <- createDataPartition(data$score, times = 1, p = 0.7, list = FALSE)
train <- data[train_index,]
test <- data[-train_index,]
```

```
#RMSE calculate after taking all variables into consideration
RMSE(test$score,test$pred_score)
```

```
## [1] 5.471756
```

```
RMSE_GLM <- RMSE(test$score,test$pred_score)
```

```
#After removing publications as one of the variables
fit <- glm(score ~ national_rank
           + quality_of_education
           + alumni_employment
```

```
            + quality_of_faculty
            + influence
            + citations
            + patents,
            data  = train)

test <- test %>% mutate(pred_score = predict.glm(fit,newdata = test))
```

```
#RMSE after eliminating the variable "publications"
RMSE(test$score,test$pred_score)
```

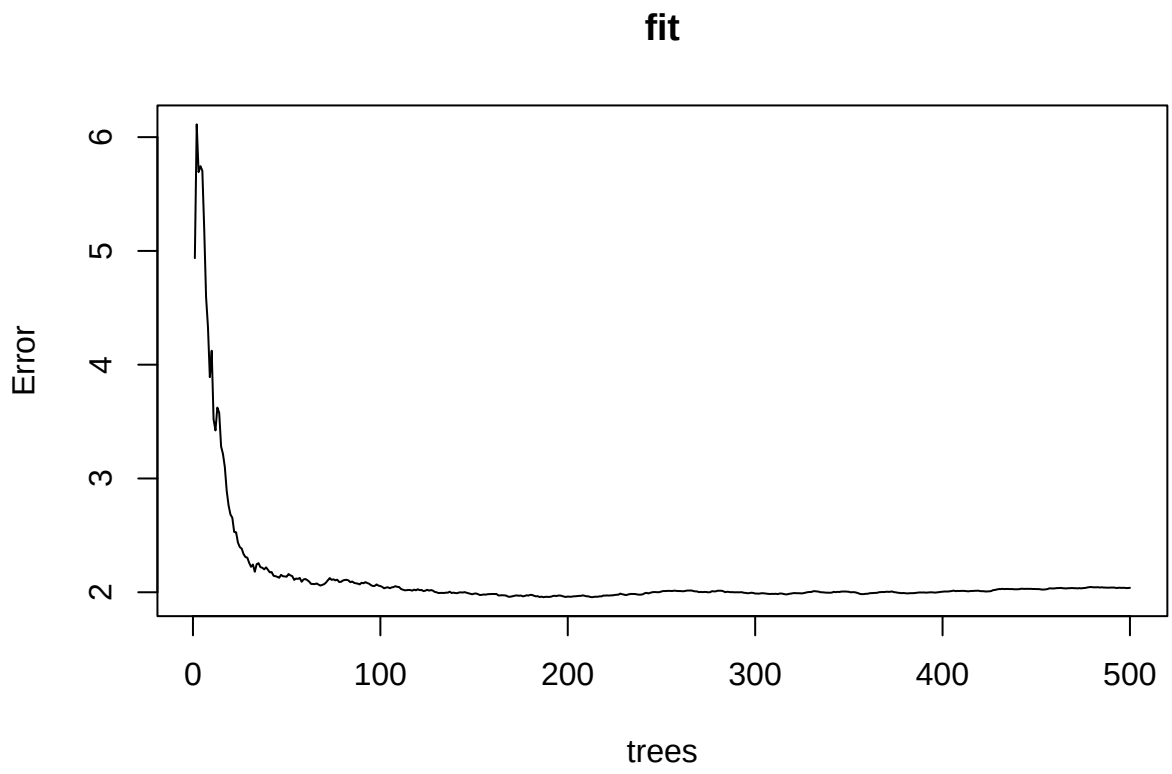
```
## [1] 5.471333
```

After calculating the RMSE by removing every individual variable, we can observe that the RMSE improves i.e. reduces after eliminating the "Publications" variable. The same does not hold true for the other variables since the RMSE increases. A decrease in RMSE indicates an improvement in the model since it minimizes the error generated.

**2. Random Forests** The purpose for using the random forests model is to mitigate instabilities and improve the accuracy of the prediction. This is done by taking the average of numerous decision trees i.e. a forest which is constructed by randomness and thereby optimizing our prediction results. Furthermore, random forests method is a quicker and simpler approach to generate predictive algorithms.

```
#Data partitioning
set.seed(1)
train_index <- createDataPartition(data$score, times = 1, p = 0.5, list = FALSE)
train <- data[train_index,]
test <- data[-train_index,]

#fitting the random forest model to the train set
fit <- randomForest(score ~ national_rank
                    + quality_of_education
                    + alumni_employment
                    + quality_of_faculty
                    +influence
                    + citations
                    + patents,
                    data = train)
plot(fit)
```




From the above plot, we can observe that the error reduces and further becomes stable after 300 trees. Hence, to optimize our model, we can use the random forest method by generating 300 random trees. The result can be seen below.

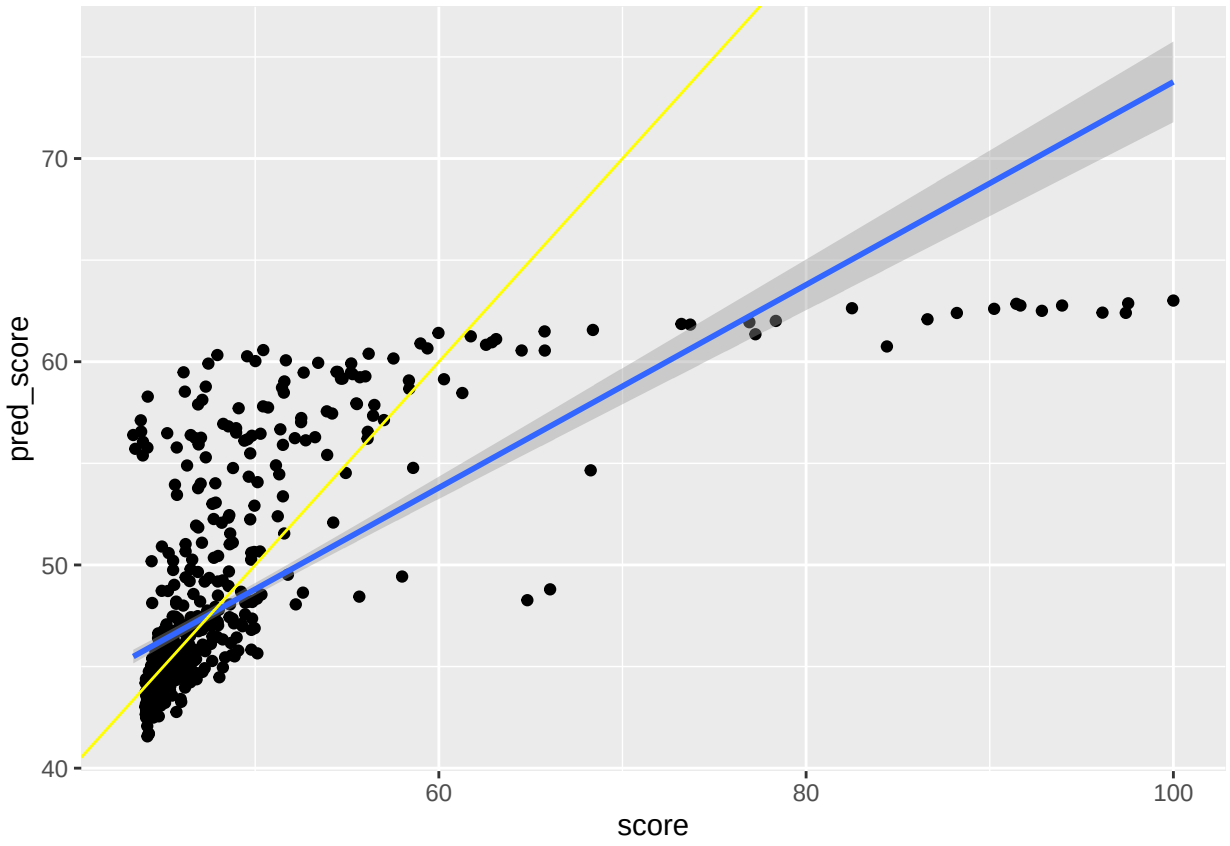
```
#Since the error reduces and stabilizes at around 300 trees, setting ntrees = 300
train_rf <- randomForest(score ~ national_rank
                         + quality_of_education
                         + alumni_employment
                         + quality_of_faculty
                         +influence
                         + citations
                         + patents,
                         data = train,
                         ntree=300)
print(train_rf)
```

```
##
## Call:
##  randomForest(formula = score ~ national_rank + quality_of_education +      alumni_employment + quali
##                Type of random forest: regression
##                      Number of trees: 300
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 2.109558
##                     % Var explained: 96.29
```

## III. Results

***Fitting the test set with the built algorithm and comparing the predicted and actual scores***

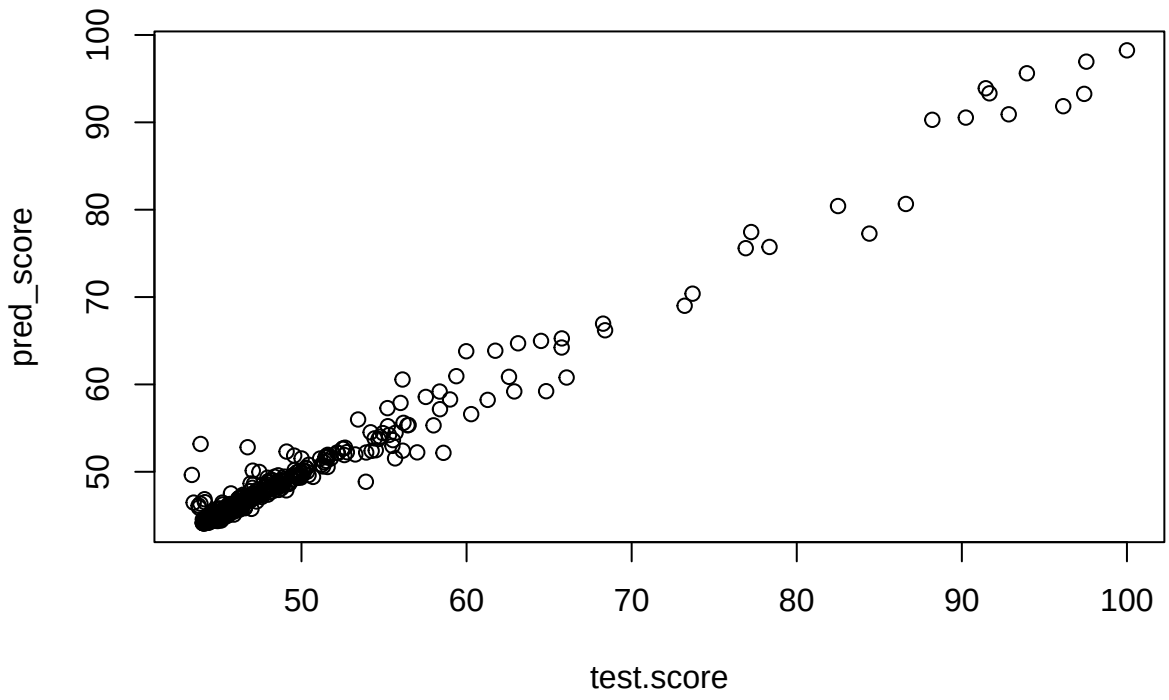
**1. GLM** Finally, the results from the GLM model can be seen below. Along with coefficients for the linear equation.



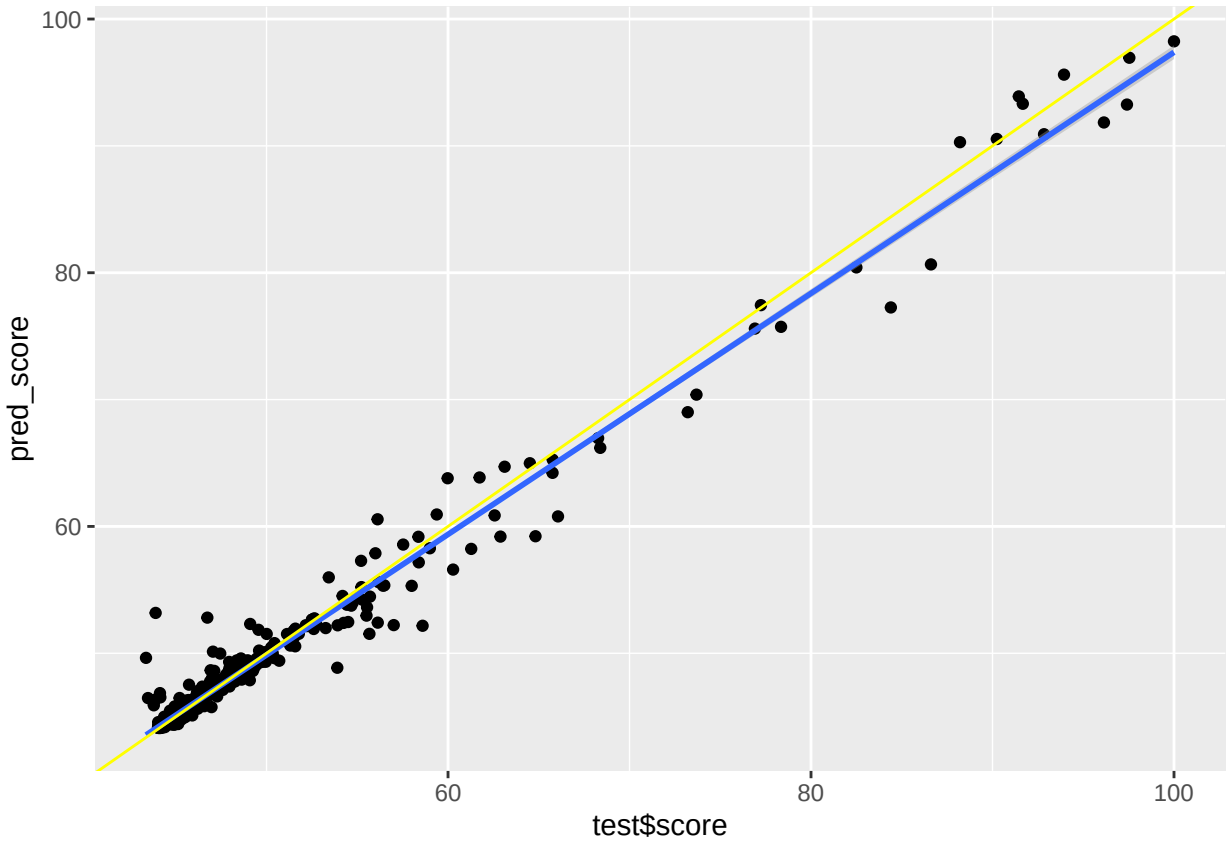

```
##          (Intercept)        national_rank quality_of_education
##         6.309210e+01        -6.838782e-03        -3.083549e-03
##    alumni_employment   quality_of_faculty            influence
##        -5.456784e-03        -6.169036e-02        -1.060100e-03
##            citations              patents
##        -5.192396e-05        -1.614380e-03
```

By analyzing the coefficients of the variables, one can observe the degree of influence that each parameter has on the dependent variable i.e. score. On observing the plotted graph, it is evident that the all the predicted scores do not match with all the actual scores. In simpler terms, the predicted scores for all the observation of the test set do not match the actual scores of all the observations of the test set. Therefore, the accuracy can be said to be low.

**2. Random Forests** The results from fitting the random forest algorithm can be seen below.




```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Checking the accuracy
RMSE(test$score,pred_score)
```

```
## [1] 1.112594
```

```
RMSE_RF <- RMSE(test$score,pred_score)
```

From the predicted scores and actual scores graph, one can observe that both the lines are close to overlapping. This indicates that the predicted results for most of the observations match the actual scores of those observations of the test set. This in turn, indicates a higher accuracy.

***Comparing the RMSEs of GLM and Random Forests Algorithms***

| Sr. No. | Algorithm Used | RMSE Generated |
|---|---|---|
| 1 | Generalized Linear Models | 5.472 |
| 2 | Random Forests | 1.113 |

As the RMSE generated from the random forests model is lower than the GLM method, it is safe to say that the random forests method proves to be the more optimal method to generate the predictive algorithm. This can be attributed to the use of multiple trees, 300 in this case, to optimize the model and provide a more accurate result.

## IV. Conclusion

With the indispensable techniques of machine learning, it has become simpler to build prediction algorithms with each technique having its own pros and cons. In this project, the GLM and Random Forests algorithms have proved beneficial in developing the prediction model. The results from the Random Forests algorithm have proved to be more promising as compared to the GLM model. Not only has Random Forests proved to be a simpler approach but it has also provided us with a higher accuracy (lower RMSE). In this way, this project has been able to develop a predictive model that predicted the scores of the universities across the globe by using regression analysis.

A stronger report can be built by further analyzing the data published by The Times Higher Education University Ranking and The Academic Ranking of World Universities and comparing the ranking parameters across these data sets. Furthermore, one can collate the data for the more recent years 2016 - 2020/2021 and build their model after taking these years into consideration to get more relevant results. Additionally, one can apply other machine learning algorithms and compare their results to identify the best model for predicting university scores which in turn will help establish their world ranks.