Agenda 2063 Progress Analysis

Hussam Hanafi

11/3/2019

# Introduction

[Agenda 2063](https://au.int/en/agenda2063/overview) is the African Unions’ Strategic framework for delivering inclusive and and sustainable development signed on May 2013.

As part of this agenda the AU has set over-arching goals and more granular targets to be achieved in the 50 years from 2013 to 2063.

The purpose of this report is to evaluate progress on two of those targets in the medium term [First Ten Year Implementation Plan](https://au.int/sites/default/files/documents/33126-doc-ten_year_implementation_book.pdf) by analyzing Indicator data from the [World Bank Open Data](https://data.worldbank.org/) Initiative.

## First Ten Year Implementation Plan Targets

The Two Targets were selected partially arbitrarily in the hope that they would be indicative of each other; however, the selection was also driven by the completeness of the data sets as there is often large gaps in data sets dealing with progress indicators due to a number of factors mostly affecting the data collection stage.

For this Report the two targets/indicators chosen were as follows:

* As part of the first Goal: A High Standard of Living, Quality of Life and Well Being for All; the FTYIP sets a target of reducing Unemployment relative 2013 by at least 25%
* As Part of the fourth Goal: Transformed Economies and Job Creation; the FTYIP sets a target of 7% Annual growth

This Report will use those two indicators to attempt to analyse progress and attempt to predict the value of these indicators by 2023.

# Data Analysis

We will analyse each of these indicators separately using the following steps:

1. Accessing the data from the World Bank Open Data web-site
2. Filtering out AU Member States
3. Wrangling the data into tidy format
4. Initial Visualization
5. Wrangling the data to minimize the effect of missing data points
6. Visualizing the Data for Analysis Purposes
7. Sub-setting a Test Set to be used with our models

## Unemployment Rate

### Data Wrangling

We Will start the with the Unemployment Rate, Loading the Data from the World Bank modeled estimate with the code below:

### Loading Unemployment Rate  
## Creating Raw Data Folder  
Raw\_Data\_Folder <- file.path(getwd(),"Raw\_Data")  
dir.create(Raw\_Data\_Folder)  
  
##Downloading and Unpacking the File  
#download.file("http://api.worldbank.org/v2/en/indicator/SL.UEM.TOTL.ZS?downloadformat=csv",   
# file.path(Raw\_Data\_Folder,"API\_SL.UEM.TOTL.ZS\_DS2\_en\_csv\_v2\_422140.zip"), mode = "wb")  
Unemployment\_zip <- "API\_SL.UEM.TOTL.ZS\_DS2\_en\_csv\_v2\_422140.zip"  
Unemployment\_csv <- "API\_SL.UEM.TOTL.ZS\_DS2\_en\_csv\_v2\_422140.csv"  
unzip(file.path(Raw\_Data\_Folder,Unemployment\_zip), exdir = file.path(Raw\_Data\_Folder,"Unemployment"))  
  
##Reading the File into R  
Unemployment <- read\_csv(file.path(Raw\_Data\_Folder,"Unemployment",Unemployment\_csv), skip = 3)

Wrangling our Data into tidy format with the following code:

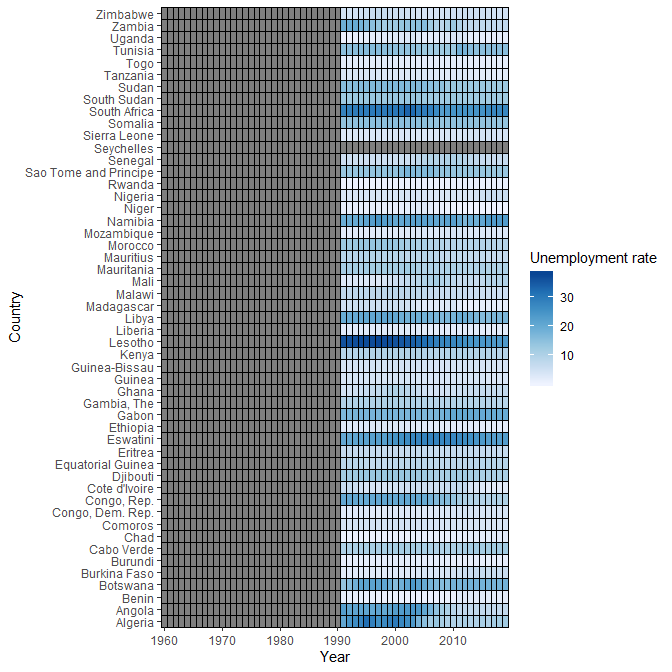
## Tidying the Data  
colnames(Unemployment) <- str\_replace(colnames(Unemployment), " ", "\_")   
Unemployment <- Unemployment %>%  
 filter(Country\_Name %in% AU) %>%   
 gather(key = "Year", value = "Unemployment", `1960`:`2019`, convert = TRUE) %>%   
 select(-c(X65,Indicator\_Code)) %>%   
 filter(Country\_Name %in% AU) %>%   
 select(-Country\_Code, -Indicator\_Name)

Below is a sample of the data frame:

Unemployment %>% sample\_n(4) %>% knitr::kable(col.names = c("Country","Year","Unemployment rate"))

|  |  |  |
| --- | --- | --- |
| Country | Year | Unemployment rate |
| Djibouti | 1962 | NA |
| Mauritius | 1964 | NA |
| Sierra Leone | 2001 | 3.52 |
| Burkina Faso | 1979 | NA |

Our Initial Visualization of the data shows some clear places where we have missing data which we’ll work to remove:



The below code achieves this:

## Removing years and countries with no data  
No\_Data <- Unemployment %>%  
 group\_by(Year) %>% summarize(No\_Data = sum(!is.na(Unemployment)))   
Unemployment <- Unemployment %>%   
 left\_join(No\_Data, by = "Year") %>% filter(No\_Data != 0) %>% select(-No\_Data)   
  
No\_Data <- Unemployment %>% group\_by(Country\_Name) %>% summarize(No\_Data = sum(!is.na(Unemployment)))   
Unemployment <- Unemployment %>%   
 left\_join(No\_Data, by = "Country\_Name") %>% filter(No\_Data != 0) %>% select(-No\_Data)

However since the stated goal is to reduce Unemployment by 25% relative to 2013 we will extract the 2013 values and use those as our Baseline; additionally, the below code will also set our test data and remove it from our training set:

## Setting Baseline and Test Sets  
# Baseline  
Unemployment %>% group\_by(Country\_Name) %>% filter(Year == c(2013)) %>%   
 summarize(Baseline = Unemployment) %>% filter(is.na(Baseline))

## # A tibble: 0 x 2  
## # ... with 2 variables: Country\_Name <chr>, Baseline <dbl>

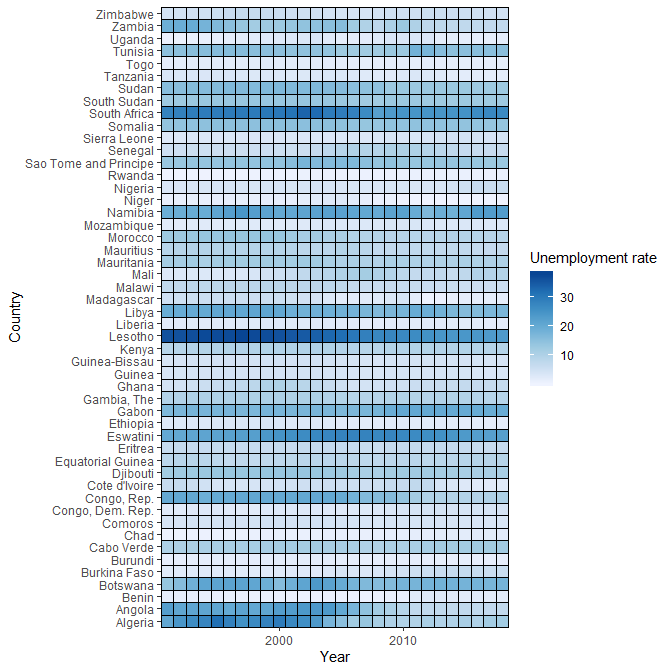
Baseline\_Unemployment <- Unemployment %>%   
 group\_by(Country\_Name) %>% filter(Year == (2013)) %>%   
 summarize(Baseline = Unemployment) %>% filter(!is.na(Baseline))  
  
# Tests Set  
Unemployment %>% group\_by(Country\_Name) %>% filter(Year == c(2019)) %>%   
 summarize(Baseline = Unemployment) %>% filter(is.na(Baseline))

## # A tibble: 0 x 2  
## # ... with 2 variables: Country\_Name <chr>, Baseline <dbl>

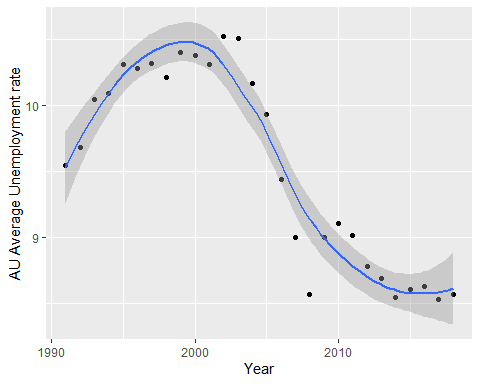
Unemployment\_test <- Unemployment %>% filter(Year == 2019)  
  
# Removing Test Data  
Unemployment <- Unemployment %>% filter(Year != 2019)

### Visualizations and Analysis

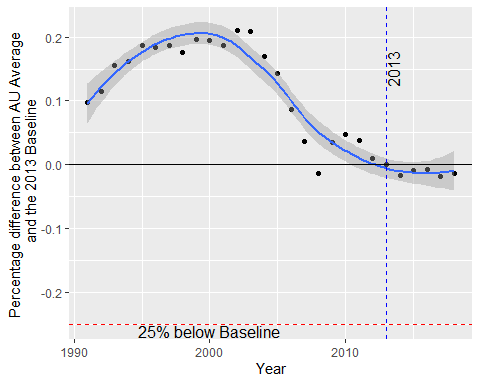
Now we are ready to Look at our data and extract some useful analyses. We repeat the initial visualization to confirm:



Next we Look at the Continent Wide Average Unemployment rate and try discern any trends:

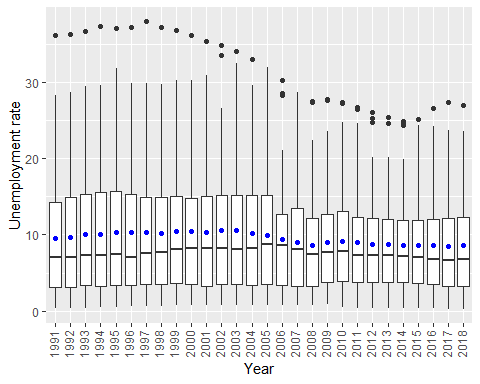


We can clearly see a downward trend with a potential leveling off, however this doesn’t provide a meaningful comparison to the stated target, instead we take a comparison between the average and the baseline we set:

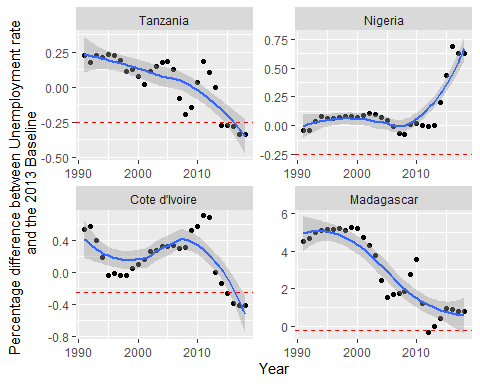


As expected we see the same trend, it is however clear that the continent wide performance leaves something to be desired with the reduction in Unemployment rate by 2019 not approaching the target.

A Box-plot showing the unemployment rate over the years paints a more nuanced picture showing large variations between Member States and a mean (in blue) that is significantly larger than the median implying that most member states are performing better than average:



To illustrate this variation we look at the two top and bottom performing countries in 2018



## Annual GDP Growth

We Repeat this process with data relating to GDP Growth, in this section of the report we will show less of the code as there are essential similarities in the process.

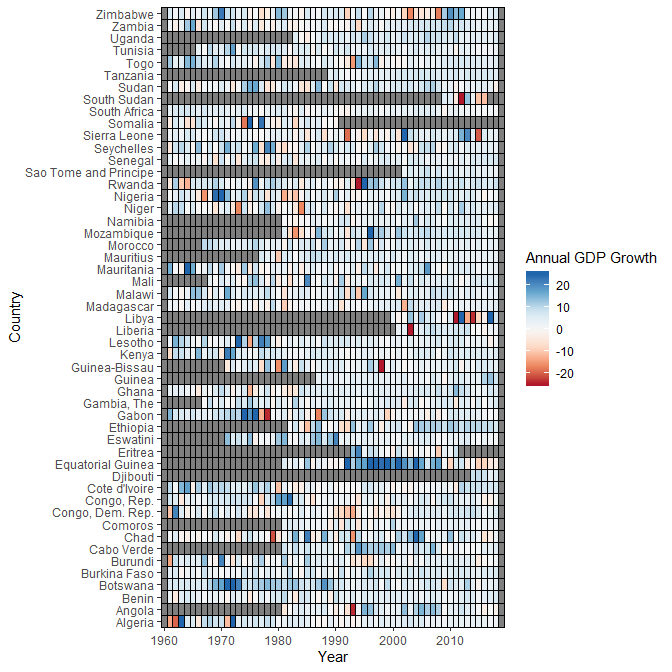
### Data Wrangling

Once we have loaded and wrangled the Annual GDP Growth Data from the World Bank Open Data website

Below is a sample of our tidied data:

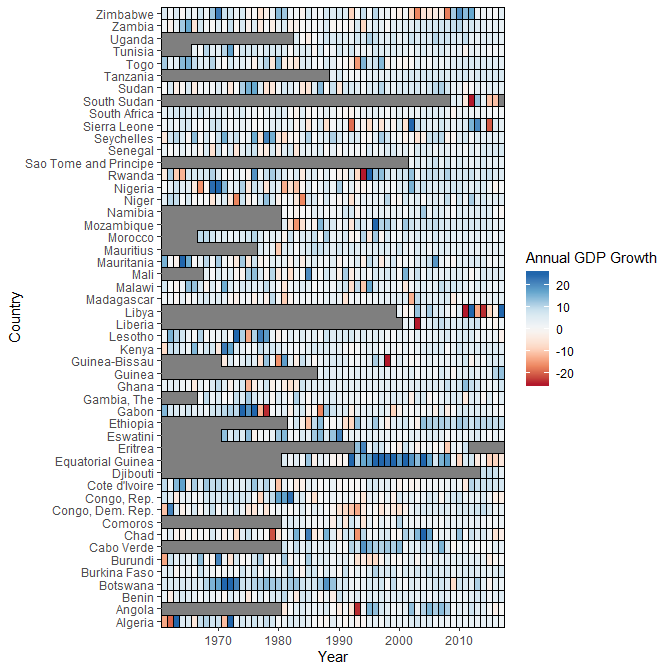
|  |  |  |
| --- | --- | --- |
| Country | Year | Unemployment rate |
| Congo, Rep. | 1973 | 8.231 |
| Sierra Leone | 2010 | 5.346 |
| Libya | 1991 | NA |
| Niger | 1989 | 0.934 |

While here we make our initial visualization; it’s important to not that for the sake of clarity outliers in both directions were scaled down to 25% and -25% Growth rate respectively:



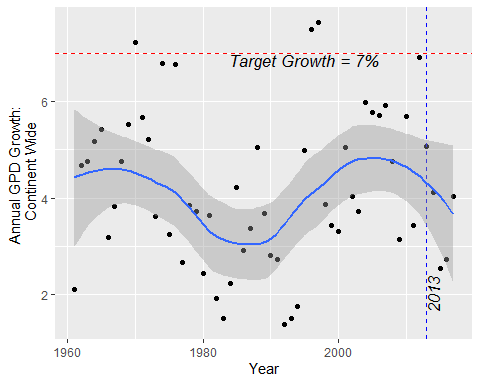
We can see that the missing data is more sporadic than with the Unemployment rate, we clean up the missing data, opting to estimate missing data points for our test set with data from the closest year. The exception to this is Somalia where the latest data is from 1990 thus we exclude it from this analysis.

We now visualize the data in our Wrangled data set after the exclusion of the Test data

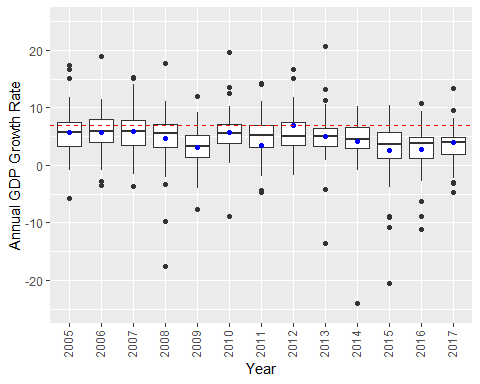


### Visualization and Analysis

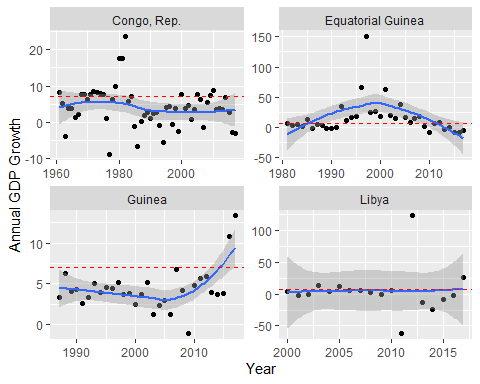
We start by looking for Continent wide trends:



The Scatter plot shows a massive variation from year to year with most of the years falling below the 7% Target. A Box-plot confirms this variation; please note that for clarity we only look at a subset of the years and do not show the largest outliers:

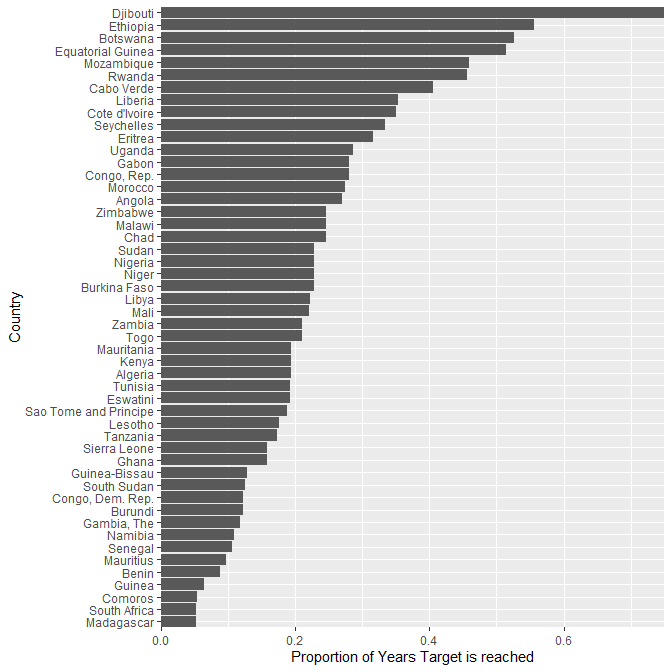


From here it is clear that the majority of Member states are not achieving the set target of 7%, a look at the countries with the highest and lowest growth rates provides a clearer picture of individual trends:

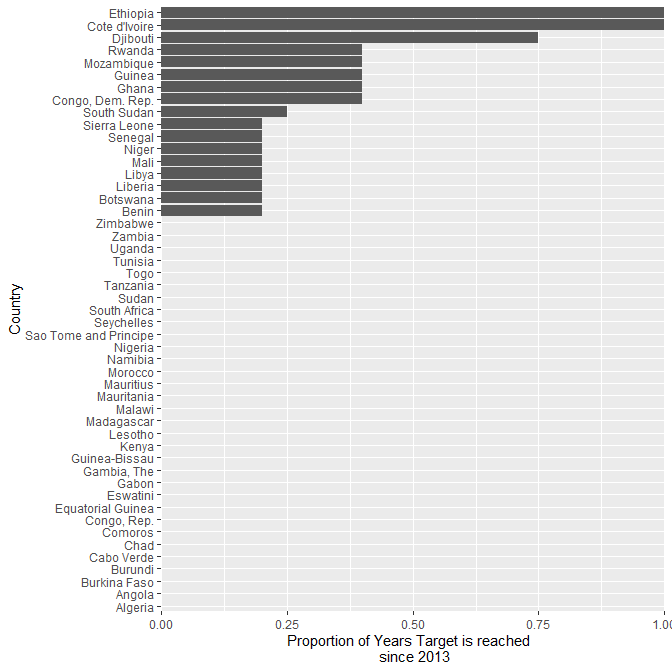


However it fails to give us an idea of what countries are performing since a consistent growth is just as if not more important than a high growth rate; additionally, countries like Libya with high variation that suffer large GDP losses can skew the trend by making equally large gains.

We propose that looking at the proportion of years the target is reached is a more indicative metric:



If we narrow our view to the years since the launch of Agenda 2063 we get yet a different look at how Member States are performing:



# Methodology

Building our model we will use 10 fold Cross-Validation across all models. For the purposes of this report we will compare four models:

1. Linear Regression
2. K-Nearest Neighbors
3. Gradient Boosting
4. Random Forest

To determine the effectiveness of our models we will use the Root Mean Square Error between our predictions and the test set as a measure; optimizing our parameters as we

## Unemployment Rate

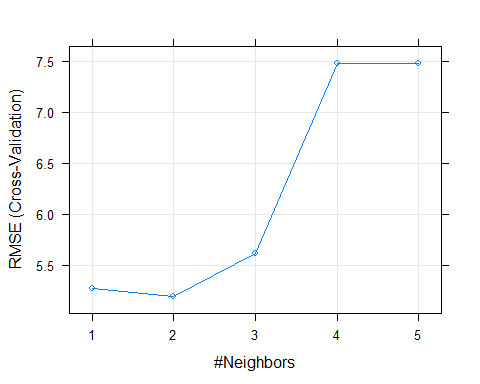
First we will Build our model for the Unemployment rate, attempting to predict the unemployment rate at the FTYIP or 2023.

### Building the Model

The below code builds the model using the four methods below.

### model  
## 10 Fold CV  
Control <- trainControl(method = "cv", number = 15, p = 0.9)  
  
##Linear Regression  
fit\_lm\_Unemployment <- train(Unemployment ~ ., data = Unemployment,   
 method = "lm", trControl = Control)  
pred\_lm\_Unemployment <- predict(fit\_lm\_Unemployment, newdata = Unemployment\_test)  
RMSE\_lm <- sqrt(mean((pred\_lm\_Unemployment - Unemployment\_test$Unemployment)^2))  
  
## K-Nearest Neighbors  
tune = data.frame(k = 1:5)  
fit\_knn\_Unemployment <- train(Unemployment ~ ., data = Unemployment,   
 method = "knn", tuneGrid = tune, trControl = Control)  
  
pred\_knn\_Unemployment <- predict(fit\_knn\_Unemployment, newdata = Unemployment\_test)  
RMSE\_knn <- sqrt(mean((pred\_knn\_Unemployment - Unemployment\_test$Unemployment)^2))  
  
## Extreme Gradient Boosting  
tune <- expand.grid(nrounds = seq(200,300,50), lambda = seq(1,3,0.5), alpha = c(0,1e-04), eta = 0.3)  
fit\_xgb\_Unemployment <- train(Unemployment ~ ., data = Unemployment,   
 method = "xgbLinear", tuneGrid = tune, trControl = Control)  
pred\_xgb\_Unemployment <- predict(fit\_xgb\_Unemployment, newdata = Unemployment\_test)  
RMSE\_xgb <- sqrt(mean((pred\_xgb\_Unemployment - Unemployment\_test$Unemployment)^2))  
  
## Random Forest  
tune = data.frame(mtry = seq(40,50,5))  
fit\_rf\_Unemployment<- train(Unemployment ~ ., data = Unemployment,   
 method = "rf", tuneGrid = tune, trControl = Control)  
pred\_rf\_Unemployment <- predict(fit\_rf\_Unemployment, newdata = Unemployment\_test)  
RMSE\_rf <- sqrt(mean((pred\_rf\_Unemployment - Unemployment\_test$Unemployment)^2))  
  
## Storing the Results  
Results\_Unemployment <- data\_frame(method = c("lm","knn","xgb","rf"),   
 RMSE = c(RMSE\_lm,RMSE\_knn,RMSE\_xgb,RMSE\_rf))  
  
fit\_Unemployment <- list(lm = fit\_lm\_Unemployment, knn = fit\_knn\_Unemployment,   
 xgb = fit\_xgb\_Unemployment, rf = fit\_rf\_Unemployment)  
Best\_fit\_Unemployment <- fit\_Unemployment[[Results\_Unemployment$method[which.min(Results\_Unemployment$RMSE)]]]

Plotting our Parameters against the RMSE allows us to ensure that our model is optimized, below is an example using the k-Nearest Neighbors Model:

 The results from each of our Models can be seen below:

|  |  |
| --- | --- |
| Method | RMSE |
| Linear Model | 2.340 |
| k-Nearest Neighbors | 3.824 |
| Gradient Boosting | 0.423 |
| Random Forest | 0.228 |

Thus the best performing methodology is:

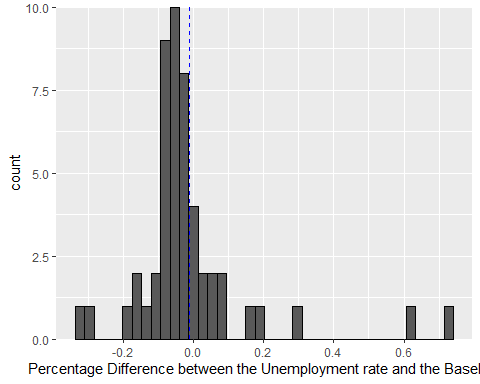
|  |  |
| --- | --- |
| Method | RMSE |
| Random Forest | 0.228 |

### FTYIP Prediction

Now we will use our model to predict the Unemployment rate for the year 2023. The Prediction can be generated with the following code:

Unemployment\_2023 <- Unemployment\_test %>%   
 mutate(Year = 2023) %>% select(-Unemployment)  
  
Unemployment\_2023 <- Unemployment\_2023 %>%   
 mutate(Unemployment = predict(Best\_fit\_Unemployment, newdata = Unemployment\_2023)) %>%   
 left\_join(Baseline\_Unemployment, by = "Country\_Name") %>%   
 mutate(Difference\_to\_Baseline = (Unemployment - Baseline)/Baseline)  
  
mu\_Unemployment\_2023 <- Unemployment\_2023 %>% summarize(mean(Difference\_to\_Baseline)) %>% pull()

Visualizing the distribution of the Percentage difference between the Prediction and the Baseline data gives us an indication of how progress towards the target is distributed:

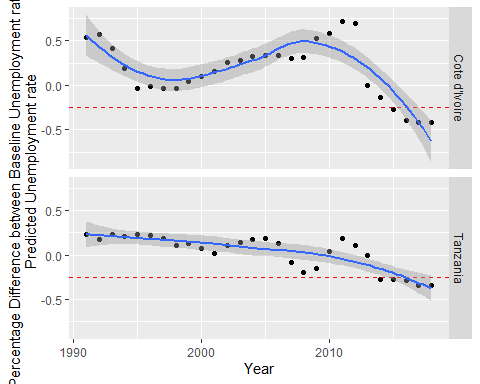


We can see a number of outliers on the high end; otherwise, the result is tightly distributed with most Member States managing to reduce unemployment performing at or better than the Continent Wide average of -1.201%. This average is much lower than the target set in the FTYIP and is such a cause for concern.

Unfortunately, our predictions indicate that only two States succeed in achieving the 2023 target of 25% reduction in the Unemployment rate:

|  |  |  |
| --- | --- | --- |
| Country | Unemployment rate | Percentage Difference to Baseline |
| Cote d’Ivoire | 2.90 | -31.9 |
| Tanzania | 2.03 | -30.6 |

Visualizing the Unemployment data for these states shows us that the prediction is consistent with the trend they have been exhibiting:



## Annual GDP Growth

We will repeat the same steps to model the Annual GDP Growth.

### Building the Model

Once again we use 10 fold cross Validation and the same four modelling methodologies as before.

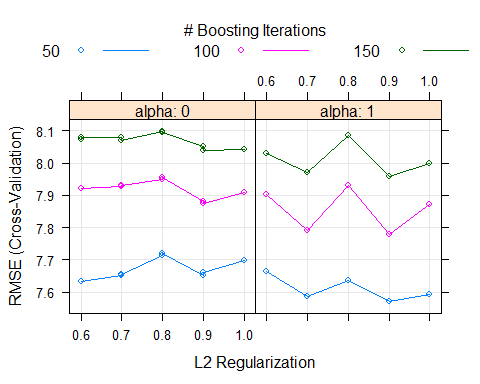
Below we can see the results for our Model

|  |  |
| --- | --- |
| Method | RMSE |
| Linear Regression | 3.81 |
| k-Nearest Neighbors | 3.50 |
| Gradient Boosting | 2.83 |
| Random Forest | 3.01 |

Thus our best performing model is:

|  |  |
| --- | --- |
| Method | RMSE |
| Gradient Boosting | 2.83 |

We can confirm the optimization of our model parameters with a plot like the below for the Gradient Boosting method:



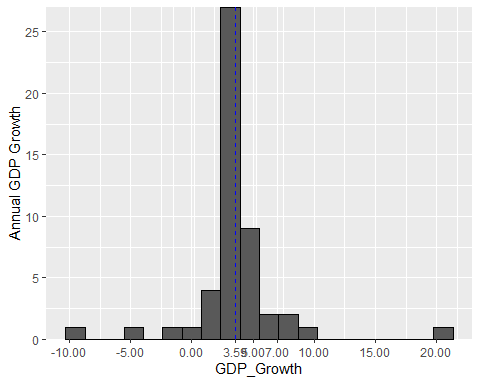
### FTYIP Prediction

Next we make predictions using our best performing model

We take a look at a sample of our predictions:

|  |  |  |
| --- | --- | --- |
| Country | Year | Annual GDP Growth |
| Gabon | 2023 | 2.71 |
| Congo, Dem. Rep. | 2023 | 3.95 |
| Togo | 2023 | 3.76 |
| Benin | 2023 | 3.52 |

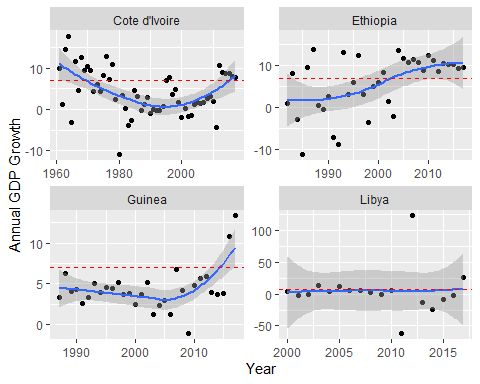
A more detailed a look at the distribution of our predictions show a large number of the Member states perform positively at or above the AU average of 3.595% even if they do not achieve the stated target of 7% Annual GDP Growth.



Below we list the states that do meet our growth target:

|  |  |  |
| --- | --- | --- |
| Country | Year | Annual GDP Growth |
| Cote d’Ivoire | 2023 | 7.48 |
| Ethiopia | 2023 | 8.90 |
| Guinea | 2023 | 8.57 |
| Libya | 2023 | 20.60 |

looking at the data for these countries we see that the apparent trend supports these predictions. The exception is Libya that has had a large variation around our growth target indicating a weakness in our model to account for them.



## Modeling the Combined Data Sets

### Wrangling the Data set

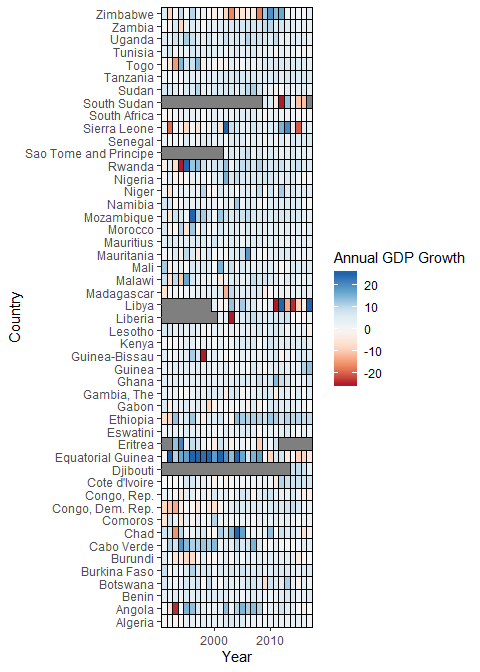
Finally we combine the data sets for both indicators and redo our predictions for both indicators using the combined data set.

We achieve this alongside preparing our data sets using the following code:

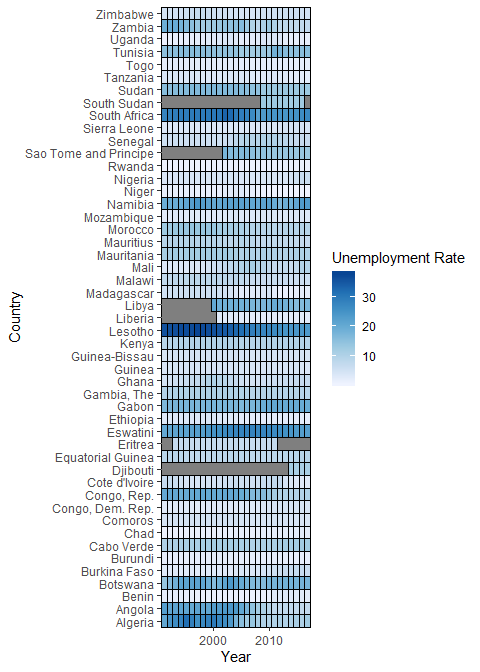
#Preparing the Data  
Model <- GDP\_Growth %>% full\_join(Unemployment)  
  
#Removing years and countries with no data  
No\_Data <- Model %>%  
 group\_by(Year) %>% summarize(No\_Data = sum(!is.na(Unemployment)))   
Model <- Model %>% left\_join(No\_Data, by = "Year") %>% filter(No\_Data != 0) %>% select(-No\_Data)   
  
No\_Data <- Model %>% group\_by(Country\_Name) %>% summarize(No\_Data = sum(!is.na(Unemployment)))   
Model <- Model %>% left\_join(No\_Data, by = "Country\_Name") %>% filter(No\_Data != 0) %>% select(-No\_Data)   
  
No\_Data <- Model %>%  
 group\_by(Year) %>% summarize(No\_Data = sum(!is.na(GDP\_Growth)))   
Model <- Model %>% left\_join(No\_Data, by = "Year") %>% filter(No\_Data != 0) %>% select(-No\_Data)   
  
No\_Data <- Model %>% group\_by(Country\_Name) %>% summarize(No\_Data = sum(!is.na(GDP\_Growth)))   
Model <- Model %>% left\_join(No\_Data, by = "Country\_Name") %>% filter(No\_Data != 0) %>% select(-No\_Data)   
  
#Removing NA's  
Model <- Model %>% filter(!is.na(GDP\_Growth))  
  
#Setting Test Data  
Unemployment\_2018 <- Unemployment %>% filter(Year == 2018)  
Model\_test <- GDP\_Growth\_test %>% left\_join(Unemployment\_2018)  
  
Model\_test <- Model\_test %>% filter(!is.na(GDP\_Growth),!is.na(Unemployment))

### Visualizations

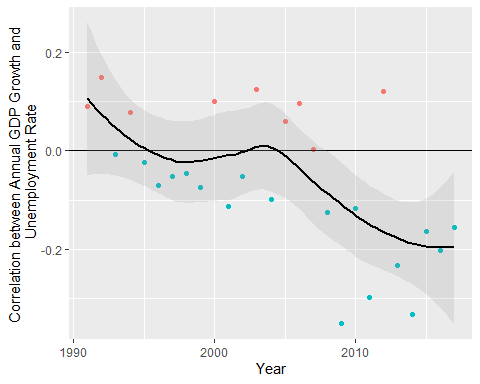
We view the new Data set first by GDP growth rate:



Then by Unemployment rate:

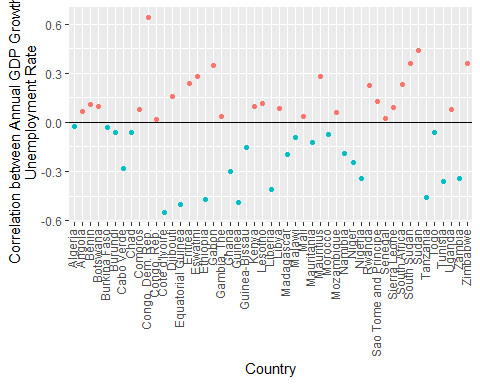


Next we attempt to visualize the correlation between the tow indicators across the years and for each member state. Ideally we would want to see a high negative correlation as we want to reduce Unemployment alongside a high GDP growth rate.



The plot of correlation against year shows us that the correlation between the two indicators is not a constant but rather follows a downward trend implying that as time goes on they are becoming better indicators of each other if only slightly.

Making a similar plot on a Country by Country basis helps explain the low continent wide correlation. There is a large variation across countries that then reduces the Continent wide average.



### New Unemployment Model

We rebuild our models using the new data sets and compare our results with our previous unemployment model:

|  |  |  |
| --- | --- | --- |
| Method | Original data set RMSE | Combined data set RMSE |
| Linear Regression | 2.340 | 2.43 |
| k-Nearest Neighbors | 3.824 | 5.28 |
| Gradient Boosting | 0.423 | 0.88 |
| Random Forest | 0.228 | 1.94 |

Our Various Model methodologies are under-performing once we add GDP growth as a predictor so we discontinue this line of inquiry.

### New Annual GDP growth model

We repeat the previous exercise, this time attempting to predict Annual GDP Growth with Unemployment as a predictor. The results are:

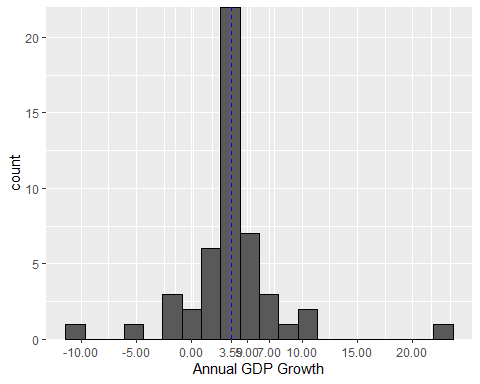
|  |  |  |
| --- | --- | --- |
| Method | Original Data set RMSE | Combined data set RMSE |
| Linear Regression | 3.81 | 4.41 |
| k-Nearest Neighbors | 3.50 | 3.30 |
| Gradient Boosting | 2.83 | 1.91 |
| Random Forest | 3.01 | 3.20 |

This time we find that adding Unemployment as a predictor of Annual GDP\_Growth leads to an improvement in our model. Therefore, we use our best performing model to make the predictions.

The below code predicts Unemployment in 2023 and then uses that to predict the Annual GDP Growth of 2023:

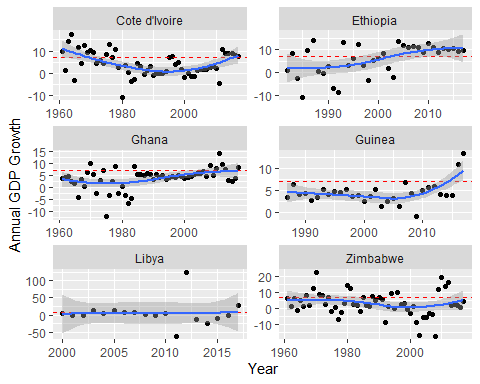
Model\_GDP\_Growth\_2023 <- Model\_test %>% mutate(Year = 2023) %>% select(Country\_Name,Year)  
  
Model\_GDP\_Growth\_2023 <- Model\_GDP\_Growth\_2023 %>%   
 mutate(Unemployment = predict(Best\_fit\_Unemployment, newdata = .)) %>%  
 mutate(GDP\_Growth = predict(Best\_fit\_Model\_GDP\_Growth, newdata = .))

A histogram of our results shows a similar trend as before although the data seems more tightly clustered with fewer outliers:



Finally by looking at the countries that achieve the targets we find that the trends do support our prediction. In spite of Cote d’Ivoire no longer predicted to succeed we have one more country predicted to succeed than in our last model.

Once again we caution at the inclusion of Libya considering the massive variance it is exhibiting in the data set.



# Conclusion

In the Course of the report we set out to assess the progress of AU member states in achieving specific targets set in the First Ten Year Implementation Plan; as well as to predict if those targets would be achieved by 2023.

## Technical findings

On the more technical side we found two important if predictable findings

1. That different indicators are better modeled with different modelling algorithms
2. That because one indicator can act as a predictor for another does not mean the opposite is true

## Analysis Results and Recommendations

In terms of the performance of member states we find that many if not most member states are performing positively in our chosen indicators; however, they are falling short of the targets with only a handful predicted to achieve it by 2023.

This can partly be blamed at how the targets are set. For example, we alluded that a more important indicator for success might be how often a target is reached rather than simply whether or not it is achieved. This would minimize the effect of large variations in our data set such as we saw with the Annual GDP Growth in Libya.

Continent Wide performance indicators; whether measured as the Mean or Median, hide the Variation within the continent with many Member States vastly over or under-performing leading to skewed distributions.

Stratifying Member states based on commonalities may help to isolate the causes of these variations guiding policy decision.