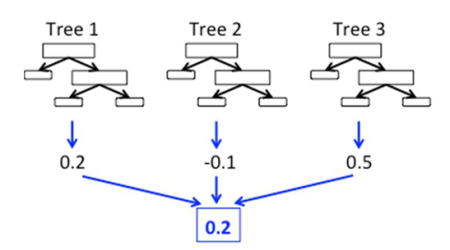
**Random Forest**; for **regression**, constructs multiple decision trees and, inferring the average estimation result of each decision tree. This algorithm is more robust to overfitting than the classical decision trees.



The random forest algorithms average these results; that is, it reduces the variation by training the different parts of the train set. This increases the performance of the final model, although this situation creates a small increase in bias.

The random forest uses bootstrap aggregating(**bagging**) algortihms. We would take for training sample, ***X* = *x1*, …, *xn*** and, ***Y = y1, …, yn*** for the outputs. The bagging process repeated **B times** with selecting a random sample by changing the training set and, tries to fit the relevant tree algorithms to the samples. This fitting function is denoted ***fb*** in the below formula.

The unseen values, ***x’***, would be fitted by ***fb*** and then all the results of **B** individuals trees are averaged. Random forest selects explanatory variables at each variable split in the learning process, which means it trains a random subset of the feature instead of all sets of features. This is called **feature bagging**. This process reduces the correlation between trees; because the strong predictors could be selected by many of the trees, and it could make them correlated.

The **permutation feature importance** method would be used to determine the effects of the variables in the random forest model. This method calculates the increase in the prediction error(**MSE**) after permuting the feature values. If the permuting wouldn’t change the model error, the related feature is considered unimportant. We will use the ***varImp*** function to calculate variable importance. It has a default parameter, ***scale=TRUE***, which scales the measures of importance up to 100.

The data we are going to use can be download [here](https://github.com/mesdi/CPI/blob/main/cpi.xlsx?raw=true). The variables we will examine:

* ***cpi***: The annual consumer price index. It is also called the inflation rate. This is our target variable.
* ***funding\_rate***: The one-week repo rate, which determined by the Turkish Central Bank. It is also called the political rate.
* ***exchange\_rate***: The currency exchange rates between Turkish Liras and American dollars.
* ***CDS***: The credit defaults swap. It kind of measures the investment risk of a country or company. It is mostly affected by foreign policy developments in Turkey. Although the most popular CDS is 5year, I will take CDS of 1 year USD in this article.

library(readxl)

library(dplyr)

library(ggplot2)

library(randomForest)

library(varImp)

#Create the dataframe

df <- read\_excel("cpi.xlsx")

#Random Forest Modelling

model <- randomForest(CPI ~ funding\_rate+exchange\_rate,

data = df, importance=TRUE)

#Conditional=True, adjusts for correlations between predictors.

i\_scores <- varImp(model, conditional=TRUE)

#Gathering rownames in 'var' and converting it to the factor

#to provide 'fill' parameter for the bar chart.

i\_scores <- i\_scores %>% tibble::rownames\_to\_column("var")

i\_scores$var<- i\_scores$var %>% as.factor()

#Plotting the bar and polar charts for comparing variables

i\_bar <- ggplot(data = i\_scores) +

geom\_bar(

stat = "identity",#it leaves the data without count and bin

mapping = aes(x = var, y=Overall, fill = var),

show.legend = FALSE,

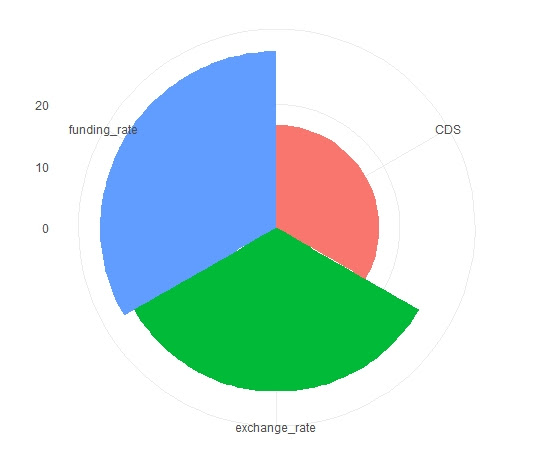
width = 1

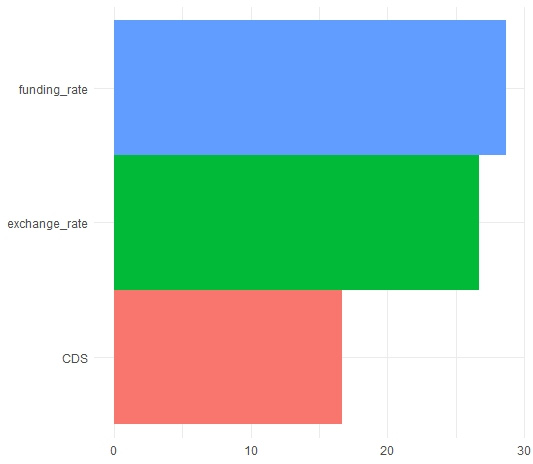
) +

labs(x = NULL, y = NULL)

i\_bar + coord\_polar() + theme\_minimal()

i\_bar + coord\_flip() + theme\_minimal()





When we examine the charts above, we can clearly see that the funding and exchange rates have similar effects on the model and, CDS has significant importance in the behavior of the CPI. So the theory of the president seems to fall short of what he claims.