CBC\_analysis\_DPI\_0

Rajnish Kumar

2024-02-22

## R Markdown

Note: This file is only for PCA analysis, and ML implementation with all the decided models on DPI ==0. Note : Make sure to drop the columns closely associated with NT Grade, and ICANs etc. in the factor analysis as well as subsequent prediction.

#load libraries  
library(plyr)  
library(tidyverse) #general data cleaning and organization

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ purrr::compact() masks plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::desc() masks plyr::desc()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ dplyr::rename() masks plyr::rename()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(Boruta) #Recursive feature elimination  
library(caret) # Logistic regression

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(MLeval)  
library(psych)

##   
## Attaching package: 'psych'  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

# Loading package   
library(e1071)   
library(caTools)   
library(class)  
library(ggcorrplot)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

## Including Plots

Read all the files:

# Day 0 is the baseline. Let us apply Boruta on DPI==0. We will later do it for multiple number of days.  
  
CBC\_DPI\_all <- read.csv("C:/Users/rajnishk/Dropbox (University of Michigan)/2023-09-25 CBC CART for Benjie and Rajnish from Mary/Only\_UM\_LTFU\_CBC\_Data\_with\_matched\_CRP\_Ferritin\_LOD\_0.csv")

Change the elements of CRS\_Grade to a numeric value for analysis later.

CBC\_DPI\_all$CRS\_Grade <-as.numeric(as.character(CBC\_DPI\_all$CRS\_Grade))

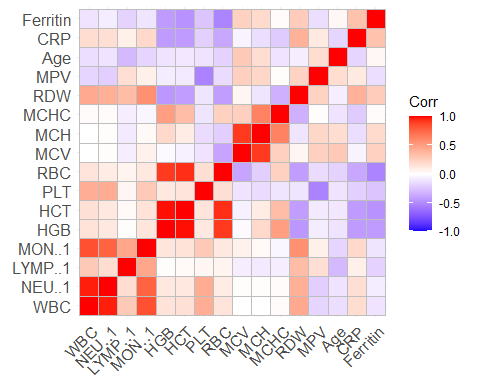
## Warning: NAs introduced by coercion

CBC\_DPI\_0\_only<- CBC\_DPI\_all%>%filter(DPI==0) # This is the dataset with all the values for DPI ==0

set.seed(1) # Will still have to check where this will be used later.   
#Boruta for DPI ==0 (baseline data)  
# Will drop a few irrelevant columns, e.g. StudyID etc. before applying feature importance.   
  
  
# drop the column Placeholder\_DPI because all the values are NA's in this case. Will see if there are real values of this variable for different values of DPIs. Same goes for During Chemo and After Chemo.  
CBC\_DPI\_0\_only <-subset(CBC\_DPI\_0\_only,select = -c(StudyID,CRS\_Incidence,CRS\_Grade, Max\_CRS,NEU.,LYMP.,MON. , DP\_CRS\_Onset,Max\_NT, NT\_Grade, Placeholder\_DPI, During.Chemo., After.Chemo, DP\_NT\_Onset)) # These variables were also dropped in agreement with the group.   
  
  
CBC\_DPI\_0\_only <- na.omit(CBC\_DPI\_0\_only) # After dropping rows with NA's in them. # Drop the NAs. You can check the values before and after this line by printing the variables.

Apply PCAs with different rotations (VARIMAX, QUARTIMAX etc.), and see the variances.

set.seed(1) # For reproducibility.   
#Apply PCA on data from DPI == 0. We will chose the dataframe that has all the irrelevant features dropped already (for implementation of Boruta).   
#Normalize all the features before implementing PCA.   
  
# Select all numeric columns that have to be normalized  
CBC\_DPI\_0\_only <-subset(CBC\_DPI\_0\_only, select = -c(DPI)) # Because normalizing DPI = 0 introduced all the NAN values into this column, so I had to remove it, and to maintain uniformity, I will have to remove DPI values for other datasets, CBC\_DPI\_0\_to\_1 and so on as well before I normalize them to apply PCA.   
CBC\_DPI\_0\_numeric <-CBC\_DPI\_0\_only%>% select\_if(~ is.numeric(.))  
# Normalize all numeric columns  
CBC\_DPI\_0\_normalized <- CBC\_DPI\_0\_numeric %>% mutate\_all(scale)  
  
corr\_matrix <- cor(CBC\_DPI\_0\_normalized)  
ggcorrplot(corr\_matrix) # visualize the correlation between different variables.



# Merge back the non-numeric columns  
CBC\_DPI\_0\_final <-bind\_cols(CBC\_DPI\_0\_only %>% select\_if( ~ !is.numeric(.)),CBC\_DPI\_0\_normalized)  
  
# Perform PCA on this data frame.   
PCA\_DPI\_0 <- principal(CBC\_DPI\_0\_normalized, nfactors = 15, rotate = "none")  
#PCA\_DPI\_0 <- principal(CBC\_DPI\_0\_normalized, nfactors = 15, rotate = "varimax")  
print(PCA\_DPI\_0)

## Principal Components Analysis  
## Call: principal(r = CBC\_DPI\_0\_normalized, nfactors = 15, rotate = "none")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11  
## WBC 0.29 0.88 0.23 -0.16 0.06 0.14 -0.05 -0.10 0.03 -0.09 -0.10  
## NEU..1 0.26 0.84 0.23 -0.24 0.07 0.19 -0.07 -0.14 -0.01 -0.11 -0.13  
## LYMP..1 0.08 0.47 0.02 0.51 -0.08 -0.56 0.22 0.06 0.36 -0.03 -0.04  
## MON..1 0.21 0.86 0.29 0.15 0.07 0.03 0.05 0.04 -0.04 0.11 -0.02  
## HGB 0.93 -0.16 0.20 0.12 -0.08 0.15 0.08 0.12 0.00 0.03 0.05  
## HCT 0.93 -0.12 0.11 0.14 0.00 0.17 0.10 0.21 -0.02 0.01 0.02  
## PLT 0.37 0.38 -0.19 -0.60 0.19 -0.37 -0.15 0.09 -0.02 0.32 0.07  
## RBC 0.93 -0.07 -0.19 0.18 -0.04 0.19 0.10 0.09 0.01 0.04 0.01  
## MCV -0.21 -0.14 0.87 -0.13 0.15 -0.16 -0.04 0.31 -0.08 -0.09 -0.01  
## MCH 0.00 -0.24 0.94 -0.13 -0.06 -0.14 -0.07 0.06 -0.01 -0.04 0.04  
## MCHC 0.39 -0.28 0.58 -0.05 -0.45 -0.02 -0.07 -0.43 0.15 0.09 0.11  
## RDW -0.46 0.70 0.12 0.25 0.05 0.07 0.27 -0.05 -0.22 0.00 0.31  
## MPV -0.24 -0.11 0.33 0.77 0.25 0.07 -0.19 -0.10 -0.14 0.25 -0.14  
## Age -0.20 -0.28 0.21 -0.15 0.75 0.31 0.15 -0.07 0.36 0.05 0.07  
## CRP -0.50 0.39 -0.05 0.14 -0.37 0.35 -0.39 0.28 0.27 0.06 0.10  
## Ferritin -0.60 -0.02 0.21 -0.31 -0.42 0.22 0.46 0.09 0.03 0.20 -0.16  
## PC12 PC13 PC14 PC15 h2 u2 com  
## WBC 0.05 0.09 -0.02 0.00 1 1.8e-09 1.7  
## NEU..1 0.12 -0.08 0.02 0.00 1 2.3e-09 1.9  
## LYMP..1 0.06 -0.01 0.00 0.00 1 1.9e-09 4.2  
## MON..1 -0.29 -0.02 0.00 0.00 1 3.5e-09 1.8  
## HGB 0.04 -0.02 -0.04 0.00 1 1.2e-04 1.3  
## HCT 0.04 -0.01 -0.02 0.00 1 1.7e-04 1.3  
## PLT 0.07 0.00 0.00 0.00 1 1.5e-09 4.9  
## RBC 0.01 0.03 0.05 0.00 1 1.1e-05 1.3  
## MCV 0.01 0.01 0.01 0.02 1 2.8e-06 1.7  
## MCH 0.01 0.01 0.01 -0.02 1 5.0e-07 1.2  
## MCHC -0.02 0.00 0.00 0.01 1 2.7e-06 4.7  
## RDW 0.09 0.00 0.00 0.00 1 2.6e-09 3.4  
## MPV 0.08 0.00 0.00 0.00 1 2.4e-10 2.6  
## Age -0.01 0.00 0.00 0.00 1 5.8e-09 2.9  
## CRP 0.02 0.00 0.00 0.00 1 3.1e-09 6.5  
## Ferritin 0.04 0.00 0.00 0.00 1 9.1e-10 4.6  
##   
## PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9 PC10 PC11  
## SS loadings 4.04 3.53 2.50 1.63 1.22 0.91 0.61 0.50 0.43 0.26 0.20  
## Proportion Var 0.25 0.22 0.16 0.10 0.08 0.06 0.04 0.03 0.03 0.02 0.01  
## Cumulative Var 0.25 0.47 0.63 0.73 0.81 0.86 0.90 0.93 0.96 0.98 0.99  
## Proportion Explained 0.25 0.22 0.16 0.10 0.08 0.06 0.04 0.03 0.03 0.02 0.01  
## Cumulative Proportion 0.25 0.47 0.63 0.73 0.81 0.86 0.90 0.93 0.96 0.98 0.99  
## PC12 PC13 PC14 PC15  
## SS loadings 0.13 0.02 0.01 0  
## Proportion Var 0.01 0.00 0.00 0  
## Cumulative Var 1.00 1.00 1.00 1  
## Proportion Explained 0.01 0.00 0.00 0  
## Cumulative Proportion 1.00 1.00 1.00 1  
##   
## Mean item complexity = 2.9  
## Test of the hypothesis that 15 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0   
## with the empirical chi square 0 with prob < NA   
##   
## Fit based upon off diagonal values = 1

summary(PCA\_DPI\_0)

##   
## Factor analysis with Call: principal(r = CBC\_DPI\_0\_normalized, nfactors = 15, rotate = "none")  
##   
## Test of the hypothesis that 15 factors are sufficient.  
## The degrees of freedom for the model is -15 and the objective function was 0.39   
## The number of observations was 75 with Chi Square = 22.64 with prob < NA   
##   
## The root mean square of the residuals (RMSA) is 0

Next section has Biplots, Scree plots and correlation plots relating features to principal components.

<https://www.datacamp.com/tutorial/pca-analysis-r>

data.pca <- princomp(corr\_matrix)  
summary(data.pca)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 0.9910653 0.8223532 0.44202724 0.38960113 0.30472295  
## Proportion of Variance 0.4468774 0.3076809 0.08889593 0.06905963 0.04224683  
## Cumulative Proportion 0.4468774 0.7545583 0.84345424 0.91251386 0.95476069  
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Standard deviation 0.21861414 0.14976759 0.122538494 0.09817442 0.052430061  
## Proportion of Variance 0.02174404 0.01020515 0.006831702 0.00438511 0.001250675  
## Cumulative Proportion 0.97650474 0.98670989 0.993541590 0.99792670 0.999177376  
## Comp.11 Comp.12 Comp.13 Comp.14  
## Standard deviation 0.040843772 1.103242e-02 4.233209e-03 4.706592e-04  
## Proportion of Variance 0.000758989 5.537651e-05 8.153108e-06 1.007852e-07  
## Cumulative Proportion 0.999936365 9.999917e-01 9.999999e-01 1.000000e+00  
## Comp.15 Comp.16  
## Standard deviation 1.022915e-04 0  
## Proportion of Variance 4.760612e-09 0  
## Cumulative Proportion 1.000000e+00 1

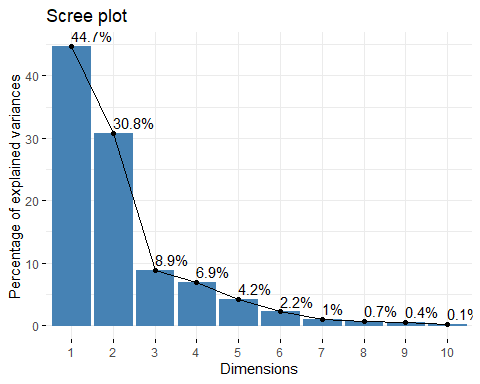
Loadings of each principal component

data.pca$loadings[, 1:2]

## Comp.1 Comp.2  
## WBC 0.010993885 0.40984054  
## NEU..1 0.001774644 0.39167397  
## LYMP..1 -0.026922161 0.22936636  
## MON..1 -0.031033076 0.37638461  
## HGB 0.453290691 -0.03545012  
## HCT 0.453443870 0.00192495  
## PLT 0.141069680 0.27384285  
## RBC 0.462543747 0.09067225  
## MCV -0.128379198 -0.27515056  
## MCH -0.020414806 -0.31734925  
## MCHC 0.192949218 -0.21740556  
## RDW -0.320820799 0.26087617  
## MPV -0.124927174 -0.16459669  
## Age -0.073991184 -0.21617344  
## CRP -0.293770479 0.14071783  
## Ferritin -0.298028534 -0.12046449

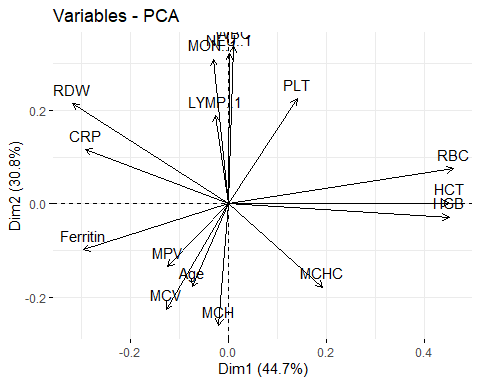
Visualization of the principal components Scree Plots:

fviz\_eig(data.pca, addlabels = TRUE)

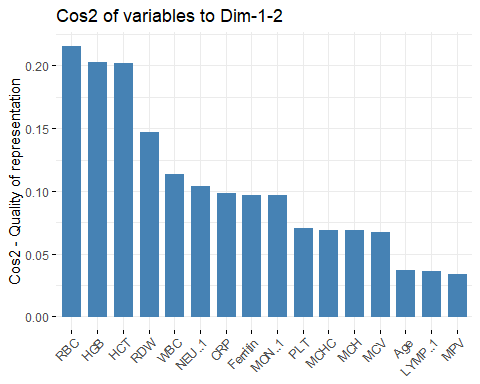


Biplot of the attributes:

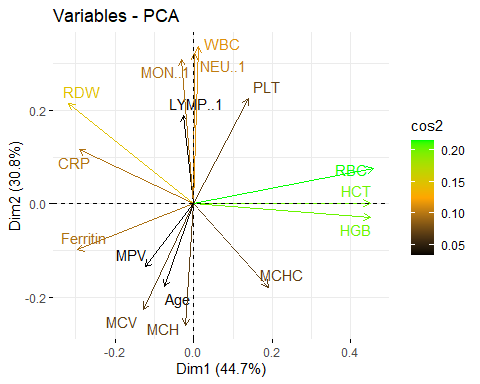
# Graph of the variables  
fviz\_pca\_var(data.pca, col.var = "black")

 Contribution of each variable:

fviz\_cos2(data.pca, choice = "var", axes = 1:2)

 Biplot combined with cos2

fviz\_pca\_var(data.pca, col.var = "cos2",  
 gradient.cols = c("black", "orange", "green"),  
 repel = TRUE)

 For now I am arbitrarily selecting first 5 principal components for the ML prediction. The values used are already normalized. Models of choice were KNN, SVM, BLR(?), ……….

This means, we need to project the test data set (from John Hopikins) on the same principal components. Later on we will create PCs based on combined dataset of Umich and JH for the purposes of generalization. FIND THE REFERENCE FOR THIS IDEA (OF GENERALIZATION) as well.