NBA MVP 2019-2020 PREDICTION.R

Hasan Hindi

2020-06-07

rm(list = ls())  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(odbc)

## Warning: package 'odbc' was built under R version 3.6.3

library(DBI)

## Warning: package 'DBI' was built under R version 3.6.3

con <- dbConnect(odbc(),  
 Driver = "SQL Server",  
 Server = "DESKTOP-K12S3HQ",  
 Database = "NBA MVP"  
)  
  
years <- seq(2012,2019,1)  
id = seq(1,length(years)-1)  
  
data = data.frame()  
for (i in id){  
 data <- rbind(data,dbReadTable(con,paste0("vv\_",years[i],"\_",years[i+1],"\_additional\_stats")))  
}  
num\_mvp\_candidates <- c()  
for (i in id){  
 num\_mvp\_candidates <- append(num\_mvp\_candidates,nrow(dbReadTable(con,paste0("vv\_",years[i],"\_",years[i+1],"\_additional\_stats"))))  
}  
  
data$begin <- NA  
data$end <- NA  
  
begin <- c()  
for (i in id){  
 begin <- append(begin,rep(years[i],num\_mvp\_candidates[i]))  
}  
data$begin <- begin  
  
end <- c()  
for (i in id){  
 end <- append(end,rep(years[i+1],num\_mvp\_candidates[i]))  
}  
data$end <- end  
  
data$begin <- as.factor(data$begin)  
data$end <- as.factor(data$end)  
data$Player\_season <- paste0(data$Player\_Name,"\_",data$begin,"\_",data$end)  
data <- data %>% filter(Tm != "TOT")

data2 = data.frame()  
for (i in id){  
 data2 <- rbind(data2,dbReadTable(con,paste0("vv\_MVP\_VOTING\_",years[i],"\_",years[i+1])))  
}  
  
num\_mvp\_candidates <- c()  
for (i in id){  
 num\_mvp\_candidates <- append(num\_mvp\_candidates,nrow(dbReadTable(con,paste0("vv\_MVP\_VOTING\_",years[i],"\_",years[i+1]))))  
}  
  
  
data2$begin <- NA  
data2$end <- NA  
  
begin <- c()  
for (i in id){  
 begin <- append(begin,rep(years[i],num\_mvp\_candidates[i]))  
}  
data2$begin <- begin  
  
end <- c()  
for (i in id){  
 end <- append(end,rep(years[i+1],num\_mvp\_candidates[i]))  
}  
data2$end <- end  
data2$X2P. <- NULL  
data2$begin <- as.factor(data2$begin)  
data2$end <- as.factor(data2$end)  
  
  
data2$team\_wins <- NA  
data2$team\_losses <- NA  
for (i in 1:nrow(data2)){  
 data2[i,"team\_wins"] <- as.numeric(unlist(sub("-.\*","",data2[i,"Record"])))  
 data2[i,"team\_losses"] <- as.numeric(unlist(sub(".\*-","",data2[i,"Record"])))  
}  
data2$team\_win\_percentage <- data2$team\_wins/(data2$team\_losses+data2$team\_wins)  
data2$percent\_games\_played <- data2$G/(data2$team\_losses+data2$team\_wins)  
data2$Player\_season <- paste0(data2$Player\_Name,"\_",data2$begin,"\_",data2$end)  
  
output <- data %>% group\_by(Player\_season) %>% count() %>% arrange(desc(n)) %>% filter(n == 1)  
output2 <- data2 %>% group\_by(Player\_season) %>% count() %>% arrange(desc(n)) %>% filter(n == 1)  
data <- data[which(data$Player\_season %in% output$Player\_season),]  
data2 <- data2[which(data2$Player\_season %in% output2$Player\_season),]  
  
data3 <- merge(data[,c("TRB.PER.36","PTS.PER.36","AST.PER.36","ORtg","DRtg","NetRtg","Player\_season")],data2,by = "Player\_season")  
data3$Share <- ifelse(is.na(data3$Share),0,data3$Share)  
data3 <- data3 %>% filter(PTS >= 20 & MP >= 30 & percent\_games\_played >= 0.75)  
data3 <- na.omit(data3)  
  
additional\_2019\_2020 <- dbReadTable(con,"vv\_2019\_2020\_additional\_stats")  
additional\_2019\_2020 <- additional\_2019\_2020 %>% filter(Tm != "TOT")  
mvp\_2019\_2020 <- dbReadTable(con,"vv\_MVP\_VOTING\_2019\_2020")  
  
data\_2019\_2020 <- merge(additional\_2019\_2020[,c("Player\_team","Tm","TRB.PER.36","AST.PER.36","PTS.PER.36","ORtg","DRtg","NetRtg")],  
 mvp\_2019\_2020[,c("Player\_team","Record","PTS","AST","BLK","FT.","G","MP","TRB","PER","TS.","USG.","VORP","WS.48","BPM")],by = "Player\_team")

data\_2019\_2020$team\_wins <- NA  
data\_2019\_2020$team\_losses <- NA  
for (i in 1:nrow(data\_2019\_2020)){  
 data\_2019\_2020[i,"team\_wins"] <- as.numeric(unlist(sub("-.\*","",data\_2019\_2020[i,"Record"])))  
 data\_2019\_2020[i,"team\_losses"] <- as.numeric(unlist(sub(".\*-","",data\_2019\_2020[i,"Record"])))  
}  
data\_2019\_2020$team\_win\_percentage <- data\_2019\_2020$team\_wins/(data\_2019\_2020$team\_losses+data\_2019\_2020$team\_wins)  
data\_2019\_2020$percent\_games\_played <- data\_2019\_2020$G/(data\_2019\_2020$team\_losses+data\_2019\_2020$team\_wins)  
  
data3$trifecta <- data3$PTS+data3$TRB+data3$AST  
data\_2019\_2020$trifecta <- data\_2019\_2020$PTS + data\_2019\_2020$TRB+data\_2019\_2020$AST   
  
  
cols\_selected <- c("TRB.PER.36","PTS.PER.36","AST.PER.36","ORtg","DRtg","NetRtg","PTS",  
 "AST","BLK","FT.","TRB","PER","TS.","USG.","VORP","WS.48","BPM",  
 "team\_win\_percentage","percent\_games\_played","MP","trifecta")  
  
test <- na.omit(data\_2019\_2020[,c(cols\_selected,"Player\_team")])  
test <- test %>% filter(PTS >= 20 & MP >= 30 & percent\_games\_played >= 0.75)  
test <- test[,!colnames(test) %in% c("percent\_games\_played","TRB.PER.36","AST.PER.36","ORtg","FT.","BLK","DRtg","MP","TRB","AST","TS.","NetRtg","PTS","USG.")]  
  
x.train <- data3[,c(cols\_selected,"Share")]  
x.train <- x.train[,!colnames(x.train) %in% c("percent\_games\_played","TRB.PER.36","AST.PER.36","ORtg","FT.","BLK","DRtg","MP","TRB","AST","TS.","NetRtg","PTS","USG.")]

library(gbm)

## Loaded gbm 2.1.5

set.seed(123)  
boost.nba <- gbm(Share~.,data=x.train,distribution= "gaussian",n.trees=5000,interaction.depth=3,shrinkage = 0.1,cv.folds = 4)  
mvp\_2019\_2020\_predicted\_results <- predict(boost.nba,test,n.trees = 5000)  
results <- cbind(test,mvp\_2019\_2020\_predicted\_results)  
head(results %>% arrange(desc(mvp\_2019\_2020\_predicted\_results)),10)

## PTS.PER.36 PER VORP WS.48 BPM team\_win\_percentage trifecta  
## 1 34.5 31.6 6.0 0.282 11.5 0.8153846 49.1  
## 2 26.5 26.0 5.7 0.218 8.7 0.7777778 44.2  
## 3 33.7 28.4 6.3 0.245 9.0 0.6250000 48.2  
## 4 28.0 28.2 5.0 0.262 8.5 0.7777778 39.2  
## 5 31.0 27.7 4.7 0.215 8.4 0.5970149 46.7  
## 6 22.5 25.0 5.1 0.209 7.6 0.6615385 37.3  
## 7 30.0 26.7 4.6 0.222 9.0 0.6875000 39.2  
## 8 25.2 21.4 2.7 0.214 4.4 0.8153846 31.4  
## 9 30.2 23.9 3.2 0.134 4.0 0.2985075 43.2  
## 10 24.0 18.7 2.0 0.133 2.2 0.7187500 34.7  
## Player\_team mvp\_2019\_2020\_predicted\_results  
## 1 Giannis Antetokounmpo\_MIL 0.56877562  
## 2 LeBron James\_LAL 0.38868450  
## 3 James Harden\_HOU 0.33276905  
## 4 Anthony Davis\_LAL 0.29670251  
## 5 Luka Doncic\_DAL 0.26183039  
## 6 Nikola Jokic\_DEN 0.17164547  
## 7 Kawhi Leonard\_LAC 0.13152928  
## 8 Khris Middleton\_MIL 0.06006667  
## 9 Trae Young\_ATL 0.05248018  
## 10 Pascal Siakam\_TOR 0.03405378