> setwd("/Users/jeongeunson/Rstudio/팀프로젝트")

> pubg <- read.csv("train\_V2.csv")

> library(ggplot2)

> library(dplyr)

> library(caret)

> library(VIM)

> library(randomForest)

> library(RColorBrewer)

> library(highcharter)

> library(xgboost)

> install.packages('ggplot2')

> install.packages('dplyr')

> install.packages('caret')

> istall.packages('VIM')

> install.packages('VIM')

> install.packages('ramdomForest')

> install.packages('RColorBrewer')

> install.packages('xgboost')

> library(ggplot2)

> library(dplyr)

> library(caret)

> install.packages('lattice')

> library(lattice)

> library(carat)

> library(caret)

> library(VIM)

> install.packages(colorspace)

> install.packages('colorspace')

> install.packages('grid')

> library(VIM)

> library(randomForest)

> library(randomForest)

> library(RColorBrewer)

> library(highcharter)

> library(xgboost)

> library(xgboost)

> dim(pubg) #데이터 갯수

[1] 4446966 29

> colSums(is.na(bubg)) #각 변수마다 결측값의 갯수 구하기 > #winPlacePerc에 1개의 결측값 있는 것 알 수 있다

> Pubg <- na.omit(pubg$winPlacePerc) #winPlacePerc에 결측값이 있는 행을 아예 날려버리기

> colSums(is.na(pubg)) #각 변수마다 결측값의 갯수 구하기

Id groupId matchId assists boosts damageDealt DBNOs headshotKills heals killPlace

0 0 0 0 0 0 0 0 0 0

killPoints kills killStreaks longestKill matchDuration matchType maxPlace numGroups rankPoints revives

0 0 0 0 0 0 0 0 0 0

rideDistance roadKills swimDistance teamKills vehicleDestroys walkDistance weaponsAcquired winPoints winPlacePerc

0 0 0 0 0 0 0 0 1

> Pubg <- na.omit(pubg)

> colSums(is.na(Pubg)) #각 변수마다 결측값의 갯수 구하기

Id groupId matchId assists boosts damageDealt DBNOs headshotKills heals killPlace

0 0 0 0 0 0 0 0 0 0

killPoints kills killStreaks longestKill matchDuration matchType maxPlace numGroups rankPoints revives

0 0 0 0 0 0 0 0 0 0

rideDistance roadKills swimDistance teamKills vehicleDestroys walkDistance weaponsAcquired winPoints winPlacePerc

0 0 0 0 0 0 0 0 0

> dim(Pubg)

[1] 4446965 29

> install.packages('devtools')

> library(devtools)

> str(Pubg) #결측값 행을 날려버린 Pubg라는 데이터셋의 구성요소 보기

'data.frame': 4446965 obs. of 29 variables:

$ Id : Factor w/ 4446966 levels "00000160737ebf",..: 2216110 4151026 532103 1216708 856522 4437931 2598153 852047 458161 3990589 ...

$ groupId : Factor w/ 2026745 levels "00000c08b5be36",..: 613776 827824 843522 1340463 1757850 321871 351332 1858645 570313 933194 ...

$ matchId : Factor w/ 47965 levels "0000a43bce5eec",..: 30087 32753 3144 45262 20533 34913 31456 7685 41676 11378 ...

$ assists : int 0 0 1 0 0 0 0 0 0 0 ...

$ boosts : int 0 0 0 0 0 0 0 0 0 0 ...

$ damageDealt : num 0 91.5 68 32.9 100 ...

$ DBNOs : int 0 0 0 0 0 1 0 0 0 0 ...

$ headshotKills : int 0 0 0 0 0 1 0 0 0 0 ...

$ heals : int 0 0 0 0 0 0 0 0 0 0 ...

$ killPlace : int 60 57 47 75 45 44 96 48 64 74 ...

$ killPoints : int 1241 0 0 0 0 0 1262 1000 0 0 ...

$ kills : int 0 0 0 0 1 1 0 0 0 0 ...

$ killStreaks : int 0 0 0 0 1 1 0 0 0 0 ...

$ longestKill : num 0 0 0 0 58.5 ...

$ matchDuration : int 1306 1777 1318 1436 1424 1395 1316 1967 1375 1930 ...

$ matchType : Factor w/ 16 levels "crashfpp","crashtpp",..: 16 16 3 16 14 16 16 14 15 15 ...

$ maxPlace : int 28 26 50 31 97 28 28 96 28 29 ...

$ numGroups : int 26 25 47 30 95 28 28 92 27 27 ...

$ rankPoints : int -1 1484 1491 1408 1560 1418 -1 -1 1493 1349 ...

$ revives : int 0 0 0 0 0 0 0 0 0 0 ...

$ rideDistance : num 0 0.0045 0 0 0 ...

$ roadKills : int 0 0 0 0 0 0 0 0 0 0 ...

$ swimDistance : num 0 11 0 0 0 ...

$ teamKills : int 0 0 0 0 0 0 0 0 0 0 ...

$ vehicleDestroys: int 0 0 0 0 0 0 0 0 0 0 ...

$ walkDistance : num 244.8 1434 161.8 202.7 49.8 ...

$ weaponsAcquired: int 1 5 2 3 2 1 1 6 4 1 ...

$ winPoints : int 1466 0 0 0 0 0 1497 1500 0 0 ...

$ winPlacePerc : num 0.444 0.64 0.775 0.167 0.188 ...

- attr(\*, "na.action")= 'omit' Named int 2744605

..- attr(\*, "names")= chr "2744605"

> hchart(cor(Pubg[,-c(1,2,3,16)]))

> #숫자형식으로 된 변수들끼리만 상관계수 그림그리기

> #walkDistance, killPlcae, boosts, weaponsAcquired가 winPlacePerc와 높은 관계가 있음을 알 수 있다

> #그럼 그 변수들만 뽑아서 비교해보자

> install.packages('hexbin')

> library(hexbin)

#각 변수 값을 게임시간별로 나누기

actualBoost <- Pubg$boosts/Pubg$matchDuration

actualAssists<- Pubg$assists/Pubg$matchDuration

actualHeals <- Pubg$heals/Pubg$matchDuration

actualWalkDistance <- Pubg$walkDistance/Pubg$matchDuration

actualWeapondAcquired<- Pubg$weapondAcquired/Pubg$matchDuration

> #생존확률과 오래 걷는것과의 상관관계

> ggplot(Pubg,aes(actualWalkDistance,winPlacePerc)) +

+ stat\_binhex() +

+ scale\_fill\_gradient(low = “pink”,high = “black”)

> #killPlace과 오래 걷는것과의 상관관계

> ggplot(Pubg,aes(as.factor(killPlace),winPlacePerc)) +

+ stat\_binhex() +

+ scale\_fill\_gradient(low = “pink”,high = “black”) +

+ labs(title = "Point for killPalce",x = "killPlace")

> #killplace가 낮다면(숫자가 크다면), winPlacePrex의 가장 큰 값이 0.75=winPlaePrex가 낮다

> #killPlace 25이상(숫자로는 25이하)는 보통 0.8이 넘는다 (평균적으로)

> #Boost아이템과과 오래 걷는것과의 상관관계

> ggplot(Pubg,aes(as.factor(actualBoosts),winPlacePerc)) +

+ geom\_boxplot(fill = col) +

> #boosts를 많이 사용했을수록 winPlacePerc가 높다

> ggplot(Pubg,aes(as.factor(actualWeaponsAcquired),winPlacePerc)) +

+ stat\_bin2d() +

+ scale\_fill\_gradient(low = "pink",high = "black") +

> #무기와 winPlacePerc는 딱히 상관관계가 없으며, 보통 10개 미만의 무기를 가진다 #= 한번 좋은 무기를 가지면 굳이 무기를 바꿀 필요가 없다

> ggplot(TOP100,aes(reorder(Id,winPoints),assists,fill = assists)) + geom\_bar(stat = 'identity') + geom\_label(aes(label = assists)) +labs(x = "Id",y = "assists",title = "Top 100 players in the number of assists") + coord\_flip()

> #도움받는게 적다 - 팀이지만 대부분 도움받기보다는 스스로 해결한다 - 좀더 나은 도표로 나타낼 수 있는 방법 찾기

> ggplot(TOP100,aes(reorder(Id,winPoints),DBNOs,fill = DBNOs)) +

+ geom\_bar(stat = 'identity') +

+ geom\_label(aes(label = DBNOs))

> #관계X

> ggplot(TOP100,aes(reorder(Id,winPoints),kills,fill = kills)) +

+ geom\_bar(stat = 'identity') +

+ geom\_label(aes(label = kills)) +

+ labs(x = "Id",y = "kills",title = "Top 100 players in the number of kills")

> #킬수랑 큰 상관관계 없다

> #선형회귀분석 그리기

> Lm1 <- lm(winPlacePerc ~ actualWalkDistance + killPlace + actualBoosts + actualWeaponsAcquired,data = Pubh)

> summary(Lm1)

Call:

Lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts +

weaponsAcquired, data = Pubg)

Residuals:

Min 1Q Median 3Q Max

-3.4352 -0.0929 -0.0083 0.0913 0.7660

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.152e-01 2.719e-04 1527.0 <2e-16 \*\*\*

walkDistance 1.307e-04 8.878e-08 1472.6 <2e-16 \*\*\*

killPlace -3.503e-03 3.454e-06 -1014.3 <2e-16 \*\*\*

boosts 1.591e-02 5.646e-05 281.8 <2e-16 \*\*\*

weaponsAcquired 1.525e-02 3.568e-05 427.4 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1504 on 4446960 degrees of freedom

Multiple R-squared: 0.7607, Adjusted R-squared: 0.7607

F-statistic: 3.534e+06 on 4 and 4446960 DF, p-value: < 2.2e-16

> Lm6 <- lm(winPlacePerc ~ killPlace + actualBoosts +actualDamageDealt + actualHeals, data = Pubg)

> summary(Lm6)

Call:

lm(formula = winPlacePerc ~ killPlace + actualBoosts + actualDamageDealt +

actualHeals, data = Pubg)

Residuals:

Min 1Q Median 3Q Max

-1.99734 -0.10849 -0.00811 0.12519 2.97120

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.568e-01 3.141e-04 2409.7 <2e-16 \*\*\*

killPlace -6.829e-03 4.618e-06 -1478.9 <2e-16 \*\*\*

actualBoosts 9.497e+01 1.096e-01 866.3 <2e-16 \*\*\*

actualDamageDealt -4.358e-01 1.025e-03 -425.3 <2e-16 \*\*\*

actualHeals 1.301e+01 6.441e-02 202.0 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1906 on 4446960 degrees of freedom

Multiple R-squared: 0.6155, Adjusted R-squared: 0.6155

F-statistic: 1.78e+06 on 4 and 4446960 DF, p-value: < 2.2e-16

> #단계적방법

Lm7 <- lm(winPlacePerc ~ +actualBoosts + actualHeals + killPlace, data=Pubg)

summary(Lm7)

vif(Lm7)

full <- lm(winPlacePerc ~ +actualBoosts + actualHeals + killPlace, data=Pubg)

null <- lm(winPlacePerc ~1, data=Pubg)

both <- step(null, direction="both", scope=(list(lower=null, upper=full)))

summary(both)

AIC(both)

BIC(both)

----------------

> fit(mod1, cok = "blue")

Error in fit(mod1, cok = "blue") : could not find function "fit"

> abline(mod1, col="blue")

Error in int\_abline(a = a, b = b, h = h, v = v, untf = untf, ...) :

plot.new has not been called yet

In addition: Warning message:

In abline(mod1, col = "blue") :

only using the first two of 7 regression coefficients

> mod1 <- lm(winPlacePerc ~ walkDistance + killPlace + boosts + weaponsAcquired + damageDealt + heals,data = Pubg)

mod1 <- lm(winPlacePerc ~ walkDistance + killPlace + boosts + weaponsAcquired + damageDealt + heals,data = Pubg)

> lm(winPlacePerc ~ walkDistance + killPlace + boosts + weaponsAcquired + damageDealt + heals,data = Pubg)

Call:

lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts +

weaponsAcquired + damageDealt + heals, data = Pubg)

Coefficients:

(Intercept) walkDistance killPlace boosts weaponsAcquired damageDealt heals

0.4797455 0.0001259 -0.0043167 0.0210424 0.0155966 -0.0002212 0.0012726

> summary(mod1)

Call:

lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts +

weaponsAcquired + damageDealt + heals, data = Pubg)

Residuals:

Min 1Q Median 3Q Max

-3.5151 -0.0867 -0.0064 0.0894 1.3893

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.797e-01 3.179e-04 1509.34 <2e-16 \*\*\*

walkDistance 1.259e-04 8.845e-08 1423.63 <2e-16 \*\*\*

killPlace -4.317e-03 4.030e-06 -1071.12 <2e-16 \*\*\*

boosts 2.104e-02 6.100e-05 344.95 <2e-16 \*\*\*

weaponsAcquired 1.560e-02 3.517e-05 443.43 <2e-16 \*\*\*

damageDealt -2.212e-04 5.809e-07 -380.82 <2e-16 \*\*\*

heals 1.273e-03 3.147e-05 40.44 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.148 on 4446958 degrees of freedom

Multiple R-squared: 0.7683, Adjusted R-squared: 0.7683

F-statistic: 2.457e+06 on 6 and 4446958 DF, p-value: < 2.2e-16

> mod <- lm(winPlacePerc ~ walkDistance + killPlace + boosts + weaponsAcquired,data = Pubg)

> pred <- predict(mod, newdata = Pubg)

> submit <- data.frame(Id = pubG$Id, winPlacePerc = pred)

Error in data.frame(Id = pubG$Id, winPlacePerc = pred) :

object 'pubG' not found

> submit <- data.frame(Id = Pubg$Id, winPlacePerc = pred)

[ reached 'max' / getOption("max.print") -- omitted 4396966 rows ]

> mod <- lm(winPlacePerc ~ walkDistance + killPlace + boosts + weaponsAcquired,data = Pubg)

> summary(mod)

Call:

lm(formula = winPlacePerc ~ walkDistance + killPlace + boosts +

weaponsAcquired, data = Pubg)

Residuals:

Min 1Q Median 3Q Max

-3.4352 -0.0929 -0.0083 0.0913 0.7660

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.152e-01 2.719e-04 1527.0 <2e-16 \*\*\*

walkDistance 1.307e-04 8.878e-08 1472.6 <2e-16 \*\*\*

killPlace -3.503e-03 3.454e-06 -1014.3 <2e-16 \*\*\*

boosts 1.591e-02 5.646e-05 281.8 <2e-16 \*\*\*

weaponsAcquired 1.525e-02 3.568e-05 427.4 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1504 on 4446960 degrees of freedom

Multiple R-squared: 0.7607, Adjusted R-squared: 0.7607

F-statistic: 3.534e+06 on 4 and 4446960 DF, p-value: < 2.2e-16

> summary(Pubg$winPlacePerc)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.2000 0.4583 0.4728 0.7407 1.0000

> nPlayersInMatch <- group\_by(Pubg, matchId) %>%

+ summarize(nPlayersInMatch = n())

> Pubg1 <- left\_join(Pubg, nPlayersInMatch, by = 'matchId')

> Pubg1 <- mutate(Pubg, totalDist = rideDistance + swimDistance + walkDistance) %>%

+ mutate(healsAndBoosts = heals + boosts) %>%

+ mutate(kills = kills \* 100 / nPlayersInMatch)

Error: Column `kills` is of unsupported class data.frame

In addition: Warning message:

In Ops.factor(left, right) : ‘/’ not meaningful for factors

>

>library(rpart)

install.packages("partykit")

library(grid)

library(libcoin)

library(mvtnorm)

library(partykit)

tree.pros <- rpart(winPlacePerc~killPlace + actualBoosts + actualDamageDealt + actualHeals , data=Pubg)

#분할횟수 정하기 위해 분할 횟수당 오차 계산

print(tree.pros$cptable)

plotcp(tree.pros)

#xerror중 최소값 고르기

cp <- min(tree.pros$cptable[5,])

#회귀나무 자르기

prune.tree.pros <- prune(tree.pros, cp=cp)

#나무그리기

plot(as.party(tree.pros))

plot(as.party(prune.tree.pros))