P8106 - Final Project

Mingkuan Xu (mx2262)

5/9/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

Data Preprocessing

```
df_salary = read_csv("NBA_season2122_player_salary.csv") %>%
  janitor::clean_names() %>%
  select(Player=x2,Team=x3,Salary=salary_4) %>%
  na.omit()
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * Salary -> Salary...4
## * Salary -> Salary...5
## * ...
```

```
## Rows: 578 Columns: 11

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): ...1, ...2, ...3, Salary...4, Salary...5, Salary...6, Salary...7, ...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_salary = df_salary[-1,]

df_stats = read_csv("NBA_season2122_player_stats.csv") %>%
  rename(Team=Tm) %>%
  select(-Rk)
```

```
## Rows: 784 Columns: 30

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Player, Pos, Tm
## dbl (27): Rk, Age, G, GS, MP, FG, FGA, FG%, 3P, 3PA, 3P%, 2P, 2PA, 2P%, eFG%...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df_players = inner_join(x=df_salary,y=df_stats,by=c("Player","Team")) %>%
  janitor::clean_names() %>%
  distinct()

df_players = df_players %>%
  arrange(player,desc(g)) %>%
  distinct(player,.keep_all = TRUE)

# Removed variables with missing data and resulted from division of other variables
df_players = df_players %>%
  select(-x3p_percent, -ft_percent, -fg_percent,-x2p_percent,-e_fg_percent)

# The final generated dataset for use: df_player.
```

```
# Convert salary from characters to numbers.
# Convert categorical variables to factors

df_players = df_players %>%
  separate(salary,into = c("symbol", "salary"),1) %>%
  select(-symbol)%>%
  mutate(salary = as.numeric(salary)/1000000,
         team = factor(team),
         pos = factor(pos)) %>%
  relocate(salary, .after = last_col())

colnames(df_players) = c("player", "team", "position", "age", "game","game_starting" ,"minute","field_g
```

```r
df_players = df_players %>%
  mutate(field_goal = field_goal/minute,
         fg_attempt = fg_attempt/minute,
         x3p = x3p/minute,
         x3p_attempt = x3p_attempt/minute,
         x2p = x2p/minute,
         x2p_attempt = x2p_attempt/minute,
         free_throw = free_throw/minute,
         ft_attempt = ft_attempt/minute,
         offensive_rb = offensive_rb/minute,
         defenssive_rb = defenssive_rb/minute,
         total_rb = total_rb/minute,
         assistance = assistance/minute,
         steal = steal/minute,
         block = block/minute,
         turnover = turnover/minute,
         personal_foul = personal_foul/minute,
         point = point/minute)
```

```r
# Data partition
set.seed(8106)

indexTrain <- createDataPartition(y = df_players$salary, p = 0.75, list = FALSE, times = 1)
ctrl1 <- trainControl(method = "cv", number = 10, repeats = 5)
```

```
## Warning: ‘repeats‘ has no meaning for this resampling method.
```

Blackbox