Final Project

Cady Stringer, James Kistner, Miranda Heyer, Justo Garica

11/21/2019

Countries and continent data from <https://github.com/dbouquin/IS_608/blob/master/NanosatDB_munging/Countries-Continents.csv> FIFA data from <https://towardsdatascience.com/exploratory-analysis-of-fifa-18-dataset-using-r-ba09aa4a2d3c>

CLEANING:

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ───────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':  
##   
## expand

## Loading required package: foreach

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loaded glmnet 2.0-18

library(glmnetUtils)

##   
## Attaching package: 'glmnetUtils'

## The following objects are masked from 'package:glmnet':  
##   
## cv.glmnet, glmnet

library(readr)  
library(dplyr)  
library(textclean)  
library(coefplot)  
library(hdm)  
library(leaps)  
library(useful)  
library(forcats)  
library(varhandle)  
library(regclass)

## Loading required package: bestglm

## Loading required package: VGAM

## Loading required package: stats4

## Loading required package: splines

##   
## Attaching package: 'VGAM'

## The following object is masked from 'package:tidyr':  
##   
## fill

## Loading required package: rpart

## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

## Important regclass change from 1.3:  
## All functions that had a . in the name now have an \_  
## all.correlations -> all\_correlations, cor.demo -> cor\_demo, etc.

library(RColorBrewer)  
library(maps)

##   
## Attaching package: 'maps'

## The following object is masked from 'package:purrr':  
##   
## map

library(geosphere)  
library(randomForest)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'lattice'

## The following object is masked from 'package:regclass':  
##   
## qq

##   
## Attaching package: 'caret'

## The following object is masked from 'package:VGAM':  
##   
## predictors

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

library(randomForestExplainer)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

fifa\_data <- read.csv(here::here("datasets","fifa\_data.csv"))   
countries <- read.csv(here::here("datasets","CountriesContinents.csv"))   
colnames(countries)[2] <- "Nationality"  
fifa\_data <- right\_join(countries, fifa\_data)

## Joining, by = "Nationality"

## Warning: Column `Nationality` joining factors with different levels,  
## coercing to character vector

fifa\_data$Wage <- sub('K', '', (fifa\_data$Wage))   
fifa\_data$Wage <- sub('€', '', (fifa\_data$Wage))   
fifa\_data$Weight <- sub('lbs', '', fifa\_data$Weight)  
  
fifa\_data$Weight <- as.numeric(fifa\_data$Weight)  
fifa\_data$Wage <- as.numeric(fifa\_data$Wage)  
  
fifa\_data <- fifa\_data %>% mutate(Club = fct\_lump(Club, n = 10))  
  
cols <- c("X", "ID", "Name", "Nationality", "Value", "Release.Clause", "Photo", "Jersey.Number", "Joined", "Flag", "Club.Logo", "Special", "Body.Type", "Real.Face", "Loaned.From", "Contract.Valid.Until")  
  
fifa\_data <- fifa\_data %>% dplyr::select(-cols)  
fifa\_data <- fifa\_data[complete.cases(fifa\_data), ]  
fifa\_data <- na.omit(fifa\_data)  
  
fifa\_data <- fifa\_data[-(15:40)]  
  
for (i in 1:dim(fifa\_data)[2]){  
 count = 0  
 for (j in 1:nrow(fifa\_data))  
 {  
 if ((fifa\_data[j - count, i] == "")){  
 fifa\_data %<>% slice(-(j - count))  
 count = count + 1  
 }  
 }  
}  
  
fifa\_data$Height <- strip(fifa\_data$Height, digit.remove = FALSE,  
 apostrophe.remove = TRUE, lower.case = TRUE)  
  
fifa\_data$Height <- as.numeric(fifa\_data$Height)  
  
height\_to\_inches <- function(height){  
 if(height >= 500){  
 inches\_left = floor((height/100))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 500  
 inches = inches\_left + inches\_right  
 }  
 else if(height >= 60){  
 inches\_left = floor((height/10))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 60  
 inches = inches\_left + inches\_right  
 }  
 else if(height >= 50){  
 inches\_left = floor((height/10))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 50  
 inches = inches\_left + inches\_right  
 }  
 return(inches)  
}  
  
for (i in 1:nrow(fifa\_data)){  
 fifa\_data$Height[[i]] <- height\_to\_inches(fifa\_data$Height[[i]])  
}

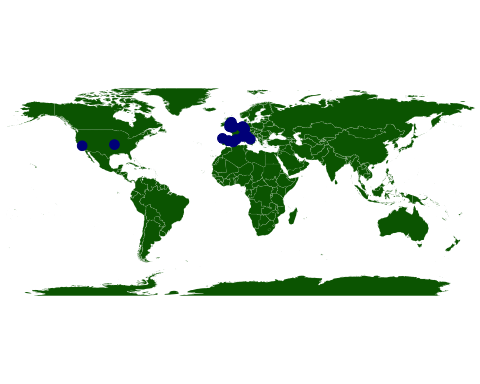
TEST/TRAIN SPLIT:

set.seed(9999)  
  
train\_size <- 0.7  
train\_idx <- sample(1:nrow(fifa\_data), size = train\_size\*(nrow(fifa\_data)))  
fifa\_train <- fifa\_data %>% slice(train\_idx)  
fifa\_test <- fifa\_data %>% slice(-train\_idx)  
  
options(scipen = 5)  
summary(fifa\_data)

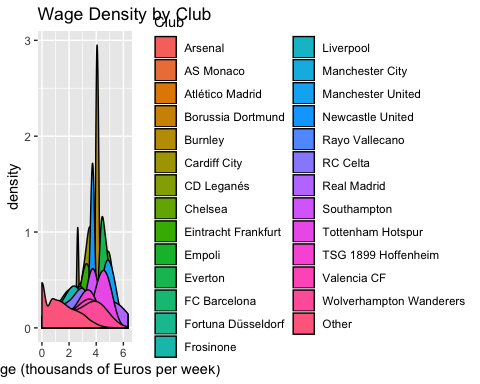
## Continent Age Overall Potential   
## Africa :1090 Min. :16.00 Min. :47.00 Min. :48.00   
## Asia :1189 1st Qu.:21.00 1st Qu.:63.00 1st Qu.:68.00   
## Europe :7707 Median :25.00 Median :67.00 Median :72.00   
## North America: 547 Mean :25.18 Mean :66.98 Mean :71.93   
## Oceania : 269 3rd Qu.:28.00 3rd Qu.:71.00 3rd Qu.:76.00   
## South America:3108 Max. :45.00 Max. :94.00 Max. :95.00   
##   
## Club Wage Preferred.Foot  
## Other :13270 Min. : 1.00 : 0   
## Atlético Madrid: 33 1st Qu.: 1.00 Left : 3303   
## CD Leganés : 33 Median : 3.00 Right:10607   
## FC Barcelona : 33 Mean : 10.56   
## Frosinone : 33 3rd Qu.: 10.00   
## RC Celta : 33 Max. :565.00   
## (Other) : 475   
## International.Reputation Weak.Foot Skill.Moves   
## Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:3.00 1st Qu.:2.000   
## Median :1.000 Median :3.00 Median :2.000   
## Mean :1.129 Mean :2.96 Mean :2.398   
## 3rd Qu.:1.000 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :5.000 Max. :5.00 Max. :5.000   
##   
## Work.Rate Position Height Weight   
## Medium/ Medium:7645 ST :1628 Min. :61.00 Min. :110.0   
## High/ Medium :2383 GK :1528 1st Qu.:69.00 1st Qu.:154.0   
## Medium/ High :1235 CB :1358 Median :71.00 Median :165.0   
## Medium/ Low : 696 LB :1019 Mean :71.31 Mean :166.2   
## High/ High : 690 CM : 971 3rd Qu.:73.00 3rd Qu.:176.0   
## High/ Low : 578 RB : 963 Max. :81.00 Max. :236.0   
## (Other) : 683 (Other):6443   
## Crossing Finishing HeadingAccuracy ShortPassing   
## Min. : 5.00 Min. : 2.00 Min. : 4.0 Min. :11.00   
## 1st Qu.:39.00 1st Qu.:30.00 1st Qu.:45.0 1st Qu.:55.00   
## Median :55.00 Median :49.00 Median :56.0 Median :63.00   
## Mean :50.46 Mean :46.02 Mean :52.8 Mean :59.54   
## 3rd Qu.:64.00 3rd Qu.:62.00 3rd Qu.:65.0 3rd Qu.:69.00   
## Max. :93.00 Max. :95.00 Max. :94.0 Max. :93.00   
##   
## Volleys Dribbling Curve FKAccuracy   
## Min. : 4.00 Min. : 4.00 Min. : 6.00 Min. : 3.00   
## 1st Qu.:31.00 1st Qu.:50.00 1st Qu.:35.00 1st Qu.:31.00   
## Median :45.00 Median :62.00 Median :49.00 Median :42.00   
## Mean :43.59 Mean :56.14 Mean :47.94 Mean :43.47   
## 3rd Qu.:58.00 3rd Qu.:69.00 3rd Qu.:63.00 3rd Qu.:57.00   
## Max. :90.00 Max. :97.00 Max. :94.00 Max. :94.00   
##   
## LongPassing BallControl Acceleration SprintSpeed   
## Min. : 9.0 Min. : 5.00 Min. :12.00 Min. :12.00   
## 1st Qu.:44.0 1st Qu.:55.00 1st Qu.:57.00 1st Qu.:57.00   
## Median :57.0 Median :64.00 Median :67.00 Median :68.00   
## Mean :53.4 Mean :59.25 Mean :64.77 Mean :64.86   
## 3rd Qu.:65.0 3rd Qu.:70.00 3rd Qu.:75.00 3rd Qu.:75.00   
## Max. :93.0 Max. :96.00 Max. :97.00 Max. :96.00   
##   
## Agility Reactions Balance ShotPower   
## Min. :15.00 Min. :30.0 Min. :16.00 Min. : 3.00   
## 1st Qu.:56.00 1st Qu.:57.0 1st Qu.:56.00 1st Qu.:47.00   
## Median :66.00 Median :63.0 Median :66.00 Median :60.00   
## Mean :63.67 Mean :62.6 Mean :63.96 Mean :56.37   
## 3rd Qu.:74.00 3rd Qu.:69.0 3rd Qu.:74.00 3rd Qu.:69.00   
## Max. :96.00 Max. :96.0 Max. :96.00 Max. :95.00   
##   
## Jumping Stamina Strength LongShots   
## Min. :15.00 Min. :12.00 Min. :17.00 Min. : 3.00   
## 1st Qu.:58.00 1st Qu.:56.00 1st Qu.:58.00 1st Qu.:34.00   
## Median :66.00 Median :67.00 Median :67.00 Median :52.00   
## Mean :65.21 Mean :63.36 Mean :65.56 Mean :47.88   
## 3rd Qu.:73.00 3rd Qu.:74.00 3rd Qu.:74.00 3rd Qu.:63.00   
## Max. :95.00 Max. :96.00 Max. :95.00 Max. :94.00   
##   
## Aggression Interceptions Positioning Vision   
## Min. :11.0 Min. : 3.00 Min. : 2.00 Min. :10.00   
## 1st Qu.:44.0 1st Qu.:26.00 1st Qu.:39.00 1st Qu.:44.00   
## Median :59.0 Median :53.00 Median :56.00 Median :55.00   
## Mean :56.1 Mean :47.12 Mean :50.48 Mean :53.88   
## 3rd Qu.:70.0 3rd Qu.:65.00 3rd Qu.:65.00 3rd Qu.:65.00   
## Max. :94.0 Max. :92.00 Max. :95.00 Max. :94.00   
##   
## Penalties Composure Marking StandingTackle   
## Min. : 5.00 Min. : 3.00 Min. : 3.00 Min. : 6.00   
## 1st Qu.:39.00 1st Qu.:52.00 1st Qu.:30.00 1st Qu.:27.00   
## Median :50.00 Median :60.00 Median :53.00 Median :55.00   
## Mean :49.04 Mean :59.34 Mean :47.68 Mean :48.05   
## 3rd Qu.:61.00 3rd Qu.:67.00 3rd Qu.:64.00 3rd Qu.:67.00   
## Max. :92.00 Max. :96.00 Max. :94.00 Max. :93.00   
##   
## SlidingTackle GKDiving GKHandling GKKicking   
## Min. : 6.00 Min. : 1.00 Min. : 1.00 Min. : 1.00   
## 1st Qu.:24.00 1st Qu.: 8.00 1st Qu.: 8.00 1st Qu.: 8.00   
## Median :53.00 Median :11.00 Median :11.00 Median :11.00   
## Mean :45.99 Mean :16.57 Mean :16.34 Mean :16.15   
## 3rd Qu.:64.00 3rd Qu.:14.00 3rd Qu.:14.00 3rd Qu.:14.00   
## Max. :91.00 Max. :90.00 Max. :92.00 Max. :91.00   
##   
## GKPositioning GKReflexes   
## Min. : 1.00 Min. : 1.00   
## 1st Qu.: 8.00 1st Qu.: 8.00   
## Median :11.00 Median :11.00   
## Mean :16.34 Mean :16.66   
## 3rd Qu.:14.00 3rd Qu.:14.00   
## Max. :90.00 Max. :94.00   
##

EXPLORATION PLOTS:

Arsenal <- c(-0.1, 51.5)  
AS\_monaco <- c(7.4, 43.7 )  
Madrid <- c(-3.7, 40.4)  
Borussia <- c(7.4, 51.51)  
Burnley <- c(-2.2, 53.7)  
Cardiff <- c(-3.1, 51.4)  
Leganes <- c(-3.7, 40.3)  
Chelsea <- c(-0.17, 51.4)  
Frankfurt <- c(8.6, 50.1)  
Empoli <- c(10.9, 43.7)  
Everton <- c(-93.7, 37.3)  
Barcelona <- c(2.1, 41.3)  
Dusseldorf <- c(6.7, 51.2)  
Frosinone <- c(13.3, 41.6)  
Liverpool <- c(-2.9, 53.4)  
Manchester <- c(-2.2, 53.4)  
Newcastle <- c(-1.6, 54.9)  
Rayo <- c(-119.1, 36.4)  
Vigo <- c(-8.7, 42.2)  
Southhampton <- c(-1.4, 50.9)  
Tottenham <- c(-0.06, 51.6)  
Hoffenheim <- c(8.8, 49.2)  
Valencia <- c(-0.3, 39.4)  
Wolverhampton <- c(-2.1, 52.5)  
  
# Data frame  
data <- rbind(Arsenal, AS\_monaco, Madrid, Borussia, Burnley, Cardiff, Leganes, Chelsea, Frankfurt, Empoli,  
 Everton, Barcelona, Dusseldorf, Frosinone, Liverpool, Manchester, Newcastle,   
 Rayo, Vigo, Southhampton, Tottenham, Hoffenheim, Valencia, Wolverhampton) %>% as.data.frame()  
colnames(data) <- c("long","lat")  
  
# Show the cities on the map  
map('world',  
col="darkgreen", fill=TRUE, bg="white", lwd=0.05,  
mar=rep(0,4),border=0, ylim=c(-80,80)   
)  
points(x=data$long, y=data$lat, col="darkblue", cex=2, pch=20)

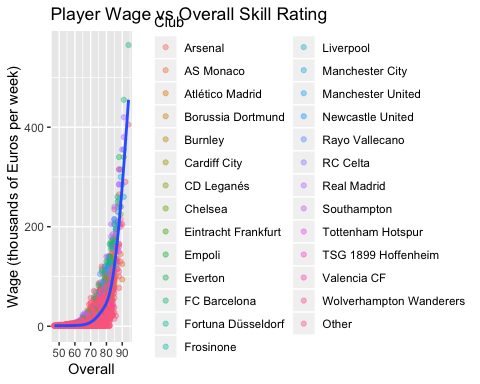


#Density Plot  
ggplot(data=fifa\_data, aes(x=log(Wage),fill=Club)) + geom\_density(adjust=1.5) +labs(x = "Log Wage (thousands of Euros per week)", title="Wage Density by Club")

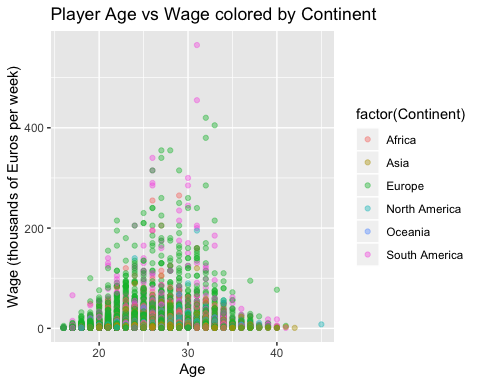


#Wage vs Overall  
ggplot(fifa\_data, aes(x = Overall, y = Wage)) + geom\_point(alpha = 0.4, aes(color = Club)) + geom\_smooth() + labs(y = "Wage (thousands of Euros per week)", title = "Player Wage vs Overall Skill Rating")

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

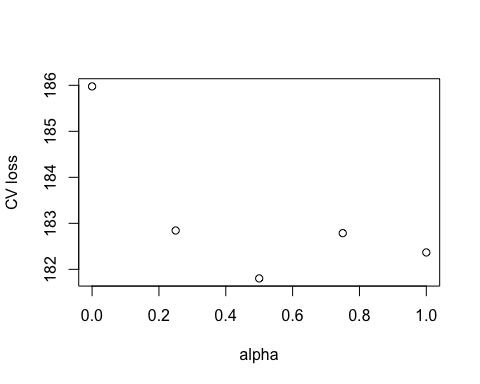


#Wage vs Age  
ggplot(fifa\_data, aes(x = Age, y = Wage)) + geom\_point(alpha = 0.4,aes(color = factor(Continent))) + labs(y = "Wage (thousands of Euros per week)", title = "Player Age vs Wage colored by Continent")

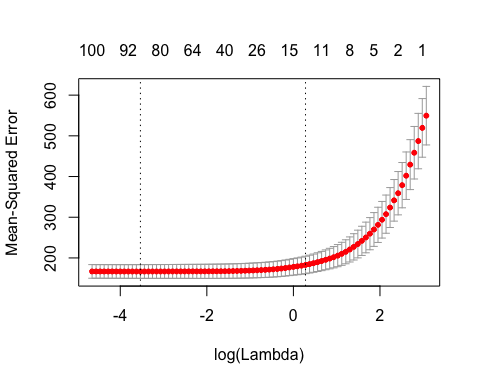


ELASTIC NET:

alpha\_grid <- seq(0,1, length = 5)  
fifa\_enet\_mod <-cva.glmnet(Wage ~.,  
 data = fifa\_train,  
 alpha = alpha\_grid)  
  
fifa\_enet\_coef <- data.frame(  
 varname = rownames(coef(fifa\_enet\_mod$modlist[[1]])),  
 ridge = as.matrix(coef(fifa\_enet\_mod$modlist[[1]])) %>% round(3),  
 coef\_0.25 = as.matrix(coef(fifa\_enet\_mod$modlist[[2]])) %>% round(3),  
 coef\_0.5 = as.matrix(coef(fifa\_enet\_mod$modlist[[3]])) %>% round(3),  
 coef\_0.75 = as.matrix(coef(fifa\_enet\_mod$modlist[[4]])) %>% round(3),  
 lasso = as.matrix(coef(fifa\_enet\_mod$modlist[[5]])) %>% round(3)  
) %>% remove\_rownames() %>% rename(varname = 1, ridge = 2, coef\_0.25 = 3, coef\_0.5 = 4, coef\_0.75 = 5, lasso = 6)  
  
minlossplot(fifa\_enet\_mod)



plot(fifa\_enet\_mod$modlist[[4]])



preds\_DF <- data.frame(  
 fifa\_enet\_preds <- round(predict(fifa\_enet\_mod, alpha = 0.75, fifa\_test,  
 s = fifa\_enet\_mod$modlist[[4]]$lambda.1se), 1), fifa\_test  
)  
  
coef(fifa\_enet\_mod$modlist[[4]])

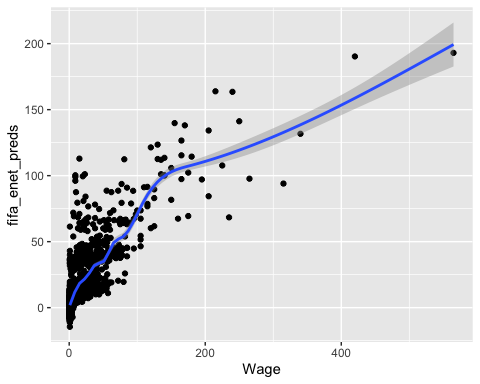
## 118 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -76.83014389  
## ContinentAfrica .   
## ContinentAsia .   
## ContinentEurope .   
## ContinentNorth America .   
## ContinentOceania .   
## ContinentSouth America .   
## Age .   
## Overall 0.73084550  
## Potential 0.30201354  
## Club .   
## ClubArsenal 16.74818930  
## ClubAS Monaco .   
## ClubAtlético Madrid .   
## ClubBorussia Dortmund .   
## ClubBurnley .   
## ClubCardiff City .   
## ClubCD Leganés .   
## ClubChelsea 25.41936456  
## ClubEintracht Frankfurt .   
## ClubEmpoli .   
## ClubEverton 8.18744693  
## ClubFC Barcelona 61.33685382  
## ClubFortuna Düsseldorf .   
## ClubFrosinone .   
## ClubLiverpool 17.94026005  
## ClubManchester City 56.32824984  
## ClubManchester United 49.79157363  
## ClubNewcastle United .   
## ClubRayo Vallecano .   
## ClubRC Celta .   
## ClubReal Madrid 83.93217504  
## ClubSouthampton .   
## ClubTottenham Hotspur 2.56812281  
## ClubTSG 1899 Hoffenheim .   
## ClubValencia CF .   
## ClubWolverhampton Wanderers .   
## ClubOther -9.49017682  
## Preferred.Foot .   
## Preferred.FootLeft .   
## Preferred.FootRight .   
## International.Reputation 22.28914109  
## Weak.Foot .   
## Skill.Moves .   
## Work.Rate .   
## Work.RateHigh/ High 0.07481775  
## Work.RateHigh/ Low .   
## Work.RateHigh/ Medium .   
## Work.RateLow/ High .   
## Work.RateLow/ Low .   
## Work.RateLow/ Medium .   
## Work.RateMedium/ High .   
## Work.RateMedium/ Low .   
## Work.RateMedium/ Medium .   
## Position .   
## PositionCAM .   
## PositionCB .   
## PositionCDM .   
## PositionCF .   
## PositionCM .   
## PositionGK .   
## PositionLAM .   
## PositionLB .   
## PositionLCB .   
## PositionLCM .   
## PositionLDM .   
## PositionLF .   
## PositionLM .   
## PositionLS .   
## PositionLW .   
## PositionLWB .   
## PositionRAM .   
## PositionRB .   
## PositionRCB .   
## PositionRCM .   
## PositionRDM .   
## PositionRF .   
## PositionRM .   
## PositionRS .   
## PositionRW .   
## PositionRWB .   
## PositionST .   
## Height .   
## Weight .   
## Crossing .   
## Finishing .   
## HeadingAccuracy .   
## ShortPassing .   
## Volleys .   
## Dribbling .   
## Curve .   
## FKAccuracy .   
## LongPassing .   
## BallControl .   
## Acceleration .   
## SprintSpeed .   
## Agility .   
## Reactions .   
## Balance .   
## ShotPower .   
## Jumping .   
## Stamina .   
## Strength .   
## LongShots .   
## Aggression .   
## Interceptions .   
## Positioning .   
## Vision .   
## Penalties .   
## Composure .   
## Marking .   
## StandingTackle .   
## SlidingTackle .   
## GKDiving .   
## GKHandling .   
## GKKicking .   
## GKPositioning .   
## GKReflexes .

postResample(pred = fifa\_enet\_preds, obs = fifa\_test$Wage)

## RMSE Rsquared MAE   
## 14.5301080 0.6303753 6.3339803

ggplot(fifa\_test, aes(x = Wage, y = fifa\_enet\_preds)) + geom\_point() + geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

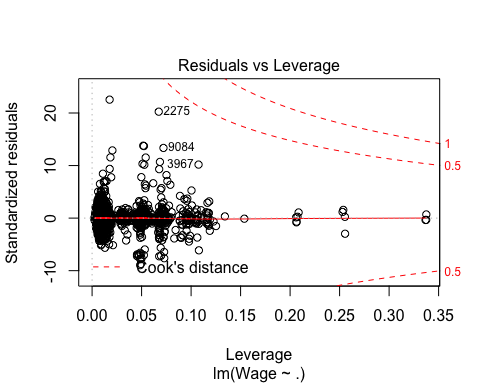
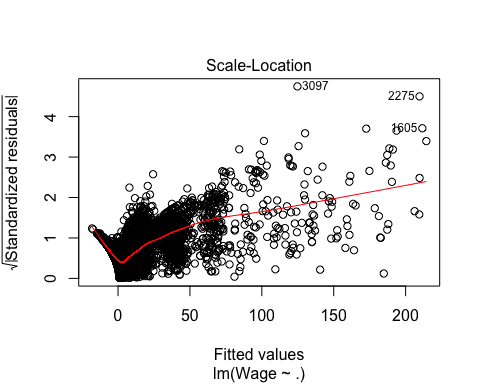
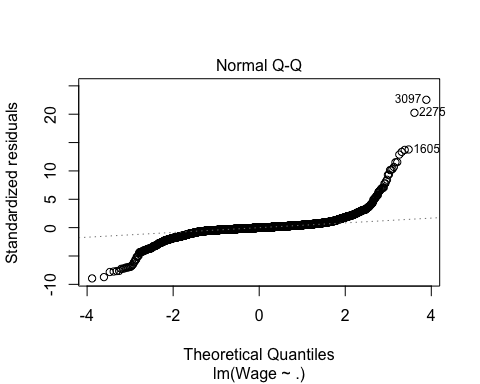
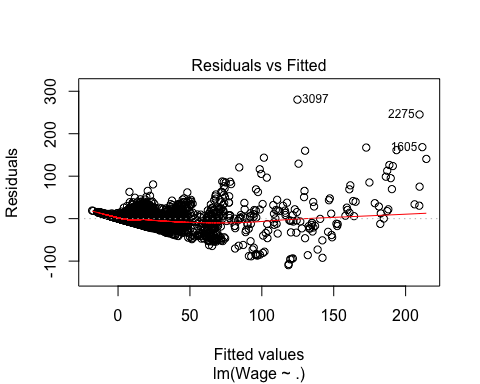


LINEAR MODEL:

fifa\_lm\_mod <- lm(Wage ~ .,  
 fifa\_train)  
VIF(fifa\_lm\_mod)

## GVIF Df GVIF^(1/(2\*Df))  
## Continent 1.544405 5 1.044422  
## Age 5.037130 1 2.244355  
## Overall 15.725947 1 3.965595  
## Potential 6.325490 1 2.515053  
## Club 1.529359 26 1.008204  
## Preferred.Foot 1.393621 1 1.180517  
## International.Reputation 1.655758 1 1.286763  
## Weak.Foot 1.237984 1 1.112647  
## Skill.Moves 4.500746 1 2.121496  
## Work.Rate 2.575030 8 1.060899  
## Position 4262.871844 26 1.174362  
## Height 4.689245 1 2.165467  
## Weight 3.479866 1 1.865440  
## Crossing 7.835982 1 2.799282  
## Finishing 11.637939 1 3.411442  
## HeadingAccuracy 7.193412 1 2.682054  
## ShortPassing 13.425909 1 3.664138  
## Volleys 7.145719 1 2.673148  
## Dribbling 17.870834 1 4.227391  
## Curve 7.384014 1 2.717354  
## FKAccuracy 5.299520 1 2.302069  
## LongPassing 8.045780 1 2.836508  
## BallControl 20.766490 1 4.557026  
## Acceleration 9.878835 1 3.143061  
## SprintSpeed 8.012344 1 2.830608  
## Agility 5.236423 1 2.288323  
## Reactions 4.396829 1 2.096862  
## Balance 5.076261 1 2.253056  
## ShotPower 6.857336 1 2.618651  
## Jumping 1.582404 1 1.257936  
## Stamina 4.131309 1 2.032562  
## Strength 3.516996 1 1.875365  
## LongShots 9.923917 1 3.150225  
## Aggression 4.069168 1 2.017218  
## Interceptions 12.453396 1 3.528937  
## Positioning 10.703887 1 3.271680  
## Vision 4.742451 1 2.177717  
## Penalties 5.021855 1 2.240950  
## Composure 3.863305 1 1.965529  
## Marking 7.269665 1 2.696232  
## StandingTackle 28.752228 1 5.362110  
## SlidingTackle 23.188632 1 4.815458  
## GKDiving 30.804282 1 5.550161  
## GKHandling 29.015448 1 5.386599  
## GKKicking 22.917465 1 4.787219  
## GKPositioning 26.927147 1 5.189137  
## GKReflexes 31.328573 1 5.597193

plot(fifa\_lm\_mod)

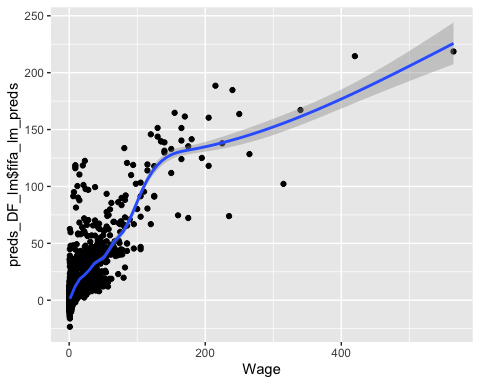


preds\_DF\_lm <- data.frame(fifa\_lm\_preds = predict(fifa\_lm\_mod, newdata=fifa\_test))  
postResample(preds\_DF\_lm$fifa\_lm\_preds,fifa\_test$Wage)

## RMSE Rsquared MAE   
## 13.8387950 0.6587853 6.3945620

ggplot(fifa\_test, aes(x = Wage, y = preds\_DF\_lm$fifa\_lm\_preds)) + geom\_point() + geom\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



RANDOM FOREST:

fifa\_train\_noclub <- fifa\_train %>% select(-"Club")  
rf\_fifa\_fit <- randomForest(Wage ~.,  
 data = fifa\_train\_noclub,  
 type = regression,  
 ntree = 100,  
 importance = TRUE,   
 localImp = TRUE   
 )  
  
fifa\_rf\_test\_preds <- predict(rf\_fifa\_fit, newdata = fifa\_test)  
fifa\_rf\_train\_preds <- predict(rf\_fifa\_fit)  
  
fifa\_train\_resids <- (fifa\_train$Wage - fifa\_rf\_train\_preds)  
  
fifa\_train\_resids <- sort(fifa\_train\_resids)  
bottom10 <- head(fifa\_train\_resids, n = 10)  
top10 <- tail(fifa\_train\_resids, n = 10)  
  
fifa\_data <- read.csv(here::here("datasets","fifa\_data.csv"))   
countries <- read.csv(here::here("datasets","CountriesContinents.csv"))   
colnames(countries)[2] <- "Nationality"  
fifa\_data <- right\_join(countries, fifa\_data)

## Joining, by = "Nationality"

## Warning: Column `Nationality` joining factors with different levels,  
## coercing to character vector

fifa\_data$Wage <- sub('K', '', (fifa\_data$Wage))   
fifa\_data$Wage <- sub('€', '', (fifa\_data$Wage))   
fifa\_data$Weight <- sub('lbs', '', fifa\_data$Weight)  
  
fifa\_data$Weight <- as.numeric(fifa\_data$Weight)  
fifa\_data$Wage <- as.numeric(fifa\_data$Wage)  
  
fifa\_data <- fifa\_data %>% mutate(Club = fct\_lump(Club, n = 10))  
  
cols <- c("X", "ID", "Nationality", "Value", "Release.Clause", "Photo", "Jersey.Number", "Joined", "Flag", "Club.Logo", "Special", "Body.Type", "Real.Face", "Loaned.From", "Contract.Valid.Until")  
  
fifa\_data <- fifa\_data %>% dplyr::select(-cols)  
fifa\_data <- fifa\_data[complete.cases(fifa\_data), ]  
fifa\_data <- na.omit(fifa\_data)  
  
fifa\_data <- fifa\_data[-(15:40)]  
  
for (i in 1:dim(fifa\_data)[2]){  
 count = 0  
 for (j in 1:nrow(fifa\_data))  
 {  
 if ((fifa\_data[j - count, i] == "")){  
 fifa\_data %<>% slice(-(j - count))  
 count = count + 1  
 }  
 }  
}  
  
fifa\_data$Height <- strip(fifa\_data$Height, digit.remove = FALSE,  
 apostrophe.remove = TRUE, lower.case = TRUE)  
  
fifa\_data$Height <- as.numeric(fifa\_data$Height)  
  
height\_to\_inches <- function(height){  
 if(height >= 500){  
 inches\_left = floor((height/100))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 500  
 inches = inches\_left + inches\_right  
 }  
 else if(height >= 60){  
 inches\_left = floor((height/10))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 60  
 inches = inches\_left + inches\_right  
 }  
 else if(height >= 50){  
 inches\_left = floor((height/10))  
 inches\_left = inches\_left \* 12  
 inches\_right = height - 50  
 inches = inches\_left + inches\_right  
 }  
 return(inches)  
}  
  
for (i in 1:nrow(fifa\_data)){  
 fifa\_data$Height[[i]] <- height\_to\_inches(fifa\_data$Height[[i]])  
}# returning sorted x (the last evaluated value inside the body   
  
top10row <- c(8236, 8869, 2946, 6650, 614, 5195, 3097, 1974, 1605, 2275)  
top10names <- c(toString(fifa\_data$Name[8236]), toString(fifa\_data$Name[8869]), toString(fifa\_data$Name[2946]), toString(fifa\_data$Name[6650]), toString(fifa\_data$Name[614]), toString(fifa\_data$Name[5195]), toString(fifa\_data$Name[3097]), toString(fifa\_data$Name[1974]), toString(fifa\_data$Name[1605]), toString(fifa\_data$Name[2275]))  
  
toString(fifa\_data$Name[8236])

## [1] "R. Marchizza"

top10names

## [1] "R. Marchizza" "F. García" "J. Foyth"   
## [4] "L. Comas" "O. Al Soma" "D. Lemos"   
## [7] "M. van der Werff" "Bebé" "Pedraza"   
## [10] "J. Stryger Larsen"

bottom10row <- c(7261, 1608, 448, 2282, 1885, 9349, 2004, 5813, 1996, 9498)  
bottom10names<- c(toString(fifa\_data$Name[7261]), toString(fifa\_data$Name[1608]), toString(fifa\_data$Name[448]), toString(fifa\_data$Name[2282]), toString(fifa\_data$Name[1885]), toString(fifa\_data$Name[9349]), toString(fifa\_data$Name[2004]), toString(fifa\_data$Name[5813]), toString(fifa\_data$Name[1996]), toString(fifa\_data$Name[9498]))  
  
bottom10names

## [1] "M. Wostry" "A. Limbombe" "V. Tsygankov"   
## [4] "F. Aguilar" "Kévin Rodrigues" "Luan"   
## [7] "G. Burdisso" "M. Herrera" "M. Gulde"   
## [10] "Paulo Vitor"

top4 <- fifa\_data$Name[8869]  
  
caret::RMSE(fifa\_rf\_test\_preds,fifa\_test$Wage)

## [1] 11.42859

caret::R2(fifa\_rf\_test\_preds,fifa\_test$Wage)

## [1] 0.7864205

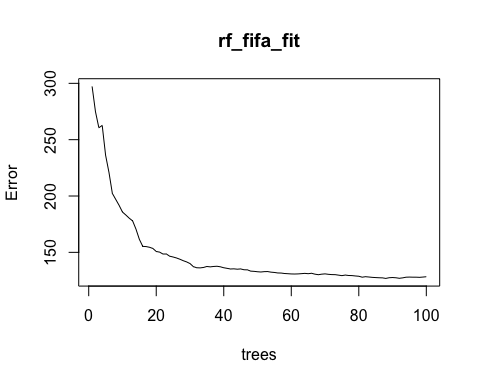
caret::MAE(fifa\_rf\_test\_preds,fifa\_test$Wage)

## [1] 4.578531

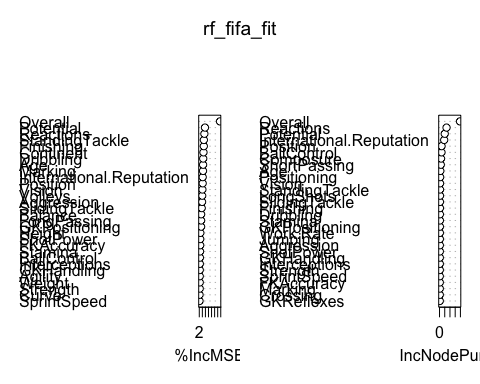
postResample(fifa\_rf\_test\_preds,fifa\_test$Wage)

## RMSE Rsquared MAE   
## 11.4285929 0.7864205 4.5785313

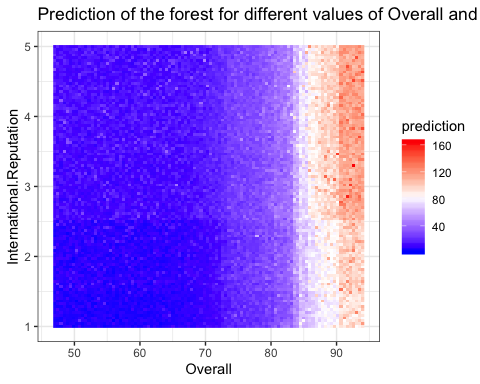
plot(rf\_fifa\_fit)



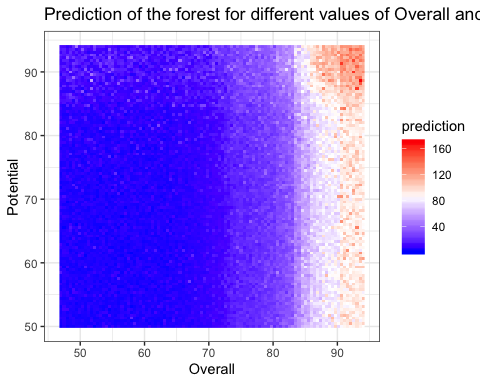
varImpPlot(rf\_fifa\_fit)



plot\_predict\_interaction(rf\_fifa\_fit, fifa\_train, "Overall", "International.Reputation")



plot\_predict\_interaction(rf\_fifa\_fit, fifa\_train, "Overall", "Potential")



LASSO MODEL:

lasso\_mod\_fifa <- cv.glmnet(Wage ~ .,   
 data = fifa\_train,  
 alpha = 1,  
 nfolds = 10)  
  
fifa\_lm\_preds = predict(lasso\_mod\_fifa, newdata=fifa\_test)  
postResample(fifa\_lm\_preds,fifa\_test$Wage)

## RMSE Rsquared MAE   
## 14.5956578 0.6270768 6.3506677

fifa\_lm\_mod\_lassopicks <- lm(Wage ~ International.Reputation + Overall + Potential + Club,  
 fifa\_train)  
preds\_DF\_lm\_lassopicks <- data.frame(fifa\_lm\_preds = predict(fifa\_lm\_mod\_lassopicks, newdata=fifa\_test))  
postResample(preds\_DF\_lm\_lassopicks$fifa\_lm\_preds,fifa\_test$Wage)

## RMSE Rsquared MAE   
## 14.03633 0.64896 6.26896

coef(lasso\_mod\_fifa, lasso\_mod\_fifa$lambda.min)

## 118 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -85.2707068552  
## ContinentAfrica -0.9859501900  
## ContinentAsia 2.4653324934  
## ContinentEurope -0.8177556971  
## ContinentNorth America 1.4123361774  
## ContinentOceania 0.8849611797  
## ContinentSouth America .   
## Age -0.6216034526  
## Overall 1.3661181964  
## Potential -0.1155728993  
## Club .   
## ClubArsenal 48.1678800989  
## ClubAS Monaco 5.9471464812  
## ClubAtlético Madrid 6.4270365225  
## ClubBorussia Dortmund -5.1490814621  
## ClubBurnley 16.1040876824  
## ClubCardiff City 7.3661851732  
## ClubCD Leganés .   
## ClubChelsea 56.8814949325  
## ClubEintracht Frankfurt -3.9104935688  
## ClubEmpoli -4.3964375927  
## ClubEverton 41.0498951546  
## ClubFC Barcelona 85.7912653188  
## ClubFortuna Düsseldorf -0.9385551735  
## ClubFrosinone -5.3925191920  
## ClubLiverpool 46.0036076623  
## ClubManchester City 84.3570536004  
## ClubManchester United 76.4567948336  
## ClubNewcastle United 9.8056464655  
## ClubRayo Vallecano .   
## ClubRC Celta -5.4125826795  
## ClubReal Madrid 108.3164540359  
## ClubSouthampton .   
## ClubTottenham Hotspur 37.3587799723  
## ClubTSG 1899 Hoffenheim -2.8856744660  
## ClubValencia CF -1.3495728096  
## ClubWolverhampton Wanderers 10.7241853497  
## ClubOther -4.5342479387  
## Preferred.Foot .   
## Preferred.FootLeft .   
## Preferred.FootRight .   
## International.Reputation 22.6773844645  
## Weak.Foot 0.1365707547  
## Skill.Moves -0.2605493079  
## Work.Rate .   
## Work.RateHigh/ High 3.1884960090  
## Work.RateHigh/ Low -0.2763160846  
## Work.RateHigh/ Medium -0.1258069808  
## Work.RateLow/ High -0.3084853834  
## Work.RateLow/ Low -1.4267023235  
## Work.RateLow/ Medium .   
## Work.RateMedium/ High 0.2906151536  
## Work.RateMedium/ Low -2.0755475965  
## Work.RateMedium/ Medium .   
## Position .   
## PositionCAM -0.0465339188  
## PositionCB -0.7278100402  
## PositionCDM -0.4465201718  
## PositionCF .   
## PositionCM 0.6982930727  
## PositionGK 6.6855917094  
## PositionLAM 2.3098690438  
## PositionLB .   
## PositionLCB -1.2701746837  
## PositionLCM 0.0410147040  
## PositionLDM -0.5351205110  
## PositionLF 2.1588215859  
## PositionLM 0.0197181685  
## PositionLS 0.0456016491  
## PositionLW 2.5404279257  
## PositionLWB 0.1628556235  
## PositionRAM 2.5863345271  
## PositionRB -0.3516000514  
## PositionRCB -1.2087911130  
## PositionRCM .   
## PositionRDM .   
## PositionRF .   
## PositionRM -0.2065094331  
## PositionRS 0.6137951376  
## PositionRW 0.2530433410  
## PositionRWB -0.5155174522  
## PositionST .   
## Height 0.1417077173  
## Weight 0.0304194001  
## Crossing 0.0014206945  
## Finishing .   
## HeadingAccuracy 0.0192840803  
## ShortPassing -0.0488691961  
## Volleys 0.0328855014  
## Dribbling .   
## Curve -0.0140064903  
## FKAccuracy .   
## LongPassing .   
## BallControl .   
## Acceleration .   
## SprintSpeed 0.0074977806  
## Agility .   
## Reactions 0.0017239101  
## Balance .   
## ShotPower -0.0231931389  
## Jumping -0.0049346189  
## Stamina .   
## Strength -0.0390484550  
## LongShots .   
## Aggression -0.0003185207  
## Interceptions -0.0073082446  
## Positioning 0.0171186055  
## Vision .   
## Penalties .   
## Composure -0.0412936931  
## Marking .   
## StandingTackle .   
## SlidingTackle 0.0006079635  
## GKDiving -0.0396288840  
## GKHandling -0.0155899155  
## GKKicking -0.0396723633  
## GKPositioning -0.0531417057  
## GKReflexes -0.0214541141

coef(lasso\_mod\_fifa, lasso\_mod\_fifa$lambda.1se)

## 118 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -75.9033307  
## ContinentAfrica .   
## ContinentAsia .   
## ContinentEurope .   
## ContinentNorth America .   
## ContinentOceania .   
## ContinentSouth America .   
## Age .   
## Overall 0.7307959  
## Potential 0.2887261  
## Club .   
## ClubArsenal 13.6957950  
## ClubAS Monaco .   
## ClubAtlético Madrid .   
## ClubBorussia Dortmund .   
## ClubBurnley .   
## ClubCardiff City .   
## ClubCD Leganés .   
## ClubChelsea 22.4012352  
## ClubEintracht Frankfurt .   
## ClubEmpoli .   
## ClubEverton 4.8901128  
## ClubFC Barcelona 59.7974119  
## ClubFortuna Düsseldorf .   
## ClubFrosinone .   
## ClubLiverpool 15.3064459  
## ClubManchester City 54.1538845  
## ClubManchester United 47.5254351  
## ClubNewcastle United .   
## ClubRayo Vallecano .   
## ClubRC Celta .   
## ClubReal Madrid 82.6532319  
## ClubSouthampton .   
## ClubTottenham Hotspur .   
## ClubTSG 1899 Hoffenheim .   
## ClubValencia CF .   
## ClubWolverhampton Wanderers .   
## ClubOther -9.7955366  
## Preferred.Foot .   
## Preferred.FootLeft .   
## Preferred.FootRight .   
## International.Reputation 22.6075431  
## Weak.Foot .   
## Skill.Moves .   
## Work.Rate .   
## Work.RateHigh/ High .   
## Work.RateHigh/ Low .   
## Work.RateHigh/ Medium .   
## Work.RateLow/ High .   
## Work.RateLow/ Low .   
## Work.RateLow/ Medium .   
## Work.RateMedium/ High .   
## Work.RateMedium/ Low .   
## Work.RateMedium/ Medium .   
## Position .   
## PositionCAM .   
## PositionCB .   
## PositionCDM .   
## PositionCF .   
## PositionCM .   
## PositionGK .   
## PositionLAM .   
## PositionLB .   
## PositionLCB .   
## PositionLCM .   
## PositionLDM .   
## PositionLF .   
## PositionLM .   
## PositionLS .   
## PositionLW .   
## PositionLWB .   
## PositionRAM .   
## PositionRB .   
## PositionRCB .   
## PositionRCM .   
## PositionRDM .   
## PositionRF .   
## PositionRM .   
## PositionRS .   
## PositionRW .   
## PositionRWB .   
## PositionST .   
## Height .   
## Weight .   
## Crossing .   
## Finishing .   
## HeadingAccuracy .   
## ShortPassing .   
## Volleys .   
## Dribbling .   
## Curve .   
## FKAccuracy .   
## LongPassing .   
## BallControl .   
## Acceleration .   
## SprintSpeed .   
## Agility .   
## Reactions .   
## Balance .   
## ShotPower .   
## Jumping .   
## Stamina .   
## Strength .   
## LongShots .   
## Aggression .   
## Interceptions .   
## Positioning .   
## Vision .   
## Penalties .   
## Composure .   
## Marking .   
## StandingTackle .   
## SlidingTackle .   
## GKDiving .   
## GKHandling .   
## GKKicking .   
## GKPositioning .   
## GKReflexes .

plot(lasso\_mod\_fifa)

