FIFA Data Cleaning

Akshit Jain

3/28/2020

```
fifa20 <- fread("~/Desktop/Northeastern-University/SML/FIFA-Player-Assessment-Model-and-Analytics/Datas
fifa20 <- as_tibble(fifa20)
```

Data cleaning

```
# remove columns that are not required for modeling
updated_fifa20 <- fifa20 %>% select(-player_url, -long_name, -dob, -real_face, -player_tags,
                                    -loaned_from, -joined, -player_positions, -contract_valid_until,
                                    -nation_position, -nation_jersey_number, -player_traits, -gk_diving
                                    -gk_handling, -gk_kicking, -gk_reflexes, -gk_speed, -gk_positioning
                                    -goalkeeping_diving, -goalkeeping_handling, -goalkeeping_kicking,
                                    -goalkeeping_positioning, -goalkeeping_reflexes,
                                    -ls, -st, -rs, -lw, -lf, -cf, -rf, -rw, -lam, -cam, -ram,
                                    -lm, -lcm, -cm, -rcm, -rm, -lwb, -ldm, -cdm, -rdm, -rwb,
                                    -lb, -lcb, -cb, -rcb, -rb)

# remove oberservations that have missing values (NOT processing Goalkeepers)
clean_fifa20 <- na.omit(updated_fifa20)
clean_fifa20
```

```
## # A tibble: 15,077 x 55
##    sofifa_id short_name   age height_cm weight_kg nationality club  overall
##        <int> <chr>      <int>     <int>     <int> <chr>       <chr>   <int>
##  1    158023 L. Messi      32       170        72 Argentina   FC B~      94
##  2     20801 Cristiano~    34       187        83 Portugal    Juve~      93
##  3    190871 Neymar Jr     27       175        68 Brazil      Pari~      92
##  4    183277 E. Hazard     28       175        74 Belgium     Real~      91
##  5    192985 K. De Bru~    28       181        70 Belgium     Manc~      91
##  6    203376 V. van Di~    27       193        92 Netherlands Live~      90
##  7    177003 L. Modrić     33       172        66 Croatia     Real~      90
##  8    209331 M. Salah      27       175        71 Egypt       Live~      90
##  9    231747 K. Mbappé     20       178        73 France      Pari~      89
## 10    201024 K. Koulib~    28       187        89 Senegal     Napo~      89
## # ... with 15,067 more rows, and 47 more variables: potential <int>,
## #   value_eur <int>, wage_eur <int>, preferred_foot <chr>,
## #   international_reputation <int>, weak_foot <int>, skill_moves <int>,
## #   work_rate <chr>, body_type <chr>, release_clause_eur <int>,
## #   team_position <chr>, team_jersey_number <int>, pace <int>,
## #   shooting <int>, passing <int>, dribbling <int>, defending <int>,
## #   physic <int>, attacking_crossing <int>, attacking_finishing <int>,
## #   attacking_heading_accuracy <int>, attacking_short_passing <int>,
## #   attacking_volleys <int>, skill_dribbling <int>, skill_curve <int>,
## #   skill_fk_accuracy <int>, skill_long_passing <int>,
## #   skill_ball_control <int>, movement_acceleration <int>,
```

```
## #   movement_sprint_speed <int>, movement_agility <int>,
## #   movement_reactions <int>, movement_balance <int>,
## #   power_shot_power <int>, power_jumping <int>, power_stamina <int>,
## #   power_strength <int>, power_long_shots <int>,
## #   mentality_aggression <int>, mentality_interceptions <int>,
## #   mentality_positioning <int>, mentality_vision <int>,
## #   mentality_penalties <int>, mentality_composure <int>,
## #   defending_marking <int>, defending_standing_tackle <int>,
## #   defending_sliding_tackle <int>
```

Feature Selection: Forward Stepwise Selection

```
df <- sample_n(clean_fifa20, nrow(clean_fifa20))
df <- df %>% select(-short_name, -nationality, -club, -body_type, -team_jersey_number, -team_position)

library(leaps)
regfit.fwd=regsubsets(value_eur~.,df, method="forward")
reg.summary <- summary(regfit.fwd)
coef(regfit.fwd, 8)
```

```
##              (Intercept)                      age                  overall
##            -7.374862e+05           -1.498995e+04             3.074348e+04
##                 potential                wage_eur international_reputation
##            -2.130298e+04            3.320752e+00             5.062111e+05
##       release_clause_eur           power_stamina defending_sliding_tackle
##             4.935300e-01            4.779582e+03            -2.253148e+03
```

```
par(mfrow=c(2,2))

plot(reg.summary$rss,xlab="Number of Variables",ylab="RSS",type="l")

plot(reg.summary$adjr2,xlab="Number of Variables",ylab="Adjusted RSq",type="l")
which.max(reg.summary$adjr2)
```

```
## [1] 8
```

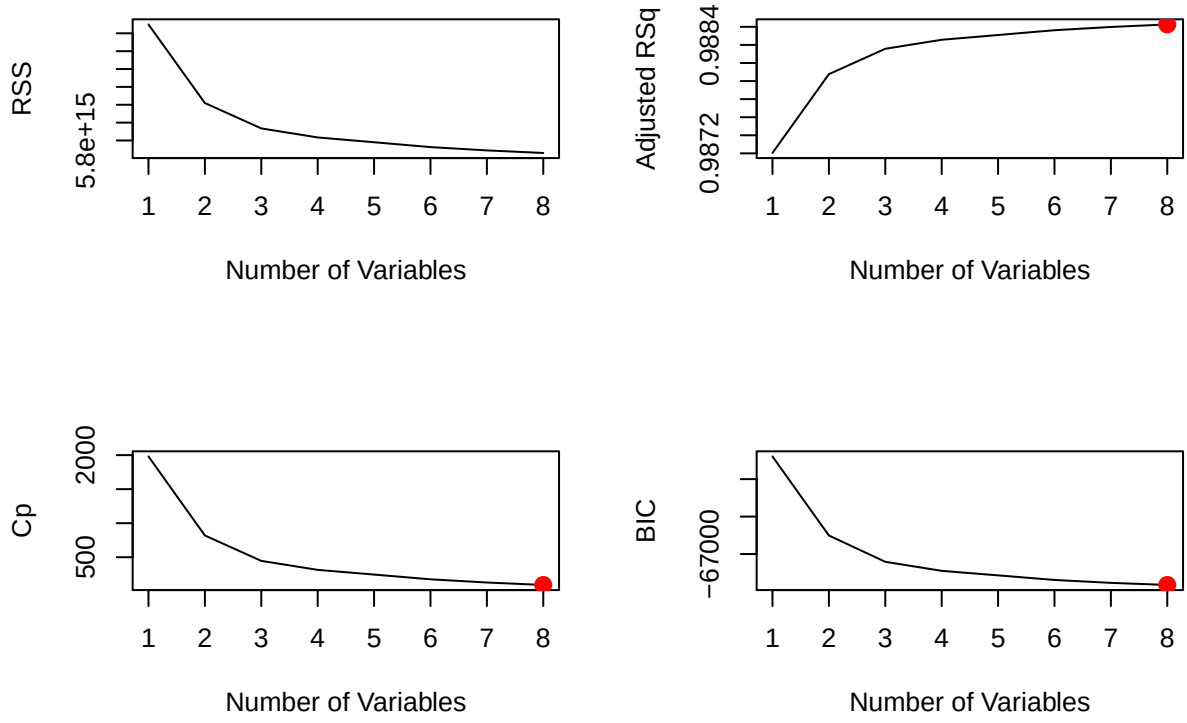
```
points(8,reg.summary$adjr2[8], col="red",cex=2,pch=20)

plot(reg.summary$cp,xlab="Number of Variables",ylab="Cp",type='l')
which.min(reg.summary$cp)
```

```
## [1] 8
```

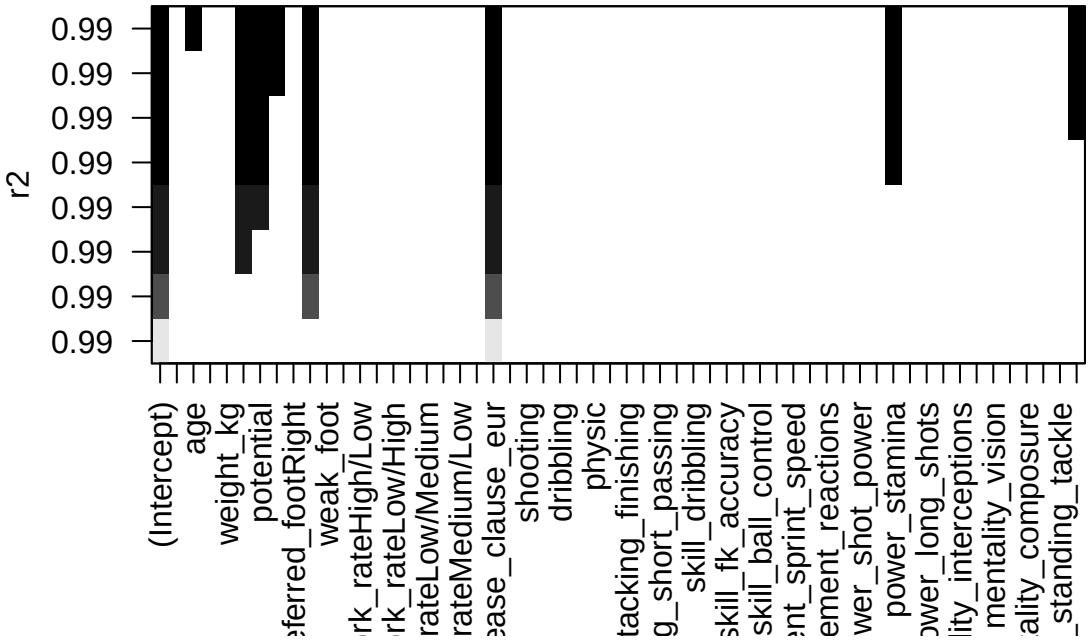
```
points(8,reg.summary$cp[8],col="red",cex=2,pch=20)

which.min(reg.summary$bic)
```

```
## [1] 8
```

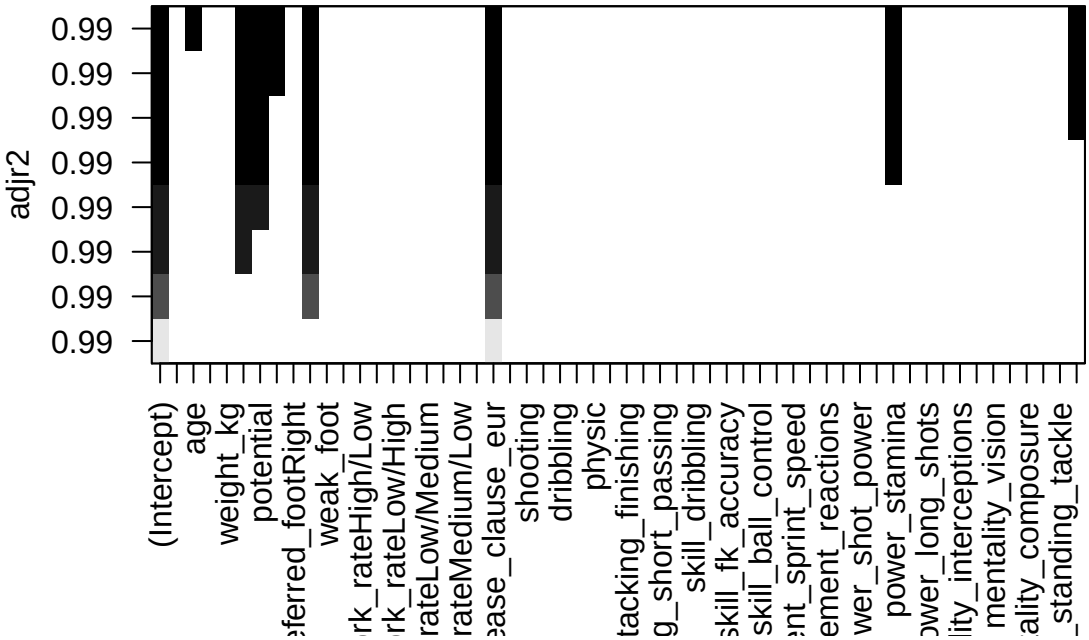
```
plot(reg.summary$bic,xlab="Number of Variables",ylab="BIC",type='l')
points(8,reg.summary$bic[8],col="red",cex=2,pch=20)
```



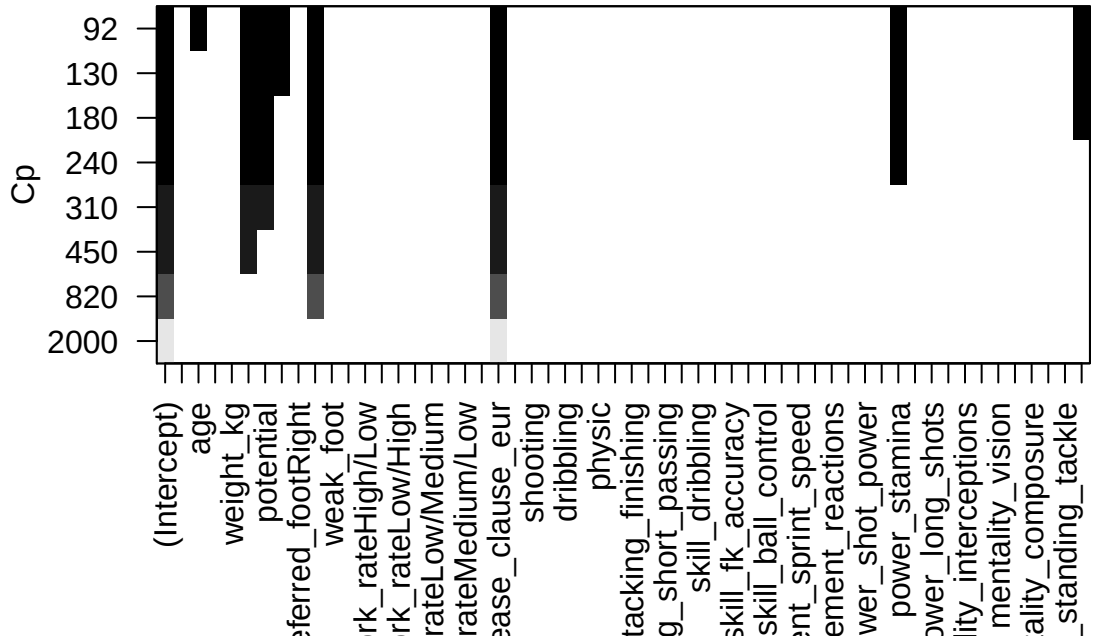

```
plot(regfit.fwd,scale="r2")
```



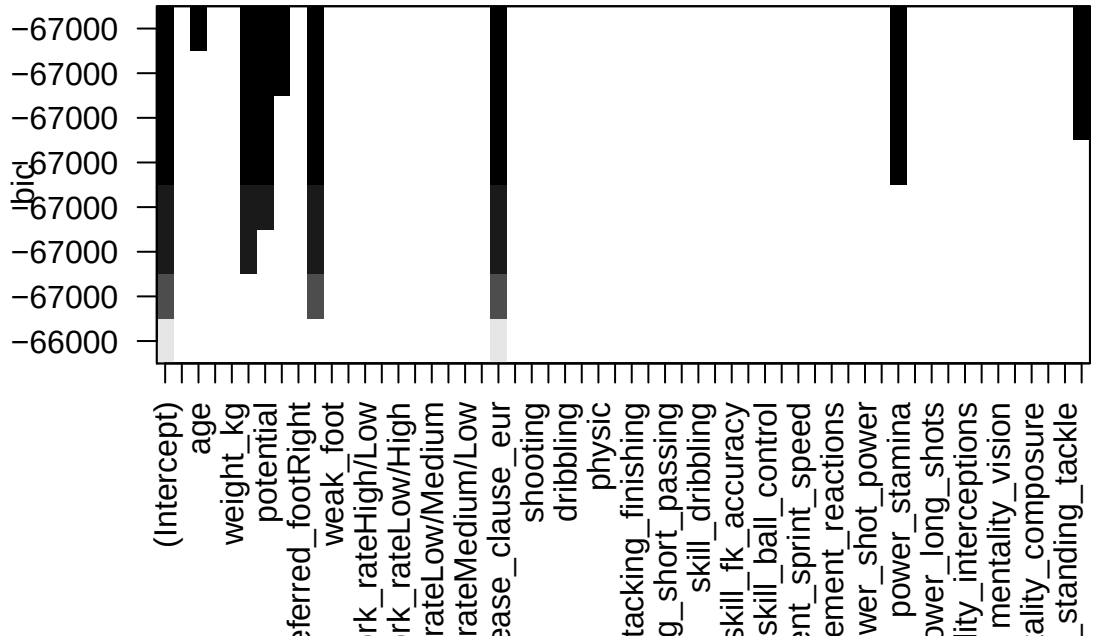

```
plot(regfit.fwd,scale="adjr2")
```




```
plot(regfit.fwd,scale="Cp")
```




```
plot(regfit.fwd,scale="bic")
```


−67000
−67000
−67000
−67000
−67000
−67000
−66000
bic
(Intercept)
age
weight_kg
potential
eferred_footRight
weak_foot
rk_rateHigh/Low
rk_rateLow/High
rateLow/Medium
rateMedium/Low
ease_clause_eur
shooting
dribbling
physic
tacking_finishing
g_short_passing
skill_dribbling
skill_fk_accuracy
skill_ball_control
nt_sprint_speed
ement_reactions
wer_shot_power
power_stamina
ower_long_shots
lity_interceptions
mentality_vision
ality_composure
_standing_tackle