HW 7: NHL stats

Stats and sports class

Fall 2019

Preliminary notes for doing HW

1. All files should be knit and compiled using R Markdown. Knit early and often! I do not recommend waiting until the end of the HW to knit.
2. All questions should be answered completely, and, wherever applicable, code should be included.
3. If you work with a partner or group, please write the names of your teammates.
4. Copying and pasting of code is a violation of the Skidmore honor code

Homework questions

Part I: Readings

1. Read the summary model by the Evolving Wild twins:

https://rpubs.com/evolvingwild/395136/

Describe five unique hockey features that were implemented in their model. That is, look through their code, and highlight various ways that hockey-specific knowledge changed how they approached the problem.

2. Compare the three variable importance plots. Which variables were more important during even-strength play? Which were more important (relatively speaking) when a team was shorthanded or at uneven strength?

Part II: Implementation

We can access recent shot data here:

```r
library(RCurl); library(tidyverse)
gitURL<- "https://raw.githubusercontent.com/statsbylopez/StatsSports/master/Data/pbp_data_hockey.rds"
pbp_data <- readRDS(gzcon(url(gitURL)))
names(pbp_data)
dim(pbp_data)
```

Question 1

Identify which NHL players have passes who have led to the highest expected goals. Note – the passing player is `event_player_2`. Hint: Connor McDavid should finish first.

Question 2

McDavid is credited with 14.3 expected goals off of his passes. In reality, he finished the 20182019 season with 75 assists. How come McDavid has so few expected assists in this data set? Think carefully about how NHL data is collected. If you aren't sure, give a guess! But also explore the data to see if you pick anything up.

Question 3

A coach wants to know if players can consistently overperform the average finishing of an NHL players. Make a plot of each players' goals above expectation in the 2018 season (+ or -) versus his goals above expectation in the 2019 data set. As a reminder, below is code from class which gets you started. You'll need to make two new variables that correspond to goals above or below expectation in each season, and compare them graphically.

```
season_2018 <- pbp_data %>%
  filter(season == 20172018) %>%
  group_by(event_player_1, season) %>%
  summarise(n_goals_18 = sum(event_type == "GOAL"),
            n_xGs_18 = sum(shot_prob),
            n_shots_18 = n()) %>%
  filter(n_shots_18 >= 100) %>%
  select(-season)

season_2019 <- pbp_data %>%
  filter(season == 20182019) %>%
  group_by(event_player_1, season) %>%
  summarise(n_goals_19 = sum(event_type == "GOAL"),
            n_xGs_19 = sum(shot_prob),
            n_shots_19 = n()) %>%
  filter(n_shots_19 >= 100) %>%
  select(-season)

season_combine <- season_2018 %>% inner_join(season_2019)
```

Question 4

Is there any positive link between the two new variables you created for performance above expectation? Use a smoothed trend curve and/or linear regression to make your conclusion.

Question 5

Put your conclusion above into one to two sentences for a coach to understand.

Question 6

Identify the one player who was 10 goals better than expectation in 2018 and 10 goals better than expectation in 2019.