Pittsburgh Penguins Project

ok title for drafty draft, but you can do much better for final paper. Feature the main question, or main result, of your work, in a short title.

Minyue Fan, Steve Kim, Kexiong Shen, Linda Yang

April 2021

1 Abstract

After being drafted, a player can train in different leagues before they they can play for the famous National Hockey League (NHL). Such training through different leagues is called a developmental path. However, the transition does not happen as often in some league s as others. We address the question of whether different developmental paths for hockey players affect their success. For this specific project, we evaluate their success by whether the player transitions into the NHL and by the number of games they play ~~he plays~~ in the NHL. We examine data from season 2000 to 2020 for players from various leagues. From exploratory data analysis, it appears that more players from the NCAA succeed than those from other leagues. However, the NCAA is competitive to get into in the first place, and the player is naturally better than those from other leagues. Therefore, we perform a causal inference analysis using propensity score rating and Bayesian Additive Regression Tree to isolate the treatment effect by the league, and evaluate whether taking different developmental path affects their success. From the propensity score analysis, we found that the treatment effect is real and different paths can affect a player's success.

this is great. please say what the treatment is here.

2 Introduction

Hockey players play in various developmental leagues as part of their professional development before entering the National Hockey League (NHL). There are multiple traditional development paths, determined by the leagues participated in, that hockey prospects can take to get to the NHL. Some of the notable ones include going from US High School Leagues such as the United States Hockey League (USHL) to College, and then to the NHL, and for International players, going from the Kontinental Hockey League (KHL) to the NHL. Unlike some other major professional sports in North America such as football and basketball, where most prospects enter the professional leagues right after they are drafted, most hockey prospects do not immediately go to the NHL when they become eligible for the NHL draft, even if they are drafted (**Need Reference**). Instead, they stay in or move to different developmental leagues before playing professionally for an NHL team.

good, thanks

Which developmental league to play in is an important career decision to make for hockey prospects because it could potentially impact their chances at getting into and performing well in the NHL. Therefore, the effect of playing in different developmental leagues on prospects future career in the NHL is crucial information when making this decision. Professional hockey teams in the NHL also find this information valuable because when they are making draft decisions, they usually know which developmental leagues the prospects are going to play in next season and knowing this information could impact the teams' decisions. Our study investigates the effect ~~if~~ of development paths in projecting a prospect's chances at having success at the NHL level. Our research question is:

this is also a crucial question for NHL recruiters / scouts /GM's, isn't it?

- How do players' development paths impact their performance and success in the NHL?

One obstacle when attempting to examine this effect is that ~~players taking different~~ development path~~s~~ a player takes is correlated with their quality level~~s~~. For example, one common observation is that American players who take the NCAA path have higher success rates than other paths, but the NCAA player pool is already better than most of the other developmental leagues in terms of quality. Our study, then, intends to establish a causal relationship between taking any particular development path and players future success in the NHL.

(a reference for this would be really good here)

The client for this project is Sam Ventura, who is the Director of Hockey Operations Hockey Research at Pittsburgh Penguins. His team wants to utilize the findings of this project to learn more about the development of prospective hockey players and to potentially guide drafting decisions in the NHL draft.

3 Data

The data for this project is web scrap~~p~~ed from Elite Prospects's website [1]. We maintain two datasets: one is players' biographical information, the other is their performance in every season from 2001-2020. For the biography datasets, 15,786 players were included. ~~The following~~ Table 1 lists the 7 variables that were kept~~.~~:

Variable	Definition
Player	Player name
Position	player's position, which includes C (captain), LW (Left Wing), RW (Right Wing),LD(Left Defense), RD(Right Defense) and G(Goalie). Note that a player can be a captain and at other positions at the same time.
DateofBirth	In MM/DD/YYYY format.
Height	player's height in both inches and meters.
Weight	players' weight in both pounds and kilograms.
Nation	Nationality of player. Dual citizen-ships are included
Shoots	Handedness of the player. Either L (left) or R (right).

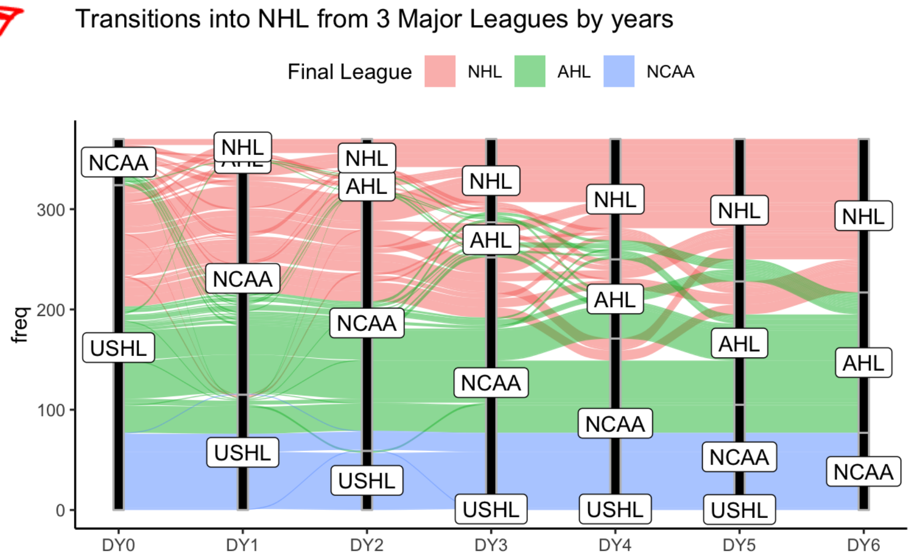
Table 1: blah blah blah

The performance dataset has 266,326 entries with information on 15,220 players. ~~The following~~ Table 2 lists the 19 variables that were kept~~:~~.

Variable	Definition
Player	Player's Name
Season	The season in which the player played. Example: 2001-02.
Team	Player's team during the season
League	Player's league during the season
Games	Number of games played during the season
Goals	Number of goals during the season
Assists	Number of times player enables the goal to a scoring teammate. There is a maximum of two assists per goal
TotalPoints	Total points achieved during the season
PenaltyMinutes	Total minutes the player was on penalty
PlusMinus	Evaluation on performance relative to other teammates

Table 2: blah blah blah

There should be a nice paragraph here discussing the very interesting Figure 1.



I'm not understanding this figure as well as I should. It may help to rotate it 90 degrees and make it larger. It would also help a lot to explain all of the white-boxed labels. For example, why does NCAA appear in several places ir the AHL stream? Etc.

Figure 1: We plot player's path from their first draft year (DY0) to 5 years onward. Notet the reason for a high transition rate from ANL H to NHL is that the AHL serves as the top minor professional league below the NHL in North America. If a player is not playing well, he will be sent from the NHL to the AHL. Then, if his play improves, he will be called back up from the AHL to the NHL.

Players' Transition into NHL in 5 years

move or delete

4 Methods

4.1 Propensity Scores

Hockey ~~NHL~~ players are eligible to be drafted into the NHL at age 18. We are interested in understanding the causal effects of developmental path during this time of a young professionals career on his success in the NHL. We decided to look at players that were in the USHL, a popular developmental league, at the age of 18. We then assessed whether they go into the NCAA, another common developmental league, the following year or stay in the USHL. We are particularly interested in These types of players and the different paths that they can pursue before making it to the NHL. Thus, we looked specifically at the impact of developmental leagues at age 19 on a player's success in the NHL.

this is fine, but explain why it's a good idea to focus on the USHL.

bad transition? editing problem?

Since The data that we have on players and their career paths are observational, we decided to use propensity score matching to assess the direct causal effects. The treatment ~~effect~~ was being in the NCAA at age 19 while the control ~~effect~~ was being in the USHL at age 19. We used success metrics when the player is at age 18 such as assists, goals, plus minus, penalty minutes to predict ~~the~~ treatment ~~effect~~ at age 19. This allows us to remove any selection effects of the treatment and control. The probability of being in the NCAA at age 19 served as our propensity scores. Players were then matched by common scores and those who were unmatched were discarded from the analysis. To ensure we are matching defense players together and offense players together, we performed this matching exercise separately for each type of position.

great

To ensure that our matching was done accurately, we ran the logistic regression model in the matched data with treatment as the response and assessed the coefficient estimates. If our estimates are close to zero at some statistically significant degree, then we can confirm that matching was done correctly and selection bias was removed.

estimate of what? probability? log-odds? etc?

Once selection effects were removed, we modeled the probability of playing in 10 or more games in the NHL on the matched data based on player success metric at age 18 and compared the coefficient in front of the treatment effect (NCAA) for the same model on the unmatched data.

4.2 Bayesian Additive Regression Trees

To investigate the causal effect of developmental paths further and model how the developmental paths affected the potential success at NHL further, we aimed to investigate the treatment effect (of taking different developmental paths) on the success at NHL for the hockey prospects by implementing Bayesian additive regression trees - BART [2]. Similar to the propensity scores, we looked at the players that became just eligible to be drafted (draft year 1, at age 18). We also filtered for those prospect players whom we know that did not play on the NHL in the following season.

We chose BART because it provides precise modeling of the response, as we can get more control for the confounding variables than with the usual parametric model. From the publicly available data, we calculated the response variable being the average NHL games played in a season for each player's career span. Using the bartMachine package in Rstudio, we fitted BART model, treating the developmental league as the treatment, and all other variables as the confounding variables. We extracted the posterior distribution from the fitted model to make predictions on how many NHL games on average the prospect would play, given the stat line and the biographical information, as well as the information on the developmental league.

citation

this is fine.

an intermediate step would be try the same model with multiple regression or poisson regression (predict games in NHL from NCAA treatment, controlling for all other available confounders). you don't have to do this for the project, but it is good to know that you could do much the same thing (and should expect somewhat similar results) with multiple regression.

5 Results

5.1 Propensity Scores

After modeling the treatment effect at age 19 based on predictors at age 18, I assessed its coefficient estimates on the matched data. For all of the coefficient estimates, we do not observe enough evidence to reject the null hypothesis that the coefficient estimates are zero. Thus, while we observe non-zero coefficients, we cannot conclude that they are non-zero based on the large SE's of the estimates.

?

After matching players that have forward positions, I fit a logistic regression model on the matched data set. The logistic regression model predicts the probability of a player playing 10 or more games in the NHL. The predictors are assists per game, goals per game, league at age 19 (NCAA or USHL), plus minus per game, penalty minutes per game, weight, and height. Unfortunately, league at age 19 was an insignificant predictor for forward positions in both the matched data and the unmatched data.

?

where above did you say you restricted the data to just these two developmental leagues? I though it was NCAA vs all others...

please refer reader to page in appx where the details are

On the other hand, when we fit the same logistic model on the defense players, we see that league at age 19 is a significant predictor in the matched data set and the unmatched data set. In the matched data set we see that the coefficient estimate for league increases compared to the estimate in the unmatched data set. With the removal of selection bias, the effects of developmental path and league increases.

5.2 Bayesian Additive Regression Trees

how did you select the subset?

Due to the limit~~ing~~ed computing power we have, only a subset of the data containing 1,017 lines was fitted at this point. Although the BART model converges, it shows high variance. The plot of posterior error variance (top left plot) shows a high variance of 1,800. Percent acceptance (top right graph) shows only about 60 percent trees were accepted. Both of the average number of nodes and average depth of trees show a small fitted tree, indicating not much information is retained in the model.

Does this mean the BART model didn't yield any useful results? If so, then you should say so.

Even if the results aren't conclusive it would be useful to try to summarize what the BART model was able to capture

But from the discussion it sounds like you're just not done with BART yet. That's fine - looking forward to seeing it in the final paper.

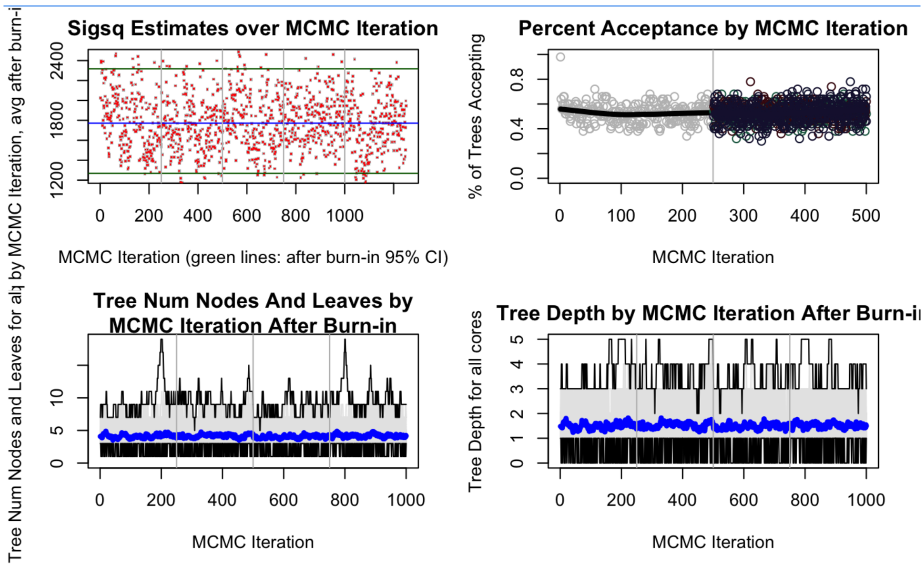


Figure 2: BART convergence plots

6 Discussion

Looking at the results of the propensity score results, the fact that we have non-significant coefficient estimates after modeling the treatment effect at age 19 based on predictors at age 18 suggests that the matching process was successful - we can conclude that the prospects who went to play in the NCAA was matched successfully with the prospects with similar stats and background who went to play in the USHL. This allows us to analyze the ~~ca~~ causal effects of taking USHL instead of NCAA.

We arrive at different conclusions for ~~the~~ forwards, vs. ~~and the~~ defensemen. Whether the prospects player plays in the USHL or NCAA in the following season after entering the draft year was not significant for the forwards in terms of eventually making it to the NHL. On the contrary, the choice of which development path (USHL vs. NCAA) resulted in statistically significant increase in the likelihood of playing in the NHL for the defensemen. It is interesting to note that, perhaps aligning with the common conception, playing at the NCAA (even at such an early stage of the developmental phase) does lead to eventual appearance at NHL for the defensemen. However, results suggest that participating in the NCAA does not particularly guarantee higher chances of reaching the NCAA, compared to playing in the USHL, for the forwards; forwards prospects, at the very early stage of the developmental phase, would not necessarily have to rush in to the NCAA to guarantee some success at the NHL.

indent paragraph

rewrite more clearly: choosing the NCAA developmental path, rather than USHL,

with probability 1? if not, choose a better word ("more likely" etc.)

We suspect that such difference may occur from the different nature of the two positions; for the attacking players, there may be skills (shooting technique, positioning, pass vision) that can be improved throughout the developmental phase. Keeping in mind that we investigated the early stages of the developmental phase (draft year 0 to draft year), in future analysis we could investigate how the same set of the forwards fared in the latter stages of the developmental path - for example by comparing those who kept staying at the USHL against those who moved to the NCAA later on in their developmental phase.

5?

We're still in the process of refining and processing the BART results for further analysis and discussion.

great! Looking forward to it

References

good start on references; be sure to put the references here, as well as the citations in the main part of the paper, in ASA format.

[1] *Elite Prospects*. URL: `http://eliteprospects.com/`.

[2] Jennifer L. Hill. "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240. DOI: `10.1198/jcgs.2010.08162`.

Technical Appendix

Minyue Fan, Steve Kim, Kexiong Shen, Linda Yang

good start. remember to

(a) put enough english here that someone can follow what you are doing and why, and what the results are, for each step in your analyses

(b) references from main report to specific pages here for details lef out of the main report.

Propensity Score Method

```
nhl = read.csv("merged_data_updated.csv")
nhl$age_actual = nhl$Season - nhl$BirthYear
nhl$Goals = as.numeric(nhl$Goals)
nhl$PlusMinus = as.numeric(nhl$PlusMinus)
nhl$PenaltyMin = as.numeric(nhl$PenaltyMinutes)
nhl$GamesPlayed = as.numeric(nhl$Games)
nhl$Assists = as.numeric(nhl$Assists)
library(tidyr)
library(tidyverse)
nhl = nhl %>% separate(Season, c("Season", "other_season"), sep = "-")
nhl$Season = as.numeric(nhl$Season)
nhl = nhl[which(nhl$Season != ""), ]
updated_nhl = nhl
# predictors : Scaled per game , goals per game, plus minus per game , penalty
# per game
updated_nhl$goals_pergame = updated_nhl$Goals/updated_nhl$GamesPlayed
updated_nhl$plusminus_pergame = updated_nhl$PlusMinus/updated_nhl$GamesPlayed
updated_nhl$PenaltyMin_pergame = updated_nhl$PenaltyMin/updated_nhl$GamesPlayed
updated_nhl$Assists_pergame = updated_nhl$Assists/updated_nhl$GamesPlayed
```

filter for USHL players of age 18

```
players_ushl_18 = updated_nhl[which(updated_nhl$age_actual == 18 & updated_nhl$League ==
    "USHL"), ]$Player
updated_nhl = updated_nhl[which(updated_nhl$Player %in% players_ushl_18), ]
updated_nhl$plusminus_pergame_draft_year = updated_nhl$plusminus_pergame
updated_nhl[which(updated_nhl$age_actual != 18), ]$plusminus_pergame_draft_year = 0
updated_nhl[which(updated_nhl$age_actual == 18 & updated_nhl$League != "USHL"), ]$plusminus_pergame_dra
updated_nhl$goals_pergame_draft_year = updated_nhl$goals_pergame
updated_nhl[which(updated_nhl$age_actual != 18), ]$goals_pergame_draft_year = 0
updated_nhl[which(updated_nhl$age_actual == 18 & updated_nhl$League != "USHL"), ]$goals_pergame_draft_y
updated_nhl$penalty_pergame_draft_year = updated_nhl$PenaltyMin_pergame
updated_nhl[which(updated_nhl$age_actual != 18), ]$penalty_pergame_draft_year = 0
updated_nhl[which(updated_nhl$age_actual == 18 & updated_nhl$League != "USHL"), ]$penalty_pergame_draft_
updated_nhl$assists_pergame_draft_year = updated_nhl$Assists_pergame
updated_nhl[which(updated_nhl$age_actual != 18), ]$assists_pergame_draft_year = 0
updated_nhl[which(updated_nhl$age_actual == 18 & updated_nhl$League != "USHL"), ]$assists_pergame_draft_
updated_nhl$GamesPlayed_after_in_NHL = updated_nhl$GamesPlayed
updated_nhl[which(updated_nhl$League != "NHL"), ]$GamesPlayed_after_in_NHL = 0
```

Players of age 19

```r
updated_nhl$league_19 = updated_nhl$League
updated_nhl[which(updated_nhl$age_actual != 19), ]$league_19 = ""
updated_nhl = updated_nhl %>% group_by(Player) %>% mutate(total_games_in_nhl = sum(GamesPlayed_after_in_
    goals_pergame_1 = sum(goals_pergame_draft_year), plusminus_pergame_1 = sum(plusminus_pergame_draft_
    penaltymin_pergame_1 = sum(penalty_pergame_draft_year), assists_pergame_1 = sum(assists_pergame_dra
    c = max(league_19))

updated_nhl[which(updated_nhl$Player == "A.J. Drobot"), ]
```

```
## # A tibble: 6 x 40
## # Groups:   Player [1]
##        X Player Season other_season Team  League Games Goals Assists TotalPoints
##    <int> <chr>   <dbl> <chr>        <chr> <chr>  <chr> <dbl>   <dbl> <chr>
## 1 129169 A.J. ~   2014 <NA>         "Sio~ USHL   1         0       0 0
## 2 129170 A.J. ~   2015 <NA>         "Aus~ NAHL   43        6       8 14
## 3 129171 A.J. ~   2016 <NA>         "Aus~ NAHL   51       14      16 30
## 4 129172 A.J. ~   2016 <NA>         "Far~ USHL   2         0       0 0
## 5 129173 A.J. ~   2017 <NA>         "Far~ USHL   57       14      11 25
## 6 129174 A.J. ~   2018 <NA>         "Far~ USHL   57        9       8 17
## # ... with 30 more variables: PenaltyMinutes <chr>, PlusMinus <dbl>,
## #   Position <chr>, DateofBirth <chr>, Height <chr>, Weight <chr>,
## #   Nation <chr>, Shoots <chr>, BirthYear <int>, Age <int>, Performance <dbl>,
## #   age_actual <int>, PenaltyMin <dbl>, GamesPlayed <dbl>, goals_pergame <dbl>,
## #   plusminus_pergame <dbl>, PenaltyMin_pergame <dbl>, Assists_pergame <dbl>,
## #   plusminus_pergame_draft_year <dbl>, goals_pergame_draft_year <dbl>,
## #   penalty_pergame_draft_year <dbl>, assists_pergame_draft_year <dbl>,
## #   GamesPlayed_after_in_NHL <dbl>, league_19 <chr>, total_games_in_nhl <dbl>,
## #   goals_pergame_1 <dbl>, plusminus_pergame_1 <dbl>,
## #   penaltymin_pergame_1 <dbl>, assists_pergame_1 <dbl>, c <chr>
```

```r
updated_nhl$position_new = "backward"
updated_nhl[which(updated_nhl$Position %in% c("C", "F", "LW", "RW")), ]$position_new = "forward"
updated_nhl$more_than_10_games = 0
updated_nhl[which(updated_nhl$total_games_in_nhl > 10), ]$more_than_10_games = 1
updated_nhl = updated_nhl[which(updated_nhl$age_actual == 18), ]
updated_nhl = updated_nhl[which(updated_nhl$c %in% c("USHL", "NCAA")), ]
updated_nhl$temp = 1
updated_nhl = updated_nhl %>% group_by(Player) %>% mutate(temp_2 = cumsum(temp))
updated_nhl = updated_nhl[which(updated_nhl$temp_2 == 1), ]
updated_nhl[order(updated_nhl$Player, updated_nhl$Season), c("Player", "Season",
    "age_actual", "League", "goals_pergame_1", "penaltymin_pergame_1", "c", "PenaltyMin",
    "GamesPlayed")]
```

```
## # A tibble: 996 x 9
## # Groups:   Player [996]
##    Player Season age_actual League goals_pergame_1 penaltymin_perg~ c
##    <chr>   <dbl>      <int> <chr>            <dbl>            <dbl> <chr>
##  1 A.J. ~   2000         18 USHL            0.225             0.4   USHL
##  2 A.J. ~   2016         18 NAHL            0                 0     USHL
##  3 A.J. ~   1997         18 USHL            0.0294            2.35  USHL
##  4 A.J. ~   2003         18 USHL            0.0769            1.13  USHL
##  5 Aaron~   2009         18 USHL            0                0.364 USHL
##  6 Aaron~   1998         18 USHL            0.179            0.625 USHL
##  7 Aaron~   2008         18 USHL            0.25             0.5   NCAA
##  8 Aaron~   2000         18 QMJHL           0.0690           0.690 USHL
##  9 Aaron~   2006         18 USHL            0.0769           0.615 NCAA
## 10 Aaron~   2010         18 USHL            0.2              1.2   USHL
## # ... with 986 more rows, and 2 more variables: PenaltyMin <dbl>,
## #   GamesPlayed <dbl>
```

```
updated_nhl = updated_nhl %>% separate(Height, c("blah", "blah1", "height", "blah3"),
    sep = " ")
updated_nhl$height = as.numeric(updated_nhl$height)
updated_nhl = updated_nhl %>% separate(Weight, c("Weight", "blah1h", "blahh2", "blahh3",
    "blahh4"), sep = " ")
updated_nhl$Weight = as.numeric(updated_nhl$Weight)
```

Set the response variable

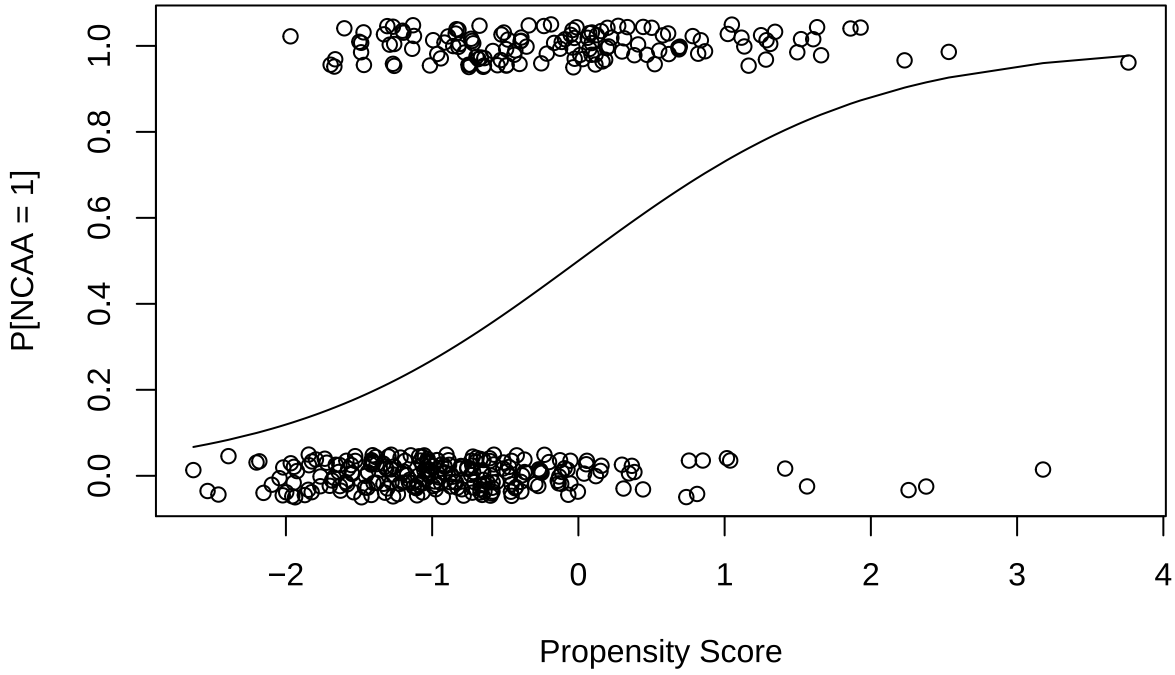
```
updated_nhl$y = 0
updated_nhl[which(updated_nhl$c == "NCAA"), ]$y = 1
updated_nhl_forward = updated_nhl[which(updated_nhl$position_new == "forward"), ]
updated_nhl_backward = updated_nhl[which(updated_nhl$position_new == "backward"),
    ]
```

Match the simialr players

```
library(Matching)
library(arm)
mod1 = glm(y ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 + goals_pergame_1 +
    Weight + height, data = updated_nhl_forward, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = y ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height, family = "binomial",
##     data = updated_nhl_forward)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5370  -0.8856  -0.6471   1.0974   2.0490
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -10.07424    4.11113  -2.450  0.01427 *
## penaltymin_pergame_1  -0.11710    0.14869  -0.788  0.43097
## plusminus_pergame_1    0.36856    0.47344   0.778  0.43629
## assists_pergame_1      2.28537    0.85689   2.667  0.00765 **
## goals_pergame_1        3.39375    1.08122   3.139  0.00170 **
## Weight                -0.02188    0.01219  -1.794  0.07278 .
## height                 0.06758    0.02925   2.311  0.02086 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 495.41  on 375  degrees of freedom
## Residual deviance: 434.33  on 369  degrees of freedom
##   (131 observations deleted due to missingness)
## AIC: 448.33
##
## Number of Fisher Scoring iterations: 4
```

```
complete_Cases = updated_nhl_forward[complete.cases(updated_nhl_forward[, c("y",
    "assists_pergame_1", "plusminus_pergame_1", "goals_pergame_1", "penaltymin_pergame_1",
    "Weight", "height")]), ]
p.scores = predict(mod1, type = "link")
plot(p.scores, jitter(updated_nhl_forward[complete.cases(updated_nhl_forward[, c("y",
    "assists_pergame_1", "plusminus_pergame_1", "goals_pergame_1", "penaltymin_pergame_1",
    "Weight", "height")]), ]$y, amount = 0.05), xlab = "Propensity Score", ylab = "P[NCAA = 1]")
o.scores = sort(p.scores)
lines(o.scores, exp(o.scores)/(1 + exp(o.scores)))
```



```
matches = matching(z = complete_Cases$y, score = p.scores)
matched = complete_Cases[matches$match.ind, ]
dim(matched)
```

```
## [1] 278  52
```

```r
dim(complete_Cases)
```

```
## [1] 376  52
```

Fit logistic regression model

```r
library(arm)
library(foreign)
display(glm(y ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height, family = "binomial", data = matched))
```

```
## glm(formula = y ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height, family = "binomial",
##     data = matched)
##                      coef.est coef.se
## (Intercept)          -3.22     4.34
## penaltymin_pergame_1  0.01     0.15
## plusminus_pergame_1   0.60     0.48
## assists_pergame_1     0.82     0.87
## goals_pergame_1       1.65     1.06
## Weight               -0.01     0.01
## height                0.02     0.03
## ---
##   n = 278, k = 7
##   residual deviance = 372.8, null deviance = 385.4 (difference = 12.6)
```

compare estimates of the effect of "watched" on post-test score, controlling for all the factors we have been interested in the unmatched regression analysis.
First, try on the forwards.

```r
# total forwards dadta
mod1 = glm(more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height + c, data = updated_nhl_forward, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height + c,
##     family = "binomial", data = updated_nhl_forward)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2370  -0.2870  -0.1921  -0.1268   3.0437
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -28.791147   9.826547  -2.930  0.00339 **
## penaltymin_pergame_1  -0.186416   0.356135  -0.523  0.60067
## plusminus_pergame_1   -0.997637   0.927203  -1.076  0.28194
## assists_pergame_1      5.417314   1.774567   3.053  0.00227 **
## goals_pergame_1        0.079054   1.950958   0.041  0.96768
## Weight                 0.002966   0.025712   0.115  0.90816
## height                 0.128146   0.065756   1.949  0.05132 .
## cUSHL                 -0.513290   0.597541  -0.859  0.39034
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 132.33  on 375  degrees of freedom
## Residual deviance: 108.07  on 368  degrees of freedom
##   (131 observations deleted due to missingness)
## AIC: 124.07
##
## Number of Fisher Scoring iterations: 7
```

```
coef(mod1)
```

```
##          (Intercept) penaltymin_pergame_1  plusminus_pergame_1
##        -28.791146868         -0.186416132         -0.997636826
##    assists_pergame_1      goals_pergame_1               Weight
##          5.417313997          0.079054015          0.002966152
##               height                cUSHL
##          0.128145748         -0.513289764
```

```
# the matched regression analysis
mod1 = glm(more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height + c, data = matched, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height + c,
##     family = "binomial", data = matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1331  -0.3356  -0.2377  -0.1834   2.9903
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -24.542114   9.858195  -2.490   0.0128 *
## penaltymin_pergame_1  -0.345586   0.422956  -0.817   0.4139
## plusminus_pergame_1   -0.821305   0.956083  -0.859   0.3903
## assists_pergame_1      4.454883   1.792971   2.485   0.0130 *
## goals_pergame_1        0.660398   1.946832   0.339   0.7344
## Weight                -0.005677   0.027734  -0.205   0.8378
## height                 0.116132   0.066506   1.746   0.0808 .
## cUSHL                 -0.457749   0.610607  -0.750   0.4535
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.76  on 277  degrees of freedom
## Residual deviance: 101.11  on 270  degrees of freedom
## AIC: 117.11
##
## Number of Fisher Scoring iterations: 6
```

```
coef(mod1)
```

```
##          (Intercept) penaltymin_pergame_1  plusminus_pergame_1
##        -24.542113503         -0.345586149         -0.821305129
##    assists_pergame_1      goals_pergame_1               Weight
##          4.454883070          0.660397767         -0.005677335
##               height                cUSHL
##          0.116131506         -0.457748907
```

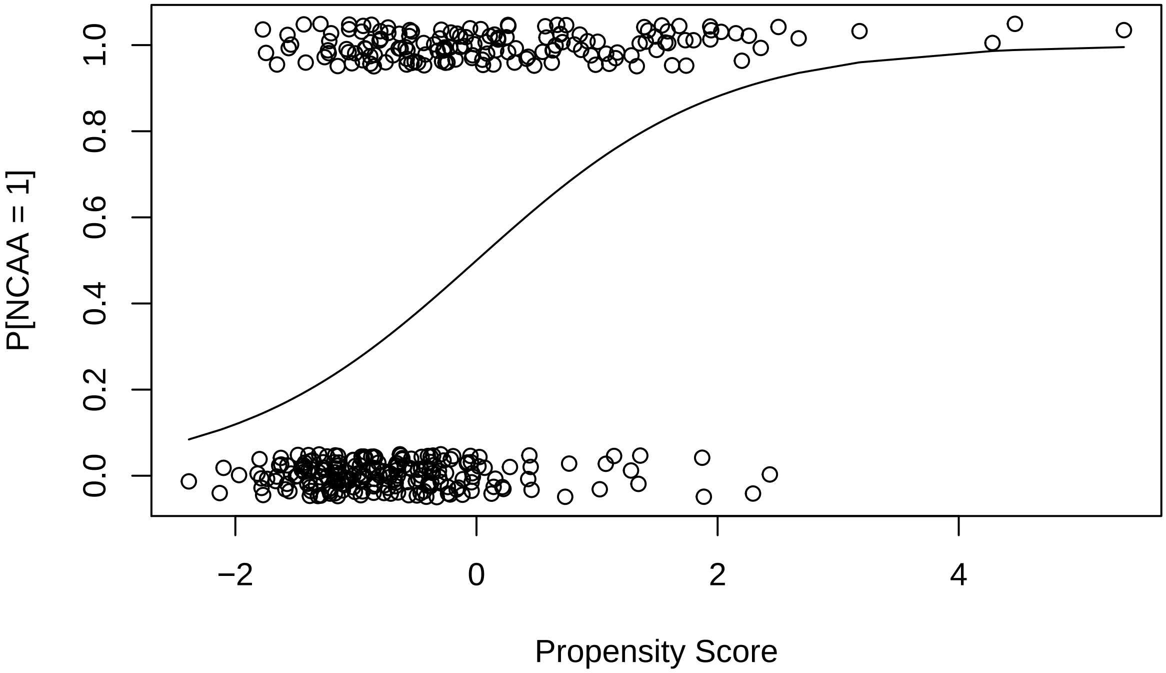
Same process for defensemen

```
library(Matching)
library(arm)
mod1 = glm(y ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 + goals_pergame_1 +
    Weight + height, data = updated_nhl_backward, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = y ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height, family = "binomial",
##     data = updated_nhl_backward)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2438  -0.8936  -0.6653   1.0679   1.9639
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -9.657988   4.488178  -2.152  0.03141 *
## penaltymin_pergame_1 -0.001568   0.151679  -0.010  0.99175
## plusminus_pergame_1   0.664847   0.410876   1.618  0.10564
## assists_pergame_1     3.153731   0.763546   4.130 3.62e-05 ***
## goals_pergame_1       3.456424   1.119641   3.087  0.00202 **
## Weight               -0.009368   0.011603  -0.807  0.41943
## height                0.053741   0.030198   1.780  0.07513 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 491.85  on 361  degrees of freedom
## Residual deviance: 420.00  on 355  degrees of freedom
##   (127 observations deleted due to missingness)
## AIC: 434
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
complete_Cases = updated_nhl_backward[complete.cases(updated_nhl_backward[, c("y",
    "assists_pergame_1", "plusminus_pergame_1", "goals_pergame_1", "penaltymin_pergame_1",
    "Weight", "height")]), ]
p.scores = predict(mod1, type = "link")
plot(p.scores, jitter(updated_nhl_backward[complete.cases(updated_nhl_backward[,
    c("y", "assists_pergame_1", "plusminus_pergame_1", "goals_pergame_1", "penaltymin_pergame_1",
        "Weight", "height")]), ]$y, amount = 0.05), xlab = "Propensity Score", ylab = "P[NCAA = 1]")
o.scores = sort(p.scores)
lines(o.scores, exp(o.scores)/(1 + exp(o.scores)))
```



```
matches = matching(z = complete_Cases$y, score = p.scores)
matched = complete_Cases[matches$match.ind, ]
dim(matched)
```

```
## [1] 302  52
```

```
dim(complete_Cases)
```

```
## [1] 362  52
```

```
display(glm(y ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height, family = "binomial", data = matched))
```

```
## glm(formula = y ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height, family = "binomial",
##     data = matched)
##                      coef.est coef.se
## (Intercept)          -7.54     4.66
## penaltymin_pergame_1 -0.03     0.15
## plusminus_pergame_1   0.27     0.43
## assists_pergame_1     2.16     0.75
## goals_pergame_1       2.75     1.07
## Weight               -0.01     0.01
## height                0.04     0.03
## ---
##   n = 302, k = 7
##   residual deviance = 386.1, null deviance = 418.7 (difference = 32.5)
```

```
# total defensemen data
mod1 = glm(more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height + c, data = updated_nhl_backward, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height + c,
##     family = "binomial", data = updated_nhl_backward)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4341  -0.5341  -0.3799  -0.2645   2.6543
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.11858    6.14704  -1.158  0.24684
## penaltymin_pergame_1 -0.31727    0.22631  -1.402  0.16094
## plusminus_pergame_1  -0.05027    0.55703  -0.090  0.92808
## assists_pergame_1     1.89130    0.86579   2.184  0.02893 *
## goals_pergame_1       1.77792    1.20137   1.480  0.13890
## Weight                0.05259    0.01730   3.040  0.00237 **
## height               -0.02768    0.04271  -0.648  0.51691
## cUSHL                -0.88895    0.37290  -2.384  0.01713 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 283.30  on 361  degrees of freedom
## Residual deviance: 243.53  on 354  degrees of freedom
##   (127 observations deleted due to missingness)
## AIC: 259.53
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
coef(mod1)
```

```
##          (Intercept) penaltymin_pergame_1  plusminus_pergame_1
##          -7.11858289          -0.31727432          -0.05027432
##    assists_pergame_1      goals_pergame_1               Weight
##           1.89130296           1.77792401           0.05259252
##               height                cUSHL
##          -0.02768135          -0.88895334
```

```
# the matched regression analysis
mod1 = glm(more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 + assists_pergame_1 +
    goals_pergame_1 + Weight + height + c, data = matched, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = more_than_10_games ~ penaltymin_pergame_1 + plusminus_pergame_1 +
##     assists_pergame_1 + goals_pergame_1 + Weight + height + c,
##     family = "binomial", data = matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7125  -0.5528  -0.3775  -0.2589   2.8267
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -5.24261    6.60455  -0.794  0.42732
## penaltymin_pergame_1 -0.27671    0.24484  -1.130  0.25842
## plusminus_pergame_1   0.67874    0.60736   1.118  0.26377
## assists_pergame_1     1.93790    0.90706   2.136  0.03264 *
## goals_pergame_1       1.31346    1.23155   1.067  0.28620
## Weight                0.05184    0.01875   2.765  0.00569 **
## height               -0.03728    0.04652  -0.801  0.42289
## cUSHL                -1.04623    0.40937  -2.556  0.01060 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 247.20  on 301  degrees of freedom
## Residual deviance: 211.05  on 294  degrees of freedom
## AIC: 227.05
##
## Number of Fisher Scoring iterations: 5
```

```
coef(mod1)
```

```
##          (Intercept) penaltymin_pergame_1  plusminus_pergame_1
##          -5.24260685          -0.27670765           0.67873741
##    assists_pergame_1      goals_pergame_1               Weight
##           1.93790321           1.31345889           0.05184324
##               height                cUSHL
##          -0.03728300          -1.04622639
```

BART

Minyue Fan

4/21/2021

```r
library(readr)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
options(java.parameters = "-Xmx5g")
library(bartMachine)
```

```
## Loading required package: rJava

## Loading required package: bartMachineJARs

## 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

## Loading required package: missForest

## Loading required package: foreach

## Loading required package: itertools

## Loading required package: iterators

## Welcome to bartMachine v1.2.6! You have 5.37GB memory available.
##
## If you run out of memory, restart R, and use e.g.
## 'options(java.parameters = "-Xmx5g")' for 5GB of RAM before you call
## 'library(bartMachine)'.
```

```r
set_bart_machine_num_cores(4)
```

```
## bartMachine now using 4 cores.
```

```r
nhl_data <- read_csv("merged_data_updated_with_nextLeague.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X1 = col_double(),
##   X = col_double(),
##   Season = col_double(),
##   PlusMinus = col_double(),
##   BirthYear = col_double(),
##   DraftYear = col_double(),
##   Performance = col_double(),
##   Height_cm = col_double(),
##   Weight_kg = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 2 parsing failures.
##    row       col               expected actual                                      file
## 186298 PlusMinus a double                   -  'merged_data_updated_with_nextLeague.csv'
## 266933 PlusMinus no trailing characters     -7 'merged_data_updated_with_nextLeague.csv'
```

```r
nhl_data <- nhl_data %>% select(-X1, -X)
nhl_data <- nhl_data[!duplicated(nhl_data),]


# Calculate NHL games per season
average_nhl <- nhl_data %>%
  filter(League == "NHL") %>%
  group_by(Player) %>%
  summarize(gp = sum(as.integer(Games)),
            num.seasons = n(),
    nhl.games.per.season = gp/num.seasons, na.rm = TRUE) %>%
  select(Player, nhl.games.per.season)
```

```
## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion
```

```
## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion
## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion

## Warning in mask$eval_all_summarise(quo): NAs introduced by coercion
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
final_nhl_data <- nhl_data %>% left_join(average_nhl)
```

```
## Joining, by = "Player"
```

```r
final_nhl_data %>%
  mutate(PenaltyMinutes = as.numeric(gsub(PenaltyMinutes, "-", "0")))
```

```
## Warning: Problem with `mutate()` input `PenaltyMinutes`.
## i argument 'pattern' has length > 1 and only the first element will be used
## i Input `PenaltyMinutes` is `as.numeric(gsub(PenaltyMinutes, "-", "0"))`.

## Warning in gsub(PenaltyMinutes, "-", "0"): argument 'pattern' has length > 1 and
## only the first element will be used
## # A tibble: 177,486 x 21
##    Player Season Team  League Games Goals Assists TotalPoints PenaltyMinutes
##    <chr>   <dbl> <chr> <chr>  <chr> <chr> <chr>   <chr>                <dbl>
##  1 Jarom~   1996 "Cal~ NHL    82    21    29      50                       0
##  2 Jarom~   1996 "Can~ WC     11    2     3       5                        0
##  3 Jarom~   1997 "Cal~ NHL    70    13    19      32                       0
##  4 Jarom~   1998 "Cal~ NHL    82    28    23      51                       0
##  5 Jarom~   1999 "Cal~ NHL    77    29    34      63                       0
##  6 Jarom~   2000 "Cal~ NHL    77    31    40      71                       0
##  7 Jarom~   2001 "Cal~ NHL    82    52    44      96                       0
##  8 Jarom~   2001 "Can~ OG     6     3     1       4                        0
##  9 Jarom~   2002 "Cal~ NHL    75    35    32      67                       0
## 10 Jarom~   2003 "Cal~ NHL    81    41    32      73                       0
## # ... with 177,476 more rows, and 12 more variables: PlusMinus <dbl>,
## #   Position <chr>, DateofBirth <chr>, Nation <chr>, Shoots <chr>,
## #   BirthYear <dbl>, DraftYear <dbl>, Performance <dbl>, Height_cm <dbl>,
## #   Weight_kg <dbl>, NextLeague <chr>, nhl.games.per.season <dbl>
```

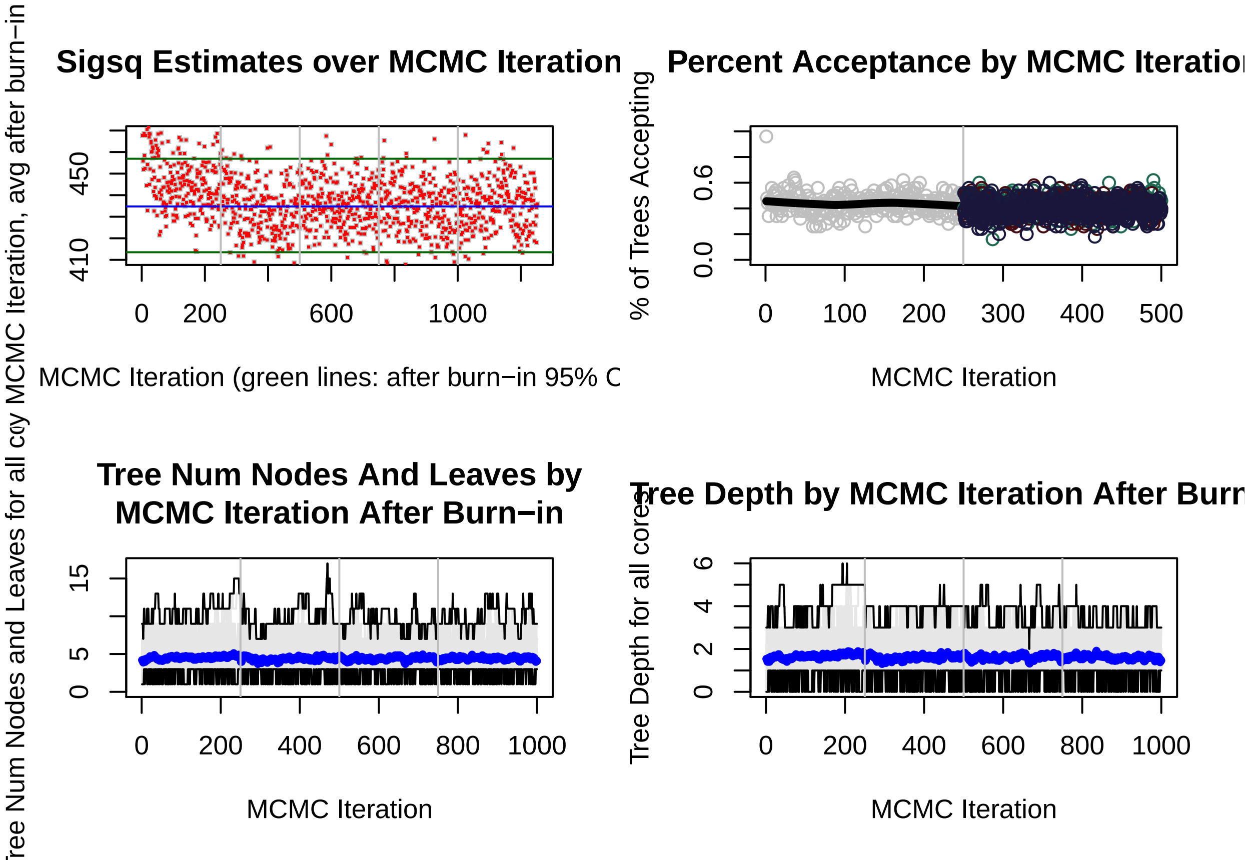
```
bart.dat <- final_nhl_data %>%
  filter(DraftYear == 1 & NextLeague != "NHL") %>%
  group_by(Player) %>%
  select(Player,  Games, Goals, Assists, PenaltyMinutes, PlusMinus,
         Position, Nation, Shoots, Performance, Height_cm, Weight_kg,
         nhl.games.per.season) %>%
  filter(nhl.games.per.season != Inf) %>%
  as.data.frame()


X <- bart.dat %>%
  select( Games, Goals, Assists, PenaltyMinutes,
                         Position, Nation, Shoots, Performance, Height_cm,
                         Weight_kg) %>%
  as.data.frame()
Y <- bart.dat$nhl.games.per.season


nhl_bart <- bartMachine(X[,names(X)!="Player"], Y, verbose = FALSE,
                        serialize = TRUE, use_missing_data = TRUE)
```

```
## serializing in order to be saved for future R sessions...done
```

```
plot_convergence_diagnostics(nhl_bart)
```


Sigsq Estimates over MCMC Iteration
Sigsq for all cores by MCMC Iteration, avg after burn−in
410
450
0
200
600
1000
MCMC Iteration (green lines: after burn−in 95% C
Percent Acceptance by MCMC Iteration
% of Trees Accepting
0.0
0.6
0
100
200
300
400
500
MCMC Iteration
Tree Num Nodes And Leaves by MCMC Iteration After Burn−in
Tree Num Nodes and Leaves for all cores
0
5
15
0
200
400
600
800
1000
MCMC Iteration
Tree Depth by MCMC Iteration After Burn
Tree Depth for all cores
0
2
4
6
0
200
400
600
800
1000
MCMC Iteration