Stat468 Final Project

Trevor Steunenberg

2025-07-12

Table of contents

1 Index

1.1 Abstract

In the days of and leading up to the 2025 NHL Entry Draft there were a total of 18 trades which only included draft picks. This report aims to use player contribution data to determine the relative value of selections in the NHL Entry Draft. Knowing the relative value of picks allows NHL teams to both determine whether they should accept trade offers they have received as well as propose favourable trades to other teams.

Depending on the context, there can be countless approaches one could take to quantify the value of a draft pick. As one example, one could estimate the fairness of a trade by comparing the assets given up and acquired to previous trades. In contrast, this report will estimate value of pick n by utilizing the point shares and games played by previous players selected at pick n along with a k-nearest neighbours algorithm.

One critical point that it is common knowledge in ice hockey circles and confirmed by the resources listed below, along with the Model chapter of this report is that NHL draft picks do **not** decrease in value linearly. In particular, the difference in value pick 1 and 30 is much greater than between pick 101 and 130.

Note that if picks did decrease linearly in value linearly then it would be very easy to create a model of draft pick value since we would have

$$v_1 = v_2 + c = v_3 + 2c = ... = v_{224} + 223c$$

where $c > 0$ and v_i is the value of the i^{th} selection, meaning we would only have to find the value of c.

1.2 Data

The data used by this report is imported from Hockey Reference, which has data on the NHL Draft and player games played and point share counts dating back to 1963, though we will only use a subset of this data as will be explained later. Each row on Hockey Reference is one player selected, and the columns included on the site are:

- `Overall`: the selection where the player was selected
- `Team`: the team that selected the player.
- `Name`, `Nat`, `Pos`, `Age`: the player's name, nationality, position, at age at the time of the draft.
- `To`: the last year a player played in the NHL. For players who never played in the NHL this will be the empty string, for those who are still playing it will be `2025`.
- `Amateur Team`: the team the player was drafted from (confusingly this could be a pro team in Europe).
- `GP`, `G`, `A`, `PTS`, `+/-`, `PIM`: the player's career games played, goals, assists, points (goals plus assists), plus minus, and penalty minutes. For players who never played in the NHL this will be the empty string.
- `GP`, `W`, `L`, `T/O`, `SV%`, `GAA`: the goalie's career games played, wins, losses, ties plus overtime losses, save percentage, and goals against average. For skater and goalies who never played in the NHL all of these columns will be the empty string.
- `PS`: the player's estimated point share, or career points addedto their team (here we mean points in the standings, not goals and assists). There is more info on point share here. \end{itemize} "' Note that we will only use a subset of these columns, as will be explained in the Tidy chapter.

1.3 Previous Work

Some work in this area has been done before, such as:

- Valuation of NHL Draft Picks using Functional Data Analysis
- Examining the value of NHL Draft picks
- NHL draft: What does it cost to trade up?

This report will most closely follow the work done in the first paper listed. As an interesting aside, Eric Tulsky, who wrote the last article listed above in 2013, was hired as General Manager of the Carolina Hurricanes in 2024.

2 Question

This report will estimate the relative value of selections in the NHL Entry Draft. I am not sure what else I am supposed to put here.

3 Import

3.1 Introduction

As mentioned before, we will be importing data from Hockey Reference. Before we import any data, it's important to consider *which* and *how many* years we want to include in this analysis. Since the NHL has changed dramatically over the years, care must be taken to ensure we do not include drafts from too long ago. The primary concern with including data from too many years ago is that teams have likely changed their drafting approach over time. For example, teams may have become better at evaluating prospects as more advanced statistics have been developed, meaning that there are likely fewer late round draft "steals" in the 2020s than there were in the 1980s. Thus including drafts from the 1980s would skew our calculations because it would overestimate contributions by players who were drafted in the later rounds, since those players would potentially have been drafted sooner if the teams of the 1980s had the resources available to teams today. This would make our model a poor estimator of draft pick value for drafts occurring in the 2020s. That being said, players drafted in recent years have not had sufficient time to contribute to their teams, so we should not include drafts from too recently either. Ideally, we would wait until all players from a draft class have retired before including it in our analysis . Practically speaking, this is not feasible since players can have very long careers (for example, Alex Ovechkin was drafted in 2004 and is still playing) which would force us to include older drafts to maintain the same sample size, which is also not ideal as explained above.

Having considered this, we make the somewhat arbitrary decision to use the 25 drafts between and 1996 and 2020 (inclusive). Note that a significant portion of the players in our dataset are still active, so we will have to make an adjustment to account for this. Additionally, it makes sense to give more recent drafts more weight for the reasons described above. We will make both of these adjustments in the Transform chapter.

3.2 Setup

We install and load the necessary packages.

```
# install.packages("rvest")
# install.packages("tidyverse")
# install.packages("janitor")
library(rvest)
library(tidyverse)
library(janitor)
```

3.3 Code

We start off by creating a function to import data from Hockey Reference.

```
start_year <- 1996
end_year <- 2020

import_draft <- function(year){
  url <- str_glue("https://www.hockey-reference.com/draft/NHL_{year}_entry.html")
  html <- read_html(url)
  draft_year_table <- html |>
    html_element("table") |>
    html_table() |>
    janitor::row_to_names(1) |>
    janitor::clean_names()
  draft_year_table
}

head(import_draft(start_year), 10)
```

```
# A tibble: 10 x 21
   overall team    player nat   pos   age   to    amateur_team gp    g     a
   <chr>   <chr>   <chr>  <chr> <chr> <chr> <chr> <chr>        <chr> <chr> <chr>
 1 1       Ottawa~ Chris~ CA    D     18    2015  Prince Albe~ 1179  71    217
 2 2       San Jo~ Andre~ RU    D     18    2008  Salavat Yul~ 496   38    82
 3 3       New Yo~ J.P. ~ CA    RW    18    2011  Val-d'Or Fo~ 822   214   309
 4 4       Washin~ Alexa~ RU    C     18    2000  Barrie Colt~ 3     0     0
 5 5       Dallas~ Ric J~ CA    D     18    2007  Soo Greyhou~ 231   19    58
 6 6       Edmont~ Boyd ~ CA    C     18    2009  Kitchener R~ 627   67    112
 7 7       Buffal~ Erik ~ US    LW/C  19    2007  Minnesota (~ 545   52    76
 8 8       Boston~ Johna~ CA    D     18    2004  Medicine Ha~ 44    0     1
 9 9       Anahei~ Rusla~ BY    D     21    2011  Las Vegas T~ 917   45    159
10 10      New Je~ Lance~ CA    D     18    2004  Red Deer Re~ 209   4     12
```

```
# i 10 more variables: pts <chr>, x <chr>, pim <chr>, gp_2 <chr>, w <chr>,
#   l <chr>, t_o <chr>, sv_percent <chr>, gaa <chr>, ps <chr>
```

We compare the first 10 rows of the 1996 draft table shown above with the table on Hockey Reference It seems that the function we created does what we want it to do.

4 Tidy

4.1 Introduction

Now that we have imported the data, we must clean it. Despite the table from the previous chapter *looking* fairly clean, further inspection reveals some issues:

```
import_draft(start_year)[23:30,]
```

```
# A tibble: 8 x 21
  overall   team   player nat   pos   age   to    amateur_team gp    g     a    
  <chr>     <chr>  <chr>  <chr> <chr> <chr> <chr> <chr>        <chr> <chr> <chr>
1 "23"      Pitts~ Craig~ "CA"  "G"   ""    ""    "Ottawa 67'~ ""    ""    ""   
2 "24"      Phoen~ Danie~ "CA"  "C"   "18"  "201~ "Drummondvi~ "973" "307" "389"
3 "25"      Color~ Peter~ "US"  "D"   "19"  "200~ "Shattuck-S~ "32"  "1"   "1"  
4 "26"      Detro~ Jesse~ "CA"  "D"   "18"  "200~ "Red Deer R~ "49"  "0"   "2"  
5 ""        Round~ Round~ ""    ""    ""    ""    ""           "NHL~ "NHL~ "NHL~
6 "Overall" Team   Player "Nat~ "Pos" "Age" "To"  "Amateur Te~ "GP"  "G"   "A"  
7 "27"      Buffa~ Cory ~ "CA"  "D"   "18"  "201~ "Saskatoon ~ "969" "21"  "137"
8 "28"      Pitts~ Pavel~ "CZ"  "D"   "18"  "200~ "HC Kladno ~ "12"  "0"   "0"  
# i 10 more variables: pts <chr>, x <chr>, pim <chr>, gp_2 <chr>, w <chr>,
#   l <chr>, t_o <chr>, sv_percent <chr>, gaa <chr>, ps <chr>
```

Two problems that immediately come up are numbers being used as strings, the two rows that get inserted at the end of every round, and the fact that (at least) one player is missing everything except for their pick number, name, team, position, nationality, and amateur team. By doing a little bit of detective work with some of the other players with missing values elsewhere in the dataset, we notice that players who never played in the NHL have NAs listed for everything except for the values attributes listed above. We will have to deal with this in the tidy step. Note that Hockey Reference begins listing player's ages in the 2001 draft, but we aren't going to use ages for our analysis so we won't bother coming up for a remedy for the players drafted between 1996 and 2000. Finally, it would be helpful to remove the columns we don't care about.

4.2 Setup

```
# install.packages("tidyverse")
library(tidyverse)
```

4.3 Code

We build a function to tidy the data. In particular, we want it to:

- Remove the rows added between rounds.
- Correct the types of each column so we can use numeric columns in calculations.
- Change `gp` and `ps` values to 0 for players who never played in the NHL or have a negative `ps`.
- If `is.na(to)`, then the player never played in the NHL, so set it to the draft year.
- Add a `year` column so we can adjust the stats of players drafted more recently.
- Select the columns we care about (`year`, `overall`, `pos`, `to`, `gp`, and `ps`) in that order.

Note that there were originally two `gp` columns (one for games played and one for games played as a goalie, goalies have the same number in both), but when we used `janitor::clean_names()` it changed them to `gp` and `gp_2`. Additionally, we cannot remove the round separating rows by removing a specified row number since many of the drafts in our dataset have different numbers of picks per round, and some rounds within the same draft have even had a different numbers of picks.

```
tidy_draft <- function(year){
  draft_year_table <- import_draft(year) |>
    filter(overall != "Overall" & overall != "")|> # remove extra rows
    type_convert() |>
    mutate("year" = year, "ps" = pmax(coalesce(ps, 0), 0),
           "gp" = coalesce(gp, 0), "to" = coalesce(to, year)) |>
    select(year, overall, to, pos, gp, ps)
  draft_year_table
}

tidy_draft(1996)
```

```
# A tibble: 241 x 6
    year overall    to pos      gp    ps
   <dbl>   <dbl> <dbl> <chr> <dbl> <dbl>
 1  1996       1  2015 D      1179  64.6
 2  1996       2  2008 D       496  25.8
 3  1996       3  2011 RW      822  56.6
 4  1996       4  2000 C         3   0  
 5  1996       5  2007 D       231   8.8
 6  1996       6  2009 C       627  12.5
 7  1996       7  2007 LW/C    545   9.2
 8  1996       8  2004 D        44   0  
 9  1996       9  2011 D       917  46.9
10  1996      10  2004 D       209   2.7
# i 231 more rows
```

5 Transform

This is a placeholder for Ch3 – Transform.

6 Visualize

This is a placeholder for Ch4 – Visualize.

7 Model

This is a placeholder for Ch5 – Model.

8 Communicate

This is a placeholder for Ch6 – Communicate.