Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection

Other Sports
1559
© Michael E. Schuckers, Statistical Sports Consulting, LLC

1. Introduction

One of the most important tasks for the general manager of any sports team is the efficient acquisition of player talent. Often one relatively inexpensive ways to accomplish this is through a league draft. In this paper we use historical data available when players were eligible to be selected in the National Hockey League (NHL) Player Entry Draft to build a statistical prediction model for their performance in the NHL. The data that we use is demographic (heights and weights), pre-draft performance (points per game and goals against average) and scouting (rankings from the NHL's own Central Scouting Service (CSS)). We focused on two cohorts of players: those drafted in the 1998 to 2002 drafts and those eligible to be taken in the 2004 to 2008 drafts. In both cohorts, we train our model on the first three draft years and test our model's performance on the remaining (out of sample) two years. Additionally, we train our data on data from the 1998 to 2000 draft and use that model to predict outcomes from the 2007 and 2008 NHL drafts. We find that our statistical model consistently orders players for selection in a way that is more highly correlated with how they eventually perform in the NHL. Simply stated, our statistical model is substantially better at ordering players for the NHL draft than NHL teams.

There has been some previous statistical analyses of the NHL draft. Many of these analyses such as [1] and [2] focus on the value of a draft pick using outcome metrics. [3] investigated league equivalencies which are designed to predict how a player would perform (usually using points) via a multiplier if they moved from a league to the NHL. [4] has reported age dependent league equivalencies. In a different vein, [5] estimated the value of team scouting over CSS and found that team draft performance exceeds that of CSS by millions of dollars. Recently, Lawrence and Weissbock in a series of articles [6, 7, 8, 9] proposed the Prospect Cohort Success (PCS) model which matches players to other players with similar metrics and then uses the success of those players to predict the probability that the original player will reach the NHL. The PCS selects players from the Canadian Hockey League (the Ontario Hockey League (OHL), the Western Hockey League (WHL) and the Quebec Major Junior Hockey League (QMJHL)) and matches those players based upon height, age and points per game. Here we will take a broader approach using a wider variety of inputs. We are interested in predicting the future NHL performance of all players drafted into the NHL or ranked by CSS. Below we summarize the data that we use, describe our statistical model for making these predictions, assess the model's ability to predict NHL performance and discuss the implications of our results.

2. Data

In order to derive a model for future player performance, we collected a series of relevant metrics on all of the players listed in both the central scouting service final report and those taken in the

NHL draft. We considered data in two groups or cohorts. The first group is players eligible for the draft in the years 1998 through 2002. The second group is for players from the 2003 to 2008 drafts. Data were chose from these two periods for two reasons. First it is necessary to get historical data so that we can get an appropriate representation of a player's career trajectory. Second, we used two groups of data to ensure that our approach was replicable across playing eras. Broadly there were three classes of measurements we had for each player: outcome metrics, demographics/physical metrics and historical performance metrics. These metrics were meticulously and painstakingly collected from a variety of websites including eliteprospects.com [10], hockeydb.com[11], hockey-reference.com[12], and nhl.com[13] and thedraftanalyst.com[14]. The difference in the two cohorts is that we have data in the first cohort only on players who were drafted while in the second cohort we have data on all of the players ranked by CSS.

Our outcome metrics were time on ice (TOI) for a player's first seven years in the NHL and the number of games played during that same time interval which follows previous work by [5] Our selection of these metrics is based upon their applicability across positions (centers, wings, defensemen and goalies). We use a players first seven years in the NHL since generally that is the time in which team's retains a players rights before they become an unrestricted free agent [15]. An alternative such as player points, goals plus assists, would predominantly focus attention on forwards (centers and wings). For both our demographic and historical performance metrics, every effort was made to collect data that was contemporaneous to the draft year. For demographics, we focused on a player's height, weight and birthdate as well as the league in which they played in the year prior to their being draft eligible. For historical performance, we accumulated data on how a player performed using metrics such as the number of games played, the points scored, and their goals against average. From these, we also calculated an individual's points per game. All of the historical data is taken from the season prior to a player being drafted.

Table 1: Number of Players at Each Position by Draft Group

Draft Classes	Centers(C)	Defensemen(D)	Forwards(F)	Goalies(G)
1998-2002	275	458	512	153
2003-2007	388	567	676	228

In addition to the above metrics, we obtained the CSS player rankings of each player. Because the CSS rankings are stratified by player location (North American versus Europe) and by position (Skaters versus Goalies), we used the Central Scouting Integratinator, or Cescin, to obtain a unified CSS ranking for each player, [16]. Based upon historical performance within each position group, Cescin takes the ranking within their strata and multiplies that by a fixed constant. For example, prior to the 2007 NHL Draft Jakub Voracek and Lars Eller were ranked 7th and 3rd within their respective categories: North American Skaters and European Skaters. Since the Cescin for North American skaters is 1.35 and for European skaters was 6.27, in our data the Cescin values for Voracek and Eller are 7x1.35 =9.45 and 3 x 6.27 = 18.81, respectively. Using Cescin will allow us to incorporate the information from the CSS rankings into our statistical model below. For those players who were not rated by CSS we use a Cescin of 1500 which is larger than the maximum value from any individual ranked by CSS.

A couple of additional notes about player data are worth mentioning. Our first group contains 1398 players for the 1998 through 2002 drafts while our second group contains 1863 players. Not all players had complete information. In a small number of cases (less than 10), we were not able to find complete information about a player. Players for whom our information was not complete were not included in the results reported below. These cases were primarily for undrafted players who were not rated highly by CSS. Each time a player was ranked by CSS we included them in our data. Likewise, each time a player was drafted we included them in our data and, consequently, some players appear multiple times. Disambiguation of players was sometimes difficult[1]. Below we will look at some summaries of our data in order to give some understanding of the data that will be analyzed by our models.

Table 2 Statistical Summaries of Outcome Metrics in the First Seven Years post-Draft by Draft Group

Draft Classes	% of first 210 players drafted with zero GP	25th percentile of GP, players with at least 1 GP	75th percentile of GP, players with at least 1 GP	25th percentile of TOI, players with at least 1 GP	75th percentile of TOI, players with at least 1 GP
1998-2002	54.6%	18	255	78	3167
2003-2007	55.3%	18	222	194	3809

The breakdown of players by position and by draft group is given in Table 1. For ease of modelling we classified players into the four positions given in Table 1. For players whose position was listed as some combination, e.g. "C/RW" we used the first position listed. Additionally, we combined players listed as forwards, right wings and left wings into Forwards (F). Table 2 below has some statistical summaries for our response variables, TOI and GP, broken down by draft class group. We can see that the 25th and 75th percentiles for GP is quite similar across the groups while the same percentiles for TOI has been shifted slightly upward. Since the length of drafts in the two groups varied from 211 (in 2008) to 293 in (2000), for comparison of the percent of players who play at least one game in the NHL, we limited our analysis to the first 210, 30 teams times 7 rounds, selections. We see a slight but not substantial difference in the number of players who do not play a single game in the NHL between the two cohorts.

Table 3: Counts of Players by Previous Season's Team and Draft Group

Draft Classes	Finland	NCAA	OHL	QMJHL	Russia	USHL	WHL
1998-2002	53	163	216	130	54	51	212
2003-2007	44	87	278	207	66	159	268

[1] For example, there are five Robin Olsson's, two forwards, two defensemen and a goalie, from Sweden born in either 1989 or 1990 and two Jakub Cech's born in 1985.

Next we consider the leagues in which player's played in the season that they were evaluated by CSS or drafted. Counts for some of the more common leagues from which players are drafted are listed in Table 3. While the overall counts are very different here, you can see there are some trends. More players are being ranked from the USHL and fewer from the NCAA. Note that totals for Finland and Russia represent the SM-Liiga and Russian first division, respectively. Also noteworthy though not demonstrated here is that the number of players selected from US high schools increased from the first draft group to the second.

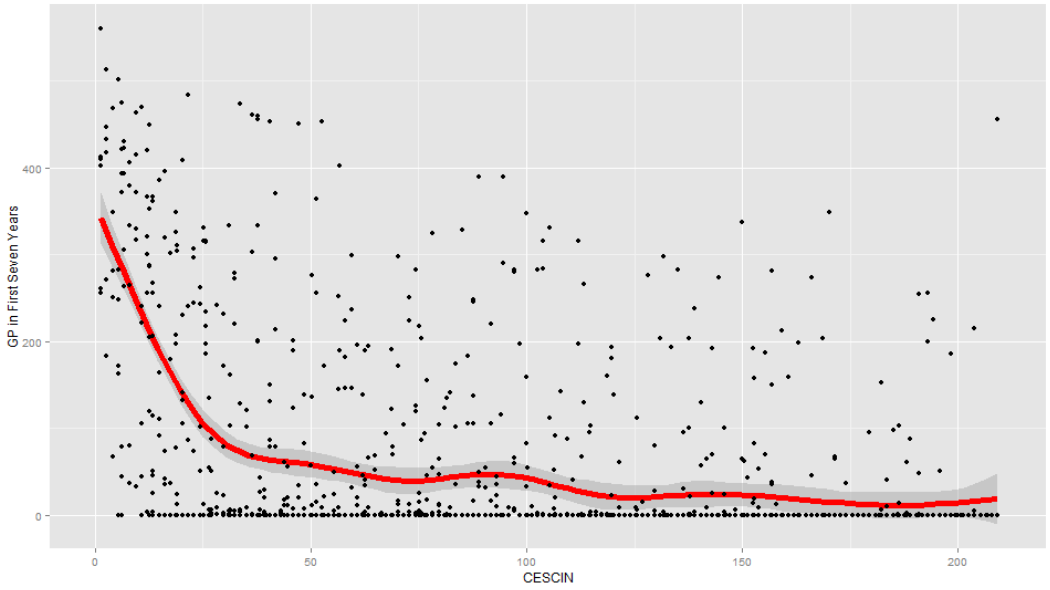

Figure 1: Plot of GP versus Cescin

3. Statistical Modelling

In order to predict player performance at the NHL level, we use a generalized additive model due to [17]to combine predictors on from players eligible for the NHL Player Entry Draft to predict player performance metrics. The data used for prediction was available about the players when they were drafted. Among the variables we use are player demographics (e.g. height, weight, nationality) and player performance (e.g. goals, assists, save percentage) in the league from which they were drafted. Our response variables are the number of games played (GP) in a player's first seven seasons after they have been drafted (roughly the period when a team controls the rights to a player) as well as their time on ice (TOI) over that same period.

For two separate draft groups, we use data from three consecutive NHL drafts to build our model and then we predict each player's performance in the two subsequent out of sample years. For example, in our first cohort we use data from players selected in the 1998, 1999 and 2000 NHL drafts to develop our model and we then predict the rank order for players in the 2001 and 2002 drafts. The second cohort covers the NHL drafts from 2004 to 2008 and uses the 2004, 2005 and 2006 drafts as training data and the 2007 and 2008 drafts as test data.

We chose a Poisson Generalized Additive Model to model our integer responses and to be able to incorporate some non-linear effects into a single predictive model, [17]. The models were fit using the gam package in R [18, 19]. As mentioned above, our model includes demographic predictors, Cescin, predictors that are a function of a potential draftee's position, predictors that are functions of the league in which the player played during the prior year and appropriate combinations of these predictors. We think of the functional relationship between Cescin and the outcome variables as the baseline for our prediction and the other predictors are adjustments to that based upon other attributes of the individual. Figure 1 illustrates the non-linear relationship between GP and Cescin. The generalized additive model allows us to easily incorporate such non-linear functional relationships between our responses (GP and TOI) and our predictors.

Table 4: Comparison of Rank Correlation Magnitude with NHL Performance Among Drafted Players

Training Data NHL Draft Years	Out of Sample Draft Year	NHL Performance Metric	NHL Draft Order	Draft by Numbers
1998, 1999, 2000	2001	TOI	0.366	0.603
1998, 1999, 2000	2001	GP	0.383	0.532
1998, 1999, 2000	2002	TOI	0.282	0.587
1998, 1999, 2000	2002	GP	0.348	0.536
2004, 2005, 2006	2007	TOI	0.403	0.642
2004, 2005, 2006	2007	GP	0.401	0.694
2004, 2005, 2006	2008	TOI	0.398	0.685
2004, 2005, 2006	2008	GP	0.401	0.708

4. Results

In this section we describe the results of two experiments. In the first we build statistical models for only drafted players for both of our cohorts. In the second we build statistical models for all players eligible for the NHL draft, specifically those drafted and those ranked by CSS. Because we are interested in ordering of players, we use the Spearman rank correlation as our measure of association and predictive ability. We analyzed the predictions for both cohorts and for both of our performance metrics. Table 4 below gives a comparison between the rank correlation between our performance metrics (TOI and GP) and the two approaches here. The second to last column has the correlation between the actual order in which players were drafted and their respective performance metrics. The last column has the correlation between the predicted values using our statistical model and their respective performance metrics. We will refer to our predictive statistical approach as Draft by Numbers. The results in Table 4 are based only upon players who were drafted. The typical correlation for the NHL draft ordering is around 0.4 while the typical correlation for DbN is close to 0.6. It is clear from this table that our predictive model orders players in a manner that is significantly more correlated with their performance in the NHL.

Table 4 gives results for the out of sample prediction for the players taken in each of the four out of sample drafts using both first seven TOI and GP as responses. To evaluate and compare our methodologies we use Spearman's rank correlation instead of the more common Pearson correlation coefficient because we are most interested in the ordering (or rank) that each methodology predicts. We can see from Table 4 that the average rank correlation between the

NHL draft order and the NHL Performance metrics is about 0.4 while our Draft by Numbers generalized additive model approach averages about 0.6 for the first draft group and about 0.7 for the second draft group. This strongly suggests that the model outperforms the draft ordering by player selection.

Table 5: Comparison of Rank Correlation with NHL Performance Among All Players

Training Data NHL Draft Years	Out of Sample Draft Year	NHL Performance Metric	NHL Draft Order	Draft by Numbers Order
2004, 2005, 2006	2007	TOI	0.547	0.667
2004, 2005, 2006	2007	GP	0.547	0.670
2004, 2005, 2006	2008	TOI	0.553	0.670
2004, 2005, 2006	2008	GP	0.557	0.655
1998, 1999, 2000	**2007**	**TOI**	**0.547**	**0.650**
1998, 1999, 2000	**2007**	**GP**	**0.547**	**0.659**
1998, 1999, 2000	**2008**	**TOI**	**0.553**	**0.619**
1998, 1999, 2000	**2008**	**GP**	**0.557**	**0.616**

While the results in Table 4 are strong, they do not mirror the process of scouting players eligible for the NHL Draft. When teams are scouting for the draft do not know the complete list of those drafted. To address this, we analyzed all those players who were either drafted or ranked by CSS and repeated the model building process above. We were only able to build a complete dataset for the second cohort and results from this analysis are given in Table 5. On average the correlation between the Draft by Numbers model prediction and a player's performance in the first seven years is about 0.66 while the NHL draft order has a rank correlation of about 0.55. Thus the statistical model improves the player ordering by about 20%. Further, we looked using the same strategy to predict the results for seven years beyond the data. This closely approximates the actual approach one would take when predicting future player performance using out outcome metric of first seven years of GP or first seven years of TOI. The results for this seven years out of sample prediction can be found in the last four rows of Table 5. Though the results are not as strong when the gap between training and test data is shorter, we find that the DbN order is still better than the NHL Draft ordering of players.

5. Discussion

In this paper we have built and assessed a variety of statistical models for the prediction of future NHL performance for players eligible for the NHL draft. Using data from two groups of NHL draft data, data from1998 to 2002 and from 2004 to 2008, we evaluated our statistical models by predicting out of sample player performance. All of the observations in our generalized additive regression models were available to teams at the time of the respective drafts. To make these predictions, we used demographic information about players as well as performance information. The performance metrics we chose to use were a player's NHL time on ice and games played through the first seven years after their draft eligibility. These response metrics allow for the comparison of player value across positions which is critical for their usage in drafting. To incorporate scouting information into our prediction models, we used multipliers of the NHL's Central Scouting Service rankings to the Cescin of Fyffe [16]. To facilitate the prediction of our

response variable using the data from these multiple sources we build generalized additive regression models. Out of sample predictions based upon these statistical models outperformed the draft ordering that NHL teams have historical chosen. These models worked on two groups of players from different draft years and both for the prediction of drafted players and for all draft eligible players. These results mean that hockey scouting is not just an 'eyeball business' but rather it is a numbers game. Our general approach of building a model that accounts for a player's physical and demographic characteristics, their performance in their respective league and other relevant factors to predict future performance is something that potentially has broad applicability for the ranking and drafting of players across sports.

References

[1] Tulsky, E. (2013, April 25). NHL draft: What does it cost to trade up? Retrieved December 14, 2015 from http://www.broadstreethockey.com/2013/4/25/4262594/nhl-draft-pick-value-trading-up.

[2] Schuckers, M. (2011 June 20) What's What‟s An NHL Draft Pick Worth?A Value Pick Chart for the National Hockey League Retrieved December 14, 2015 from http://myslu.stlawu.edu/~msch/sports/Schuckers_NHL_Draft.pdf

[3]Desjardins, G. (2004, December 30), League Equivalancies. Retrieved June 11, 2013 from http://hockeyanalytics.com/2004/12/league-equivalencies/.

[4] Vollman (2015, November 8) Updated Translation Factors. Retrieved on December 13, 2015 from http://www.hockeyabstract.com/thoughts/updatedtranslationfactors

[5]Schuckers, M. and Argeris, S. (2014, November 2014) You Can Beat the Market: Estimating the Return on Investment for National Hockey League (NHL) Team Scouting using a Draft Value Pick Chart for the NHL. Retrieved December 14, 2015 from http://arxiv.org/abs/1411.5754

[6] Lawrence, C., (2015, January 22). The NHL Draft: Maybe Size Does Matter. Retrieved December 13, 2015, from http://canucksarmy.com/2015/1/22/the-nhl-draft-maybe-size-does-matter.

[7] Lawrence, C., (2015, April 11). The Draft Files: The Historical Cohort Based Approach Gets Its Sham On. Retrieved December 13, 2015, from http://canucksarmy.com/2015/4/11/the-draft-files-the-historical-cohort-based-approach-gets-its-sham-on

[8] Weissbock, J. (2015, May 26). Draft Analytics: Unveiling The Prospect Cohort Success Model. Retrieved December 13, 2015 from http://canucksarmy.com/2015/5/26/draft-analytics-unveiling-the-prospect-cohort-success-model.

[9] Lawrence, C. (2015, September 21). Prospect Cohort Success – Evaluation of Results. Retrieved from http://hockey-graphs.com/2015/09/21/prospect-cohort-success-evaluation-of-results/

[10] Elite Hockey Prospects (2015) www.eliteprospects.com Last accessed December 10, 2015.

[11]hockeyDB.com (2015) www.hockeydb.com Last accessed December 10, 2015.

[12] Hockey-Reference (2015) www.hockey-reference Last accessed December 10, 2015.

[13] NHL.com (2015) NHL Draft - Historical Data. Retrieved December 2, 2015 from http://www.nhl.com/ice/draftsearch.htm

[14] The Draft Analyst (2015) www.thedraftanalyst.com. Last accessed December 13, 2015.

[15]National Hockey League (2013, February15) Collective Bargaining Agreement Between National Hockey League And National Hockey League Players' Association. Retrieved December 13, 2015 from http://www.nhl.com/nhl/en/v3/ext/CBA2012/NHL_NHLPA_2013_CBA.pdf

[16]Fyffe, I. (2011, January 19). Evaluating Central Scouting, January 19, 2011. Retrieved April 1, 2013 from http://www.hockeyprospectus.com/article.php?articleid=777, accessed April 1, 2013.

[17]Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman & Hall/CRC.

[18] Hastie, T. (2015) gam: Generalized Additive Models. R package version 1.12. Retrieved from http://CRAN.R-project.org/package=gam

[19] R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria. http://www.R-project.org/},

[20]Macdonald, B. (2012) An Expected Goals Model for Evaluating NHL Teams and Players. Proceedings of 2012 MIT Sloan Sports Analytics Conference: Boston, MA.