Utah Hockey Club Hackathon Challenge

Ryan Chase and Anup De

Introduction

With the NHL draft recently taking place and being students at Boston University, we have been very intrigued by what goes into the behind-the-scenes work of a player being scouted and ultimately drafted. With the hockey world being so large and diverse with many different leagues and styles, we wanted to analyze and model how a player's performance in their juniors league (pre-draft) will impact their draft stock and furthermore, their NHL stats.

Firstly, we looked through the Kaggle database to find possible datasets to be used. We found a few that used web scraping to collect data from "Elite Hockey Prospects" which had the most encompassing data from leagues around the world and the NHL. Once we had a rough dataset to work with, we began the cleaning process by scanning the data using Power BI to get initial visuals on the data and find what information we lacked to connect players' juniors league performance to their NHL draft picks and NHL production. The first issue we came across was that most of the data did not fit the specific column data type. For example, some players' names had been shifted over into the year column. These players often lacked all necessary data, so we needed to remove them from the dataset. The other main issue was in our junior league data set, we were missing approximately half of all draft picks, especially the most recent drafts. To solve this problem, we got a list of every single draft pick since 1980 and their NHL stats to pair with this data. Once we tried to combine this data we ran into an issue of repeating names, so to resolve this we made unique player IDs based on a player's name and the amateur league they played in. This allowed us to match almost 80% of the players. Now that we had two main data sets which had a many to one, meaning we had multiple years of junior stats for every player, connection from their juniors' stats to their singular draft position, we were almost done with our data cleaning. In our juniors' stats dataset, players' team and league names were different from our draft data set, so we needed to alter the league and teams based on a mapping system in Python. At this point, we were ready to start analyzing the data, but we wanted to narrow down our scope by limiting what drafts we analyzed. We decided to remove any players who were drafted before the 2000 NHL draft as this was the first year that the draft had thirty teams participating. Very few players from the 2000-2005 draft are still playing in the league so it allows us to look at players who were drafted and had a long career and retired along with all active players.

Before we dive into our analysis, we want to cover what exactly we were looking for in this data which helped to guide our analysis. The NHL draft is integral to team success, whether it is continuing to improve on an already good roster or in the rebuilding process. From trying to fill gaps in a roster with cheap entry-level contracts or looking for the next franchise cornerstone, a team needs to make sure they have all of the information available to them while making a selection. While this may not determine exactly who to pick, it adds to the information a team has while making a pivotal decision towards the success of the team. A big part of what we aimed to do is to find a way to level the playing field between leagues. For example, a player from the NCAA will have different competition and standards than someone coming out of the Swedish league. Players being drafted out of professional leagues will have different statistics than players from leagues where the age is capped. This allows us to look at different junior leagues to determine what stats correspond with what draft picks, leveling the playing field between a player with lower metrics in a tougher league and a player with better juniors numbers but in a lower caliber league. With that, we will first look at players' careers in the NHL and create a model to determine a player's performance relative to their draft pick as well as which leagues have the best NHL performance.

Draft Positions and NHL Production Analysis

One indicator of a player's performance in the NHL is how long they were able to stay in the league, so we first looked at how many games a player had in the NHL compared to where they were drafted. As shown below, we can see about 50% of players drafted in the first round played over 200 games, representing about two and a half full seasons. The top 50% of players drafted in the second round played 0 or more games, a big decrease from the first round. The third round follows a similar trend as the second, but fewer than 50% of players playing a game in the NHL. This demonstrates that in most situations the best players will come out of the first round followed by a severe

drop in talent. The second and third round have significant drops as well, and in any subsequent round only outliers have a meaningful impact in the league, as displayed in the boxplots adjacent.

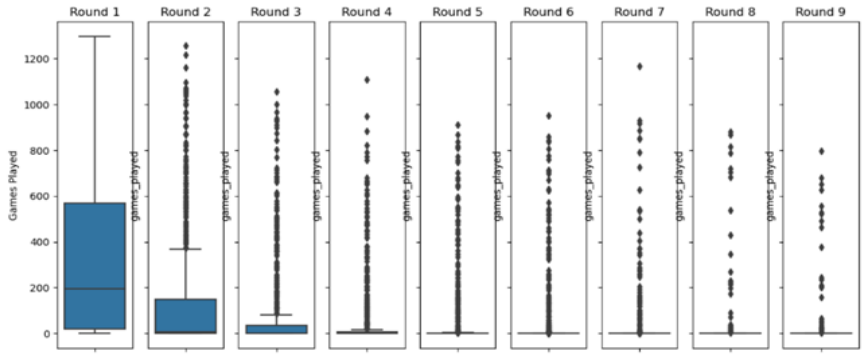

Figure 1: Box plots of all 9 rounds of the draft and their respective games played in the NHL

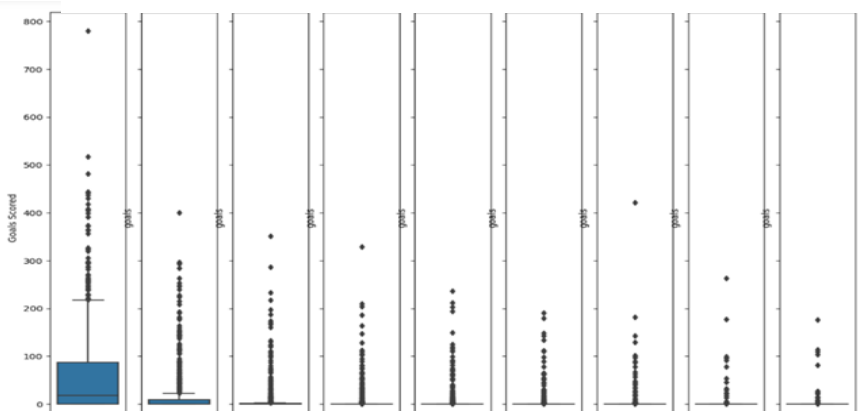

Figure 2: Box plots of all nine rounds of the draft and their respective goals scored in the NHL

Next, we will take a look at how goals scored in the NHL changes depending on the round players were drafted in. This is very similar to the games played statistics but likely favors forwards and offensive defenseman more than other players. This continues to show a large disparity between the first round and the second round and gets even bigger as time goes on. This shows that about the top 25% of players in the first round will score more than 100 goals in their career but the second round only has severe outliers who score more than 100 goals. This further illustrates the need for the first-round picks to be cornerstones of a franchise and look for possible steals or role players with the remaining rounds.

Next, we decided to create a model to determine which players have performed above their expected value in each overall pick. We ran a linear regression model on the goals, games played, assists, and plus minus of every single player in their respective overall pick. This would give us a general idea of what value you should be expecting to get with the pick you have, with this value being measured in eventual production at the NHL level. Next, we took the residuals to find the difference between the players predicted performance and how they actually performed. Once this was done, we gave a player a score in each of the main statistical categories to determine their value above the expected. After totaling all four categories up, this gives a Pick Value Residual which we used to determine a player's overall performance relative to their pick. The idea of this is to level the playing field between players drafted across the board as it does not make sense to compare the career of a first overall pick to the last overall pick as they started from different points. Let us look at which players performed the best compared to their peers in the same pick in each category. For games played, Joe Pavelski received a 1117 residual meaning he played over a thousand more games than expected for his pick. Alexander Ovechkin, one of the most unanimous first overall picks, had the highest residual in goals with 642 more goals than predicted. Even though he is being compared against other first overall picks, he is still a remarkably good goal scorer. Sidney Crosby was one of the greatest draft talents of all time and has also played above expected with an assist residual of 672. In plus minus, Brad Marchand shined above everyone with a residual of 263 proving him to be one of the most impactful players on the ice. Finally, combining all of a players residuals gave them their overall pick value and one player reigned supreme in this as Patrice Bergeron scored a 1925 in overall pick residual. These players represent the tops, middle, and bottom of their draft classes, showing the model can accurately compare careers relative to all overall picks.

Once we compiled every player's residuals in the main statistical categories as well as their total pick value, we wanted to analyze how different leagues performed against each other. To do this we created a data frame with the

leagues with the most draft picks and then all of their players residuals in each category summed up. This creates this below data frame:

	games_played residuals	assists residuals	goals residuals	+/- residuals	Pick Value Residual
OHL	15037.057377	3152.327383	3494.862491	-1864.632892	16060.350015
WHL	-5017.07438	-4649.499659	-3424.795995	-661.158388	-12498.259827
QMJHL	-6890.831238	11.727251	-271.84254	799.132161	-4629.106556
BCHL	689.386737	283.055203	235.85641	-292.057063	743.894603
USHL	4661.798345	1522.333881	868.109711	74.720598	5961.512949
NCAA	6001.385108	481.045679	19.378419	-254.12013	4747.342799
Finland	1028.666076	425.392845	-13.375078	69.576377	1253.093702
USDP/USHL	-802.994483	-585.673093	-405.165192	-248.878513	-1841.96266
Russia Jr.	-1009.141524	-79.431391	-24.711916	129.233424	-731.766026
Sweden	3612.300854	1360.854373	298.28443	177.180272	4545.544715
Sweden Jr.	-1583.225785	-320.41692	-492.58507	352.512111	-1647.909217
Swiss	-112.221293	208.78755	-101.279567	-17.400674	5.941339
Norway	22.517453	-24.192435	-12.140071	-24.2204	-43.664816

Figure 3: Data frame with four major statistical categories and their sum with respective top juniors leagues

In this data frame, we can see the best overall league is the OHL with a considerable gap over all of the other leagues. With OHL players performing above and beyond expectations this would make their players more likely to turn out to be producers in the league. The other leagues following behind are the USHL in second followed closely by the NCAA. An interesting note is that Joe Pavelski had a pick value of 1895 which makes up a large portion of the USHL's 5961 total pick value residual. The WHL was by far the worst performer in almost every single category. This illustrates that players from the WHL may be getting picked higher than they the numbers suggest they should be. The top European league is by far the Swedish league, where players compete against other older professionals with vast experiences around the world, and so these junior players are much more likely to transfer their talents directly to the NHL. These insights can help to determine which players would be best to take a chance on and see if they could turn out to be a much better player for where they are selected. This does not mean that if you draft only OHL players that you will have all of them perform above expectations but does say these players are more likely to perform well in the league. If deciding between two different players from drastically different leagues and different performances, this helps to scale their performances for a more accurate idea of how this player will perform in the NHL. Our further analysis will help find players who are going to become the next Joe Pavelski by building a model looking at juniors statistics and again solving the problem of numbers coming out of different leagues.

Exploring Connections Between Junior League Performance and Draft Status

To mitigate the problem of comparing different leagues, we compared players to others in that league creating a percentile statistic that is controlled for position and league. We now have information that informs us of how a player must perform in their junior league in order to get drafted. For example, the average winger drafted out of the USHL was in the 53rd percentile of all wingers to play in the USHL since 1998. This analysis can be further

Figure 3: PowerBI Dashboard that shows for wingers out of QMJHL, drafted in the first round, their average percentile is 0.82 in goals. Chart shows percentile by draft pick with the constant line highlighting that average, making residual analysis easier. Sample size, player identification, and further slicers are also present.

applied amongst specific draft rounds, so this number jumps to the 64th percentile for those drafted in the first round. To compare with other major leagues, the QMJHL has its average at the 69th percentile amongst wingers with a jump to the 82nd percentile for those drafted in the first round. This allows us to conclude about the strength of these leagues letting us know that players must perform significantly better in the QMJHL to get drafted compared to players out of the USHL. This PowerBI analysis only includes players out of the USHL, QMJHL, BCHL, WHL, NCAA, and OHL, as these maintained a large sample size and were comparable in the demographics playing in each league.

Using historical trends of draft picks, we can understand what we expect to get out of a certain draft pick. For the purposes of maintaining sample sizes, draft picks were considered in groupings of 10, considered in the PowerBI report as 'overall_pick (bins).' Given that a player was drafted in a certain grouping, we can find what percentile of key statistics (goals, assists, games played, and plus-minus) we expect, once again controlled for what league they came from and what position they play. Once we have gathered these expected values (which are average percentiles of all players drafted in that specific grouping, out of that league, and playing that position), we can find players' residual values for their draft position. Large residual values suggest that a player's juniors performance was well above where they were drafted.

For a case study example, let us consider the highest residual we found for the goals statistic. Among players that have gone on to play at least one game in the NHL, Joe Pavelski achieved the highest residual. Based on his status as the 205th overall pick (for our purposes this is the grouping of picks numbered 200-209), and as a center out of the USHL, he was expected to fall in the 51st percentile amongst USHL centers (this is the average percentile of all USHL centers drafted between 200 and 209). However, his juniors performance had him in the 99th percentile, giving him a residual of 48 points. From this metric, we can see how he was one of the greatest 'steals' in any NHL entry draft.

We can further combine our residuals to create one encompassing score that shows how far above or below a player's junior performance is compared to their draft position. For the purposes of this exercise, the residuals for goals, assists, games played, and plus-minus were added together. If one of those categories was deemed more important than the others, that score could be calculated differently. That summed score informs us that Max Pacioretty (1.97) has the highest summed residual of all players with Brandon Montour (1.82) in second. They were drafted 22nd and 55th overall respectively and this shows that our statistic and percentile-based residuals does not just highlight players drafted later on like Pavelski but can also tell us about steals taken in the first and second rounds. Let us do a case study on another player with a high score in this metric: Oliver Bjorkstrand (1.44). Amongst all right wingers out of the WHL drafted between 80th and 89th overall, they average in the 47th, 52nd, 54th, and 66th percentile in goals, assists, games played, and plus-minus respectively. Bjorkstrand however, performed well above all of these marks in the WHL, earning him highly positive residuals.

This analysis can help an NHL team at the draft. Imagine the scenario at the 2013 NHL Entry Draft, where a team has a pick in that 80th to 89th grouping. While considering all players available, an NHL team could have difficulty comparing players with different stats in different leagues. A statistic could simulate if a player was to be drafted now, how much above (or below) that player is compared to what we expect from their league and their position. Bjorkstrand would have numbers well above a typical right winger from the WHL drafted in that range. Compare this, for example, with another right winger taken 8 spots ahead of Bjorkstrand: Kurtis Gabriel. For all OHL right wingers drafted in this range, we expect percentiles of 58, 60, 75, and 66 for the key stats, but Gabriel performed below these metrics and had a negative total residual of -0.35 compared to Bjorkstrand's total of 1.44.

Our draft dashboard page in the PowerBI report attempts to help an NHL team sitting at the draft table. They could run this simulation with the pick they are currently on to determine if they are getting a 'steal' for that draft position. This dashboard allows the user to put themselves in a draft year and at a certain grouping of picks and see the remaining players. A calculation is then performed to see if a player were to be selected now, what juniors' statistics are expected of them given that player's biography (position and junior league). This is reflected in the 'ExpectedGoals_CurrentPick' and 'ExpectedAssists_CurrentPick' columns for each player. Based on their actual junior league performance, the residual is then calculated. This residual represents how much of a 'steal' a team is getting, which means how much of a better player a team is getting based on what they expect to get out of that pick.

The dashboard now has the players sorted based on their residuals summed between goals and assists. Other advanced statistics that an NHL team would be interested in, could also be implemented. Here is a possible example of the current dashboard in use. The dashboard will show that for defensemen left after the second round of the 2014 draft, the players with the highest summed residual are Devon Toews, Ben Thomas, and Brandon Montour, making them all steals for this pick in the draft, in fact greater 'steals' than all other players. As a defenseman drafted out of the NCAA, the 30th to 39th overall picks are supposed to yield a player in the 66th percentile of both assists and goals. However Toews is still left on the board as a player in the 85th percentile in goals and 88th percentile in assists.

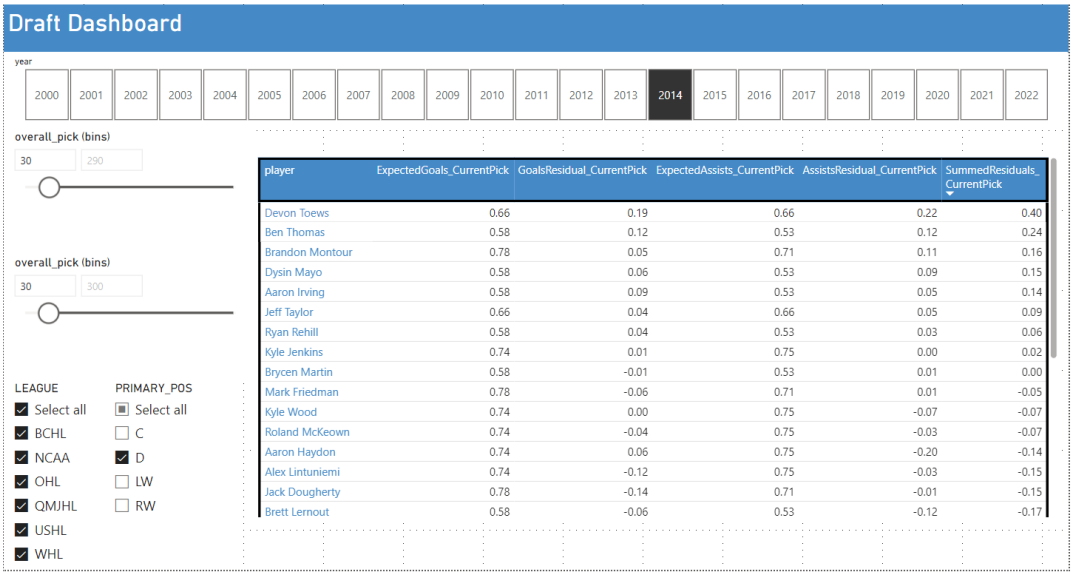

Figure 4: Draft Dashboard highlights the 2014 draft, amongst defensemen remaining after the first round. The table features expectations of goals and assists based on individual juniors league and positions and residuals based on actual performance in those juniors leagues. This expectation is reflected on if that player was drafted between 30 and 39th. SummedResiduals_CurrentPick adds these residuals, and they are displayed in decreasing order, putting Devon Toews at the top.

Another dashboard was also built to convert these percentiles into their actual numbers, so it can be easily understood what a certain percentile in a certain league at a specific position actually represents when it comes to goals, assists, games played, or plus-minus.

Conclusion

With all of this in mind, we do need to discuss some of the caveats to our analysis. The biggest issue in this is incomplete data, as we are still missing a portion of draft picks and player data due to data cleaning and general web scraping. Much of our analysis is broken down by league and position, meaning that there are cases of a small sample size of players being drafted in those groupings of 10, who were from the same league and played the same position. On the player side, we do not have some players' full NHL careers especially for those drafted recently. Goalies are also included in this data set which is unfair to them as they would not have almost any goals or points. Our dataset also includes juniors statistics for players who started playing very early on, thus bringing down certain averages. Furthermore, players with only a few games played in each junior league also skew the overall percentiles as their overall statistics are naturally lower. On the regression front, the model we built is rather simple and covers the entire data well but does not handle outliers well.

This project was done in a very short amount of time, and we have continued to ask more questions as we have gone on and we do believe that the continuation of such a project could yield remarkable conclusions. Expanding our dataset through more accurate web scraping or access to more complete juniors and draft data would increase the accuracy of our percentiles. Adding other metrics like blocks, steals, time on ice or advanced metrics could also make the models more useful to NHL teams. With more complete data, we can work on creating a better model to more accurately determine the best players to draft and their possible futures in the NHL. The draft dashboard is an outline on what we want to continue doing in the future. Our modeling leaves us room to make our own mock draft, where we find where the best place to draft every player lies, based on how their percentiles line up with historical averages. More accurate models can help enhance a team's ability to select the best players at the right time.