Predictive Modelling and the NHL Entry Draft

Alex Church

100825475

COMP 4905, FALL 2019

Supervisor: Dr. Tony White

Contents

[Introduction 1](#_Toc26649681)

[Background 1](#_Toc26649682)

[Prior Work 2](#_Toc26649683)

[Objective 7](#_Toc26649684)

[Methodology 8](#_Toc26649685)

[Data Collection 8](#_Toc26649686)

[Computing of League Equivalency Factors 10](#_Toc26649687)

[k-means Clustering 15](#_Toc26649688)

[Design and Data Pre-Processing 17](#_Toc26649689)

[Training, Testing, Tuning, and Model Selection 21](#_Toc26649690)

[Predicting Whether a Player Will Make the NHL 21](#_Toc26649691)

[Forwards 21](#_Toc26649692)

[Predicting Top 9 Forwards and Top 4 Defensemen 26](#_Toc26649693)

[Forwards 26](#_Toc26649694)

[Conclusions and Future Work 26](#_Toc26649695)

[Works Cited 27](#_Toc26649696)

# Introduction

## Background

At the conclusion of every season the NHL holds an annual Entry Draft, where teams alternatingly select the rights to junior-aged prospects, in the hopes of identifying prospective NHL players. The modern day draft is composed of seven rounds, where each of the 31 franchises are allotted one selection per round, with the annual order of selection being predetermined based on a team’s performance in the NHL season prior. Teams are also permitted to exchange future draft selections via trade at any point. The most recent iteration of the NHL draft was held on June 21-22 of this year, where 217 eligible prospects were selected collectively by the NHL’s 31 franchises.

Each year, the pool of draft eligible prospects is drawn from a number of junior and professional leagues across the world, all varying in terms of age and skill level. In order to be eligible for selection, a prospect must have turned 18 prior to September 16th of the draft year in question. A first time eligible player can be defined a prospect who either turned 18 years of age between January 1st and September 15th of the year of the draft, or who turned 18 between September 15th and December 31st of the prior calendar year. Any eligible prospects who are older than 18 (based on the above criteria) would have been eligible for a previous Entry Draft, meaning they are then only eligible for selection in subsequent years so long as their rights are not held by an NHL franchise at the time of the draft. This latter group of prospects is often referred to as over-agers.

Since the introduction of the NHL’s salary cap in 2005, which imposes an upper limit on the amount a team may spend collectively on the salaries of their players, the NHL Entry Draft has gained increasing significance as a means by which teams might be able to gain a competitive advantage. This is largely due to the fact that, upon signing with the team who drafted them, a player’s first professional contract (known as an entry level contract) is capped using a standard scale. This is significant as the end result is that players on their entry level contracts can often offer franchises better value per dollar than their more tenured counterparts. As players age and their salaries increase, teams are again forced to turn to the Entry Draft in order to replenish their stocks with younger and more cost effective options. Furthermore, once drafted and signed, NHL teams holds the exclusive rights to a player for the player’s first 7 full professional seasons, or until the player in question turns 27 (whichever of the two events occurs first), adding even greater weight to the importance of the draft, as smaller market teams may have difficulty retaining the rights to existing players, or even attracting new talent, once these exclusive negotiation rights expire.

## Prior Work

It has been widely documented that advanced statistical and predictive analysis have gained a significant foothold in the modern era of major league baseball, but more recently these methods have also become increasingly prevalent in the realms of professional basketball, professional football, and even more recently, in the world of professional hockey. Teams have leveraged this kind of advanced analysis in an attempt to gain new insights, challenge conventional wisdom, and identify areas where marginal gains may still lay unexploited. While it is difficult to say what kind of inroads have been made to-date in hockey’s proprietary domain, there has been a great deal of work published in the public domain that has re-shaped the way most interpret the modern game. In this section we will briefly touch on some of these works, highlighting those that focus predominantly on predictive analysis vis à vis the NHL draft and NHL prospects.

One of earliest such works, and perhaps the most widely cited piece of work in this domain, is the National Hockey League Equivalency model (NHLe) developed by Gabriel Desjardins.[[1]](#footnote-1) Desjardins’ model is largely based on the methodology first presented by Bill James in his 1985 book *Baseball Abstract,* where James detailed a method for predicting how a baseball prospect’s hitting statistics would carry over as they moved up through the various minor leagues to the major league level.

In the context of the NHL Entry Draft and evaluation of NHL prospects, this type of analysis is especially important, as prospects often hail from a multitude of feeder leagues, most of which vary greatly in terms of age range, and skill level. These feeder leagues can be divided primarily into two main subsets. The first is the North American feeder leagues, where the top developmental leagues in terms of producing the highest volume of NHL talent include the National Collegiate Athletic Association (NCAA), the United States Hockey League (USHL/USDP), and the Canadian Hockey League (CHL), which itself is further broken down into three regional leagues – the Ontario Hockey League (OHL), the Western Hockey League (WHL), and the Quebec Major Junior Hockey League (QMJHL). The CHL ranges in age from 16-20, the USHL from ages 16-21, and the NCAA mostly from ages 19-25. North America is also host to the NHL’s two professional farm leagues - the AHL and ECHL. However, in order to be eligible to play in either of these leagues, a player must be 18 years of age or older and must be signed to a professional contract, meaning no draft eligible prospects can be found in either league. Among European leagues, the top junior leagues primarily include those in the Czech Republic (Czech U20), Sweden (SuperElit), Finland (Jr. A SM-liiga), and Russia (MHL). Player ages in these leagues again range primarily from ages 16-20. A small portion of European prospects also hail from certain professional leagues, which unlike their counterparts in North America, do allow for players under the age of 18 to participate. The list of Europe’s primary professional leagues includes the first and second divisions in each of Sweden (SHL/Allsvenskan), Russia (KHL/VHL), Finland (Liiga/Mestis), and the Czech Republic (Czech/Czech2), along with the first division leagues in Switzerland (NLA), and Germany (DEL). Together these aforementioned leagues present only a small subset of the numerous feeder leagues from which NHL teams may draw draft eligible players in any given year.

In order to compare prospects from across this multitude of leagues, we must first normalize player production in order to account for the variance in skill level across leagues. The NHLe approach offers a way to do so by looking at all players who played significant time in a given league one year, and who then went on to play significant time in a different league in the year following that, allowing for direct comparison of their production rates between leagues in order to determine how much a point in one league would hypothetically be worth in another. Desjardins chose the NHL as his baseline league for which every other league would be compared, using points per game as the chosen metric for evaluating production rates. This resulted in an NHLe translation factor for each league, which offers an estimate of the portion of a player’s production rate we can expect them to retain when transitioning from the league in question directly to the NHL. These translation factors not only allow us to compare the strength of any given league in direct comparison to the NHL, but also allow us to indirectly compare the feeder leagues amongst themselves, resulting in the formulation of a basic hierarchy that defines the relative strength of every league. This hierarchy as defined by Desjardins’ model is depicted in Figure 1, along with his model’s resulting translation factors.



Figure 1

Unsurprisingly, Desjardins found that the various professional leagues are much closer in strength to the NHL than the junior leagues. Based on the chart in Figure 1, we would estimate a draft eligible prospect transitioning directly from the OHL to the NHL would retain only 30% of their draft year production, whereas a prospect drafted out of the Czech league would be expected to retain 74% of their production. It’s easy to see how adding this type of context can be useful when dealing with the analysis of prospects.

Building off of Desjardins’ work, Josh Weissbock and Cam Lawrence later developed the Prospect Cohort Success model (PCS),[[2]](#footnote-2) which aside from the NHLe model, is likely the second most widely cited piece of work in the domain of NHL Draft analysis. The idea behind Weissbock and Lawrence’s model was to generate a cohort of historical comparables for each prospect in their dataset, based on a small number of inputs they had found to be statistically significant. The inputs selected for the PCS model included age (defined as the exact age of a player, down to the day), both unadjusted and NHLe adjusted points per game from a player’s 18 year old draft season, league of origin, and height. Using these inputs, they then computed the Euclidian distance between each point in their dataset, and used these distances to group prospects into cohorts based on their nearest comparables. Then, using these cohorts in combination with known knowledge about the careers of the players in each cohort, Weissbock and Lawrence were able to compute both the probability that any new player assigned to that cohort would go on to play 200 NHL games, along with their expected career points per game rate at the NHL level if they were to reach that 200 game threshold. The combination of these two metrics provides an expected value of sorts for each prospect. For example, if a prospect is assigned to a cohort that dictates they have a 25% chance of reaching the 200 game threshold, and an expected career points per game rate of 0.45 (or 36.9 points per 82 game season) based on their nearest comparables, then that player’s expected value would be equal to 0.25 \* 36.9, or 9.225. Once these probabilities were computed, Weissbock and Lawrence ran tests to compare the predictive ability of their model in comparison to a simple multivariate linear regression model using the same inputs (age, height, and points per game rates). The results showed that their model tended to be a significantly better predictor of whether or not a player would go on to play 200 NHL games if that player originated from the CHL, but as can be seen in Figure 2, using the r-squared values provided as a metric for evaluating predictvity, the PCS model fared no better than the basic regression model, and in some cases even acted as a worse predictor.[[3]](#footnote-3)

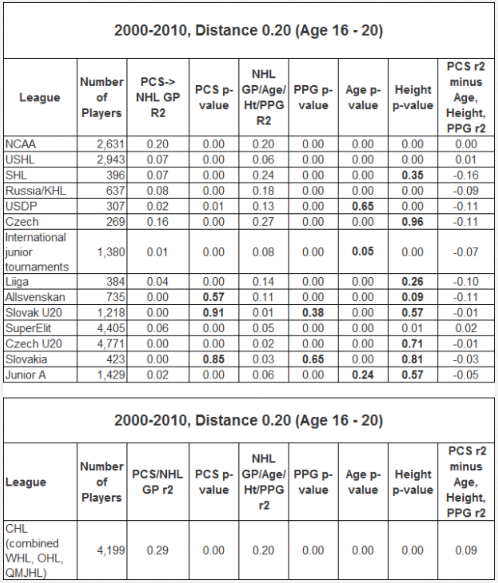


Figure 2

# Objective

The goal of this study is to both replicate, and also to build upon some of the aforementioned works. To begin, we wish to improve upon Desjardins’ original NHLe model, by not only exploring players who transition directly from any given league to the NHL in subsequent seasons, but also by exploring transitions between intermediate feeder leagues as well. This should offer a greater sample size to base our computations off of, allowing us to not only compute more accurate equivalency factors for any given league, but also to compute equivalency factors for a number of additional leagues which Desjardins’ method was unable to cover due to the sheer lack of direct links between those leagues and the NHL.

We will then use these newly computed league equivalency factors to normalize the pre-draft production rates for our sample of players, which we will in turn use to reproduce an unsupervised clustering model similar to the PCS model developed by Weissbock and Lawrence. By replicating their methodology, we hope to reassess the validity of whether such an approach can offer any additional predictive power that goes beyond traditional model inputs such as height, points per game, and age.

Finally, we will train and test various supervised classifiers using different combinations of inputs, in an attempt to establish a possible upper bound on the predictivity of future success at the NHL level, defined as some combination of total number of NHL games played, and/or career points per game rates.

# Methodology

## Data Collection

All data collection was done in large part thanks to the API made available by eliteprospects.com. Their extensive database not only offers stats from the NHL level, but also from the various junior and intermediate feeder leagues as well. By querying their API repeatedly and parsing the various json responses, we were able to construct a database containing 493,590 unique players spanning 1,321 different leagues.

The database includes basic information such as a player’s name, birthdate, nationality, height, weight, and preferred playing position. It also contains player statistics, broken down both by season and career totals. Figures 3, 4, and 5 offer an example of how the data is stored.



Figure 3

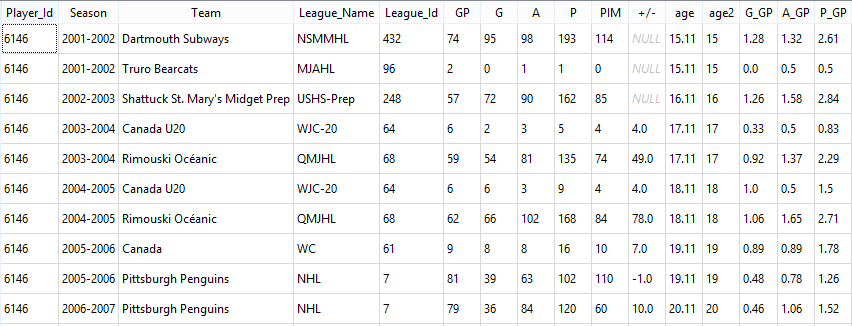


Figure 4

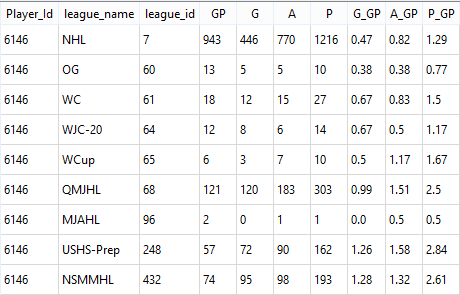


Figure 5

The database also includes additional details regarding each player’s career accolades, such as how many NHL All-Star teams they were named to over the course of their careers, and how many times they participated in the Under-18 or Under-20 World Championships prior to their first draft eligible season. The idea behind tracking the latter is that it might possibly act as a good indicator for identifying players at the top of their age groups at the junior level. Finally, the database includes historical information pertaining to the NHL Entry Draft itself, including draft year, round & overall selection, along with the selecting team, and the player selected with each pick.

Additional statistics that could potentially be of use, such as shot totals and ice time figures, are unfortunately not tracked by eliteprospects, as only a small number of junior leagues track and make these figures available to the public. As a result we decided to omit them entirely for the purposes of this study. This alludes to the first major roadblock standing in the way predictive analysis pertaining to the NHL Draft, which is that the predictive ability of any such models is hindered by the availability (or lack thereof) of model inputs beyond the rudimentary games played, goals, and assists.

## Computing of League Equivalency Factors

With the data collection step out of the way, we have what is required to compute our own version of Desjardins’ league equivalency factors, which will be used to normalize the production rates of prospects across leagues. Before detailing the methodology behind these computations, it is first import to acknowledge a few of the shortcomings presented by the original NHLe model. While Desjardins’ methodology focused only on players who jumped directly from any given feeder league straight to the NHL, in reality many players travel up the hierarchy of leagues, transitioning between multiple intermediate leagues before finally making the transition to the NHL. At lower levels of the hierarchy, such as the junior level, by looking only at players who jump directly to the NHL we limit our sample by ignoring these intermediate links between the feeder leagues themselves. By factoring in intermediate transitions, not only does it improve our sample size, allowing us to compute a more representative set of equivalency factors, but it also allows us to capture additional leagues that Desjardins’ model was unable to account for, simply due to the inexistence of direct links between those leagues and the NHL.

The first step in computing our equivalency factors is to identify all records in our database for which a player changed leagues in subsequent seasons. Then for each pair of records (*league n, league n+1*) for which players transitioned from one league to the other in consecutive seasons, we calculated how much of the sample’s cumulative points per game rates from year *n* were retained in year *n + 1* when moving between leagues. This offers an estimate of how much a point in *league n* might be worth in *league n + 1*.

In order to ensure that our sample was not skewed by any outliers, we restricted the query to the subset of leagues for which a minimum of 50 players historically transitioned from one league to the other in subsequent seasons.

Next, we constructed a directional graph using each league as a node, and the initial LE factors we previously computed as edge weights between nodes. Here is where our methodology encounters the same issue as the NHLe model, however. After constructing the graph, we are left with only 13 of the 1,321 leagues for which an out edge exists leading directly to the NHL node. Figure 6 depicts a subset of these 13 leagues, including many of the feeder leagues encompassed by Desjardins’ original work. While the sample may be small, one positive takeaway is that the results seem to corroborate the prior findings of Desjardins, as we observe a similar hierarchy to that depicted in Figure 1.



Figure 6

In order to overcome the issue of sample size, we would like to incorporate all intermediate transitions between feeder leagues into our calculations as well, so that we are no longer limited only to direct links between the NHL node and other nodes. To accomplish this, we will look at all paths in our graph of length less than or equal to size 3 which end at the NHL node. Then we take the product of each edge weight along these paths, weighting paths of size 1 more heavily than those of size 2, and paths of size 2 more heavily than those of size 3. As an example, while an edge between the USHL (one of the aforementioned premier North American junior leagues) and the NHL may not exist, an edge may exist in our graph between the USHL and the AHL. If we assume this edge has weight 0.5, telling us that players from the USHL generally retain 50% of their production when transitioning to the AHL, and we know from our chart in Figure 6 that AHL players typically retain 48% of their production when moving to the NHL ranks, we can thus surmise players from the USHL should be expected to retain 0.5 \* 0.48 = 24% of their production if they were to jump directly to the NHL level. We would then repeat this process for all other paths of at most length three that lead from the USHL node to the NHL node, and then calculate a weighted average using the resulting products, leaving us with a final LE factor.

Using this approach, our equivalency factors were re-computed, this time producing a much larger resulting set of 432 leagues for which we now have a means of comparison. A small snapshot of these updated LE factors is outlined in Figure 7.



Figure 7

With our equivalency factors now computed, the next step is to normalize the pre-draft production rates for all players in our database. Because the draft eligibility cut-off specifies an 18 year old birthdate of September 15th for each draft year, meaning players born between September 16th and December 31st will have played three full seasons in the junior ranks before their first eligible draft year, and due to the fact that most prospects do not graduate to the major junior ranks until two years prior to their draft year, in an attempt to bypass this added complexity we will only include a player’s 17 and 18 year old seasons as inputs to our model. This brings us to another one of the main challenges facing this type of predictive modeling of the NHL draft, which is that inputs are generally restricted to only one or two years’ worth of meaningful data.

After normalizing each player’s 17 and 18 year old goal and assist rates using our newly generated league equivalency factors, there is still the remaining challenge of normalizing production across eras within each league. The motivation in doing so is that even within the same league, scoring rates can vary greatly by era, so one point scored in one season may not be equivalent to a point scored in another year, or decade. For example, the QMJHL’s leading scorer this past year finished the season with 111 points, whereas in the 1999-2000 season the leading scorer finished with 186 total points, with a dozen players matching or eclipsing the 111 point mark. It is possible that the talent crop of 1999-2000 was simply more skilled than that of the modern era, but it is even more likely that there are additional underlying factors causing these large variations in scoring rates, an example of such being the significant advances of the goaltending discipline over the years.

We can adjust for this type of variance in league scoring rates between years, by first computing the average points per game rate for each league across all seasons available in our dataset, and then comparing the yearly average scoring rates against this all-time average, in order to identify historical trends in scoring rates across each league. As an example, using this methodology we can compute that QMJHL scoring in the 1999-2000 season was 10-11% higher than the all-time average. Based on this we would regress each player’s production from that season back toward the all-time mean. Using our aforementioned leading scorer from the 1999-2000 season as an example, his adjusted point total would then be equal to 186 \* 0.89 = 166.

k-means Clustering

With our pre-draft production rates now league and era adjusted, we can move on to constructing a replication of the PCS model. In order to do so, we will use an unsupervised k-means clustering approach to identify cohorts of comparable players, based only off of each player’s pre-draft normalized production rates. The reason for selecting the k-means clustering algorithm (aside from the fact that it is relatively easy to implement) is that, similarly to Weissbock and Lawrence’s model, it too uses the Euclidian between points to define similarity. The normalized age 17 and age 18 goal and assist rates of each player will serve as our model inputs in this case, and to add an additional layer of complexity, the age 18 production inputs will be weighted twice as heavily as the age 17 production inputs in order to add additional emphasis to the more recent of the observations.

Prior to running the k-means algorithm, we must first evaluate whether our data lends itself to clustering in the first place, along with estimating an optimal number of clusters based on the underlying composition of our dataset. By mapping our age 17 and age 18 production rates to the 2D plane (Figure 8), it appears as though we cannot generate any meaningful insights at a glance.

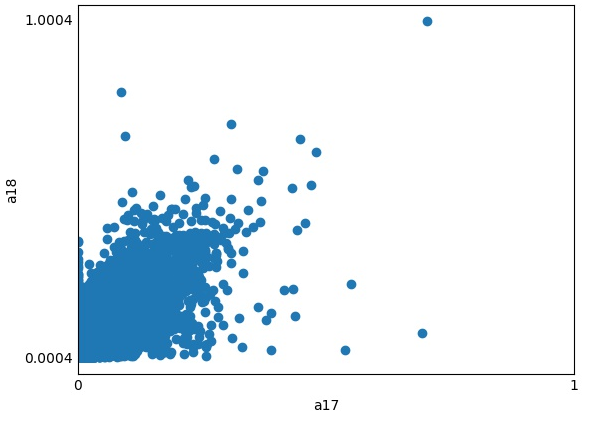


Figure 8

Thankfully there are heuristics that can be used to estimate the optimal number of clusters instead of relying on our intuition. One of the most popular such methods is the elbow method, which computes variance within clusters based on their SSE, and then maps a cost function based on those values, where the optimal number of clusters is said to be the point on the x axis around which a visible “elbow” appears. Figure 9 demonstrates the resulting elbow method when applied to our dataset.

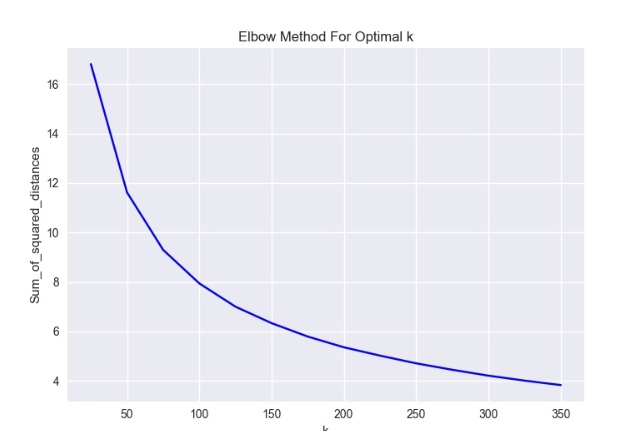


Figure 9

Unfortunately the results yield no visible elbow, but rather a relatively smooth curve, which again does not bode well for a cluster based approach. However, this does not necessarily mean that our dataset does not lend itself to clustering. Instead we can experiment with various levels of clustering to see if any level of partitioning can yield significant results when training our models.

## Design and Data Pre-Processing

With our clustering model built and our pre-draft production rates fully normalized, we have all the inputs required to put the finishing touches on the design of our classifiers before moving on to training, testing, and tuning the models in order to analyze the results. For our purposes, six classification models were hand-picked (all from Python’s sklearn library), all of which take varying approaches to the problem of classification. The six selected models include the standard Gradient Boosting Classifier, Support Vector Classifier, Multilayer Perceptron, Random Forrest Classifier, K-Nearest Neighbours classifier, and the Naïve Bayes classifier. We will compare model performance across these various approaches, and select from among the best performers. Once these top performers have been identified, we may then choose to further refine our approach by trying out various ensembles from among a subset of our best individual performers, to see if we can possibly improve accuracy further. Separate classifiers will be trained and tested for both forwards and defensemen, with goalies being set aside for further study at a later date.

The inputs selected for our classifiers include traditional metrics such as height, weight, age at time of draft, country of origin, league of play (age 17 and 18), goals and assists per game (age 17 and 18), along with the binary identifiers indicating whether or not a player appeared in the World Under-18 or World Under-20 Championships at any point prior to their draft year. We will first train our models separately using adjusted and unadjusted goal and assist rates, which will hopefully allow us to assess if any additional predictive capability is being added by our league and era adjustments. Next, we will introduce our cluster IDs as an additional input feature, using separate runs of 50, 100, and 200 clusters to see if incorporating the cluster data at any level of partitioning adds any gains in performance to our base models. When training each model, we will apply feature selection to weed out any inputs which may not act as meaningful predictors, and would only serve to add noise if included.

Some data pre-processing of our inputs is required before we can split our sample into train and test sets. Any ordinal inputs, such as goal and assist rates, height, weight, and age have to be standardized in order to normalize their distributions, and reduce variance. Thankfully, the sklearn library in Python offers a Standard Scaler function that allows for easy normalization of features, which we will make use of for this part of our pre-processing. Next, the one hot encoding method must be applied to any categorical inputs (country of origin, league, cluster ID, etc.) that don’t lend themselves to standardization. If this method is not applied to our inputs, we risk having the machine learning algorithms treat this categorical data as ordinal instead. One hot encoding avoids this error by transforming categorical data to a numeric representation where necessary (most algorithms do not possess the capacity to process text inputs), and then mapping the data to multiple fields (one for each category), where membership to any individual category is indicated using binary classification.

Using our selected inputs, each model will be trained to predict the probability of the following binary classifiers:

1. The probability a given prospect will play at least 1 NHL game.
2. The probability a given prospect will play more than 2 full NHL seasons (164 games)

*and* achieve a career points per game average greater than or equal to 0.33 in the case of forwards, or 0.25 in the case of defensemen.\*

\*These point per game thresholds were heuristically selected as a means of classifying “top 9” production among forwards and “top 4” production among defensemen, largely based on last season’s distribution of scoring for each position at the NHL level.

Our classifiers will be trained and tested on a sample of 12,149 forwards and xxx defensemen aged 25 and over. A quick analysis of the sample sets shows that of our sample of 12,149 forwards aged 25 and older, only 1,206 (or just shy of 10%) went on to play only a single game, and of those only 354 (under 3%) went on to become top 9 forwards based on the criteria specified previously. Among our sample of xxx defensemen aged 25 and older, only xxx (or just shy of xxx %) went on to play only a single NHL game, and of those only xxx (under xxx %) went on to produce at the rate expected of a top 4 defenseman.

Based on these figures, it is evident that an imbalance exists in our sample sets between classes. When data sets exhibit this type of imbalance in classification problems, we can no longer rely on a standard accuracy, which uses the rate of true positives (TP) and true negatives (TN), defined as as a measure of evaluation. Take our extreme case of imbalance as an example, where less than 3% of our observed sample of forwards develop into top 9 producers at the NHL level. If our classifier were to simply predict that no forward would ever be classified as a top 9 producer, it would exhibit an accuracy of over 97%, but we would not be left with any meaningful results. Instead, when dealing with imbalance we must look to alternative metrics of evaluation, namely precision and recall. Precision refers to the ratio of positive predictions that are observed positives, and is defined as whereas recall refers to the ratio of observed positives that a model classifies as true positives, defined as The F1 score offers a harmonic mean of these two metrics, and is defined as

As a result of the underlying imbalance in our sample set, we will rely largely on the F1 score when evaluating the performance of each model.

Along with alternative methods for evaluating model performance on imbalanced data sets, there exist certain unconventional sampling methods that can help improve model performance in the face of imbalance. These approaches include purposefully oversampling the underrepresented class in order to gain an equal representation for each class, where existing instances are either duplicated, or used as a basis in order to artificially generate new instances,. On the other hand, undersampling can also be employed in an attempt to counteract imbalance. With undersampling, only a subset of the overrepresented class is taken into account in an attempt to introduce balance to the class structure. We will test both of these methods when building our models to see if they can offer any additional performance gains by combatting the imbalance in our sample sets. To implement these methods, we will make use of the Python imblearn library, using the NearMiss approach for our undersampling method, and the SMOTE approach for our oversampling method, which achieves oversampling through interpolation.

Before proceeding to the train/test step, we must first split our samples into separate train and test sets. In this case, we opted for a test size of 30% of the overall sample set, as well as to split in a stratified fashion in order to ensure the test and train sets are equally representative. The decision to stratify our test and training sets is due to the fact that our data set exhibits a strong imbalance, as was alluded to earlier. Stratifying will ensure that a representative portion of the underrepresented class is contained within our test set.

# Training, Testing, and Model Selection

## Predicting Whether a Player Will Make the NHL

### Forwards

First, we trained and test each model using our unadjusted scoring rates, yielding the results from Figure 10. The Naïve Bayes classifier is the early top performer, but none of the models showed much in the way of predictive capacity.



Figure 10

Re-training our models using our adjusted scoring figures yields much better results (Figure 11). Again the Naïve Bayes classifier performs best, but all classifiers exhibit similar performance in this case.



Figure 11

Next we retrain/test our models with the cluster groupings added in as inputs. Based on the results shown in Figure 12, we see that adding in the clustering does not seem to affect predictivity.



Figure 12

Finally, we explore whether or not it is possible to improve our model performance further by applying both undersampling and oversampling methods to counteract our sample imbalance.

Figure 13

Figure 14

At first glance the results seem promising, with both methods outperforming our base model. As we would expect, undersampling manages to capture a slightly larger portion of true positives, but this gain comes at the expense of a significantly larger portion of false positives, leading to lower accuracy. Based on this we can conclude that the oversampling method using our normalized scoring rates produces the best results. A small number of ensemble methods from among the best performing models (Gradient Boosting, Multilayer Perceptron, Random Forest, KNN) were tested to see if we might be able improve our model’s performance further, but none seemed to offer significant gains.

Using our model’s results, the top 5 ranked players from our test are listed in Figure 15. This list is composed of three former 1st overall draft choices (Lecavalier, Thornton, Legwand), along with a 2nd (Van Riemsdyk), and 7th overall selection (Voracek). At a glance it seems to pass the sniff test.

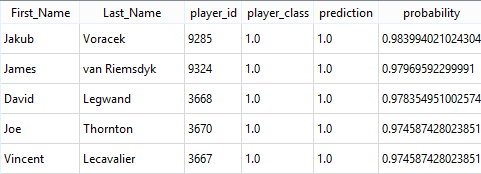


Figure 15

Of course, the model is not without its misses, though admittedly they appear to be few and far between. Figure 16 highlights some of the most notable false negatives. What’s interesting NHL teams also seemed to misevaluate the same players in their draft seasons. Langenbrunner and Karlsson were the highest selections of the group, both going in the 2nd round, whereas Gaudreau was not selected until the 4th round, Shaw and Versteeg the 5th round, and Marchessault went unselected entirely.

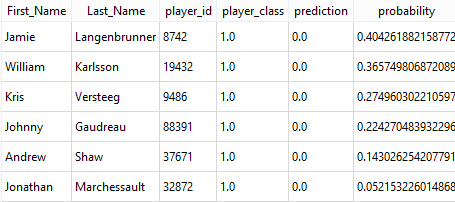


Figure 16

## Identifying Top 9 Forwards and Top 4 Defensemen

### Forwards

Here we shift our focus, now not only trying to predict the probability a prospect will meet a minimum games played threshold (now 165 games), but also that they will post a career points per game rate of 0.33, or greater. Training and testing our models with the raw (Figure 17) and adjusted data (Figure 18), we again see a gain a marked improvement in predictive power by using the league and era adjusted figures. 

Figure 17



Figure 18

Despite the improvement provided by using our adjusted figures, the rate or false negatives is still significantly high. Before we test our over and undersampling methods to try to correct for this, we first want to have a look to see if including our clusters will add any additional predictive gains. Figure 19 shows that again we don’t receive any added performance gains by introducing our cluster values into the model.



Figure 19

Finally, we try oversampling and undersampling to see if they might fare any better than our base model.

# Conclusions and Future Work

# Works Cited

Desjardins, G. (n.d.). *Projecting Junior Hockey Players and Translating Performance to the NHL*. Retrieved from Behind the Net: http://www.behindthenet.ca/projecting\_to\_nhl.php

Lawrence, C. (2015). *Prospect Cohort Success – Evaluation of Results*. Retrieved from Hockey Graphs: https://hockey-graphs.com/2015/09/21/prospect-cohort-success-evaluation-of-results/

Saxena, S. (2018). *Precision vs Recall*. Retrieved from Towards Data Science: https://towardsdatascience.com/precision-vs-recall-386cf9f89488

Weissbock and Lawrence. (2015). *Draft Analytics: Unveiling the Prospect Cohort Success Model*. Retrieved from Canucks Army: https://canucksarmy.com/2015/05/26/draft-analytics-unveiling-the-prospect-cohort-success-model/

1. (Desjardins) [↑](#footnote-ref-1)
2. (Weissbock and Lawrence, 2015) [↑](#footnote-ref-2)
3. (Lawrence, 2015) [↑](#footnote-ref-3)