Predictive Modelling and the NHL Entry Draft

Alex Church

100825475

COMP 4905, FALL 2019

Supervisor: Dr. Tony White

Contents

Introduction

Background

At the conclusion of every season, the NHL holds its annual Entry Draft, where teams alternatingly select the rights to junior-aged prospects in the hopes of identifying prospective NHL talent. The modern day draft is composed of seven rounds, where each of the 31 franchises are allotted one selection per round, and the annual order of selection is 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 spanning 12 nationalities were selected collectively by the NHL's 31 franchises.

Each year, the pool of NHL draft eligible prospects is drawn from a number of junior and professional feeder leagues from across the world, all varying in terms of age range and skill level. This group of feeder leagues can be broken down into two main subsets – those based in North America, and those based in Europe. The top North American developmental leagues in terms of producing the highest volume of draft eligible talent include the National Collegiate Athletic Association (NCAA), the United States Hockey League (USHL), 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 from ages 16-20, the USHL from ages 16-21, and the NCAA mostly from ages 19-25. North America is also home to the NHL's two affiliated professional farm leagues - the AHL and the ECHL. However, in order to be eligible to play in either the AHL or ECHL, a player must be signed to a professional contract, meaning teams in these leagues are not permitted to roster draft eligible players.

Overseas in Europe, the top junior leagues in terms of producing draft eligible talent include those located in the Czech Republic (Czech U20), Sweden (SuperElit), Finland (Jr. A SM-liiga), and Russia (MHL). Player ages in these junior leagues again range mostly from ages 16-20. A portion of European draft eligible prospects also hail from a select group of professional leagues, who unlike their counterparts in North America, share no direct affiliation to the NHL and its franchises. As a result, teams in these leagues are permitted to roster players eligible for selection in the NHL draft. The European professional leagues that tend to produce the highest volume of draft eligible talent include 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 represent only a fraction of the dozens of junior and professional leagues around the world from which NHL teams may identify draft-worthy talent in any given year.

In order to be eligible for selection in the NHL's Entry Draft, a prospect must have turned 18 years of age 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 years of age 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 thus been considered eligible for a previous year's draft, meaning they are only considered 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. For the purposes of this study, we will focus only on first time eligible prospects.

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 attempt 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 (also 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 tend to increase, teams are again forced to turn to the Entry Draft in order to replenish their stocks with younger and more cost effective options, creating a cycle whereby sustained success is often heavily tied to a team's success at the draft table. Furthermore, once drafted and signed, NHL teams hold the exclusive negotiation 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). This adds even greater weight to the importance of the draft, as once these exclusive negotiation rights expire, smaller market teams may have difficulty retaining the rights to existing players, along with attracting new talent.

Related Work

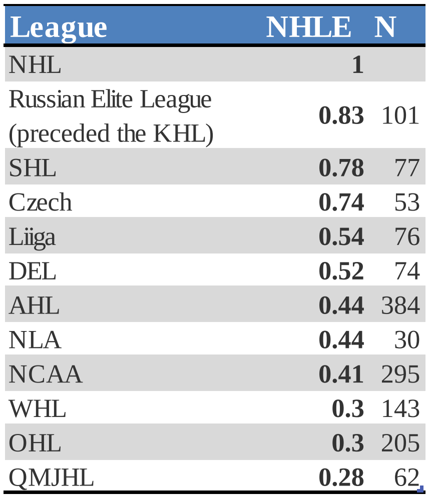
It has been widely documented that advanced statistical and predictive analysis have gained a significant foothold in the modern era of major league baseball, but more recently these methods have also become increasingly prevalent in the realms of professional basketball, professional football, and even more recently, in the world of professional hockey. Teams have begun to leverage this kind of advanced analysis more frequently 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 speculate on 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 many interpret the modern game. In the section to follow, we will briefly touch on some of these works, highlighting those that have focused predominantly on predictive analysis vis à vis the NHL draft.

One of earliest such works, and perhaps the most widely cited piece of work in this domain to-date, is the National Hockey League Equivalency model (NHLe) developed by Gabriel Desjardins.[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 projecting how a baseball prospect's hitting statistics would carry over as they graduate through the various minor league ranks to the major league level.

In the context of the NHL Entry Draft, where the aim is to compare and rank prospects from across the NHL's wide number of feeder leagues, a method for normalizing player production rates to account for the variance in skill level from league to league is of critical importance. The NHLe approach proposed by Desjardins 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 another league in the year following that, allowing for direct comparison of their production rates across leagues from one season to the next. Desjardins chose the NHL as his baseline league for which every other league would be measured, and by looking only at the sample of players who jumped from one league directly to the NHL in subsequent seasons, he was able to compute a list of translation factors that represent an estimate of how much a point in

[1] (Desjardins)

each league would roughly be worth at the NHL level. These NHLe translation factors not only allow us to compare the strength of any given league directly to that of the NHL, but also allow us to indirectly compare the feeder leagues amongst themselves, resulting in the formation of a basic hierarchy that defines the relative strength of every league in the sample. This hierarchy, as defined by Desjardins' model, is depicted in Figure 1.



League	NHLE	N
NHL	**1**	
Russian Elite League (preceded the KHL)	**0.83**	101
SHL	**0.78**	77
Czech	**0.74**	53
Liiga	**0.54**	76
DEL	**0.52**	74
AHL	**0.44**	384
NLA	**0.44**	30
NCAA	**0.41**	295
WHL	**0.3**	143
OHL	**0.3**	205
QMJHL	**0.28**	62

Figure 1

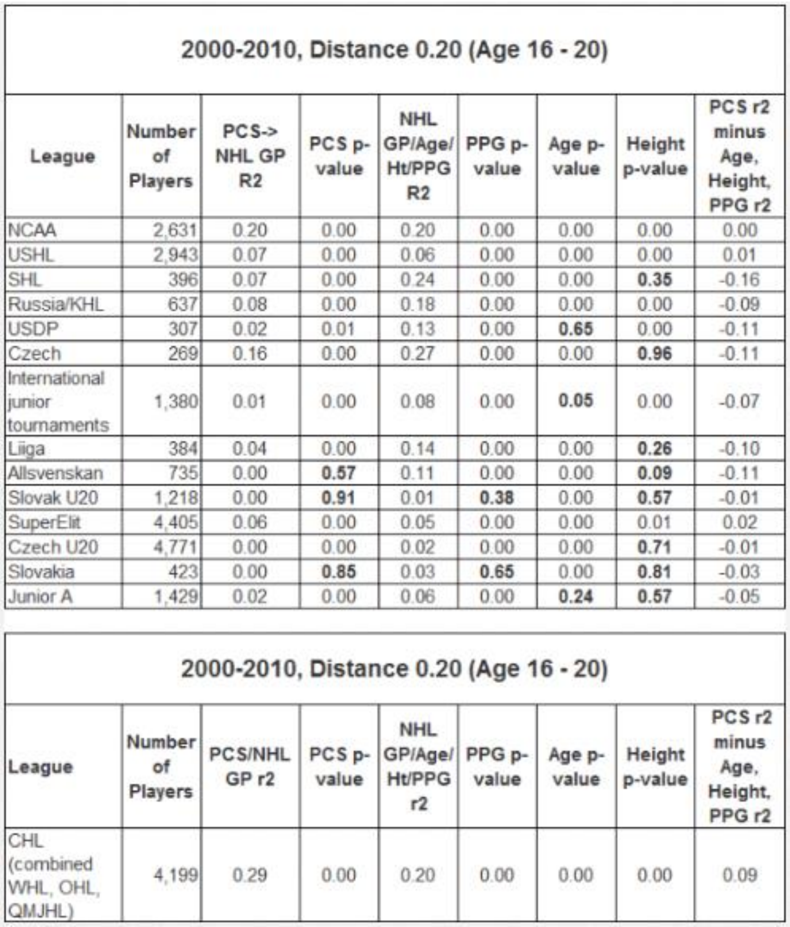
Unsurprisingly, Desjardins found that the various professional leagues (Russian Elite League, SHL, Czech, Liiga, DEL, AHL, NLA) are much closer to the NHL in terms of relative skill level than the top junior leagues (NCAA, WHL, OHL, QMJHL). As an example of how these translation factors might be applied, by using 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 pre-draft production if they were to make the jump to the NHL the following year, whereas a prospect drafted out of the Czech league would be expected to retain 74% of their production under the same circumstances.

Aside from the NHLe model, the second most widely cited piece of work in the public domain of NHL Draft analysis is most likely the Prospect Cohort Success model (PCS), developed by Josh Weissbock and Cam Lawrence.[2] The idea behind Weissbock and Lawrence's model revolves around generating a cohort of historical comparables for each prospect, which in turn can be used to predict the probability that any given prospect will go on to enjoy success at the NHL level. The predictors Weissbock and Lawrence deemed to be statistically significant in terms of predicting future NHL success included age on draft day (defined as the exact age of a player, down to the day), league of origin, height, and both unadjusted and NHLe adjusted point per game rates from a player's 18 year old pre-draft season. Using these inputs, they were able to compute the Euclidian distance between each point in their dataset, and then use those distances to group prospects into cohorts based on their nearest neighbours. Then, 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 a cohort would go on to play 200 NHL games, as well as 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 can be used in unison to provide an expected value for each prospect. As an 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 NHL season) based on their closest historical comparables, then that player's expected value would simply be computed as 0.25 * 36.9 = 9.3.

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, so long as that player

[2] (Weissbock and Lawrence, 2015)

originated from one of the CHL feeder leagues. However, based on the provided r-squared values outlined in Figure 2, in most cases the PCS model fared no better than the basic regression model in terms of predicting how many career NHL games any given prospect would go on to play, and in some cases even acted as a worse predictor.[3]



2000-2010, Distance 0.20 (Age 16 - 20)

League	Number of Players	PCS-> NHL GP R2	PCS p-value	NHL GP/Age/ Ht/PPG R2	PPG p-value	Age p-value	Height p-value	PCS r2 minus Age, Height, PPG r2
NCAA	2,631	0.20	0.00	0.20	0.00	0.00	0.00	0.00
USHL	2,943	0.07	0.00	0.06	0.00	0.00	0.00	0.01
SHL	396	0.07	0.00	0.24	0.00	0.00	**0.35**	-0.16
Russia/KHL	637	0.08	0.00	0.18	0.00	0.00	0.00	-0.09
USDP	307	0.02	0.01	0.13	0.00	**0.65**	0.00	-0.11
Czech	269	0.16	0.00	0.27	0.00	0.00	**0.96**	-0.11
International junior tournaments	1,380	0.01	0.00	0.08	0.00	**0.05**	0.00	-0.07
Liiga	384	0.04	0.00	0.14	0.00	0.00	**0.26**	-0.10
Allsvenskan	735	0.00	**0.57**	0.11	0.00	0.00	**0.09**	-0.11
Slovak U20	1,218	0.00	**0.91**	0.01	**0.38**	0.00	**0.57**	-0.01
SuperElit	4,405	0.06	0.00	0.05	0.00	0.00	0.01	0.02
Czech U20	4,771	0.00	0.00	0.02	0.00	0.00	**0.71**	-0.01
Slovakia	423	0.00	**0.85**	0.03	**0.65**	0.00	**0.81**	-0.03
Junior A	1,429	0.02	0.00	0.06	0.00	**0.24**	**0.57**	-0.05

2000-2010, Distance 0.20 (Age 16 - 20)

League	Number of Players	PCS/NHL GP r2	PCS p-value	NHL GP/Age/ Ht/PPG r2	PPG p-value	Age p-value	Height p-value	PCS r2 minus Age, Height, PPG r2
CHL (combined WHL, OHL, QMJHL)	4,199	0.29	0.00	0.20	0.00	0.00	0.00	0.09

Figure 2

The PCS model did not remain public domain for long, as it was subsequently purchased and made proprietary by the NHL's Florida Panthers. Similar models have been developed in the wake of PCS, a small number of which have also been purchased and employed by NHL

[3] (Lawrence, 2015)

franchises.[4] Unfortunately, without being privy to the current state of these now proprietary models, or to what degree they are being leveraged by teams as part of their decision making process, it is difficult to assess if the teams employing them have been able to gain a significant leg up on their competition when it comes to identifying prospective NHL talent at the draft table.

Objective

The goal of our study is to build upon some of the aforementioned works by incorporating their methodologies into a new approach toward predicting which NHL draft eligible prospects are most likely to go on to enjoy success at the NHL level.

As a first step, we wish to construct a feeder league equivalency model, similar to the original NHLe model proposed by Desjardins. The resulting league equivalency factors will be used to normalize the pre-draft production rates of all players in our sample, allowing for direct comparison of production rates across leagues.

Next, we wish to build an unsupervised clustering model, similar to the PCS model developed by Weissbock and Lawrence. This will provide us with a list of the closest historical comparables for each prospect in our set. By replicating the PCS methodology, we hope to reassess the validity of whether such an approach can offer any additional insights as to which prospects are most likely to go on to succeed at the NHL level, based only on the career trajectories of their nearest neighbours.

[4] (Columbus Blue Jackets, 2019)

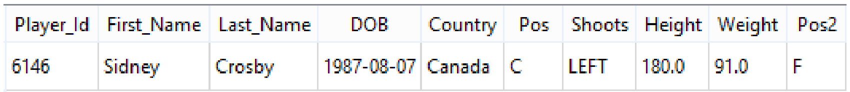
Finally, using our normalized production rates from step one, and our cluster values from step two, along with a short list of additional features, including each player's height, age, and country of origin, we wish to train and test various supervised classifiers in an attempt to establish an upper bound on our ability to predict which draft eligible prospects are most likely to go on to succeed at the NHL level, with success in this case being defined using metrics such as the probability any given prospect will go on to meet a minimum games played threshold, or eclipse a minimum career points per game mark.

Methodology

Data Collection

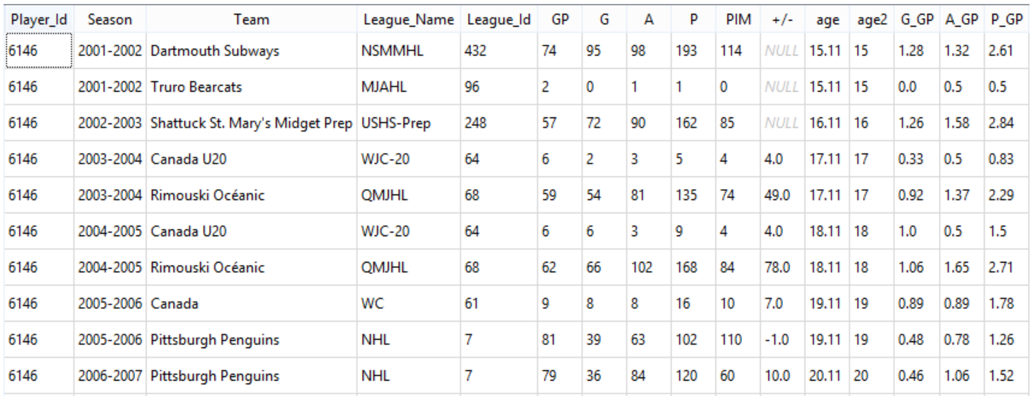
All data collection was done in large part thanks to the API made available by eliteprospects.com. Their extensive database not only offers a wide array of statistics from the NHL level, but also from the various junior and intermediate feeder leagues as well. By querying their API repeatedly and parsing the various 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.



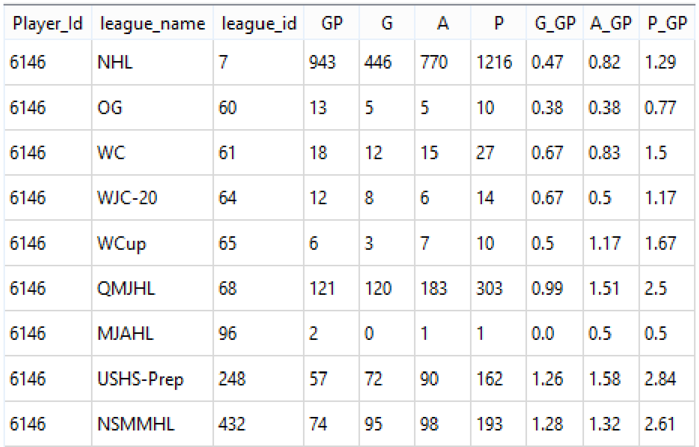
Player_Id	First_Name	Last_Name	DOB	Country	Pos	Shoots	Height	Weight	Pos2
6146	Sidney	Crosby	1987-08-07	Canada	C	LEFT	180.0	91.0	F

Figure 3



Player_Id	Season	Team	League_Name	League_Id	GP	G	A	P	PIM	+/-	age	age2	G_GP	A_GP	P_GP
6146	2001-2002	Dartmouth Subways	NSMMHL	432	74	95	98	193	114	NULL	15.11	15	1.28	1.32	2.61
6146	2001-2002	Truro Bearcats	MJAHL	96	2	0	1	1	0	NULL	15.11	15	0.0	0.5	0.5
6146	2002-2003	Shattuck St. Mary's Midget Prep	USHS-Prep	248	57	72	90	162	85	NULL	16.11	16	1.26	1.58	2.84
6146	2003-2004	Canada U20	WJC-20	64	6	2	3	5	4	4.0	17.11	17	0.33	0.5	0.83
6146	2003-2004	Rimouski Océanic	QMJHL	68	59	54	81	135	74	49.0	17.11	17	0.92	1.37	2.29
6146	2004-2005	Canada U20	WJC-20	64	6	6	3	9	4	4.0	18.11	18	1.0	0.5	1.5
6146	2004-2005	Rimouski Océanic	QMJHL	68	62	66	102	168	84	78.0	18.11	18	1.06	1.65	2.71
6146	2005-2006	Canada	WC	61	9	8	8	16	10	7.0	19.11	19	0.89	0.89	1.78
6146	2005-2006	Pittsburgh Penguins	NHL	7	81	39	63	102	110	-1.0	19.11	19	0.48	0.78	1.26
6146	2006-2007	Pittsburgh Penguins	NHL	7	79	36	84	120	60	10.0	20.11	20	0.46	1.06	1.52

Figure 4



Player_Id	league_name	league_id	GP	G	A	P	G_GP	A_GP	P_GP
6146	NHL	7	943	446	770	1216	0.47	0.82	1.29
6146	OG	60	13	5	5	10	0.38	0.38	0.77
6146	WC	61	18	12	15	27	0.67	0.83	1.5
6146	WJC-20	64	12	8	6	14	0.67	0.5	1.17
6146	WCup	65	6	3	7	10	0.5	1.17	1.67
6146	QMJHL	68	121	120	183	303	0.99	1.51	2.5
6146	MJAHL	96	2	0	1	1	0.0	0.5	0.5
6146	USHS-Prep	248	57	72	90	162	1.26	1.58	2.84
6146	NSMMHL	432	74	95	98	193	1.28	1.32	2.61

Figure 5

The database also includes additional details regarding each player's career accolades, such as how many times they were selected to participate in the Under-18 or Under-20 World Junior Championships prior to their first draft eligible season. The idea behind tracking a feature such as this is that it serves as an identifier for draft eligible players who rank at the top of their junior-aged peer group, which might offer an indication of prospects who are more likely to go on to greater success as they graduate from the junior ranks. Finally, the database includes historical information pertaining to the NHL Entry Draft itself, dating back to the 1995 draft, including all draft selections broken down by year, round, overall selection, as well as 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 made available by eliteprospects, as only a small number of feeder 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 of predictive analysis pertaining to the NHL Draft, which is that the predictive ability of any such model is hindered by the availability (or lack thereof) of model inputs beyond the standard games played, goals, and assists, which are available for all leagues.

League Equivalency Factors

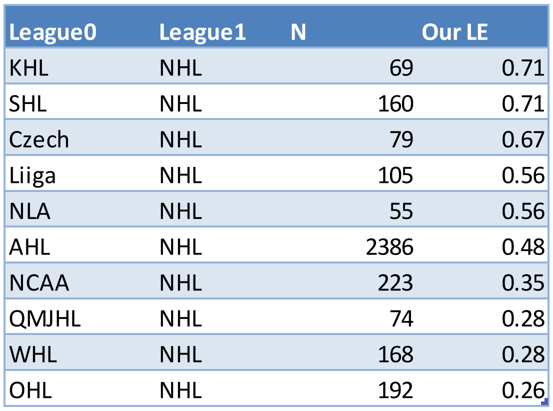
With our data collection complete, we have what is required to begin computing our own version of Desjardins' league equivalency factors. Before detailing the methodology behind these computations, it is first important to acknowledge a few of the shortcomings presented by the original NHLe model. While Desjardins' methodology focused only on players who jumped directly from any given feeder league straight to the NHL, in reality, many players transition between multiple intermediate leagues before finally making the transition to the NHL level. At lower levels of the hierarchy, such as the junior ranks, 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, which often represent the paths most commonly travelled. 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 level.

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 tuple of the form (league *n*, league *n+1*), representing a player who transitioned from one league to another in consecutive seasons, we can group our sample by each pairwise distinct set of leagues in order to calculate how much of our sample's cumulative points per game rates from year *n* were retained in year *n + 1* when moving between leagues *n* and *n + 1*. This provides an estimate for how much a point in league *n* might be worth in league *n + 1*, computed as follows:

$$LE_{(league\ n, league\ n+1)} = \frac{\left(\frac{\sum points\ year\ n}{\sum games\ played\ year\ n}\right)}{\left(\frac{\sum points\ year\ n+1}{\sum games\ played\ year\ n+1}\right)}$$

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 league *n* to the league *n + 1* in subsequent seasons. Each player must have also played a minimum of 20 games in both seasons *n*, and *n + 1* in order to be included in the sample.

Next, we construct a directional graph, where each league is represented by a node, and the initial LE factors we previously computed are represented as edge weights between nodes. It is at this point where our methodology begins to encounter some of the issues that plagued the NHLe model, as after constructing the graph, we are left with only 13 of a possible 1,320 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.

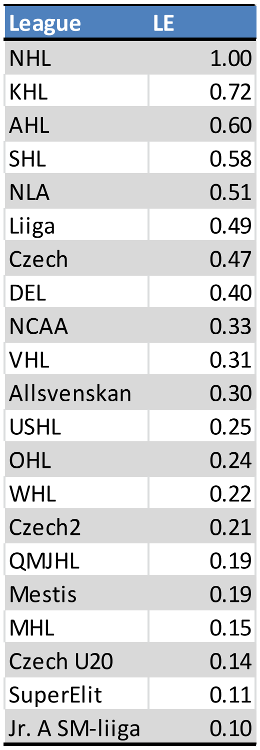


League0	League1	N	Our LE
KHL	NHL	69	0.71
SHL	NHL	160	0.71
Czech	NHL	79	0.67
Liiga	NHL	105	0.56
NLA	NHL	55	0.56
AHL	NHL	2386	0.48
NCAA	NHL	223	0.35
QMJHL	NHL	74	0.28
WHL	NHL	168	0.28
OHL	NHL	192	0.26

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 to the NHL node. To accomplish this, we look at the neighbourhood of paths in our graph of length less than or equal to size three, which conclude at the NHL node. For every league in that neighbourhood, we first compute the product of all edge weights along each path leading from the league in question to the NHL node, and then compute a weighted average of all paths for that league, weighting paths of size one more heavily than those of size two, and paths of size two more heavily than those of size three. 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 NHL's farm league, 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 length three or less that lead from the USHL node to the NHL node, and then calculate a weighted average based on each path, leaving us with a final league equivalency (LE) factor for the USHL.

Using this approach, our equivalency factors are 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. The observed hierarchy is again quite similar to that discovered by Desjardins initially, only now our hierarchy is made up of a much larger subset of feeder leagues.



League	LE
NHL	1.00
KHL	0.72
AHL	0.60
SHL	0.58
NLA	0.51
Liiga	0.49
Czech	0.47
DEL	0.40
NCAA	0.33
VHL	0.31
Allsvenskan	0.30
USHL	0.25
OHL	0.24
WHL	0.22
Czech2	0.21
QMJHL	0.19
Mestis	0.19
MHL	0.15
Czech U20	0.14
SuperElit	0.11
Jr. A SM-liiga	0.10

Figure 7

With our equivalency factors now refined, 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 regarding the NHL draft and draft eligible prospects, 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, allowing us to account for some of the variance that exists in skill level between leagues, there is still the remaining challenge of further normalizing production for the variance that exists among scoring rates within each league. The motivation in doing so is that even within the same league, scoring rates can vary greatly by era. 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 on average 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 across era by first computing the all-time average points per game rate for each league in our sample, and then comparing the yearly scoring rates (grouped by league and season) against the all-time average for each league, in order to plot historical trends in scoring rates within each league. As an example, using this methodology we can compute that scoring in the QMJHL 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 QMJHL season as an example, his era adjusted point total would thus be equal to 186 * 0.89 = 166.

Moving forward, this methodology may require some refining itself, as depending on the normality of the distribution in league scoring rates over time, it is possible that the median may lend itself better to these types of adjustments. Though, it should be noted that after some initial testing using both the historical mean and median scoring rates, there did not appear to be any discernable difference in results.

K-Means Clustering

With our pre-draft production rates both league and era normalized, next we move on to our goal of constructing a replica of the PCS model. In order to do so, an unsupervised k-means clustering approach was selected as a means by which we can partition our sample into cohorts of comparable players, where similarity is based on 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 distance between points as a measure of 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 in an attempt to place additional emphasis to the more recent of the observations, the age 18 production inputs will be weighted twice as heavily as the age 17 production inputs,.

It should be noted that this methodology differs slightly from that of Weissbock and Lawrence, who used only a player’s age 18 production rates based on points per game, along with additional inputs such as height and age. While Weissbock and Lawrence found that height and age acted as significant predictors of future NHL success based on their resulting p-values, age 18 points per game rates remained far and away the primary predictor of the three, with age and height adding sharing only a moderate correlation to future NHL points per game rates among their test set.[5] In order to avoid inadvertently over-weighting age and height as inputs in our model, they were decidedly omitted. Based on the strong correlation observed between points per game rates at the junior level and points per game rates observed at the NHL level, we have good reason to believe we will observe similar results to that of the PCS model, despite the slightly different approach taken.

Prior to running the k-means algorithm, it is important to first evaluate whether our data lends itself to clustering, along with estimating an optimal number of clusters to be used based on the underlying composition of the dataset. By mapping our age 17 and age 18 normalized production rates to the 2D plane (Figure 8), it appears as though we cannot generate any meaningful insights at a quick glance, as no apparent clusters are visible.

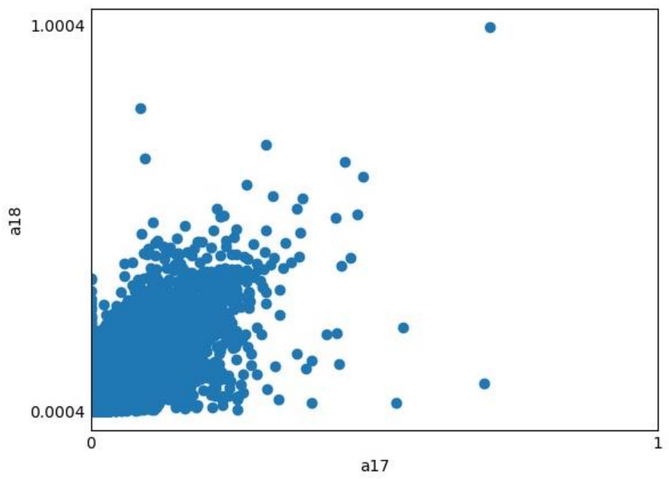


Figure 8

[5] (Hohl, 2015)

Thankfully there are heuristics that can be used to estimate the optimal number of clusters, instead of relying solely 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.

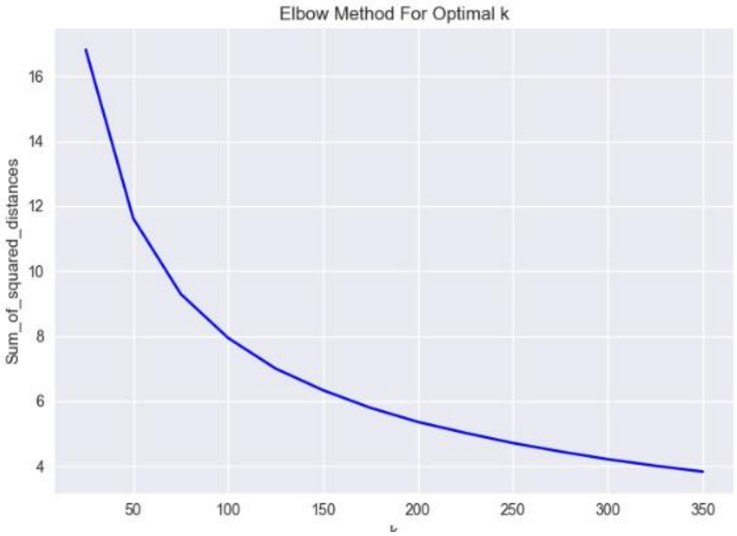


Figure 9

Unfortunately in our case, the results yield no visible elbow, but rather a relatively smooth curve, which again does not bode well for a clustering approach. However, this does not necessarily mean that our dataset does not lend itself to clustering at all. Instead we can experiment with various levels of clustering to see if any such level of partitioning can yield improved results when training our models, treating the number of clusters as a hyperparameter of sorts.

Model Design and Data Pre-Processing

With our k-means clustering model ready, and our pre-draft production rates fully normalized for league and era variance, 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 our models in order to analyze their performance. For our purposes, six popular classification models will be trained and tested, 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 (all from Python's sklearn library). Once we have tested and identified the top performing model(s) from among this set, we may choose to further refine our approach by testing various ensembles, or by tuning certain hyperparameters to see if we can possibly further improve on our model's performance. 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 features chosen to be used as inputs for our classifiers include:

- Player height
- Player weight
- Age (as of September 15th of the draft year in question)
- Country of origin
- League of origin (age 17 and 18)
- Goals and assists per game (age 17 and 18)

- Binary identifiers indicating whether or not a prospect played in the World Under-18 or World Under-20 Championships at any point prior to their draft year
- The cluster values produced by our k-means model

We will train our models first using unadjusted goal and assist rates, followed by our normalized goal and assist totals, which will hopefully allow us to assess if any additional predictive capability is being added by applying league and era standardization to scoring totals. Next, we will introduce our k-means cluster values as an additional model feature, using separate runs of 50, 100, and 200 clusters in an attempt to determine if incorporating the results into our models at any level of partitioning adds any gains in performance to the base model.

When training each model, we will apply feature selection where necessary in order to crop out any features which may not act as meaningful predictors, and would only serve to add noise if included. At this point, some data pre-processing of our inputs is required as well. Any ordinal inputs, such as goal and assist rates, height, weight, and age must first be standardized in order to normalize their distributions and ensure all ordinal data is scaled in the range 0 to 1. Next, we must apply one hot encoding to any categorical inputs (country of origin, league, cluster values, etc.) that don’t lend themselves to standardization. If this method is not applied to our categorical inputs, we risk having the machine learning algorithms treat the categorical data as it would ordinal data. One hot encoding avoids this error by transforming categorical data to a numeric representation where necessary (most machine learning algorithms do not possess the capacity to process text inputs), and then by mapping the data to multiple fields (one for each category), where membership to any individual category is indicated using binary classification.

Using our previously defined features, each model will be trained to predict the following binary classifiers:

1) The probability a given prospect will play at least one NHL game.
2) The probability a given prospect will play more than two full NHL seasons (164 GP) *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 points per game thresholds from our second classifier were heuristically selected as a means of classifying "top 9" production among forwards and "top 4" production among defensemen, based on last season's distribution of scoring by position at the NHL level. I.e., being that the league is composed of 31 teams, we would expect there to be 31 * 9 = 279 top 9 forwards, and 31 * 4 = 124 top 4 defensemen in any given year. Based on the points per game rates of all forwards and defensemen who dressed in at least 20 games in the 2018-19 NHL season, the 279th forward and 124th ranked defensemen scored at approximately a rate of 0.33 and 0.25 points per game, respectively, providing us with our thresholds for defining "top 9" and "top 4" production respectively.

Both classifiers are to be trained and tested on a sample from our database composed of 12,149 forwards and 6,109 defensemen, all 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 just a single game at the NHL level, and of those only 354 (under 3%) went on to become top 9 forwards based on the criteria specified previously. Among our sample of 6,109 defensemen aged 25 and older, only 586 (again, just shy of 10%) went on to play only a single

NHL game, and of those only 89 (under 1.5%) went on to produce at the rate expected of a top 4 defenseman.

Based on these figures, we can clearly discern that an imbalance exists between classes in our sample sets. Classic examples of some more common classification problems where imbalance is prevalent include credit card fraud detection, and cancer diagnosis. When sample sets exhibit this type of imbalance, we can no longer rely on a standard accuracy score as a primary measure of evaluation (defined as $Accuracy = \frac{TN+TP}{all\ observations}$, where TN represents true negative predictions, and TP represents true positives). Using our case as an example, where fewer than 3% of our observed sample of forwards go on to 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 score of over 97%, but we would not be left with any meaningful insights. Instead, when dealing with such extreme imbalance we must look to alternative metrics of evaluation, namely precision and recall. Precision refers to the ratio of positive predictions that represent observed positives, and is defined as $\frac{TP}{TP+FP}$, whereas recall refers to the ratio of observed positives that a model classifies as true positives, defined as $\frac{TP}{TP+FN}$. The F1 score provides a harmonic mean of these two metrics, and is computed as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + \ Recall}.$$[6]

As a result of the underlying imbalance observed in our sample set, we will rely largely on recall and F1 scores when evaluating the performance of each of our models.

[6] (Saxena, 2018)

Along with using alternative metrics for evaluating model performance on imbalanced data sets, there exist certain unconventional sampling methods that can help improve performance in the face of imbalance. The first of these approaches is to oversample the underrepresented class in order to gain an equal representation of each class in the train and test sets, where existing instances are either duplicated, or used as a basis in order to artificially generate new instances. Undersampling can also be employed in an attempt to counteract imbalance. With undersampling, only a subset of the overrepresented class is sampled in an attempt to introduce balance into the class structure. We will test both of these methods when building our models to see if either might be able to offer any additional performance gains above and beyond that of our base models. To implement these methods, we will make use of the Python imblearn library, using the NearMiss approach as a method of undersampling, and the SMOTE approach as a method of oversampling. The SMOTE method achieves oversampling by way of interpolation (i.e. by artificially generating new instances of the underrepresented class, rather than simply duplicating existing instances).

Before training and testing our models, the final step required is to split our samples into separate train and test sets. In this case, we opted for a standard test size of 30% of the overall sample set. We also opted to split our train/test sets in a stratified fashion in order to ensure both the train and test sets are equally representative of each of our classes.

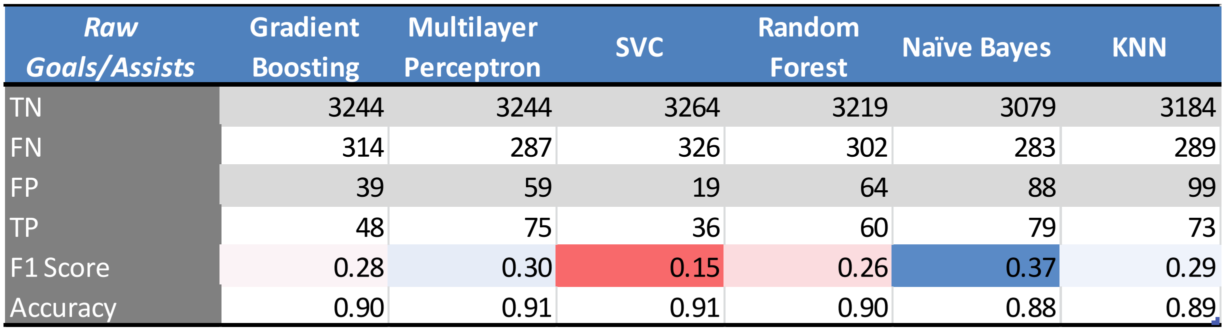
Training, Testing, and Model Selection

Predicting Whether a Player Will Make the NHL

First we want to train our classifiers to predict the probability that a draft eligible prospect will go on to play at least a single game at the NHL level.

Forwards

As a first step, we train and test each model using our raw unadjusted scoring rates, yielding the results shown in Figure 10. These results will serve as a baseline for the performance of our models, which we can reference as we refine our classifiers at each step.

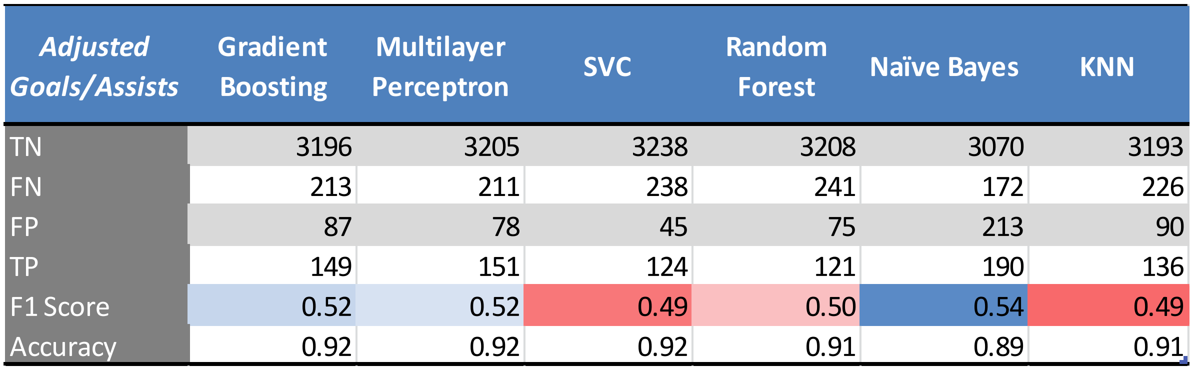


Raw Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3244	3244	3264	3219	3079	3184
FN	314	287	326	302	283	289
FP	39	59	19	64	88	99
TP	48	75	36	60	79	73
F1 Score	0.28	0.30	0.15	0.26	0.37	0.29
Accuracy	0.90	0.91	0.91	0.90	0.88	0.89

Figure 10

As can be seen in Figure 10, the Naïve Bayes classifier is the early top performer, correctly identifying only 79 of a total 362 players from our test set who went on to play in the NHL in any capacity. However, none of the models demonstrate much in the way of predictive capability, with the highest F1 score coming in at 37%.

Re-training our models using our adjusted scoring figures yields significantly better results, which are displayed in Figure 11.

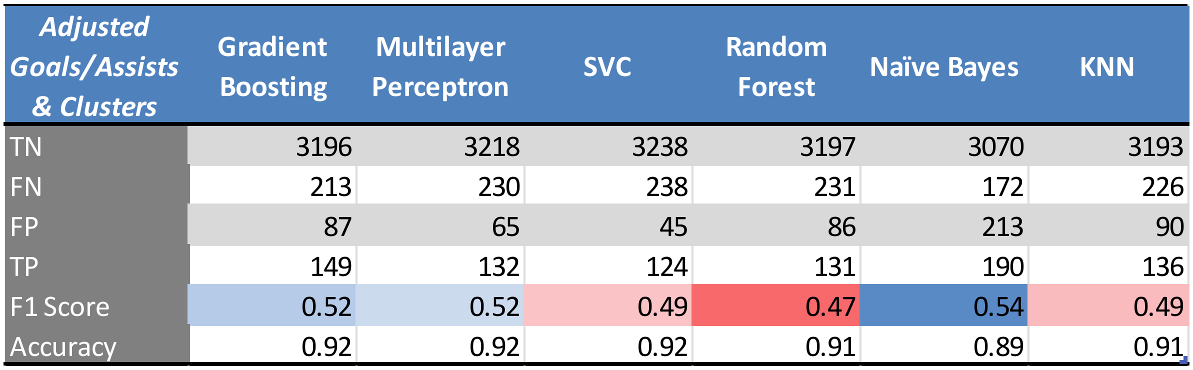


Adjusted Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3196	3205	3238	3208	3070	3193
FN	213	211	238	241	172	226
FP	87	78	45	75	213	90
TP	149	151	124	121	190	136
F1 Score	0.52	0.52	0.49	0.50	0.54	0.49
Accuracy	0.92	0.92	0.92	0.91	0.89	0.91

Figure 11

Again we observe that the Naïve Bayes classifier performs best, but this time all of our classifiers exhibit similar performance across the board. Comparing the results of the Naïve Bayes classifier to those observed using our unadjusted scoring figures, we see that it is now able to correctly predict 190 of the 362 observed positives, though the number of false positives has increased as well. Despite this, we can consider this a marked improvement over our initial results, and conclude that our league and era adjustment do seem to greatly improve our model's performance.

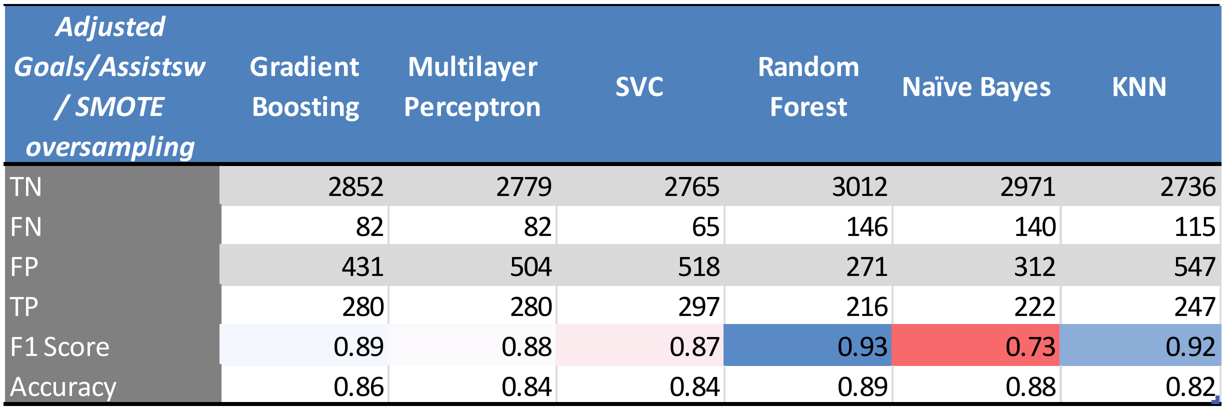
Next, we retrain and test our models with the cluster groupings added in as additional inputs. In trying out various levels of clustering (n=50, n=100, n=200), we observe that adding in the cluster values does not seem to improve performance from the previous step in any significant manner, and in certain cases may just add noise. A snapshot of these results is shown in Figure 12 (n=100 clusters).



Adjusted Goals/Assists & Clusters	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3196	3218	3238	3197	3070	3193
FN	213	230	238	231	172	226
FP	87	65	45	86	213	90
TP	149	132	124	131	190	136
F1 Score	0.52	0.52	0.49	0.47	0.54	0.49
Accuracy	0.92	0.92	0.92	0.91	0.89	0.91

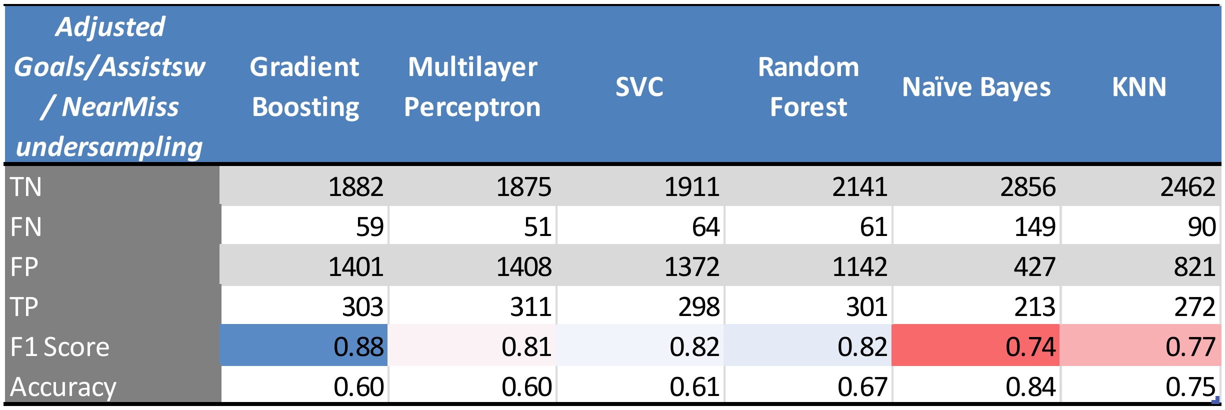
Figure 12

Finally, we want to explore whether or not it is possible to improve our model performance even further by applying both our undersampling and oversampling methods in order to counteract the effects of imbalance in our sample. Figure 13 shows the results after applying the SMOTE oversampling method, while Figure 14 shows the results of applying the NearMiss undersampling method.



Adjusted Goals/Assistsw / SMOTE oversampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	2852	2779	2765	3012	2971	2736
FN	82	82	65	146	140	115
FP	431	504	518	271	312	547
TP	280	280	297	216	222	247
F1 Score	0.89	0.88	0.87	0.93	0.73	0.92
Accuracy	0.86	0.84	0.84	0.89	0.88	0.82

Figure 13



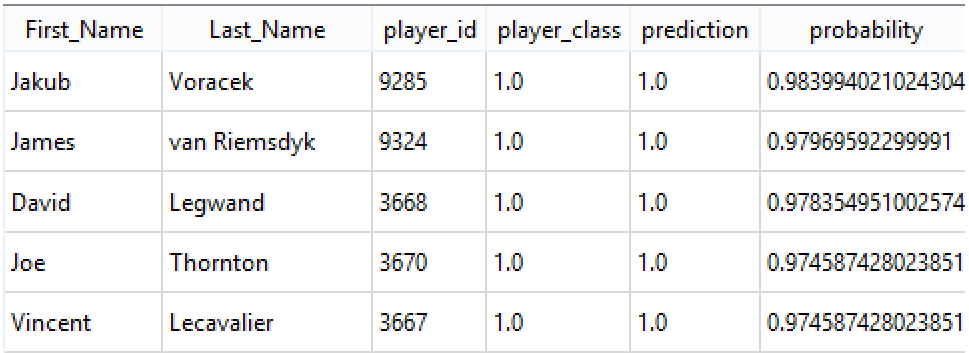
Adjusted Goals/Assistsw / NearMiss undersampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1882	1875	1911	2141	2856	2462
FN	59	51	64	61	149	90
FP	1401	1408	1372	1142	427	821
TP	303	311	298	301	213	272
F1 Score	0.88	0.81	0.82	0.82	0.74	0.77
Accuracy	0.60	0.60	0.61	0.67	0.84	0.75

Figure 14

At first glance the results seem promising, with both methods outperforming our base model. As we might expect, undersampling manages to capture a slightly larger portion of true positives, but this gain comes at the cost of a significantly larger portion of false positives, leading to lower accuracy across the board, despite exhibiting similar F1 scores to the oversampling method.

At this stage a small number of ensemble methods from among the best performers (Gradient Boosting, Multilayer Perceptron, Random Forest, KNN) were also tested to see if we might be able improve our model’s performance further, but none seemed to offer significant gains. In the end we elect for the league/era adjusted Gradient Booster with SMOTE oversampling applied. While both our KNN and Random Forest classifiers did grade out better in terms of F1 score, we elect for the Gradient Boosting model due to its higher level of recall. The chosen model registered a final F1 score of 89%, recall score of 77%, and accuracy score of 86%. In the end our model was able to correctly identify 280 of the 362 forwards in the test set who went on to play a single NHL game.

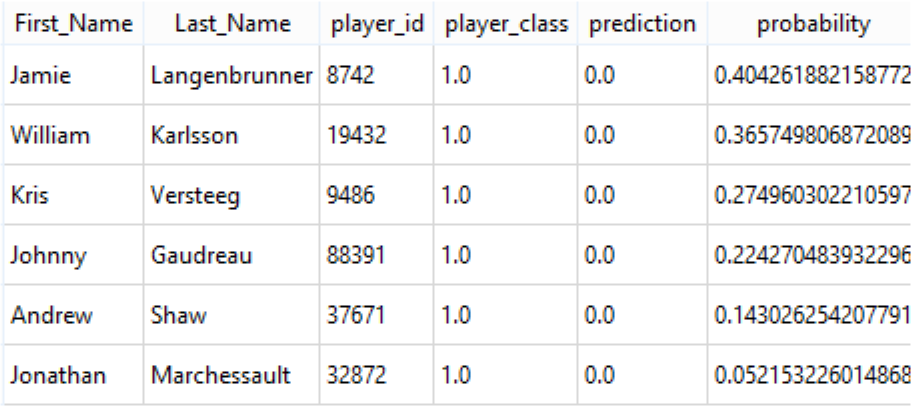
Using our model's results, the top 5 ranked players from our test set (based on the predicted probability that they will go on to play a single game at the NHL level) are listed in Figure 15. We note that this list is composed of three former 1st overall draft choices (Lecavalier, Thornton, Legwand), along with a 2nd (Van Riemsdyk), and 7th overall selection (Voracek). Here, "probability" represents the model's confidence level that the player in question will go on to play in the NHL.



First_Name	Last_Name	player_id	player_class	prediction	probability
Jakub	Voracek	9285	1.0	1.0	0.983994021024304
James	van Riemsdyk	9324	1.0	1.0	0.97969592299991
David	Legwand	3668	1.0	1.0	0.978354951002574
Joe	Thornton	3670	1.0	1.0	0.974587428023851
Vincent	Lecavalier	3667	1.0	1.0	0.974587428023851

Figure 15

Of course, the model is not without its misses, though at a glance they appear to be few and far between. Figure 16 highlights some of the most notable false negatives. What is interesting is that 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, indicating that this group is largely made up of late bloomers whose development trajectory may have been unconventional in the years following their first draft eligible seasons.

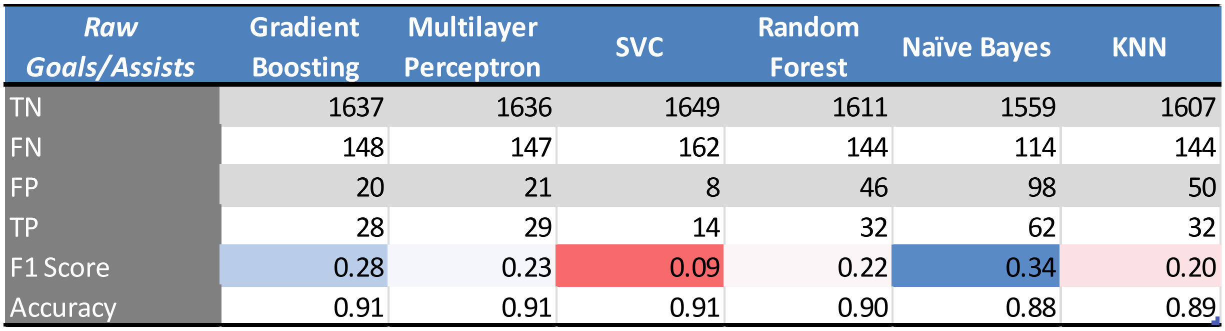


First_Name	Last_Name	player_id	player_class	prediction	probability
Jamie	Langenbrunner	8742	1.0	0.0	0.404261882158772
William	Karlsson	19432	1.0	0.0	0.365749806872089
Kris	Versteeg	9486	1.0	0.0	0.274960302210597
Johnny	Gaudreau	88391	1.0	0.0	0.224270483932296
Andrew	Shaw	37671	1.0	0.0	0.143026254207791
Jonathan	Marchessault	32872	1.0	0.0	0.052153226014868

Figure 16

Defensemen

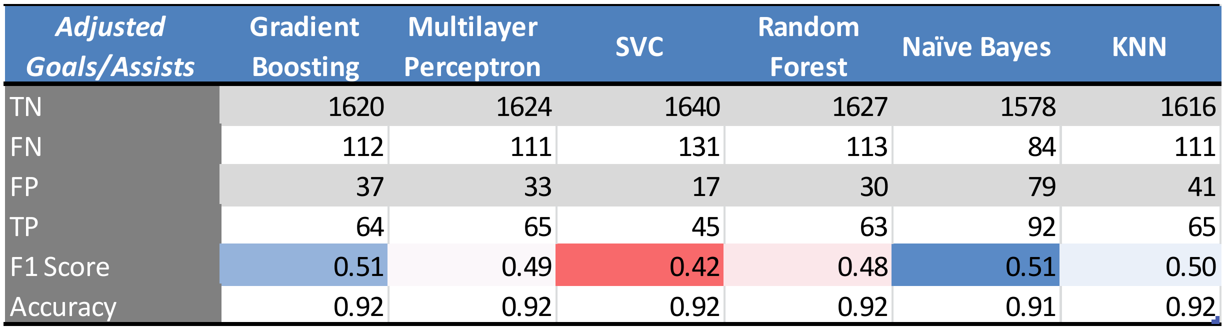
Next we apply the same methodology to predicting whether any given prospect at the position of defense will go on to play in at least one NHL game. Figure 17 shows the initial baseline results for our model using unadjusted scoring rates.



Raw Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1637	1636	1649	1611	1559	1607
FN	148	147	162	144	114	144
FP	20	21	8	46	98	50
TP	28	29	14	32	62	32
F1 Score	0.28	0.23	0.09	0.22	0.34	0.20
Accuracy	0.91	0.91	0.91	0.90	0.88	0.89

Figure 17

The initial results are similar to those observed among forwards, with the Naïve Bayes classifier grading out as the top performer, registering an F1 score of 34%. Next, we apply league and era normalization to see if we observe any increase in model performance as a result.

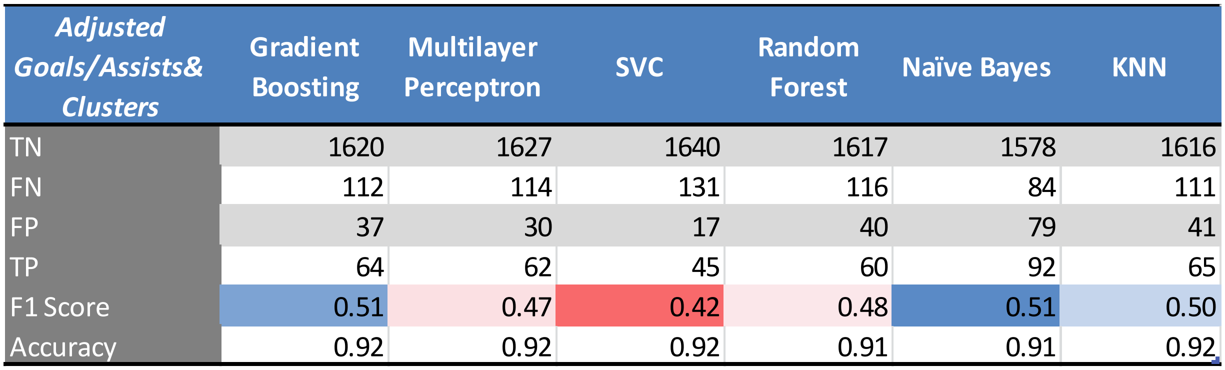


Adjusted Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1620	1624	1640	1627	1578	1616
FN	112	111	131	113	84	111
FP	37	33	17	30	79	41
TP	64	65	45	63	92	65
F1 Score	0.51	0.49	0.42	0.48	0.51	0.50
Accuracy	0.92	0.92	0.92	0.92	0.91	0.92

Figure 18

Figure 18 shows that again by applying our league and era adjustments we greatly improve the collective performance of our models. Whereas the best base model (Naïve Bayes) correctly identified 62 of a possible 176 observed positives, after normalizing scoring rates that total jumps to 92 of a possible 176.

As a next step, we re-train and test our models including our cluster values for cluster levels n=50, n=100, and n=200, in order to assess if including this additional feature might offer any gains in model performance. Figure 19 displays these results in detail for n=100 clusters.

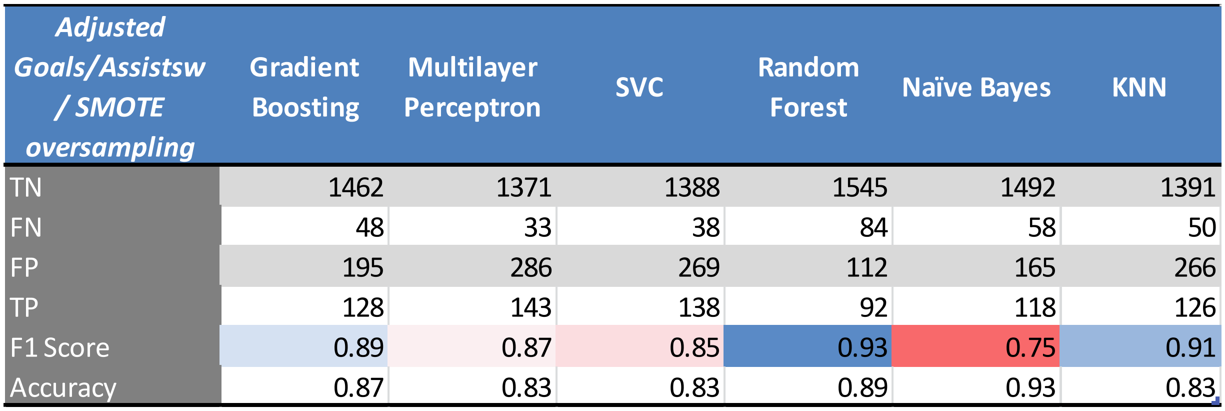


Adjusted Goals/Assists& Clusters	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1620	1627	1640	1617	1578	1616
FN	112	114	131	116	84	111
FP	37	30	17	40	79	41
TP	64	62	45	60	92	65
F1 Score	0.51	0.47	0.42	0.48	0.51	0.50
Accuracy	0.92	0.92	0.92	0.91	0.91	0.92

Figure 19

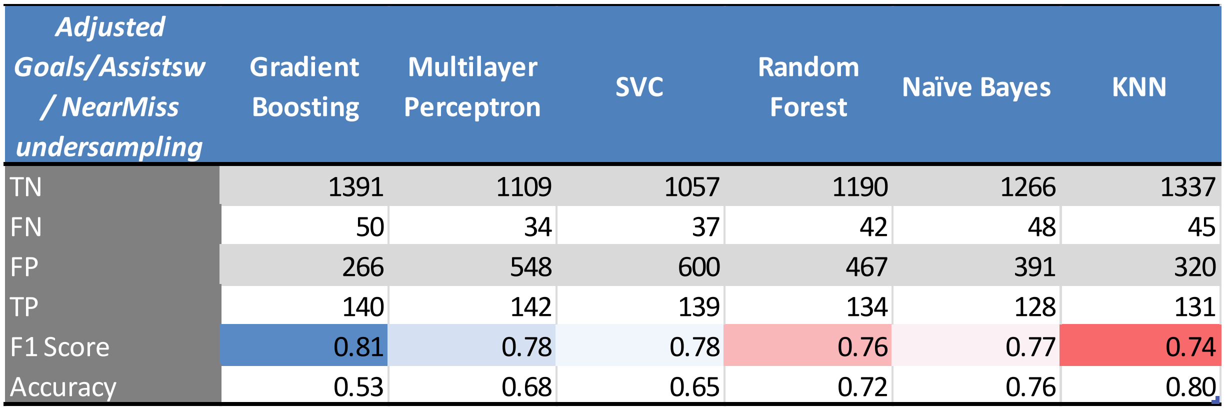
Similarly to what was observed with forwards, including our cluster values as an additional parameter does not seem to generate any gains in terms of predicting which defensemen will go on to play at the NHL level in any capacity.

Finally, we move on to applying our oversampling and undersampling methods. These results are shown in Figure 20 (oversampling), and Figure 21 (underampling).



Adjusted Goals/Assistsw / SMOTE oversampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1462	1371	1388	1545	1492	1391
FN	48	33	38	84	58	50
FP	195	286	269	112	165	266
TP	128	143	138	92	118	126
F1 Score	0.89	0.87	0.85	0.93	0.75	0.91
Accuracy	0.87	0.83	0.83	0.89	0.93	0.83

Figure 20

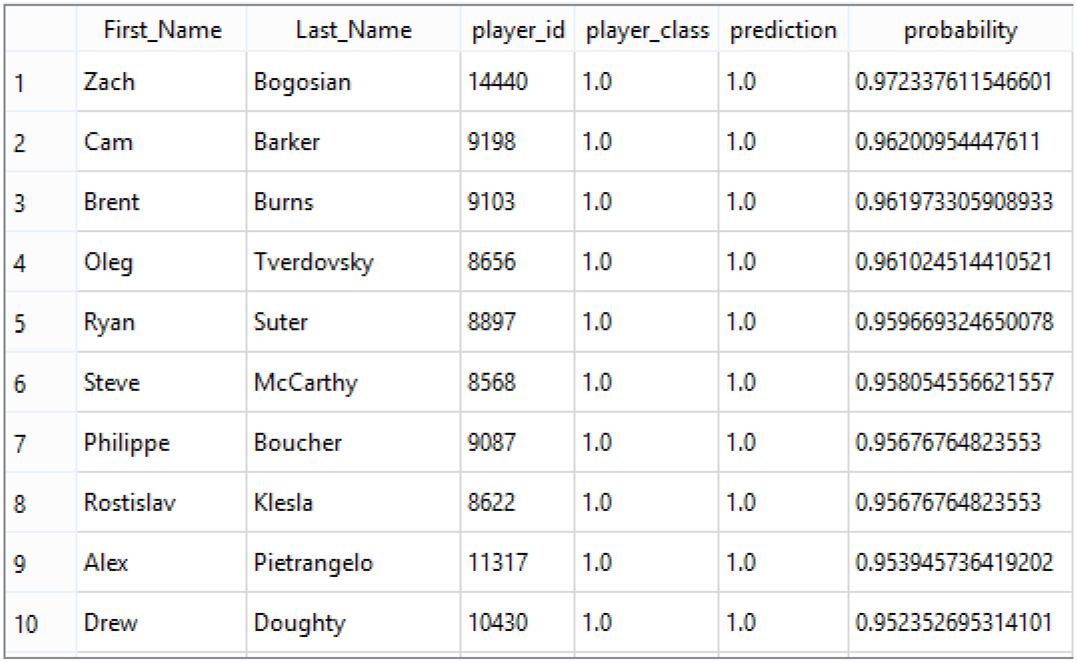


Adjusted Goals/Assistsw / NearMiss undersampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1391	1109	1057	1190	1266	1337
FN	50	34	37	42	48	45
FP	266	548	600	467	391	320
TP	140	142	139	134	128	131
F1 Score	0.81	0.78	0.78	0.76	0.77	0.74
Accuracy	0.53	0.68	0.65	0.72	0.76	0.80

Figure 21

These results are again very similar to those observed with our sample group of forwards. The oversampling method appears to outperform the undersampling method, both based on F1 score, and accuracy. The Random Forest and KNN models achieve the highest F1 scores of the bunch, however we yet again observe that the Gradient Boosting classifier exhibits a higher recall score, and lower ratio of false positives. Based on this we conclude that the Gradient Booster seems to offer the best tradeoff of all the tested models. The chosen Gradient Boosting classifier with adjusted scoring rates and SMOTE oversampling, registers an F1 score of 89%, a recall score of 73% and an accuracy score of 87%.

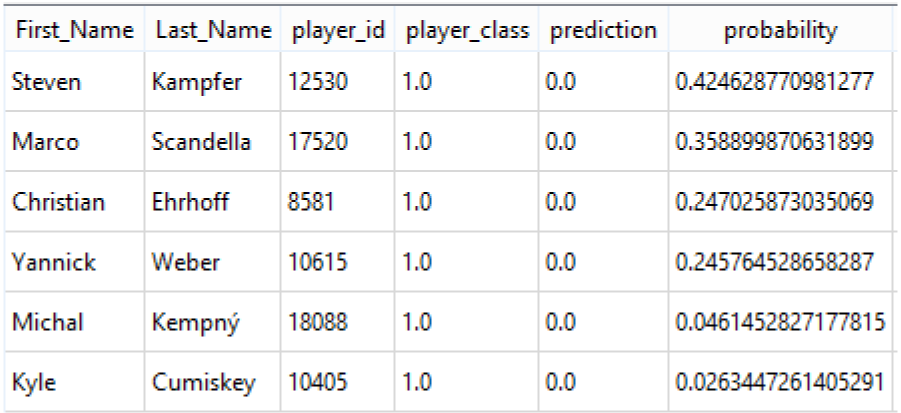
Taking a quick glance at the model's output from our test set (Figure 22), the results appear to be promising. The top 10 defensemen from our test set, which the model predicted would go on to play in the NHL in any capacity, all managed to play at least a single NHL game, and most went on to enjoy successful NHL careers to varying degrees.



	First_Name	Last_Name	player_id	player_class	prediction	probability
1	Zach	Bogosian	14440	1.0	1.0	0.972337611546601
2	Cam	Barker	9198	1.0	1.0	0.96200954447611
3	Brent	Burns	9103	1.0	1.0	0.961973305908933
4	Oleg	Tverdovsky	8656	1.0	1.0	0.961024514410521
5	Ryan	Suter	8897	1.0	1.0	0.959669324650078
6	Steve	McCarthy	8568	1.0	1.0	0.958054556621557
7	Philippe	Boucher	9087	1.0	1.0	0.95676764823553
8	Rostislav	Klesla	8622	1.0	1.0	0.95676764823553
9	Alex	Pietrangelo	11317	1.0	1.0	0.953945736419202
10	Drew	Doughty	10430	1.0	1.0	0.952352695314101

Figure 22

Again, we'd be remiss if we didn't also cover some of the model's notable false negatives (Figure 23). The good news is that, similar to what was observed with forwards, these misses appear to be few and far between, and in the case of defensemen, include a list of players who mostly went on to carve out fringe careers at the NHL level.



First_Name	Last_Name	player_id	player_class	prediction	probability
Steven	Kampfer	12530	1.0	0.0	0.424628770981277
Marco	Scandella	17520	1.0	0.0	0.358899870631899
Christian	Ehrhoff	8581	1.0	0.0	0.247025873035069
Yannick	Weber	10615	1.0	0.0	0.245764528658287
Michal	Kempný	18088	1.0	0.0	0.0461452827177815
Kyle	Cumiskey	10405	1.0	0.0	0.0263447261405291

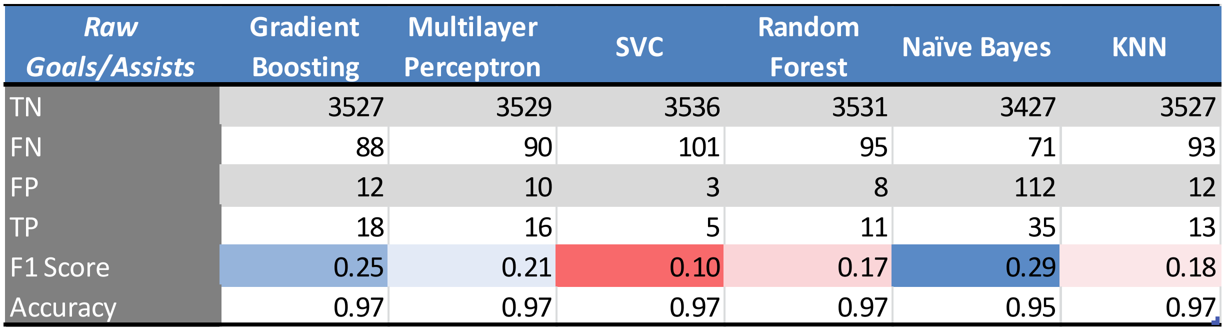
Figure 23

Identifying Top 9 Forwards and Top 4 Defensemen

In this section we refine the logic of our classifier to make things ever more restrictive. Instead of predicting the probability a prospect simply goes on to play a single game at the NHL level, we now wish to compute the probability they go on to play more than two seasons (> 164 games), and achieve a career points per game rate of at least 0.33 in the case of forwards (signifying what we had previously defined as “top 9” production), or 0.25 in the case of defenseman (signifying “top 4” production).

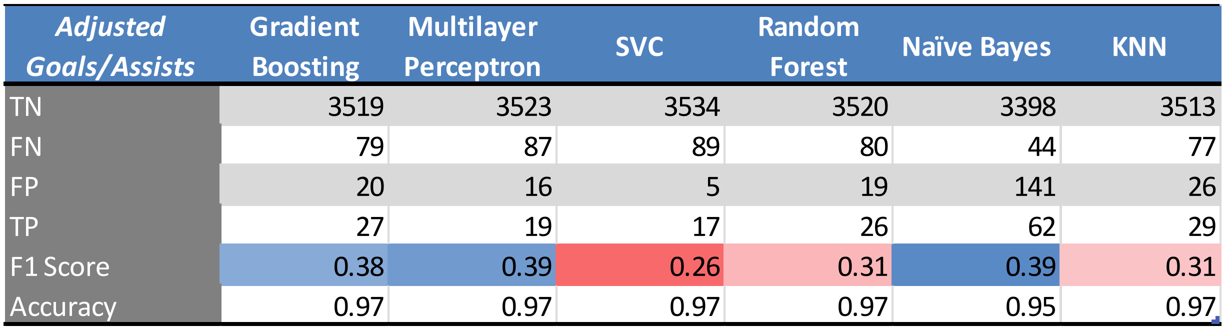
Forwards

After generating new train/test sets for our second classifier (a step made necessary by the application of stratified sampling earlier on), we begin by testing our models with both our unadjusted (Figure 24) and adjusted scoring figures (Figure 25).



Raw Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3527	3529	3536	3531	3427	3527
FN	88	90	101	95	71	93
FP	12	10	3	8	112	12
TP	18	16	5	11	35	13
F1 Score	0.25	0.21	0.10	0.17	0.29	0.18
Accuracy	0.97	0.97	0.97	0.97	0.95	0.97

Figure 24

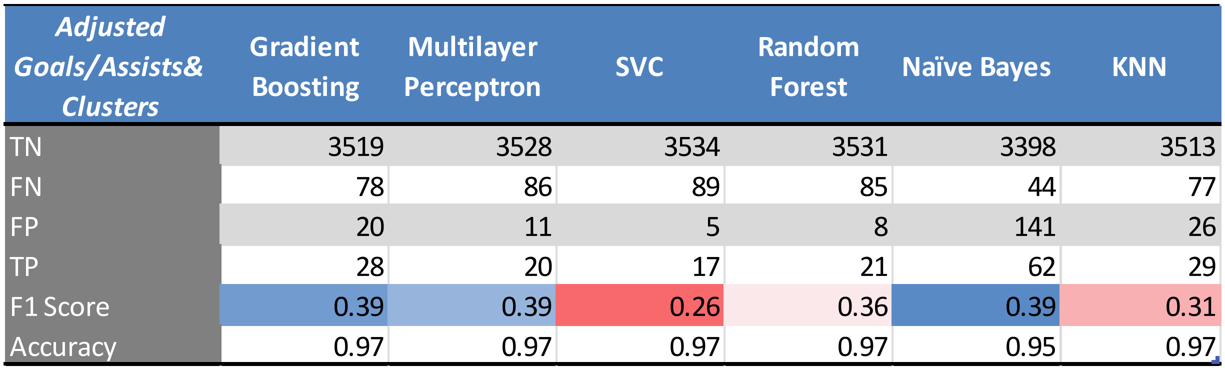


Adjusted Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3519	3523	3534	3520	3398	3513
FN	79	87	89	80	44	77
FP	20	16	5	19	141	26
TP	27	19	17	26	62	29
F1 Score	0.38	0.39	0.26	0.31	0.39	0.31
Accuracy	0.97	0.97	0.97	0.97	0.95	0.97

Figure 25

Again, we observe significant gain in performance when applying the league and era adjustments to scoring figures. However, despite the improvement in performance provided by using our adjusted figures, the rate or false negatives is still significantly high, with the Naïve Bayes classifier being the only model able to successfully identify more than 50% of the forwards from our test set who went on to become top 9 producers at the NHL level.

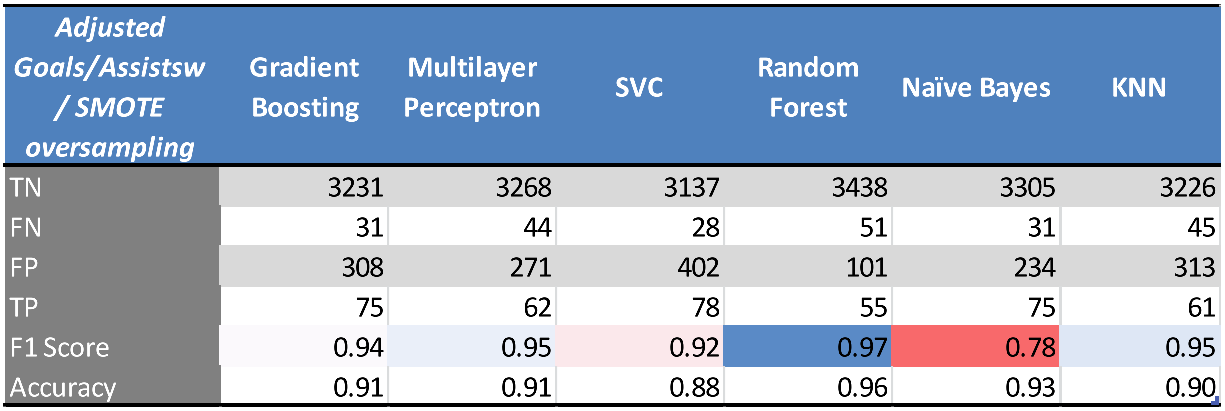
Next we want to assess if including our cluster values will add any additional predictive gains to our new classifier. Figure 26 (displaying the results for n=100 clusters) shows that again we don't appear to receive any added performance gains by introducing the cluster values into our models at any level of clustering (n=50, n=100, n=200).



Adjusted Goals/Assists& Clusters	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3519	3528	3534	3531	3398	3513
FN	78	86	89	85	44	77
FP	20	11	5	8	141	26
TP	28	20	17	21	62	29
F1 Score	0.39	0.39	0.26	0.36	0.39	0.31
Accuracy	0.97	0.97	0.97	0.97	0.95	0.97

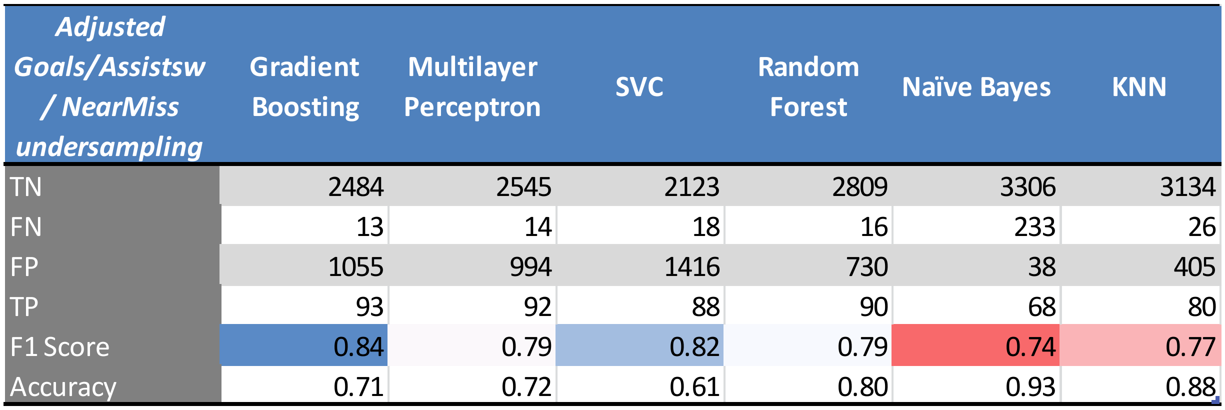
Figure 26

Finally, we employ our oversampling and undersampling methods to see if they might fare any better than our base model. Figures 27 and 28 show that both outperform the base model, with oversampling again appearing to be the most successful predictor, identifying nearly as many top 9 forwards as the undersampling method, while registering far fewer false positives.



Adjusted Goals/Assistsw / SMOTE oversampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	3231	3268	3137	3438	3305	3226
FN	31	44	28	51	31	45
FP	308	271	402	101	234	313
TP	75	62	78	55	75	61
F1 Score	0.94	0.95	0.92	0.97	0.78	0.95
Accuracy	0.91	0.91	0.88	0.96	0.93	0.90

Figure 27

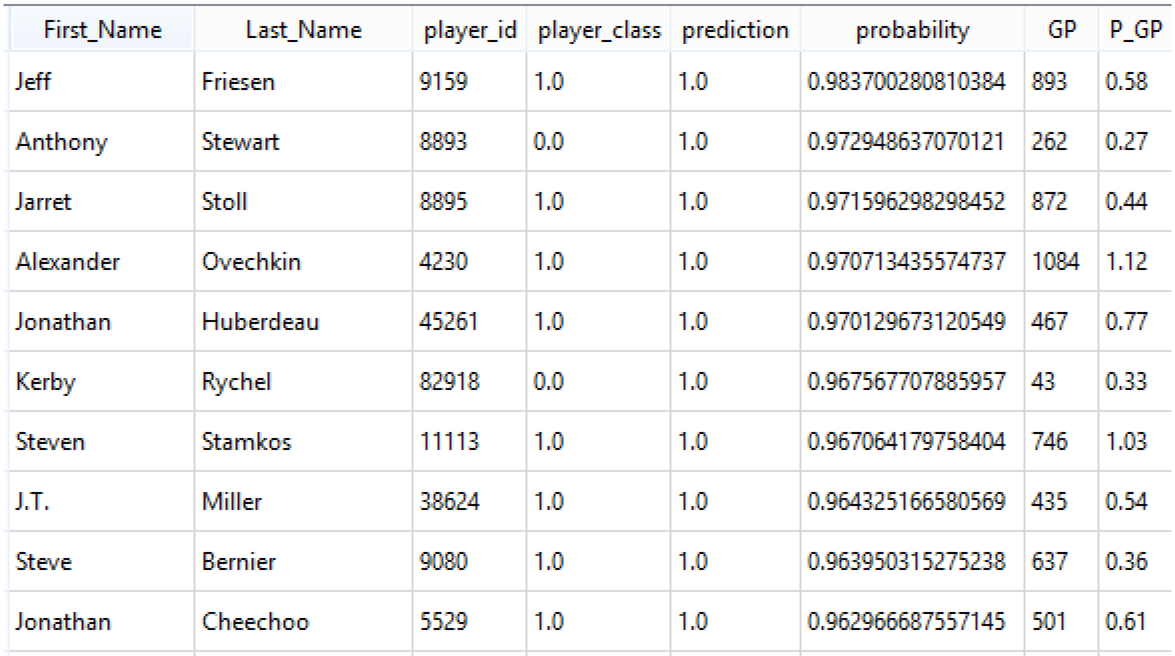


Adjusted Goals/Assistsw / NearMiss undersampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	2484	2545	2123	2809	3306	3134
FN	13	14	18	16	233	26
FP	1055	994	1416	730	38	405
TP	93	92	88	90	68	80
F1 Score	0.84	0.79	0.82	0.79	0.74	0.77
Accuracy	0.71	0.72	0.61	0.80	0.93	0.88

Figure 28

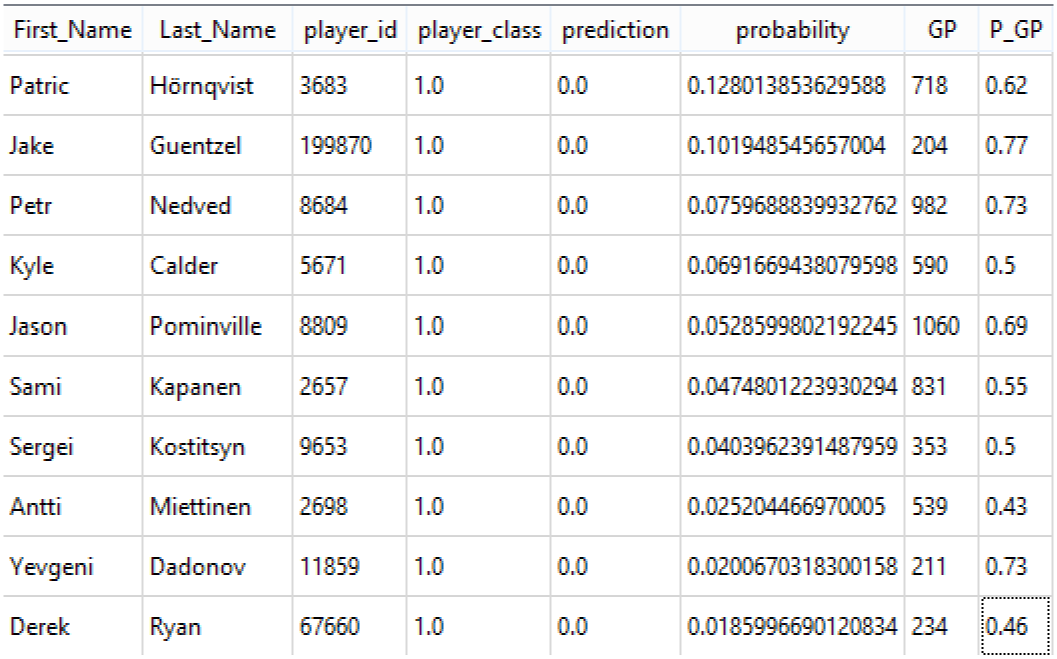
Here we observe that again the Gradient Boosting model seems to offer the best tradeoff between its rate of true and false positives, registering an F1 score of 94%, a recall score of 71%, and an accuracy score of 91%. The model successfully manages to identify 75 of the 106 forwards from our test set who achieved the minimum games played threshold of 165 games, and the minimum career points per game threshold of 0.33. Different ensembles were tested as well, but none offered any significant gains over the performance of the Gradient Boosting model.

A snippet of the top 10 ranked players from our test set, based on the predicted probability that they would go on to become top 9 producers at the NHL level is shown in Figure 29, whereas Figure 30 shows a snippet of the top 10 false negatives. Both lists present a slightly more mixed bag than we observed with our initial classifier, but that is to be expected seeing as our restrictions are more stringent in this case.



First_Name	Last_Name	player_id	player_class	prediction	probability	GP	P_GP
Jeff	Friesen	9159	1.0	1.0	0.983700280810384	893	0.58
Anthony	Stewart	8893	0.0	1.0	0.972948637070121	262	0.27
Jarret	Stoll	8895	1.0	1.0	0.971596298298452	872	0.44
Alexander	Ovechkin	4230	1.0	1.0	0.970713435574737	1084	1.12
Jonathan	Huberdeau	45261	1.0	1.0	0.970129673120549	467	0.77
Kerby	Rychel	82918	0.0	1.0	0.967567707885957	43	0.33
Steven	Stamkos	11113	1.0	1.0	0.967064179758404	746	1.03
J.T.	Miller	38624	1.0	1.0	0.964325166580569	435	0.54
Steve	Bernier	9080	1.0	1.0	0.963950315275238	637	0.36
Jonathan	Cheechoo	5529	1.0	1.0	0.962966687557145	501	0.61

Figure 29



First_Name	Last_Name	player_id	player_class	prediction	probability	GP	P_GP
Patric	Hörnqvist	3683	1.0	0.0	0.128013853629588	718	0.62
Jake	Guentzel	199870	1.0	0.0	0.101948545657004	204	0.77
Petr	Nedved	8684	1.0	0.0	0.0759688839932762	982	0.73
Kyle	Calder	5671	1.0	0.0	0.0691669438079598	590	0.5
Jason	Pominville	8809	1.0	0.0	0.0528599802192245	1060	0.69
Sami	Kapanen	2657	1.0	0.0	0.0474801223930294	831	0.55
Sergei	Kostitsyn	9653	1.0	0.0	0.0403962391487959	353	0.5
Antti	Miettinen	2698	1.0	0.0	0.025204466970005	539	0.43
Yevgeni	Dadonov	11859	1.0	0.0	0.0200670318300158	211	0.73
Derek	Ryan	67660	1.0	0.0	0.0185996690120834	234	0.46

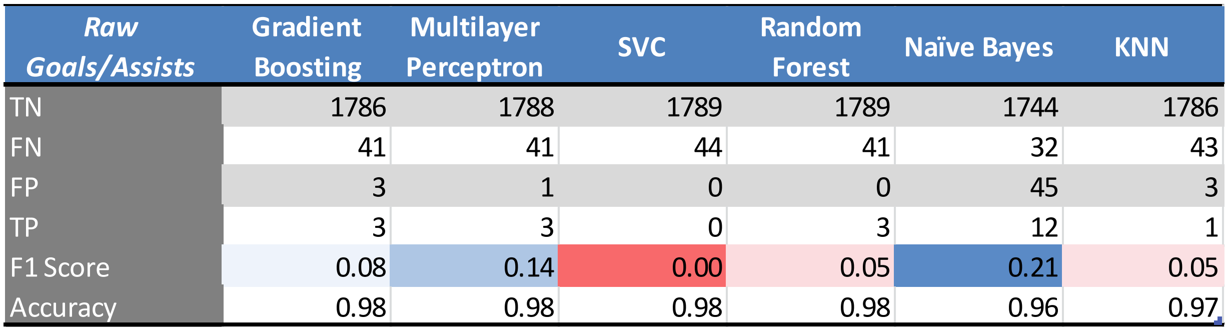
Figure 30

We note that our top 10 most likely prospects to go on to become top 9 forwards at the NHL level includes eight 1st round picks, and two 2nd round selections (Stoll, Cheechoo).

Furthermore, only two of the ten players identified went on to be misclassified. Among the biggest misses depicted in Figure 30, again we note that the list is largely made up of what can be classified as late bloomers, including a large number of players who either have late birthdays, making them ineligible to be drafted following their 18 year old seasons, and/or who would go on to be selected no earlier than the mid-to-late rounds of the draft following their 18 year old seasons. In some cases we even observe players who went undrafted entirely (Dadonov/Ryan).

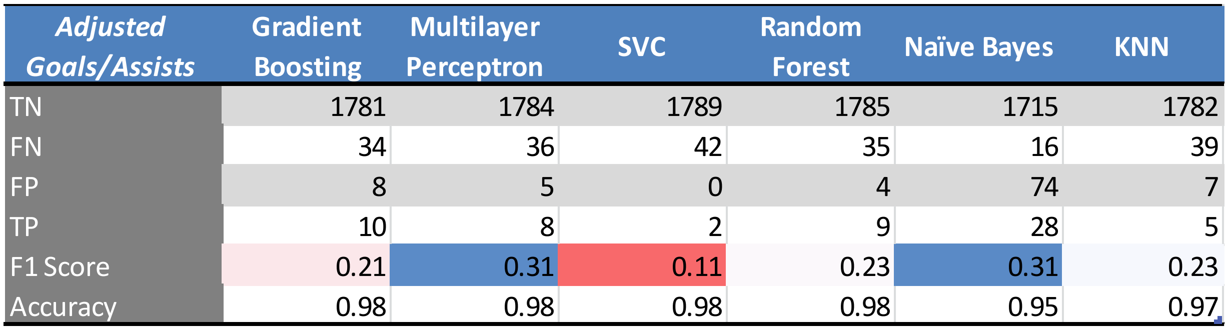
Defensemen

Finally, we move on to modeling the probability that any given defensive prospect will develop into a "top 4" defenseman at the NHL level. Figures 31 and 32 show the results of our base models using unadjusted scoring rates (Figure 31), and our adjusted rates (Figure 32).



Raw Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1786	1788	1789	1789	1744	1786
FN	41	41	44	41	32	43
FP	3	1	0	0	45	3
TP	3	3	0	3	12	1
F1 Score	0.08	0.14	0.00	0.05	0.21	0.05
Accuracy	0.98	0.98	0.98	0.98	0.96	0.97

Figure 31

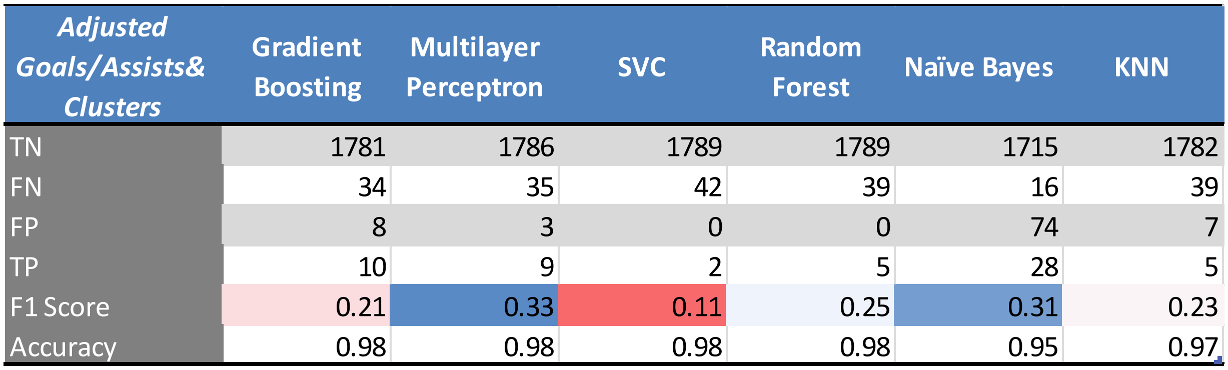


Adjusted Goals/Assists	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1781	1784	1789	1785	1715	1782
FN	34	36	42	35	16	39
FP	8	5	0	4	74	7
TP	10	8	2	9	28	5
F1 Score	0.21	0.31	0.11	0.23	0.31	0.23
Accuracy	0.98	0.98	0.98	0.98	0.95	0.97

Figure 32

The best predictor using our base unadjusted scoring rates (our Naïve Bayes classifier) manages to successfully identify only 12 of the 44 top 4 defensemen contained within our test set. Again, adding in our league and era adjustments appears to offer a great improvement, allowing us to now successfully identify up to 28 of the total 44.

Next we add in our cluster identifiers to see if they might offer any increase in predictive power. Figure 33 shows the results for clustering at level n=100.

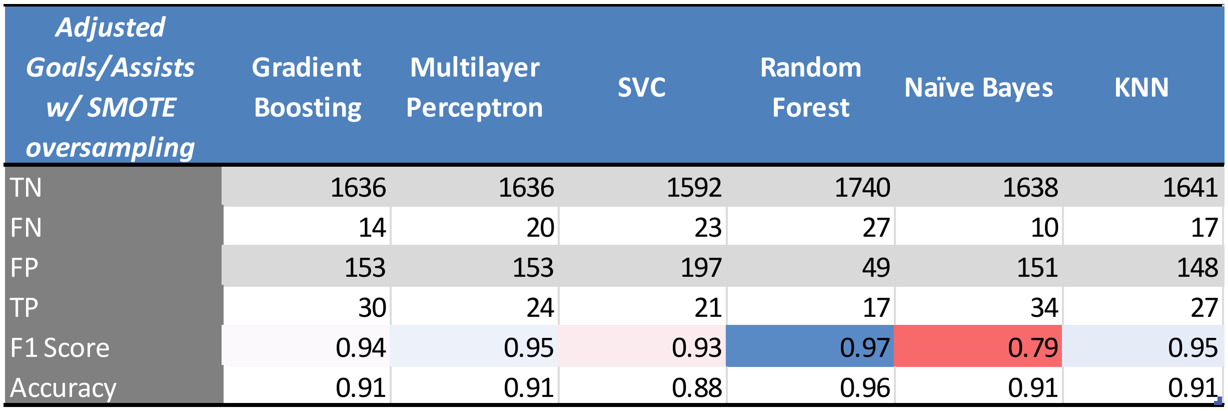


Adjusted Goals/Assists& Clusters	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1781	1786	1789	1789	1715	1782
FN	34	35	42	39	16	39
FP	8	3	0	0	74	7
TP	10	9	2	5	28	5
F1 Score	0.21	0.33	0.11	0.25	0.31	0.23
Accuracy	0.98	0.98	0.98	0.98	0.95	0.97

Figure 33

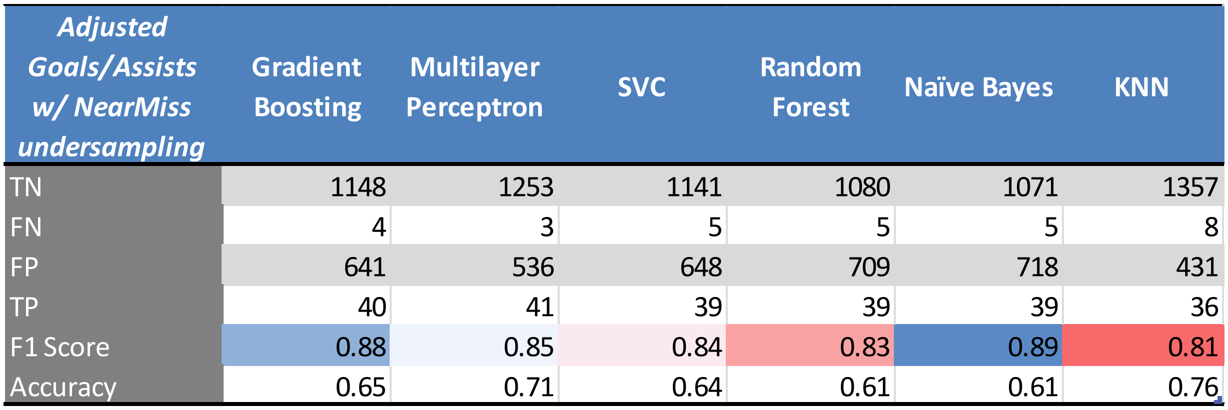
Again we observe no improvement by including our cluster values at any level of partitioning (n=50, n=100, or n=200).

As a final step, we apply oversampling and undersampling to introduce balance into our train/test sets. Figure 34 displays the results of applying oversampling, while Figure 35 displays the results of applying undersampling.



Adjusted Goals/Assists w/ SMOTE oversampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1636	1636	1592	1740	1638	1641
FN	14	20	23	27	10	17
FP	153	153	197	49	151	148
TP	30	24	21	17	34	27
F1 Score	0.94	0.95	0.93	0.97	0.79	0.95
Accuracy	0.91	0.91	0.88	0.96	0.91	0.91

Figure 34

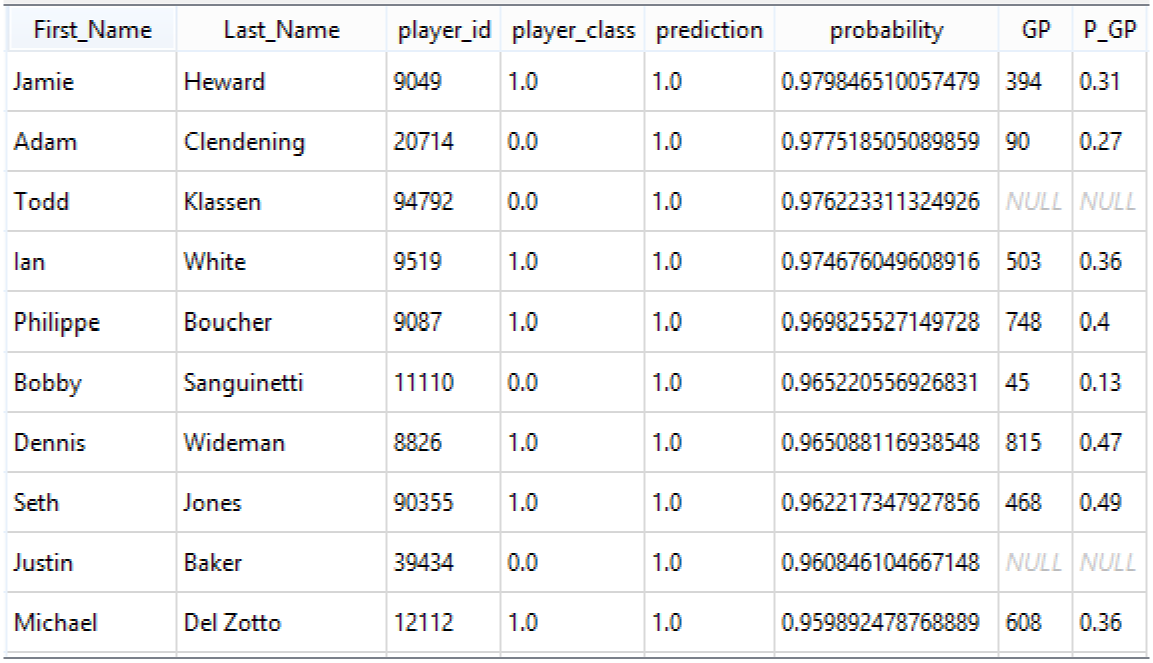


Adjusted Goals/Assists w/ NearMiss undersampling	Gradient Boosting	Multilayer Perceptron	SVC	Random Forest	Naïve Bayes	KNN
TN	1148	1253	1141	1080	1071	1357
FN	4	3	5	5	5	8
FP	641	536	648	709	718	431
TP	40	41	39	39	39	36
F1 Score	0.88	0.85	0.84	0.83	0.89	0.81
Accuracy	0.65	0.71	0.64	0.61	0.61	0.76

Figure 35

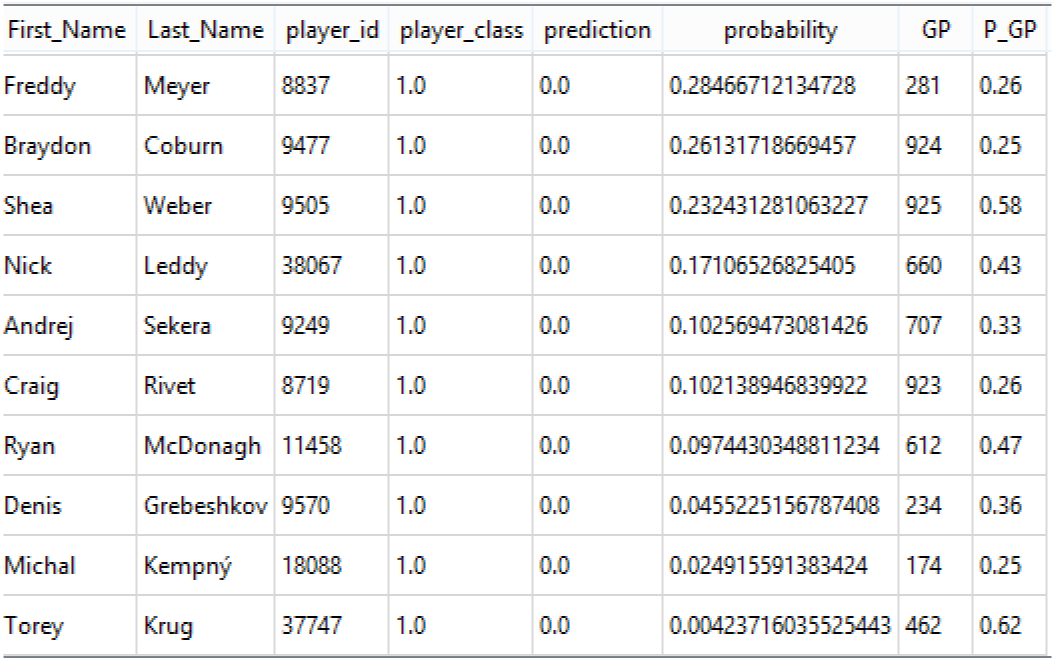
Based on these results, the Gradient Boosting model employing SMOTE oversampling and using our adjusted scoring rates appears to produce the best results once again. This model was able to successfully predict 30 of the 44 top 4 defensemen, registering an F1 score of 94%, a recall score of 68%, and an accuracy score of 91%.

Looking at the top 10 predictions (Figure 36) and top 10 worst misses (Figure 37), as predicted by our model, we again experience more mixed results than we did when simply classifying which defensemen would make the NHL in any capacity. Of the top 10 players predicted to develop into top 4 defensemen, we observe 4 false positives, whereas our list of biggest misses now includes a number of prestigious names. Though, out of these misses we observe a mix of late bloomers, along players who were drafted with high selections, but out of low strength leagues (McDonagh, Leddy), where our model likely discounts their chances based on league strength alone.



First_Name	Last_Name	player_id	player_class	prediction	probability	GP	P_GP
Jamie	Heward	9049	1.0	1.0	0.979846510057479	394	0.31
Adam	Clendening	20714	0.0	1.0	0.977518505089859	90	0.27
Todd	Klassen	94792	0.0	1.0	0.976223311324926	NULL	NULL
Ian	White	9519	1.0	1.0	0.974676049608916	503	0.36
Philippe	Boucher	9087	1.0	1.0	0.969825527149728	748	0.4
Bobby	Sanguinetti	11110	0.0	1.0	0.965220556926831	45	0.13
Dennis	Wideman	8826	1.0	1.0	0.965088116938548	815	0.47
Seth	Jones	90355	1.0	1.0	0.962217347927856	468	0.49
Justin	Baker	39434	0.0	1.0	0.960846104667148	NULL	NULL
Michael	Del Zotto	12112	1.0	1.0	0.959892478768889	608	0.36

Figure 36

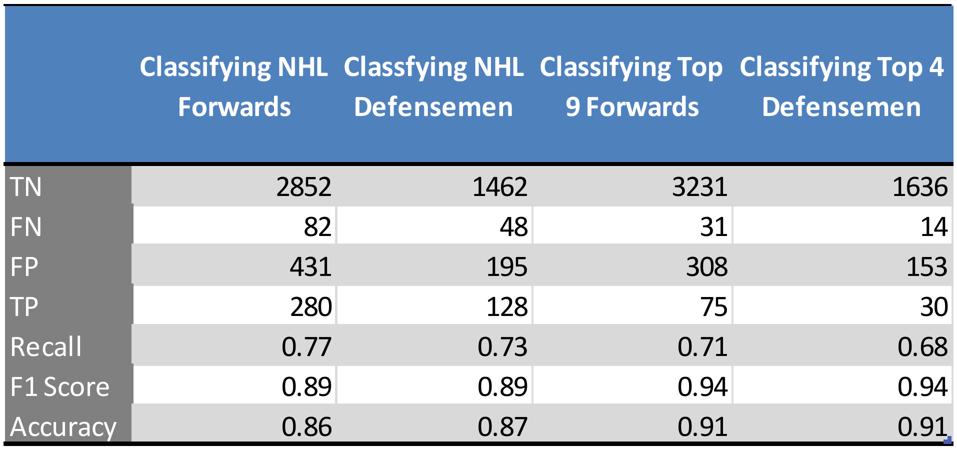


First_Name	Last_Name	player_id	player_class	prediction	probability	GP	P_GP
Freddy	Meyer	8837	1.0	0.0	0.28466712134728	281	0.26
Braydon	Coburn	9477	1.0	0.0	0.26131718669457	924	0.25
Shea	Weber	9505	1.0	0.0	0.232431281063227	925	0.58
Nick	Leddy	38067	1.0	0.0	0.17106526825405	660	0.43
Andrej	Sekera	9249	1.0	0.0	0.102569473081426	707	0.33
Craig	Rivet	8719	1.0	0.0	0.102138946839922	923	0.26
Ryan	McDonagh	11458	1.0	0.0	0.0974430348811234	612	0.47
Denis	Grebeshkov	9570	1.0	0.0	0.0455225156787408	234	0.36
Michal	Kempný	18088	1.0	0.0	0.024915591383424	174	0.25
Torey	Krug	37747	1.0	0.0	0.00423716035525443	462	0.62

Figure 37

Conclusions and Future Work

Figure 38 outlines the full performance breakdown of our chosen Gradient Boosting classifiers, all of which employ SMOTE oversampling and league/era normalized scoring rates. At a glance, it appears that our classifiers' performance is relatively even across the board. As we would expect, predicting which players will go on to play a single game at the NHL level proves to be slightly more reliable than predicting those who will go on to become top 9 forwards or top 4 defensemen, but the drop off in predictivity between the two classifiers is not extreme. Furthermore, it appears that predicting which forwards will go on to experience greater success at the NHL level is slightly easier than predicting which defensemen will go on to become impact players. This likely has to do with the fact that metrics such as goal and assist rates encapsulate more of what makes a forward successful, with defensemen often being evaluated more on traits that might not necessarily translate directly to the scoresheet.



	Classifying NHL Forwards	Classfying NHL Defensemen	Classifying Top 9 Forwards	Classifying Top 4 Defensemen
TN	2852	1462	3231	1636
FN	82	48	31	14
FP	431	195	308	153
TP	280	128	75	30
Recall	0.77	0.73	0.71	0.68
F1 Score	0.89	0.89	0.94	0.94
Accuracy	0.86	0.87	0.91	0.91

Figure 38

In the future we would like to extend this work to possibly look at classifying the likelihood that any given goaltending prospect will go on to enjoy success at the NHL level as well. We would also like to include additional data beyond a player's age 18 season in order to incorporate over-agers into our projections. This would also allow us to adjust expectations for each player for every subsequent season following their first eligible draft year, potentially helping to identify some of the aforementioned late bloomers, or alternatively flagging prospects whose post-draft development may be stagnating. Finally, we would also like to experiment with various additional classifiers, an example of which being classifying the likelihood any given prospect might go on to become an All-Star at the NHL level.

Works Cited

Columbus Blue Jackets. (2019). *Blue Jackets announce several changes to Hockey Operations department*. Retrieved from NHL.com: https://www.nhl.com/bluejackets/news/blue-jackets-hockey-operations-department-changes/c-308249804

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

Hohl, G. (2015). *Draft Theory: Height matters, but maybe due to bias*. Retrieved from Canucks Army: https://canucksarmy.com/2015/02/10/draft-theory-height-matters-but-maybe-due-to-bias/

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/