Applications of Artificial Intelligence to the NHL Entry Draft

by

Brandon Alcox

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2019

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This thesis investigates the application of various fields of artificial intelligence to the domain of sports management and analysis. The research in this thesis is primarily focussed on the entry draft for the National Hockey League, though many of the models proposed may be applied to other sports and leagues with minimal adjustments. A utility model is proposed to define which players are preferred by which teams for a given draft. This model allows for the consideration of how teams acting in a multiagent system may reason about each other's preferences, as well as how they might strategize and interact with one another through trades. A trading scheme where agents may trade picks with each other to change the picking order is established and an algorithm is proposed to find optimal trade offers to propose under an imperfect knowledge setting. Through simulations based on the National Hockey League Entry Draft data, the algorithms provide mutually beneficial trades that also increase the social utility of the league over the course of the draft.

Machine learning classifiers are proposed to suggest which prospects will be successful at the highest level of the sport over various metrics using statistics and scouting reports from their draft year as features. The classifiers out-perform conventional draft selections in the NHL and provide insights into which attributes of a player are important in development. Clustering techniques are used to determine playstyles in the NHL and these clusters are fed as annotations into additional classifiers to project which prospects will fall into certain clusters later in their careers. These latter classifiers demonstrated promising results but were ultimately limited by the availability of data.

A discussion of future avenues of artificial intelligence research in the young but growing field of sports analytics is carried throughout this thesis.

Acknowledgements

I would like to express my appreciation for my thesis supervisor Professor Pascal Poupart for his guidance, patience, and support on this research work. I would also like to thank Professor Philip Curry for his insight into hockey analytics and his help in developing the early ideas that form the basis of this work. In addition, I would like to thank Professor Kate Larsson for her help as I refined the work regarding multiagent systems in this thesis.

I wish to extend gratitude to HockeyTech Inc. for access to the data that made much of this research possible.

Finally, I would like to thank my family for their support throughout my academic career.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This Master's thesis investigates the results of several systems created with varying elements of artificial intelligence for the purpose of solving open questions concerning amateur drafts in sports.

An amateur draft is an opportunity for teams to select young rising stars in the sport to have exclusive rights to sign them to their team and bring them into the league. In general a draft is broken down into rounds where each team gets one selection, where the order within each round is usually related to the reverse standings in the previous season. This allows worse teams to select before better teams and hopefully add better players to their future rosters. As only a small fraction of players selected in any given year will make an impact over the course of their careers and teams on average have seven chances to select each year it is very important that teams efficiently use their picks to find future players that will not only make an impact in general, but also that will work in the systems and play-styles of their specific team. With new ways to study and understand the draft at a deeper level, managers will be able to make decisions that more accurately reflect the best interests of their franchises.

The findings for this thesis will be split into two main chapters which concern two important topics in the draft. The first will look at modeling the draft in terms of which types of players each individual team is likely to select. This topic goes deeper into how strategy can be developed from this information when teams are around the drafting table. The second chapter concerns player projections. Specifically, machine learning techniques will be used to classify prospective players into those which will be successful or not across various metrics, as well as a classification for developing into one of eight common types of National Hockey League (NHL) players.

1.1 Formalizing the Draft

From a mathematical perspective the most basic form of the draft requires a set of available players P, a set of teams T, and a draft ordering $O = \langle o_1, o_2, ..., o_n \rangle$ such that $o_i \in T \; \forall i \in [1..n], n \leq |P|$. The draft is defined recursively such that if o_1 selects player $p \in P$ the draft continues with $\{P \backslash p\}$, T, and $\langle o_2, ..., o_n \rangle$.

This simple drafting framework is also studied as the "Picking Sequences" game in the literature, where agents take turns selecting objects from a finite set and receive an associated utility for each item they select [7]. The connection to sports drafts is made in [13], though the game framework is simplified in that case with ordinal preferences and perfect knowledge settings.

In Chapter 2 this framework will be extended and studied in further depth, including ways to define how o_1 comes to the decision of selecting player p based on the specific needs of the team. The framework is also made more sophisticated by allowing for teams to rearrange the ordering O through trades by negotiating with each other.

1.2 Contributions

This thesis intends to provide the following contributions:

1. Exploring several applications of artificial intelligence techniques to the field of sports analytics, with specific focus on the sport of hockey. This field is still in its infancy — especially with respect to leveraging artificial intelligence — and hopefully the methods and results provided can form the ground work for future study. Throughout the work, there are suggestions for future directions of research that go deeper into specific topics.

2. An approach to calculating agents' utility functions over a set of objects — or in this case hockey prospects — from past decisions using linear programming is proposed. Since linear programs can be solved relatively quickly, this method allows for generation of tailored utility functions quickly.

3. A pruning algorithm for trade offers in the picking sequences game is proposed to limit the number of possible sequences that must be evaluated before the optimal trade offer is identified. This algorithm can have applications outside of sports analytics where the utility functions of agents with respect to picking order are positive and monotonically decreasing. That is, any application where agents want to pick as many times as possible and as early in the order as possible.

4. Classifiers are provided for determining the probability a player will play in the NHL and in what capacity they might play. These classifiers can help predict if a future prospect is likely to play professionally, which is important for management staff making draft selections. There is also insight from the algorithms themselves as to which attributes are likely predictors of future success, which can be incredibly valuable information for development and coaching staff that are trying to make their players the best that they can be.

1.3 Statistical and Scouting Data

Throughout the experiments in this thesis there will be many types of data used to describe each individual player. This data will mainly fall into either primary statistical data that could be seen on a game scorecard or scouting ratings collected by the International Scouting Services (ISS) scouts over the course of several trips throughout a season. Ideally the scouting information provides an accurate and unbiased depiction of each player's strengths and weaknesses, whereas a player's statistics show how productive a player was with their current team and opponents.

The scouting information provided is over 10 attributes (skating, puck skills, shot, offensive, defensive, physical, compete, sense, strength, explosiveness) which is too broad to consider all of them in a meaningful analysis. Thus the data from this source is condensed into 6 "tools" to more concisely describe a player's skill set.

1. Skating combines skating, compete, and explosiveness.

2. Shooting combines shot, puck skills, offensive, and strength.

3. Passing combines puck skills, offensive, and sense.

4. Defensive Ability combines defensive, compete, and sense.

5. Grit, or Toughness, combines physical, compete, and strength.
6. Hockey Sense, which is a player's general understanding of the game and the ability to be in the right place at all times, combines offensive, defensive, compete, and sense.

Data points representing a player's skill in each specific tool were normalized to their z-score among data points for that tool.

While scouting ratings show a direct description of a player's skill set during the particular games they were scouted, the statistics they obtain over the course of a season can provide evidence supporting high or low skill in each of the tools. Such relationships would be that a player with a good shot is likely to score more goals, and similarly a good passer with good hockey sense will likely set up other teammates very well leading to more opportunities to be credited with assists. Another example is a tougher player that is hitting opponents often or getting into fights is likely to receive more penalty minutes. There are benefits to using statistics over scouting data to make up for its indirect description of player skill. Namely there are a limited number of scouting visits possible during a year so the sample size of statistics that are tracked every game is far greater. In addition there can be bias effects that appear in the human scouting reports, so statistics can show an objective picture of how a player played during a season if their context is accounted for.

1.3.1 Additional Data Specification

When comparing statistical data from players in different leagues, it is important to account for that context and degree of difficulty that each league poses. To account for this effect in our experiments we use two different approaches. The first method is to simply group leagues together by similarities. For example the Canadian major junior leagues (Ontario Hockey League, Western Hockey League, and Quebec Major Junior Hockey League) are very comparable to each other in difficulty for young players, but might be easier than European professional leagues such as Finnish Liiga or the Swedish Hockey League where prospects play against grown men. Then for classification using this approach a discrete value could be used to denote 1 for Canadian major junior leagues, 2 for European professional leagues, and so on.

Another commonly used way to establish the context of a league's difficulty is the statistic known as NHLe. This statistic is calculated as the product of a difficulty modifier and a prospect's points per game to result in an expected number of points scored by that player had they been playing that year in the NHL. Thus the modifier can be viewed as a quantitative value for $\frac{\text{Difficulty of prospect league}}{\text{Difficulty of NHL}}$.

For established players in the NHL there are additional pieces of data tracked that can be of use when analyzing the style and performance of players. Of particular note for Chapter 5 will be the statistic known as Corsi or simply "shot attempts". This value tracks the ratio of times a player's team or the opposition makes an attempted shot that is on goal, blocked, or a missed shot. It is a indicator of how often a player's team is in the offensive zone compared to the defensive zone, and is often used in favour of other options such as plus minus (+/-) due to a larger sample size. For comparing player types Corsi is useful to identify players that might not score many points, but contribute to the team by keeping the opposition from having scoring chances of their own.

One appealing trove of information that is available for considering prospect players are the qualitative descriptions that scouts provide. The distinction between this source and the scouting ratings that will be used in this thesis is that when scouts write a report they provide numerical values for each skill, but also a story in text explaining what stood out about each player scouted in that game. By collecting all stories about a player over all scouting visits, one can have a sizable corpus to use as input for various natural language processing schemes. This idea has been touched on for some player projection work similar to what is done in Section 4.2, but using data from the National Football League [18].

1.4 Literature Overview

The framing of the draft for game theory is comparable to past works in picking sequences where agents select objects out of a pool of items [7], without perfect information in the draft setting. Work by [7, 9] investigate how manipulations affect the picking sequences game. It has been shown that teams acting strategically with perfect information concerning ordinal preferences may cause Prisoner's Dilemma situations [19, 8] in picking sequences and drafts. The setting of the draft analysis of strategic picking covered in Section 2.5 differs from [19, 8] in that agents have imperfect knowledge of the others' intended selections. Some work exists comparing different orderings for picking strategies which ties into the trade modelling done in this thesis [13].

Some work has been done in the direction of assigning utilities to draft picks, as Schuckers proposes assigning the expected career games played by players drafted at each pick position [17]. If one considers each team assigning utility to pick positions in the draft, extending on the work by [17], then the problem of pick trading could be framed as an application of the resource allocation game, with work in [2, 16] being of relevance to this framing of the problem. A final paper considers bidding strategies in fantasy drafts which considers a draft as an auction where each team has a budget they may spend [1].

The application of artificial intelligence to sports analytics topics is a relatively new though growing field of interest. Coupled with the infancy of hockey analytics as a whole, most research pertaining to artificial intelligence in sports concerns sports such as soccer. Success has been found using neural networks to isolate goal events in soccer videos [22, 23], which could have applications for the media such as creating highlight reels. These solutions tend to rely heavily on audio cues such as crowd noise as opposed to generating a deeper understanding of the sport, though they have incredible application value. One work focussing on hockey used machine learning combined with natural language processing techniques to first predict the winner of a single hockey game, then extend the idea to predict a best-of-seven series of games between the opponents [21]. The analysis in [21] touches on the important issue of bias with respect to textual input. In their case the input was pre-game expert analysis stories, though the issue of bias could easily be present if using sentiment analysis on textual scouting reports.

With respect to hockey analytics, the bulk of the work available has been done using pure statistical methods as opposed to artificial intelligence techniques. While to our knowledge there is no prior published work on applying artificial intelligence to the NHL Draft, there are many pieces of research available on analyzing the draft in using other tools. A common topic for research in the draft can be characterized as a search for "market inefficiencies." These are characteristics that cause prospects to be overlooked by some teams during the draft, allowing players with higher chances of succeeding in the NHL to be available "cheaper with later picks than their probability of success would suggest. Two such papers [3, 11] identified relative age as a characteristic causing a market inefficiency in that players that were a few months younger than their draft peers were chosen significantly later given their future production in the NHL. In fact [3] took the idea one step further by indentifying a second market inefficiency of birthplace population size where prospects born in smaller cities or towns were less commonly drafted or drafted later, possibly due to less scouting exposure from a young age.

1.5 Overview

Chapter 2 will investigate creating a multiagent model for framing the draft. This includes generating utility functions for agents based on past decisions, as well as mechanics for greedy and strategic drafting. Chapter 3 extends these notions by allowing agents to trade picks resulting in a reordering of the draft. Processes for determining the best trade to offer and the utility benefits to these trades is determined. Chapter 4 focusses on utilizing decision trees to predict if a given prospect will have success at the NHL level given stats and scouting data from their draft year. The findings from these decision trees are taken a step further in Chapter 5 by using clustering techniques to annotate existing professional players into similar groups of playstyles. These data points are then used to form training data to enable predictions of which cluster a given prospect would fall into if they end up playing professionally.

Chapter 2

Multiagent Models for the Entry Draft

This chapter focuses on developing a way to look at how individual teams will make selections and develop their strategy during a draft. This is achieved by defining a utility function for each of the 30 current NHL teams — not including Vegas — over the set of all players available. These functions will be the driving factor for each decision that a team makes in the simulations that will follow. While this chapter goes through generating utility functions and determining player selection policies, Chapter 3 will explore an expanded model that allows teams to interact with one another through trades.

2.1 Utility Function Specification

The ideal utility functions for describing teams in this model will be able to accurately reflect how teams have made their decisions in the past, while also having the flexibility to account for deviations from the standard type of selection. We propose three main factors in determining how much a team t will like a certain player p:

1. Team's preferences w_t over the player's skillset s_p, denoted $w_t \cdot s_p$. This calculation considers the player p's prowess at each skill multiplied by the importance of that skill to the team t for each skill represented in the vectors w_t and s_p. Each element in the vectors w_t and s_p maps to one of the 6 tools specified in 1.3.

2. Balance of team t's depth at each position pos, denoted $d_{t,pos} \in \mathbb{R}^+$. For example, if t has particularily few forwards (F) then $d_{t,F}$ should be a larger value than average to indicate that this is a position the team needs more players in.

3. A random aspect to encompass utility changes from other sources, denoted $r_{t,p} \in \mathbb{R}^+$

This results in the following definition of the utility gained by t for drafting p:

$$U_t(p) = (w_t \cdot s_p) d_{t,p} r_{t,p} \tag{E1}$$

The first term is the core of the idea for team utilities. Specifically, teams have inherent likes and dislikes for types of players and playstyles. These profiles will be established as the weight vector w_t which is combined with the player's skill vector s_p with a dot product to return a team's base skill preference for that player. Due to differences in the skills required for forwards and defensive players, teams will have a different weight vector for each position. The values for the s_p vector are found by taking scouting rankings from ISS and combining them into six tools as described in Section 1.3. Determining the values that make up w_t will be covered in Section 2.1.1.

The second term exists to acknowledge the fact that a team will only play a certain number of players at each position. In general this breakdown is 12 forwards and 6 defensemen for a 2:1 ratio. By looking at how many prospects a team already has at each position it is possible to find if the team will be unbalanced in the future, and managers may look to draft players to fill scarcity to avoid a problem in the future. This can be a minor factor in the decision making process, as drafting the player you feel is the best regardless of the position can still lead to trades where a team can balance their depth. The effect of $d_{t,pos}$ can be seen as tipping the balance between two otherwise equally preferred players at different positions. For the implementation, prospect depth was acquired from hockeysfuture.com to find the number of forward ($n_{t,F}$) and defensive ($n_{t,D}$) prospects. Then if p is a forward:

$$d_{t,F} = \frac{2n_{t,D}}{n_{t,F} + 1}$$

with a corresponding definition for defensive prospects:

$$d_{t,D} = \frac{n_{t,F}}{2n_{t,D} + 1}$$

To limit the extent of this factor in extreme cases where teams may have far more of one position than the other a value of $d'_{t,pos} = (d_{t,pos})^{\frac{1}{3}}$ is used in the following experiments.

Since it is unlikely that $d_{t,pos}$ would ever be greater than 9, the factor $d'_{t,pos}$ should not allow either position to be more than twice as preferred as the other.

The purpose of the $r_{t,p}$ term is to account for all the other factors influencing the draft decision that are not found in the data. For example, each team has its own scouting staff that only has time to visit each prospect a limited number of times over the course of the year. It is entirely possible that over a small sample size a player performed uncharacteristically well or poorly, causing the team to have a skewed opinion of them. Other factors could be that the franchise has ties to relatives of a player or the team they were recently playing in, amoung other qualitative issues that could alter a team's decision. The parameters defining the distribution of $r_{t,p}$ were set to balance the importance of teams having their own unique preferences from their weights and depth against allowing the utility functions the flexibility to sometimes lead to other selections. In the implementation for this thesis, $r_{t,p} \sim Gaussian(1, 0.1)$.

2.1.1 Determining Weights

The goal for the team skill weights are to fairly represent what skills matter most to the team using past draft selections as evidence. This will result in a vector $w = (w_{\text{speed}}, w_{\text{shot}}, w_{\text{pass}}, w_{\text{defensive}}, w_{\text{grit}}, w_{\text{sense}})$ where the sum of the elements is equal to 1. We also insist that all terms are non-negative. That is, no team will view any skill as a bad quality. For example no team will actively prefer a slower player, but such effects might be seen anyway if they prefer types of players with skills that negatively correlate with speed. In such a case, the weight of speed may be low or even zero, but the reasoning for the selection will be high rating in other skills, not specifically because the player is slow.

The data for identifying the team weights will be how that team has made draft decisions in the past. We consider all past draft selections in the following way:

If t selects player p while player q is still available, we conclude that $U_t(p) \geq U_t(q)$. (Rule 2.1)

If we consider pairs of players such that p and q play the same position (suppose forwards, F without loss of generality) we can eliminate the effect of $d_{t,F}$, and focus on the weights specific to that position. When looking for the weights, we assume as much utility as possible is driven by the team preferences and player skill, and we want to minimize the effect of the random factor $r_{t,p}$. By modifying the definition of U_t to use $r'_{t,p}$ as an additive factor we can set up a linear program by finding constraints using (Rule 2.1) and then minimizing the summed values of $r'_{t,p}$ for each player pairing, calculated as $r'_{t,p} + (-r'_{t,q})$ to demonstrate positive random utility associated with p and negative random utility associated with q. If $r'_{t,p}$ was negative or $r'_{t,q}$ was positive, this would only strengthen the need for utility from the weight vector w_t to ensure that $U_t(p) \geq U_t(q)$, so the worst case is assumed. Pairs are formed between each player p selected by a team since 2012 and the five players $(q_1, ..., q_5)$ that were next selected at that position. For example, the Edmonton Oilers selected forward Conner McDavid first overall in the 2015 draft, so if p is Conner McDavid and t is the Edmonton Oilers then players $(q_1, ..., q_5)$ would be forwards Jack Eichel, Dylan Strome, Mitchell Marner, and Pavel Zacha who were selected second, third, fourth, and sixth respectively in the 2015 draft (a defenseman was selected fifth). Consider the simplified linear program for just p, q_1, and q_2:

$$
\begin{aligned}
&\min\ r'_{t,p} + (-r'_{t,q_1}) + (-r'_{t,q_2}) \\
&\text{such that:} \\
&w_t \cdot (p - q_1) + r'_{t,p} - r'_{t,q_1} \geq 0 \\
&w_t \cdot (p - q_2) + r'_{t,p} - r'_{t,q_2} \geq 0 \\
&r'_{t,p} \geq 0 \\
&r'_{t,q_1} \leq 0 \\
&r'_{t,q_2} \leq 0 \\
&\sum w_t = 1 \\
&(w_t)_i \geq 0\ \forall (w_t)_i \in w_t
\end{aligned}
\tag{E2}
$$

This can be simplified slightly by noting that the most efficient way to assign value to the slack variables such as $r'_{t,p}$ and r'_{t,q_1} will be to load as much utility as is required into $r'_{t,p}$ leaving all of the r'_{t,q_i} values as 0. Therefore we only need to include the r' terms that represent additional utility given to p instead of any utility that might be taken away from any q_i.

To put things into a concrete example, consider the following data from the 2015 draft:

Player	Skating	Shot	Passing	Defensive	Grit	Sense
MM	6.63	3.81	7.65	5.76	0.54	8.41
PZ	5.05	7.01	5.08	4.97	5.46	5.05
TM	1.33	3.65	2.32	2.31	3.67	2.71
MR	4.09	5.99	5.17	4.47	5.58	6.36
LC	3.30	5.33	1.42	3.96	6.69	2.81
DG	3.20	4.10	3.22	0.69	4.27	3.04

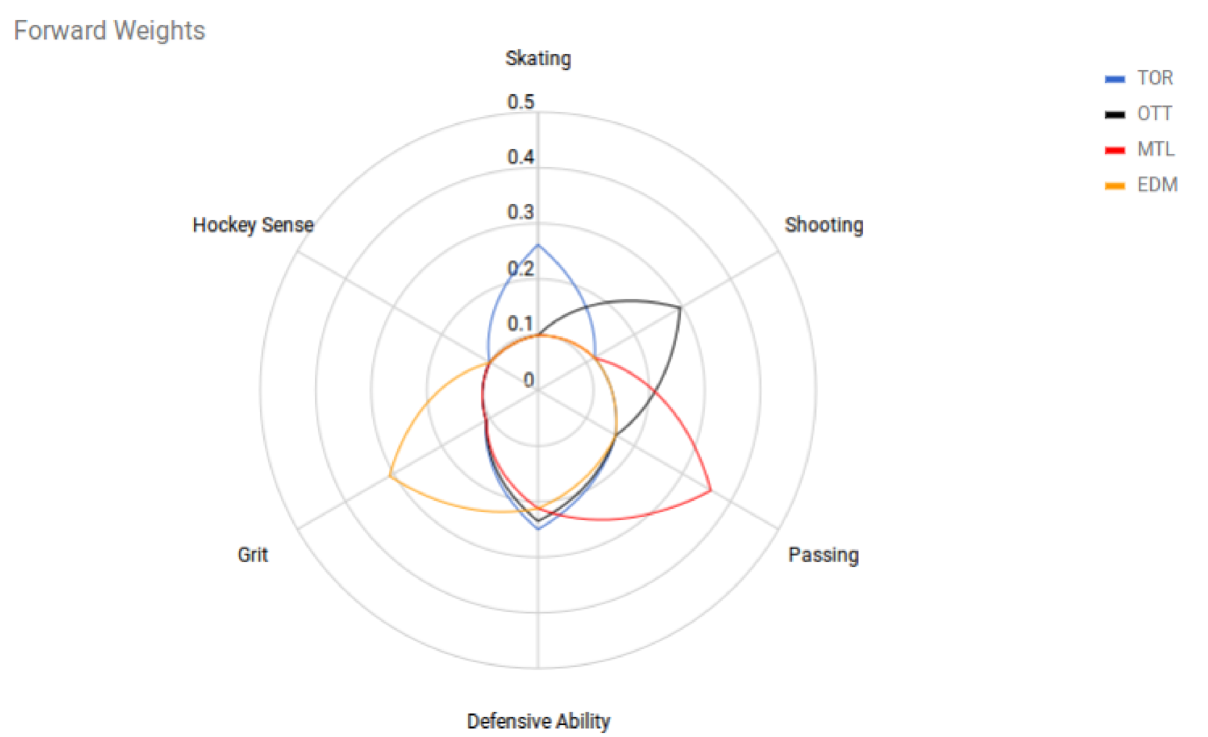
Table 2.1: Skill values of players in 2015 NHL Draft

For the team that selected MM while PZ, TM, MR, LC, and DG were still available, a potential weight vector that satisfies all contraints could be $(\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{4})$ which would require a value of $r_{t,\mathrm{MM}} = 0$.

Another set of weights w_0 was found by considering all players taken by all teams to represent the general preferences of the league by using the linear program described with all players drafted, not only those selected by a single team. To combat extreme weights that would arise for some teams with relatively few historic picks to draw from, all weight vectors were balanced with the general league weights. For the purposes of the experiments performed, a value of $w'_t = \frac{2w_t + w_0}{3}$ was used in place of w_t in the utility calculation (E1).

2.1.2 Bringing it All Together

With all of the parameters for the utility functions set, it is possible to evaluate how the functions compare with two core ideas in mind. First of all the functions should reflect individual preferences so that teams do not have identical values for their selections. If this were the case the order of players selected would be identical regardless of the teams drafting at each pick. Secondly, the individual differences in utility functions should not be so great as to break implicit correlations in utility that can be found in reality. A way to think about this point is that if team t_1 likes player p a lot then it is likely that p is actually a good player. If that is the case then it is also likely that team t_2 likes p because teams should tend to like good players. Of course the amount by which t_1 and t_2 prefer p may vary, but it should not be the case that a player near the top of t_1's ordinal preferences list is near the bottom of the list for t_2.



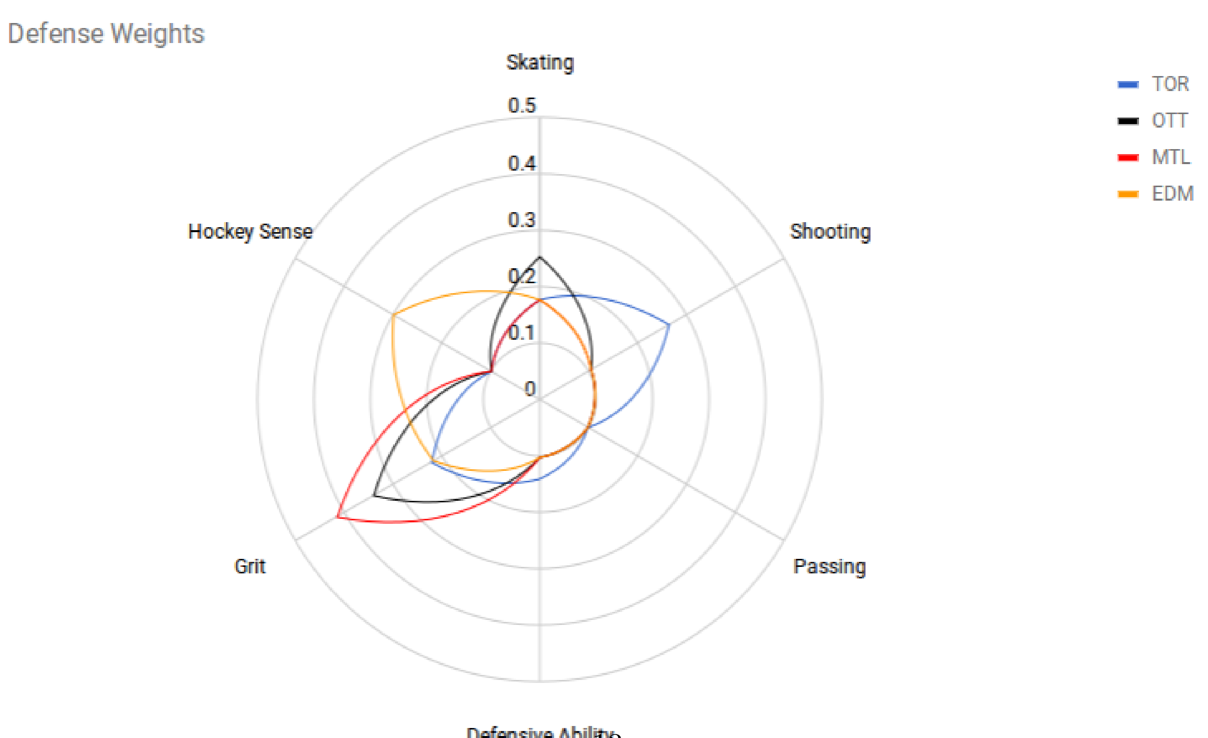


Figure 2.1: Skill preference weights assigned to Toronto, Ottawa, Montreal, and Edmonton.

Figure 2.1 shows how several Canadian teams have valued skills of forwards and defencemen historically. It is clear that there are definite differences in what teams look for in draftees, and even differences in how teams evaluate players depending on their position. For example with forwards the Toronto Maple Leafs value skating while the Montreal Canadiens look for passing skill most among potential forwards. For defencemen, the Maple Leafs value shooting most and the Canadiens value grit. While this is suggestive that teams have unique utility functions, further evidence is required to demonstrate this and to also show that these utility functions do not go too far in weighting team preferences to the point where teams would be willing to pass on top quality players. One noticeable aspect from Figure 2.1 is that aside from Edmonton there is little to no value placed on the attribute of hockey sense for either forward of defencemen. This was common in the full results as well with very few utility functions putting any significant weight on hockey sense.

Next we consider both the ordinal preferences and the numerical utilities of these same five teams for their favourite 10 players for the 2015 NHL Entry Draft.

TOR	OTT	MTL	WPG	EDM
CM 69.9	CM 57.7	CM 48.1	CM 48.7	CM 68.1
JE 41.6	JE 41.9	JE 44.8	NH 35.6	JE 44.1
NH 37.7	NH 40.9	NH 38.9	JE 33.7	NH 34.5
PZ 26.3	ZW 28.5	MR 26.2	ZW 26.2	PZ 27.1
MM 25.5	MB 27.5	ZW 25.1	MM 25.5	MM 26.3
OK 24.4	MM 27.2	OK 24.8	PZ 25.3	MB 26.1
MB 24.2	MR 25.6	MM 23.5	OK 25.0	OK 24.7
MR 24.0	OK 23.6	MB 23.2	MB 21.5	MR 23.4
ZW 23.3	PZ 23.3	TK 22.6	CW 20.2	NM 22.4
NM 21.8	CW 22.5	CW 22.5	MR 19.9	DS 22.1

Table 2.2: Ordinal preferences and associated utilities from (E1) with weights from (E2) of 2015 draft prospects for Toronto, Ottawa, Montreal, Winnipeg, and Edmonton. Forwards are shown in green and defencemen in red.

One of the most noticeable features of Table 2.2 is that all teams rank prospect CM higher than any other, regardless of the differences seen in Figure 2.1. As this particular player was generally considered to be a once-in-a-generation talent, this is to be expected. It supports the second desired property where teams will still recognize top level skill instead of having an elite player low on their list due to their skill breakdown.

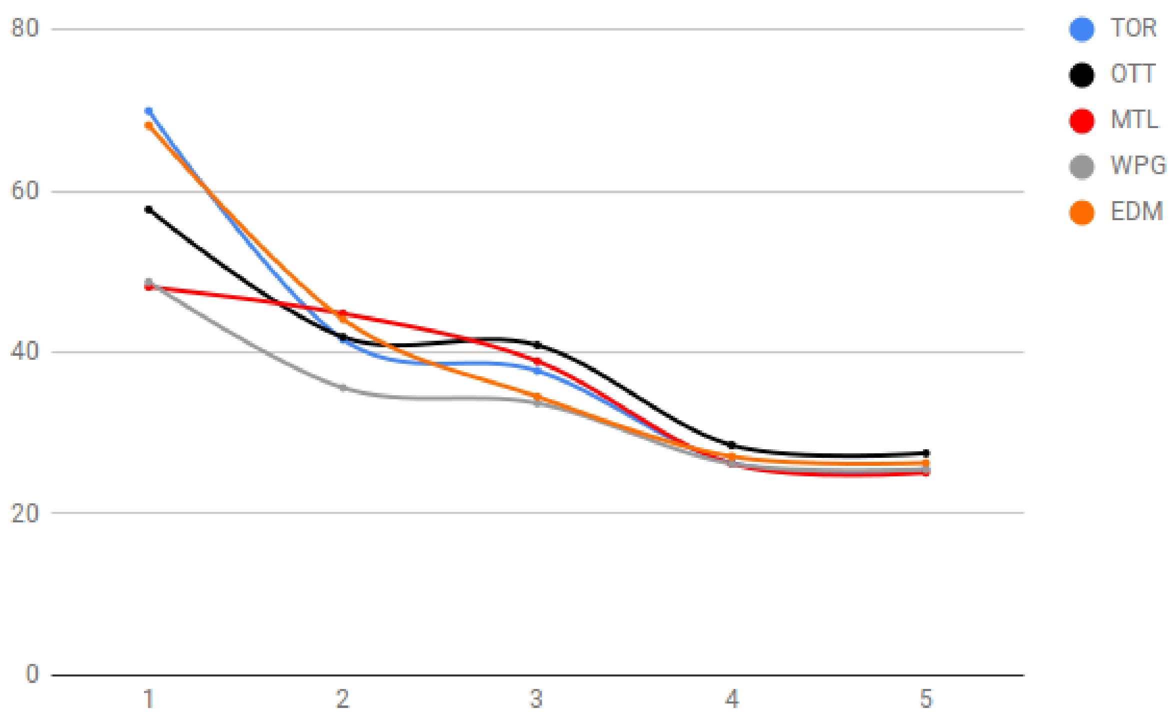


Figure 2.2: Utilities as a function of ordinal rank for Toronto, Ottawa, Montreal, Winnipeg, and Edmonton.

2.2 Expected Utility

The most interesting questions to tackle with this utilitarian approach require teams to be able to reason about the future consequences of actions they or others might take. For this reason, we must introduce the ability for teams to calculate expected utility for themselves or estimate the value for others considering the resulting draft that would occur if they took various actions. The expected utility for team t over pick order o will be denoted $EU_t(o)$. We use a Monte Carlo approach where for the values of w_t and the r values of the other 29 teams are drawn at random where $r_{t,p} \sim Gaussian(1, 0.1)$ and each element of w_t is drawn uniformly from the set of values such that the constraints of (E2) are satisfied. A draft is then simulated over the result of the action being considered where each team t's utility profile is generated from the drawn values of w_t and $r_{t,p}$. The average utility that the simulations give for team t is the expected utility of t for the action. In our simulations for expected utility we limit the agents to be myopic in the sense that they will assume all future selections will be made naively and without future pick trades occurring in order to keep runtime manageable. If a simulation of 1000 naive drafts each with 99 picks takes one second then 1000 drafts simulations each requiring 1000 subdrafts of 98 naive picks would take 16.5 minutes, and so on for 68 more levels.

2.3 Utility Functions of New Teams

On occasion leagues have decided to expand their league by adding one or more teams at the start of a future season. For the NHL, the example that will be looked at is the Vegas Golden Knights which entered at the start of the 2017 - 2018 regular season. The main problem with trying to develop a utility function for the Golden Knights is that prior to the entry draft between June 23 and 24 there has been no draft history to draw upon for setting their skill preference weights, so an alternate solution is required.

An important event that occurred before the 2017 NHL Entry Draft is known as the expansion draft. This is a special event that occurs when new teams enter the league to allow them to create a roster of players by selecting one player from each team. The established teams may all protect several of their most valuable players from the expansion draft, but what remains still gives the Golden Knights plenty of options available to determine the identity of their franchise. As such, it would be possible to generate a set of constraints similar to the ones described above where instead of comparing players selected to those selected later in the draft, the comparisons could be for each team the player that was selected compared to all other exposed players from the same team. One main drawback to this approach is that the expansion picks have more to them than judging skill sets since the player's contracts carry over to Vegas as well. Then it is entirely possible that when choosing between a favourably skilled player with a poor contract and a different player with a long term and cheap contract the decision might be made from a business standpoint rather than which player fits their system the best.

Another approach to take is to consider the people who will be making the draft selections for the Golden Knights. Each have been a part of the league with other teams for a long time, often making the same types of draft decisions. By taking the union of the draft selections made with other clubs over the years it might be possible to find a reasonable picture of what these managers will be thinking when it comes to finding prospects for their new franchise. For our purposes we have chosen to follow this approach and use the history of the Golden Knight's general manager from when he was with the Washington Capitals between 1997-2014 and the New York Islanders between 2015-2016.

2.4 Greedy Drafting

With team utility functions set, selection policies can be investigated. The first and most basic policy is greedy drafting. According to this form of selection, at each pick the team t will rank the remaining available players into ordinal preferences according to $U_t(p)$ and select the top option. Algorithm 1 is a recursive definition for how an entire draft would be run if all teams picked according to the greedy policy. It takes the standard draft parameters as input and returns the list of drafted players in the order that they were selected.

Algorithm 1 greedyDraft$(P, T, \langle o_1, o_2, ..., o_n \rangle)$

Input: set of prospects P, set of teams T, draft order O
Output: list of drafted players
  $p^* \leftarrow \text{argmax}_{p \in P} U_{o_1}(p)$
  return $\langle p^* \rangle$ + greedyDraft$(P \backslash \{p^*\}, T, \langle o_2, ..., o_n \rangle)$

With the greedy policy, teams are able to maximize the utility gained in the moment of each individual selection.

Pick Number	Team	Predicted Player	Utility	Actual Selection
1	Toronto	Auston Matthews	46.0	Auston Matthews
2	Winnipeg	Patrik Laine	32.9	Patrik Laine
3	Columbus	Jakob Chychrun	37.4	Pierre-Luc Dubois
4	Edmonton	Jesse Puljujarvi	43.5	Jesse Puljujarvi
5	Vancouver	Dante Fabbro	31.7	Olli Juolevi
6	Calgary	Max Jones	35.4	Matthew Tkachuk
7	Arizona	Matthew Tkachuk	30.6	Clayton Keller
8	Buffalo	Tyson Jost	33.4	Alexander Nylander
9	Montreal	Clayton Keller	28.5	Mikhail Sergachev
10	Colorado	Michael McLeod	25.7	Tyson Jost

Table 2.3: Predicted top 10 selections for 2016 NHL Draft

Team	Number of Picks	Total Utility
Anaheim	4	42.1
Arizona	3	58.7
Boston	3	49.4
Buffalo	5	64.9
Carolina	6	79.7
Calgary	5	69.2
Chicago	4	41.2
Colorado	3	46.9
Columbus	3	59.9
Dallas	2	23.4
Detroit	3	41.1
Edmonton	5	77.2
Florida	4	41.1
Los Angeles	1	12.0
Minnesota	1	22.7
Montreal	3	41.9
Nashville	4	44.1
New Jersey	4	53.2
NY Islanders	2	24.5
NY Rangers	2	10.6
Ottawa	2	37.1
Philadelphia	5	63.6
Pittsburgh	3	26.1
San Jose	1	9.7
St. Louis	3	40.2
Tampa Bay	5	56.1
Toronto	6	92.8
Vancouver	2	39.8
Washington	2	21.8
Winnipeg	4	65.8

Table 2.4: Predicted complete draft utilities for top 100 selections of 2016 NHL Draft

Table 2.3 shows how the model predicted the top 10 selections of the 2016 NHL Draft, and how much utility would be gained by each team by the end of the top 100 selections.

For consistency between results, the values indicating random utility between teams and players, denoted as $r_{t,p}$ in Section 2.1, will be held constant for the experiments in the remainder of the chapter.

2.5 Strategic Drafting

An important question in draft strategy is whether it is ever beneficial to select a player other than your first ordinal preference, hoping that the player will be available for selection with your next pick. This is known as "strategic drafting" and it allows for the opportunity for a team to select both of their top two favourite players at the risk that you may only be left with your second choice when you could have had a player you preferred instead if you had made the naive choice. In terms of the draft simulation, this is equivalent to picking a player that results in higher expected utility that is not the same as the player with maximal utility value. The major difference between Algorithms 1 & 2 is that the strategic draft makes use of $EU_{o_1}(p)$ in place of $U_{o_1}(p)$ where EU is the expected utility function for picking p given order o_1 and U is the utility of p in the context of order o_1.

Algorithm 2 strategicDraft$(P, T, \langle o_1, o_2, ..., o_n \rangle)$

Input: set of prospects P, set of teams T, draft order O
Output: list of drafted players

$p^* \leftarrow \text{argmax}_{p \in P} EU_{o_1}(p)$
return $\langle p^* \rangle$ + strategicDraft$(P \backslash \{p^*\}, T, \langle o_2, ..., o_n \rangle)$

Simulations were run on the first 100 picks of the 2016 NHL draft with Algorithm 2, counting instances of strategic picks where a player p was selected in the context of order o_1 when there existed some other player q ($p \neq q$) for which $U_{o_1}(q) > U_{o_1}(p)$. That is, player p was selected when another player q would have been selected in that context with Algorithm 1. It was found that the action that nearly always results in highest expected utility is to draft naively, with strategic picks occurring on 4.2% of the 70 picks where strategic drafting is possible (on a team's last pick, the naive action is trivially optimal). It is expected that this is due to the fact that for most picks the next pick owned by the team is 30 selections later. Coupled with the inherent correlation between utility functions of teams this means that if a player is one team's favourite then he is likely good enough to be drafted by another team before the team's next pick.

This raises an interesting question of how close picks need to be for strategic drafting to become a legitimate option. To investigate this, we created drafts with n many teams where picks were evenly distributed, so that each team makes a selection once every n picks. Then by varying the value of n, we see the effect of pick density on the frequency of strategic picks.

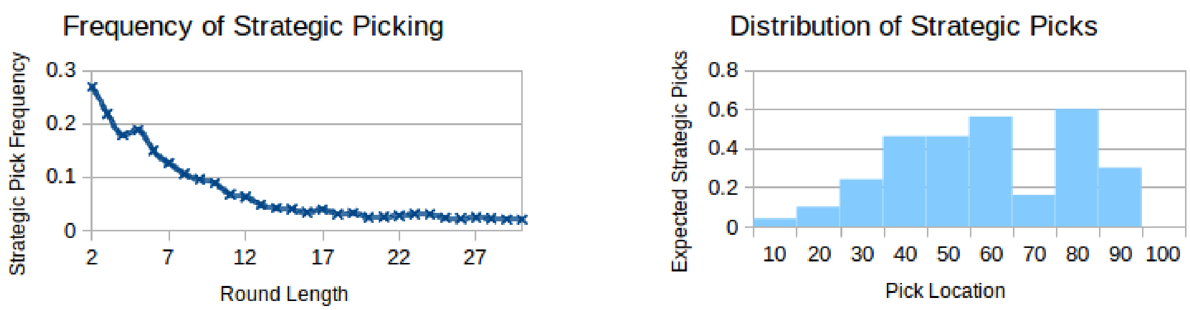


Figure 2.3: Results of Strategic Drafting Experiments. Left: Frequency of strategic picks decreasing as picks become sparser. Right: Histogram showing the distribution of strategic picking events on 2016 Entry Draft order.

As Figure 2.3 shows, strategic picking does become far more common when picks are closer together. This data accounts for the fact that final picks for teams are always naive. Also shown is that strategic picking is more common in the middle portion of the draft rather than early when there are clearcut choices or late when teams begin to run out of picks. The picks in the 61 to 70 range were owned by teams whose next picks were relatively far away compared to other stretches, which accounts for notably few instances of strategic picks.

2.6 Extensions to the Utility Model

This utility model could be expanded upon in several ways for future work. One way would be to incorporate the concept of complimentary players. For example, if a team drafted only players with high shot score but never a player known for being a strong passer, the shooters would have a harder time getting a good opportunity. Thus teams often strive to find a balance of skill sets when building their teams. A related extension would be having teams address areas of weakness in their game. For example a team with the worst-ranked defense in terms of goals allowed would have incentive to draft a player with a higher score in defensive attributes. Likewise, a team that struggles to score goals would possibly target offensive-minded prospects.

A possible way to generate player attributes that was not investigated would be to leverage natural language processing with respect to written scouting reports as opposed to the numerical grades assigned as a summary of the report. Generating the attributes through this processing step might lead to similar ratings if the report summaries tend to match the content of the report well, but otherwise could lead to interesting analysis of the differences in the results. Using the qualitative data of the report could also enable more granular utilities than the summaries provide. For example, two teams might target players with high shot scores, but perhaps one prefers players with accurate shots and the other wants players with extremely hard shots. Both attributes would lead to high shot scores, but could be entirely different subsets of players. Knowledge of the difference on this dimension could increase performance in all draft experiments — including strategic drafting in Section 2.5 and drafting with trades as will be covered in Chapter 3.
In terms of teams adopting a strategic drafting plan, an enhancement not covered in the experiments is using the knowledge of picks within the current draft to shape each agent's predicted preferences as the draft unfolds. These picks could give a depiction of each team's immediate needs and preferences that historic picks do not, so could be given greater weight.

Chapter 3

Drafting with Pick Trades

In a real amateur draft there is often more happening at the event than teams announcing who they will select next. Team managers are able to call each other and propose swapping of team assets usually to move up or down in the draft order depending on when they believe a particular player they are interested in will be available. In the actual draft these trades could involve roster players, past draftees, or even picks in future drafts — though the focus of this chapter will be on the exchange of picks for picks in the same draft. Being able to evaluate trades and expected values for individual picks will be immensely important as to judge what selections a team is willing to give up for the chance to draft next. As a simplification, teams may only attempt a trade if it involves the upcoming pick in the draft. This provides structure to the process as for each pick there can be a single round of trade offers for the selection followed by one possible order change then a continuation of the draft. The goal is that the system of agents will be able to identify and perform trades which are mutually beneficial to the parties involved. The example in Table 3.1 between teams x and y shows how a mutually beneficial trade could look in a simple 2-team 2-round draft.

Prospect	U_x	U_y
A	11	10
B	6	8
C	5	7
D	4	1

Table 3.1: Example Utilities for Trading

For the draft that follows $O = \langle y, x, x, y \rangle$, x would receive prospects {B,C} for 11 utility and y would get {A,D} for 11 utility. Now consider an $O' = \langle x, y, y, x \rangle$ that x might propose. This would result in x and y each receiving 15 utility, which is preferable for both teams. In the 30-team setting for this chapter, finding trades like this becomes considerably more complicated especially since teams will not have full knowledge of how others view utility or what offers other teams are submitting.

3.1 Utility of a Trade Offer

By looking only at pick trades each trade offer can be seen as a unique draft order. From the perspective of the agent that is receiving trade offers, it will have a choice of up to one offer for each team in the draft plus the original draft order if it chooses to keep the upcoming selection. From that point it is easy for the agent to calculate the expected utility for each draft order through Monte Carlo simulations and select the most preferred draft order. Algorithm 3 shows the general structure for the draft used in this chapter, with a single trade round prior to each pick. The real challenge will be for the agents trying

Algorithm 3 draftWithTrades(P, T, O)

Input: set of prospects P, set of teams T,
draft order $O = \langle o_1, ..., o_n \rangle$
Output: list of drafted players
offers $\leftarrow \{O\}$
for $t \in T$ such that $t \neq o_1$ **do**
offers $\leftarrow$ offers $\cup$ {createOffer$_{t,o_1}(O)$}
end for
$O' = \langle o'_1, o'_2, ..., o'_n \rangle \leftarrow \text{argmax}_{l \in \text{offers}}\ EU_{o_1}(P, T, l)$
$p^* \leftarrow \text{argmax}_{p \in P} U_{o'_1}(p)$
return $\langle p^* \rangle$ + draftWithTrades($P \backslash \{p^*\}, T, \langle o'_2, ..., o'_n \rangle$)

to propose a trade offer as the expected utility for them will incorporate the probability of their offer being accepted alongside the utility they expect to gain from the order they are proposing. For a trade offer $t_{o'}$ from x to y resulting in draft order o' from an original order of o the expected utility for x submitting the offer is

$$EU_x(t_{o'}) = Pr\left[y \text{ accepts } o'\right] * (EU_x(o') - EU_x(o)) \tag{E3}$$

Given that y is known to be a rational agent, it will accept whichever trade o^* it is offered that maximizes $EU_y(o^*)$, including the non-trade state o. If T is the set of all trades offered to y aside from x's offer of o', then $Pr[y \text{ accepts } o'] = 1$ if $EU_y(o') > \max_{o'' \in (T \cup \{o\})} EU_y(o'')$ and 0 otherwise. Since neither T nor the utility profile of y is known to the agent x at the time an offer, the probability term must be estimated by repeatedly drawing random utility profiles for each agent in the manner described in Section 2.2 followed by calculating each agent's optimal trade offer. Then $Pr[y \text{ accepts } o']$ is the average probability of o' being accepted in that Monte Carlo simulation.

3.2 Identifying Potential Trades

Since a pick trade is nothing more than a redistribution of picks between two agents, then if team x who owns n_x many selections wants to submit a trade offer to y who owns n_y picks then there are $2^{n_x+n_y}$ many ways to do that redistribution. Since the first pick owned by y is guaranteed to be in the trade offer, this leaves $2^{s_x+s_y-1}$ offers for x to potentially consider submitting. This set of trade offers will be denoted as the trade space S. In a real draft there is a tight time limit between picks, so calculating the expected utility of each offer in S will be infeasible. Fortunately each trade offer can be related to others in such a way that information gained about one offer can direct the agent towards better offers or away from worse offers without the need to calculate all expected utilities. The key idea is that some offers can be seen to be trivially preferred by one party over another related offer.

At this point, it will be beneficial to introduce a condensed notation for trade offers from x to y where only the relative positioning of picks owned by x and y are considered. For example $\langle x, y, z, x, z, y \rangle$ would become $xyxy$. The example in the previous paragraph could have been written in the notation $xxy \succ_x yxx$ where $\succ_x$ denotes "x prefers", though a more precise definition of what $\succ_x$ means will be given. Condensed notation will be written as a string whereas an absolute order will be written in the usual list notation.

Proposition 1. *Expected utilities as a function of pick position (utility expected from a specific pick position) are positive and monotonically decreasing.*

Proof. Trivially the expected utilities of all pick positions must be positive as the utility functions of all teams are positive for all prospects. Then consider picks at positions p_1, p_2 such that $p_1 < p_2$. That is, p_1 is an earlier pick than p_2. Then the expected returns for any team cannot possibly be any worse than that for p_2 as every player that is available for selection at p_2 is also available at p_1. □

There are two rules that can be defined to encapsulate how orders can be trivially preferred which follow from Proposition 1. The first is called "Addition" and states that everything else being equal, x would prefer to have a pick than not to since utilities are positive. As an example $xx \succ_x xy$. The second rule is known as "Promotion" and states that x would prefer to swap a later pick for an earlier pick if everything else remains the same. An example of this would be $xy \succ_x yx$ where x moves its pick from second in the pool of picks to first.

By combining the rules of Addition and Promotion, chains of dominance can be created between orders that also imply trivial preference due to the transitivity of the preference operation. For example $xx \succ_x xy \succ_x yx \succ_x yy$. Thus, if O_1 can be obtained from O_2 only using applications of the Addition or Promotion rules then $O_1 \succ_x O_2$. As a result from the chain above $xx \succ_x yy$.

Proposition 2. *In the context of a trade between x and y, $O_1 \succ_x O_2$ iff $O_2 \succ_y O_1$.*

Proof. Consider the chain of rules $\langle r_1, ..., r_k \rangle$ for x's benefit used to obtain O_1 from O_2. Each rule r_i has a start state s_i and an end state e_i such that $e_i = s_{i+1}, s_1 = O_1, e_k = O_2$. If r_i is Addition then s_{i+1} differs from s_i by a single pick owned by either x in the case of s_{i+1} or y in the case of s_i. Then a change from state s_{i+1} to s_i is an example of Addition from y's perspective. Similarly if r_i is a Promotion rule then the only difference between s_i and s_{i+1} is a swapping of picks such that s_i has a pick owned by y in a higher position than the corresponding pick in s_{i+1}, so transitioning from s_{i+1} to s_i. Then to obtain O_1 from O_2, transition from O_2 to s_k followed by s_{k-1} until $s_1 = O_1$ is reached. Each such transition will be either an Addition or Promotion rule as shown. □

3.3 Searching the Trade Space

Consider a graph G_S of the trade space S for offers from x to y with $n = n_x + n_y$ picks where nodes are the various trade states (2^{n-1} nodes). Then add a directed edge from node s_1 to s_2 where s_2 can be reached from s_1 by a single application of Addition or Promotion from x's perspective. Then by the definition of domination chains, some state s_2 is a descendant of s_1 iff $s_2 \succ_x s_1$. By Proposition 2, s_2 is an ancestor of s_1 iff $s_2 \succ_y s_1$. Further, since every possible application of rules is represented in the graph, so must all dominance relations through descendant relationships. To save on space required to store edges, only the transitive reduction of the graph will be required in practice.

Proposition 3. *G_S is acyclic and connected.*

Proof. First assume that G_S contains a cycle. Then there is some dominance chain $s_1 \succ_x s_2 \succ_x ... \succ_x s_k \succ_x s_1$. Then by transitivity $s_2 \succ_x s_1$, but s_2 is acquired by applying either Addition or Promotion to s_1 so this cannot possibly be the case, implying a contradiction.

Consider the state $r = xyyy..$, and some arbitrary node s_i. s_i must take the form of starting with x then a sequence of x's and y's. Then s_i can be obtained from r using a chain of Addition rules changing y's to x's at the positions they are found in s_i. Then there is a path from r to every node, so G_S is connected. $\square$

There are two crucial pieces of information for speeding up the search through G_S. The first is that for any node s_i with order o' where $EU_x(o') \leq EU_x(o)$ then any trade offer s_j with order o'' which is a descendant of s must also have an order with lower expected utility than o so $EU_x(t_{s_j}) \leq 0$ where t_{s_j} is the action of submitting the offer s_j. This is because as a descendant of s_i it is the case that $EU_x(o'') \leq EU_x(o') \leq EU_x(o)$ which implies that $(EU_x(o'') - EU_x(o)) \leq 0$. From Equation (E3), the probability term is ≥ 0 so the resulting expected utility must be non-positive. Therefore all descendants of nodes with lower expected utility than the original ordering can be removed from consideration. The second note is that for any node s_i with 0 probability of y accepting the deal, all ancestors s_j of the node must also have zero probability of acceptance. If this were not the case then there would be some situation where s_j could be accepted by y. Since s_i is trivially preferred by y to s_j as shown above through Proposition 2, in that same situation y would accept s_i meaning the probability of acceptance could not possibly be zero. This would be a contradiction. Then since the probability term of offering any such ancestor as a trade offer would be zero, so would the overall expected utility for x of such an action, meaning it would have nothing to gain by offering a trade corresponding to any of those nodes.

By starting the search in the middle of G_S and working outwards in each direction, eventually nodes will be found which cause either all ancestors or descendants to be disqualified from consideration as trade offers.

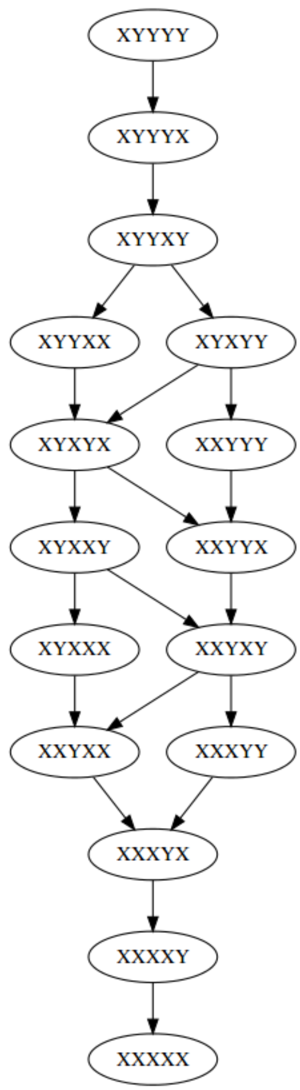


Figure 3.1: The graph G_S where S is the space of 5 picks where X receives the first pick. Arrows point in the direction of states more preferred by X and less preferred by Y.

This helps greatly reduce the amount of computation required to identify the optimal trade offer action for each team. This can be summarized neatly into a table of actions as in Table 3.2 to form an algorithm for identifying the optimal trade offer to submit.
Starting from an arbitrary node in the middle of the graph G_S, nodes can be evaluated from this point outwards to try to identify the edges of the valid space without the need

to evaluate any additional nodes in the invalid region. In the event that the algorithm is exhausted while there are still valid unviewed nodes in G_S (for example if the first node has Probability $= 0$ and $EU_x \leq 0$) then a valid unviewed node can be chosen at random to continue from.

	Probability > 0	**Probability** $= 0$
$EU_x > 0$	Recurse on valid parents and children	Ancestors invalid Recurse on valid children
$EU_x \leq 0$	Descendants invalid Recurse on valid parents	Ancestors and children invalid

Table 3.2: Actions when evaluating trade graph node

Algorithm 4 createOffer(P, T, G_S)

Input: set of prospects P, set of teams T, trade space graph G_S
Output: $s \in G_S$ of proposed draft order

```
if G_S is singleton s* then
    return s*
end if
for s ∈ G_S do
    Count number of ancestors a_s and descendants d_s for state s
end for
s′ ← argmax_{s∈G_S}(min(a_s, d_s))
p′ ← Pr[y accepts s′]
EU′_s ← p′ × EU_x(s′)
if p′ = 0 then
    Prune ancestors of s′ from G_S
end if
if EU_x(s′) ≤ 0 then
    Prune descendants of s′ from G_S
end if
s″ ← createOffer(P, T, G_S \ {s′})
return argmax_{s*∈{s′,s″}} EU_{s*}
```

Algorithm 4 differs from the function called in Algorithm 3 in that it assumes preproccessing is done to generate G_S from the initial draft order.

3.4 Results

For this section, drafts are simulated using each algorithm over the first 91 picks of the 2015 NHL draft with a set of 181 players and 30 teams. Results will consider the value of individual trades from the drafting model incorporating trades, and compare cumulative social utilities between both algorithms over the course of the draft. All results listed use aggregated data of 50 simulations of drafts. To add context to the data, recall from Table 2.2 that first overall selections of superstar players provide utility between 45 and 70 units.

3.4.1 Benefits to Trading

This experiment considers the benefits gained by each participating team when a trade occurs. Recall from Algorithm 3 that for every pick in the draft the current owner may receive trade offers from every other team. A trade occurs if one of the proposed draft orders results in a higher expected utility for the pick's owner than the current draft order. The trade chosen will be the draft order which maximizes that expected utility if multiple trades are eligible. For each instance of a trade that occurs in the simulation two greedy subdrafts are created. One uses the order after the trade and the other uses the original order. The new owner of the pick after the trade will be known as the "trading up" team and the team that originally owned the pick will be called the "trading down" team. A trade is beneficial to either party if the subdraft run on the new order results in a higher utility than the original order would have.

	Average Utility Gain	**% of Trades Beneficial**
Trade Up	0.33	64.3
Trade Down	2.02	85.0

Table 3.3: Utility Gains from Trades

Table 3.3 demonstrates how resulting orders of a trade tend to be mutually beneficial to the parties involved, though the team trading down often experiences a far larger utility gain than the team trading up. This point makes sense considering that the teams offering trades seek to maximize their peer's utility to a degree in order to increase the probability of an accepted trade, and the team trading down is able to choose its favourite of these orders, most of which are likely in its best interest to begin with. Some trades lead to utility loss compared to if they did not occur, but this is an effect of how expectation can differ from reality. For example, a trade may have occurred where a team believed they would be able to draft prospect p with a pick they were receiving, but that player ended up being drafted earlier than expected leaving the team with a less-preferred option and potentially a utility loss. The important result from this experiment though is the fact that through trades both the team trading up and the team trading down can expect positive utility changes and social utility is expected to increase as a result.

3.4.2 Comparing Model Social Utilities

In this experiment, the greedy draft model from Algorithm 1 and the drafting with trades model from Algorithm 3 are compared side-by-side in terms of the cumulative social utility they achieve over the course of the draft, where social utility is defined as the sum of utilities each team's selection by the end of the draft. Each model runs using the same set of utility functions to keep the comparisons consistent.

	Social Utility	**Average Social Utility**
Greedy	1201.1	13.20
Draft w/ Trades	1249.2	13.73

Table 3.4: Comparing Cumulative Social Utility

Table 3.4 shows the cumulative social utilities of drafting with and without trading. These results provide evidence that the model allowing pick trades is also better for the league as a whole, increasing social utility by 4.0%. This extends the fact that trades are mutually beneficial by showing that the gain in utility from the trade is not offset by a loss in utility by all third-party teams.

Chapter 4

Player Projection

Earlier chapters focused primarily on how teams should strategize and interact in a draft based on utility functions generated from evidence from past drafts. To contrast, this chapter will look at specific prospects and making predictions as to their quality and playstyle for five years later. The importance of this type of work is to show teams which types of players they should be targeting for the needs they predict they will have in the future.

The first major issue for drafting prospects is that many drafted players will never play for the franchise's NHL team, but will instead spend the majority of their careers in the minor leagues providing little or no benefit to the franchise. The first goal will be to predict which players will be of a quality to actually play in the NHL, which is investigated in Section 4.2.

Once this baseline has been set, predictions will shift towards finding impact players. The reason that this is important is that if a player is of "replacement" value they are not providing any value over one of the many players in the minors system that they could be cheaply replaced with, even if actually playing in the NHL. This leads to predictions in Section 4.3 for players which not only will play games, but will be impactful in those games.

The final investigation will take things even further by looking at how existing NHL players do not all play identically, and there are really a few styles that players can be grouped into. Going back to some of the ideas in Chapter 2, teams may value players of different styles differently. Thus, being able to predict what style a player will have if reaching the NHL will be of immense value, which is what Chapter 5 will accomplish.

4.1 Data and Method Overview

The experiments in this chapter will primarily focus on supervised learning methods. As such, annotated training data will be required. The main portion of each instance will be statistical or scouting data from the draftees of past drafts, who have now either established themselves or not in the NHL.

Statistical data for players prior to their draft year is limited by the fact that a statistic must be used by the vast majority if not all junior leagues to be included in enough player instances to be of use. Since these leagues also do not generally have as much money, they often track only minimal stats which require only a few employees or volunteers to maintain. The statistics used for player instances from junior leagues are goal (G), assist (A), points (P), penalty minutes (PIM), and plus minus (+/-). To account for some players playing more or fewer games due to injury or the schedule length of each league, each of these statistics is considered as a rate per game played. Age is also included in the data, and is considered a discrete variable. Players are eligible for the draft if they will be at least 18 years old on September 15th of that year, and no older than 21. As the draft occurs in June, it is possible for players to have age values ranging between 17 and 21, though the vast majority are 18 years old. A final element included to give context to statistical data is the league that they were in at the time. Since players come from leagues of entirely different difficulty and style, this is important to do. As there are 30 source leagues in the data, they are grouped together into similar types to form another discrete variable. The league groupings are the Canadian major junior leagues (such as the Ontario Hockey League and Quebec Major Junior Hockey League), NCAA leagues in the USA, European professional leagues (such as the Finnish Liiga and Russian Kontinental Hockey League), European junior leagues (versions of the professional leagues restricted to under-20 players), and North American midget leagues including the Alberta Junior Hockey League, British Columbia Hockey League, and various high school prep leagues. Participation in these league groupings is one-hot encoded in the features in implementation.

Other experiments will use scouting data in the form of the tools defined in Section 1.3 (skating, shooting, passing, defensive skill, grit, and hockey sense) when there are enough data instances to draw meaningful conclusions. This is not always possible in cases where statistical data would work as the instances with scouting data are a subset of those with statistical data. For a player to be scouted they must have played games, so there is statistical data available. However, not all players have relevant scouting data as there are limited resources (scouts) available during the season meaning they can not watch all possible players.

For processing the data to learn the player projections, there are many machine learning techniques that could be applied. While deciding which technique to use it was important to consider the ultimate audience for these projections. Since the use case for the recommendations made by these experiments would be assisting the decisions made by a management team in a hockey franchise, the audience is likely one without formal computer science and machine learning training. There would be a lot of value for these recommendations to not only specify which players will reach certain thresholds, but also what about them makes them particularly good or bad targets. It might not be sufficient to state that a player fell above or below some function returned by a neural network or support vector machine. However, with decision trees and their associated rules it can be possible to explain recommendations in terms of thresholds reached by players in their statistical or scouting data in a similar way that a manager might think about these decisions previously. For example a player could be recommended because "they scored 100 points in the Ontario Hockey League last season" or because "they have excellent skating ability and hockey sense." The simplicity of decision trees could also make them easy to explain to the intended audience and thereby more trustworthy to them.

The decisions trees in this chapter and Chapter 5 utilize CART as the basic algorithm with the Gini impurity metric for determining where to create splits, and a forward pruning policy which ensures all nodes have at least 50 samples to avoid over-fitting. For Sections 4.2 & 4.3 each leaf of the decision trees created will contain a probability of the player reaching a specific threshold. The leaves corresponding to the top 20% of samples — ranked by predicted probability — were defined as the "recommendations" of the resulting algorithm. This corresponds to minimum predicted success rates of 40% and 33% for forwards and defensemen respectively in Section 4.2 and a minimum threshold of 20% for Section 4.3. These recommendations will be thought of as predicted to reach the threshold and those not included will be seen as predicted to fall below the experiment's goal.

Each decision tree diagram can be found in Appendix A, but included in the body of this chapter will be the relevant decisions rules for each. Summary statistics for performance analysis of each tree against the others and against the average success rates of drafted players in the NHL were applicable. Tree performance is estimated using a n-fold cross-validation process where n is the size of the experiment's data set.

4.2 Baseline Projections

The first goal for any player drafted in the NHL Draft is to play in the big league. This is not a certainty for any prospect, not even those drafted in the top five. Thus to give a basic sense of a draftee providing value to a team, it is natural to consider which types of players are likely to play a meaningful number of games in the NHL. These games may or may not be played for the team that drafted them, but that does not matter as even if they were traded to another team they still provided value to the team that drafted them through what they received in return in the trade. The required number of games played will be 164 in this chapter, which is equivalent to two full seasons of play.

The first pair of trees (A.1 & A.2) considers the baseline projections for forwards and defensemen respectively from the basic statistical inputs including source league as a discrete value as described in Section 4.1. Below are the decision rules leading to a recommended player selection along with the proportion of draftees in the leaf who reached the 164 game threshold.

Decision Rule	**Success Rate (%)**
P/GP > 1.24, (+/-)/GP > 0, G/GP > 0.51 NOT in European Jr	56
P/GP > 1.03, (+/-)/GP > 0, G/GP < 0.51, height > 71.5" NOT in European Jr	50
0.86 < P/GP < 1.24, (+/-)/GP > 0, G/GP > 0.51 NOT in European Jr	46
P/GP > 0.86, (+/-)/GP < -0.08 NOT in European Jr	44

Table 4.1: Forward decision rules for Tree A.1

Decision Rule	**Success Rate (%)**
P/GP > 1.37 in CHL	66.7
1.37 > P/GP > 0.91 P/GP and PIM/GP > 0.94 in CHL	55.6
P/GP > 0.66 in CHL	56
P/GP > 0.43 in NCAA	36
0.43 < P/GP < 0.66 and G/GP > 0.11 in CHL	34

Table 4.2: Defense decision rules for Tree A.2

Both of these decision rule sets demonstrate very well a few important uses for this type of research. Firstly, if a manager were to ask a question such as "What should I look for in a forward to ensure they will play for my team?" one could use Table 4.1 or even Tree A.1 to quickly respond "Big players in college hockey who can score 0.86 P/GP" or "I'd recommend anybody in the CHL that can score at least 1.2 P/GP." The second important take away somebody using this type of research could have is the immense importance of point scoring in a player's draft year as an indicator of playing in the NHL later in their career.

There is a clear importance to point scoring as a standard indicator of future success across all source leagues, though with greatly differing threshold values depending on the difficulty of the league. This leads to a question of whether similar results can be found if the data associated with point scoring and source league is combined into one statistic that takes the league difficulty into account. Of course this statistic would be NHLe as described in Section 1.3 as the product of point production with a league scoring difficulty modifier to give a single value that is comparable across players from all leagues. One potential benefit of this approach rather than using discrete league competition level groupings is that some groupings contain far more prospects than others, and in projecting player success the rare future star that is in a lower tier league might be seen as a potential failure based on their participation in a league which has produced worse players on average in the past. In contrast it does not make sense to reward players for having high point production in a league where scoring is far easier. So using a single value to consider both league level and production should allow useful information for players of all leagues while giving each prospect's production an even ground to be compared along. Trees A.3 and A.4 show how games played projection probabilities can be found using statistics where goal scoring and point production are replaced with their NHLe counterparts. As NHLe incorporates source league information in its definition, the league splits seen in Trees A.1 and A.2 were not allowed. Below are decision rules for recommendations that follow from the trees.

Decision Rule	**Success Rate (%)**
eP > 0.36, height > 71.5"	69
$0.27 <$ eP < 0.36, weight > 189.5lbs	44
eP > 0.36, height < 71.5"	43

Table 4.3: Forward decision rules for Tree A.3 using NHLe

Decision Rule	Success Rate (%)
eP > 0.17, weight > 189.0lbs	60
eP > 0.17, weight < 189.0lbs	34
0.11 > eP > 0.14	33

Table 4.4: Defense decision rules for Tree A.4 using NHLe

The next consideration is how well a scout's ratings over a player's skill set can predict if that draftee will have a future in the NHL. Due to the constraints of data set containing scouting information the sample size is smaller and even too small to produce useful insights for defensemen. Instead the focus will be strictly on forwards for Tree A.5, which uses a minimum leaf size of 10 samples to overcome the smaller data set.

Decision Rule	Success Rate (%)
Sense > 3.79 in CHL	71
Sense < 3.79, 3.43 < Shot < 3.45	58
Sense > 3.79, Shot < 3.82, not in CHL	50
Sense > 3.79, Shot > 3.82, Skating > 3.99, not in CHL	40

Table 4.5: Forward decision rules for Tree A.5 with scouting data

One note that is particularly interesting is the importance of the "hockey sense" skill as a predictor of success. Recall from the analysis as part of Section 2.1 that hockey sense is not generally weighted as important historically in drafts by nearly all teams. However according to Tree A.5 from the initial split of data players with a sense rating over 3.78 have a 43% chance to reach the NHL for a relevant period of time, whereas those who fail to reach the 3.78 threshold only play in the majors 14% of the time. In fact, if the players with a high "hockey IQ" happen to also be from the CHL leagues, then they have a 71% chance of playing in the NHL. Clearly there is an inefficiency in how teams are drafting in terms of that skill. This could be leveraged by teams with this knowledge. Another interesting piece of information to be learned from this tree is how a player with a good shot needs other skills such as skating or defensive play to complement shooting ability to be successful at higher levels. A way to think about why this might be is that if a player is a one-dimensional shooter it is possible they are scoring in ways that might work in lower competition, but as the level of opposing defensemen increases the shooter may find themselves unable to score in the same ways as in junior leagues. If such players are unable to overcome this fact with strong skating speed to get past the defensemen or contribute with responsible defensive play, having a strong shot in the juniors and getting away with not developing a multi-faceted game appears to be more of a curse than a blessing at times.

Looking at the n-fold cross-validation results for the decision trees in this section it is clear that the machine learning recommendations perform far better than the comparable average success rate present in the NHL, especially for forwards. By using the recommendations formed by data utilizing the NHLe statistic, one could expect nearly half of the recommended forwards to play in the NHL. The recommendations from the classifier using scouting data (Tree A.5) led to successful picks 1.95 times more often than the NHL average success rate for the same sample, and recommendations generated by Tree A.3 were successful 45.1% of the time. Using basic statistics was the best for defensemen, though there was not enough scouting data available to run a defenseman experiment.

	Forwards		**Defensemen**	
Input Type	**Tree Diagram**	**Estimated Accuracy (%)**	**Tree Diagram**	**Estimated Accuracy (%)**
Basic Statistics	A.1	36.7%	A.2	**28.5%**
Statistics with NHLe	A.3	**45.1%**	A.4	23.2%
Scouting Ratings	A.5	36.3%	—	—
NHL Average /w Stats	—	23.4%	—	20.9%
NHL Average /w NHLe	—	25.9%	—	21.9%
NHL Average /w Scouting	—	18.6%	—	—

Table 4.6: Summary statistics for baseline projection trees.

Past work by [17] has shown that the expected number of NHL games played for a player by their draft position drops rapidly and is near to zero after the first sixty to ninety picks. Given this knowledge it can be of value to see if there is any difference in players predicted to reach the 164 game threshold based on whether they are an early or late draft selection. For this experiment, separate decision trees were generated using players picked in the first 90 selections of their draft year as "early selections" and any picks after 91 as "late selections." As the league size and draft length have changed over the years this is a more consistent definition than splitting on the first three rounds as the number of picks per round was shorter in earlier years with fewer teams.

4.3 Forward Upside Projections

It might not be enough to know that a player will play games in the NHL to believe that they will be an impactful addition to a franchise. Many players are of a skill level that is easily replaceable through ways other than the draft, such as signing an older player at that skill level during the offseason. Consider an extreme example where a manager is picking between drafting player A and player B with the following projections. The manager is certain that player A will make the team. but will always be at a replaceable level. Player B has only a 10% chance of ever playing in the NHL, but will be a superstar if he reaches that potential. The manager might be tempted to go with the higher risk player against what Section 4.2 would recommend if they viewed replacement level skill as not much better than nothing.

This creates a need for a recommendation model that accounts for performance after reaching the NHL. It would not be practical to look for superstars as that would lead to far too few "success" annotations in the dataset, so instead the target will be players that perform in the top half of the league. For forwards as primarily offensive players, performance can be simply rated by points (P) scored per game played. With 30 teams each playing with 12 forwards, a successful forward should be in the top 180 players ranked by points (P) per game. From the statistics kept on NHL.com, the threshold to achieve a ranking in the top 50% of forwards is approximately 0.6 points (P) per game. Then for the decision trees in this section, the class will be 1 if a player played at least 164 NHL games and scored at a rate of at least 0.6 points (P) per game, and 0 otherwise.

As this problem suffers from a class imbalance where only 5.7% of classes are labeled "success" in the statistics data set and 3.2% for the data set with scouting information, the trees needed to train deeper to maximize the cross-validation projected success rate. Thus Trees A.6 & A.8 continue the node expansion process to a minimum size of 10 and 5 samples respectively. Both trees use a minimum predicted success rate of 20% to generate a recommendation in the evaluation process, contrasting the 40% value used in Section 4.2.

Comparing Trees A.1 & A.6 there are some common themes such as the importance of point production which would suggest that the difference between a forward who plays in the NHL and one who plays and scores many points is producing at a higher rate in junior leagues. The clearest example of this is the initial split of 0.86 P/GP in the baseline tree (A.1) and 1.15 P/GP in the tree, which also predicts major league point production (A.6). However one difference is in how the predictive behaviour of draft year goal scoring changes when looking at point production. For the experiments in Section 4.2, goal scoring was often not very important in predicting games played at the NHL level. Coupled with the findings from Tree A.6 that goal scoring is a positive predictor of forwards with high production ceilings, this suggests that these players are high risk in the sense that there are better things to look for to ensure a draftee reaches the NHL, but high reward players that can create a major impact if they are among the ones who make it. This type of player could be attractive to certain types of decision makers that are willing to make that gamble, but perhaps not worth the price required to more conservative managers.

Similar to the trees projecting games played from NHLe statistics in Section 4.2, Tree A.7 is dominated by the point production of each forward in predicting success. There is another common split into age brackets, often as older or younger than 17 or 18. Intuitively, scoring in junior leagues is easier when a player is older and more developed physically as a 17 year old is still growing into their frame. This is shown in the data as when an age split occurs the child node corresponding to a younger group of players always has a higher proportion of successful players than the child node with older players. This notion could lead to a further extension of the NHLe methodology to place players on common ground for their age.

There are a few interesting pieces of information that might be valuable to a decision maker in Tree A.8 where scouting ratings predict success or lack thereof for forwards to score impactful points for a team. The most obvious note is the importance of passing in determining the success. Having a passing rating of at least 4.5 (where 5 is a perfect score) increases the predicted probability of success from the baseline 3% up to a 50% success rate. Only 2.8% of prospects reach that high of a rating, but more evidence to support this fact is seen one step further down the tree on the opposite side. If a player has a passing rating below 4.5 they have a 2% chance of being an impactful point producing forward in the NHL. However for the 75.4% who do not reach 3.8 passing rating — which is still a strong score — there is a 0.48% chance that they will make that type of impact. There are two notable scouting dimensions not present as predictors in Tree A.8. First recall the importance of "hockey sense" as a predictor from Section 4.2, specifically Tree A.1. A possible analysis to be derived from this information is that the skill to generally be in the right place at the right time and understanding the way the game works can be useful both offensively and defensively in terms of knowing where an opposing player might want to pass to or shoot. Then a larger proportion of players that reach the NHL based on their "hockey IQ" skill could take on more defensive forward roles than those with exceptional skill in a more offensive dimension such as passing. The other notable absence is arguably the most purely-offensive skill of the 6 tools, shooting ability. Intuitively shooting ability should be a strong predictor of scoring a goal (G), which made up 37.4% of points scored in the 2016-2017 NHL season according to statistics tracked on Nhl.com. However, following the analysis from Section 4.2 with reference to shooting ability as a predictor of playing in the NHL, perhaps it is the case that what is seen by scouts as a good shot is a player taking advantage of some situations that might not work against more skilled and mature defensemen at the professional level. This line of reasoning would suggest that a strong passer and playmaker is more likely to see that success carry over to the next level of competition than a player that relies on their shooting ability to generate points. This would appear contradictory to the findings of Tree A.6 which shows that goal scoring in one's draft year is a predictor of future success along the same definition as success is measured in Tree A.8, as shooting ability should be tied to goal scoring. However it is not necessarily the case that goals are generated solely by shooting ability. Recall that those players who are deemed future successes for this section were also seen to have scored many points in their draft year as part of Tree A.6, which might point to them being good at more than just shooting. It is entirely possible that the goals for those future successful players were scored due in part to high skating or hockey sense scores, which would also be crucial in the high point production rates which incorporate high assist totals that those players attained in their draft years.

Input Type	**Tree Diagram**	**Estimated Accuracy**	**NHL Average**
Basic Statistics	A.6	17.7%	5.70%
NHLe	A.7	**25.2%**	6.82%
Scouting Data	A.8	9.52%	3.16%

Table 4.7: Summary of approaches to predict forwards scoring many points.

The best decision tree in for this section was the one using NHLe as input, though all outperform the average NHL success rates for their data sets by at least three times according to the n-fold cross validation results. Recall from Section 4.2 that with the forwards decision trees for predicting only games played, the NHLe tree had the largest success rate, but with scouting data input there was greater gains compared to the average NHL success rate in the data set. It makes sense that when considering success in terms of offensive production the most direct view of production across all league difficulties would be the best predictor of future success. The scouting data was of particular use in the games played success predictions from Section 4.2 as that definition of success includes forwards who contribute in ways which might not be seen on the score sheet every game. Identifying these types of players will be part of the next step in player projections in Chapter 5.

Chapter 5

Projecting Player Type

The previous chapters of this thesis have focused on determining if a player will reach certain milestones or levels of performance in their career. While there is clear value in this information and it may be exactly what some decision makers are interested in, it does not cover the fact that players have different play styles and what is seen as ”successful” could be very subjective to the case of each player. It is also important to consider that a team is unlikely to have 12 identical forwards and 6 identical defensemen, but groupings which are seen to compliment each other such as a strong passer with a goal scorer. This chapter will look at how to project what type of player each draftee will develop into at the professional level if they make it that far. Due to sample size constraints, only forwards using basic statistics as input were possible for this experiment, as opposed to the settings possible in Chapter 4.

5.1 Annotations for Player Style

The primary challenge with this task is deriving a set of annotations to be placed on established players to encapsulate their style of play. Those who follow the NHL closely would likely find it easy to describe how their favourite players of league stars play, such as Steven Stamkos as a ”sniper” (or goal scorer) or Joe Thornton the ”playmaker” (somebody who often uses passes to create scoring opportunities for teammates), but labelling each of the 600 skaters in the league each year over the period of a decade is unrealistic, especially when the majority of those players are not given much thought on a day-to-day basis. Grouping players into identifiable archetypes is an appealing route to creating annotations, but must be done in an automated fashion which still seems correct to an onlooker that is knowledgeable in the sport.

The natural approach to take is through a clustering algorithm. Players can be clustered in a space where dimensions are statistics known to be correlated with styles of play. For example, players with a better shot will generate goals at a higher rate, and players with an aggressive style will accumulate more hits. For forwards, a space with 5 dimensions <points per game (P/GP), goals per point (G/P), shots per game (S/GP), hits per game (H/GP), blocks per game (B/GP)> was chosen. Though many different clustering algorithms were tested, this analysis will focus on the k-means algorithm. Distances along each dimension are determined by standard deviations for that particular dimension to account for the fact that a difference in five goals between two players is far more relevant than five penalty minutes, which could be considered a difference of one fight.

Clustering figures for forwards are projected to points per game along the x-axis, goals per point along the y-axis, and hits per game determining the size of the points to give the illusion of depth where larger points indicate a player with more hits. Defensemen clusterings are projected into shots per game along the x-axis, blocks per game along the y-axis, and hits per game for the point size. Linkage-based algorithms were considered for the annotation task, however the data tended to form one large central cluster surrounded by small outlier clusters as seen in the average-linkage results shown in Figure 5.1. This lead to annotations that would not give much meaningful information, so only the k-means clusters were used.

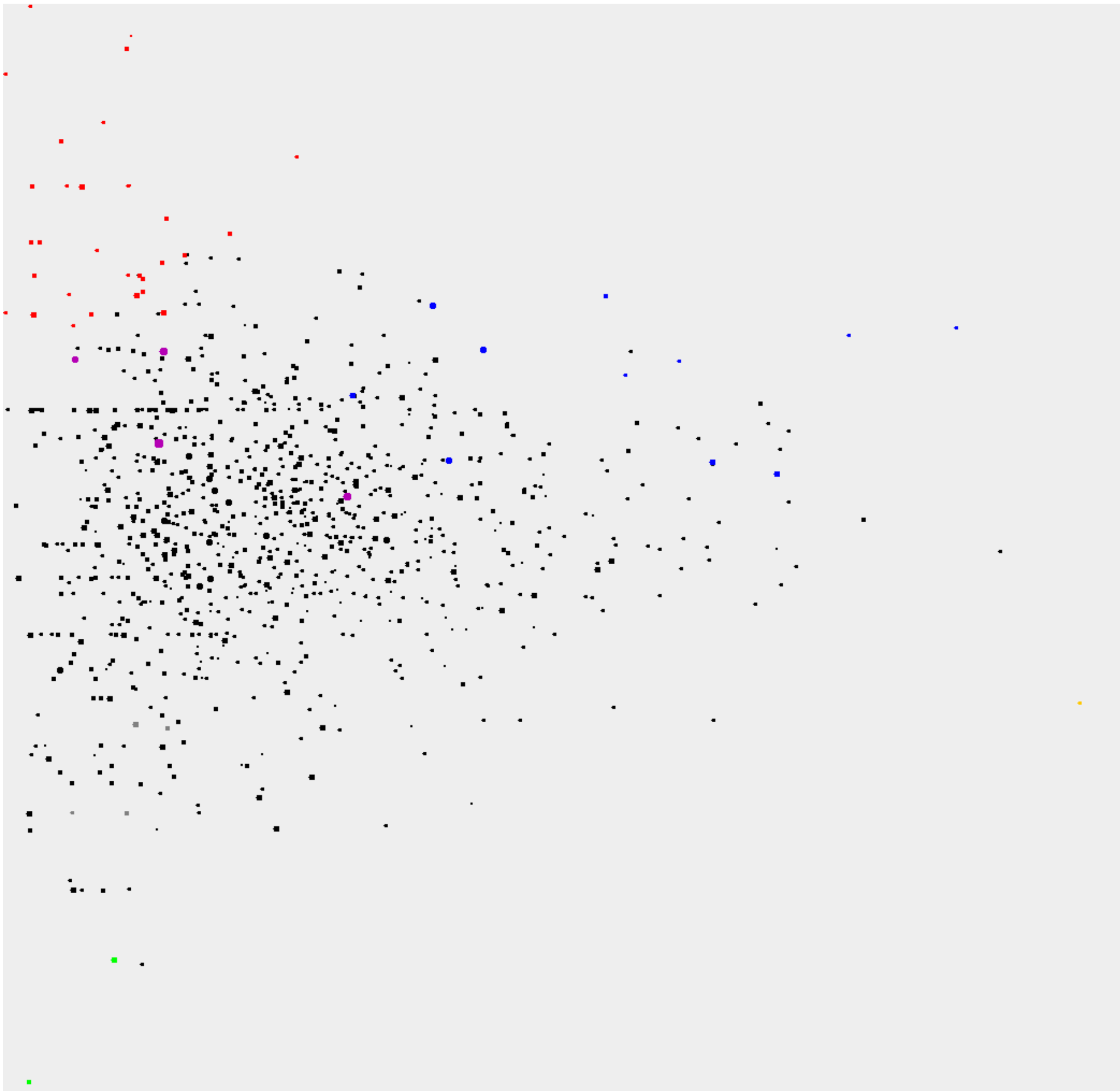


Figure 5.1: Resulting average-linkage clustering of forwards projected to a (P/GP, G/P, H/GP) space

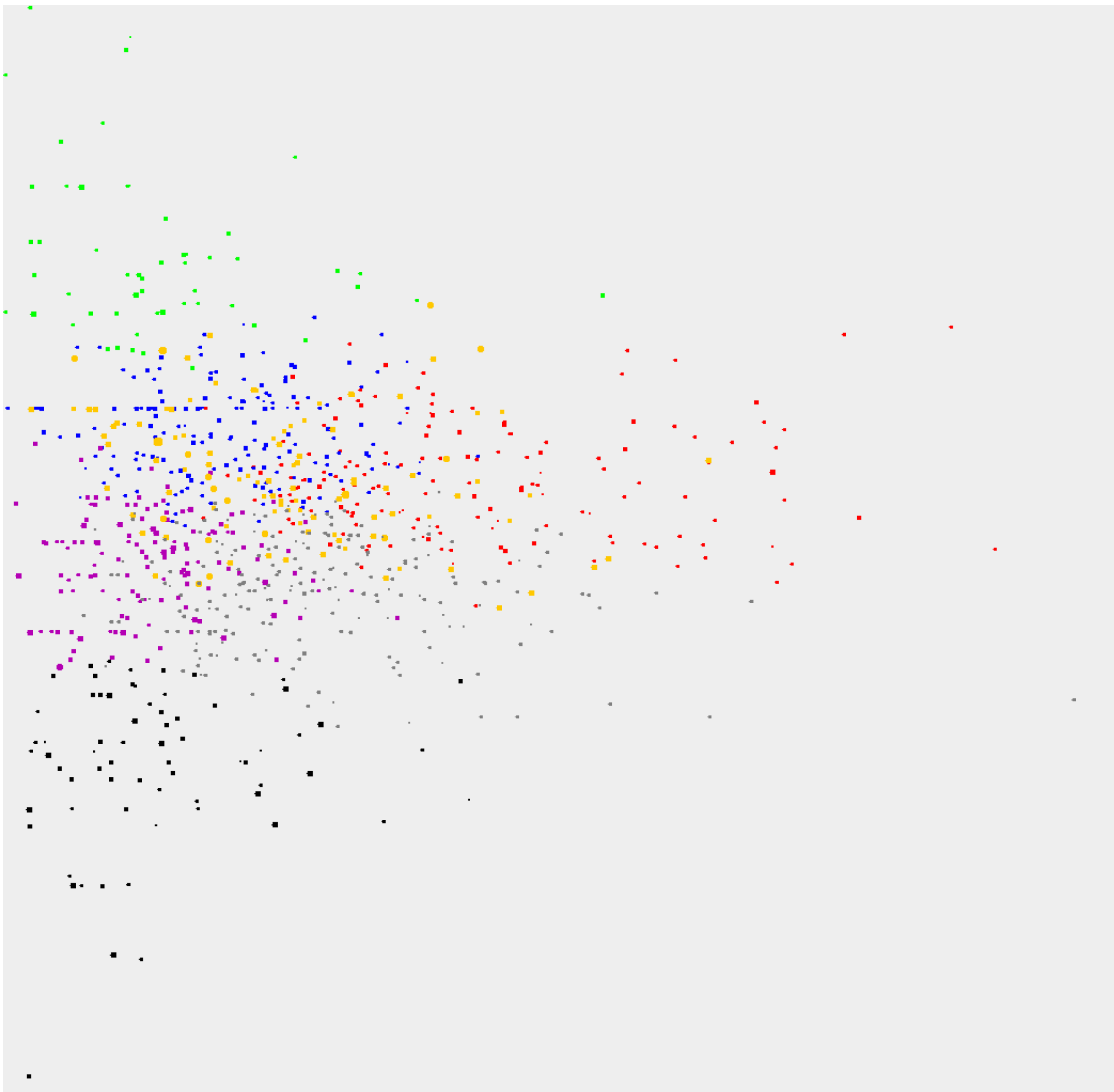


Figure 5.2: Resulting 7-means clustering of forwards projected to a (P/GP, G/P, H/GP) space

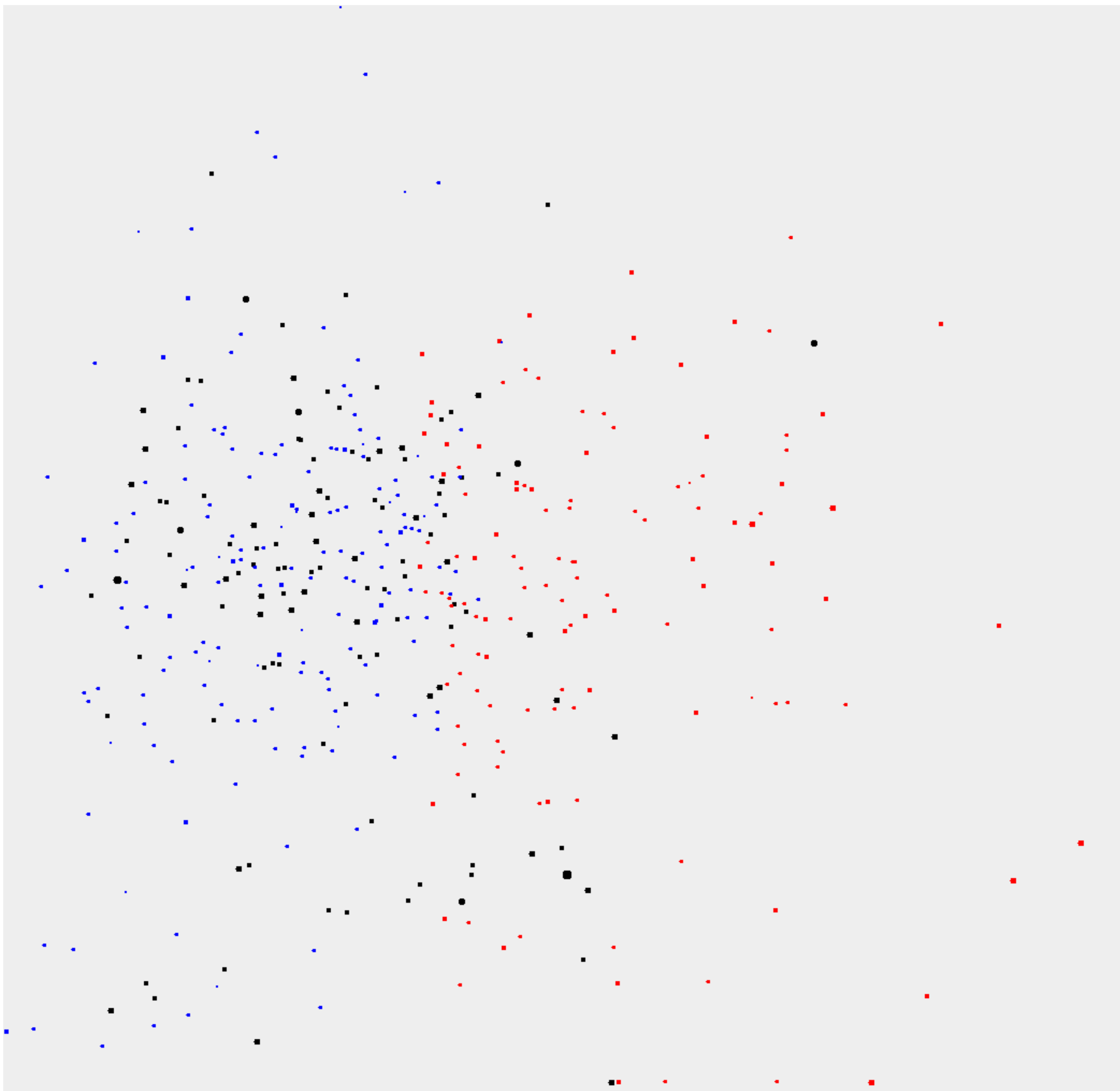


Figure 5.3: Resulting 3-means clustering of defensemen projected to a (S/GP, Blk/GP, H/GP) space

In the end it was found that the 7-means for forwards and 3-means for defensemen resulted in clusters that made the most sense and are listed below in Tables 5.1 & 5.2 along with two notable NHL players from each cluster.

#	Play Style	Fig 5.2 Colour	Notable Players
1	Playmaker	Black	Joe Thornton Henrik Sedin
2	Sniper	Red	Steven Stamkos Patrick Kane
3	Passing Power Forward	Blue	Ryan Johansen Sean Monahan
4	Two Way Forward	Green	Wendel Clark Trevor Linden
5	Grinder	Purple	Curtis Lazar Marin Hanzal
6	Shooting Power Forward	Orange	Alex Ovechkin Gabriel Landeskog
7	Playmaker 2	Grey	Ryan Nugent-Hopkins Nicklas Backstrom

Table 5.1: List of play style annotations for forwards.

#	Play Style	Fig 5.3 Colour	Notable Players
1	Heavy Hitters	Black	Erik Gudbranson Marc Methot
2	Offensive D	Red	Erik Karlsson Victor Hedman
3	Defensive D	Blue	Marc-Edouard Vlasic Jay Bouwmeester

Table 5.2: List of play style annotations for defensemen.

5.2 Style Projections

Using the annotations from the clustering process described above, the same process can be followed as the projections in Sections 4.2 & 4.3 except with either three or seven discrete categories instead of binary classification. The decision trees are built using the standard entropy formula, without any indication that certain classes are "closer" than others. One could certainly argue that errors with the two playmaker classes are smaller than "Grinder" and "Sniper", but for the following analysis — both for the entropy and error calculations — all classes are seen to be equally distant from each other. For error calculation, players are assigned the class that has the most instances in their assigned leaf.

Actual Class	**Predicted Class**							**Sum (Recall)**
	PLY	**SNP**	**PPF**	**TWF**	**GRN**	**SPF**	**PLY2**	
PLY	**0**	3	5	0	8	8	4	28 (0%)
SNP	0	**31**	20	0	7	2	22	82 (38%)
PPF	0	26	**32**	0	14	4	21	97 (33%)
TWF	0	10	4	**0**	4	9	3	30 (0%)
GRN	0	13	10	0	**35**	6	15	79 (44%)
SPF	0	14	19	0	11	**21**	6	71 (30%)
PLY2	0	18	31	0	7	2	**36**	94 (38%)
Sum (Precision)	0 (-)	115 (27%)	121 (26%)	0 (-)	86 (41%)	52 (40%)	107 (34%)	

Table 5.3: Confusion Matrix for Forward Play Style predictor

One clear issue that arises is the non-homogeneity in the leaves of the trees. Classes with smaller representation in the data set are never the majority class at a leaf and therefore are never the predicted class. This leads to a lower overall n-fold cross validation success rate of 32.2%. One obvious solution would be to expand the leaves further, but this would result in over-fitting. Instead, if the idea of a player's style is broadened to allow for mixtures of archetypes, it becomes clearer what the decision trees are demonstrating. For a leaf with 50% Snipers and 50% Power-Forwards it is possible that all players in that leaf are in fact somewhere along the border between the two clusters and were assigned their annotation by little more than luck. Investigating the confusion matrix closer supports this theory as instances of misclassification tend to exist between intuitively similar classes. For example 53% of misclassified players in "Playmaker 2" were predicted to be "Passing Power Forwards", as were 36% of misclassified points from the "Shooting Power Forward" class. Since the passing power forward class is intuitively very similar to both playmaking and the shooting power forward class, these are likely small errors involving points on the borders between clusters. With this perspective, two new approaches can be taken. One option would be to alter the annotations to describe players where they fit in the spectrum of play-styles by considering their relative distances to each cluster's centroid. A second approach would be to remove the idea of the play-styles all together and instead use a machine learning approach to group players in such a way that minimized average distance between each other in the clustering space. Then each leaf could be seen as a small cluster of players, but would also be reachable through their draft year data. This approach would not label future draftees as "snipers" or "playmakers", but would offer lists of players that have comparable careers to what can be expected of the given prospect.

Actual Class	**Predicted Class**			**Sum (Recall)**
	HH	**OFD**	**DFD**	
HH	**38**	30	9	77 (49%)
OFD	30	**34**	16	80 (43%)
DFD	26	15	**41**	82 (50%)
Sum (Precision)	94 (40%)	79 (43%)	66 (62%)	

Table 5.4: Confusion Matrix for Defenseman Play Style predictor

For the defensive projections, the actual size of each class in the data is relatively consistent. However the data available for prospect leagues is limited and it can be difficult to predict important stats such as shots, blocks, and hits at the NHL level without any of those data points from their pre-draft careers as input. As seen in the related decision tree diagram (Tree A.10), goals can be used as a strong predictor of shots as more goals would be proportional to more shots indicative of an offensive defenseman. Likewise, a larger number of penalty minutes can mean the player is making more hits leading to the prediction of the heavy hitter class. Even with these correlations, there is a limit to what the decision trees can do with the restricted inputs, so there is room to improve upon the 47.3% n-fold cross validation success rate.

5.3 Alternative Direction

While not studied as a part of this research, a viable alternative to the annotation process outlined in Section 5.1 would be to frame the problem as semi-supervised clustering. This can be done by leveraging domain knowledge to mark out particular players as prototypical of each play style. Then by using either constraint based techniques by grouping similarly annotated players together, or learning a distance function to minimize the distance between those players one could possibly find clusters which lead to stronger classifiers [4, 6, 5]. A further option would be to use domain knowledge to give feedback to the learner if clusters do not appear ideal, as done in [10]. One potential pitfall to avoid for this method is the temptation to select the prototypical players as only the most popular or "best" players of that archetype as this type of sample could make it difficult to cluster the far more numerous players of lower skill levels.

Chapter 6

Conclusions

The goal of this thesis was to explore several ways that methods in artificial intelligence can be leveraged to improve decision making in sports amateur drafts. The experiments focused on two broad areas of the draft. Chapters 2 and 3 were concerned with strategies for executing draft selections in a multiagent setting to maximize an individual or league-wide utility. Chapters 4 and 5 investigated a machine learning approach to determining a prospect's probability of success in the professional league and what style of play they would have upon reaching their prime.

The results of these experiments were very promising as they demonstrated in each field ways that artificial intelligence could out-perform the current processes teams adhere to in the draft. The ideas presented also open out new avenues to further improve the results and tackle related problems that were not covered. One such improvement would be to run strategic drafts in a world where all agents assumed a proportion of other agents were also taking a strategic drafting approach. By overcoming limitations to this research such as the availability of new forms of data like interview results, a new group of applications could be studied. One example would be the topic of the characteristics of a team captain. While this would be difficult to see in numbers such as how many goals the player scored, insights could potentially be found using Natural Language Processing techniques on certain interview responses.

With a few important changes, these ideas could even be applied to sports management concepts outside of the draft. Free agency is a concept present in many sports where teams may offer contracts to players not currently on a roster, similar to how companies might compete for the services of an unemployed skilled software developer. One could make minor changes to the utility model introduced in Section 2.1 to infer which teams would likely offer a given player a contract and even estimate how much the team would be willing to pay the player. Taking this idea one step further could lead to applications in contract negotiation models completely outside the scope of the sports industry.

One final thought is that none of the ideas, methods, or results of this work should be seen as something to replace the value and expertise of scouts and decision makers in teams. To the contrary, these ideas should be used in conjunction with existing practices to increase their effectiveness. Projecting a player's future using machine learning does not eliminate the value of a scout being able to identify a player's attributes, and in some settings the scout is a requirement. What these experiments do provide is context to the scout's information such as the relative importance of speed to grit on metrics like the probability of playing in the major leagues or the likelihood of the player being a goal-scorer. These approaches are intended to work hand-in-hand with human expertise to further our understanding of the sport.

References

[1] Aris Anagnostopoulos, Ruggiero Cavallo, Stefano Leonardi, and Maxim Sviridenko. Bidding strategies for fantasy-sports auctions. In *International Conference on Web and Internet Economics*, pages 102–115. Springer, 2016.

[2] Haris Aziz, Páter Biró, Jérôme Lang, Julien Lesca, and Jérôme Monnot. Optimal reallocation under additive and ordinal preferences. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 402–410. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[3] Joseph Baker and A Jane Logan. Developmental contexts and sporting success: Birthdate and birthplace effects in national hockey league draftees 2000-2005. *British Journal of Sports Medicine*, 2007.

[4] Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.

[5] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.

[6] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 11. ACM, 2004.

[7] Sylvain Bouveret and Jérôme Lang. Manipulating picking sequences. In *ECAI*, pages 141–146, 2014.

[8] Steven J Brams and Philip D Straffin. Prisoners' dilemma and professional sports drafts. *The American Mathematical Monthly*, 86(2):80–88, 1979.

[9] Eric Budish and Estelle Cantillon. The multi-unit assignment problem: Theory and evidence from course allocation at harvard. *The American economic review*, 102(5):2237–2271, 2012.

[10] David Cohn, Rich Caruana, and Andrew McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 4(1):17–32, 2003.

[11] Robert O Deaner, Aaron Lowen, and Stephen Cobley. Born at the wrong time: selection bias in the nhl draft. *PLoS One*, 8(2):e57753, 2013.

[12] Hadi Hosseini, Kate Larson, and Robin Cohen. Matching with dynamic ordinal preferences. In *AAAI*, pages 936–943, 2015.

[13] David A Kohler and R Chandrasekaran. A class of sequential games. *Operations Research*, 19(2):270–277, 1971.

[14] Namita Nandakumar. What does it mean to draft perfectly?

[15] Noam Peled, Sarit Kraus, et al. A study of computational and human strategies in revelation games. *Autonomous Agents and Multi-Agent Systems*, 29(1):73–97, 2015.

[16] Arvind Sathi et al. Constraint-directed negotiation of resource reallocations. 1989.

[17] Michael E Schuckers. What's an nhl draft pick worth? a value pick chart for the national hockey league. 2011.

[18] Hubert Song and Karl Pazdernik. Predicting nfl player success from analysts' diction, 2017.

[19] Philip D Straffin. *Game theory and strategy*, volume 36. MAA, 1993.

[20] Peter Tingling, Kamal Masri, and Matthew Martell. Does order matter? an empirical analysis of nhl draft decisions. 1:155–171, 07 2011.

[21] Joshua Weissbock. *Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data*. PhD thesis, Université d'Ottawa/University of Ottawa, 2014.

[22] Kasun Wickramaratna, Min Chen, Shu-Ching Chen, and Mei-Ling Shyu. Neural network based framework for goal event detection in soccer videos. 2005.

[23] Min Xu, Namunu C Maddage, Changsheng Xu, Mohan Kankanhalli, and Qi Tian. Creating audio keywords for event detection in soccer video. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–281. IEEE, 2003.

APPENDICES

Appendix A

Decision Trees for Player Projections

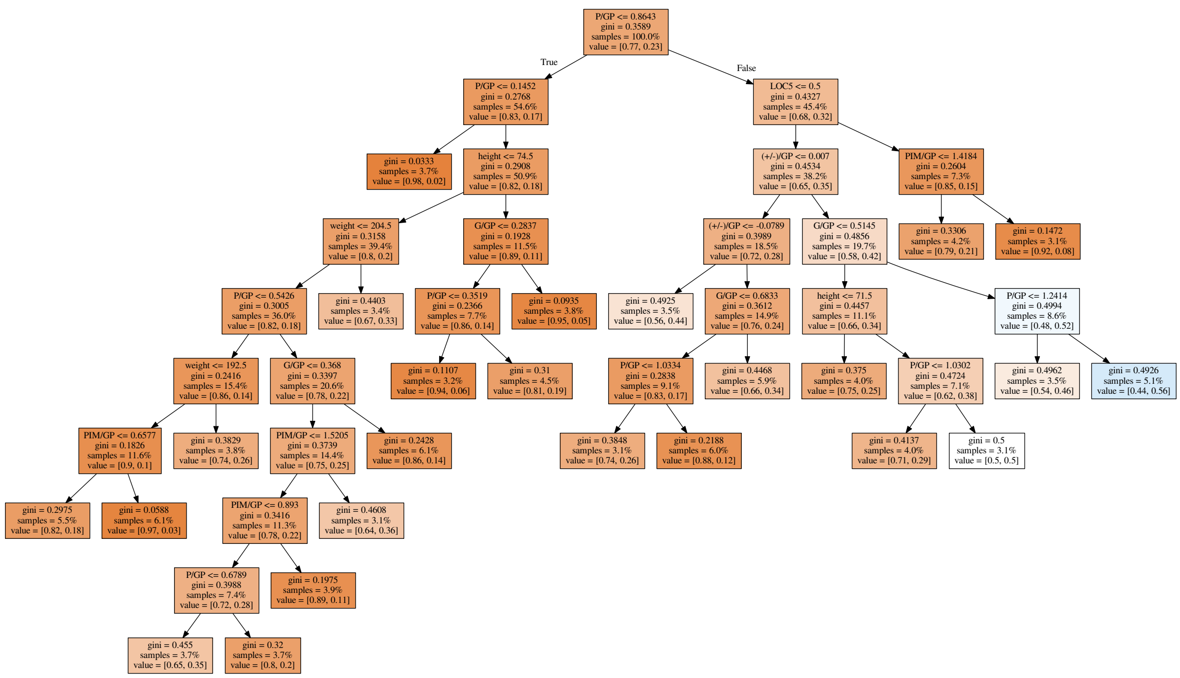


Figure A.1: Decision tree for forward baseline projections using basic statistics.

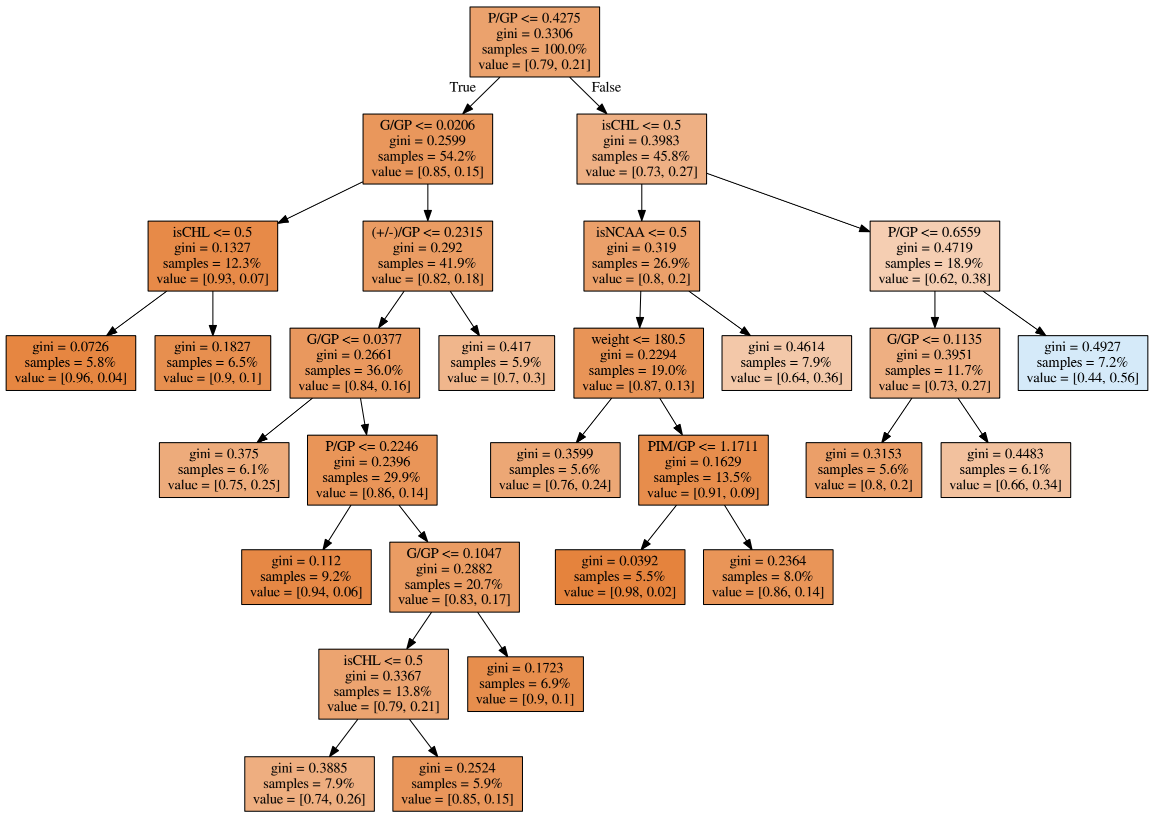


Figure A.2: Decision tree for defensemen baseline projections using basic statistics.

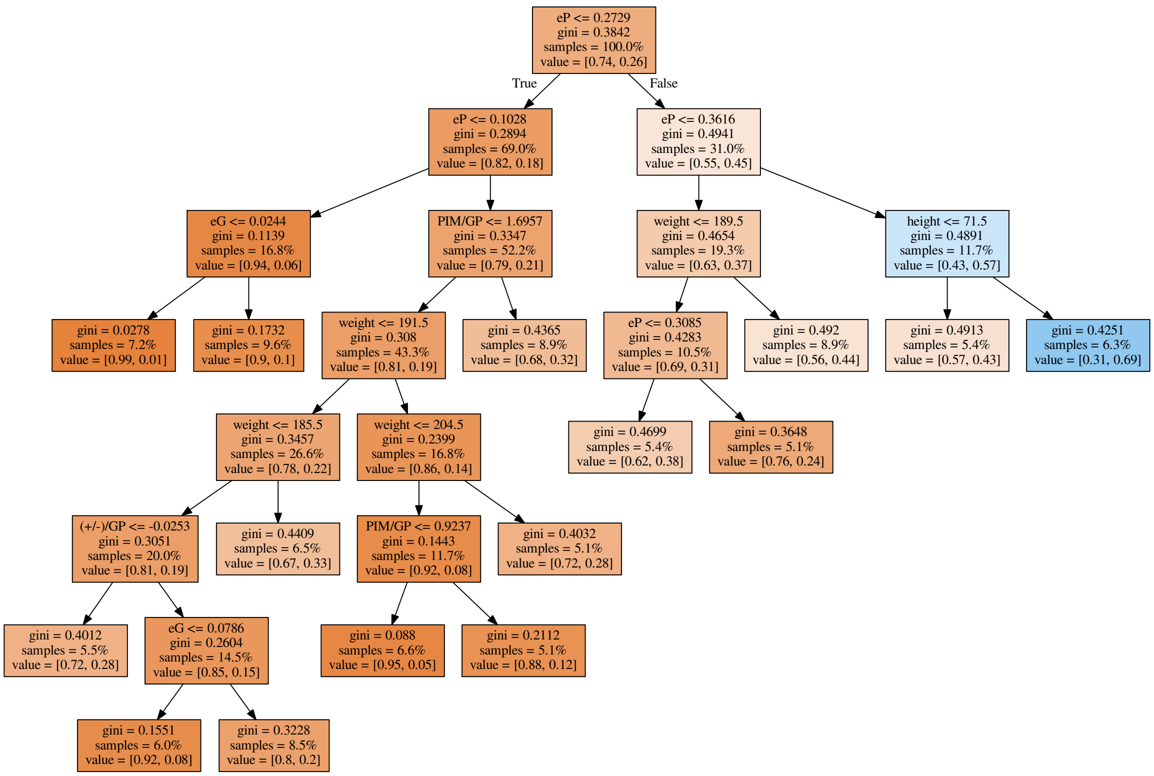


Figure A.3: Decision tree for forward baseline projections using basic statistics and NHLe.

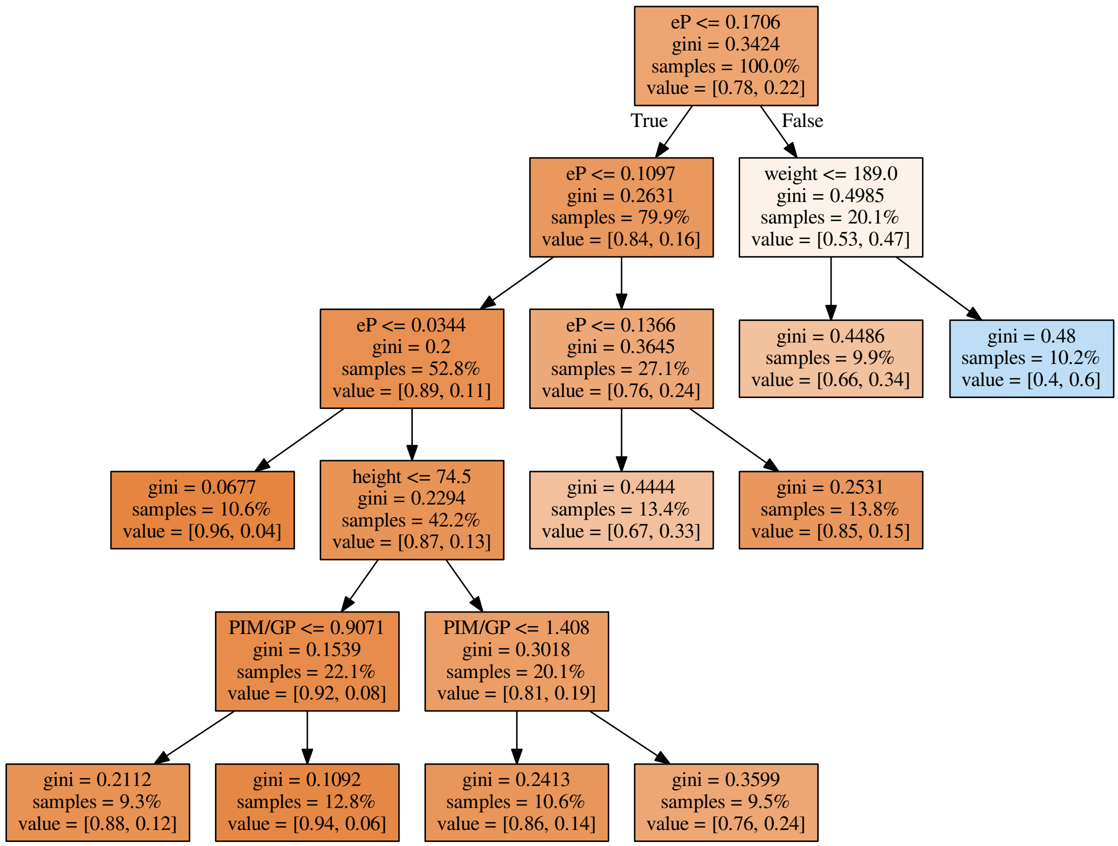


Figure A.4: Decision tree for defensemen baseline projections using basic statistics and NHLe.

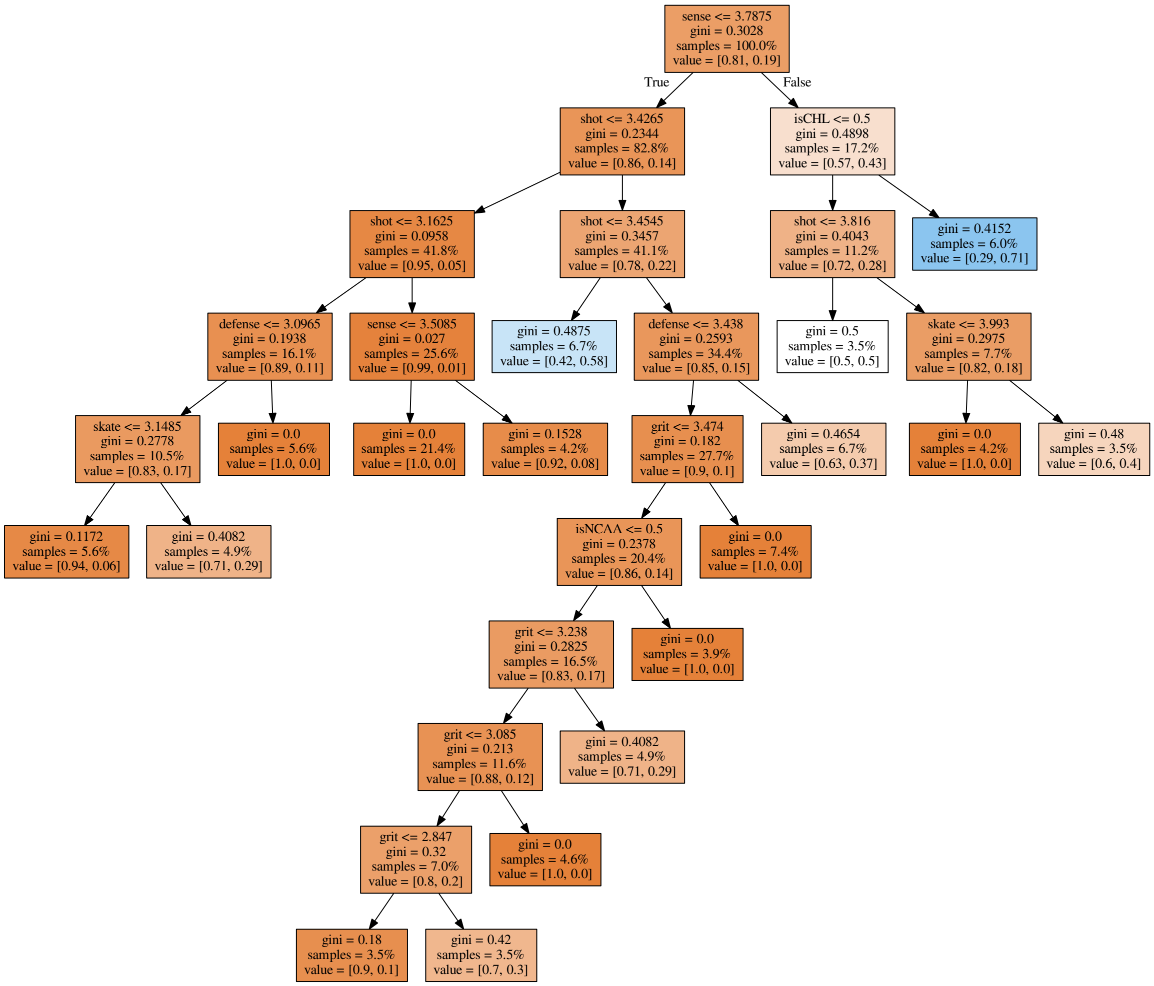


Figure A.5: Decision tree for forward baseline projections using scouting ratings.

A.1 Forward Upside Trees

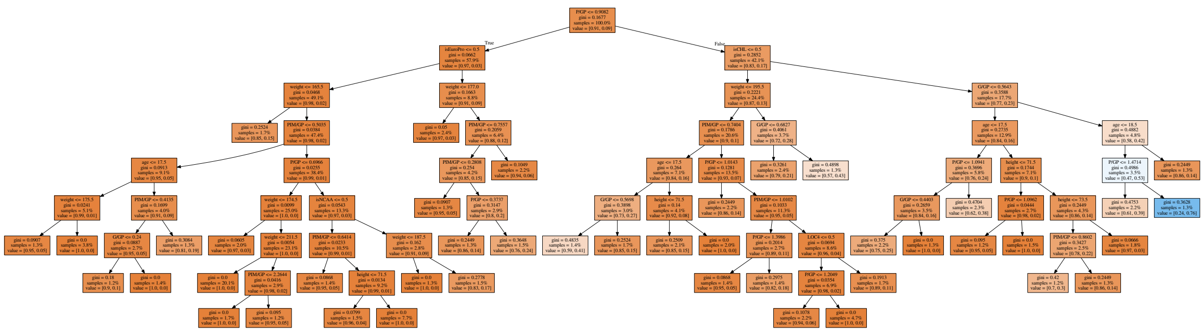


Figure A.6: Decision tree for forward point upside projections using basic statistics.

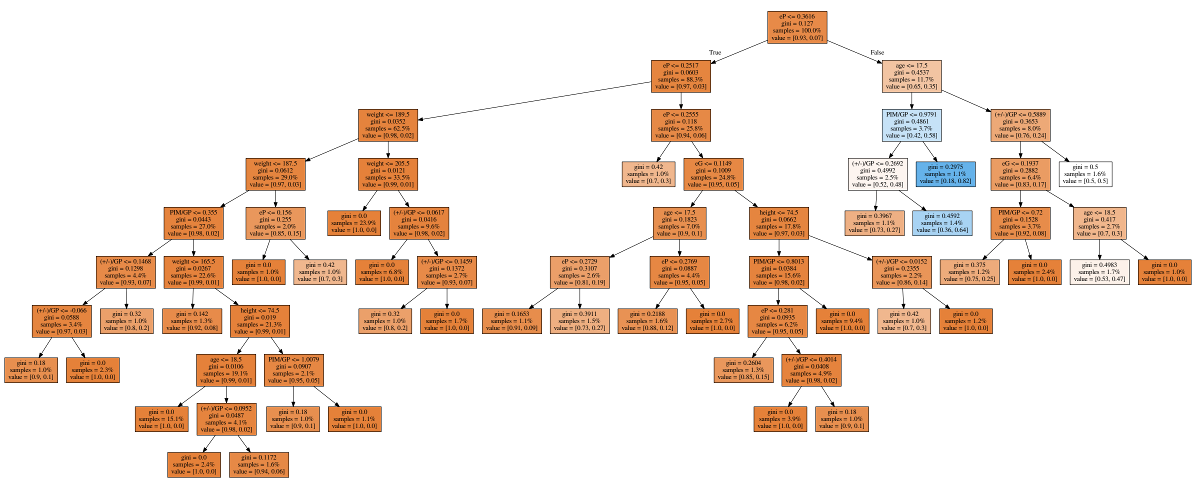


Figure A.7: Decision tree for forward point upside projections using basic statistics and NHLe.

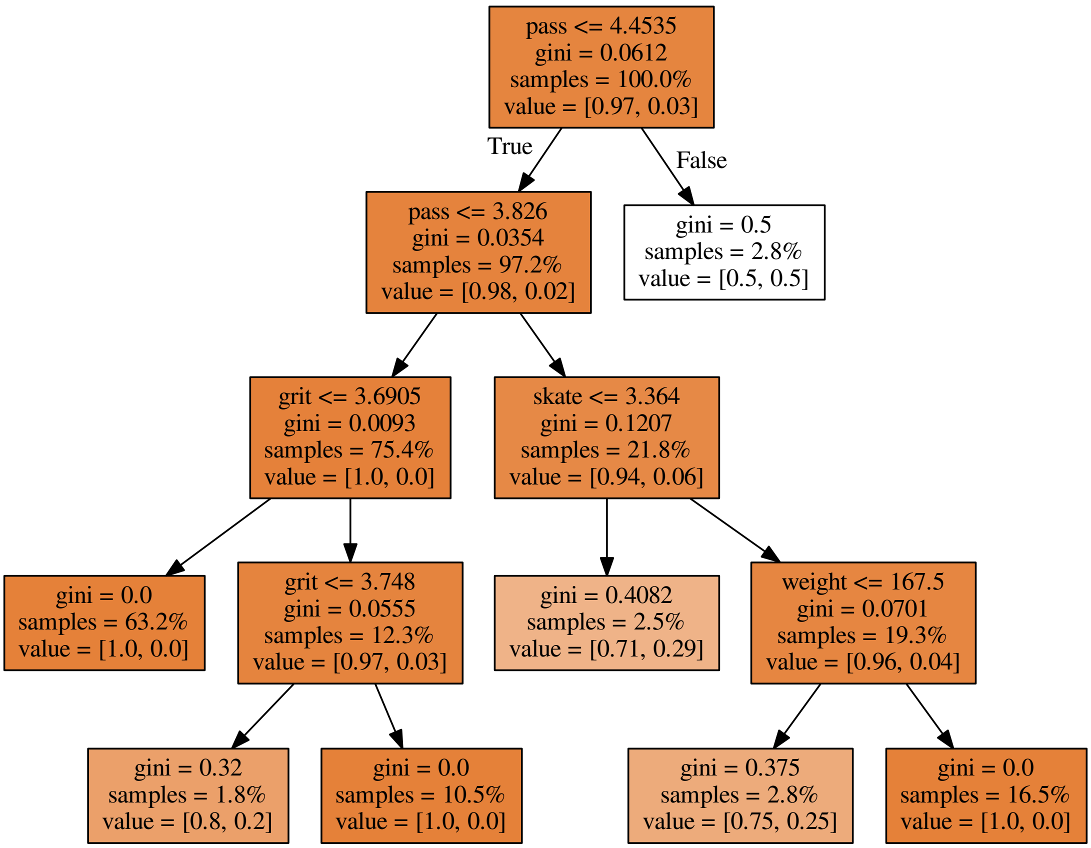


Figure A.8: Decision tree for forward point upside projections using scouting ratings.

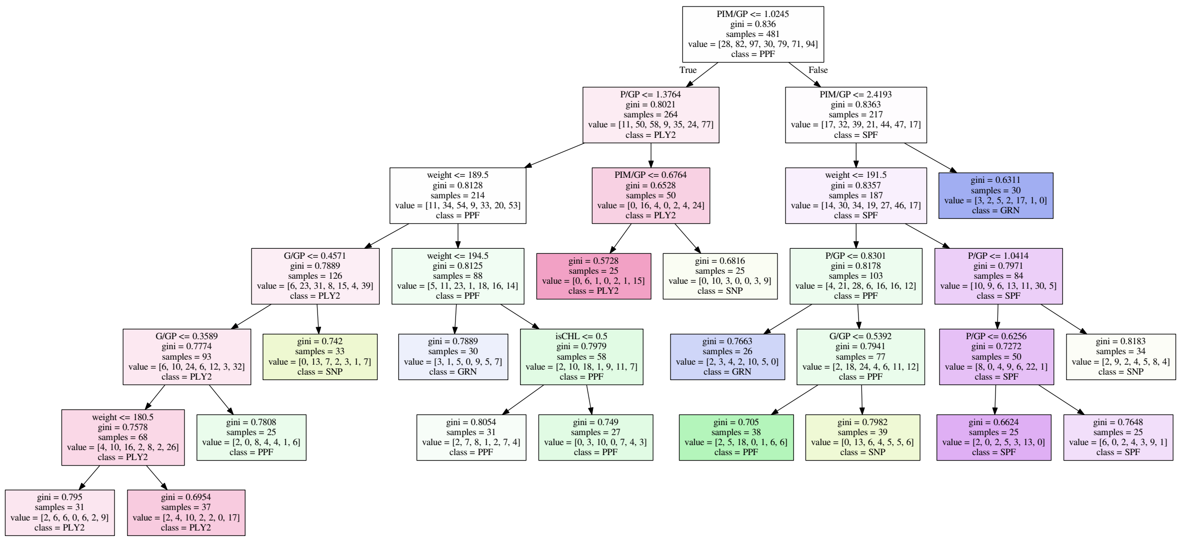


Figure A.9: Decision tree for forward play style projections using basic statistics.

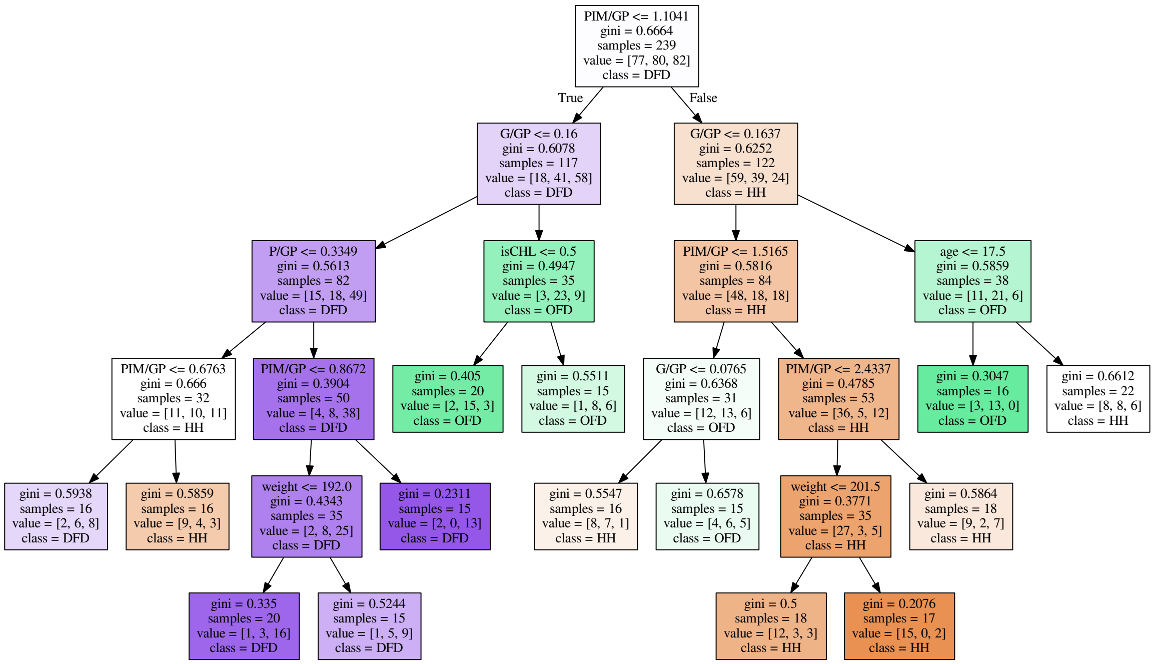


Figure A.10: Decision tree for defenseman play style projections using basic statistics.

Glossary

assist (A) Awarded to up to the last two players excluding the goal scorer to touch touch the puck prior to a goal. Assists are only given if the scoring team maintained possession of the puck between the assisting player's touch and the goal. 33

Corsi A ratio indicating the number of shot attempts for versus against while a player is on the ice. Shot attempts may be shots on goal, missed shots, or blocked shots. 5

goal (G) Awarded to the player who directed the puck into the opponent's net 33, 41

penalty minutes (PIM) Record of how many minutes a player spent penalized for various infractions such as tripping or fighting. 33

plus minus (+/-) When a goal is scored players on the ice for the scoring team are awarded +1, opposition players receive -1. Plus minus is not recorded for power play goals. 5, 33

points (P) The sum of a player's goals and assists. 33, 39

Abbreviations

ISS International Scouting Services 3, 9

NHL National Hockey League 2, 3, 5, 7, 8, 14, 17, 32, 33, 35–42, 47

NHLe NHL Equivalency viii, ix, 4, 36–38, 40, 41, 59, 60, 62