CARLETON UNIVERSITY

# Mining NHL Draft Data and A New Value Pick Chart

David R. Wilson

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of
**Master of Science**

in the
Faculty of Science
School of Mathematics and Statistics

May 2016

*"I think I'm a good Canadian, but I'm not the greatest Canadian."*

Don Cherry

Abstract

For every hockey player, getting drafted to the National Hockey League (NHL) is a dream come true, but the real goal is to reach 160 games played (GP) for pension reasons. First, a simple model was fit to NHL draft years 1998 to 2009 with an aim at predicting the proportion of players who will play a specified number of games in future years. For individual players drafted between 1998 and 2011, predictive models (Generalized Linear Model, Artificial Neural Network, Support Vector Machine and a LOESS) were created to predict a player's career GP. These models were combined to create a voting model to decide whether or not a player will reach 160 GP - a tool that can be used by any team, player or agent. Next, all players drafted in 1998 to 2008 were analysed using a non-linear multivariate model, with a modified weighted least squares, to predict a player's Time-On-Ice (TOI) for their first seven seasons. Position and Nationality dummy variables were used to distinguish between players. The seven seasons of TOI were then summed and smoothed using a LOESS. The graphs that resulted were analysed for positional advantage, then converted to a proportion and multiplied by a value of 1000 to create positional and Nationality value pick charts, which may be used by teams when choosing a draft pick or considering a trade, as well as player agents when negotiating player salaries.

*Acknowledgements*

I wish to thank my Carleton supervisor Dr. Shirley Mills for believing in me, her guidance, encouragement and for introducing me to statistics and analytics. I would also like to thank Dr. Michael Schuckers for inspiring me to follow this topic; and to both Shirley and Michael for their patience and assistance throughout this thesis journey. Also, thank-you to my classmate and friend Alex Diaz-Papkovich for his assistance, as well as introducing me to the hockey analytics community.

I would like to thank all my friends and classmates who have supported and encouraged me throughout my studies, particularly Melissa H and Alex Desforges. I'd especially like to thank the friends who offered and provided assistance throughout: Graeme K, Laura T, Matt S, and Patrick C.

A special thank-you to my parents for their never-ending support and encouragement. I am glad I make you proud. Thanks to my brother Adam for his support and the initial idea to start playing with draft data.

Most importantly, my wife and guiding light Angela, your support and encouragement topped them all. Whenever I doubted myself, you were there and when I struggled, you were there. On the dark days you gave me the strength and love that I needed to move forward. If it were not for your help and assistance with the data scrapping and programming I would still be trying to complete my thesis.

Finally, I wish to dedicate this thesis to all those who have a learning disability and/or suffer from mental illness.

# Contents


**Abstract** ii

**Acknowledgements** iii

**List of Figures** vi

**List of Tables** vii

**1 Introduction** 1
- 1.1 Introduction to Ice Hockey . . . 1
- 1.2 Literature review of hockey analytics . . . 3
- 1.3 Goal, Approach and Main Contributions . . . 8

**2 Data and Software** 10
- 2.1 Data Collection . . . 10
  - 2.1.1 Data problems and issues . . . 11
- 2.2 Data Summary . . . 12
- 2.3 Software used . . . 14

**3 Mathematical Techniques** 15
- 3.1 Career Games Played . . . 15
  - 3.1.1 Career Games Played by Draft . . . 15
  - 3.1.2 Career Games Played by Individual Player . . . 16
    - 3.1.2.1 Model Development: the Generalized Linear Model . . . 17
    - 3.1.2.2 Model Development: Artificial Neural Net . . . 18
    - 3.1.2.3 Model Development: the Support Vector Machine . . . 19
    - 3.1.2.4 Model Development: the Locally Weighted Regression . . . 22
    - 3.1.2.5 Model Development: the Voting Method . . . 22
- 3.2 Value Pick Chart . . . 23

**4 Applications** 26
- 4.1 Career Games Played . . . 26
  - 4.1.1 Career Games Played by Draft . . . 26
    - 4.1.1.1 Predicting One NHL Career Game Played . . . 26
    - 4.1.1.2 Predicting 160 NHL Career Games Played . . . 29
  - 4.1.2 Career Games Played by Individual Player . . . 31
    - 4.1.2.1 Results of the Generalized Linear Model . . . 32

4.1.2.2 Results of the Artificial Neural Net . . . . . . . . . . . . 32
4.1.2.3 Results of the Support Vector Machine . . . . . . . . . . 33
4.1.2.4 Results of the Locally Weighted Regression . . . . . . . . 34
4.1.2.5 Results of the Voting Model . . . . . . . . . . . . . . . . 35
4.2 Value Pick Chart . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
4.2.1 Value Pick Chart Creation and its Uses . . . . . . . . . . . . . . . . 44

**5 Conclusion and Future Work** **47**
5.1 Career Games Played . . . . . . . . . . . . . . . . . . . . . . . . . 47
5.2 Value Pick Chart . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
5.3 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

**References** **50**

**Bibliography** **53**

**Appendices** **54**

**A** **54**

# List of Figures


2.1 Early Visualization of TOI vs. Draft Age by Nationality grouping . . . . 13

3.1 ANN with one *Hidden Layer* consisting of 3 nodes. . . . . . . . . . . . . . 20

4.1 One NHL Game Played Proportion . . . . . . . . . . . . . . . . . . . . . 27
4.2 Fitted One NHL Game Played Proportion . . . . . . . . . . . . . . . . . 28
4.3 Residual Plot for One NHL Game Played Proportion . . . . . . . . . . . 29
4.4 160+ Game Played Proportion . . . . . . . . . . . . . . . . . . . . . . . 30
4.5 Fitted 160+ Game Played Proportion . . . . . . . . . . . . . . . . . . . 31
4.6 Residual Plot for 160 Game Played Proportion . . . . . . . . . . . . . . 31
4.7 Confusion Matrix on the training set for GLM & STEP Models . . . . . . 32
4.8 Training Set Confusion Matrix for the ANN (no *Hidden Layer*) . . . . . . 33
4.9 Test Set Confusion Matrix for ANN (no *Hidden Layer*) . . . . . . . . . . 33
4.10 Training Set Confusion Matrix for the SVM . . . . . . . . . . . . . . . . 33
4.11 Test Set Confusion Matrix for the SVM . . . . . . . . . . . . . . . . . . 34
4.12 Training Set Confusion Matrix for the LOESS . . . . . . . . . . . . . . . 34
4.13 Test Set Confusion Matrix for the LOESS . . . . . . . . . . . . . . . . . 35
4.14 Confusion Matrix for the Voting Model . . . . . . . . . . . . . . . . . . 35
4.15 Original data for all players and coloured by their first seven seasons. . . . 36
4.16 Original data, sum of first seven seasons for all players. . . . . . . . . . . . 37
4.17 Original data, smoothed. . . . . . . . . . . . . . . . . . . . . . . . . . . 37
4.18 Zoomed image of Figure 4.17. . . . . . . . . . . . . . . . . . . . . . . . 38
4.19 Monoreg plot of the Original data. . . . . . . . . . . . . . . . . . . . . . 38
4.20 Predicted TOI for a players first seven seasons in the NHL. . . . . . . . . 39
4.21 TOI for first seven seasons by Position. . . . . . . . . . . . . . . . . . . . 39
4.22 TOI for first seven seasons by Nationality. . . . . . . . . . . . . . . . . . 40
4.23 TOI for first seven seasons by Position for North American & non-North American Players. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
4.24 TOI sum of seven seasons by Position. . . . . . . . . . . . . . . . . . . . 42
4.25 All positions combined. . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
4.26 Smoothing of Figures 4.24. . . . . . . . . . . . . . . . . . . . . . . . . . 43
4.27 Smoothing of all positions combined. . . . . . . . . . . . . . . . . . . . . 43
4.28 Overlay of all positions during the draft. . . . . . . . . . . . . . . . . . . 44
4.29 Magnified image of Figures 4.28 & 4.27 between picks 150 and 200. . . . . 44
4.30 Nationality Comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . 45

# List of Tables


1.1 Draft Lottery Percentages . . . . . . . . . . . . . . . . . . . . . . . . . 3

2.1 Re-drafted Players . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

3.1 Variables for Player $i$ . . . . . . . . . . . . . . . . . . . . . . . . . . 17
3.2 Variables for Player $i$, Season $\ell$ . . . . . . . . . . . . . . . . . . . . . 23

4.1 One NHL Game Played Proportion . . . . . . . . . . . . . . . . . . . . . 27
4.2 160+ Games Played Proportion . . . . . . . . . . . . . . . . . . . . . . 30
4.3 2010 and 2011 160+ GP Prediction Based on 99 & 95 % Confidence Interval 30

A.1 **Value Pick Chart** . . . . . . . . . . . . . . . . . . . . . . . . . . . 54
A.1 **Value Pick Chart (Continued)** . . . . . . . . . . . . . . . . . . . . . 55
A.1 **Value Pick Chart (Continued)** . . . . . . . . . . . . . . . . . . . . . 56
A.1 **Value Pick Chart (Continued)** . . . . . . . . . . . . . . . . . . . . . 57
A.1 **Value Pick Chart (Continued)** . . . . . . . . . . . . . . . . . . . . . 58

# Chapter 1

# Introduction

Every athlete dreams of one day playing professionally, and for many that dream becomes closer to fruition the moment their name is called at a player entry draft. In the sport of hockey, National Hockey League(NHL) teams make over 200 dreams come true annually by selecting players from around the world. The NHL currently has 30 teams and each team owns one selection(or pick) for each round of the draft unless they trade away their pick. There are currently 7 rounds in the annual draft.

## 1.1 Introduction to Ice Hockey

Ice Hockey is a game played between two teams with sticks and a rubber puck on a skating rink.[1] The object of the game is to score the rubber puck into the opposing team's net, which is being defended by one of the teams' goaltenders. The team that scores the most goals wins the game.

[1] Although it is a sport played in the winter, its origins can be traced back 2500 years ago to Althenians who etched markings of the sport on the building of the Parthenon. [1]

In the NHL, 30 teams are divided into two conferences; 14 teams play in the Western conference and 16 play in the Eastern conference. These conferences are further broken down into divisions, namely the Pacific, Central, Atlantic and Metropolitan divisions. Each team may dress 20 players for a game, comprising of 18 skaters and two goaltenders. Most teams will break up their 18 skaters into four forward lines and three defensive pairings. A forward line consists of one centre and two wingers. A centre's job is to take face-offs and to assist the defense and forward wingers. The centre position requires slightly more skills such as skating and agility; therefore, they shall be distinguished from the other forwards.

In the NHL, teams play 82 games a season, mostly within their own conference. Each game consists of three 20 minute timed periods called regulation time. In the event of a tie at the end of regulation time, both teams now play a five minute three-on-three sudden death overtime period[2]. If teams are still tied at the end of overtime, the game is decided by a shootout. Each team elects three shooters to face off against the opposing team's goaltender. The team with the most goals wins. If it is still tied, the shootout continues one shooter at a time until there is a winner. The winning team is awarded two points in the standings; the losing team is awarded one point if they lost in overtime or a shootout and zero points if they lost during regulation time.[2]

At the end of the regular season for all 30 teams, the top three teams in each division automatically advance to the playoffs. The next two teams with the highest points in each conference are selected as wild card teams and will also compete in the playoffs. The remaining 14 teams are entered into a lottery, called the NHL Draft Lottery, to determine the first 14 picks of the draft. The team with the fewest points during the

[2]Prior to 2005, overtime was five-on-five, but after the lockout the NHL made a number of changes, including four-on-four overtime. Starting in 2015-2016 season, in order to decrease the number of games decided by a shootout, the league switched to three-on-three overtime.

season has the highest probability of winning the coveted first overall choice in the draft. Table 1.1 provides the probability of having the first overall selection by regular season ranking of points[3]:

Table 1.1: Draft Lottery Percentages

| Rank at End of Regular Season | Percentage of Selection |
|---|---|
| 30th | 20.0% |
| 29th | 13.5% |
| 28th | 11.5% |
| 27th | 9.50% |
| 26th | 8.50% |
| 25th | 7.50% |
| 24th | 6.50% |
| 23rd | 6.00% |
| 22nd | 5.00% |
| 21st | 3.50% |
| 20th | 3.00% |
| 19th | 2.50% |
| 18th | 2.00% |
| 17th | 1.00% |

The remaining 13 teams will be slotted in reverse order based on their regular season point totals.[3] As for the playoff teams, their draft spot is determined by reverse order of elimination from the playoffs. For teams that are eliminated in the same round, the team with the lower regular season point total will choose first.

## 1.2 Literature review of hockey analytics

In professional leagues, teams are creating new positions for front office staff. Qualified personnel must be capable of performing analytics (advanced statistics) on player and team data. Unlike other sports such as baseball, basketball, and football, hockey is relatively new to the game of analytics. Academics, fans, math 'geeks' and pure hobbiests are jumping on the hockey analytics bandwagon and are studying the game in

[3]For the 2016 NHL Draft Lottery, the top three draft picks will be determined by the lottery, and the remaining eleven teams will be slotted in reverse order based on their regular season point totals.

new ways. Many of these researchers rely on the NHL.com website for their data. Data such as Drafts, Play-By-Play(PBP), Career Statistics and Time-On-Ice (TOI) are often analysed and when the data is not available or is inaccurate at NHL.com, the researchers must seek out third party sites such as Hockey-Reference.com and HockeyDB.com. Unless otherwise stated, assume that any work referenced herein comes from the above mentioned sites, but primarily NHL.com.

NHL teams looking for a competitive advantage are hiring these researchers before their rivals.[4] In the old days before analytics, scouts, team personnel, and coaches would match lines or players against the opposition based on human observation. However, today's game is much quicker and it is becoming more difficult to rely on human observation alone. NHL players today can reach speeds between 25 and 30km/h and shoot the puck in excess of 165km/h. In addition, the NHL keeps track of a lot more metrics today than it did before, such as TOI, Faceoff Wins & Losses, plus-minus, blocked shots, hits and more[4]. Relying on human observation alone is no longer an option in the NHL; advanced statistics are a must if a team wishes to be competitive. The demand for statisticians will only increase as new technologies (such as chip technology) are introduced into the game by companies like SportVision, HockeyTech, SPORTLOGIQ, and PowerScout Hockey.

Since hockey began, discrete metrics such as goals, assists and points have been the main focus of traditional statistics.[5] Today, we hear terms like Corsi, Fenwick, and WOWY. The Hockey Prospectus[6] defines Corsi as a plus-minus statistic that measures shot attempts; Fenwick is the same, but does not include blocked shots. WOWY (With or Without You) is the analysis used to determine which players benefit the most from their linemates and which players are driving the play. Schuckers and Curro[7] created

[4] Although they are available today, many of these are new: TOI was not available until 1998 and Faceoff Wins & Losses, Blocked Shots, Hits were not available until 2007.

a new term in hockey analytics called the Total Hockey Rating (THoR), which rates players based on their on-ice events such as zone starts, non-shooting events, and hits that occur when a player is on the ice. Using probability, they determined the impact an individual player has on their team by their on-ice events leading to a goal for or against.

Diaz-Papkovich[5] utilized regression to look at Corsi, Fenwick and shot differential to examine their value on a team's win and goal percentage. In addition, Diaz-Papkovich created a binary transaction matrix of players and Corsi events and used association rule learning to measure player chemistry and performance, a sort of WOWY approach. Using the Toronto Maple Leafs top line wingers, Diaz-Papkovich paired different centres with Kessel and van Riemsdyk to demonstrate their overall Lift to the line. As is demonstrated throughout, Diaz-Papkovich's approach to player combinations would make an excellent tool for coaching staff at any hockey level.

Creating balanced line combinations is a challenge for team personnel, so how are NHL teams made up? What combination of players is required to have a balanced team? Vincent and Eastman[8] perform k-means clustering to categorize NHL players based on player weight, points per game, penalty minutes, and their plus-minus. Through mathematical justification, they determine that NHL forwards are categorized into three groups "scorers", "grinders" and "enforcers" whereas defensemen are categorized into two groups: "scorers" and "aggressors". A combination of these categories is required for any team to have on-ice success. However, as the study was performed with pre-lockout data, a follow-up study should be conducted in order to determine how modern players are categorized, especially given the fact that now NHL teams rely less on "enforcers".

Despite a team having the right line combinations, teams may find themselves losing the game with little time remaining. Beaudoin and Swartz[9] developed a simulator for NHL games and looked at different strategies for pulling a goaltender, giving a team an extra attacker. Parameter estimates were obtained in their various simulations through analysis using constrained Bayesian estimation by Markov Chain methods. In 5 vs. 5 situations, current NHL coaches typically pull their goaltender when trailing by one goal with 1 minute remaining on average and 1:30 remaining when trailing by two goals. However, Beaudoin and Swartz suggest that teams should be aggressive and pull their goalie with 3 minutes remaining. (NOTE: this strategy changes slightly if there is a penalty). In addition, they suggest teams should be extremely aggressive when trailing by 2 goals and pull their goalie with 6 minutes remaining, unless they receive a penalty, in which case they should wait until after the penalty is over before pulling their goaltender. Their findings suggest that the current strategy adopted by NHL coaches of pulling a goaltender when trailing by 2 goals with 1:30 remaining is "a lost cause".

However, not all games can be decided by the players and coaches of the team. Referees can drastically change the course of a game with one call. Beaudoin, Schulte & Swartz[10] investigated penalty calls in the NHL for the 2009/2010 through 2013/2014 regular seasons. Using a logistic regression model, where $y_i = 1(0)$ according to whether the $i$th penalty was called against the home (road) team, $i = 1, ..., n$ and is distributed according to $y_i \sim Bernoulli(p_i)$, they considered four covariates and using the Akaike Information Criterion (AIC), they observed that using all four covariates gave the best fitting model. For comparison, they considered a gradient boosting algorithm, called the *gbm* function in R. They concluded that "teams that have taken more penalties in a match are less likely to have the next penalty called against them and teams that are leading in a match are more likely to have the next penalty called against them". However, they did caution the reader to consider causal relationships in more detail before accepting their findings as fact.

It has been widely studied and proven that home teams have an advantage over their opponents. The works of Pollard[11] in 2005, and more recently, Doyle and Leard[12] in 2012 confirm that home teams consistently win more games. Trandel and Maxcy[13] derived a balanced league standard deviation formula of winning percentages that takes into account the home advantage. They used this new formula to recompute the standard deviation ratios for major sports leagues and they considered the competitive balance in the various leagues. Home ice advantage can also affect events and recorded statistics in the NHL. Schuckers and Macdonald[14] looked at inconsistencies in the recording of events and statistics at rinks in the National Hockey League, known as the Real Time Scoring System (RTSS). They showed how some rinks under or over estimate events such as shots, blocks, hits, etc. In order to create an equal balance across NHL rinks, they proposed a log-linear model to re-weight the recorded events.

Schuckers[15] created an alternative to the National Football League Draft Pick Value Chart based upon player performance. Schuckers looked at the first 255 draft selections from the years 1991 through 2001 and based his performance analysis on non-position dependent metrics. Using a nonparametric regression (LOESS), the various metrics (Games Played, Career Approximate Value, Game Starts, and Pro Bowls) were plotted with draft selection as the independent variable. Not surprisingly, the fitted line was monotonically decreasing (as one would expect) as Draft Selection increases. Each of the four metrics were normalized to reflect the total value currently used by the NFL. After comparison, Game Starts was selected to create the alternative Draft Pick Value Chart, as it was a better overall metric than Games Played and easier to calculate and predict than Career Approximate Value.

Similarly, Schuckers[16] followed up on his NFL Draft Pick Value Chart with a hockey version. Using the NHL Drafts from 1988 through to 1997, he analysed career performance metrics: games played, career goals, assists, points and plus-minus. Schuckers used LOESS regression on the probability that a given draft selection will play at least 200 NHL games, splined twice for the first 45 selections and the remaining to ensure monotonicity. Using a similar approach, Schuckers fitted a splined LOESS function to regress Draft Selection on mean Games Played and scaled this model to create the Pick Value Chart.

Schuckers and Argeris[17] continued to build on Schuckers' previous research by estimating the average yearly financial gain on teams' internal scouting. Using the NHL Drafts from 1998 through to 2002, they analysed career performance metrics: cumulative Games Played, cumulative TOI and Goals Versus Threshold[5] across a player's first seven years post draft. Schuckers and Argeris primarily focused on how often the Central Scouting Service (CSS) optimally or nearly optimally selected the best player available. They determined that teams, on average, outperformed the CSS and their financial gain from scouting was between \$1.8$MM$ $and$ \$5.2MM per year. As a bi-product of their work, Argeris and Schuckers created a new Draft Value Pick Chart, based on TOI.

## 1.3 Goal, Approach and Main Contributions

The goals and approaches of the thesis are:

1. To fit a model and predict the proportion of players who will play 1 (or 160) career Games Played (GP) in future draft years.

[5] A player metric created by Tom Awad.

2. Build models and combine them into a voting model, to predict if an individual player will play 160 career GP.

3. Predict the first seven seasons of TOI using a multivariate model, and then used it to create Value Pick Charts (VPCs) broken down by position and Nationality.

The first model (predicting proportion of players who will play 1(or 160) career GP) can be used by teams, the media, and fans to determine whether or not a future draft year will be talent rich or not. The voting model can be used by any team, player or agent to determine if a player will reach the 160 career GP milestone and earn a valuable pension. Finally, the VPCs are used by teams when choosing a draft pick or considering a trade, they can also be used by player agents when negotiating player salaries.

# Chapter 2

# Data and Software

## 2.1 Data Collection

The primary sources of data for this thesis are Hockey-Reference.com and NHL.com. The drafts are broken down by year, rounds (1 through 7)[1], and overall pick. The analysis presented in this thesis will use draft data and career statistics for players selected from 1998 through to 2013, although most of the analysis will be for the years 1998 to 2011.

Each player that plays an NHL game has their statistics recorded on their individual player bio. The NHL did not start keeping track of TOI data until the 1998 season, which is the primary reason for excluding earlier draft years. Each draft year and individual player statistics were downloaded using a Python script. The HTML was parsed using a Python package called Scrapy. The draft data and individual player data were merged to provide the following: name, age, season[2] , birth date, birth month, Games Played(GP), Points(Pts), TOI, Average TOI(avgTOI), Overall, Round, Draft Year, Draft Age, Draft

[1]The NHL draft from 1998 to 2004 had 9 rounds, but switched to a 7 round draft in 2005.

[2]The 'season year' (eg. 2010-11), played by a player and their respective season statistic metrics are provided.

Team, Nationality(Nat), Position(Pos), and First Eligible Season. In order to get the player information in a workable format to develop statistical models, data manipulating programs were written and applied to the original merged data set.

Players often play more than one position. For example, a player may play Centre and Right Wing (C/RW) or Right Wing and Defense (RW/D). Although some players may play more than one position, for the purpose of this thesis, a player's primary or dominant position was used.

### 2.1.1 Data problems and issues

There were a number of data issues and errors that needed to be verified and/or corrected prior to moving forward with any analysis.

- Duplicated Players: Players that are drafted into the NHL, if they are not signed by their team, may re-enter the draft at a future draft. Table 2.1 provides a list of players that needed to be removed to avoid duplication of metrics.

Table 2.1: Re-drafted Players

| | | |
|---|---|---|
| Ryan Vanbuskirk | Justin Papineau | Garrett Bembridge |
| Paul Flache | Tim Brent | Matthew Lombardi |
| Gerard Dicaire | Mike Rupp | Dany Roussin |
| Teigan Zahn | Mikko Lehtonen | Eric Hunter |
| Brett Scheffelmaier | Ramzi Abid | Jonas Fiedler |
| Brenden Kichton | Kyle Wanvig | Eric Johansson |
| Jordan Bendfeld | Brandon Nolan | Craig Brunel |
| Nathan Paetsch | Ryan Murphy | Shay Stephenson |
| Jeremy Van Hoof | Juraj Mikus | Sean Collins |
| Will Colbert | Ashton Rome | Mathieu Chouinard |
| Martin Vagner | Masi Marjamaki | Craig Anderson |
| Peter Reynolds | Alan Quine | Rob Zepp |
| Trevor Hendrikx | Fedor Fedorov | Peter Hamerlik |
| Alexandre Picard | Ryan Murphy | Mike Brown |
| Jarret Stoll | Brent Gauvreau | Frederik Andersen |
| Mike Zigomanis | Charlie Stephens | Andy Chiodo |

- Matching Draft name with an individual player bio proved challenging, as there were over 200 players whose names were either shortened or misspelled. The players that fell into this category required their data to be entered manually. It should be noted that this most often occurred with non-North American players (primarily from Russia) and players who have not played a single NHL game.

- Often-times there were missing values. Missing values were usually a parsing issue with the HTML. For example, if a player achieved a significant milestone such as scoring the most points, then their value was written in bold and therefore omitted from the initial parse. During the 2012/2013 shortened season[3], any player who played all 48 games had their GP written in bold and these were omitted from the initial parse. However, there were a number of players that had missing values from their individual bio (most commonly age) and this required manual verification with NHL.com.

## 2.2 Data Summary

For all drafts from 1998 to 2013 there were 941 Centremen selected, 1242 Defensemen, 1264 Forwards(Wingers), and 406 Goalies. As Goalies have their own metrics that are different from other players, they were excluded from portions of the analysis.

As the number of non-North American players selected is far fewer than North American-born players, they were grouped in the following manner:

- All former U.S.S.R countries such as Belarus, Ukraine, Kazakhstan, etc. were grouped as "Russia".

[3]Team owners locked out the players over a labour dispute. When an agreement was reached between the players(NHLPA) and the owners a shortened season was created.

- All Northern European countries such as Sweden, Finland, Norway, and Denmark were grouped as ”Scandinavian”.
- All European countries such as Switzerland, Germany, Czech Republic, Slovakia, etc. were grouped as ”Europe”.
- Finally all other countries were grouped as ”Other”.

Canadian-born players dominate the draft; there were 1660 Canadian (CA) players drafted in the NHL during this time frame. The next closest was the United States (US) with 883, followed by Scandinavian countries with 541. Europe had 416 players, mostly Czech and Slovakian, and Russia had 352; there was one Other player drafted (from Japan).[4] Figure 2.1 was captured early on during the visualization of our data

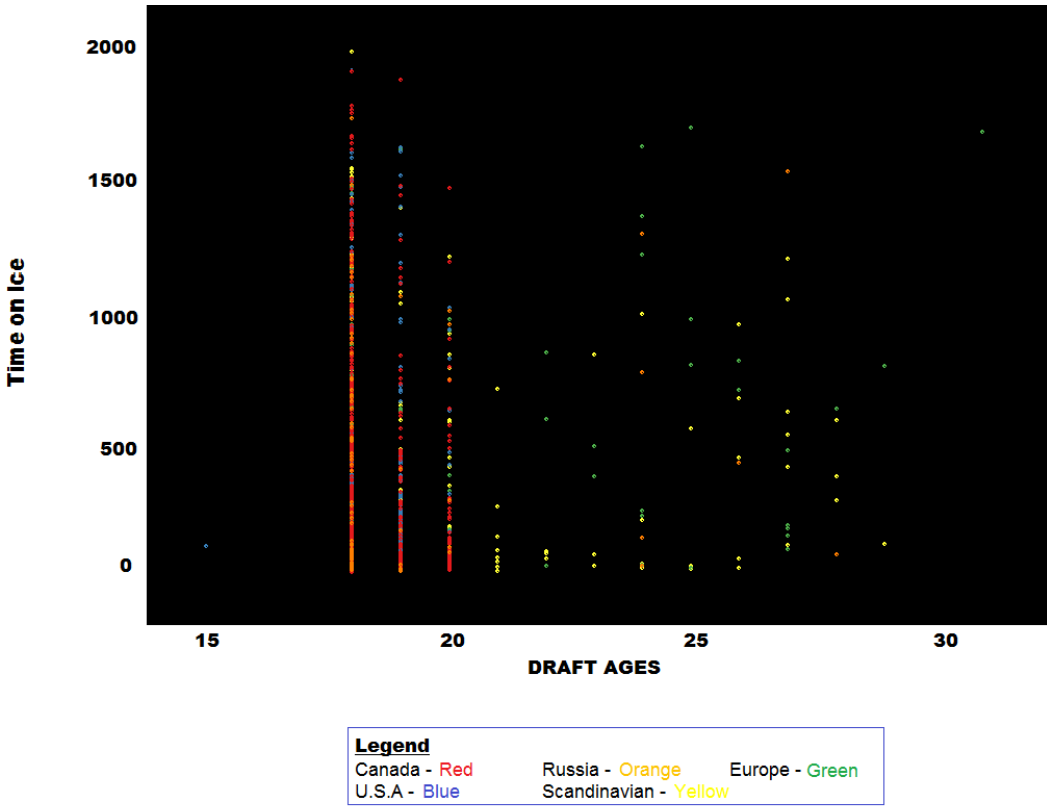



FIGURE 2.1: Early Visualization of TOI vs. Draft Age by Nationality grouping

[4]Yutaka Fukufuji is the first Japanese goaltender to be drafted in the NHL. He was drafted 238th Overall by the Los Angeles Kings in 2004.[18]

and is TOI vs. Draft Age, coloured by the Nationality groups specified.

It provides an interesting look at our data, and confirms what we already know, i.e.:

- That North American players who are not drafted by age 21 must try to enter the NHL by some other means, such as US Colleges.
- All Non-North American players must be drafted into the NHL prior to signing with a team, as seen by the various ages over 21.
- Finally, the minimum age that a player can be drafted is 18. The one outlier was an example of inaccurate data.[5]

## 2.3 Software used

As mentioned, all data scraping and parsing were done in **Python 2.7** using the package called Scrapy version 0.24.4. All data manipulation, cleansing, correction and merging were done in **R 3.1.3** and **Microsoft Excel**. All statistical analysis were carried out in **R 3.1.3**. All models were fitted using the standard conventions within their **R** libraries, which were: **stats, nnet, e1071, and rpart**. Finally, this document was created in LaTeX using the Thesis template provided by Sharelatex.com.

[5] There were two Brian Lee's Drafted, but Hockey Reference thought it was Brian Lee drafted 9th Overall by Ottawa in the 2005 NHL Draft, which would have made him 15 years old in 2002 when the first Brian Lee was drafted 71st Overall by Anaheim.[19]

# Chapter 3

# Mathematical Techniques

Data analysis is organized into two different types. First, a Career Games Played by Draft model is explained, followed by a Career GP model for individual players. It should be noted that historical draft and player data formed an integral part of predicting Career GP and TOI. Finally, models for determining and creating a Value Pick Chart are explained.

## 3.1 Career Games Played

### 3.1.1 Career Games Played by Draft

First, a simple 2-D model is considered. Let $x$ represent the Draft Year and $P_x$ represent the GP proportion for a given Draft Year $x$. The GP proportion is determined by taking the sum of all drafted players $i$ who have played a specified number of games divided by the total number of drafted players ($m$) for a given Draft Year $x$. That is,

$$P_x = \frac{\sum_{i=1}^{m} I(i)}{m} \quad , \text{ where } \quad I(i) = \begin{cases} 1 & \text{if player } i \text{ plays at least 1 (or 160) GP,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\text{for } i = 1, \ldots, m. \tag{3.1}$$

Based on our early visualization, the data appeared to be polynomial with 3 local extrema, as can be seen in Figures 4.1 and 4.4. Therefore, we assume our model is a quartic polynomial, represented by:

$$P_x = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4. \tag{3.2}$$

This model will be applied in Chapter 4.1.1.1 & 4.1.1.2 to the proportions of one Career GP and 160 Career GP.

### 3.1.2 Career Games Played by Individual Player

There were a number of data issues and errors that needed to be verified and/or corrected prior to moving forward with any analysis, as mentioned in Chapter 2. Predicting a player's Career GP after 4 seasons is a valuable tool for any team, player, or agent. A team may want to estimate how many games their asset is predicted to play. Likewise, a player and/or their agent may want to know if they will reach a minimum number of games for pension reasons.

Three models were considered to predict the Career GP of an individual player: Artificial Neural Net (ANN), Support Vector Machine (SVM), and a Locally Weighted

Scatterplot Smoother (LOESS). These three models were combined in a voting model. However, prior to examining these three models, a Generalized Linear Model(GLM) was considered.

#### 3.1.2.1 Model Development: the Generalized Linear Model

In Section 2.1, all player metrics that can be used in a model are listed. After visualizing and analysing the data with basic statistics, a subset of these metrics was used to reduce multicollinearity. For example, TOI and Pts are highly correlated so only TOI was used. Player positions (Pos) were converted to dummy variables $D_1$ and $D_2$; where

$$D_1 = \begin{cases} 1 & \text{represents Centres,} \\ 0 & \text{otherwise.} \end{cases} \qquad D_2 = \begin{cases} 1 & \text{represents Defence,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

and when both are 0 it represents the forward Wingers[1]. All variables are listed Table 3.1 with their description:

TABLE 3.1: Variables for Player $i$

| | |
|---|---|
| $Y_i$ | The response Variable, Career Games Played |
| $x_{i1}$ | Season 1 Time On Ice |
| $x_{i2}$ | Season 2 Time On Ice |
| $x_{i3}$ | Season 3 Time On Ice |
| $x_{i4}$ | Season 4 Time On Ice |
| $x_{i5}$ | Overall Pick |
| $x_{i6}$ | Players Current Age |
| $D_{i1}$ | Position Dummy Variable 1 |
| $D_{i2}$ | Position Dummy Variable 2 |

[1]In Section 3.2 so all player positions are represented in the Value Pick Chart we included Goalies in the analysis. They are represented by both $D_1$ *and* $D_2 = 1$.

Using these variables the Poisson Generalized Linear Model to model career GP considered is:

$$\begin{aligned}
\mu_i = \; & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 D_{i1} + \beta_8 D_{i2} + \beta_9 x_{1i} D_{i1} \\
& + \beta_{10} x_{i1} D_{i2} + \beta_{11} x_{i2} D_{i1} + \beta_{12} x_{i2} D_{i2} \beta_{13} x_{i3} D_{i1} + \beta_{14} x_{i3} D_{i2} + \beta_{15} x_{i4} D_{i1} + \beta_{16} x_{i4} D_{i2} \\
& + \beta_{15} x_{i5} D_{i1} + \beta_{18} x_{i5} D_{i2} + \beta_{19} x_{i6} D_{i1} + \beta_{20} x_{i6} D_{i2} + \beta_{21} x_{i1} x_{i5} + \beta_{22} x_{i2} x_{i5} \\
& + \beta_{23} x_{i3} x_{i5} + \beta_{24} x_{i4} x_{i5} + \beta_{25} x_{i1} x_{i6} + \beta_{26} x_{i2} x_{i6} + \beta_{27} x_{i3} x_{i6} + \beta_{28} x_{i4} x_{i6} + \beta_{29} D_{i1} D_{i2} \\
& i = 1, \ . \ . \ . \ , m \ (m = 3076) \text{ and } \mu_i = E\{Y_i\}.
\end{aligned} \tag{3.4}$$

Stepwise regression was run, giving an Akaike Information Criterion (AIC) of 258350 and a Deviance of 252731 and the following, slightly modified, model was produced:

$$\begin{aligned}
\mu_i = \; & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 D_{i1} + \beta_8 D_{i2} + \beta_9 x_{1i} D_{i1} \\
& + \beta_{10} x_{i1} D_{i2} + \beta_{11} x_{i2} D_{i2} \beta_{12} x_{i3} D_{i1} + \beta_{13} x_{i3} D_{i2} + \beta_{14} x_{i4} D_{i2} \\
& + \beta_{15} x_{i5} D_{i1} + \beta_{16} x_{i5} D_{i2} + \beta_{17} x_{i6} D_{i1} + \beta_{18} x_{i6} D_{i2} + \beta_{19} x_{i1} x_{i5} + \beta_{20} x_{i2} x_{i5} \\
& + \beta_{21} x_{i3} x_{i5} + \beta_{22} x_{i4} x_{i5} + \beta_{23} x_{i1} x_{i6} + \beta_{24} x_{i2} x_{i6} + \beta_{25} x_{i3} x_{i6} + \beta_{26} x_{i4} x_{i6} \\
& i = 1, \ . \ . \ . \ , m \ (m = 3076) \text{ and } \mu_i = E\{Y_i\}.
\end{aligned} \tag{3.5}$$

Models 3.4 and 3.5 form the bases for the three main models used for predicting Career GP.

#### 3.1.2.2 Model Development: Artificial Neural Net

For simplicity, let all interaction variables in Equation 3.4 be represented by the $x$ variables associated with their corresponding $\beta$ subscript. For example, $x_{i3}x_{i5}$ will be represented as $x_{i,21}$. So from Equation 3.4 we have our predictor nodes $\mathbf{x} = x_1, x_2, ..., x_{29}$ and using a simple feedforward artificial neural network (ANN) to model career GP, our model with one potential *Hidden Layer*, $H$, of size $k$ and input weights, $\mathbf{w}$, will be represented as $w_{0k}, w_{1k}, ..., w_{29k}$ For case $i$, the input to node $j$ of the *Hidden Layer* becomes:

$$h_{ij} = w_{0j}Bias_i + w_{1j}x_{i1} + ... + w_{29,j}x_{i,29} \quad , \; j = 1, \; . \; . \; . \; , k \text{ and } i = 1, \; . \; . \; . \; , m \tag{3.6}$$

where $Bias$ is the intercept term.

An example of a ANN with one *Hidden Layer* containing 3 nodes is represented in Figure 3.1.

The weights from the *Hidden Layer* to the output layer are given by:

$$\mathbf{W_n^T} = \begin{bmatrix} \mathbf{W_{0,n}} \; \mathbf{W_{1,n}} \; \mathbf{W_{2,n}} & \cdots & \mathbf{W_{k,n}} \end{bmatrix} \tag{3.7}$$

where $n$ represents the number of nodes in the output layer,

and the weights are determined by the minimum Mean Squared Error (MSE)

and the input, due to case $i$, from the *Hidden Layer* to the Output layer is given by:

$$Bias_i W_{0,n} + \sum_{j=1}^{k} h_{ij} W_{jn} \qquad , \; i = 1, \; . \; . \; . \; , m. \tag{3.8}$$

#### 3.1.2.3 Model Development: the Support Vector Machine

The Support Vector Machine (SVM) was the third model used to predict career GP. Let $\mathbf{X}$ represent a matrix of our variables, $\forall \; i = 1, \; . \; . \; . \; , m$ players. Let $f(\mathbf{X}) = Y_i$ represent our response variable i.e. career GP and $\xi_i$ represent a slack variable. Using

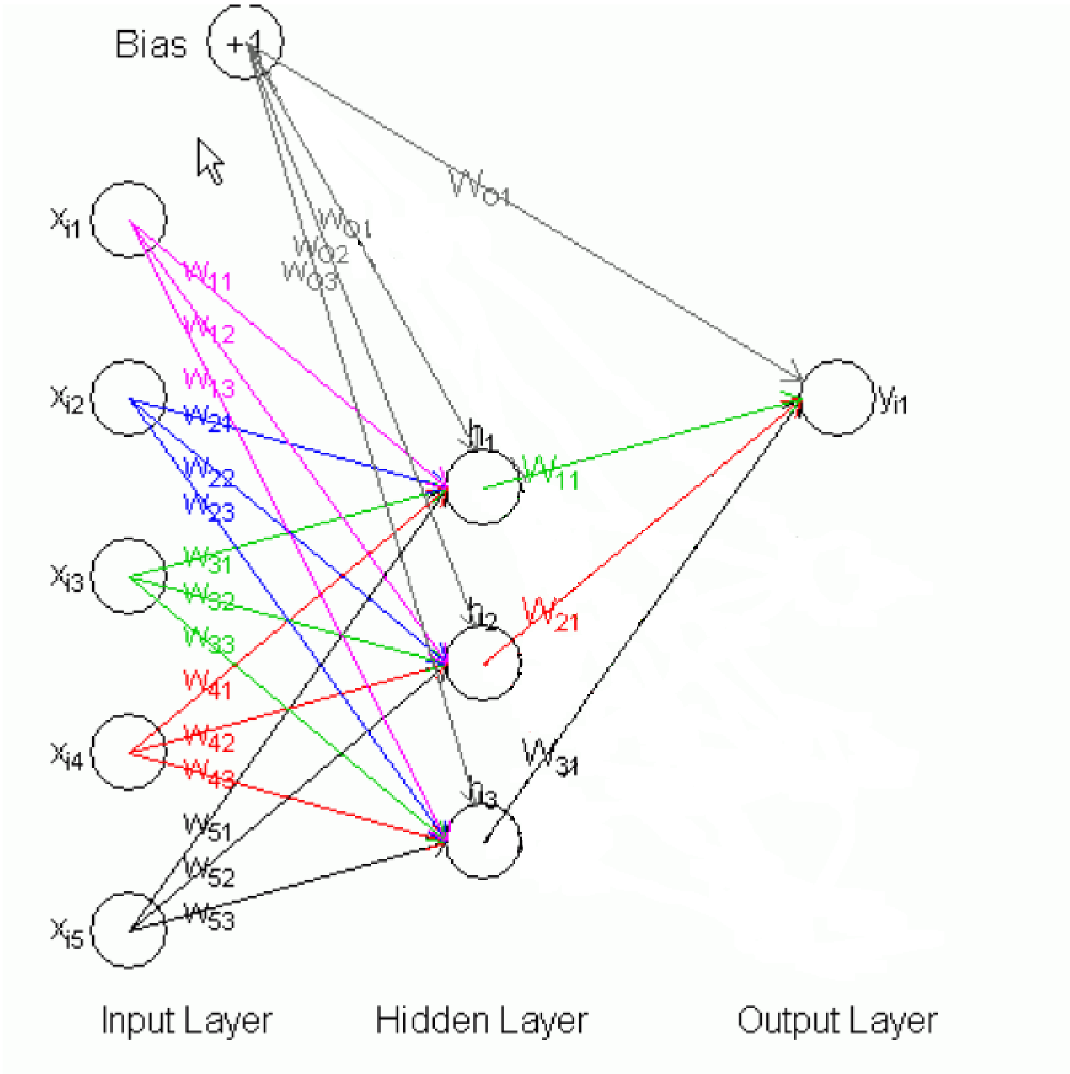



FIGURE 3.1: ANN with one *Hidden Layer* consisting of 3 nodes.

the formulation of Hastie, Tibshirani and Friedman [20]:

$$
\begin{aligned}
& f(\mathbf{X}) = \mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\beta_0} + \boldsymbol{\epsilon} \quad , \qquad\qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\
& \text{where we want} \quad \min_{\beta, \beta_0} \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^{m} \xi_i \\
& \text{subject to} \quad \xi_i \geq 0, \quad Y_i - (\mathbf{x}_i^T \boldsymbol{\beta} + \boldsymbol{\beta_0}) \leq \epsilon_i + \xi_i \quad \forall i.
\end{aligned}
\tag{3.9}
$$

With constant $C = \frac{1}{\lambda} > 0$ (where $\lambda$ is a regularization parameter), our $\epsilon$-insensitive error measure is determined by:

$$V_\epsilon(\xi) = \begin{cases} 0 & \text{if } |\xi| \leq \epsilon, \\ |\xi| - \epsilon, & \text{otherwise.} \end{cases} \tag{3.10}$$

If $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{\mathbf{0}}$ are the minimizers, then the idea is to construct a Lagrange function such that:

$$L : \min_{\alpha_i, \alpha_i^*} \epsilon \sum_{i=1}^{m} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{m} Y_i(\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,i'=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_{i'}^* - \alpha_{i'})\langle x_i, x_{i'} \rangle \tag{3.11}$$

subject to the following constraints:

$$\begin{aligned} &0 \leq \alpha_i, \alpha_i^* \leq \frac{1}{\lambda} = C \\ &\sum_{i=1}^{m} (\alpha_i^* - \alpha_i) = 0, \\ &\alpha_i \alpha_i^* = 0. \end{aligned} \tag{3.12}$$

Within R, we use the default settings, which uses a $3^{rd}$ degree polynomial kernel: $K(x_i, x_i^{'}) = \langle h(x_i), h(x_{i'}) \rangle = (1 + \langle x_i, x_{i'} \rangle)^3$ to produce our results. This alters Equation 3.11 and our new Support Vector Algorithm is as follows:

$$\text{maximize} \quad \begin{cases} -\frac{1}{2} \sum_{i,i'=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_i - \alpha_i^*) K \langle x_i, x_{i'} \rangle \\ -\epsilon \ \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) + \sum_{i=1}^{\ell} Y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{3.13}$$

$$\text{subject to} \quad \begin{cases} \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) & = & 0 \\ \alpha_i, \alpha_i^* & \in & [0, C]. \end{cases}$$

#### 3.1.2.4 Model Development: the Locally Weighted Regression

Locally Weighted Regression (LOESS) was the fourth model used to predict career GP. Again using the variables from Table 3.1, we let $\mathbf{x}_i = (x_{i3}, x_{i4}, x_{i5}, x_{i6})$. We generate our data by the most commonly used framework for regression:

$$Y_i = g(\mathbf{x}_i) + \epsilon_i \quad , \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{3.14}$$

It is assumed that $g$ is a smooth function of the $\mathbf{x}_i$ variables. The LOESS will provide an estimate $\hat{g}(\mathbf{x})$ at any value of $\mathbf{x}_i$. Essentially, our model will find the estimate, $\hat{g}$, at $\mathbf{x}_i$ whose value is closest to $g(\mathbf{x}_i)$, i.e. of all the points in the neighbourhood of $\mathbf{x}_i$, take the one closest to it.

#### 3.1.2.5 Model Development: the Voting Method

Let $N$ represent the total number of runs. $N$ training and test sets are created and applied to the models 3.1.2.2, 3.1.2.3 and 3.1.2.4 with an imposed lower bound of zero, since career GP cannot be below zero. The results are then converted to binary; 1 if the player is predicted to achieve 160 GP, 0 otherwise. A simple majority vote determines whether a player is predicted to achieve 160 GP or not. Also, the model allows us to capitalise on the strengths of the individual models.

## 3.2 Value Pick Chart

Similar to Section 3.1.2.1, all variables used in the model development for the draft Value Pick Chart need to be represented. Table 3.2 represents a list of all variables with their descriptions:

TABLE 3.2: Variables for Player $i$, Season $\ell$

| | |
|---|---|
| $Y_{i\ell}$ | The response variable, TOI for season $\ell$ |
| $x_{i1}$ | Player's Age |
| $x_{i2}$ | Player's Draft Age |
| $x_{i3}$ | Overall Pick |
| $D_{i1}$ | Position Dummy Variable 1 |
| $D_{i2}$ | Position Dummy Variable 2 |
| $D_{i3}$ | Nationality Dummy Variable 3 |
| $D_{i4}$ | Nationality Dummy Variable 4 |
| $x_{i,8}$ | Draft Age interacting with Overall Pick |
| $x_{i,9}$ | Draft Age interacting with Position Variable 1 |
| $x_{i,10}$ | Draft Age interacting with Position Variable 2 |
| $x_{i,11}$ | Draft Age interacting with Nationality Variable 3 |
| $x_{i,12}$ | Draft Age interacting with Nationality Variable 4 |
| $x_{i,13}$ | Overall Pick interacting with Position Variable 1 |
| $x_{i,14}$ | Overall Pick interacting with Position Variable 2 |
| $x_{i,15}$ | Overall Pick interacting with Nationality Variable 3 |
| $x_{i,16}$ | Overall Pick interacting with Nationality Variable 4 |

Since $Y_{i\ell}$ is multivariate, for $i = 1, 2, \cdots, m$ players and $\ell = 1, 2, \cdots, 7$ seasons, our model will also be multivariate: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & y_{16} & y_{17} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} & y_{26} & y_{27} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{m1} & y_{m2} & y_{m3} & y_{m4} & y_{m5} & y_{m6} & y_{m7} \end{bmatrix}$$

$$\mathbf{X}_{i,16} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & D_{11} & D_{12} & D_{13} & D_{14} & \cdots & x_{1,16} \\ 1 & x_{21} & x_{22} & x_{23} & D_{21} & D_{22} & D_{23} & D_{24} & \cdots & x_{2,16} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{m1} & x_{m2} & x_{m3} & D_{m1} & D_{m2} & D_{m3} & D_{m4} & \cdots & x_{m,16} \end{bmatrix} \tag{3.15}$$

$$\boldsymbol{\beta}_{16,j} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{07} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{17} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{27} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{15,1} & \beta_{15,2} & \cdots & \beta_{15,7} \\ \beta_{16,1} & \beta_{16,2} & \cdots & \beta_{16,7} \end{bmatrix}, \ \boldsymbol{\epsilon}_{i,j} = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{17} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{27} \\ \vdots & \vdots & \cdots & \vdots \\ \epsilon_{m1} & \epsilon_{m2} & \cdots & \epsilon_{m7} \end{bmatrix}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ and where $\boldsymbol{\beta}$ and $\sigma^2$ are unknown parameters.

Finally, to determine the values within the Value Pick Chart, we take the maximum value which is for the first overall pick $O_1$ and then dividing each pick ($O_t$ for draft pick $t = 1, ..., 210$) by it and then multiplying by the value we choose to assign the first overall pick (say 1000),i.e.

$$\text{Value for pick } t = \frac{O_t}{O_1} \times 1000 \qquad t = 1, ..., 210. \tag{3.16}$$

# Chapter 4

# Applications

Before producing the Value Pick Chart, career GP were analysed; first by draft and then by individual players. As previously mentioned in Chapter 3, predicting a player's career GP is a valuable tool for any team, player, or agent and using historical draft and player data formed an integral part in predicting career GP.

## 4.1 Career Games Played

### 4.1.1 Career Games Played by Draft

#### 4.1.1.1 Predicting One NHL Career Game Played

Getting drafted to the NHL does not guarantee a player will ever play a single NHL game. In fact, for the draft years 1998 to 2009 less than 48% of drafted players will ever play an NHL game. The proportions of players playing at least one NHL game for each draft year, rounded to the nearest fourth decimal place, are given in Table 4.1:

Table 4.1: One NHL Game Played Proportion

| Draft Year | Proportion |
|---|---|
| 1998 | 0.5116 |
| 1999 | 0.4118 |
| 2000 | 0.4027 |
| 2001 | 0.4360 |
| 2002 | 0.3608 |
| 2003 | 0.4452 |
| 2004 | 0.4364 |
| 2005 | 0.4826 |
| 2006 | 0.4178 |
| 2007 | 0.4408 |
| 2008 | 0.4976 |
| 2009 | 0.5000 |

Plotting proportion vs. Draft Year (Figure 4.1), we can see that there appears to be a few outliers in 2002 and 2005; otherwise, the data appears to be quartic. What happened in 2002 and 2005 to cause these outliers?

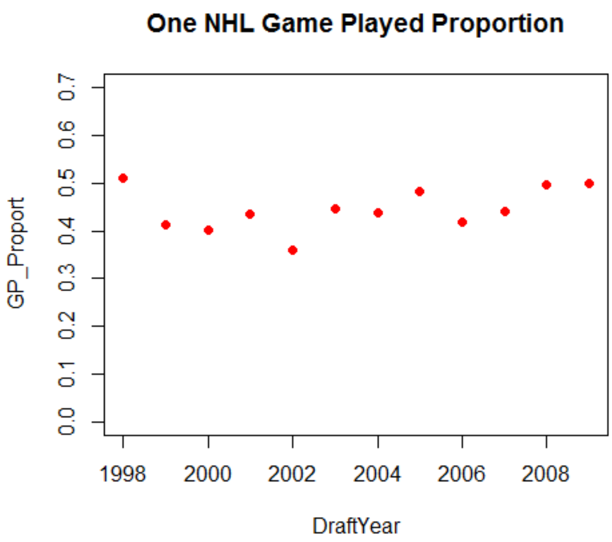



Figure 4.1: One NHL Game Played Proportion

In 2002, 123 non-North American players were drafted, one of the highest number of non-North American players drafted during the time frame of our study. It is also largely considered by the experts as one of the weakest draft years. Between 2000 and 2006, 621 non-North American players were drafted. Out of concern, the International

Ice Hockey Federation (IIHF) did a study based on non-North Americans playing in North America[21]. Of the 621, 388 were considered marginal or below average and 133 never played an NHL game. After the study, the IIHF convinced the NHL to let non-North American players continue to develop their game in Europe. The numbers show that after 2004, fewer than 60 non-North Americans are drafted each year.

In 2005, it was the Lock-out draft[1] where the NHL implemented a number of changes to the game, as well as the Draft. The NHL switched from nine rounds to seven and it was the first year where less than 30% of non-North American players were selected.

Despite the two outlier years, we applied Equation 3.2 to our data and obtained Figure 4.2.

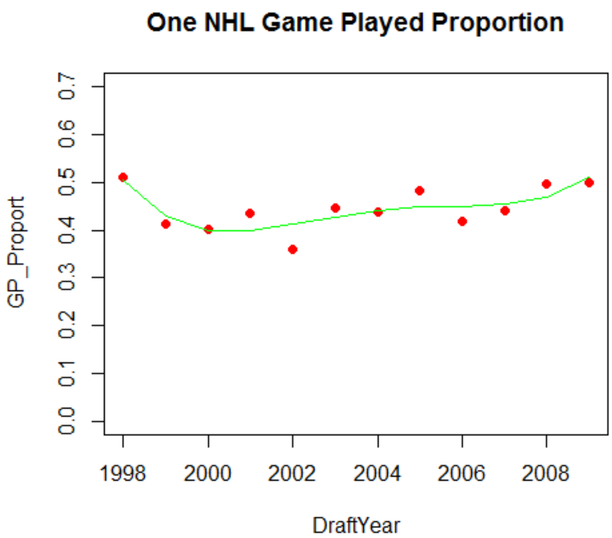



FIGURE 4.2: Fitted One NHL Game Played Proportion

Even with the two outliers, the fitted model has fairly low residual values, which are all under 0.05 as shown in Figure 4.3:

[1]The entire 2004-2005 season was cancelled due to the NHL owners locking out the players (NHLPA) in a labour dispute.

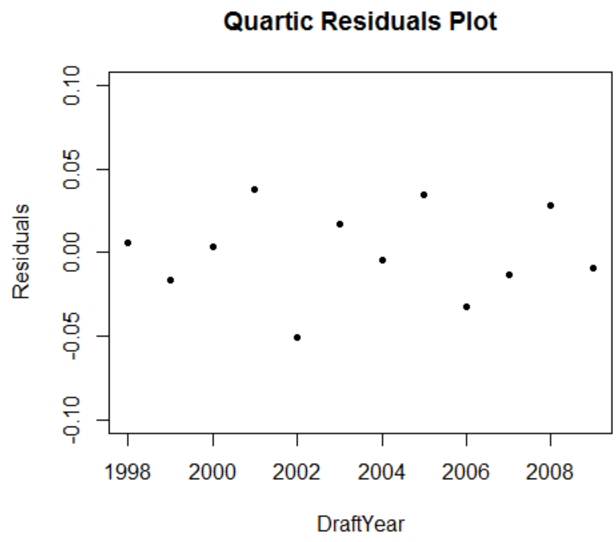



FIGURE 4.3: Residual Plot for One NHL Game Played Proportion

#### 4.1.1.2 Predicting 160 NHL Career Games Played

In 1986, the NHL switched from a Defined Benefit Pension Plan to a Defined Contribution Plan. After the Lock-out in 2005, the New Collective Bargaining Agreement(CBA) between the National Hockey League Players' Association (NHLPA) and the NHL switched back to a Defined Benefit Pension Plan, an unprecedented change in today's business world, which made business headlines across North America. Essentially, the new pension plan means "NHL players were eligible to earn a maximum (pension) of $50,000 (all currency U.S.) a year in 2012 after playing at least 160 games".[22]

Achieving 160 GP in the NHL is one goal for any newly drafted NHL player, as set-out in the new CBA. However, this goal maybe difficult to obtain, considering that less than 48% of drafted players play a single NHL game and the average percentage that play at least 160 GP is less than 21%. Again, plotting Proportion vs. Draft Year as shown in Figure 4.4, we can see that the outliers in 2002 and 2005 are less noticeable and that the data still appears to be quartic based on the three local extrema. We again apply Equation 3.2 to our data to get Figure 4.5. The fitted model performed much

TABLE 4.2: 160+ Games Played Proportion

| Draft Year | Proportion |
|---|---|
| 1998 | 0.2481 |
| 1999 | 0.1765 |
| 2000 | 0.1672 |
| 2001 | 0.2318 |
| 2002 | 0.1821 |
| 2003 | 0.2500 |
| 2004 | 0.2096 |
| 2005 | 0.2217 |
| 2006 | 0.2019 |
| 2007 | 0.1754 |
| 2008 | 0.1991 |
| 2009 | 0.1619 |

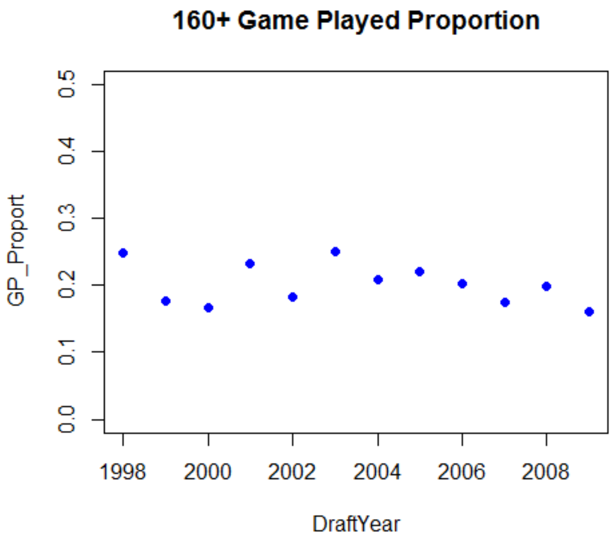


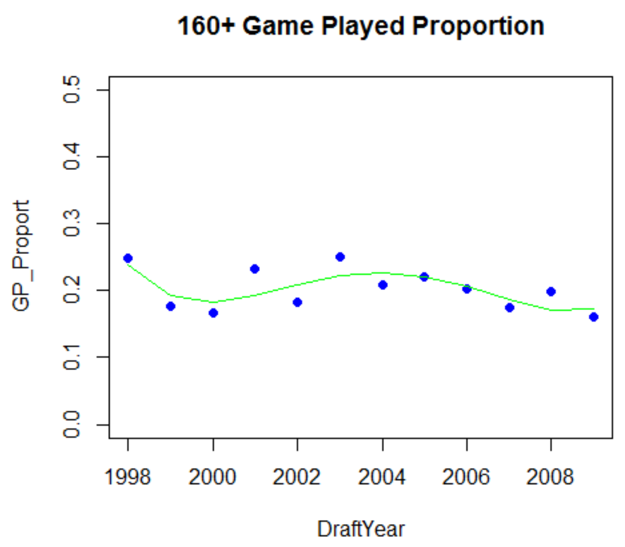

FIGURE 4.4: 160+ Game Played Proportion

better than the previous one with even lower residual values - all under 0.04, as can be seen in Figure 4.6. Using 99% and 95% confidence intervals, we apply this model to predict the proportion of players to achieve over 160 GP in 2010 and 2011 Draft Years (see Table 4.3).

TABLE 4.3: 2010 and 2011 160+ GP Prediction Based on 99 & 95 % Confidence Interval

| Draft Year | Fit | 95% CI | 99% CI |
|---|---|---|---|
| 2010 | 0.2386 | (0.1791, 0.2980) | (0.1506, 0.3266) |
| 2011 | 0.1924 | (0.1541, 0.2306) | (0.1358, 0.2490) |

FIGURE 4.5: Fitted 160+ Game Played Proportion



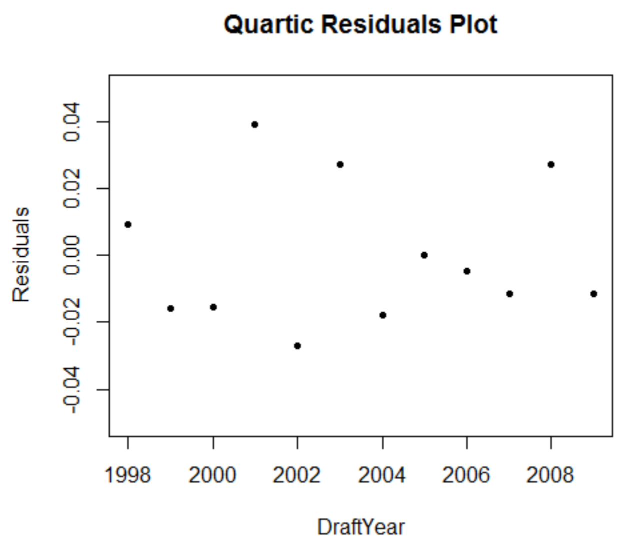




FIGURE 4.6: Residual Plot for 160 Game Played Proportion

### 4.1.2 Career Games Played by Individual Player

In Section 3.1.2, recall that predicting a player's Career GP after four seasons is a valuable tool to any team, player, or agent. In this Section, all four models developed in Chapter 3 are considered and applied to the data to predict the Career GP of an individual player.

#### 4.1.2.1 Results of the Generalized Linear Model

Using models 3.4 and 3.5 developed in 3.1.2.1, both models produced error rates close to 9.8% respectively as seen in the following confusion matrix in Figure 4.7. Object 1 represents the value if a player played at least 160 games and object 0 represents otherwise. The column values come from what the model predicted - if the true value matches with the object then it was an accurate prediction; otherwise it was an error. Therefore, the error rate is calculated by the total number of errors divided by the total number of players in the set[2].

```
          true
object    0    1    | Row Sum
  0       1619 170  | 1789
  1       33   235  | 268
  ------- ---- ---- | ----
  Col Sum 1652 405  | 2057
attr(,"error")
[1] 0.09868741
```

(a) GLM

```
          true
object    0    1    | Row Sum
  0       1620 170  | 1790
  1       32   235  | 267
  ------- ---- ---- | ----
  Col Sum 1652 405  | 2057
attr(,"error")
[1] 0.09820126
```

(b) STEP

FIGURE 4.7: Confusion Matrix on the training set for GLM & STEP Models

However, both had a complete mismatch when predicting GP (i.e. they were unable to accurately predict the correct GP for all players in the test set[3]) and therefore we abandoned the model and considered 3.1.2.2, 3.1.2.3 and 3.1.2.4.

#### 4.1.2.2 Results of the Artificial Neural Net

Using Equation 3.8 and library *nnet* in R with one *Hidden layer*, the model was run on our training set. Over-fitting occurred and the model only produced two results (266.36 GP or 21.16 GP). Thus, we tried a Neural Net with no *Hidden layer*, which produced

[2] In Figure 4.7 (GLM), the model predicted 203 errors (170 and 33) and divided by the total number of players in the training set (2057) provides the error rate of 0.09868741.

[3] Multiple training and test sets were created by randomly sampling from the data and then verified to make sure there was a proper distribution of the key variables.

much more realistic results for Career GP. As seen in the following confusion matrix in Figure 4.8, the ANN only produced an error rate of 8.56% on the training set.

```
              true
object     0    1    | Row Sum
  0        1609 133  | 1742
  1        43   272  | 315
  ------- ---- ----  | ----
  Col Sum 1652 405   | 2057
attr(,"error")
[1] 0.0855615
```

FIGURE 4.8: Training Set Confusion Matrix for the ANN (no *Hidden Layer*)

On the test set, the ANN with no *Hidden Layer* performed well and only produced an error rate of 9.22% as seen in the confusion matrix in Figure 4.9.

```
              true
object     0    1    | Row Sum
  0        797  67   | 864
  1        27   128  | 155
  ------- ---- ----  | ----
  Col Sum 824  195   | 1019
attr(,"error")
[1] 0.0922473
```

FIGURE 4.9: Test Set Confusion Matrix for ANN (no *Hidden Layer*)

#### 4.1.2.3 Results of the Support Vector Machine

Next we applied the Support Vector Machine (SVM) model using the library *e1071* in R. As indicated in 3.1.2.3, the SVM default settings were used, i.e. a polynomial kernel of degree 3. The results on the training set were equivalent to the ANN (no *Hidden Layer*) with an error rate of 8.56%, as can be seen in the confusion matrix Figure 4.10. Other degrees of the polynomial kernel were attempted but they all had the same or

```
              true
object     0    1    | Row Sum
  0        1606 130  | 1736
  1        46   275  | 321
  ------- ---- ----  | ----
  Col Sum 1652 405   | 2057
attr(,"error")
[1] 0.0855615
```

FIGURE 4.10: Training Set Confusion Matrix for the SVM

worse results.

Finally, on the test set, the SVM did not do as well as the ANN (no *Hidden Layer*). As can be seen in the confusion matrix in Figure 4.11, the SVM had an error rate of 9.52%.

```
          true
object    0    1     | Row Sum
  0       795  68    | 863
  1       29   127   | 156
  ------- ---- ----  | ----
  Col Sum 824  195   | 1019
attr(,"error")
[1] 0.09519136
```

Figure 4.11: Test Set Confusion Matrix for the SVM

#### 4.1.2.4 Results of the Locally Weighted Regression

The last model considered, prior to using a voting model, was the LOESS model developed in 3.1.2.4. The LOESS model was much more sensitive to the data than the Neural Network and the SVM, and therefore the training window had to be expanded to avoid prediction misses[4]. Essentially, the training and test sets needed an equal proportion of the max and min values for the variables. On the training set, the LOESS did the worst of the three, with an 8.85% error rate, as can be seen in the confusion matrix in Figure 4.12. It also performed the worst on the test set with an error rate of only 9.72%, as

```
          true
object    0    1     | Row Sum
  0       1585 115   | 1700
  1       67   290   | 357
  ------- ---- ----  | ----
  Col Sum 1652 405   | 2057
attr(,"error")
[1] 0.08847837
```

Figure 4.12: Training Set Confusion Matrix for the LOESS

can be seen in the confusion matrix in Figure 4.13.

[4] For example, if there was only one player with the maximum age and they were in the test set then the LOESS could not predict career GP for that player resulting in a NA value.

```
              true
object     0     1     | Row Sum
  0        785   60    | 845
  1        39    135   | 174
  ------- ---- ----    | ----
  Col Sum 824   195    | 1019
attr(,"error")
[1] 0.09715407
```

FIGURE 4.13: Test Set Confusion Matrix for the LOESS

#### 4.1.2.5 Results of the Voting Model

In Section 3.1.2.5 $N$ represents the total number of training and test sets to create and use with each of our three remaining models. We let $N = 1000$ and ran the models above. Again, as indicated in 3.1.2.5, after each run, the values were converted to binary (achieved at least 160 or not) and then a majority vote was taken.

Each player had a varying number of results, depending on the total number of times they were in the test set. If a player $x$ had mixed results, i.e. 1's and 0's, then the majority was taken as the predicted result for that player. The final result was a 9.51% error rate at predicting whether or not a player will achieve at least 160 games in their career, as can be seen in the confusion matrix in Figure 4.14[5].

```
              true
object     0     1     | Row Sum
  0        2372 194    | 2566
  1        98   406    | 504
  ------- ---- ----    | ----
  Col Sum 2470 600     | 3070
attr(,"error")
[1] 0.09511401
```

FIGURE 4.14: Confusion Matrix for the Voting Model

## 4.2 Value Pick Chart

A Value Pick Chart (VPC) places a numeric value on each draft pick 1 through to 210. As previously done in the work of Schuckers et al[16] and [17], the VPCs have always

[5] Although it may seem like the SVM performed as well as the voting model, and that the ANN (no *Hidden Layer*) outperformed the voting model, one should recognize that the test set results provided for the SVM and ANN were only for one sample, whereas the voting model was on 1000.

been monotonically decreasing. Since we are assigning a value to a draft pick, the value also represents the value to an NHL team who selects that draft pick. We followed the same way to value the performance of players after they were drafted as Schuckers et al[17], that NHL "teams have the rights to players for at least the first seven years after they are drafted". Therefore, we will focus on a player's first seven seasons after they are drafted, whether they play NHL games or not, with the exception of the 'lockout' season. (As previously noted, the 2004-2005 season was the NHL 'lockout' season and no NHL games were played so this particular season will not count towards a player's first seven seasons.) In the previous Sections, Goalies were not included due to their unique metrics. However, since Goalies do accumulate TOI minutes and since VPC is based on this, we included them in the analysis and creation of the VPC.

Prior to applying any models to the data, we plotted the TOI of the original data for all players, for their first seven seasons, as seen in Figure 4.15 (First Seven Seasons of TOI vs. Overall Draft Pick). Summing all seven seasons we obtained Figure 4.16 and

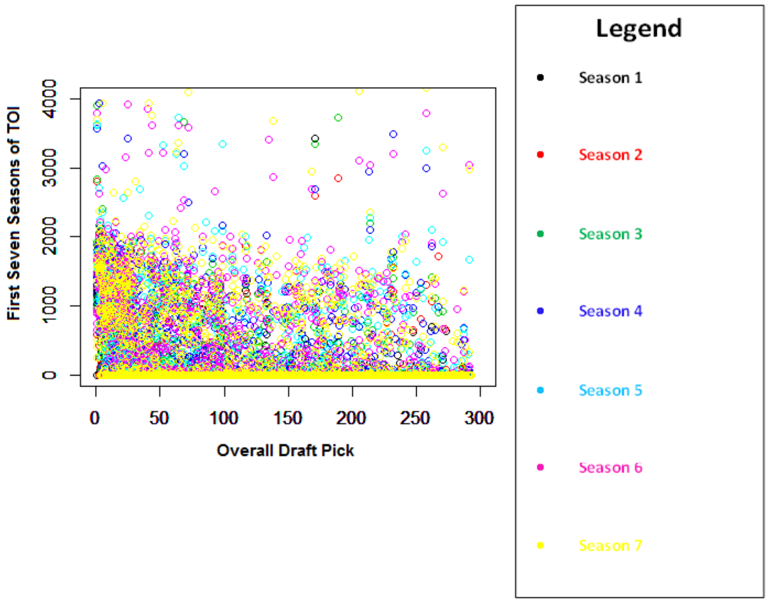



FIGURE 4.15: Original data for all players and coloured by their first seven seasons.

applied LOESS smoothing (span=0.2) on this figure to obtain Figure 4.17, but limited the Overall picks to first seven rounds (i.e. 210 total draft picks).

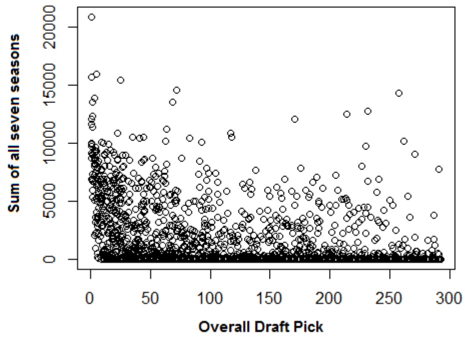



FIGURE 4.16: Original data, sum of first seven seasons for all players.

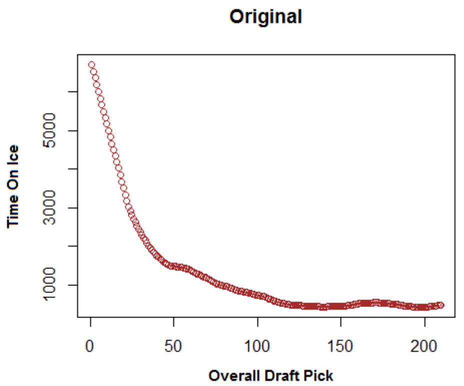



FIGURE 4.17: Original data, smoothed.

Interestingly, there appeared to be an increase or 'bump' between picks 150 and 200. Figure 4.18 zooms in to provide a better look at, and to verify, the 'bump'.

This 'bump' posed a problem since the original data showed TOI was not monotonically decreasing. Applying a *monoreg* function in R using Library *fdrtool*, as shown in Figure 4.19 we see that the data was horizontal from draft pick 100 onwards, which means each draft pick would be approximately equally valued. Therefore we proceeded as if late round draft picks were not monotonically decreasing and that there is a slight 'bump' between picks 150 and 200.

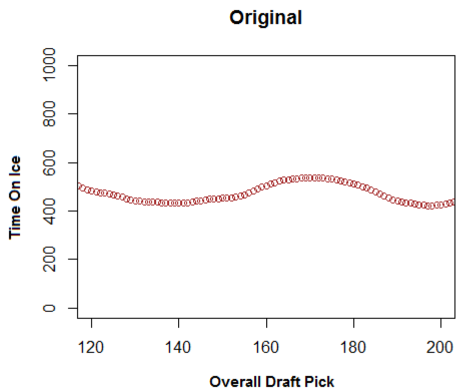



FIGURE 4.18: Zoomed image of Figure 4.17.

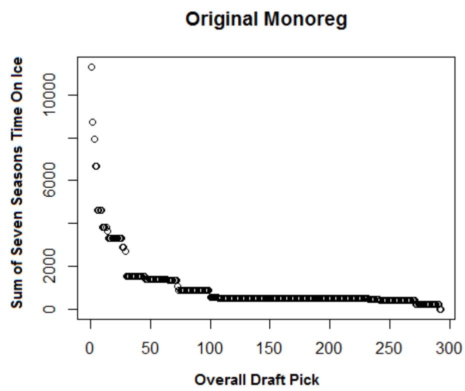



FIGURE 4.19: Monoreg plot of the Original data.

We applied a SVM using a third degree polynomial kernel and a forced lower bound of zero to predict a player's TOI for a player's first seven seasons. We also applied a weighted least squares approach depending on the value of variable $x_{i3}$ (Overall Pick). To adjust for the sampling variability we ran multiple predictions on our test sets (250 in total) and averaged the results; the results are displayed in Figure 4.20.

By using 'Dummy' variables, we separated the results by position and nationality, as seen in Figures 4.21 & 4.22. As displayed, we see that Defense have TOI that increases in the third, fourth and fifth rounds (Overall picks 60 through 150), as well as goalies TOI increases much later (Overall picks 140 to 200). Also, there does not appear to be much difference between Centres and Wingers other than Centres having slightly higher

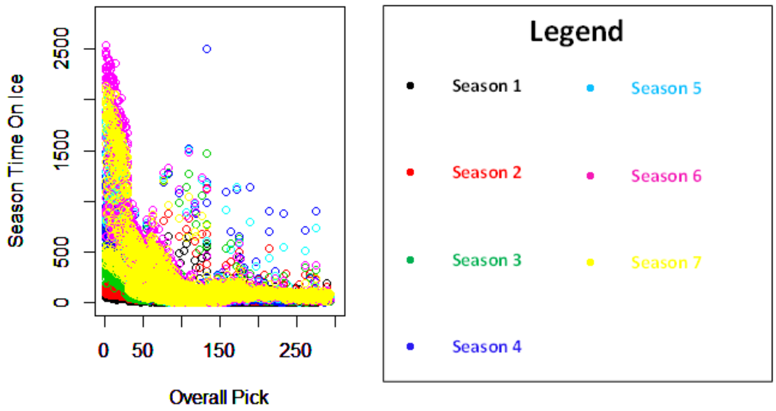



FIGURE 4.20: Predicted TOI for a players first seven seasons in the NHL.

TOI in the first round (Overall pick 1 through 30).

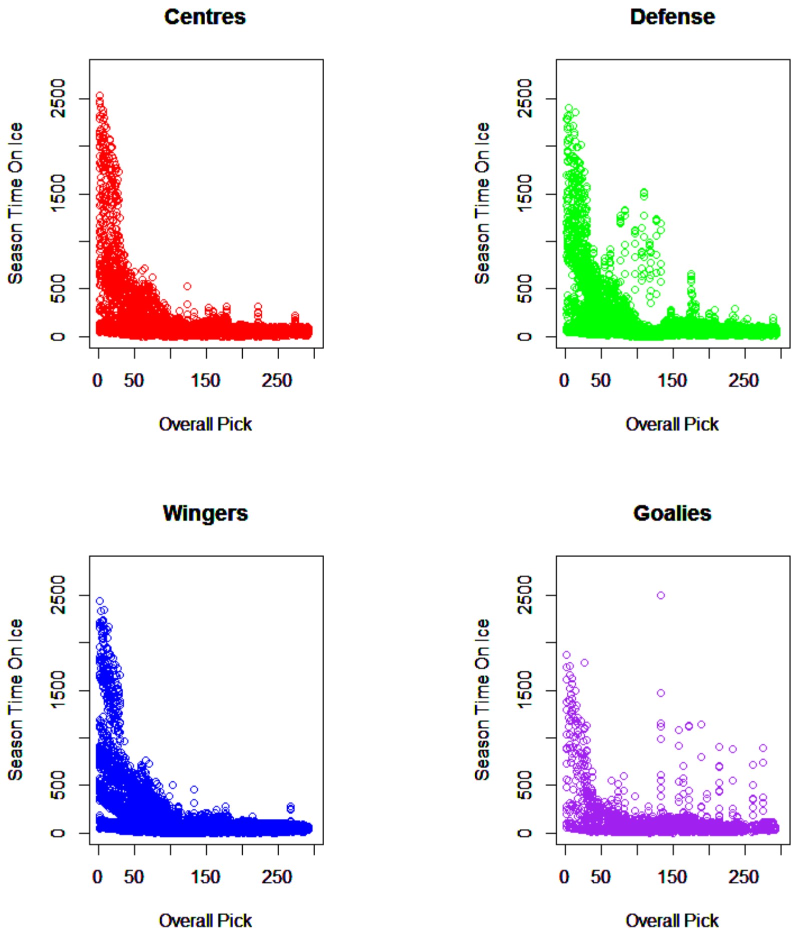



FIGURE 4.21: TOI for first seven seasons by Position.

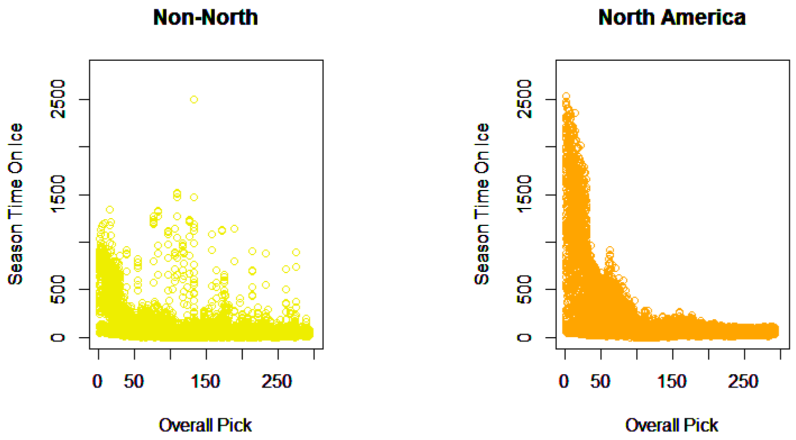



FIGURE 4.22: TOI for first seven seasons by Nationality.

We further separated non-North American and North American players by position, as shown in Figure 4.23. These figures show that North American players dominate non-North American players in the early rounds for TOI. In addition, the increases in TOI for Defense and Goalies in the later rounds appear to be attributed to the non-North American players.

Using TOI, the object was to create a VPC, therefore the seven seasons needed to be summed, as we did with the original data (Figure 4.16). The resulting plots can be seen in Figure 4.24. Finally, we combined all positions, which produced Figure 4.25.

Similar to Figure 4.17, we applied LOESS (Span=0.2) to the data in Figures 4.24 and 4.25 and set the total Overall picks to the first seven rounds (210 total draft picks) to obtain Figures 4.26 and 4.27.

Interestingly, we overlayed the four positions on the same plot (Figure 4.28), and can then see throughout the draft when one position may have an advantage over the others. Similarly to the original data in Figure 4.17 the predicted positions model also produced a 'bump' in TOI between picks 150 and 200. To gain a better look at the 'bump', we zoomed in on the image (Figure 4.29), keeping the scale the same as before

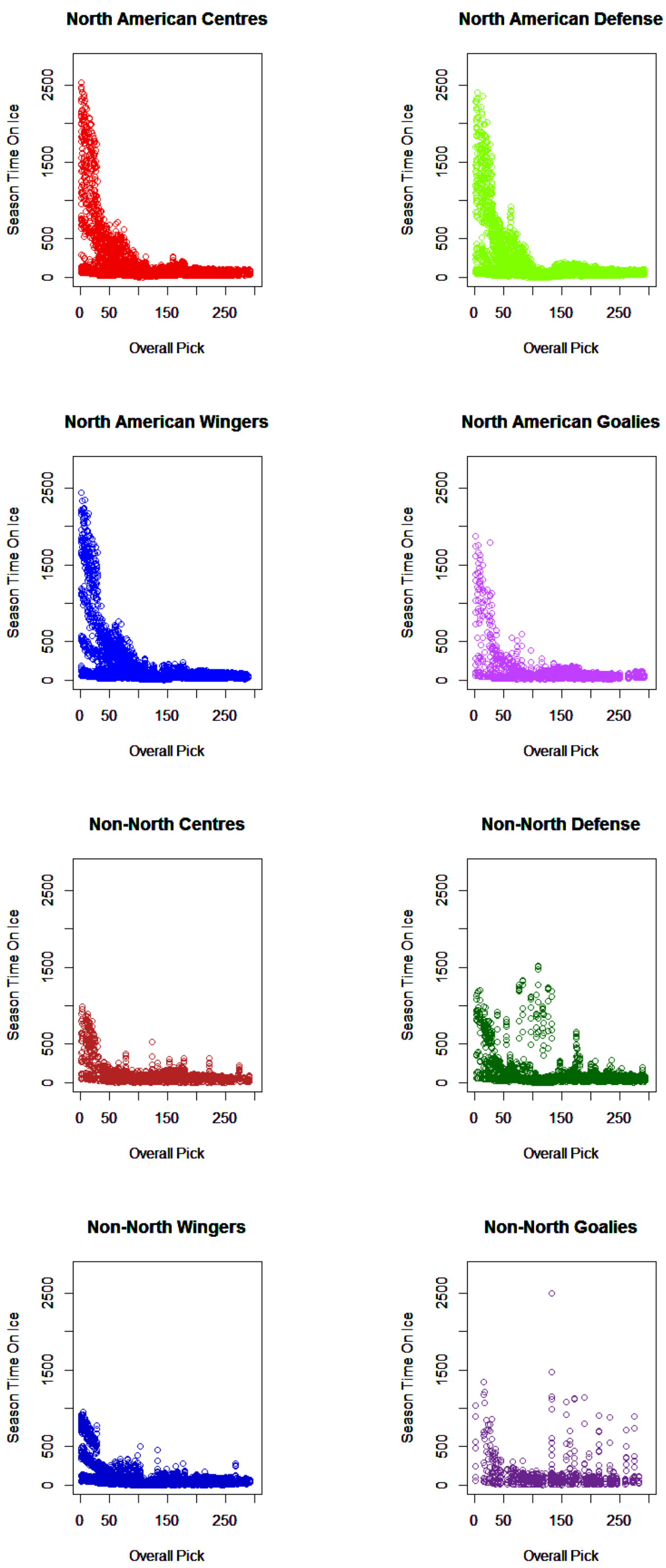



FIGURE 4.23: TOI for first seven seasons by Position for North American & non-North American Players.

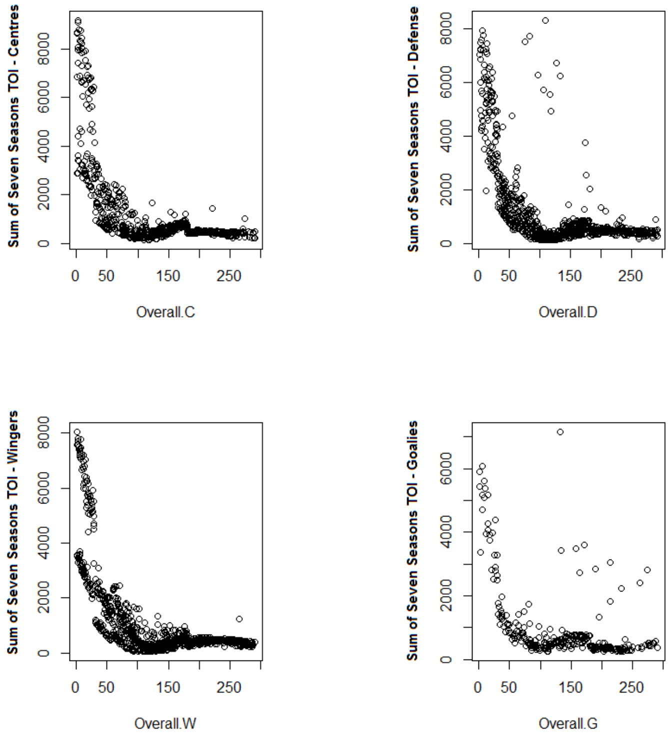



FIGURE 4.24: TOI sum of seven seasons by Position.

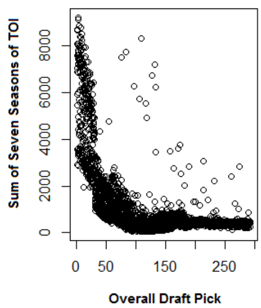



FIGURE 4.25: All positions combined.

in Figure 4.18. As can be seen with Figure 4.29, all player positions clearly contribute to this 'bump' between draft picks 150 and 200, but separating the positions clearly distinguished which position had the most influence on the anomaly (i.e. Goalies).

In Figure 4.30 we can see that there is no major difference in TOI by Nationality between picks 150 and 200.

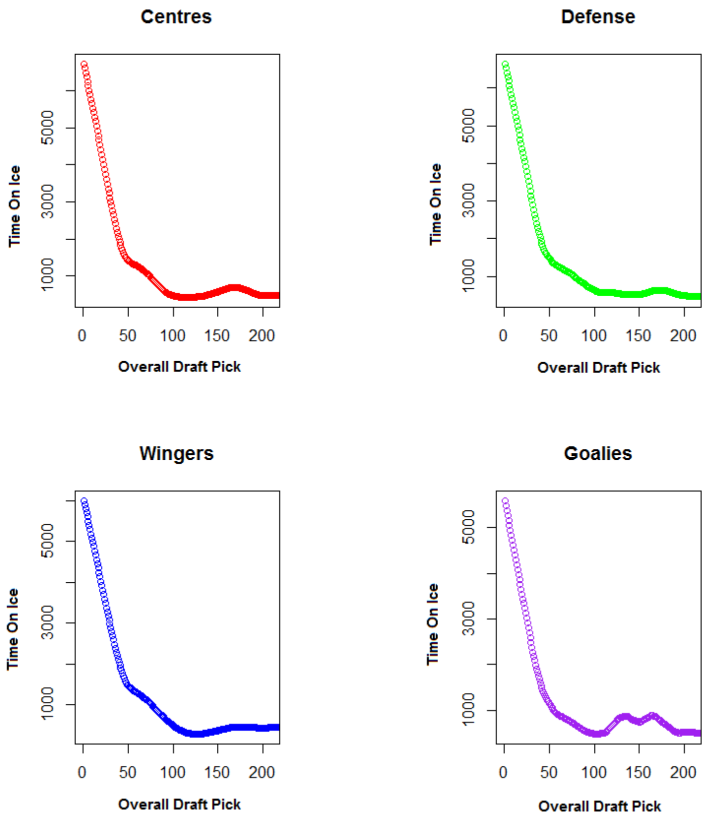



FIGURE 4.26: Smoothing of Figures 4.24.

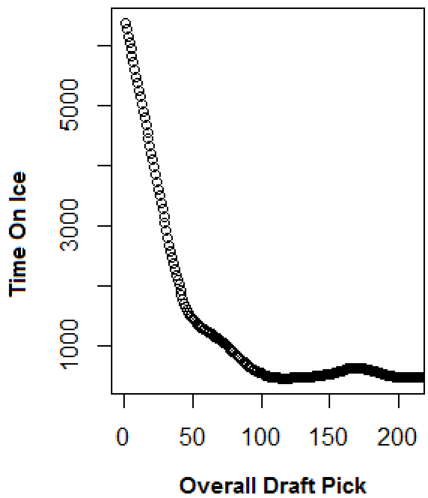



FIGURE 4.27: Smoothing of all positions combined.

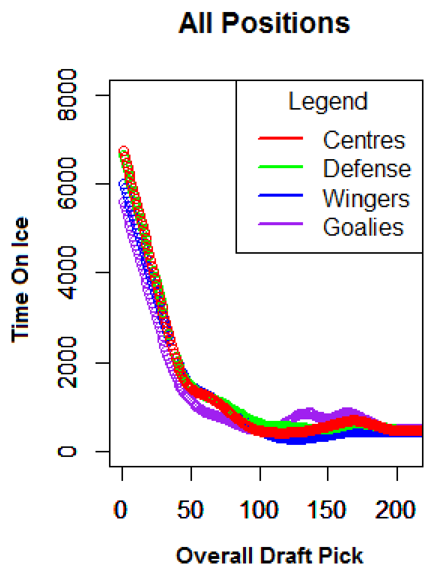



FIGURE 4.28: Overlay of all positions during the draft.

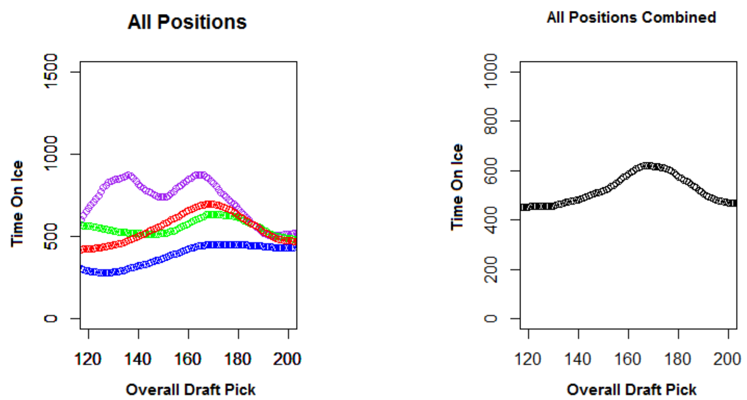



FIGURE 4.29: Magnified image of Figures 4.28 & 4.27 between picks 150 and 200.

However, what Figure 4.30 does indicate is that non-North American players have a slight advantage over North American players regarding TOI between picks 95 and 145.

### 4.2.1 Value Pick Chart Creation and its Uses

Finally, we applied equation 3.16 to each Position (Figure 4.26), Nationality (Figure 4.30) and all combined Positions(Figure 4.27), rounded to the nearest whole number,

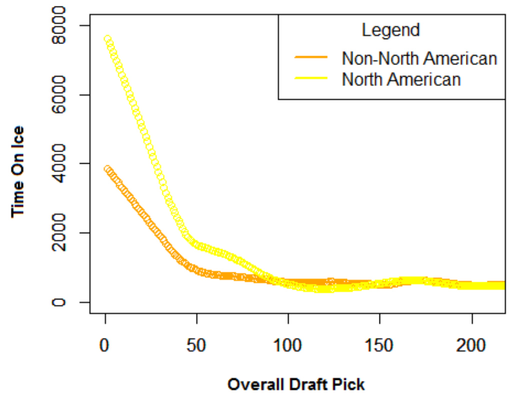



FIGURE 4.30: Nationality Comparison.

and obtained the Value Pick Charts accordingly. These are combined and are located in Appendix A.

The following is a description of the VPC in Appendix A, followed by an explanation on how to interpret the VPC:

- The first column 'Overall' represents the Overall pick within the draft.
- The second column 'All' represents the value of a draft pick when there is no position preference, which is an average of all the position values.
- Columns two through five represent the value for that particular position.
- Columns six and seven represent the value of choosing a North American player vs. a non-North American player.

If a team has no particular positional preference with their draft pick then they should refer to the 'All' column in the VPC, which is similar to what Schuckers et al[17] had previously created. However, if a team wishes to choose a Defenseman with their draft pick then they should refer to the 'Defense' column. Finally, after a team selects which column they would like to consider then they can refer to the last two columns to decide whether or not they should choose a North American or non-North American player (or they may wish to ignore Nationality all together).

We believe these VPCs can provide NHL teams and personnel options during the draft when selecting a draft pick or when considering a trade. As mentioned, if a team has no positional preference then they can refer to the 'All' column in the VPC or they may want to choose the position that offers them the most draft value. It may also provide them an opportunity to seek out a trade. For example, suppose we have two teams, the Ottawa Senators and the Toronto Maple Leafs. Let us assume the Maple Leafs have the second Overall draft pick (as well as all their other draft picks)[6] and the Ottawa Senators have the eighth Overall draft pick (as well as all their other draft picks). If the Toronto Maple Leafs want to draft a Defenseman (worth a value of 971) with their first round pick, they may want to trade that selection to the Ottawa Senators, who want to draft a Centre (worth a value of 982 with the second pick). The Senators may trade their eighth Overall pick to the Maple Leafs for their second Overall pick, and possibly a player and/or other draft pick(s). The subsequent 'pieces' of the trade should total close to 100(i.e. the difference between 912(second Overall pick) and 812(eighth Overall pick)), or greater than 100 if the Maple Leafs use the Defense column to value the trade since that is the position they wanted to select. In order to come close to the 100 value and using our VPC with no specific position, the Senators should include their 158th (sixth round) draft pick (worth a value of 91). This is a win-win for both teams. The Maple Leafs can still choose a Defenseman with their first round draft pick, and they have managed to obtain an additional player(or draft pick), whereas the Senators get to move up in the draft and choose the Centre they wanted.

[6]We note that at the time of this thesis, the Toronto Maple Leafs were in a position to finish with one of the top three draft picks, but we are merely providing a hypothetical example.

# Chapter 5

# Conclusion and Future Work

## 5.1 Career Games Played

For hockey players, getting drafted to the NHL is a dream come true, but their journey should not end there. As shown throughout Section 3.1, less than 48% of all drafted players ever play an NHL game. However, their goal should not be one GP, but a career where they achieve at least 160 GP. Reaching at least 160 GP is an even more challenging task than one GP, as on average, less than 21% of all drafted players reach this level in their careers. Using Equation 3.2 we showed that, for the draft years 2010 and 2011 respectively, only 23.86% and 19.24% will reach at least 160 GP. It would seem that 2010 players are predicted to perform slightly better than the average and 2011 slightly poorer than the average.

Four years after a player is drafted, if a player has not reached 160 GP then a team, player, or a player's agent may want to know if they will ever reach that milestone. Models were developed and combined in this thesis to create a valuable tool to predict career GP and hence provide an answer to this question. As the Voting Model demonstrated in Section 4.1.2.5, the voting model had a error rate of only 9.51% in predicting Career GP.

## 5.2 Value Pick Chart

The original data on TOI by Overall draft pick showed that the NHL draft is not completely monotonically decreasing. During the sixth round (picks 150 to 180) there is a slight increase in TOI. The predicted model suggests that Goalies are one reason for the slight increase in the sixth round. The increase in the sixth round may be due to:

- Between Rounds 1 and 5, teams are more risk averse (or conservative) with their draft picks, relying on their team scouts and other player information to make their choices, which for the most part results in a monotonically decreasing value, as can be seen in the VPCs in Appendix A.
- Between Rounds 6 and 7, teams begin to take more risk on players and perhaps they are relying on one scout's strong recommendation, as opposed to the majority. The result is that these risks are paying off more than some fourth and fifth round picks resulting in the increase ('bump') that we observe for TOI.

In addition, as shown in Figure 4.28, the predicted model demonstrates when a team should consider choosing one player position over another. Likewise, Figure 4.30 shows when a team should consider a non-North American player over a North American player (i.e. between pick 95 and 145).

Finally, we recognize that predicted TOI is not the only criteria for valuing a draft selection; criteria such as junior performance, injuries/health, team needs, financial reasons, and other considerations might adjust the values in Appendix. A.

## 5.3 Future Work

As more player metrics are introduced by the NHL, more data can be included in predictive models such as those introduced in this thesis. Some metrics such as blocked shots, Face-off Wins & Losses, Corsi, and Fenwick are just a few examples of metrics that could have improved our models, but were not used since they were not recorded prior to 2007.

Some statistical methods like error bounds could have been introduced to validate some plots, but due to research deadlines we were unable to include them. Similarly, we would have liked to increase some of the sampling techniques to further reduce variability, however that would have required much more time and computer power to produce.

The Nationality (North American vs. non-North American) within the VPC could be broken down further for comparison (i.e. Canadian, US, and non-North American).

The example with Toronto Maple Leafs and Ottawa Senators is an indication where Game Theory can be applied (i.e. A Win-Win). Trade scenarios using the VPC could be studied using Game Theory principles.

Lastly, as the current NHL regular season (2015-2016) comes to an end, the 2009 draft players can be included in the model (i.e. using a player's first seven seasons after draft). Introducing this new data would allow interesting comparisons with the results in this thesis.

# References


[1] CBC Home Video. Hockey A People's History - A simple Game Episode 1, 2006.

[2] National Hockey League. National Hockey League Official Rules 2015-2016. http://www.nhl.com/nhl/en/v3/ext/rules/2015-2016-Interactive-rulebook.pdf, 2015.

[3] NHL.com. NHL announces changes to Draft Lottery format. http://www.nhl.com/ice/news.htm?id=728795, August 20, 2014.

[4] Elliotte Friedman. In search of Vic Ferrari. http://ww.cbc.ca/sports-content/hockey/opinion/2014/08/in-search-of-vic-ferrari.html, August 6, 2014.

[5] Alexandr Diaz-Papkovich. Data Mining the Play-By-Play. Master of Science Thesis, Carleton University, April 2015.

[6] Tom Awad et al. Glossary. http://www.hockeyprospectus.com/puck/glossary/, January 21, 2015.

[7] Michael Schuckers and James Curro. Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. http://statsportsconsulting.com/main/wp-content/uploads/Schuckers_Curro_MIT_Sloan_THoR.pdf, March 2013.

[8] Claude B. Vincent and Byron Eastman. Defining the Style of Play in the NHL: An Application of Cluster Analysis. *Journal of Quantitative Analysis in Sports 5.10*, page 1133, January 2009.

[9] David Beaudoin and Tim B. Swartz. Strategies for Pulling the Goalie in Hockey. http://people.stat.sfu.ca/~tim/papers/goalie.pdf, April 24, 2010.

[10] Oliver Schulte David Beaudoin and Tim B. Swartz. Biased Penalty Calls in the National Hockey League. *ASA Data Science Journal: Sports Analytics*, TBD.

[11] R Pollard and G Pollard. Long-term trends in home advantage in professional team sports in North America and England (1876-2003). *Journal of Sports Sciences 23*, pages 337–350, April 4, 2005. [DOI:10.1080/02640410400021559].

[12] Joanne M. Doyle and Benjamin Leard. Variations in Home Advantage Evidence from the National Hockey League. *Journal of Quantitative Analysis in Sports 8*, June 2, 2012. [DOI:10.1515/1559-0410.1446].

[13] Gregory A. Trandel and Joel G. Maxcy. Adjusting Winning-Percentage Standard Deviations and a Measure of Competitive Balance for Home Advantage. *Journal of Quantitative Analysis in Sports*, 7, 2011. [DOI:10.2202/1559-0410.1297].

[14] Brian Macdonald and Michael Schuckers. Rink Effects. http://arxiv.org/pdf/1412.1035.pdf, December 3, 2014.

[15] Michael Schuckers. An Alternative to the NFL Draft Pick Value Chart Based upon Player Performance. *Journal of Quantitative Analysis in Sports: Vol. 7: Iss. 2, Article 10*, 7, 2011. [DOI:10.2202/1559-0410.1329].

[16] Michael Schuckers. What's An NHL Draft Pick Worth? A Value Pick Chart for the National Hockey League. http://myslu.stlawu.edu/~msch/sports/Schuckers_NHL_Draft.pdf, 2011.

[17] Michael Schuckers and Steve Argeris. You Can Beat the Market: Estimating the Return on Investment for National Hockey League (NHL) Team Scouting using a Draft Value Pick Chart for the NHL. http://arxiv.org/abs/1411.5754, November 21, 2014.

[18] Hockey Hall of Fame Legends of Hockey. Yutaka Fukufuji. http://www.hhof.com/LegendsOfHockey/jsp/SearchPlayer.jsp?player=21496, February 21, 2016.

[19] NHL.com. Player Profile. https://www.nhl.com/player/brian-lee-8470311, January 11, 2016.

[20] Robert Tibshirani Tevor Hastie and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, 2nd edition, 2009. [DOI:10.1007/b94608].

[21] Rene Fasel. IIHF study opens eyes to development dilemma. *IceTimes*, 10(6), November 2006.

[22] Janet McFarland and David Ebner. Still the 'poor cousin', new CBA improves pension plan for average NHLers. *The Globe and Mail*, January 10, 2013.

# Bibliography

[Cleveland and Devlin, 1988] Cleveland, W. S. and Devlin, S. J. (Sep., 1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610.

[Johnson and Wichern, 2007] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson Prentice Hall, 6th edition. [ISBN:0-13-187715-1].

[Michael H. Kutner and Li, 2004] Michael H. Kutner, Christopher J. Nachtsheim, J. N. and Li, W. (2004). *Applied Linear Statistical Models.* McGraw-Hill/Irwin, 5th edition. [ISBN:0-07-238688-6].

[Mills, 2015] Mills, S. M. (2015). *SECTION 9 Classification (Cont'd) - Neural Networks.* [page 377].

[Smola and Scholkopf, 1998] Smola, A. J. and Scholkopf, B. (October 1998). *A Tutorial on Support Vector Regression.*

# Appendix A

TABLE A.1: **Value Pick Chart**

| **Overall** | All | Centres | Defense | Wingers | Goalies | Non-North | NorthAm |
|---|---|---|---|---|---|---|---|
| **1** | 928 | 1000 | 988 | 893 | 832 | 504 | 1000 |
| **2** | 912 | 982 | 971 | 878 | 816 | 495 | 983 |
| **3** | 895 | 965 | 953 | 863 | 800 | 487 | 966 |
| **4** | 879 | 947 | 936 | 848 | 784 | 478 | 949 |
| **5** | 862 | 929 | 919 | 833 | 767 | 470 | 932 |
| **6** | 846 | 912 | 902 | 818 | 751 | 461 | 915 |
| **7** | 829 | 894 | 885 | 802 | 735 | 452 | 897 |
| **8** | 812 | 876 | 867 | 787 | 719 | 444 | 880 |
| **9** | 795 | 858 | 850 | 771 | 702 | 435 | 862 |
| **10** | 779 | 840 | 832 | 756 | 686 | 426 | 844 |
| **11** | 762 | 822 | 814 | 740 | 670 | 417 | 826 |
| **12** | 745 | 804 | 796 | 725 | 654 | 408 | 808 |
| **13** | 728 | 786 | 778 | 709 | 638 | 399 | 790 |
| **14** | 711 | 768 | 761 | 694 | 622 | 390 | 772 |
| **15** | 695 | 750 | 743 | 678 | 607 | 382 | 754 |
| **16** | 677 | 731 | 725 | 662 | 591 | 373 | 735 |
| **17** | 660 | 712 | 707 | 646 | 575 | 364 | 717 |
| **18** | 643 | 694 | 689 | 630 | 559 | 355 | 699 |
| **19** | 626 | 675 | 670 | 615 | 543 | 346 | 682 |
| **20** | 609 | 656 | 653 | 599 | 528 | 337 | 664 |

Table A.1: **Value Pick Chart (Continued)**

| **Overall** | All | Centres | Defense | Wingers | Goalies | Non-North | NorthAm |
|---|---|---|---|---|---|---|---|
| **21** | 592 | 637 | 635 | 583 | 512 | 328 | 646 |
| **22** | 575 | 617 | 618 | 567 | 496 | 319 | 628 |
| **23** | 557 | 597 | 601 | 551 | 480 | 311 | 609 |
| **24** | 540 | 577 | 583 | 534 | 464 | 301 | 588 |
| **25** | 522 | 557 | 565 | 518 | 449 | 292 | 568 |
| **26** | 504 | 536 | 545 | 502 | 433 | 283 | 549 |
| **27** | 486 | 516 | 525 | 487 | 417 | 274 | 530 |
| **28** | 469 | 498 | 505 | 473 | 401 | 265 | 512 |
| **29** | 453 | 480 | 486 | 459 | 385 | 256 | 494 |
| **30** | 436 | 463 | 467 | 445 | 369 | 248 | 474 |
| **31** | 420 | 446 | 449 | 430 | 353 | 239 | 454 |
| **32** | 403 | 429 | 430 | 416 | 338 | 231 | 433 |
| **33** | 387 | 412 | 412 | 401 | 323 | 222 | 414 |
| **34** | 371 | 395 | 393 | 385 | 309 | 213 | 397 |
| **35** | 354 | 376 | 375 | 369 | 296 | 204 | 380 |
| **36** | 339 | 358 | 358 | 355 | 284 | 196 | 365 |
| **37** | 324 | 340 | 342 | 340 | 272 | 187 | 350 |
| **38** | 310 | 324 | 327 | 327 | 260 | 180 | 336 |
| **39** | 297 | 309 | 314 | 315 | 249 | 172 | 322 |
| **40** | 285 | 295 | 301 | 303 | 239 | 166 | 307 |
| **41** | 273 | 282 | 288 | 292 | 229 | 159 | 293 |
| **42** | 262 | 270 | 277 | 282 | 220 | 153 | 280 |
| **43** | 253 | 260 | 266 | 273 | 212 | 148 | 267 |
| **44** | 244 | 251 | 257 | 264 | 205 | 143 | 256 |
| **45** | 236 | 242 | 249 | 255 | 198 | 138 | 245 |
| **46** | 228 | 234 | 241 | 246 | 191 | 134 | 237 |
| **47** | 221 | 226 | 235 | 237 | 185 | 130 | 231 |
| **48** | 215 | 220 | 229 | 229 | 180 | 126 | 226 |
| **49** | 209 | 215 | 224 | 222 | 175 | 122 | 221 |
| **50** | 205 | 211 | 220 | 217 | 170 | 119 | 217 |
| **51** | 201 | 207 | 216 | 214 | 165 | 117 | 213 |
| **52** | 197 | 204 | 211 | 211 | 161 | 114 | 210 |
| **53** | 194 | 202 | 207 | 209 | 156 | 112 | 208 |
| **54** | 190 | 199 | 203 | 206 | 152 | 110 | 206 |
| **55** | 187 | 197 | 200 | 204 | 148 | 109 | 204 |
| **56** | 185 | 196 | 197 | 201 | 145 | 107 | 201 |
| **57** | 182 | 195 | 194 | 199 | 141 | 106 | 199 |
| **58** | 180 | 193 | 192 | 196 | 138 | 105 | 196 |
| **59** | 178 | 191 | 190 | 194 | 136 | 104 | 194 |
| **60** | 175 | 189 | 187 | 192 | 133 | 103 | 192 |
| **61** | 173 | 187 | 185 | 190 | 131 | 102 | 190 |
| **62** | 171 | 184 | 183 | 188 | 129 | 101 | 188 |
| **63** | 169 | 181 | 181 | 186 | 128 | 101 | 186 |
| **64** | 167 | 179 | 180 | 183 | 127 | 100 | 183 |
| **65** | 165 | 177 | 178 | 181 | 125 | 100 | 181 |
| **66** | 163 | 174 | 176 | 178 | 124 | 99 | 178 |
| **67** | 161 | 172 | 174 | 176 | 122 | 99 | 175 |
| **68** | 159 | 169 | 172 | 173 | 121 | 98 | 171 |
| **69** | 157 | 166 | 170 | 171 | 119 | 97 | 168 |

TABLE A.1: **Value Pick Chart (Continued)**

| **Overall** | All | Centres | Defense | Wingers | Goalies | Non-North | NorthAm |
|---|---|---|---|---|---|---|---|
| **70** | 154 | 163 | 168 | 168 | 117 | 97 | 165 |
| **71** | 152 | 160 | 166 | 166 | 116 | 96 | 162 |
| **72** | 149 | 156 | 164 | 163 | 114 | 96 | 159 |
| **73** | 147 | 152 | 162 | 160 | 112 | 95 | 156 |
| **74** | 143 | 148 | 159 | 156 | 110 | 94 | 153 |
| **75** | 141 | 144 | 157 | 153 | 108 | 94 | 149 |
| **76** | 137 | 139 | 154 | 149 | 107 | 93 | 146 |
| **77** | 135 | 136 | 151 | 146 | 105 | 92 | 142 |
| **78** | 131 | 132 | 148 | 142 | 103 | 91 | 137 |
| **79** | 129 | 129 | 145 | 139 | 101 | 90 | 133 |
| **80** | 125 | 125 | 142 | 135 | 99 | 89 | 129 |
| **81** | 122 | 121 | 139 | 132 | 97 | 88 | 124 |
| **82** | 119 | 117 | 136 | 128 | 95 | 88 | 120 |
| **83** | 116 | 113 | 133 | 125 | 93 | 87 | 116 |
| **84** | 113 | 109 | 131 | 121 | 91 | 87 | 112 |
| **85** | 110 | 105 | 128 | 118 | 88 | 86 | 108 |
| **86** | 107 | 101 | 126 | 115 | 86 | 85 | 104 |
| **87** | 104 | 98 | 123 | 111 | 84 | 85 | 100 |
| **88** | 102 | 95 | 121 | 108 | 83 | 84 | 97 |
| **89** | 99 | 91 | 118 | 105 | 81 | 83 | 93 |
| **90** | 97 | 89 | 115 | 102 | 80 | 82 | 90 |
| **91** | 94 | 86 | 112 | 99 | 78 | 81 | 87 |
| **92** | 92 | 83 | 110 | 96 | 77 | 80 | 84 |
| **93** | 89 | 81 | 107 | 92 | 76 | 79 | 81 |
| **94** | 87 | 78 | 105 | 90 | 75 | 78 | 78 |
| **95** | 85 | 76 | 102 | 87 | 74 | 77 | 76 |
| **96** | 83 | 75 | 100 | 84 | 73 | 76 | 74 |
| **97** | 81 | 73 | 99 | 81 | 72 | 76 | 72 |
| **98** | 80 | 72 | 97 | 78 | 72 | 75 | 70 |
| **99** | 78 | 70 | 96 | 76 | 71 | 75 | 68 |
| **100** | 77 | 69 | 94 | 73 | 71 | 74 | 66 |
| **101** | 76 | 69 | 92 | 71 | 70 | 73 | 65 |
| **102** | 74 | 68 | 91 | 68 | 70 | 73 | 63 |
| **103** | 73 | 67 | 89 | 66 | 70 | 72 | 61 |
| **104** | 73 | 67 | 88 | 64 | 71 | 72 | 60 |
| **105** | 71 | 66 | 87 | 61 | 71 | 72 | 59 |
| **106** | 71 | 65 | 86 | 60 | 71 | 72 | 57 |
| **107** | 70 | 65 | 86 | 58 | 72 | 72 | 56 |
| **108** | 70 | 64 | 86 | 56 | 73 | 72 | 55 |
| **109** | 69 | 64 | 85 | 54 | 73 | 72 | 54 |
| **110** | 69 | 63 | 85 | 53 | 74 | 72 | 53 |
| **111** | 69 | 63 | 85 | 52 | 75 | 73 | 52 |
| **112** | 69 | 63 | 85 | 50 | 76 | 73 | 52 |
| **113** | 69 | 63 | 85 | 49 | 78 | 73 | 51 |
| **114** | 69 | 63 | 84 | 48 | 80 | 73 | 51 |
| **115** | 69 | 62 | 84 | 47 | 83 | 73 | 50 |
| **116** | 70 | 62 | 84 | 46 | 86 | 74 | 50 |
| **117** | 71 | 63 | 84 | 45 | 90 | 74 | 50 |
| **118** | 71 | 63 | 84 | 44 | 93 | 74 | 49 |
| **119** | 72 | 63 | 84 | 44 | 96 | 75 | 49 |
| **120** | 72 | 63 | 84 | 43 | 99 | 75 | 49 |

TABLE A.1: **Value Pick Chart (Continued)**

| **Overall** | All | Centres | Defense | Wingers | Goalies | Non-North | NorthAm |
|---|---|---|---|---|---|---|---|
| **121** | 73 | 63 | 83 | 43 | 102 | 76 | 49 |
| **122** | 73 | 63 | 83 | 42 | 105 | 76 | 50 |
| **123** | 74 | 64 | 83 | 42 | 108 | 76 | 50 |
| **124** | 75 | 64 | 82 | 42 | 111 | 76 | 50 |
| **125** | 76 | 64 | 82 | 41 | 115 | 75 | 51 |
| **126** | 76 | 65 | 81 | 41 | 118 | 75 | 51 |
| **127** | 77 | 65 | 81 | 41 | 121 | 74 | 51 |
| **128** | 77 | 65 | 80 | 41 | 123 | 73 | 51 |
| **129** | 78 | 66 | 80 | 42 | 124 | 72 | 52 |
| **130** | 78 | 66 | 79 | 42 | 125 | 72 | 52 |
| **131** | 79 | 67 | 79 | 42 | 126 | 72 | 53 |
| **132** | 79 | 67 | 78 | 43 | 126 | 71 | 53 |
| **133** | 79 | 68 | 78 | 43 | 127 | 71 | 54 |
| **134** | 80 | 69 | 78 | 44 | 128 | 71 | 55 |
| **135** | 80 | 69 | 78 | 44 | 129 | 70 | 55 |
| **136** | 81 | 70 | 78 | 45 | 129 | 70 | 56 |
| **137** | 81 | 71 | 77 | 46 | 129 | 69 | 57 |
| **138** | 81 | 72 | 77 | 46 | 127 | 69 | 58 |
| **139** | 80 | 73 | 77 | 47 | 124 | 68 | 59 |
| **140** | 80 | 74 | 77 | 48 | 121 | 68 | 60 |
| **141** | 80 | 75 | 77 | 48 | 119 | 68 | 61 |
| **142** | 80 | 76 | 77 | 49 | 118 | 68 | 62 |
| **143** | 80 | 77 | 76 | 49 | 116 | 68 | 63 |
| **144** | 80 | 79 | 76 | 50 | 115 | 68 | 64 |
| **145** | 80 | 80 | 76 | 51 | 113 | 67 | 65 |
| **146** | 80 | 81 | 76 | 51 | 112 | 66 | 66 |
| **147** | 80 | 82 | 76 | 52 | 111 | 65 | 67 |
| **148** | 81 | 83 | 77 | 53 | 110 | 65 | 68 |
| **149** | 81 | 84 | 77 | 54 | 110 | 64 | 69 |
| **150** | 82 | 85 | 77 | 55 | 110 | 64 | 70 |
| **151** | 82 | 86 | 77 | 55 | 110 | 63 | 71 |
| **152** | 83 | 88 | 78 | 56 | 111 | 63 | 72 |
| **153** | 85 | 89 | 79 | 57 | 113 | 64 | 73 |
| **154** | 86 | 90 | 79 | 58 | 115 | 64 | 74 |
| **155** | 87 | 91 | 80 | 59 | 117 | 65 | 75 |
| **156** | 88 | 93 | 81 | 60 | 119 | 67 | 76 |
| **157** | 90 | 94 | 83 | 60 | 121 | 68 | 77 |
| **158** | 91 | 95 | 84 | 61 | 122 | 70 | 78 |
| **159** | 92 | 96 | 85 | 62 | 124 | 71 | 79 |
| **160** | 93 | 98 | 86 | 63 | 126 | 73 | 79 |
| **161** | 94 | 99 | 87 | 64 | 127 | 74 | 80 |
| **162** | 95 | 100 | 88 | 64 | 129 | 76 | 80 |
| **163** | 96 | 101 | 89 | 65 | 129 | 77 | 81 |
| **164** | 97 | 101 | 90 | 65 | 130 | 78 | 81 |
| **165** | 97 | 102 | 91 | 66 | 130 | 79 | 81 |
| **166** | 98 | 103 | 93 | 66 | 129 | 79 | 82 |
| **167** | 98 | 103 | 94 | 66 | 128 | 80 | 82 |
| **168** | 98 | 104 | 94 | 67 | 126 | 80 | 82 |
| **169** | 97 | 104 | 94 | 67 | 124 | 81 | 82 |
| **170** | 97 | 103 | 94 | 67 | 122 | 81 | 82 |

Table A.1: **Value Pick Chart (Continued)**

| **Overall** | All | Centres | Defense | Wingers | Goalies | Non-North | NorthAm |
|---|---|---|---|---|---|---|---|
| **171** | 96 | 103 | 94 | 67 | 119 | 82 | 81 |
| **172** | 95 | 102 | 94 | 67 | 116 | 82 | 81 |
| **173** | 94 | 101 | 94 | 67 | 114 | 81 | 81 |
| **174** | 93 | 101 | 93 | 67 | 112 | 81 | 80 |
| **175** | 93 | 100 | 93 | 67 | 110 | 80 | 79 |
| **176** | 92 | 99 | 93 | 67 | 108 | 79 | 78 |
| **177** | 91 | 98 | 93 | 67 | 106 | 79 | 77 |
| **178** | 90 | 96 | 92 | 67 | 103 | 78 | 76 |
| **179** | 89 | 95 | 92 | 67 | 101 | 78 | 75 |
| **180** | 88 | 93 | 91 | 67 | 99 | 77 | 75 |
| **181** | 86 | 91 | 90 | 67 | 96 | 77 | 74 |
| **182** | 85 | 90 | 89 | 67 | 94 | 77 | 73 |
| **183** | 83 | 89 | 87 | 66 | 91 | 76 | 72 |
| **184** | 82 | 87 | 86 | 66 | 89 | 75 | 70 |
| **185** | 81 | 86 | 85 | 66 | 87 | 75 | 69 |
| **186** | 80 | 85 | 85 | 66 | 85 | 74 | 68 |
| **187** | 79 | 83 | 84 | 66 | 83 | 72 | 67 |
| **188** | 78 | 82 | 83 | 66 | 81 | 71 | 67 |
| **189** | 76 | 80 | 81 | 65 | 79 | 69 | 66 |
| **190** | 75 | 79 | 80 | 65 | 77 | 68 | 65 |
| **191** | 74 | 77 | 79 | 65 | 76 | 67 | 64 |
| **192** | 73 | 76 | 77 | 65 | 75 | 66 | 64 |
| **193** | 73 | 75 | 76 | 65 | 75 | 66 | 63 |
| **194** | 72 | 74 | 75 | 64 | 75 | 65 | 62 |
| **195** | 72 | 73 | 75 | 64 | 75 | 64 | 62 |
| **196** | 71 | 72 | 74 | 64 | 75 | 64 | 62 |
| **197** | 71 | 71 | 74 | 64 | 75 | 64 | 61 |
| **198** | 71 | 71 | 73 | 64 | 76 | 63 | 61 |
| **199** | 71 | 70 | 73 | 64 | 76 | 63 | 61 |
| **200** | 71 | 70 | 72 | 64 | 76 | 63 | 61 |
| **201** | 71 | 70 | 72 | 64 | 77 | 63 | 61 |
| **202** | 71 | 70 | 72 | 64 | 77 | 62 | 61 |
| **203** | 71 | 70 | 71 | 65 | 77 | 62 | 61 |
| **204** | 71 | 70 | 71 | 65 | 77 | 62 | 60 |
| **205** | 71 | 70 | 71 | 65 | 78 | 63 | 60 |
| **206** | 71 | 70 | 71 | 65 | 78 | 63 | 60 |
| **207** | 71 | 70 | 71 | 65 | 78 | 63 | 60 |
| **208** | 71 | 70 | 71 | 65 | 78 | 63 | 60 |
| **209** | 71 | 70 | 71 | 66 | 77 | 63 | 60 |
| **210** | 71 | 70 | 71 | 66 | 76 | 63 | 60 |