Estimating Player Survivability in the NHL

Chris Shoniker
chris.shoniker@gmail.com

1.0 Introduction

Every year, the National Hockey League (NHL) draft brings a revitalization to NHL teams and their fanbases in hopes that they will pick a good player for their team. In this paper, I use data from the 2002-2015 NHL drafts to calculate the probability that a draft player will reach a specific number of games. Survival analysis techniques based on total games played were used to adjust for censored players who had not yet retired and allowed for a much larger dataset. A training model was built on 80% of the given data and a the remaining 20% was used as a test group to determine the probability that those players would reach a specific number of games. The model performed well when detecting whether a given drafted player would ever play in the NHL.

There have been many papers written about the NHL draft and player variables associated with their future performance. According to Schuckers [1], a Poisson Generalized Additive Model is able to outperform the NHL draft selections of teams and their scouts. Additional studies have been conducted on the NHL draft such as the work by Weissbock [2] and Lawrence [3] in which their Prospect Cohort Success (PCS) evaluated the probability that amateur players would play in the NHL by comparing them to a cohort peer group. While not many papers have been written about using survival analysis to determine the expected number of games a drafted player might play, Scully [4] used survival analysis to estimate coaching longevity in three major North American sports: Major League Baseball (MLB), the National Basketball Association (NBA), and the National Football League (NFL). Additionally, several papers have discussed some of the variables used in my model. [5] found the relative age effects (RAEs) existed in the NHL draft, and that players born in the third and fourth quarter of months were twice as likely to reach career milestones as those born in the first quarter (January-April) even though the younger players were drafted later than their productivity warranted. Survival analysis was used in [6] to test the hypothesis that early career achievement may result in early retirement as shown in the MLB, but the hypothesis was not supported in the NHL study. Furthermore, [7] found that height was a significant factor in whether a player was drafted, but was not significant in predicting his career success.

2.0 Data

Data was gathered and cleaned from a variety of sources on all players drafted in the NHL Entry Draft. The primary source of data for this paper was gathered from http://www.hockey-reference.com [8], but http://www.hockeydb.com [9] was also used to gather missing data and cross reference the data from http://www.hockey-reference.com. Additionally, player height and weight were gathered from http://www.nhl.com [10]. Only players drafted in the 2002 draft (June 22-23, 2002) until the 2015 draft (June 26-27, 2015) were include in the analysis.

The data gathered for each player included draft year, draft position, draft round selected, career games played, nationality, amateur team and league, birth month, day and year, and NHL games played for each player. Player nationality was divided into eight groups: Canada, U.S.A., Sweden, Finland, Czech Republic, Slovakia, Russia, and Other. The Other category considers all players with a nationality differing from the seven countries also listed and allowed for a larger number for each factor. This category contains players from smaller NHL player producing European countries and, for one player, Japan.

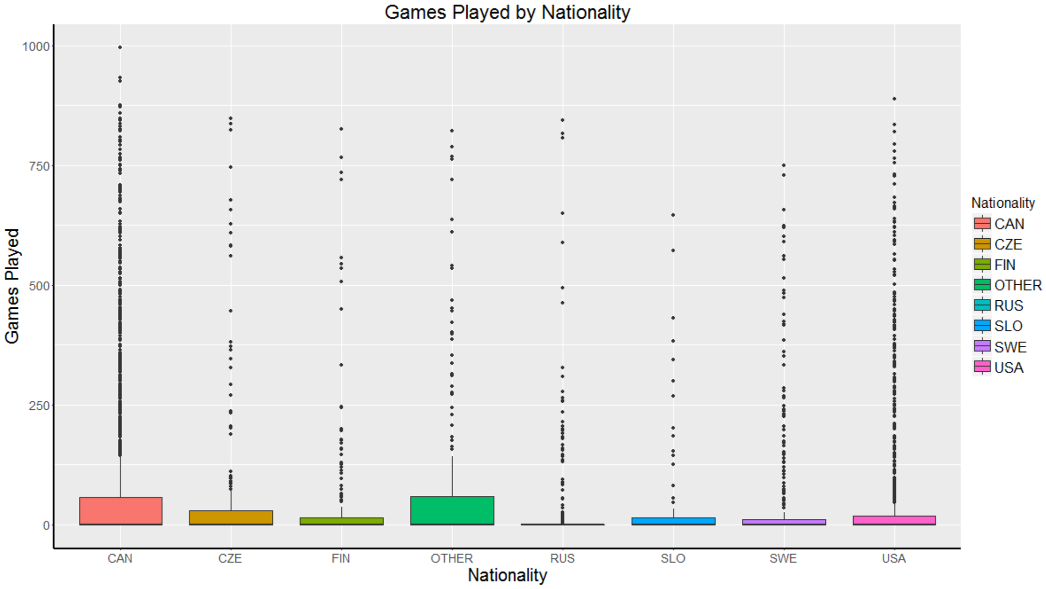

Figure 1: Games Played by Nationality

The players' position was also gathered for each player and was divided into three groups: forward (F), defenseman (D), and goalie (G). Lastly, the age of each player was calculated by subtracting the player's birthdate from June 25th of his draft year thus giving his age at the time he was drafted.

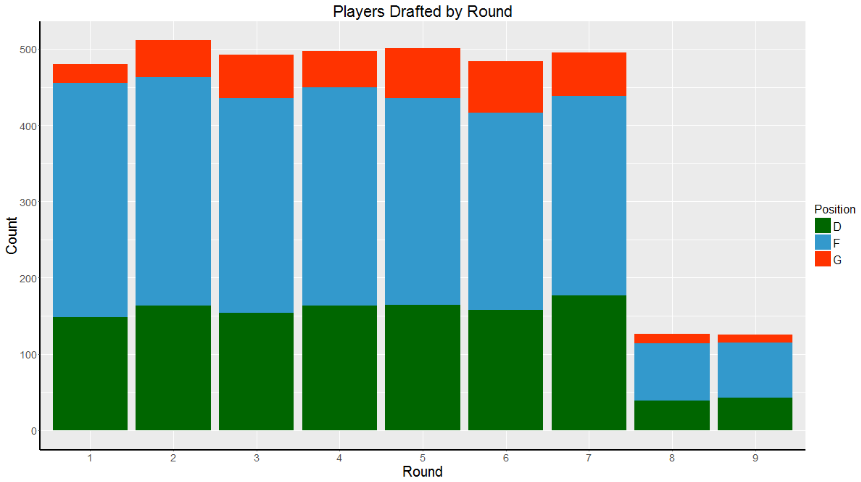

Figure 2: Players Drafted by Position and Round

The response variable was total games played in the NHL, but all 0s were replaced with 0.0001 to allow for logarithmic distributions. A player was censored if the player had played in the 2015-16 season or if the player was drafted in the 2011-2015 drafts. Since almost all players made their first NHL start within the first five years after being drafted (as shown in Figure 3 below), it was safe to assume that any player who was drafted before 2011 was almost certainly never going to make the NHL. Players who were drafted before 2011, but did not play in the 2015 season were assumed to have retired and were not censored.

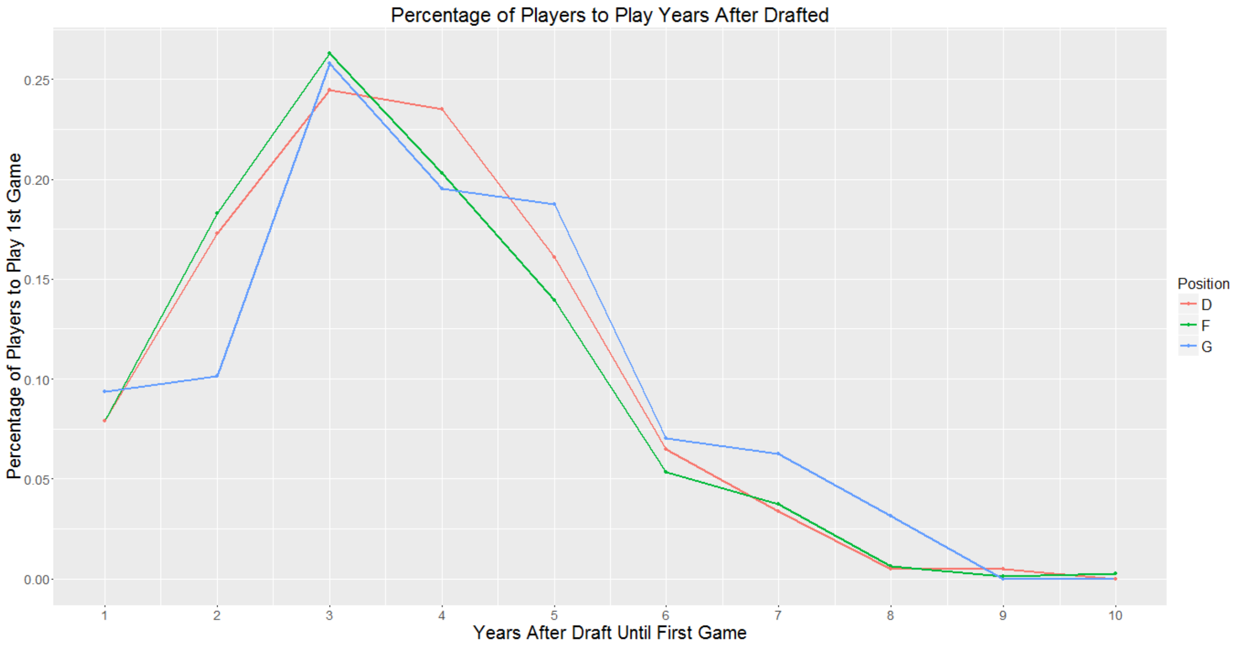

Figure 3: Proportion of Years Until First Game is Played

When a drafted player does not sign with the team that drafted him, he may re-enter the draft two years later. The re-drafted players were kept in the model since I am evaluating how a specific draft pick will perform. Since the players never signed with the original drafting teams, but have not retired, they are still capable of making the NHL; some players decide to re-enter the NHL draft while others, such as Jimmy Vesey and Justin Schultz elect to play collegiate hockey and sign with teams as an unregistered free agent.

3.0 Statistical Modelling

Survival analysis often incorporates censored data and is used to determine the length of time until an event happens. This process is translated to the NHL data by using games played as the time variable, retirement as the event variable, and right-censoring players if they are still playing in the NHL.

Using this format, I am able to model the probability that a player will retire after playing "X" number of games. The survival package in R was used for the analysis [11].

A Kaplan-Meier (K-M) estimate was fit to the data first as a non-parametric approach. Additional K-M estimates were run on subsets of players based on the rounds in which they were drafted in. This approach allowed for visual analysis of the basic data before moving on to a parametric alternative.

An Accelerated Failure Time Model (AFT) was fit to the data to predict the survival times of each drafted player. The AFT works similarly to the Cox Proportional Hazard model in that it uses a set of covariates; however, it measures the effect of the covariates on the survival time instead of the hazard. The hazard function $h(t)$ measures the instantaneous rate of retirement at a time "t" rather than the survival rate which measures the probability that a player will retire by time "t". The AFT model allows for different distributions of the survival times, T_i, and the error term, ε_i. A log-logistic distribution was chosen to model the survival times and error term since it best fit the data. Furthermore, the covariates in the AFT model are constant yet multiplicative on the time scale which allows for an acceleration factor for the hazard function so it may decrease or increase monotonically. Since there is likely to be an increase in retirements within the first few games, an acceleration factor is mostly likely relevant so the AFT model was chosen over the conventional Cox-PH model.

4.0 Results

A survival analysis approach was taken to determine whether a drafted player will play a certain number of games before they retire. The censoring method has been previously noted and was used throughout this analysis without change.

The first approach was to fit the data using the nonparametric, Kaplan-Meier approach. The survivability plot below shows that only about 45% of drafted players will survive until their first game. There is a steep drop-off in the first 20 games and the survival curve becomes relatively flat indicating that if a player reaches the 200 game milestone then their ability to continue to survive further is much more likely. Because of this effect, the accelerated failure time model is likely to provide a better fit.

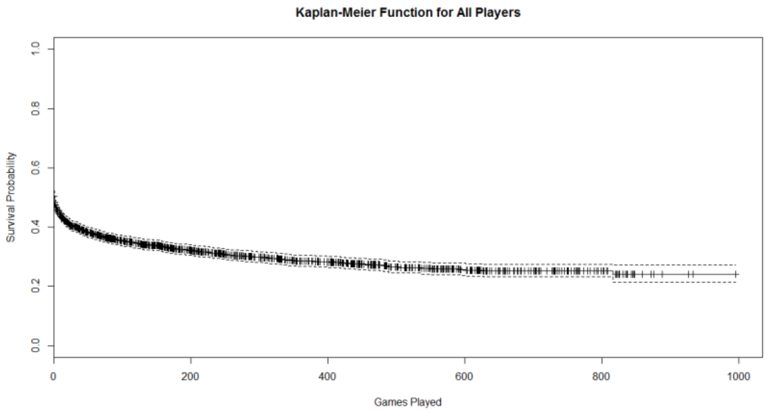

Figure 4: Kaplan-Meier Estimate Curve with Confidence Interval

The Kaplan-Meier approach was, next, repeated for players drafted in the first seven rounds. Rounds eight and nine were excluded from the model since those rounds were eliminated from the draft after 2004. By plotting each round against each other, I can determine if there is a difference between players drafted in each round. It is clear from Figure 5 that players drafted in the first round have a very high chance of playing many games in the NHL. Also, players drafted in the second round have a significantly higher probability than those drafted in later rounds. While players drafted in the third round still have a higher probability to survive than those drafted in the later rounds, the difference in probabilities between players drafted in the third and fourth round is not as large as the probability differences between those drafted in the second and third rounds. After the third round, players' probability to survive past 700 games tends to converge together. The probability to survive until game "X" is almost equal for players drafted from the fifth to seventh rounds.

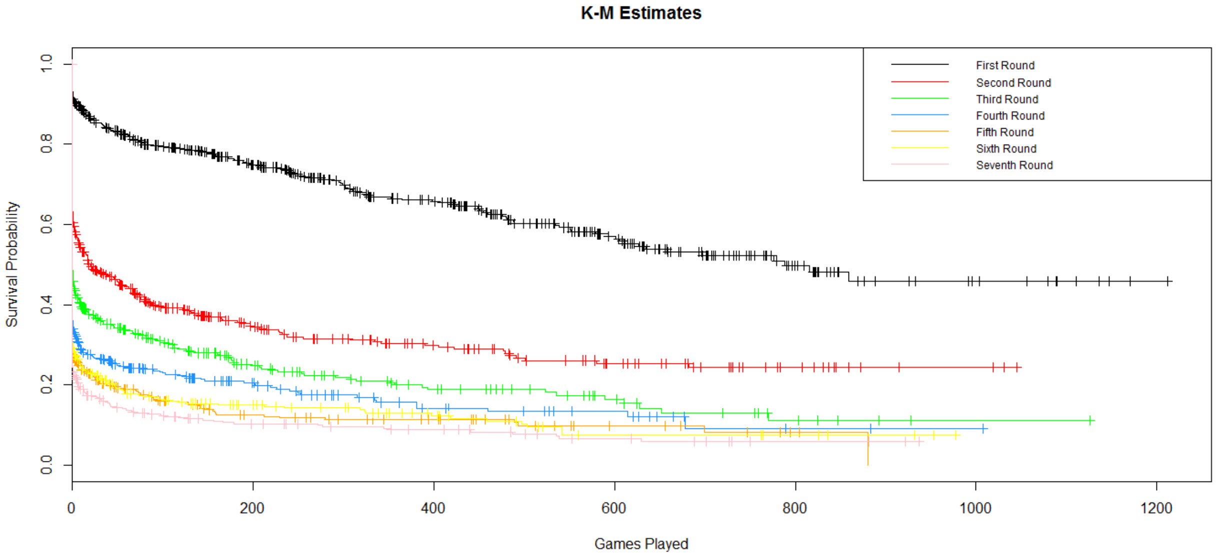

Figure 8: Kaplan-Meier Curve for Each Round

Lastly, the AFT model with a log-logistic distribution was fit to the full data set to determine the survivability of a drafted player. All available predictor variables were added into the model and were removed one by one until only those with a significance level less than α=0.10 were left. Player position was not significant nor were the associated interaction terms with player position such as the Overall Draft Order*Position and Age*Position, so these terms were dropped from the model. While player position may be important in determining whether a player would play in the NHL or whether they may take longer until they reach the NHL, it was not an important factor in determining the survivability of each player.

To test the model, a simple approach of estimating whether a drafted player will ever play an NHL game was used as a classifier. The model was run on the remaining 20% of data allotted for the test set and returned probabilities that each draft player would survive until game one. Table 1 below shows the results.

	Observed (Will Play)	Observed (Won't Play)
Predicted (Will Play)	163	56
Predicted (Won't Play)	111	371
	Accuracy: 76.18%	

Table 1: Confusion Matrix to Survive Until Game 1

The AFT model was relatively accurate with only an error rate of 23.82%. This is most likely from players drafted late in the draft who were able to beat expectations and make their way into the NHL. Overall, the model does quite well in returning an expectation of how a drafted player will perform.

5.0 Discussion

In this paper, I used survival analysis techniques to model the career expectancy of drafted NHL players. Using data from 14 previous NHL drafts, I evaluated the AFT model by predicting out of sample player survivability, namely, whether a drafted player will ever play in the NHL. A small set of predictor variables such as height, weight, age, nationality, and draft order were used and are available to anyone online. The model was tested to determine whether a draft player would play in the NHL and was 76% accurate in its prediction. While the model returned positive results, I expect this to be the tip of the iceberg in using survival analysis in NHL and sports. Further analysis on this methodology could be used to determine more accurate draft pick values for team evaluations, signings, and trades between teams. Furthermore, by removing the predictor variable of overall draft position, and applying new variables, I hope to use this methodology to predict the career longevity of undrafted players. This approach and the results show that player survivability can be modelled and could be of use to NHL teams, players, player agents, and of course, casual fans.

References

[1] Michael Schuckers. Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection. http://www.sloansportsconference.com/wp-content/uploads/2016/02/1559-Draft-by-Numbers.pdf, 2016.

[2] Josh Weissbock. Draft Analytics: Unveiling the Prospect Cohort Success Model. https://canucksarmy.com/2015/5/26/draft-analytics-unveiling-the-prospect-cohort-success-model/, May 26, 2015.

[3] Cam Lawrence. Draft Theory: On Risk and Reward. https://canucksarmy.com/2015/5/18/draft-theory-on-risk-and-reward-54e5db85-f1c1-4291-9268-830220dff34a/, May 18,2015.

[4] Gerald W. Scully. Managerial Efficiency and Survivability in Professional Team Sports. *Managerial and Decision Economics.* 15(5), Sept.-Oct. 1994, p.403-411.

[5] Robert O. Deaner, Aaron Lowen, & Stephen Cobley. Born at the Wrong Time: Selection Bias in the NHL Draft. *PLoS ONE*. 8(2), Feb.2013, p.1-7.

[6] Srdjan Lemez. The Precocity-Longevity Hypothesis Re-Examined: Does Career Start Age in Canadian National Hockey League Players Influence Length of Lifespan? *Journal of Sports Science and Medicine.* 13, 2014, p.969-970.

[7] Garrett Hohl. Draft Theory: Height Matters, But Maybe Due to Bias. https://canucksarmy.com/2015/2/10/draft-theory-height-matters-but-maybe-due-to-bias/, Feb.10, 2015.

[8] Hockey-Reference (2016) Last accessed January 16, 2017.

[9] HockeyDB.com (2016) Last accessed January 16,2017.

[10] NHL.com (2016) Last accessed March 28, 2017.

[11] Terry M Thereau (2017) gam. Survival. R package version 2.41-3. Retrieved from https://cran.r-project.org/web/packages/survival/survival.pdf.