AN ANALYSIS OF FACTORS CONTRIBUTING TO WINS

IN THE NATIONAL HOCKEY LEAGUE

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Joseph Michael Roith

In Partial Fulfillment
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

An Analysis of Factors Contributing to Wins

In the National Hockey League

By

Joseph Michael Roith

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Ronald Degges

Dr. Megan Orr

Dr. James Council

Approved:

4/5/2013
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

This thesis looks at common factors that have the largest impact on winning games in the NHL. Data was collected from regular season games for all teams in the NHL over seven seasons. Logistic and least squares regressions were performed to create a win probability model and a goal margin model to predict the outcome of games. Discriminant analysis was also used to determine significant factors over the course of an entire season. Save percentage margin, shot margin, block margin, short-handed shot margin, short-handed faceoff percentage, and even-handed faceoff percentage were found to be significant influences on individual game wins. Total goals, total goals against and takeaway totals for a season were enough to correctly predict whether a team made the playoffs 87% of the time. The accuracies of the models were then tested by predicting the outcome of games from the 2012 NHL regular season.

ACKNOWLEDGMENTS

I would like to thank Dr. Rhonda Magel for advising me throughout the process of writing and reviewing this thesis. And also for encouraging and supporting a subject matter that so deeply interests me. Finally, I want to thank Dr. Magel for all of the opportunities offered to me during my time at NDSU.

I also want to thank the rest of my committee members, Dr. Ronald Degges, Dr. Megan Orr, and Dr. James Council for the time they took and the input they contributed towards my thesis. They, along with the rest of the statistics faculty, provided the base knowledge and direct along with indirect inspiration I needed to complete this project.

I would especially like to thank my fellow graduate students for letting me use them as a sounding board and for finding the flaws in my own reasoning. Many of the thoughts presented in my thesis are a combination of multiple points of views and revelations achieved only after hours of collecting data and writing code. It would have been impossible to finish this without you guys.

DEDICATIONS

I would like to dedicate this thesis to all of my family and friends that have supported me throughout my whole college career. The decision for me to go back to school was fully accepted by everyone and never opposed by anyone.

I want to especially thank my mom Joanne Roith, you are a terrific inspiration and help me through any difficulties I encounter. Thanks to my sisters, Rachael, Becky, Katie, and Molly, the drive to complete my education is directly derived from the sheer humiliation and degradation I would have received had I not. But in fact, you guys have all helped me become the person I am now.

To the rest of my family, particularly Grandma Hauer, all grandparents not with us anymore and my godfather Pete, thank you for always being there for me and providing an extra level of support, I know how proud you all are.

Finally, thank you to all of my close friends, those of twenty years and those more recent. You are always able to take my mind off of school and make me remember what is really important in life. No matter if it has been three weeks or three years, we can always pick up right where we left off and I know it will always be that way. Thanks to all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

The National Hockey League (NHL) is a $3 billion a year industry with thirty franchises throughout the United States and Canada. Each season, clubs fight for their share of that revenue by committing millions of dollars into scouting, player development, and coaching staffs. In ice hockey, particularly the NHL, as in many other sports there is a copy-cat mentality towards what makes teams successful. Front offices for franchises will look to successful teams in recent history and try to build their own team with similar attributes, such as a strong goalie or a high priced star forward that will provide a lot of offense. Ice hockey has also been the beneficiary of an increase in available data and data collection devices in recent years. With all of the available information, it can be harder to pinpoint what makes a team successful and what areas are perhaps emphasized too much. In this thesis we develop models to explain the underlying reasons teams win and lose games.

We will show what makes a team successful over the course of an entire season, and we will also introduce two equations that predict short-term success, mainly the outcome of individual NHL games. One equation will estimate the overall probability that a given team will win the game, the other will try to predict the actual goal margin at the end of the game. To do this, sixty different variables were initially considered and narrowed down to the ones that most accurately and efficiently describe the data. We will also analyze the most important aspects towards making the playoffs as a team.

In the NHL currently, there are thirty teams that are split into two conferences, East and West. The conferences are then divided into three divisions of five teams each. All clubs play eighty-two games in a typical regular season. Every time a team wins, they receive two points that accumulate over the course of the year. Whenever a team loses in regulation play, they do not receive any points. If a team loses during overtime play, or during the round of shootouts, they will claim one point towards their total. At the end of the season, the teams are ranked by the total number of points they are able to collect and from these rankings are determined eligible for playoffs or not. First, the team with the most points in each division receives an automatic playoff bid. After that, the next five teams in each conference with the highest point total receive a playoff berth, regardless of the division they belong to. This will constitute the 16 team playoff for the Stanley Cup, or league championship.

This paper will focus on the results of regular season games and not playoffs since the circumstances that apply to the regular season are more general and have a much larger sample size to procure games. Over the course of this thesis, we will refer to teams as having a “successful” season; this will imply that they have played well enough over the course of eighty-two games to amass ample points to make the playoffs. Since the goal of every team is to win the Stanley Cup, the first logical step is to be in the postseason. Teams that do not make the playoffs are not eligible to win the championship.

It should be noted that all data was collected from 2005 to 2012. Seasons previous to these years were played under slightly different rules. After the labor strike in 2004, rules were changed with the specific intention to increase scoring in the game. Therefore any assumptions of identical distribution of outcomes from game to game are violated when using those pre-strike observations. Another labor strike that was resolved in early 2013 was primarily based in the economic foundations of the game and focused on financial agreements between players and owners. The settlement of that dispute did not result in the change of any rules regarding to game play. This suggests that all finding of this thesis can be confidently used for future games and seasons until the time that such further rule changes require additional adjustment and modifications.

Over the course of this paper, we will divide my analysis into two separate areas; the examination of the outcome of a single game, and that of the entire season. We will determine significant factors in winning a head to head matchup and also those that contribute to the overall success of a team. In the end, we will show that there are common factors for both situations that will lend to a general philosophy that can be employed for short term and long term accomplishment. We will also show that areas currently being heralded as vital aspects to winning, are overemphasized and do not add as much to achievement.

CHAPTER 2. LITERATURE REVIEW

There has been extensive research into the game of hockey. However, the majority of these explorations have been focused on goal scoring and the distribution of goal scoring. Ryder [2004a] suggests that goal scoring in the NHL follows a Poisson distribution. This results in looking at competing Poisson processes when trying to predict the outcome of the games. He shows that by breaking down the scoring in hockey into short time intervals, we can accurately predict goals, except in the last two minutes when scoring is greatly increased due to the occasional strategy of pulling the goalie for an extra attacker when a team is down by one or two goals. In the same effect, the opponents' goals can be predicted with a Poisson distribution, which is why he tries to model the outcome of games with the competing processes. This is the most theoretical approach we have been able to find and it actually does a very good job at predicting wins, more specifically, winning percentage.

Ryder takes a similar approach to this paper by focusing on individual games and also the bigger picture in terms of the whole season. If we know the average goals scored for both teams playing in a contest, using Ryder's method, we can create a chart that give the probabilities of one team winning by *z* amount of goals. We can then add these probabilities to give a total value for the likelihood of that team winning. For the entire season, goals per game and opponents' goals per game are used again. This is an extension of his previous system, calculating a table of theoretical values for goals and goals against and deriving the probability of winning. For instance, if a team averages 2.5 goals per game throughout the season, and also gives up an average of 2.5 goals per game, we can expect their winning percentage to be 0.500 on the season. However, if a team averages 4.0 goals and gives up 3.5 goals per game, they should have a winning percentage of about 0.566.

In another paper by Ryder [2004b], he looks at some more empirical methods towards predicting the outcome of games. Again in each case he looks only at goals scored and goals against, starting with a linear regression approach and moving towards more non-linear approaches. Ryder borrows some from Bill James, a notorious number cruncher that has worked primarily with baseball data. In the mid-90's James developed a Pythagorean relationship between wins and runs scored and runs allowed. The most logical analog for hockey then would be to use goals and goals against, which is exactly what Ryder does. The equation shows the general form of his results using Goals For (GF), Goals Against (GA), Goals For per Game (GFg), and Goals Against per Game (GAg).

$$P(\text{Win}) = GF^{E} / (GF^{E} + GA^{E})$$

$$\text{Where } E = (GFg + GAg)^{0.458}$$

With this formula, Ryder was able to model games in the post WWII NHL with an R-square of 0.941 [2004a].

These are very strong results that accurately reflect reality. However, it is our view that while this research is based in solid theory and the outcomes are obviously powerful; we do not gain much insight into the game by limiting ourselves to only looking at goals scored and goal against. The unmistakable strategy for winning at hockey is to score more goals than the opponent. Ryder's formulae gives us a much better understanding of the extent to which more scoring will increase the likelihood of winning, but it does nothing to further our tactics towards achieving those goals. To be fair, Ryder admits to only setting out to laying down a more theoretical approach towards describing ice hockey. But one of the aims of this thesis is to go beyond goal information, specifically for individual games where predicting how many goals will be scored is a much harder task. We want to find out if there are more subtle indicators that will lead to the success of a team.

Thomas [2007] takes a different approach towards the prediction of hockey games. While he agrees that goals follow a Poisson distribution, he is more interested in the simulation of games to calculate results. Once we know how often goals are scored and the probability of them occurring, we can simulate the probability of a certain team winning throughout different moments in the game for separate instances. For example, through simulation Thomas found that the probability of a team winning when leading by two goals with forty minutes remaining is 0.8077. The probability of winning when leading by one goal with one minute left is 0.9461. You can basically find the probability for any such situation the game may present. These results from simulation are very close to the real world results from hockey games.

While this method also only looks at the goal scoring figures, we believe it provides more useful information to the coach and team relating to what kind of strategies should be used in certain situations. Now, teams can have tangible numbers to look at when they are down by two goals with twenty minutes left. If they score the next goal their chance of winning increases by 10%, however if they give up the next goal, their chances of winning are almost 0%. This strategy based analysis is more in line to what we would like to accomplish. A more delicate way to look at the game than simply saying we need to score more often and not let the other team score so much.

An even more inventive method is discussed by Thomas [2006] looking at the Harvard ice hockey team. Here, Thomas makes the argument that hockey can be described as a continuous time semi-Markov process. He separates the game into 19 distinct states such as; offensive team with the puck in defensive zone, defensive team with the puck in the offensive zone, faceoff at center ice, defensive takeaway, and so on. From there, the expected number of goals scored in each state is calculated as time increases. The expected value of goals is much higher for each state as the time in that state increases, as soon as the situation shifts to a different state, the time is reset. It is noted that not all states are equal. At 40 seconds, the expected value of goals while in the offensive giveaway state is higher than the expected number of goals 40 seconds into a defensive possession state.

Again, the implications here are more strategy based. How do different aspects of the game differ considering goals? In fact, Thomas does actually compare common defensive plays to see which result in more frequent scoring, the "dump and chase" strategy is superior to "carry-in" [2006]. We believe this lends itself well to the NHL despite being a smaller sample of an intercollegiate team. The fact that an abstract scheme can be quantified shows that there are alternative methods towards evaluating team and player performance.

One thing that gets a lot of attention in all sports is the magnitude of importance assigned to offense and defense. It has long been the consensus from analysts and coaches that defense wins championships. Moskowitz and Wertheim [2011] disagree with that statement. They looked at multiple sports, including hockey, and tried to determine if teams who were ranked as a top defensive team won championships more often than those who were ranked as a top offensive team. In every sport they looked at, there were just as many offensive teams winning as there were defensive. The main point here though is that teams were categorized based only on rankings from the seasonal performances, and only playoffs and championships are considered. We still believe that in a game by game setting, defense has a more important role and that if a team would like to make the playoffs, having a defensive mindset during the season will be a better way to achieve that.

While there are many papers and books that seek to explain a game that is inherently hard to explain, they are mainly focused on points and goals. This thesis will look at more subtle aspects of the game that may be indicators of a style of play that leads to more success along with goals in some cases. If these indicators lead to more success, we can assume that it means they are leading to more goals being scored, less goals being given up, or a combination of the two to some degree. The extent to which those variables contribute towards success will be discussed later in the paper.

CHAPTER 3. DESIGN OF STUDY

The purpose of this research is to analyze what variables may influence winning in the NHL. In particular, what aspects of the game should be emphasized to win an individual match and to be successful over the course of a season. For example, is it more important for a team to take more shots and try to push play in the offensive zone in order to score more goals? Or, is blocking opponents' shots a better approach to win a game? Another example would be to look at overall scheme for the season. If a coach teaches an aggressive style of play that scores a lot of goals, but also gives up more goals, will that lead to a better record at the end of the season than one that preaches defense?

3.1. Seasonal Analysis

This study examines the effects of variables on overall season success, whether or not a team makes the playoffs. Twenty-five initial variables were selected for each of the thirty teams over seven seasons. These variables are commonly collected data throughout the league and can be accessed on the NHL website. Among the data obtained were; total goals scored, total goals against, total shots for/against, penalty minutes, and power play results. A full list of initial variables is listed in Appendix A. The data collected consisted of the season totals for those twenty-five initial variables for all thirty teams over the course of seven seasons, from 2005 – 2012. This results in 210 total observations for each variable.

The goal for this research is not to so much to predict which teams will make the playoffs, but to analyze what components are most important to making them. Therefore, a stepwise discriminant analysis approach was used on the data for playoffs. First, for teams that made the playoffs, this was coded as a "1", and secondly for teams that did not make the playoffs, notated as "0". All data used was found to have a multivariate normal distribution, and all variables chosen individually had univariate normal distributions. A quick check of histograms for all variables also confirms approximately normal characteristics.

The stepwise procedure will then choose those variables that contribute the most towards making the playoffs, and conversely not making them. The selection criterion chosen was an entry alpha level of 0.25, and a significance level of 0.20 to stay in the model. With the discriminant function, we will be able to determine the order and degree of influence each indicator has. Classification analysis will also be performed on the data. This will give a better idea of how well the chosen variables can assign teams as playoff contenders or not. One thing we can take advantage of in this case is the fact that each year, there is a fixed number of teams that make the playoffs. Sixteen out of the total thirty teams will eventually make the playoffs, so we will assign a prior probability of 0.5333 for the made playoffs level, and 0.4667 for the missed playoffs level. This will give a more accurate report of the error rate in classifications.

From these analyses, we hope to pinpoint the main factors towards achieving playoff status for teams over the long term period of an NHL season. We can then compare the results found with those acquired performing tests on the short term success and failure of individual games.

3.2. Game Analysis

The majority of the work done here is based on trying to predict the winner of a single game. The data for the initial sixty variables were collected from individual games from the 2009-10 season and also the 2010-11 season. Each season was divided into four quarters to ensure an even sampling of the games. The first quarter of the season is determined to be one of the first twenty games played by that team. The second quarter would then be considered games 21-40, 41-60 for the third, and games 61-80 for the fourth quarter. Every team in the league had one game recorded for each quarter, for both seasons. This results in eight total games for each team, or 240 total observations. A 1-in-k systematic sampling approach was used to determine the games selected. A random number from 1 to 20 was chosen for each system and that game was selected for each quarter. For the 2009-10 season, the random number generator produced 16. Therefore, the 16th, 36th, 56th, and 76th games were collected for each team. In the event that this system requires a game to be recorded twice, the first prior game was used. When both the initial game and the first prior game were already recorded, the next available game was taken. For the 2010-11 season, the random number generator produced 5, so the 5th, 25th, 45th, and 65th games were used in that season. Cases where replicate games would have been recorded were treated the same way as the previous season. All statistics documented were taken from the official scorer's game log [NHL.com].

The analysis of game data is further divided into two methods. One is to determine the probability that the team of interest will win, and another that will try to predict the actual goal margin at the end of the game. The first method will be achieved through a logistic regression analysis. In this case, the variable win has two levels; "1" will indicate that the team of interest has won that game, while "0" means that the team lost that game. Ties at the end of regulation play and overtime will be coded as "1" since shootouts do not follow regular in game rules. This will be our dichotomous response variable. Stepwise selection was used to determine the most efficient model, an entry alpha level of 0.25 was used along with a significance level of 0.20 required to stay in the model. Alternative methods such as receiver operating characteristic curves which measure the ratio of false positives and true positives were also used to simplify the model. The final model will be in the form of the equation below where X will be the design matrix of selected variables and β will be a vector of coefficients.

$$\text{P(Win)} = \frac{e^{X\beta}}{(1 - e^{X\beta})}$$

To create a goal margin model, the least squared means multiple regression method was used with goal margin as the dependent variable. Again, stepwise selection was performed to determine significant independent variables with criteria the same as was used in the logistic approach. R-squared was assessed for potential models to select one that was simple and also effective in explaining the variation in the data.

Once the models for each method were selected, another data set was collected in order to test the accuracy of the models. This was a much smaller group since the purpose is to simply verify that the selected variables do an adequate job predicting the outcomes. Games were selected from the 2011-12 season, outside of the timeline for the original data set the models were based on. Sixty games were randomly collected; two for each team, the 23rd and 48th games which were selected randomly. Through this data set, we should be able to confirm the selection of the variables for both models by applying them using the statistics from games that had already occurred. If the models are good, then they should perform well in predicting the outcomes of these games.

Finally, a third data set was collected to test the models ability to predict games that have not occurred. To do this, sixty target games were chosen. The data from these games were not collected; rather the averages for the significant variables were taken over the previous three games. This was done for both teams of interest and in the case of marginal statistics the difference of the averages for both teams was used in the models. Here the point is to see if the models can be used to predict the outcome of games solely based on team trends up to that date, not considering the information from the game after the fact. The predictive capability of the models will be compared to a basic method of choosing a winner such as picking the home team, choosing the team with a better record, and comparing it to a sports handicapping website. We will be able to tell whether the models can outperform these common benchmarks for prediction.

CHAPTER 4. RESULTS

4.1. Seasonal Analysis

The results for the seasonal analysis of the data did include the goal information, similar to the previous studies mentioned before. This is because, while still difficult, you can get a more accurate idea of how many goals a team will score over the course of a season compared with trying to predict how many goals they will score in a single game. Through the stepwise discriminant method as described earlier, the most influential season factors contributing to making the playoffs are total goals scored, total goals given up, and the total number of takeaways for the season. Table 4.1 shows the stepwise selection process along with the partial R-squared vales associated with each of the variables.

Table 4.1. Stepwise Selection for Seasonal Model

Step	Variable Entered	Partial R^2	F – value	Pr > F
1	Goals Against	0.3654	119.77	<0.0001
2	Goals	0.3045	90.64	<0.0001
3	Takeaways	0.0178	3.74	0.0546

From the subset selection procedure, we can now look at the linear discriminant functions for making and not making the playoffs based on the standardized variables. Table 4.2 gives these results. As you can see in both cases, the magnitude of scoring goals is less than that of allowing goals. This would lead us to believe that it is more important to give up fewer goals than it is to score an abundance of them in order to make the playoffs. Takeaways, while shown to be significant in the previous step, are still not as important as the two goal statistics.

Now with the significant factors and also their discriminant functions, we can perform a classification analysis to see how well the data can be grouped using our new indicators. To refit the model the holdout procedure was used, that is each observation was classified using all remaining observations excluding itself. In addition, the prior probabilities stated in Chapter 3 were used. We will use the covariance matrices for classification and assume that they are equal, since this assumption has been checked and verified. It should be noted that using the correlation matrices will provide similar results. The error rates for classification can be seen in Table 4.3. The outcomes show a misclassification rate of 0.1286. This means that teams were correctly grouped as a playoff team or non-playoff team 87.14% of the time when only considering their season totals for goals scored, goals against, and takeaways. Furthermore, only 8.93% of teams classified as missing the playoffs actually made them, while 17.35% of teams were grouped as a playoff team, but did not succeeded in reaching that goal.

Table 4.2. Linear Discriminant Functions

Variable	Did Not Make Playoffs	Made Playoffs
Constant	-1.50183	-1.19494
Goals Scored	-1.05916	0.92677
Goals Against	1.50968	-1.32097
Takeaways	-0.21979	0.19231

Table 4.3. Crossvalidation Classification Table

Playoffs	Did Not Make	Made	Total
Did Not Make	81	10	91
Made	17	102	119
Total	98	112	210
Error Rate	0.1735	0.0893	**0.1286**
Prior Probability	0.4667	0.5333	

4.2. Game Analysis

4.2.1. Win Probability Model

We will first look at the results from the logistic regression analysis. Appendix B lists the 62 initial factors considered. From the stepwise selection process shown in Table 4.4, the model came back with eight variables; save percentage margin, shot margin, even strength faceoff percentage, short-handed faceoff percentage, block margin, short-handed shot margin, power play time margin, and giveaway margin. The last two variables entered; power play time margin and giveaway margin will actually be removed in order to simplify the model. This is justified by noting the p-values of these parameters are higher than those of the other variables which are shown in Table 4.4. The final model will contain the six variables, save percentage margin, shot margin, even strength faceoff percentage, short-handed faceoff percentage, short-handed shot margin, and block margin. Diagnostics for the model fit are very good, a Hosmer-Lemeshow goodness of fit test, where the null hypothesis is that the model fits the data and the alternative hypothesis concludes that there is not a good fit, returns a p-value of 0.9431 meaning we cannot reject the null hypothesis. In addition to this, the receiving operating characteristic curve, seen in Figure 4.1, has an area of 0.9888 under the curve of the current model, meaning there are few false positives and many true positives. The significance of the intercept was negligible with a p-value of 0.1218, so it was removed. This makes sense because with all variables being equal for both teams, the probability of winning should be 50% and an intercept would affect that percentage.

Table 4.4. Stepwise Selection for Win Probability Model

Step	Variable Entered	Score χ^2	Pr > χ^2
1	Save Percentage Margin	121.7044	<0.0001
2	Shot Margin	40.7660	<0.0001
3	Block Margin	7.4259	0.0064
4	Short-handed Faceoff Percentage	7.0031	0.0081
5	Short-handed Shot Margin	8.6487	0.0033
6	Even-handed Faceoff Percentage	3.7971	0.0513
7	Power Play Time Margin	3.3372	0.0677
8	Giveaway Margin	2.4351	0.1186

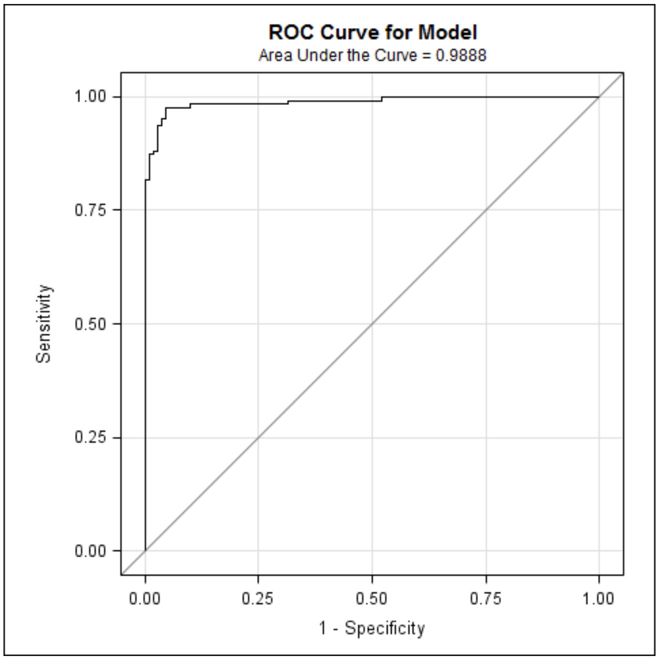

Figure 4.1. ROC Curve for Win Probability Model

Now we can present the parameter estimates along with their odds ratio estimates, shown in Table 4.5. These estimates will be used in testing the predictive ability of our new model. To do this, we look at how the model performs on predicting games that have already happened, but are outside of the original data set. As stated in the previous chapter, sixty games were selected from the 2011-12 season and the outcome was predicted from the six factors in the model. Of the initial sixty games, only fifty-two were used because eight games were decided in a shootout. Shootout games occur when there is a tie after the original 60 minutes of play and after 5 minutes of sudden death overtime. Since shootouts do not consist of team play and only one-on-one situations between player and goalie, these games were not considered. Using the model, 51 out of 52 games were correctly predicted using the in game statistics from those events. While this is not necessarily surprising, it does assure us that the model works for games other than those that it was formulated from which is crucial for our next step in predicting.

Table 4.5. Parameter Estimates and Odds Ratios

Variable	Parameter Estimate	Standard Error	Odds Ratio Estimate
Save Percentage Margin	0.8355	0.1482	2.306
Shot Margin	0.2691	0.0512	1.309
Block Margin	0.1511	0.0568	1.163
Short-handed Faceoff Percentage	0.0437	0.0161	1.045
Short-handed Shot Margin	-0.4814	0.2070	0.618
Even-handed Faceoff Percentage	-0.0325	0.0155	0.968

The next step is to try and predict the outcome of games that have not yet occurred. To do this, we will look at the averages of the in game statistics from the previous three games prior to the contest of interest for both teams. Initially, the averages from the previous five, four, three, and also the medians were considered, but three games seems to provide the most accurate information. For the marginal variables, save percentage margin, shot margin, block margin, and short-handed shot margin, the difference was taken from the average number of shots for each team over the previous three games. Again sixty initial games were of interest, and we will use the model that gives a probability that the team of interest will win. If the probability comes back as greater than 0.5, we will consider that as the model telling us it favors that team as winning the game. If it comes back less than 0.5, we predict the team will lose the game.

Shootout wins and losses in this case will be considered in the same class as wins. Losses will only be deemed a loss if the team is defeated during regulation time or overtime. With that information, the outcome of a game was correctly predicted for 39 out of the 60 games, or a success rate of about 65%. If we consider a few basic scenarios of how we could choose the winner of a hockey game, a few obvious ones come to mind; choosing the home team, choosing the team with the better record, and looking at a sports handicapper. The home team in the NHL has won about 55% of the time over the last five seasons. A simple one-sample proportion test will show that if we consider 55% our null hypothesis, 65% is higher with marginal significance and a p-value of 0.052. When the team with the better record was chosen to win for the sixty games, the success rate was also 55%, meaning our model was marginally better here as well. A quick look on the website covers.com, which supplies historical handicapping data, indicates they only predicted the correct winner of those games 52% of the time, which means the model's accuracy of 65% is significantly higher with a p-value of 0.013.

4.2.2. Goal Margin Model

The second model that was derived will try to estimate the actual goal margin at the end of the game. Once more stepwise selection was used with an entry level of alpha equal to 0.25 and a significance level in the model of 0.20 to stay. The initial results are listed in Table 4.6. While there are seven variables chosen from this procedure, only the first two will be used in the goal margin model. When you consider the partial R-square associated with each variable, save percentage margin and shot margin have the two largest values and all other factors are considerably lower. In addition to this, the cumulative R-square for the model from these two variables is over 0.93 and does not significantly increase when the others are included. To simplify the equation, only save percentage margin and shot margin will be used from here on. Table 4.7 gives the parameter estimates which are very similar to the estimates calculated with all seven variables originally selected. The intercept proved also to be non-significant with a p-value of 0.3641 and is not included. This can be expected since two teams that have the exact same save percentage and the same number of shots should have a goal differential of zero.

Table 4.6. Stepwise Selection for Goal Margin Model

Step	Variable Entered	Partial R^2	F – value	Pr > F
1	Save Percentage Margin	0.8204	1018.7	<0.0001
2	Shot Margin	0.1160	405.15	<0.0001
3	Power Play Goal Margin	0.0019	6.87	0.0094
4	Even Shot Margin	0.0016	5.88	0.0161
5	Hits Margin	0.0010	3.89	0.0497
6	Power Play Time Margin	0.0009	3.29	0.0711
7	Missed Shot Margin	0.0007	2.64	0.1056

Table 4.7. Parameter Estimates for Goal Margin Model

Variable	Parameter Estimate	Standard Error
Save Percentage Margin	0.27101	0.00479
Shot Margin	0.09043	0.00465

The final model has an adjusted R-squared value of 0.9306. We will now look at how well the goal margin model can predict the winner of a hockey game after it has already occurred. The same sixty games were taken from the 2011-12 season as above, the 23rd and 48th games for each team. If our model results in a positive number, then we are predicting the team of interest will win, or lose if the model returns a negative value. Again, we will not consider the shootout games for prediction using stats from a game that has already occurred. So out of 52 games, the model correctly predicted the winner 52 times. The model had a 100% accuracy and only missed the actual goal margin by an average of 0.43 goals per game. This is a very encouraging sign as we move forward to test the model on games that have not occurred.

Taking the same sixty games of interest that were used to test the predictive power of the logistic model, the three games prior averages were taken for save percentage and shots for each team. Once again in this case, shootout games will be grouped with wins since we do not expect to have an output of exactly zero in the model. Forty out of the sixty games were correctly predicted. This is a little better than the logistic model presented earlier and still significantly better than picking winners based off of home teams, better records, or the handicapping website. It should be noted that when using the three prior games, the goal margin model did well at choosing the winner, but was not very accurate when considering what the final goal margin of the game actually was. We believe this can be attributed to the fact that while looking at recent games can give a better impression of how a particular team is playing, there are so many other random factors that influence any given game. The best one can hope for is to get a better understanding of how a team is trending, whether they seem to be playing better or not.

CHAPTER 5. CONCLUSIONS

The goal of this thesis was not to just be able to predict the outcome of ice hockey games, but rather to get a better understanding of the factors that contribute towards winning. If you look at what significant variables were used to determine whether a team was successful over the course of an entire season and made the playoffs, you see that goals scored, goals against and takeaways can give you a very good idea. A further look into the results will also show which of these has a more significant value. The determinant function indicates that goals against has a larger magnitude than goals scored. This would lead us to believe that it is more important for a team that is striving to make the postseason to keep their opponents from scoring an abundance of goals. It boils down to a comparison of strategies, a strong defensive team that may not score a lot of goals may have an advantage in making the playoffs over a team that has a lot of offensive capabilities, but is lacking in a good defensive scheme. Scoring goals is still important obviously because that is how you win, but the evidence suggests that over the long course of a season, preventing scoring is more vital. In addition to this, the fact that takeaways are an indication of a club’s defensive mindset gives further evidence that defense is more crucial than offense, even if it is less glamorous or publicized.

Even over the short term, defense shows it has a larger impact than offense. When trying to predict the winner of an individual game, we can see that above all else save percentage is the most important factor. For every percentage point better your team is at stopping shots, the probability of winning increases by a factor of 2.3. Shots are also important because they eventually lead to goals, but not to the same degree as stopping shots. One subtle variable that is included is short-handed faceoff win percentage and also the fact that power play faceoff win percentage was left out. Most experts, including Jones [2012] will agree that the benefit of winning a faceoff during a power play is so the players on the ice can set up their positions and really take advantage of the extra skater. In other words power plays generally lead to more scoring especially if you can control the puck. However, when a team is short-handed they rarely score any goals and winning possession of the puck off of a faceoff is even more critical in order to keep the opponent from setting up plays and having that advantage. It says a lot just from the fact that power play faceoff win percentage was not selected for the individual game model.

Furthermore, when you look at the goal margin model, save percentage margin has a larger influence than shot margin. In addition to this, the fact that all of these models can accurately describe the outcome of games should be even more proof of their worth and that they should be emphasized throughout the league. Prediction for future games may be better than guessing or taking a rudimentary approach, but it still seems to hover around that 65% mark. We believe that a team that really coaches to the points illustrated in this paper and creates a philosophy towards defense could greatly improve that prediction rate along with the success of the organization.

The applications go beyond coaching and performing as well though. Front offices and decisions makers can put a concrete number on how they value certain skillsets. Many times a franchise will have to choose between trying to sign different free agents. In today’s NHL, a flashy player that can score a lot of goals certainly seems tempting to have on your team, but could a mediocre player that has a much more solid all-around game including defense be a better choice? The models presented here can give a starting point to try and empirically define the two players to make a more informed decision. Maybe the offensive player is so good he overcomes any deficits in defense, but perhaps the mediocre player is just good enough on both offense and defense that he is a much better deal, especially since the super-star is likely to be overvalued in the market and would demand a lot more money.

In this sense, basically any aspect of the game can be assisted with the extra knowledge and correct interpretation. Entire teams that are built from the draft, trades, and by signing free agents should be, and currently are thoroughly looked over and scrutinized as to whether they fit into the team’s core ideas. Nevertheless you have to wonder sometimes with the overabundance of data that is collected and available these days, whether people are focusing on the areas and preaching parts of the game that just are not essential to winning. All of the figures are great and bring us into a deeper level of the game, but perspective is also necessary when you consider the millions of dollars that are thrown around and spent from year to year on these players. In the end, any coach from bantam to collegiate to the NHL will tell you that defense wins championships, but the ability to quantify and prove that statement should be a much more powerful prospect and create a better game for the future.

REFERENCES

Jones, Joe. [2012]. “The Importance of Faceoffs”. Seeing the Ice, www.seeingtheice.wordpress.com.

NHL Scores. [2005-2012]. Retrieved February 28, 2013, from NHL.com: http://www.nhl.com/ice/scores

Moskowitz, Tobias, and Wertheim, Jon L. [2012]. *Scorecasting: The Hidden Influences Behind How Sports Are Played and Games Are Won*. Crown Publishing Group. ISBN-13: 9780307591807.

Ryder, Alan. [2004a]. “Win Probabilities: A tour through win probability models for hockey”. Hockey Analytics, www.hockeyanalytics.com.

Ryder, Alan. [2004b]. “Poisson Toolbox: A review of the application of the Poisson Probability Distribution in hockey”. Hockey Analytics, www.hockeyanalytics.com.

Thomas, Andrew C. [2006]. “The Impact of Puck Possession and Location on Ice Hockey Strategy”. *Journal of Quantitative Analysis in Sports*, Vol 2, Issue 1.

Thomas, Andrew C. [2007]. “Inter-arrival Times of Goals in Ice Hockey*”. Journal of Quantitative Analysis in Sports*, Vol 3, Issue 3.

APPENDIX A. LIST OF INITIAL SEASONAL VARIABLES

- Blocked shots
- Faceoff win percentage
- Five on five goals/goals against ratio
- Giveaways
- Goal plus/minus
- **Goals**
- **Goals against**
- Hits
- Missed shots
- Penalty kill percentage
- Penalty minutes
- Power play goals
- Power play goals against
- Power play percentage
- Save percentage
- Shot plus/minus
- Shots against
- Shots for
- **Takeaways**
- Winning percentage when leading after 1st period
- Winning percentage when leading after 2nd period
- Winning percentage when outshooting opponent
- Winning percentage when outshot by opponent
- Winning percentage when scoring first
- Winning percentage when trailing first

Bold indicates a selected variable

APPENDIX B. LIST OF INITIAL GAME VARIABLES

- Assist margin
- Assists
- Assists against
- **Block margin**
- Blocks
- Down 1 goals against
- Down 2 goals against
- **Even-handed faceoff win percentage**
- Even-handed goal margin
- Even-handed goals
- Even-handed goals against
- Even-handed shot margin
- Even-handed shots
- Faceoff losses
- Faceoff win percentage
- Faceoff wins
- First goal scored
- Giveaway margin
- Giveaways
- Goal margin
- Goals
- Goals against
- Hits
- Hits margin
- Missed shot margin
- Missed shots
- Opponent Blocks
- Opponent even-handed shots
- Opponent giveaways
- Opponent hits
- Opponent Missed shots
- Opponent power play shots
- Opponent save percentage
- Opponent short-handed shots
- Opponent takeaways
- Penalty kill percentage
- Penalty minutes
- Power play faceoff win percentage
- Power play goal margin
- Power play goals
- Power play goals against

- Power play percentage
- Power play shot margin
- Power play shots
- Power play time margin
- **Save percentage**
- Save percentage margin
- Second goal scored
- **Short-handed faceoff win percentage**
- Short-handed goals
- Short-handed goals against
- **Short-handed shot margin**
- Short-handed shots
- **Shot margin**
- Shots against
- Shots for
- Takeaway margin
- Takeaways
- Time on power play
- Up 1 goals
- Up 2 goals

Bold indicates a selected variable

APPENDIX C. SAS CODE

```
proc import datafile='C:\Users\Joe Roith\Desktop\NDSU Stats\NHL data\game
data\nhlgamedata.xlsx' out=nhlgames replace;
run;

proc import datafile='C:\Users\Joe Roith\Desktop\NDSU Stats\NHL data\game
data\nhlgamedata.xlsx' out=nhlstandard replace;
run;

proc import datafile='C:\Users\Joe Roith\Desktop\NDSU Stats\NHL
data\NHL.xlsx' out=nhl replace;
  run;

* Seasonal Analysis *;

proc stepdisc data=nhl slentry=0.25 slstay=0.20;
  class playoffs;
  var season goals goals_against ppga--bks faceoffwinperc shotpls_mns;
run;

proc discrim data=nhlstandard crossvalidate pool=yes;
  class playoffs;
  var goals goals_against takaw;
  priors '0'=0.4667 '1'=0.5333;
run;

* Game Analysis *;

proc logistic data=nhlgames outest=betas covout plots=all;
  model win(event='1')= home firstgoal secondgoal ppperc--pkperc fowins--
shfoperc pptimemargin shotmargin--savemargin evenshotmargin--blkmargin
      / selection=s slentry=0.25 slstay=0.20 details lackfit scale=none
      rsquare;
  output out=pred p=phat lower=lcl upper=ucl
  predprob=(individual crossvalidate);
run;

proc logistic data=nhlgames covout plots=all;
  model win(event='1')= evenfoperc shfoperc shotmargin savemargin
shrtshtmargin blkmargin
      / details rsquare lackfit noint;
  output out=pred p=phat;
run;

proc reg data=nhlgames plots=all;
  model goalmargin = home firstgoal secondgoal ppperc--pkperc fowins--
shfoperc pptimemargin shotmargin--savemargin evenshotmargin--blkmargin
      / selection=stepwise aic bic rsquare slentry=0.25 slstay=0.20;
run;

proc reg data=nhlgames;
  model goalmargin = savemargin shotmargin
      / noint aic bic rsquare;
  plot residual.*predicted. / cmallows  h cookd;
run;
```