2018

Predicting the Current Season's Win Percentages in the National Hockey League Using Data from the Previous Season: Can Game-Level Data Help?

Suyash Sharma

Claremont McKenna College

Predicting the Current Season's Win Percentages in the National Hockey League Using Data from the Previous Season: Can Game-Level Data Help?

Submitted To

Prof. Darren Filson

By

Suyash Sharma

For

Senior Thesis

Spring 2018

23rd April, 2018

This page is intentionally left blank

Acknowledgments

I want to thank everyone who helped me in completing my thesis. I would specially like to thank Professor Darren Filson. This thesis would not have been possible without his help and support. I would also particularly like to thank Professor Yong Kim. His recommendations greatly helped me in completing my thesis. I would also really like to thank my parents and my relatives (Mr. Suresh Lodha and Mrs. Madhu Lodha) for their unconditional love and support.

Table of Contents

Abstract

Researchers have tried to predict winning percentages for the National Hockey League (NHL) teams based on their performance in the previous seasons. However, these predictions have not been very accurate. This study hypothesizes that incorporating pair-wise game-level data with season-level data can be useful in improving the prediction of a team's win percentage. Season-level data and pair-wise game-level data from the 2005-2006 season to the 2015-2016 season has been used to predict winning percentages for the pairs in each of the following seasons. Significant results were not found for any of the pair-wise game-level data variables except for two pair-wise variables. This helps establish the idea that including more granular information does not necessarily increase the predictive power of models. One of the pair-wise variables found to be significant (at the 10% level of significance) was when high goal differential was observed in the interaction term between high goal differential for a team in its home games against the other pair-wise team and the goal differential for a team in its home games against the other pair-wise team. This provides marginal support for the claim that extreme game-level outcomes from the previous season can help in predicting a team's win percentage in the following season. Another pair-level variable found to be significant (at the 5% level of significance) was when high goal differential was observed and at least 4 games played was not observed in the interaction term between at least 4 games played against the other pair-wise team and high goal differential for a team in its home games against the other pair-wise team. This suggests that only in the games a team plays outside its own division, the extreme game-level data helps in predicting a team's win percentage in the following season.

I. Introduction

A number of studies have tried to predict winning percentages for NHL teams based on their performance in the previous season. None of these studies have made use of game-level data in order to predict win percentages for teams. Also, these studies have not been able to accurately predict win percentages for teams. This study makes use of both season-level data and pair-wise game-level data in a season in order to predict win percentages for the pairs of teams in the following seasons. I hypothesize that using pair-wise game-level data along with season-level data will improve the prediction of win percentages for NHL teams.

The study employs both linear and non-linear regression models to predict win percentages for NHL teams. It uses the season-level data and pair-wise game-level data from the 2005-2006 season to the 2015-2016 season in order to predict win percentages for the pairs of teams in each of the following seasons. It tests a number of models to see if incorporating pair-wise game-level data along with season-level data helps better predict win percentages for NHL teams. All of the models test for robustness. The 2004-2005 season was a lockout (the entire season was cancelled) and the rules changed then and have remained stable since, so starting from the 2005-2006 season was ideal. Logically, it makes sense that including pair-wise game-level data along with the season-level data for teams should improve the prediction of their win percentages. This is because now we are not only taking into account how the team performed against all the other teams in the previous season, but are also specifically looking at how a particular team performed against another team in the previous season in order to predict this pair's performance in the next season.

I did not find any significant results for the pair-wise game-level data variables except for two pair-wise variables. This helps establish the idea that including more

granular information does not necessarily increase the predictive power of models. One of the pair-wise variables found to be significant (at the 10% level of significance) was when high goal differential was observed in the interaction term between high goal differential for a team in its home games against the other pair-wise team and the goal differential for a team in its home games against the other pair-wise team (high_home_goal_diff#c.pair_home_goal_diff). This provides marginal support for the claim that extreme game-level outcomes from the previous season can help in predicting a team's win percentage in the following season. Another pair-level variable found to be significant (at the 5% level of significance) was when high goal differential was observed and at least 4 games played was not observed in the interaction term between at least 4 games played against the other pair-wise team and high goal differential for a team in its home games against the other pair-wise team (at_least_4_gp#c.high_home_goal_diff). This suggests that only in the games a team plays outside its own division, the extreme game-level data helps in predicting a team's win percentage in the following season. This study also concluded that pair-wise game-level data won't help predict aggregate win percentages for teams as it did not help predict game-level outcomes in the following seasons.

Unlike Laffey and Ames (2016), this study does not use statistics split over multiple scenarios such as when the team is leading, trailing, the team is shorthanded etc. as well as statistics such as average player ages to estimate regular season and playoff wins for NHL teams. Including this information may help in improving win percentages as teams adjust their strategies depending on the situation. Additionally, unlike Schulte et al (2017), I will not make use of data that includes location information about where an action took place (except for home versus away games). Including location information about where an action took place may have been of help in

predicting game outcomes and thus in predicting winning percentages for teams. Furthermore, the study does not include information on the quality of the players playing in a game for both the teams. In ice hockey players often get injured. If the star player of a team gets injured and thus is unable to play a few games, then this could have direct consequences on the results of those games for that team.

This paper will proceed as follows: first, it will provide general background information about the NHL, then it will review the previous literature on ice hockey and layout the data used in this study. Next, the methodology used to incorporate the pair-wise game-level data with the season-level data in order to better predict win percentages for NHL teams will be discussed. Discussion of the results shall follow along with a brief discussion of further research which can be done in this field.

II. Industry Background

The National Hockey League (NHL) is a professional ice hockey league in North America. As of now, the NHL consists of 31 teams: 24 in the United States of America and 7 in Canada (NHL). The NHL has divided the teams into two conferences namely the Eastern Conference and the Western Conference (NHL). Both of these conferences are further divided into two divisions (NHL). The Eastern Conference is divided into the Metropolitan Division, which consists of eight teams and the Atlantic Division, which also consists of eight teams (NHL). The Western Conference is divided into the Central Division, which consists of seven teams and the Pacific Division, which consists of eight teams (NHL). There were actually 30 NHL teams for about seventeen years until the league decided to expand by adding the Vegas Golden Knights in 2017 (NHL).

The NHL season is separated into a postseason (the Stanley Cup playoffs) and a regular season (from early October through about middle of April). Every team plays a total of 82 games (41 away games and 41 home games) during the regular season. Each team in the Eastern Conference plays four games against each of the seven teams in its own division. So each team in the Eastern Conference ends up playing twenty-eight games in its own division. Each team in the Eastern Conference also plays three games against every team in the other division of its conference. So each team in the Eastern Conference plays twenty-four games against the other eight teams in the other division of the Eastern Conference. Finally, each team in the Eastern Conference plays every team in the Western Conference twice (once at home and once away). So each team in the Eastern Conference plays thirty games against the fifteen teams in the Western Conference.

Each team in the Western Conference plays four or five games against each of the six or seven teams in its division. So each team in the Western Conference plays

twenty-six or twenty-nine games in its own division. Each team in the Western Conference also plays three games against six or seven of the teams in the other division of the Western Conference. So each team in the Western Conference plays twenty-one or twenty-four games against the six or seven teams in the other division of the Western Conference. Finally, each team in the Western Conference plays every team in the Eastern Conference twice (once at home and once away). So each team in the Western Conference plays thirty-two games against the sixteen teams in the Eastern Conference.

Apart from the regular season, the NHL has a postseason (the Stanley Cup playoffs). This is basically an elimination tournament where two teams play against each other to win a best-of-seven series in order to go to the next round. The top three teams in each of the four divisions and the two conference teams with the next highest number of points qualify for the playoffs. So basically eight teams from each of the two conferences qualify for the playoffs.

All the NHL ice hockey games are 60 minutes long. Each game consists of three twenty-minute periods with an interval between periods. At the end of the three periods, the team that has scored more goals wins the game. Overtime occurs in case the game is tied at the end of the three periods. Overtime is a five-minute, three-on-three sudden-death period during the regular season. Sudden-death period means that whichever team scores a goal first wins the game. In the regular season, if at the end of overtime the game is still tied then the game enters a shootout. For each team three players in turn take a penalty shot. During the three-round shootout the team with the most goals wins the game. If after the three shootout rounds the game is still tied then the shootout is continued but it becomes sudden-death. It is important to note that unlike the regular season, during the playoffs there are no shootouts. Rather during the playoffs, multiple sudden-death, twenty minute five-on-five periods are played till a team scores a goal. In

the regular season, if a team wins a game it is awarded two points, if it loses in overtime or shootout it is awarded one point, and it is awarded zero points if it loses within the three twenty minute periods.

III. Literature Review

There has been some research on evaluating team performances in ice hockey. However, none of them make use of game-level data in order to predict team wins. This study hypothesizes that using pair-wise game-level data along with season-level data can be useful in improving the prediction of a team's win percentage in the following seasons.

Laffey and Ames (2016) develop generalized linear models based on OLS and Poisson regression with elastic net regularization for estimating the number of regular season and playoff wins for ice hockey teams from a wide variety of regular season team statistics. Laffey and Ames (2016) state that shot counts may actually be a better measure to estimate future team performance than puck possession and goal counts. Laffey and Ames (2016) state so because there are frequent changes in puck possession in hockey, which makes it hard to estimate dominance of puck possession, and goals are scarce in hockey making robust estimate of future team performance difficult. Laffey and Ames (2016) use 53 statistics split over multiple scenarios, such as when the team is leading, trailing, the team is shorthanded etc. as well as statistics such as average player ages to estimate regular season and playoff wins for NHL teams as teams adjust strategy depending on the situation. Laffey and Ames (2016) face the issue of small sample size (less than 500 team observations) due to the expansion of predictor variables. Furthermore, Laffey and Ames (2016) state that in their elastic net Poisson regression models, statistics such as "shooting percentage" (the percentage of shots on net taken by a player that result in a goal) and "goals for" (a player scores against the opposing team) have little influence over the model's prediction of playoff performance, while heavily positively influencing the estimation of regular season wins.

My study tests whether a team's win percentage next season can be predicted using variables such as its current season goal differential and shot differential. Like Laffey and Ames (2016), my study uses generalized linear models based on OLS regression to estimate

win percentages for ice hockey teams. My study uses season-level and pair-wise shot differential as key independent variables to estimate win percentages for ice hockey teams. My study uses less predictor variables than Laffey and Ames (2016), and does not account for the situation of the game to predict win percentages for NHL teams. This is why, unlike Laffey and Ames (2016), my study doesn't face the issue of small sample size. My study only predicts regular season performance for teams. My study hypothesizes that both "shooting percentage" and "goals for" will have a considerable positive influence over my model's prediction of team performance.

Schulte et al (2017) also develop a model to estimate ice hockey team performance. A novel aspect of their data set is that it includes location information about where an action took place. Schulte et al (2017) take into account the context of the action (represented by the Markov game state) and model the medium-term impact of an action by propagating its effect to future states. Using AI techniques they apply their model to evaluate the performance of teams in terms of their actions' total impact on which team scores the next goal.

Unlike Schulte et al (2017), my study does not make use of location information about where an action took place (except for home versus away games). My study also uses regression analysis instead of AI techniques. Additionally, my study predicts something more important than what Schulte et al (2017) predict. My study predicts win percentages for ice hockey teams, whereas Schulte et al (2017) only predict which team scores the next goal in a game. Finally, out of the 13 action types used by Schulte et al (2017), my study uses 3 action types, which are "shot" (a player shoots on opposing team's goal), "shot against" (opposing team's player shoots on goal), and "goal" (a player scores a goal against the opposition), to compute some of its key independent variables. My study uses only these 3 action types because they are much better predictors of wins as compared to the other action types such as

"pass" (the player attempts a pass to a teammate) and "block" (a block attempt on the puck's trajectory).

Kaplan et al (2014) develop a model that produces win probabilities given the goal differential ("goals for" minus "goals against") and manpower differential (caused by penalties) at any point in a game. Additionally, Kaplan et al (2014) show how their real-time win probability scorecard can be used to evaluate a hockey player's individual contribution to the probability of winning (win probability added).

Like Kaplan et al (2014), this study uses goal differential as one of the key independent variables. However, unlike Kaplan et al (2014), this study predicts win percentages for ice hockey teams in the next season based on the data from the previous season. This study does not evaluate or predict the individual performance of hockey players.

Papers on estimating ice hockey team performances are fairly limited. Most of the papers on ice hockey are on player evaluation. One such paper is Schuckers and Curro (2013). They consider various events such as shots, hits, and takeaways to estimate the probability that a goal arises within a 20 second window of the event. They account for the home team advantage and advantage of beginning a shift in the offensive zone in their model.

Like Schuckers and Curro (2013), this study takes into account events such as "shot for" and "shot against." Like Schuckers and Curro (2013), this study also controls for home versus away games to improve forecast. However, unlike Schuckers and Curro (2013) this study predicts win percentages for ice hockey teams and does not evaluate ice hockey players.

Gramacy, Taddy, and Tian (2017) develop a model to better evaluate hockey players. They considered goals (either for or against the home team) as the dependent variable in their model. Gramacy, Taddy, and Tian (2017) also account for the home team effect, team-seasons effect, manpower effect and playoff effects in their model. Lastly, Gramacy, Taddy,

and Tian (2017) use multiple seasons of data because ice hockey teams do not score many goals per match (roughly 5.5 goals per match).

Unlike Gramacy, Taddy, and Tian (2017), this study will use "goals for" and "goals against" to compute some of the key independent variables of the models. I will account for the home team effect in the models. This study also uses multiple seasons of data, so that it can be accurately checked whether extreme goal differentials in games help predict outcomes of the same pairs of teams in the following season. However, unlike Gramacy, Taddy, and Tian (2017), this study predicts win percentages for ice hockey teams and does not evaluate ice hockey players.

Another paper on player evaluation is Smith (2016), which is about creating a better plus-minus statistic for evaluating players that unlike the traditional plus-minus metric takes into account the quality of the other players on the ice with an individual. This paper builds of previous research (Gramacy et al (2013)), which focused on goals to build an adjusted plus-minus statistic. Goals are rare in hockey thus this paper uses shots instead of goals, which allows it to get much more information per game and thus come up with a more robust adjusted plus-minus statistic for players.

My study uses "shots for" and "shots against" to compute some of the key independent variables in models to predict win percentages for ice hockey teams, as like Smith (2016), I too feel that using shots will allow me to get more information per game and thus will help me make a more accurate prediction. However, unlike Smith (2016), this study predicts win percentages for ice hockey teams and does not evaluate ice hockey players.

Finally, Macdonald (2012) states that one of the main disadvantages of the OLS regression models is that the estimates have large error bounds. As certain pairs of teammates often play together, collinearity is present in the data and is one reason for the large errors. The relative lack of scoring in hockey is the second reason for the large errors. Macdonald

(2012) uses the ridge regression method instead of OLS, which is often the case when collinearity is present in the data. Macdonald (2012) also creates models which use not only goals but also shots, Fenwick rating (shots plus missed shots), and Corsi rating (shots, missed shots, and blocked shots) as shots are more common in hockey (more data) and so the resulting estimates have smaller error bounds. The results of Macdonald's (2012) ridge regression models are estimates of the offensive and defensive contributions of forwards and defensemen during even strength, power play, and shorthanded situations, in terms of goals per 60 minutes. These estimates are independent of strength of teammates, strength of opponents, and the zone in which a player's shift begins.

My study does not evaluate player abilities. However, like Macdonald (2012), my study uses not only goals but also shots as independent variables in my regression model so that even my resulting estimates have smaller error bounds.

There are a few other papers on ice hockey player evaluations as well but because they do not relate to my research, my study does not discuss them. For this very reason, my study will also not be discussing two other papers on ice hockey, one of which examines the effect of age on scoring performance and on plus minus statistic for NHL players, and the other one classifies puck possession events in ice hockey.

To summarize, none of the studies have made use of game-level data in order to predict win percentages for NHL teams. This study will make use of both season-level data and pair-wise game-level data in a season in order to predict win percentages for the pairs of teams in the following seasons. I think that doing so instead of using only season-level data will improve the prediction of win percentages for NHL teams. As previously stated, papers on estimating NHL team performances are pretty limited. One such paper is by Laffey and Ames (2016). Their paper estimates the number of regular season and playoff wins from a wide variety of regular season team statistics.

IV. Hypothesis Development

Prior studies find that a team's win percentage next season can be predicted using its current-season goal differential and current-season shot differential. Thus, one of the hypotheses of this paper is that:

- H0: A team's win percentage next season can be predicted using its current-season goal differential and shot differential.

Predicting winning percentages for the NHL teams in the following seasons based on their performance in the previous seasons has not been very accurate. I think it is not accurate because it ignores pair-specific factors in making the prediction. Thus, one of the hypotheses of this paper is that:

- H1: Game-level data during a season can be useful for improving the prediction of a team's win percentage in the following season.

The motivation of the following hypothesis (H2) is similar to the hypothesis stated above (H1). However, the following hypothesis is more specific compared to H1:

- H2: The game-level data that is useful pertains to games with extreme outcomes in either goal differential or shot differential or both (this has been expressed as multiple hypotheses below):

    H2a: games with extreme goal differential.

    H2b: games with extreme shot differential.

    H2c: games with extreme goal and shot differentials.

Games with extreme outcomes are more helpful to improve the prediction of a team's win percentage if there is a large sample of opponent interactions during a season. Thus, one of the hypotheses of this paper is that:

- H3: The effects in H2 are stronger if there is a larger sample of opponent interactions during a season (basically there is a larger sample leading to the extreme outcomes) and perhaps only hold in such cases.

A number of papers on estimating team wins and on player evaluations in ice hockey control for the home edge afforded by visiting teams. Thus, one of the hypotheses of this paper is that:

- H4: Controlling for home vs. away games improves the forecast.

V. Data

The pair-wise NHL game-level data which was collected was from the 2005-2006 season to the 2016-2017 season. The pair-wise game-level data before October 2005 was not collected because the 2004-05 season was a lockout (the whole season was cancelled). Also, the salary cap and rules changed in October 2005 and have remained stable since then, so starting in October 2005 was ideal for this study. Details of how the pair-wise game-level data was obtained from the NHL stats website can be found in "Appendix 1: Getting the pair-wise game-level data from the NHL stats website." The game-level data variables are described under "Appendix 2: Game Level Data Variables."

The season-level data, which was collected, was from the 2005-2006 season to the 2016-2017 season. The season-level data was collected from the NHL stats website. The season-level data variables are described under "Appendix 3: Season-Level Data Variables." The season-level data obtained for this study suffered from one of the same challenges as the pair-wise game-level data. This challenge was that Atlanta Thrashers changed its name to Winnipeg Jets in the 2011-12 season and Phoenix Coyotes changed its name to Arizona Coyotes in the 2014-2015 season. The old names of these two teams were changed to their new names in STATA.

The pair-wise game-level data and the season-level/ aggregate data were then combined using code. The variables in the combined dataset are described under "Table I: Combined Data Variable Names and their Description." The summary statistics for these combined data variables are described under "Table 2: Summary Statistics for the Combined Data Variables."

The independent variables which were used for data analytics in this study were created using the variables listed in "Table 1: Combined Data Variable Names and their Description." The independent variables which were used for data analytics in this study are

described under “Table 3: Independent Variables used for Data Analytics in STATA.” The dependent variable which was used for data analytics in STATA is described under “Table 4: Dependent Variable used for Data Analytics in STATA.”

The summary statistics for the independent variables which were used for data analytics in this study are described under “Table 5: Summary Statistics for the Independent Variables used for Data Analytics in STATA.” The summary statistics for the dependent variable which was used for data analytics in this study is described under “Table 6: Summary Statistics for the Dependent Variable used for Data Analytics in STATA.”

I required many full seasons of data as the main focus was on assessing whether extreme imbalances in games (scores of 7 to 2 as opposed to 3 to 2, for example, or very large shot differentials) help us predict outcomes of the same pair of teams in the following season. Games with extreme imbalances are relatively scarce, so we needed many full seasons in order to have precise estimates (statistically significant effects). This study’s models used every game but this study hypothesized that the useful information would most probably be only in games with extreme outcomes with the possible additional condition that the teams faced each other multiple times in the season.

There are a few shortcomings of the data used in this study. Laffey and Ames (2016) use statistics split over multiple scenarios, such as when the team is leading, trailing, the team is shorthanded etc. as well as statistics such as average player ages to estimate regular season and playoff wins for NHL teams as teams adjust strategy depending on the situation. Unlike Laffey and Ames (2016), my study does not account for the situation of the game. Additionally, Schulte et al (2017) use a dataset that includes location information about where an action took place. Compared to previous studies that assign a single value to action, Schulte et al (2017) take into account the context of the actions and then accordingly assign values to them. Schulte et al (2017) apply AI techniques for their research. My study does not

make use of location information about where an action took place (except for home versus away games). Furthermore, my paper's dataset does not include information on the quality of the players playing in a game for both the teams. In ice hockey players often get injured. If the star player of a team gets injured and thus is unable to play a few games, then this could have direct consequences on the results of those games for that team. In conclusion, including information on the situation of the game, location of actions such as shots, and quality of the players of both the teams may be of value in order to better predict win percentages for ice hockey teams. However, because the main focus of my paper is to see if including pair-wise game-level data with the season-level/ aggregate data helps to better predict win percentages for ice hockey teams, my paper does not make use of that information. Also, getting information on the situation of the game, location of actions, and quality of the players playing in a game is not straightforward and requires the use AI and machine learning techniques, which is another reason why this paper does not make use of that information.

Finally, a possible reason why other researchers may not have used game-level data to predict winning percentages for ice hockey teams is that they might have felt that looking at how a team performed in a game against another team is not of much value in itself and it is just better to look at how a team performed against all the other teams in a season. I agree with these researches to a certain extent as I think that all of the game-level data is probably not of much value in predicting winning percentages for ice hockey teams. However, I think that some of the game-level data for instance for those games with high shot differential and high goal differential may actually be valuable and so should be used along with the season-level/ aggregate-level data in order to predict win percentages for ice hockey teams.

VI. Empirical Method

This study uses a cross-sectional regression: each pair of items in a particular season is a cross-sectional unit, and current season performance is used to predict next season's win percentage. This study uses both pair-wise game-level data and season-level data in the previous season to predict the win-percentages for the pairs of teams in the following season. This study tests for its hypotheses using fourteen different models. One of the models uses all of the pair-wise game-level data along with the season-level data to predict winning percentages for NHL teams. A second model uses pair-wise game-level data to condition on extreme shot differential and uses this along with the season-level data to predict winning percentages for NHL teams. A third model uses pair-wise game-level data to condition on number of games played in the prior season along with the season-level data to predict winning percentages for NHL teams. Ten of the fourteen models are linear regression models, two are probit, and two are logit models. I use a number of different models in order to see if in any of these models the pair-wise game-level data helps in predicting winning percentages for NHL teams. The models predict individual game-level outcomes. If the pair-wise variables are significant then I will aggregate these pair-wise game-level predictions to compute a season-level prediction.

Model 1 is a linear regression model and tests for HO. It uses season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins in the next season. This study used the following regression equation for model 1:

(1) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\;i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \varepsilon_{i\,t}$

where $Y_{i\,t}$ is a dummy variable that takes the value of 1 if the home team wins in the current game and 0 otherwise,

β_0 is the constant on the regression,

$X_{1\,i\,t-1}$ is the goal differential for the home team against all the other teams it played in the previous season. It has been constructed by dividing goals h (home team) scored in all games by the total of goals h (home team) scored in all games and goals a (away team) scored in all games.

$X_{2\,i\,t-1}$ is the goal differential for the away team against all the other teams it played in the previous season. It has been constructed by dividing goals a (away team) scored in all games by the total of goals a (away team) scored in all games and goals h (home team) scored in all games,

$X_{3\,i\,t-1}$ is the shot differential for the home team against all the other teams it played in the previous season. It has been constructed by dividing shots h (home team) scored in all games by the total of shots h (home team) scored in all games and shots a (away team) scored in all games,

$X_{4\,i\,t-1}$ is the shot differential for the away team against all the other teams it played in the previous season. It has been constructed by dividing shots a (away team) scored in all games by the total of shots a (away team) scored in all games and shots h (home team) scored in all games,

$\varepsilon_{i\,t}$ is the error term on the regression

Model 2 is a linear regression model and it tests for H1 and H4. It uses pairwise total goal differential, pair-wise total shot differential, pair-wise goal differential for only the home games of one of the pair-wise teams and pair-wise shot differential for only the home games of one of the pair-wise teams in the previous season along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 2:

(2) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} +$

$\beta_6 X_{6\,i\,t-1} + \beta_7 X_{7\,i\,t-1} + \beta_8 X_{8\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the goal differential for a team against the other pair-wise team it played in the previous season. It has been constructed by dividing goals h (home team) scored when the pair played by the total of goals h (home team) scored when the pair played and goals a (away team) scored when the pair played,

$X_{2\,i\,t-1}$ is the goal differential for a team in its home games against the other pair-wise team it played in the previous season. It has been constructed by dividing goals h

(home team) scored in its home games when the pair played by the total of goals h (home team) scored in its home games when the pair played and goals a (away team) scored against h (home team) when a was away,

$X_{3\,i\,t-1}$ is the shot differential for a team against the other pair-wise team it played in the previous season. It has been constructed by dividing shots h (home team) scored against a (away team) when the pair played by the total of shots h (home team) scored against a (away team) when the pair played and shots a (away team) scored against h when the pair played,

$X_{4\,i\,t-1}$ is the shot differential for a team in its home games against the other pair-wise team it played in the previous season. It has been constructed by dividing shots h (home team) scored in its home games when the pair played by the total of shots h (home team) scored in its home games when the pair played and shots a (away team) scored against h (home team) when a was away,

$X_{5\,i\,t-1}$, $X_{6\,i\,t-1}$, $X_{7\,i\,t-1}$, $X_{8\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as
$X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 3 is a linear regression model and it tests for H2b and H4. It uses pair-level data to condition on extreme shot differentials and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 3:

(3) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the interaction term between (1) the shot differential for a team in its home games against the other pair-wise team it played in the previous season and (2) the high shot differential (2 SDs above the mean of pair_home_shot_differential, which is explained above in (1)) for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 for games in which there is high shot differential and 0 otherwise,

$X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as
$X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 4 is a linear regression model and it tests for H3, H4. It uses pair-level data to condition on number of games played in the prior season and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and

the away team in the previous season in order to predict wins next season. This study uses the following equation for model 4:

(4) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

> $X_{1\,i\,t-1}$ is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the shot differential for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 if the number of games played is equal to or greater than 4 and takes the value of 0 otherwise. The reason I take the cut-off as being greater than or equal to four is that during the regular season, each team in the Eastern Conference plays four games against each of the seven teams in its own division and each team in the Western Conference plays four or five games against each of the six or seven teams in its own division,
>
> $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 5 is a logit model and it tests for H1, H4. It uses pair-wise goal differential for only the home games of one of the pair-wise teams along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 5:

(5) $F\,(\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}) =$

$(e^{\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}})$ /

$(1+e^{\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}})$

Where $F\,(\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1})$ and β_0 are the same as $Y_{i\,t}$ and β_0 in Equation 1 respectively,

> $X_{1\,i\,t-1}$ is the same as $X_{2\,i\,t-1}$ in Equation 2,
>
> $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $X_{5\,i\,t-1}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$ and $X_{4\,i\,t-1}$ in Equation 1 respectively.

Model 6 is a linear regression model and it tests for H2a, H4. It uses pair-level data to condition on extreme goal differentials and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 6:

(6) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

> $X_{1\,i\,t-1}$ is the interaction term between (1) the goal differential for a team in its home games against the other pair-wise team it played in the previous season and (2) the high goal differential (2 SDs above the mean of pair_home_goal_differential, which is explained above in (1)) for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 for games in which there is high goal differential and takes the value of 0 otherwise,
>
> $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 7 is a linear regression model and it tests for H3, H4. It uses pair-level data to condition on number of games played in the prior season and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 7:

(7) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

> $X_{1\,i\,t-1}$ is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the goal differential for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 if the number of games played is equal to or greater than 4 and takes the value of 0 otherwise. The reason I take the cut-off as being greater than or equal to four is that during the regular season, each team in the Eastern Conference plays four games against each of the seven teams in its own division and each team in the Western Conference plays four or five games against each of the six or seven teams in its own division,

$X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 8 has the same setup and tests the same hypothesis as model 5, but it uses a probit model rather than a logit model.

Model 9 is a logit model and it tests for H1, H4. It uses pair-wise shot differential for only the home games of one of the pair-wise teams along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 9:

(9) $F\,(\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}) =$

$(e^{\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}})$ / (1

$+ e^{\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1}})$

where $F\,(\beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1})$ and β_0 are the same as $Y_{i\,t}$ and β_0 in Equation 1 respectively,

$X_{1\,i\,t-1}$ is the same as $X_{4\,i\,t-1}$ in Equation 2,

$X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $X_{5\,i\,t-1}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$ and $X_{4\,i\,t-1}$ in Equation 1 respectively.

Model 10 has the same setup and tests the same hypothesis as model 9, but it uses a probit model rather than a logit model.

Model 11 is a linear regression model and it tests for H2C, H4. It uses pair-level data to condition on extreme goal differentials and extreme shot differentials, and uses this along with the season-level total goal differential and season-level total shot differential for both the

home team and the away team in the previous season in order to predict wins next season. This study uses the following Equation for model 11:

(11) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \beta_6 X_{6\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the same as $X_{1\,i\,t-1}$ in Equation 3,

$X_{2\,i\,t-1}$ is the same as $X_{1\,i\,t-1}$ in Equation 6,

$X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$, $X_{6\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 12 is a linear regression model and it tests for H3, H4. It uses pair-level data to condition on number of games played in the prior season and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. This study uses the following equation for model 12:

(12) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \beta_6 X_{6\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the same as $X_{1\,i\,t-1}$ in Equation 4,

$X_{2\,i\,t-1}$ is the same as $X_{1\,i\,t-1}$ in Equation 7,

$X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$, $X_{6\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as $X_{1\,i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 13 is a linear regression model and it tests for H2a, H3. It examines cases that have both a high goal differential and at least 4 games in the prior season. This study uses the following equation for model 13:

(13) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the high goal differential (2 SDs above the mean of pair_home_goal_differential) for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 if both the conditions above are met and takes the value of 0 if condition 1 above is not satisfied and condition 2 above is satisfied.

$X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as
$X_{1\ i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

Model 14 is a linear regression model and it tests for H2b, H3. It examines cases that have both a high shot differential and at least 4 games in the prior season. This study uses the following equation for model 14:

(14) $Y_{i\,t} = \beta_0 + \beta_1 X_{1\,i\,t-1} + \beta_2 X_{2\,i\,t-1} + \beta_3 X_{3\,i\,t-1} + \beta_4 X_{4\,i\,t-1} + \beta_5 X_{5\,i\,t-1} + \varepsilon_{i\,t}$

Where $Y_{i\,t}$ and β_0 are the same as in Equation 1,

$X_{1\,i\,t-1}$ is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the high shot differential (2 SDs above the mean of pair_home_shot_differential) for a team in its home games against the other pair-wise team it played in the previous season. $X_{1\,i\,t-1}$ is a dummy variable that takes the value of 1 if both the conditions above are met and takes the value of 0 if condition 1 above is not satisfied and condition 2 above is satisfied.

$X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$, $X_{5\,i\,t-1}$ and $\varepsilon_{i\,t}$ are the same as
$X_{1\ i\,t-1}$, $X_{2\,i\,t-1}$, $X_{3\,i\,t-1}$, $X_{4\,i\,t-1}$ and $\varepsilon_{i\,t}$ in Equation 1 respectively.

VII. Results

In all of the 14 models, almost all of the season-level variables are significant at the 1% level of significance. The only season-level variables which are not significant at the 1% level of significance and instead are significant at the 5% level of significance are shot differential for the home team against all the other teams it played in the previous season (s_h_shot_diff) in model 3 (as it can be seen from Table 9), in model 4 (as it can be seen from Table 10), in model 9 (as it can be seen from Table 15), in model 10 (as it can be seen from Table 16), in model 11 (as it can be seen from Table 17) and in model 12 (as it can be seen from Table 18). Also, only in model 12 (as it can been from Table 18) shot differential for the away team against the other teams it played in the previous season (s_a_shot_diff) is significant at the 5% level and not at the 1% level.

In all of the 14 models, almost all of the pair-level variables are insignificant even at the 10% level of significance. Only two pair-level variables were found to be significant. One of the pair-wise variable found to be significant was when high goal differential was observed in the interaction term between high goal differential for a team in its home games against the other pair-wise team and the goal differential for a team in its home games against the other pair-wise team (high_home_goal_diff#c.pair_home_goal_diff) (as it can be seen in model 6 from Table 12 and as it can be seen in model 11 from Table 17). It was significant at the 10% level of significance. This provides marginal support for the claim that extreme game-level outcomes from the previous season can help in predicting a team's win percentage in the following season. Another pair-level variable found to be significant was when high goal differential was observed and at least 4 games played was not observed in the interaction term between at least 4 games played against the other pair-wise team and high goal differential for a team in its home games against the other pair-wise team (at_least_4_gp#c.high_home_goal_diff) (as it can be seen in model 13 from Table 19). It was

significant at the 5% level of significance. This suggests that only in the games a team plays outside its own division, the extreme game-level data helps in predicting its win percentage in the following season.

The signs of the season-level variables make sense in all the models. Season-level total goal differential for the home team and the season level total shot differential for the home team in the current season have positive effects on wins of the home team in the following season. Also, season-level total goal differential for the away team and the season level total shot differential for the away team in the current season have negative effects on wins of the home team in the following season. Additionally, all the linear regression models have an R-squared value below 0.0187, and all the probit and logit models have a Pseudo R2 value below 0.0133. This shows that the predictive power of the models for team wins in the following season is pretty low. Low R-squared values were expected as the models predict individual game outcomes and not season-level outcomes.

Model 1 is a linear regression model and tests for HO. It uses season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins in the next season. The results of the regression for model 1 are summarized in Table 7. All of the independent variables used in this model, which are all season-level-variables, are significant at the 1% level of significance. However, the predictive power of these variables for team wins in the following season is pretty low. This can be seen from the R-squared value of 0.0171. Thus, model 1 only somewhat confirms H0.

Model 2 is a linear regression model and it tests for H1 and H4. It uses pair-wise total goal differential, pair-wise total shot differential, pair-wise goal differential for only the home games of one of the pair-wise teams and pair-wise shot differential for only the home games of one of the pair-wise teams in the previous season along with the season-level total goal

differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 2 are summarized in Table 8. Only the season-level variables used in this model are significant. All of the season-level variables are significant at the 1% level of significance. The pair-level variables used in this model are not even significant at the 10% level of significance. This suggests that in this model, pair-level variables do not help in predicting wins for a team next season. The predictive power of all of the variables of this model for team wins in the following season is also pretty low. This can be seen from the R-squared value of 0.0184. Thus, model 2 is unable to confirm H1 and H4.

Model 3 is a linear regression model and it tests for H2b and H4. It uses pair-level data to condition on extreme shot differentials and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 3 are summarized in Table 9. Only the season-level variables used in this model are significant. All of the season-level variables except for s_h_shot_diff (shot differential for the home team against all the other teams it played in the previous season) are significant at the 1% level of significance. s_h_shot_diff is significant at the 5% level of significance. The pair-level variables used in this model are not even significant at the 10% level of significance. This suggests that in this model pair-level variables do not help in predicting wins for a team next season. The predictive power of all of these variables in this model for team wins in the following season is also pretty low. This can be seen from the R-squared value of 0.0183. Thus, model 3 is unable to confirm H2b and H4.

Model 6 is a linear regression model and it tests for H2a and H4. It uses pair-level data to condition on extreme goal differentials and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away

team in the previous season in order to predict wins next season. The results of the regression for model 6 are summarized in Table 12. All of the season-level variables used in this model are significant. All of the season-level variables are significant at the 1% level of significance. The pair-level variable used in this model is the interaction term between (1) the goal differential for a team in its home games against the other pair-wise team it played in the previous season and (2) the high goal differential for a team in its home games. The pair-level variable is significant when there is high goal differential. It is significant at the 10% level of significance. This provides marginal support for the claim that extreme game level outcome from the previous season can help in predicting a team's win percentage in the following season. The predictive power of all these variables for team wins in the following season is pretty low. This can be seen from the R-squared value of 0.0184. Thus, model 6 provides only marginal support for H2a and H4.

Model 4 and model 7 are linear regression models and test for H3 and H4. Both these models use pair-level data to condition on number of games played in the prior season and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The only difference between these two models is that model 4 conditions on number of games played in the prior season using pair-level shot differential, whereas model 7 conditions on number of games played in the prior season using pair-level goal differential. The results of the regression for model 4 are summarized in Table 10 and the results of the regression for model 7 are summarized in Table 13. In both these models, only the season-level variables are significant. All of the season-level variables are significant at the 1% level of significance except for s_h_shot_diff (shot differential for the home team against all the other teams it played in the previous season) in model 4. s_h_shot_diff in model 4 is significant at the 5% level of significance. In both these models, the pair-level

variables used are not even significant at the 10% level of significance. This suggests that in these models pair-level variables do not help in predicting wins for a team next season. The predictive power of the variables of both these models is also pretty low. Both model 4 and model 7 have an R-squared value of 0.0181. Both model 4 and model 7 are unable to confirm H3 and H4.

Model 5 is a logit model and model 8 is a probit model. Both of these models test for H1 and H4. Both these models use pair-wise goal differential for only the home games of one of the pair-wise teams along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 5 are summarized in Table 11 and the results of the regression for model 8 are summarized in Table 14. In both these models, only the season-level variables are significant. All of the season-level variables are significant at the 1% level of significance. In both these models, the pair-level variables are not even significant at the 10% level of significance. This suggests that in these two models the pair-level variable does not help in predicting wins for a team next season. The predictive power of the variables of both these models is also pretty low. Both model 5 and model 8 have a Pseudo R2 value of 0.0133. Both model 5 and model 8 are unable to confirm H1 and H4.

Model 9 is a logit model and model 10 is a probit model. Both of these models test for H1 and H4. Both these models use pair-wise shot differential for only the home games of one of the pair-wise teams along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 9 are summarized in Table 15 and the results of the regression for model 10 are summarized in Table 16. In both these models, only the season-level variables are significant. All of the season-level variables

are significant at the 1% level of significance except for s_h_shot_diff (shot differential for the home team against all the other teams it played in the previous season), which is significant at the 5% level of significance in both the models. In both these models, the pair-level variable is not even significant at the 10% level of significance. This suggests that in these models the pair-level variable does not help in predicting wins for a team next season. The predictive power of the variables of both these models is also pretty low. Both model 9 and model 10 have a Pseudo R2 value of 0.0132. Both model 9 and model 10 are unable to confirm H1 and H4.

Model 11 is a linear regression model and it tests for H2C, H4. It uses pair-level data to condition on extreme goal differentials and extreme shot differentials, and uses this along with the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 11 are summarized in Table 17. All of the season-level variables are significant at the 1% level of significance, except for s_h_shot_diff (shot differential for the home team against all the other teams it played in the previous season), which is significant at the 5% level of significance. Two pair-level variables were used in this model. One of the pair-level variables used in this model is an interaction term between (1) the shot differential for a team in its home games against the other pair-wise team it played in the previous season and (2) the high shot differential for a team in its home games. The other pair-level variable used in this model is an interaction term between (1) the goal differential for a team in its home games against the other pair-wise team it played in the previous season and (2) the high goal differential for a team in its home games. The pair-level variable which is significant in this model is high goal differential for a team in its home games against the other pair-wise team it played in the previous season (high_home_goal_diff). It is significant at the 10% level of significance. This provides marginal support for the claim that extreme

game level outcomes from the previous season can help in predicting a team's win percentage in the following season. The predictive power of all these variables for team wins in the following season is also pretty low. This can be seen from the R-squared value of 0.0187. Thus, model 11 provides marginal support for H2c and H4.

Model 12 is a linear regression model and it tests for H3, H4. It uses pair-level data to condition on number of games played in the prior season using pair-level shot differential and to condition on number of games played in the prior season using pair-level goal differential. Along with these variables, this model also uses the season-level total goal differential and season-level total shot differential for both the home team and the away team in the previous season in order to predict wins next season. The results of the regression for model 12 are summarized in Table 18. The two goal differential season level variables used in this model are significant at the 1% level of significance and the two shot differential season level variables used in this model are significant at the 5% level of significance. The pair-level variables used in this model are not even significant at the 10% level of significance. This suggests that in this model pair-level variables do not help in predicting wins for a team next season. The predictive power of all these variables for team wins in the following season is also pretty low. This can be seen from the R-squared value of 0.0182. Thus, model 12 is unable to confirm H3 and H4.

Model 13 is a linear regression model and it tests for H2a, H3. It examines cases that have both a high goal differential and at least 4 games in the prior season. The results of the regression for model 13 are summarized in Table 19. All of the season-level variables used in this model are significant at the 1% level of significance. The pair-level variable used in this model is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the high goal differential (2 SDs above the mean of pair_home_goal_differential) for a team in its home

games against the other pair-wise team it played in the previous season. The pair-level variable is significant when condition 1 above is not satisfied and condition 2 above is satisfied. It is significant at the 5% level of significance. The pair-level variable is not significant even at the 10% level of significance if both the conditions above are satisfied. The results of the pair-level variables suggest that only in the games a team plays outside its own division, the extreme game-level data helps in predicting a team's win percentage in the following season. The predictive power of all these variables for team wins in the following season is pretty low. This can be seen from the R-squared value of 0.0176.

Model 14 is a linear regression model and it tests for H2b, H3. It examines cases that have both a high shot differential and at least 4 games in the prior season. The results of the regression for model 14 are summarized in Table 20. All of the season-level variables used in this model are significant at the 1% level of significance. The pair-level variable used in this model is the interaction term between (1) the total games a team played against the other pair-wise team in the previous season being greater than or equal to four and (2) the high shot differential (2 SDs above the mean of pair_home_shot_differential) for a team in its home games against the other pair-wise team it played in the previous season. The pair-level variables are not even significant at the 10% level of significance. The predictive power of all these variables for team wins in the following season is pretty low. This can be seen from the R-squared value of 0.0174.

VIII. Conclusion

This study hypothesized that using pair-wise game-level data along with the season-level data from the previous season would be helpful in improving the prediction of a team's win percentage in the current season. In majority of the models this study did not find any significant results, even at the 10% level of significance, for the pair-wise game-level data variables in predicting the pair's outcome next season. This helps establish the idea that including more granular information does not necessarily increase the predictive power of models. Only two pair-level variables were found to be significant. One of the pair-wise variable found to be significant was when high goal differential was observed in the interaction term between high goal differential for a team in its home games against the other pair-wise team and the goal differential for a team in its home games against the other pair-wise team (high_home_goal_diff#c.pair_home_goal_diff) (as it can be seen in model 6 from Table 12 and as it can be seen in model 11 from Table 17). It was significant at the 10% level of significance. This provides marginal support for the claim that extreme game-level outcomes from the previous season can help in predicting a team's win percentage in the following season. Another pair-level variable found to be significant was when high goal differential was observed and at least 4 games was not observed in the interaction term between at least 4 games played against the other pair-wise team and high goal differential for a team in its home games against the other pair-wise team (at_least_4_gp#c.high_home_goal_diff) (as it can be seen in model 13 from Table 19). It was significant at the 5% level of significance. This suggests that only in the games a team plays outside its own division, the extreme game-level data helps in predicting a team's win percentage in the following season. This study also concludes that pair-wise game-level data won't help predict aggregate win percentages for teams as it did not help predict game-level outcomes in the following seasons.

Across all of its 14 models this study found most of the season-level variables to be significant at the 1% level of significance and a few season-level variables to be significant at the 5% level of significance in predicting a team's winning percentage next season. However, in all of the models the predictive power of the variables for team wins in the following season was pretty low.

Further research could use the same methodology used in this study but with different pair-wise game level variables in order to see if these new game-level variables help in predicting a team's winning percentage next season. For instance, further research could use pair-wise power play goal differentials and pair-wise penalty kill percentage to see if these game-level variables improve the predictive power of the models. I leave this for future research.

References

Brander, James A., Edward J. Egan, and Louisa Yeung. "Estimating the Effects of Age on NHL Player Performance." *Journal of Quantitative Analysis in Sports* 10, no. 2 (2014). doi:10.1515/jqas-2013-0085.

Gramacy, Robert B., Shane T. Jensen, and Matt Taddy. "Estimating Player Contribution in Hockey with Regularized Logistic Regression." *Journal of Quantitative Analysis in Sports* 9, no. 1 (2013). doi:10.1515/jqas-2012-0001.

Jensen, Shane. "A Statistician Reads the Sports Pages: Measuring Player Contributions in Hockey." *Chance* 26, no. 3 (2013): 34-38. doi:10.1080/09332480.2013.845449.

Kaplan, Edward H., Kevin Mongeon, and John T. Ryan. "A Markov Model for Hockey: Manpower Differential and Win Probability Added." *INFOR: Information Systems and Operational Research* 52, no. 2 (2014): 39-50. doi:10.3138/infor.52.2.39.

Laffey, Nicholas, and Brendan Ames. "A Sparse Regression Approach For Evaluating and Predicting NHL Results." December 1, 2016.

Macdonald, Brian. "Adjusted Plus-Minus for NHL Players Using Ridge Regression with Goals, Shots, Fenwick, and Corsi." *Journal of Quantitative Analysis in Sports* 8, no. 3 (2012). doi:10.1515/1559-0410.1447.

Macdonald, Brian, Christopher Weld, and David Arney. "Quantifying Playmaking Ability in Hockey." July 5, 2013.

"Official Site of the National Hockey League." NHL.com. https://www.nhl.com/.

Routley, Kurt. "A Markov Game Model for Valuing Player Actions in Ice Hockey." Spring 2015.

Schuckers, Michael, and James Curro. "Total Hockey Rating (THoR): A Comprehensive Statistical Rating of National Hockey League Forwards and Defensemen Based upon All On-ice Events." March 1-2, 2013.

Schulte, Oliver, and Zeyu Zhao. "Apples-to Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact." March 3-4, 2017.

Schulte, Oliver, Mahmoud Khademi, Sajjad Gholami, Zeyu Zhao, Mehrsan Javan, and Philippe Desaulniers. "A Markov Game Model for Valuing Actions, Locations, and Team Performance in Ice Hockey." *Data Mining and Knowledge Discovery* 31, no. 6 (2017): 1735-757. doi:10.1007/s10618-017-0496-z.

Schulte, Oliver, Zeyu Zhao, and Kurt Routley. "What Is the Value of an Action in Ice Hockey? Learning a Q-function for the NHL."

Smith, Gerald. "A SHOT QUALITY ADJUSTED PLUS-MINUS FOR THE NHL." Fall 2016.

"Stats." NHL.com. http://www.nhl.com/stats/team?reportType=game&dateFrom=2017-05-01&dateTo=2017-05-01&gameType=2&gameLocation=H&filter=gamesPlayed%2Cgte%2C1&sort=gameDate.

Swartz, Tim B. "Hockey Analytics." *Wiley StatsRef: Statistics Reference Online*, 2017, 1-10. doi:10.1002/9781118445112.stat07965.

"Teams." NHL.com. https://www.nhl.com/info/teams.

Thomas, A. C., Samuel L. Ventura, Shane T. Jensen, and Stephen Ma. "Competing Process Hazard Function Models for Player Ratings in Ice Hockey." *The Annals of Applied Statistics* 7, no. 3 (2013): 1497-524. doi:10.1214/13-aoas646.

Tora, Moumita Roy, Jianhui Chen, and James J. Little. "Classification of Puck Possession Events in Ice Hockey." *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. doi:10.1109/cvprw.2017.24.

Tables

Table 1: Combined Data Variable Names and their Description

No.	Variable Name	First-Row Value	Definition	More information on the variables
1	wins	1	In the current game: 1 if the home team (h_team) wins and 0 otherwise	Data for the current season
2	season_start	2006	Year the current season started (Seasons are Oct-April). Season_start ranges from 2006 to 2016	Data for the current season
3	h_team_id	1	An ID code for the home team (h_team). Each team has been given a unique code. The ID code for the home team (h_team) ranges from 1 to 30. The NHL currently comprises of 31 teams (the Vegas Golden Knights joined in 2017).	Data for the current season
4	a_team_id	2	An ID code for the away team (a_team). The same code has been used for teams which was used in h_team_id. For example Anaheim Ducks was coded 1 in h_team_id and so it is also 1 in a_team_id.	Data for the current season
5	h_team	Anaheim Ducks	Name of the h_team (home team). There are a total of 30 teams in the dataset. The NHL currently comprises of 31 teams (the Vegas Golden Knights joined in 2017).	Data for the current season
6	a_team	Arizona Coyotes	Name of the a_team (away team). There are a total of 30 teams in the dataset. The NHL currently comprises of 31 teams (the Vegas Golden Knights joined in 2017).	Data for the current season
7	home_gamesplayed	4	Number of games the pair played in the prior season in which the h_team (home team) was at home	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
8	away_gamesplayed	4	Number of games the pair played in the prior season in which the h_team (home team) was away	This data is from the previous season. Beginning with pair-level (h

				and a) data. H is at home. A is away. Total is total of home plus away
9	total_gamesplayed	8	Total games the pair played in the prior season. (This is the total of home_games played and away_games played).	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
10	total_home_wins	4	Total number of home games between the pair in the prior season in which h (home team) won	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
11	total_home_goalsfor	19	Total goals h (home team) scored in the home games	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
12	total_home_goalsagainst	8	Total goals a (away team) scored against h when a was away	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
13	total_home_shotsfor	135	Total shots by h (home team) in the home games	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
14	total_home_shotsagainst	102	Total shots by a (away team) in the home games (when a was away)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
15	total_away_wins	3	Similar to total_home_wins, etc.; but now h is playing at	This data is from the previous

			a's rink	season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
16	total_away_goalsfor	13	All data is still in terms of h, so this is total goals h scored while playing a in a's rink	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
17	total_away_goalsagainst	8	Total goals a scored against h while playing h in its rink (a's rink).	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
18	total_away_shotsfor	127	Total shots h scored against a while playing a in a's rink.	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
19	total_away_shotsagainst	101	Total shots a scored against h while playing h in its rink (a's rink).	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
20	total_wins	7	Total times h won when the pair played (adding total_away_wins and total_home_wins)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
21	total_goalsfor	32	Total goals h scored when the pair played (adding total_away_goalsfor and total_home_goalsfor)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away

22	total_goalsagainst	16	Total goals a scored against h when the pair played (adding total_away_goalsagainst and total_home_goalsagainst)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
23	total_shotsfor	262	Total shots h scored against a when the pair played (total_away_shotsfor + total_home_shotsfor)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
24	total_shotsagainst	203	Total shots a scored against h when the pair played (total_away_shotsagainst + total_home_shotsagainst)	This data is from the previous season. Beginning with pair-level (h and a) data. H is at home. A is away. Total is total of home plus away
25	s_h_home_games	41	Total games h played at home in the prior season	This data is season-level (from the prior season)
26	s_h_home_wins	26	Total home games h won in the prior season	This data is season-level (from the prior season)
27	s_h_home_gf	146	Total goals h scored in home games	This data is season-level (from the prior season)
28	s_h_home_ga	113	Total goals against h in h's home games	This data is season-level (from the prior season)
29	s_h_home_sf	1378	Total shots from h in h's home games	This data is season-level (from the prior season)
30	s_h_home_sa	1221	Total shots against h in h's home games	This data is season-level (from the prior season)
31	s_h_away_games	41	Total games h played away in the prior season	This data is season-level (from the prior season)
32	s_h_away_ga	109	Total goals against h in h’s away games	This data is season-level (from the prior season)
33	s_h_away_gf	105	Total goals h scored in away games	This data is season-level (from the prior season)
34	s_h_away_sa	1210	Total shots against h in h's away games	This data is season-level (from the prior season)

35	s_h_away_sf	1207	Total shots from h in h's away games	This data is season-level (from the prior season)
36	s_h_away_wins	17	Total away games h won in the prior season	This data is season-level (from the prior season)
37	s_a_home_games	41	Total games a played at home in the prior season	This data is season-level (from the prior season)
38	s_a_home_wins	19	a's wins while a was at home.	This data is season-level (from the prior season)
39	s_a_home_gf	129	Total goals a scored in home games	This data is season-level (from the prior season)
40	s_a_home_ga	134	Total goals against a in a's home games	This data is season-level (from the prior season)
41	s_a_home_sf	1206	Total shots from a in a's home games	This data is season-level (from the prior season)
42	s_a_home_sa	1190	Total shots against a in a's home games	This data is season-level (from the prior season)
43	s_a_away_games	41	Total games a played away in the prior season	This data is season-level (from the prior season)
44	s_a_away_ga	134	Total goals against a in a’s away games	This data is season-level (from the prior season)
45	s_a_away_gf	113	Total goals a scored in away games	This data is season-level (from the prior season)
46	s_a_away_sa	1290	Total shots against a in a's away games	This data is season-level (from the prior season)
47	s_a_away_sf	1112	Total shots from a in a's away games	This data is season-level (from the prior season)
48	s_a_away_wins	19	Total away games a won in the prior season	This data is season-level (from the prior season)
49	s_h_total_games	82	h's total games (sums s_h_away_games and s_h_home_games)	This data is season-level (from the prior season)
50	s_h_total_wins	43	h's total wins (sums s_h_home_wins and s_h_away_wins)	This data is season-level (from the prior season)
51	s_h_total_gf	251	Total goals h scored in all games (sums s_h_away_gf + s_h_home_gf)	This data is season-level (from the prior season)
52	s_h_total_ga	222	Total goals scored by a in all games (sums s_h_away_ga +	This data is season-level (from the prior season)

			s_h_home_ga)	
53	s_h_total_sf	2585	Total shots scored by h in all games (sums s_h_away_sf + s_h_home_sf)	This data is season-level (from the prior season)
54	s_h_total_sa	2431	Total shots scored by a in all games (sums s_h_away_sa + s_h_home_sa)	This data is season-level (from the prior season)
55	s_a_total_games	82	a's total games (sums a_h_away_games and a_h_home_games)	This data is season-level (from the prior season)
56	s_a_total_wins	38	a's total wins (sums a_h_home_wins and a_h_away_wins)	This data is season-level (from the prior season)
57	s_a_total_gf	242	Total goals a scored in all games (sums s_a_away_gf + s_a_home_gf)	This data is season-level (from the prior season)
58	s_a_total_ga	268	Total goals scored by h in all games (sums s_a_away_ga + s_a_home_ga)	This data is season-level (from the prior season)
59	s_a_total_sf	2318	Total shots scored by h in all games (sums s_a_away_sf + s_a_home_sf)	This data is season-level (from the prior season)
60	s_a_total_sa	2480	Total shots scored by h in all games (sums s_a_away_sa + s_a_home_sa)	This data is season-level (from the prior season)

Table 2: Summary Statistics for the Combined Data Variables

No.	Variable	Obs	Mean	Std. Dev.	Min	Max
1	wins	12,255	.5459812	.4979016	0	1
2	season_start	12,255	2010.952	3.241435	2006	2016
3	h_team_id	12,255	15.54761	8.659943	1	30
4	a_team_id	12,255	15.44047	8.649044	1	30
5	h_team	0				
6	a_team	0				
7	home_gamesplayed	12,255	2.012893	1.021395	0	4
8	away_gamesplayed	12,255	2.100775	.9227791	0	4
9	total_gamesplayed	12,255	4.113668	1.879569	1	8
10	total_home_wins	12,255	1.09898	.9126595	0	4
11	total_home_goalsfor	12,255	5.860302	3.933103	0	24
12	total_home_goalsagainst	12,255	5.312199	3.646372	0	22
13	total_home_shotsfor	12,255	62.15251	33.40103	0	175
14	total_home_shotsagainst	12,255	58.18809	31.1661	0	165
15	total_away_wins	12,255	.9521828	.8475225	0	4
16	total_away_goalsfor	12,255	5.541656	3.492846	0	22
17	total_away_goalsagainst	12,255	6.125908	3.737716	0	24
18	total_away_shotsfor	12,255	60.69621	28.5655	0	165
19	total_away_shotsagainst	12,255	64.86348	30.60304	0	175
20	total_wins	12,255	2.051163	1.442916	0	8
21	total_goalsfor	12,255	11.40196	6.550154	0	40
22	total_goalsagainst	12,255	11.43811	6.517448	0	40
23	total_shotsfor	12,255	122.8487	58.79993	10	333
24	total_shotsagainst	12,255	123.0516	58.57346	11	333
25	s_h_home_games	12,255	40.0151	3.971728	24	41
26	s_h_home_wins	12,255	21.87067	4.636413	8	32
27	s_h_home_gf	12,255	116.2094	18.97832	52	163
28	s_h_home_ga	12,255	104.9262	18.38032	43	153
29	s_h_home_sf	12,255	1234.515	159.3875	630	1540
30	s_h_home_sa	12,255	1158.209	149.9161	546	1450
31	s_h_away_games	12,255	40.0151	3.971728	24	41
32	s_h_away_ga	12,255	116.2376	19.23558	43	160
33	s_h_away_gf	12,255	104.9206	17.25695	45	157
34	s_h_away_sa	12,255	1234.665	155.8768	562	1532
35	s_h_away_sf	12,255	1157.986	141.237	586	1432
36	s_h_away_wins	12,255	18.13847	4.644099	4	31
37	s_a_home_games	12,255	40.0151	3.971728	24	41
38	s_a_home_wins	12,255	21.88576	4.649034	8	32
39	s_a_home_gf	12,255	116.1575	18.95197	52	163
40	s_a_home_ga	12,255	104.7951	18.32124	43	153
41	s_a_home_sf	12,255	1233.859	159.2678	630	1540
42	s_a_home_sa	12,255	1157.03	149.8149	546	1450
43	s_a_away_games	12,255	40.0151	3.971728	24	41
44	s_a_away_ga	12,255	116.1193	19.2386	43	160
45	s_a_away_gf	12,255	104.8007	17.23009	45	157
46	s_a_away_sa	12,255	1233.556	155.8756	562	1532
47	s_a_away_sf	12,255	1157.257	141.2789	586	1432
48	s_a_away_wins	12,255	18.13537	4.653932	4	31
49	s_h_total_games	12,255	80.03019	7.943455	48	82

50	s_h_total_wins	12,255	40.00914	8.22887	15	58
51	s_h_total_gf	12,255	221.13	33.73807	109	313
52	s_h_total_ga	12,255	221.1639	35.07718	97	310
53	s_h_total_sf	12,255	2392.501	291.7653	1244	2965
54	s_h_total_sa	12,255	2392.875	298.4439	1110	2945
55	s_a_total_games	12,255	80.03019	7.943455	48	82
56	s_a_total_wins	12,255	40.02113	8.24323	15	58
57	s_a_total_gf	12,255	220.9582	33.67337	109	313
58	s_a_total_ga	12,255	220.9144	34.99102	97	310
59	s_a_total_sf	12,255	2391.115	291.6806	1244	2965
60	s_a_total_sa	12,255	2390.585	298.299	1110	2945

Table 3: Independent Variables used for Data Analytics in STATA

Independent Variables	**Formula to generate these variables from the variables in the combined dataset**
pair_total_goal_diff	total_goalsfor / (total_goalsfor + total_goalsagainst)
pair_home_goal_diff	total_home_goalsfor / (total_home_goalsfor + total_home_goalsagainst)
pair_total_shot_diff	total_shotsfor / (total_shotsfor + total_shotsagainst)
pair_home_shot_diff	total_home_shotsfor / (total_home_shotsfor + total_home_shotsagainst)
high_home_shot_diff#c.pair_home_shot_diff	1) pair_home_shot_diff = total_home_shotsfor / (total_home_shotsfor + total_home_shotsagainst) 2) obtain the mean and SD of pair_home_shot_diff 3) cutoff = .5159517 + 2*.0665424 (2 SDs above the mean of pair_home_shot_diff) 4) high_home_shot_diff = pair_home_shot_diff > cutoff
high_home_goal_diff#c.pair_home_goal_diff	1) pair_home_goal_diff = total_home_goasfor / (total_home_goasfor + total_home_goalsagainst) 2) obtain the mean and SD of pair_home_goal_diff 3) cutoff_for_high_home_goal_diff = .5236112 + 2 * .1772372 (2 SDs above the mean of pair_home_goal_diff) 4) high_home_goal_diff = pair_home_goal_diff > cutoff
at_least_4_gp#c.pair_home_shot_diff	1) pair_home_shot_diff = total_home_shotsfor / (total_home_shotsfor + total_home_shotsagainst) 2) at_least_4_gp = (total_gamesplayed >= 4)
at_least_4_gp#c.pair_home_goal_diff	1) pair_home_goal_diff = total_home_goalsfor / (total_home_goalsfor + total_home_goalsagainst) 2) at_least_4_gp = (total_gamesplayed >= 4)
at_least_4_gp#c.high_home_shot_diff	1) obtain the mean and SD of pair_home_shot_diff 2) cutoff = .5159517 + 2*.0665424 (2 SDs above the mean of pair_home_shot_diff) 3) high_home_shot_diff = pair_home_shot_diff > cutoff 4) at_least_4_gp = (total_gamesplayed >= 4)
at_least_4_gp#c.high_home_goal_diff	1) obtain the mean and SD of pair_home_goal_diff 2) cutoff_for_high_home_goal_diff = .5236112 + 2 * .1772372 (2 SDs above the mean of pair_home_goal_diff) 3) high_home_goal_diff = pair_home_goal_diff > cutoff

	4) at_least_4_gp = (total_gamesplayed >= 4)
s_h_goal_diff	s_h_total_gf / (s_h_total_gf + s_h_total_ga)
s_a_goal_diff	s_a_total_gf / (s_a_total_gf + s_a_total_ga)
s_h_shot_diff	s_h_total_sf / (s_h_total_sf + s_h_total_sa)
s_a_shot_diff	s_a_total_sf / (s_a_total_sf + s_a_total_sa)

Table 4: Dependent Variable used for Data Analytics in STATA

Variable Name	Definition	More information on the variables
wins	In the current game: 1 if the home team (h_team) wins and 0 otherwise	Data for the current season

Table 5: Summary Statistics for the Independent Variables used for Data Analytics in STATA

Variable	Obs	Mean	Std. Dev.	Min	Max
pair_total_goal_diff	12,255	.4973014	.1348512	0	1
pair_home_goal_diff	11,451	.5236112	.1772372	0	1
pair_total_shot_diff	12,255	.4987062	.055408	.2222222	.754717
pair_home_shot_diff	11,458	.5159517	.0665424	.2142857	.7555556
high_home_shot_diff	12,255	.0862505	.2807451	0	1
high_home_shot_diff#c.pair_home_shot_diff					
0	11,458	.5006665	.0982887	0	.6489362
1	11,458	.0152853	.1003703	0	.7555556
high_home_goal_diff	12,255	.0864137	.2809854	0	1
high_home_goal_diff#c.pair_home_goal_diff					
0	11,451	.5017533	.1798213	0	.875
1	11,451	.0218579	.1449612	0	1
at_least_4_gp	12,255	.6756426	.4681534	0	1
at_least_4_gp#c.pair_home_shot_diff					
0	11,458	.1423699	.233763	0	.7555556
1	11,458	.3735819	.2369929	0	.720339
at_least_4_gp#c.pair_home_goal_diff					
0	11,451	.1449623	.2645759	0	1
1	11,451	.3786488	.2668367	0	1

at_least_4_gp#c.high_home_goal_diff					
0	12,255	.0814361	.2735149	0	1
1	12,255	.0049776	.0703789	0	1
at_least_4_gp#c.high_home_shot_diff					
0	12,255	.0767034	.2661311	0	1
1	12,255	.0095471	.0972458	0	1
s_h_goal_d~f	12,255	.500241	.0411756	.3625592	.6056911
s_a_goal_d~f	12,255	.5003253	.041307	.3625592	.6056911
s_h_shot_d~f	12,255	.5000619	.0266734	.4045677	.5936842
s_a_shot_d~f	12,255	.5001586	.0267474	.4045677	.5936842

Table 6: Summary Statistics for the Dependent Variable used for Data Analytics in STATA

Variable	Obs	Mean	Std. Dev.	Min	Max
wins	12,255	.5459812	.4979016	0	1

Table 7: Results for Model 1

Model 1 tests for HO. The R-squared value for model 1 is 0.0171. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P> | t |	95% Confidence Interval	
s_h_goal_diff	.8538761***	.1350008	6.32	0.000	.5892533	1.118499
s_a_goal_diff	-.7375015***	.1339363	-5.51	0.000	-1.000038	-.4749651
s_h_shot_diff	.6626108***	.2078358	3.19	0.001	.2552198	1.070002
s_a_shot_diff	-.58422***	.2071057	-2.82	0.005	-.9901798	-.1782603
constant	.4486843***	.1215265	3.69	0.000	.2104732	.6868955

Table 8: Results for Model 2

The R-squared value for model 2 is 0.0184. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>| t |	95% Confidence Interval	
pair_total_goal_diff	-.0673315	.0563194	-1.20	0.232	-.1777272	.0430641
pair_home_goal_diff	.0510786	.0368311	1.39	0.166	-.0211166	.1232738
pair_total_shot_diff	-.1672632	.166058	-1.01	0.314	-.4927653	.1582388
pair_home_shot_diff	.1264728	.1125571	1.12	0.261	-.0941583	.347104
s_h_goal_diff	.9296193***	.1454045	6.39	0.000	.6446016	1.214637
s_a_goal_diff	-.7526232***	.144302	-5.22	0.000	-1.03548	-.469766
s_h_shot_diff	.6423394***	.2446473	2.63	0.009	.1627888	1.12189
s_a_shot_diff	-.6643747***	.2433942	-2.73	0.006	-1.141469	-.187280
constant	.4928018***	.1413548	3.49	0.000	.2157221	.7698814

Table 9: Results for Model 3

The R-squared value for model 3 is 0.0183. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>| t |	95% Confidence Interval	
high_home_shot_diff# c.pair_home_shot_diff						
0	-.001431	.0875668	-0.02	0.987	-.1730769	.1702148
1	.0847782	.0845212	1.00	0.316	-.0808978	.2504542
s_h_goal_diff	.9208151***	.1391152	6.62	0.000	.6481255	1.193505
s_a_goal_diff	-.7340466***	.1383817	-5.30	0.000	-1.005298	-.462794
s_h_shot_diff	.5460408**	.2287084	2.39	0.017	.0977332	.9943483
s_a_shot_diff	-.6000849***	.2281537	-2.63	0.009	-1.047305	-.152864
constant	.4780709***	.1338305	3.57	0.000	.2157401	.7404016

Table 10: Results for Model 4

Model 4 tests H3. The R-squared value for model 4 is 0.0181. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>| t |	95% Confidence Interval	
at_least_4_gp# c.pair_home_shot_diff						
0	.0530145	.0836061	0.63	0.526	-.1108678	.2168967
1	.0449656	.0829273	0.54	0.588	-.1175861	.2075173
s_h_goal_diff	.914482***	.1390984	6.57	0.000	.6418253	1.187139
s_a_goal_diff	-.7341311***	.1383867	-5.30	0.000	-1.005393	-.4628694
s_h_shot_diff	.5496714**	.228732	2.40	0.016	.1013175	.9980253
s_a_shot_diff	-.5901715***	.2280588	-2.59	0.010	-1.037206	-.1431373
constant	.4507413***	.1328719	3.39	0.001	.1902896	.7111929

Table 11: Results for Model 5

Model 5 tests for H1 and H4. The Pseudo R2 value for model 5 is 0.0133. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	z	P>| z |	95% Confidence Interval	
pair_home_goal_diff	.0895248	.1135535	0.79	0.430	-.13303	.31208
s_h_goal_diff	3.651478***	.5862511	6.23	0.000	2.5024	4.8005
s_a_goal_diff	-2.955167***	.5822029	-5.08	0.000	-4.0962	-1.814
s_h_shot_diff	2.478078***	.8873037	2.79	0.005	.73899	4.2171
s_a_shot_diff	-2.562506***	.8835839	-2.90	0.004	-4.2942	-.8307
constant	-.1690229	.5215752	-0.32	0.746	-1.1912	.85324

Table 12: Results for Model 6

Model 6 tests H2A. The R-squared value for model 6 is 0.0184. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>| t |	95% Confidence Interval	
high_home_goal_diff# c.pair_home_goal_diff						
0	.0002913	.0301181	0.01	0.992	-.05874	.0593
1	.0637815*	.0355164	1.80	0.073	-.00583	.1333
s_h_goal_diff	.9057059***	.142403	6.36	0.000	.62657	1.184
s_a_goal_diff	-.7231443***	.1407431	-5.14	0.000	-.9990	-.4472
s_h_shot_diff	.5956577***	.2150133	2.77	0.006	.1741	1.0171
s_a_shot_diff	-.6308187***	.2147761	-2.94	0.003	-1.0518	-.2098
constant	.4697605***	.126628	3.71	0.000	.2215	.7179

Table 13: Results for Model 7

Model 7 tests H3. The R-squared value for model 7 is 0.0181. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>|t|	95% Confidence Interval	
at_least_4_gp# c.pair_home_goal_ diff						
0	.0255249	.0291042	0.88	0.380	-.0315243	.082574
1	.0180238	.0292349	0.62	0.538	-.0392815	.0753292
s_h_goal_diff	.8941504***	.1423018	6.28	0.000	.6152145	1.173086
s_a_goal_diff	-.7179209***	.1407674	-5.10	0.000	-.9938492	-.4419927
s_h_shot_diff	.6004778***	.2149665	2.79	0.005	.1791066	1.021849
s_a_shot_diff	-.6238729***	.2146788	-2.91	0.004	-1.04468	-.2030657
Constant	.4580594***	.1264389	3.62	0.000	.2102175	.7059014

Table 14: Results for Model 8

Model 8 tests H2. The Pseudo R2 value for model 8 is 0.0133. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	z	P>| z|	95% Confidence Interval	
pair_home_goal_diff	.0552279	.070810	0.78	0.435	-.08355	.19401
s_h_goal_diff	2.278726***	.364580	6.25	0.000	1.56416	2.9932
s_a_goal_diff	-1.83666***	.362421	-5.07	0.000	-2.54700	-1.12633
s_h_shot_diff	1.53892***	.551838	2.79	0.005	.457335	2.62050
s_a_shot_diff	-1.59684***	.550507	-2.90	0.004	-2.67581	-.517868
constant	-.106294	.324430	-0.33	0.743	-.742166	.5295782

Table 15: Results for Model 9

Model 9 tests H2. The Pseudo R2 value for model 9 is 0.0132. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	z	P>| z|	95% Confidence Interval	
pair_home_shot_diff	.1963445	.338812	0.58	0.562	-.467715	.860404
s_h_goal_diff	3.743244***	.573695	6.52	0.000	2.618822	4.86766
s_a_goal_diff	-3.026981***	.572690	-5.29	0.000	-4.14943	-1.9045
s_h_shot_diff	2.265304**	.942846	2.40	0.016	.41736	4.11324
s_a_shot_diff	-2.420666***	.937805	-2.58	0.010	-4.25873	-.58260
constant	-.1979272	.548052	-0.36	0.718	-1.27209	.876236

Table 16: Results for Model 10

Model 10 tests H2. The Pseudo R2 value for model 10 is 0.0132. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	z	P>| z |	95% Confidence Interval	
pair_home_shot_diff	.1209214	.2114218	0.57	0.567	-.29345	.5353005
s_h_goal_diff	2.335138***	.3566952	6.55	0.000	1.6360	3.034248
s_a_goal_diff	-1.881056***	.3564091	-5.28	0.000	-2.5796	-1.18250
s_h_shot_diff	1.407913**	.5866474	2.40	0.016	.25810	2.557721
s_a_shot_diff	-1.509792***	.5845949	-2.58	0.010	-2.6555	-.364007
constant	-.1237733	.341035	-0.36	0.717	-.79218	.5446431

Table 17: Results for Model 11

Model 11 tests H2C. The R-squared value for model 11 is 0.0187. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>| t|	95% Confidence Interval	
high_home_shot_diff# c.pair_home_shot_diff						
0	.0082827	.087779	0.09	0.925	-.16377	.180345
1	.0869229	.084612	1.03	0.304	-.07893	.252778
high_home_goal_diff# c.pair_home_goal_diff						
0	.000259	.030132	0.01	0.993	-.05880	.059323
1	.0617537*	.035559	1.74	0.082	-.00794	.131457
s_h_goal_diff	.9127724***	.142431	6.41	0.000	.63358	1.19196
s_a_goal_diff	-.7225126***	.140763	-5.13	0.000	-.99843	-.44659
s_h_shot_diff	.5425112**	.228850	2.37	0.018	.09392	.991097
s_a_shot_diff	-.590945***	.228306	-2.59	0.010	-1.0384	-.14342
constant	.4671351***	.135008	3.46	0.001	.20249	.731774

Table 18: Results for Model 12

Model 12 tests H3. The R-squared value for model 12 is 0.0182. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>|t|	95% Confidence Interval	
at_least_4_gp# c.pair_home_shot_ diff						
0	.0531052	.0880082	0.60	0.546	-.1194059	.2256164
1	.0516174	.0862845	0.60	0.550	-.1175151	.2207498
at_least_4_gp# c.pair_home_goal_ diff						
0	.0257637	.0372307	0.69	0.489	-.0472148	.0987422
1	.0192207	.037673	0.51	0.610	-.0546248	.0930662
s_h_goal_diff	.8940184***	.1428752	6.26	0.000	.6139585	1.174078
s_a_goal_diff	-.7163164***	.1410353	-5.08	0.000	-.9927698	-.439863
s_h_shot_diff	.5514326**	.2292014	2.41	0.016	.1021586	1.000707
s_a_shot_diff	-.5752119**	.2283899	-2.52	0.012	-1.022895	-.1275286
constant	.4301828***	.1338365	3.21	0.001	.1678404	.6925253

Table 19: Results for Model 13

Model 13 tests H2a and H3. The R-squared value for model 13 is 0.0176. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>|t|	95% Confidence Interval	
at_least_4_gp# c.high_home_goal_ diff						
0	.0345542**	.0162596	2.13	0.034	.0026829	.0664256
1	-.0796312	.0662539	-1.20	0.229	-.2094993	.0502369
s_h_goal_diff	.8566745***	.1351132	6.34	0.000	.5918313	1.121518
s_a_goal_diff	-.7396786***	.1340421	-5.52	0.000	-1.002422	-.476935
s_h_shot_diff	.6628194***	.2078213	3.19	0.001	.2554568	1.070182
s_a_shot_diff	-.5806325***	.2071552	-2.80	0.005	-.9866894	-.1745756
constant	.4440575***	.1215134	3.65	0.000	.2058721	.6822429

Table 20: Results for Model 14

Model 14 tests H2b and H3. The R-squared value for model 13 is 0.0174. *, **, and *** measure significance at the 10%, 5%, and 1% level respectively.

Independent Variables	Coefficient	Robust Standard Error	t	P>|t|	95% Confidence Interval	
at_least_4_gp# c.high_home_goal_ diff						
0	.0246166	.0167298	1.47	0.141	-.0081764	.0574095
1	.049742	.0437039	1.14	0.255	-.0359246	.1354086
s_h_goal_diff	.8577171***	.1350216	6.35	0.000	.5930533	1.122381
s_a_goal_diff	-.7372755***	.1339532	-5.50	0.000	-.9998449	-.474706
s_h_shot_diff	.6327267***	.2090789	3.03	0.002	.2228992	1.042554
s_a_shot_diff	-.562706***	.2077476	-2.71	0.007	-.969924	-.1554879
constant	.4484702***	.1215396	3.69	0.000	.2102333	.686707

Appendix

Appendix 1: Getting the pair-wise game-level data from the NHL stats website

I collected the game level data from the following NHL stats website: http://www.nhl.com/stats/team?reportType=game&dateFrom=2005-04-30&dateTo=2017-05-01&gameType=2&gameLocation=H&filter=gamesPlayed,gte,1&sort=gameDate. Season - level data and pair-wise game-level data from the 2005-2006 season to the 2015-2016 season has been used to predict winning percentages for the pairs in each of the following seasons.

In order get the game level data, change the setting to the following on the NHL stats website:

1) Select "TEAMS"
2) Select "GAME BY GAME"
3) Change the dates of "GAME BY GAME" to what is required (I did 1st October 2005 to 30th April 2017)
4) Under "GAME TYPE" select "Regular Season"
5) Under "REPORT" select "Team Summary"
6) Click on "Refine Results" (this will allow you to further customize your search)
7) Click on "Run Report"
8) Under "Team" select "All Teams"
9) Under "Game" select "Home" (in order to avoid repetition of games)
10) Under "Game" also select "All Decisions"
11) Under "Opponent" select "All Teams"
12) Under "Filter results by" select "Games Played >= 1"
13) Click on "Game" to arrange by ascending/ descending order by date

The first attempt to get this data from the NHL stats website into Excel was in December 2017 that encountered the following challenges. Four different data pulls had to be

done for: (a) Penalty Kill % = 0, Power Play % >= 1, (b) Penalty Kill % = 0, Power Play % = 0, (c) Penalty Kill >= 1, Power Play % = 0, (d) Power Play % >= 1, Penalty Kill % >= 1. Four different data pulls had to be done because until December 2017 the NHL stats website left those cells blank for Power Play % and Penalty Kill % if Power Play % and Penalty Kill % were 0. This was creating issues when the OFFSET function was being used in the Excel spreadsheet to convert the pasted data from the NHL stats website, which all came in one column, into different rows. However, as of January 2018 the NHL Stats website has fixed this issue. Thus, if Power Play % or Penalty Kill % is 0, then the website lists it as 0 and not as a blank, which it previously did.

Now I am going to describe the step-by-step methodology to obtain the data from the NHL website into excel, which I have used for this thesis:

1) Copy 1 page (50 games) of data from the NHL stats website and paste it in Excel. The data for the 50 games will come in one single row in Excel. You can keep pasting until Excel allows you to. Just to ensure my Excel did not crash I pasted no more than 50 sheets (that is 2500 games) in one Excel file.
2) In order to get the data in different columns from one single row, use the following formula in Excel and then drag as necessary. The formula is: "=OFFSET (A1,(ROW()-1)*24+INT((COLUMN()-3)),0)." After using this formula I dragged from cell "C1" to "Z1" as I wanted 24 columns. Then I selected C1 to Z1 and dragged it down to create multiple rows of 24 columns each.
3) After this, in Excel click on "Data," then on "Text to Columns," then on "Delimited" in order to separate data in "Game" to "Date" and "Away Team."
4) Then, open another Excel workbook. In this workbook type the full names of all ice hockey teams in one column and next to this columns also type out the abbreviated names of all the ice hockey teams.

5) Now from your data file (in which you have all the game level data) copy and paste the column with the abbreviated away team names in the file you created above (in bullet 4).
6) Now use the Vlookup function in the bullet 4 file to get away team full names. Then paste the full names in your master data file (where you have all the data).
7) Convert the master data file into a table. This will make it easier to arrange the data in the manner you want. For instance in order of ascending/ descending order by date.

This methodology still encountered one problem. I had got most of the data from the NHL stats website into Excel during winter break when I was in India (which is one day ahead of USA). This is why initially there was a difference of one day between the dates in my dataset and the dates on the NHL stats website (when I access the website in USA). This may sound extremely strange because one would expect to see the local time of games irrespective of the geographic location. To confirm this anomaly, I video called one of my friends in India and told him to open the NHL stats website. When one opens the NHL stats website in India versus in USA there is actually a difference of one day. I then corrected the dates in my dataset such that they showed the dates as per the time in USA. After making this change my dataset matches the data one would have obtained by downloading the dataset in USA.

There was still one challenge with this dataset. The difficulty was that Atlanta Thrashers changed its name to Winnipeg Jets in the 2011-12 season and Phoenix Coyotes changed its name to Arizona Coyotes in the 2014-2015 season. The old names of these two teams were changed to their new names in STATA.

Appendix 2: Game Level Data Variables (Only those variables are described which have been used either directly or indirectly (that is in order to create other variables) for this study):

Variables	Description
Goals for	Total goals scored by a team against the opposition team
Goals against	Total goals scored by the opposition team in a game
Shots for	Total shots scored by a team against the opposition team
Shots against	Total shots scored by the opposition team in a game
Home team	Team that's playing in the usual area they play in
Road team	Travelling team

Appendix 3: Season-Level Data Variables (In bold are those variables which have been used either directly for this study or which were used in order to create other variables for this study):

Variables	Description
GP	**games played**
W	**wins**
L	losses
T	ties
OT	overtime losses
P	points
ROW	regulation plus overtime wins
P_perc	point percentage
GF	**goals for**
GA	**goals against**
S/O Wins	shootout games won
GF/GP	goals for per game
GA/GP	goals against per game
PP%	power play percentage
PK%	penalty kill percentage
Shots/GP	**shots per game played**
SA/GP	**shots against per game played**
FOW%	faceoff win %

Appendix 4: Some of the rules of ice hockey

Official NHL (National Hockey League) rules can be found on the following website: http://www.nhl.com/nhl/en/v3/ext/rules/2017-2018-NHL-rulebook.pdf. However, I am describing below some of the rules of ice hockey that may help the reader understand and appreciate the game better.

The game is one hour long with three periods of 20 minutes each. Each team has six players. Two are defenders, three are forwards, and one is a goalkeeper. The three forwards are called left wing, center, and right wing. Out of the forwards, the center takes faceoff (which is like kick off in soccer). Two people are involved in a faceoff. There are different types of faceoff and there are 4 circles for penalties / if game needs to be stopped.

Substitutions happen on the fly (without any time off). One can pass the puck behind the goal. Can rotate (change) players as many times as one wants. If puck goes really high then one cannot bring it down with his stick. If the goalie stops the puck for a few seconds within the crease then faceoff takes place at that goalie's end of the court.

Icing occurs if a player hits the puck from behind his defense line to behind the goal of the other team and the other team's player touches it. The point of this is to waste time. Icing is allowed during power play. If not in power play then icing results in faceoff from the circle close to the goal of the team whose player committed the foul.

During playoffs, overtime periods are repeated until a team scores a goal to win the game. Shootout takes place if a game is tied and no one scores in overtime. Overtime cannot last more than 5 minutes and ends as soon as someone scores. Power play occurs if a player commits a foul and is taken out as a result. Team who committed the foul is called shorthand and the other team is called power play. The whole idea of power play is to bring out your best player to try and score. Penalties result in player sitting in penalty box for 2, 4, 5 or 10 minutes. Penalized team plays shorthanded.

A shorthanded goal is if a team manages to score when shorthanded. Assist: last two players to touch the puck before someone scores. Empty net: when the team removes it goalie for an extra player (desperation move to score a goal). Take away: steal puck. Give away: loose puck. Major penalties for fighting, it's still a 5 on 5, and it doesn't end even after scoring. Penalty results in overplay (2 minute power play). 2 minutes power play ends if either someone scores or if time ends.