Forecasting Real-Time Win Probability in NHL Games

Author: Christophe Bernier

Forecasting Real-Time Win Probability in NHL Games

Christophe Bernier
Senior Honors Thesis
Department of Economics
Boston College

Advisor: Christopher Maxwell, Ph.D.

Table of Contents

Abstract

Uncertainty is a key part of any sports game; without it, there is little reason to be interested in the outcome. This thesis attempts to quantify the uncertainty inherent in NHL hockey games by building a real-time win probability model that estimates both teams' likelihood of winning based on what has happened in the game so far. The model is built using historical data from the 2009-2010 season all the way to the 2016-2017 season. Given the differential and the time left, the model evaluates historical data for that specific game-state and calculates a win probability. The model also uses a multi-regression approach to incorporate pre-game Vegas odds as a way to factor the strength of both teams; to my knowledge, this is the first publicly available hockey win probability model to do so. Finally, the model also factors in elements unique to the sport of hockey, like power plays and shootout periods.

Acknowledgments

I owe this thesis to Professor Christopher Maxwell, my thesis advisor and my mentor. His guidance, patience, and trust were instrumental in completing this thesis. I am especially thankful for his help with data scraping to build my datasets and with STATA coding. I want to thank Professor Bob Murphy and the Boston College Economics Department for allowing me to pursue this passion of mine. I want to thank Stephen Pettigrew, whose model gave me the inspiration to build my own. Finally, I want to thank my family and friends for the continuous support and genuine interest they displayed for my work.

1. Introduction

Your favorite hockey team is up by one goal at the end of the second period. Obviously you're feeling good about the game, but you know a lot could happen in the third period. What are the odds that your team actually holds the lead and wins the game? Should you accept the weighted bet that your friend just proposed you, or is the return unfavorable?

These are the questions that this thesis attempts to address by developing a real-time win probability model for NHL hockey games. Though prevalent in other sports like baseball, football, and basketball, win probability models have taken longer to adapt to the sport of hockey. This paper uses methods developed in other hockey win probability models as well as more advanced methods used in basketball and football models.

2. Win Probability Models in Other Sports

Win probability models began with baseball, a sport that can be easily analyzed as distinct sequential events rather than one continuous event. Each moment of the game can be characterized by a game-state, which typically captures at least the following: the score differential, the current inning, the number of outs and which bases are taken. Given the relatively low number of distinct game-states and the large dataset of historical games, it is possible to accurately predict the win expectancy based on the historical final results of that game-state.

Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
9	Bottom	0	0				0.634
9	Bottom	0	0	1st			0.715
9	Bottom	0	0	1st	2nd		0.816
9	Bottom	0	0	1st	2nd	3rd	0.934
9	Bottom	0	0	1st		3rd	0.929
9	Bottom	0	0		2nd		0.807
9	Bottom	0	0		2nd	3rd	0.928
9	Bottom	0	0			3rd	0.914
Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
9	Bottom	0	1				0.577
9	Bottom	0	1	1st			0.637
9	Bottom	0	1	1st	2nd		0.711
9	Bottom	0	1	1st	2nd	3rd	0.835
9	Bottom	0	1	1st		3rd	0.829
9	Bottom	0	1		2nd		0.703
9	Bottom	0	1		2nd	3rd	0.839
9	Bottom	0	1			3rd	0.830
Inning	Top/Bottom	Score	Outs	1B	2B	3B	WE
9	Bottom	0	2				0.532
9	Bottom	0	2	1st			0.562
9	Bottom	0	2	1st	2nd		0.613
9	Bottom	0	2	1st	2nd	3rd	0.662
9	Bottom	0	2	1st		3rd	0.642
9	Bottom	0	2		2nd		0.610
9	Bottom	0	2		2nd	3rd	0.639
9	Bottom	0	2			3rd	0.637

Figure 1: Win expectancy for every game-state of a tied game in the bottom of the ninth inning, compiled by Tom Tango (2007)

In Figure 1, we can see the difference that no bases and loaded bases have on win expectancy is considerable, especially when there are no outs. We also see that fully loaded bases are considerably more favorable when there are no outs than when there are two outs. Such win expectancies are calculable for every game-state.

Approaches inspired by this one were created to analyze other sports. However, defining the game-state in other sports is not as clear and definite as it is in baseball. Most sports, like football, basketball, and hockey, run on continuous time, and score differentials can be extremely variable. Models are therefore adapted to incorporate the complexities associated with the sport they analyze.

For basketball, two things to consider are the large amount of differentials possible and the considerable variance in strengths of the two teams. Tackling these challenges, Phil Everson

and Jimmy Charite (2013) developed a model that estimates the distribution of the final margin of victory (MOV) of the home team at the end of regulation. Theoretically, the probability that the MOV falls above 0 is the win probability. This approach allows flexibility and avoids the need to have win expectancy tied to a specific game-state; given that differentials often reach twenty points or higher, there would often be a lack of available data to accurately calculate the win expectancy. Their model is based on three factors: the amount of time remaining in the game, the betting spreads at the beginning of the game, or the predicted MOV, and the current score differential. Understandably, at the beginning of the game, the betting spread will have more weight than the score differential, which can be variable. However, at the end of the game, the score differential will have much more weight than the betting spread. Figure 2 demonstrates the relationship between the regression coefficient and minutes remaining, and we can see how the model changes as time goes on. We also see that the decrease in the spread coefficient is not quite linear; it will therefore be interesting to see if this is the same for hockey. At each point in the game, the mean and variance of the distribution are be fitted using the results of Figure 2, and from the distribution, the probability that the MOV is greater than 0 is calculated. This percentage is equivalent to the win probability.

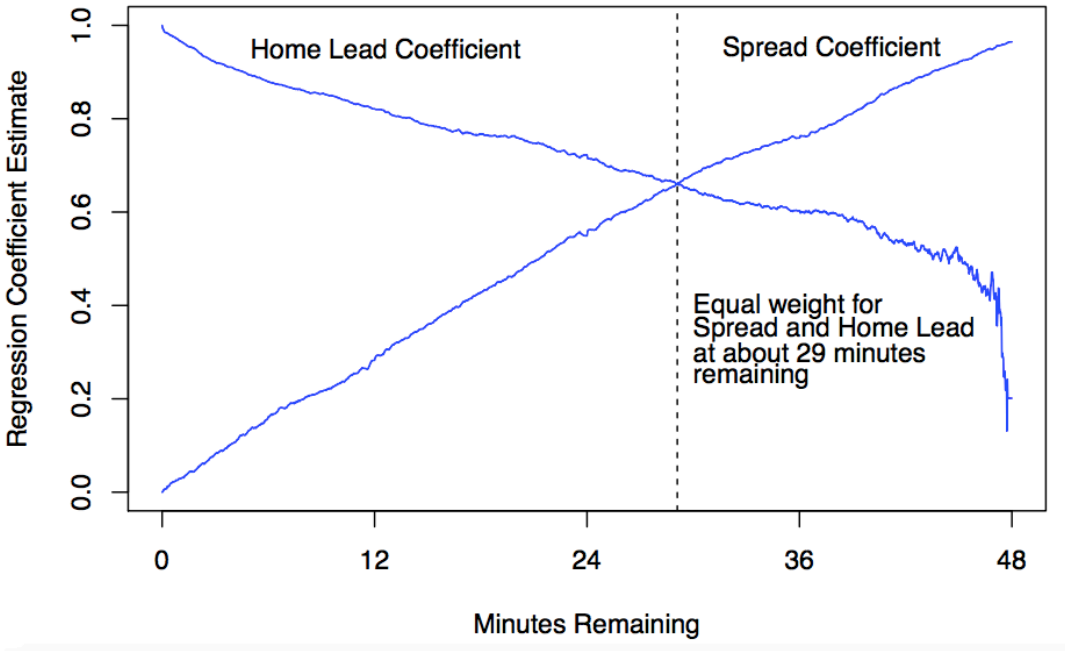


Figure 2: Regression of Home MOV and Current Home Lead, taken from Everson and Charite's model (2013). "Spread Coefficient" refers to the Vegas expected spread, and "Home Lead Coefficient" refers to the actual score

Just like basketball models, football win expectancy models are usually based on three major factors: score differential, time remaining, and betting spreads prior to the game. However, there are other key factors that need to be considered: which team is in possession of the ball, the number of downs, and the field position. For example, the win probability of a team that is up by one with five seconds left will be extremely different if the ball is on their own 10-yard line or on their opponent's 10-yard line. To address this caveat, the "Pro-Football-Reference" model first calculates the win probability model assuming neutral possession and field position by normalizing the standard deviation based on the time left in the game—see Figure 3. Then, the model calculates the expected points: based on historical data of field position and number of

downs, what is the expected net point from scoring drive for the offensive team? The model then recalculates the win probability based on the expected points of the next scoring play.

```
(1-NORMDIST(((away_margin)+0.5),(-home_vegas_line*(45/60)),
(13.45/SQRT((60/45))),TRUE))+(0.5*(NORMDIST(((away_margin)+0.5),
(-home_vegas_line*(45/60)),(13.45/SQRT((60/45))),TRUE)-
NORMDIST(((away_margin)-0.5),(-home_vegas_line*(45/60)),
(13.45/SQRT((60/45))),TRUE)))
```

Figure 3: Example of the Excel formula used to calculate win expectancy after one quarter. The formula assumes neutral possession and position (Pro-Football-Reference)

For each sport, win probability models are influenced by the chronology of the game and the score differential. Then, models are adapted to take the specific characteristics of the sports into account. Hockey is no different. Given that differentials are not very variable—generally the differential is within 3 goals—our approach will be closer to the game-state one that baseball uses, with the game-state being defined as a two-value vector consisting of time and score differential. However, just like in Everson and Charite's model, the model will need to account for the tradeoff between betting odds and score differential. The model will also include elements specific to hockey, like power plays, overtime, and shootouts.

3. Stephen Pettigrew's Win Probability Model

As a starting point for this paper, I analyzed a model developed by Stephen Pettigrew, a doctoral student at Harvard University (Pettigrew, 2014). He published demonstrations of the model and its methodology on his website, Rink Stats. Pettigrew bases his model on four metrics: time remaining, score differential, power plays, and shootout percentages. As previously discussed, he defines the game-state as a two-dimensional vector consisting of time and score

differential. He therefore calculated historical frequencies of wins based on the current game-state, which are represented in Figure 3.

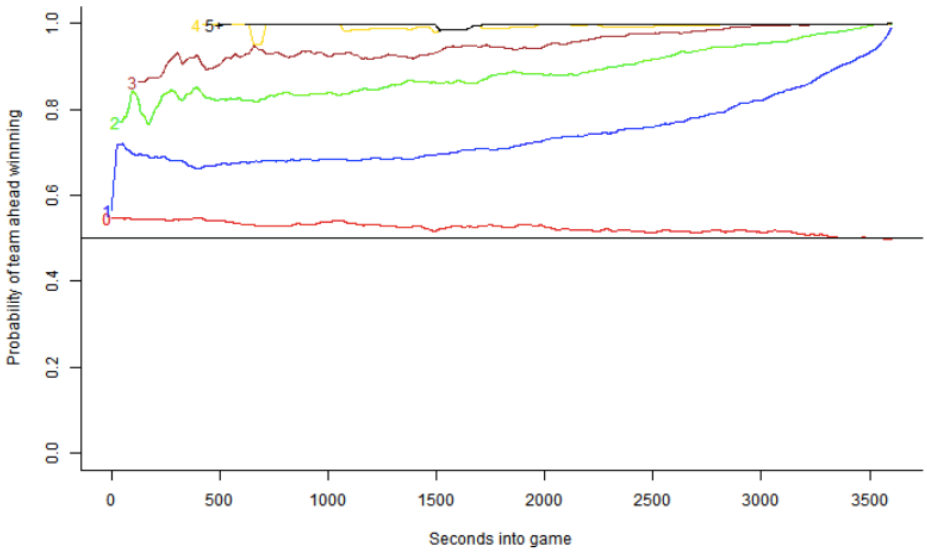


Figure 4: Smoothed empirical probabilities of home team winning conditional on goal differential, taken from Pettigrew's Rink Stats (2014)

These results are very intuitive. We see that the home team win expectancy begins at 55% and converges to 50% as time goes on. For all the other differentials, the win expectancy approaches 100% with increasing marginal expectancy. Therefore, the score will determine which curve is used, and the time will determine where on the curve the value is taken. When a goal is scored, the model will "shift" one curve up or one curve down depending on which team scored the goal.

This game-state represents the bulk of the model. But the model also takes into account power plays by using conditional probabilities. Indeed, the model calculates the probability that a goal will be scored based on the time left in the penalty, and then incorporates the effect that such a goal would have on the win expectancy. Figure 5 shows the formula used.

$$
\begin{aligned}
P(home\ win) \quad = \quad & P(home\ win|current\ score) * P(no\ SHG) * P(no\ PPG) + \\
& P(home\ win|current\ score) * P(1\ SHG) * P(1\ PPG) + \\
& P(home\ win|current\ score + 1) * P(no\ SHG) * P(PPG) + \\
& P(home\ win|current\ score - 1) * P(SHG) * P(no\ PPG)
\end{aligned}
$$

Figure 5: Pettigrew's formula to account for power plays (Pettigrew, 2014)

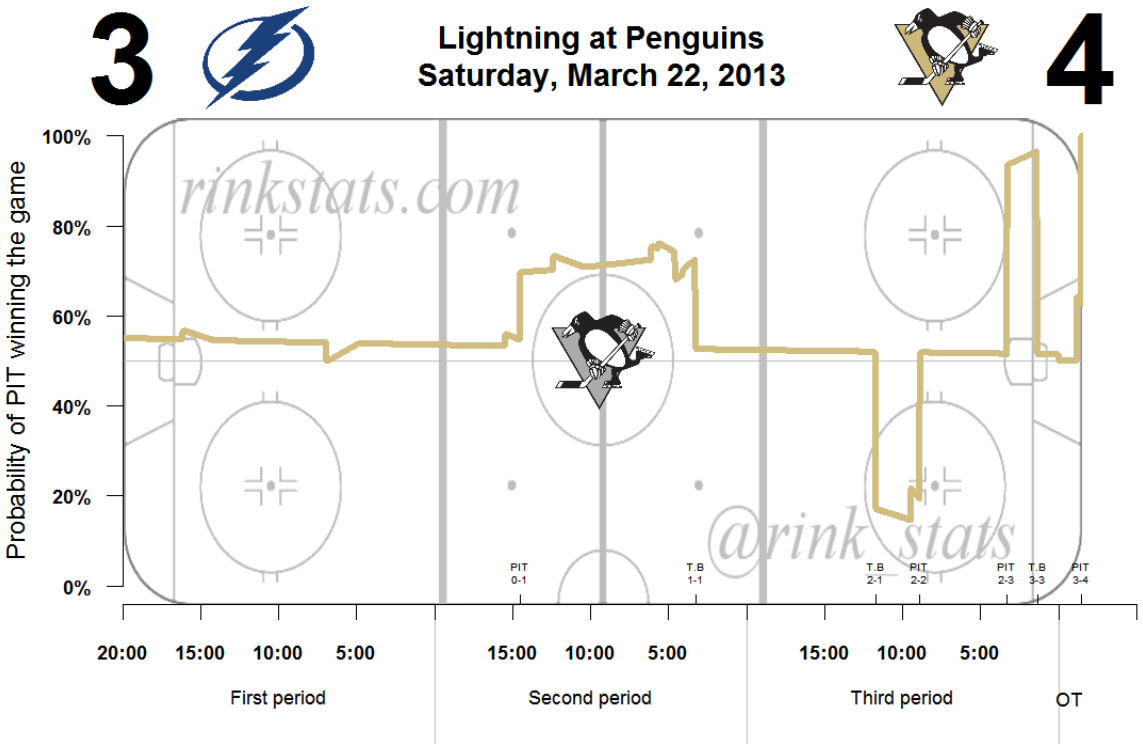


Figure 6: Pettigrew's model applied to a game, taken from Rink Stats (Pettigrew, 2014)

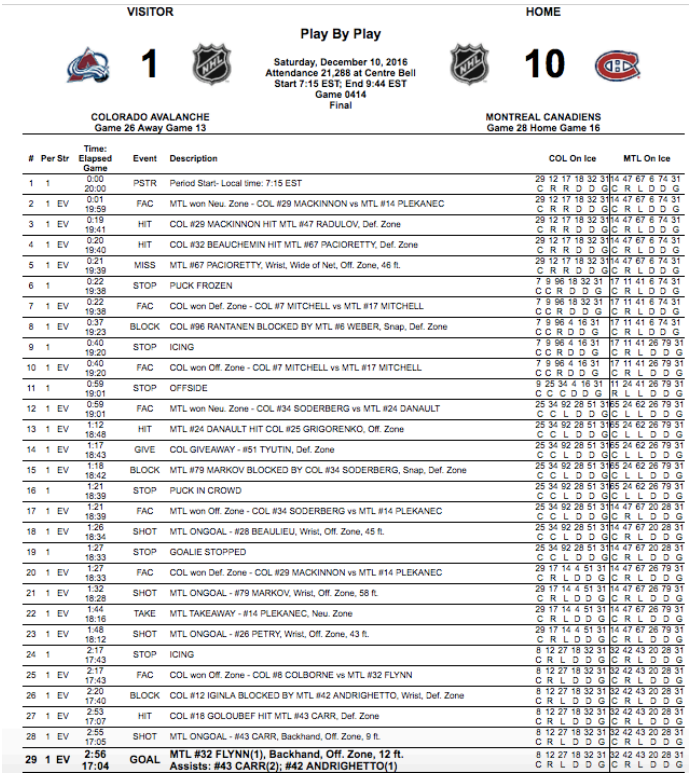
In Figure 6, we can see the model applied retrodictively to a hockey game played on March 22nd, 2013. On the time axis, we can see when goals were scored, and we can see the associated shift in the win probability model. We can also observe smaller shifts—these represent power plays opportunities. We notice that for power plays, the curve follows a saw

tooth shape: a jump at the beginning of the power play, followed by a gradual decline as time goes on.

Though well constructed, one main shortcoming that my model will attempt to address is the omission of betting odds as the starting point. Indeed, the model always gives the home team a 55% win probability at the beginning of the game, which assumes equal strength. Using betting odds would account for the difference in strength as well as home-ice advantage. To incorporate Vegas odds, we will need to model the diminishing weight that they have as the game goes on, similar to what we saw in Figure 2.

4. Data Collection

To build my own model, I first wanted to compile the widest dataset possible. I decided to look at games starting with the 2009-2010 season, which constitutes almost 10,000 games. I first needed the score at every second of all these games, and who ended up winning the game. Using this, I would be able to reconstruct the game-state graph created by Pettigrew (Figure 4). Given that I did not know exactly what data I would need later on, I searched for the data source that would give me the most information possible, and I would narrow it down from there. This led me to NHL Play-By-Play reports, which are official reports in HTML format posted by the NHL after every game. Figure 7 displays one example of these reports: each line represents a different event. Events can be anything from goals to shots to face-offs to whistles. The line incorporates the event type, the description, the time at which the event occurred, and who was on the ice at the time of the event.



VISITOR

HOME

Play By Play

1

Saturday, December 10, 2016
Attendance 21,288 at Centre Bell
Start 7:15 EST; End 9:44 EST
Game 0414
Final

10

COLORADO AVALANCHE
Game 26 Away Game 13

MONTREAL CANADIENS
Game 28 Home Game 16

#	Per	Str	Time: Elapsed Game	Event	Description	COL On Ice	MTL On Ice
1	1		0:00 20:00	PSTR	Period Start- Local time: 7:15 EST	29 12 17 18 32 31 C R R D D G	14 47 67 6 74 31 C R L D D G
2	1	EV	0:01 19:59	FAC	MTL won Neu. Zone - COL #29 MACKINNON vs MTL #14 PLEKANEC	29 12 17 18 32 31 C R R D D G	14 47 67 6 74 31 C R L D D G
3	1	EV	0:19 19:41	HIT	COL #29 MACKINNON HIT MTL #47 RADULOV, Def. Zone	29 12 17 18 32 31 C R R D D G	14 47 67 6 74 31 C R L D D G
4	1	EV	0:20 19:40	HIT	COL #32 BEAUCHEMIN HIT MTL #67 PACIORETTY, Def. Zone	29 12 17 18 32 31 C R R D D G	14 47 67 6 74 31 C R L D D G
5	1	EV	0:21 19:39	MISS	MTL #67 PACIORETTY, Wrist, Wide of Net, Off. Zone, 46 ft.	29 12 17 18 32 31 C R R D D G	14 47 67 6 74 31 C R L D D G
6	1		0:22 19:38	STOP	PUCK FROZEN	7 9 96 18 32 31 C C R D D G	17 11 41 6 74 31 C R L D D G
7	1	EV	0:22 19:38	FAC	COL won Def. Zone - COL #7 MITCHELL vs MTL #17 MITCHELL	7 9 96 18 32 31 C C R D D G	17 11 41 6 74 31 C R L D D G
8	1	EV	0:37 19:23	BLOCK	COL #96 RANTANEN BLOCKED BY MTL #6 WEBER, Snap, Def. Zone	7 9 96 4 16 31 C C R D D G	17 11 41 6 74 31 C R L D D G
9	1		0:40 19:20	STOP	ICING	7 9 96 4 16 31 C C R D D G	17 11 41 26 79 31 C R L D D G
10	1	EV	0:40 19:20	FAC	COL won Off. Zone - COL #7 MITCHELL vs MTL #17 MITCHELL	7 9 96 4 16 31 C C R D D G	17 11 41 26 79 31 C R L D D G
11	1		0:59 19:01	STOP	OFFSIDE	9 25 34 4 16 31 C C C D D G	11 24 41 26 79 31 R L L D D G
12	1	EV	0:59 19:01	FAC	MTL won Neu. Zone - COL #34 SODERBERG vs MTL #24 DANAULT	25 34 92 28 51 31 C C L D D G	65 24 62 26 79 31 C L L D D G
13	1	EV	1:12 18:48	HIT	MTL #24 DANAULT HIT COL #25 GRIGORENKO, Off. Zone	25 34 92 28 51 31 C C L D D G	65 24 62 26 79 31 C L L D D G
14	1	EV	1:17 18:43	GIVE	COL GIVEAWAY - #51 TYUTIN, Def. Zone	25 34 92 28 51 31 C C L D D G	65 24 62 26 79 31 C L L D D G
15	1	EV	1:18 18:42	BLOCK	MTL #79 MARKOV BLOCKED BY COL #34 SODERBERG, Snap, Def. Zone	25 34 92 28 51 31 C C L D D G	65 24 62 26 79 31 C L L D D G
16	1		1:21 18:39	STOP	PUCK IN CROWD	25 34 92 28 51 31 C C L D D G	65 24 62 26 79 31 C L L D D G
17	1	EV	1:21 18:39	FAC	MTL won Off. Zone - COL #34 SODERBERG vs MTL #14 PLEKANEC	25 34 92 28 51 31 C C L D D G	14 47 67 20 28 31 C R L D D G
18	1	EV	1:26 18:34	SHOT	MTL ONGOAL - #28 BEAULIEU, Wrist, Off. Zone, 45 ft.	25 34 92 28 51 31 C C L D D G	14 47 67 20 28 31 C R L D D G
19	1		1:27 18:33	STOP	GOALIE STOPPED	25 34 92 28 51 31 C C L D D G	14 47 67 20 28 31 C R L D D G
20	1	EV	1:27 18:33	FAC	COL won Def. Zone - COL #29 MACKINNON vs MTL #14 PLEKANEC	29 17 14 4 51 31 C R L D D G	14 47 67 26 79 31 C R L D D G
21	1	EV	1:32 18:28	SHOT	MTL ONGOAL - #79 MARKOV, Wrist, Off. Zone, 58 ft.	29 17 14 4 51 31 C R L D D G	14 47 67 26 79 31 C R L D D G
22	1	EV	1:44 18:16	TAKE	MTL TAKEAWAY - #14 PLEKANEC, Neu. Zone	29 17 14 4 51 31 C R L D D G	14 47 67 26 79 31 C R L D D G
23	1	EV	1:48 18:12	SHOT	MTL ONGOAL - #26 PETRY, Wrist, Off. Zone, 43 ft.	29 17 14 4 51 31 C R L D D G	14 47 67 26 79 31 C R L D D G
24	1		2:17 17:43	STOP	ICING	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G
25	1	EV	2:17 17:43	FAC	COL won Off. Zone - COL #8 COLBORNE vs MTL #32 FLYNN	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G
26	1	EV	2:20 17:40	BLOCK	COL #12 IGINLA BLOCKED BY MTL #42 ANDRIGHETTO, Wrist, Def. Zone	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G
27	1	EV	2:53 17:07	HIT	COL #18 GOLOUBEF HIT MTL #43 CARR, Def. Zone	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G
28	1	EV	2:55 17:05	SHOT	MTL ONGOAL - #43 CARR, Backhand, Off. Zone, 9 ft.	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G
29	**1**	**EV**	**2:56 17:04**	**GOAL**	**MTL #32 FLYNN(1), Backhand, Off. Zone, 12 ft. Assists: #43 CARR(2); #42 ANDRIGHETTO(1)**	8 12 27 18 32 31 C R L D D G	32 42 43 20 28 31 C R L D D G

Figure 7: Screenshot of the Play-By-Play report posted after a game played on December 16th 2016

To retrieve this information, I used a Python scraping program that reads every line of the report and writes the information in a text file, separating the information with pipes. The program then loops through every game of a season and appends each game to the text file. The program retrieves the data for one regular season, so the program is modified to retrieve data for a different season. After manipulating the data, I arrived to an Excel dataset containing the following: time of the event, details about the event, who was on the ice for that event, and the final score of the game associated with that event. The 2017 season alone contained almost 400,000 observations. Given that there is a total of 8 seasons, the total number of observations compiled is over 3 million. This dataset contains virtually all the information regarding what happened in every regular season game, and therefore constitutes the raw data of the model.

5. Analyzing Win-Frequencies Based on Game-States

The next step was to calculate the historical win frequencies of every game-state. To do so, I extracted all the goals from the raw dataset, resulting in a dataset of about 52,000 goals. Then, I modified the dataset into a matrix containing the score differential associated with the game ID and the number of seconds in the game. Therefore, each game is treated as a separate event, and the differential is mapped for every second of that game. This matrix can be treated as a function accepting the game ID and the time left and returning the score differential. Then, this matrix was separated into two separate matrices. The first one counted the number of occurrences of a certain goal differential at each point in time. The second matrix looked at the same criteria, but only counted instances where the home team ended up winning the game. Therefore, dividing every cell of the second matrix by the cells in the first matrix yielded the win frequency given the game-state, which can be observed in Figure 8. Note that differentials over five goals were combined in one category, and frequencies at times with less than ten occurrences were not calculated.

Figure 8 confirms Pettigrew's previous observations regarding game-state win frequencies. When the game is tied, the home ice advantage begins at around 55%, and progressively converges to 50% with time. Understandably, the curves are also roughly mirrored about the 50% line. We can still see the home ice-advantage when comparing leads for home teams to leads for visiting teams. For example, home teams with a one goal lead reach an 80% win probability about 400 seconds earlier than visiting teams do.

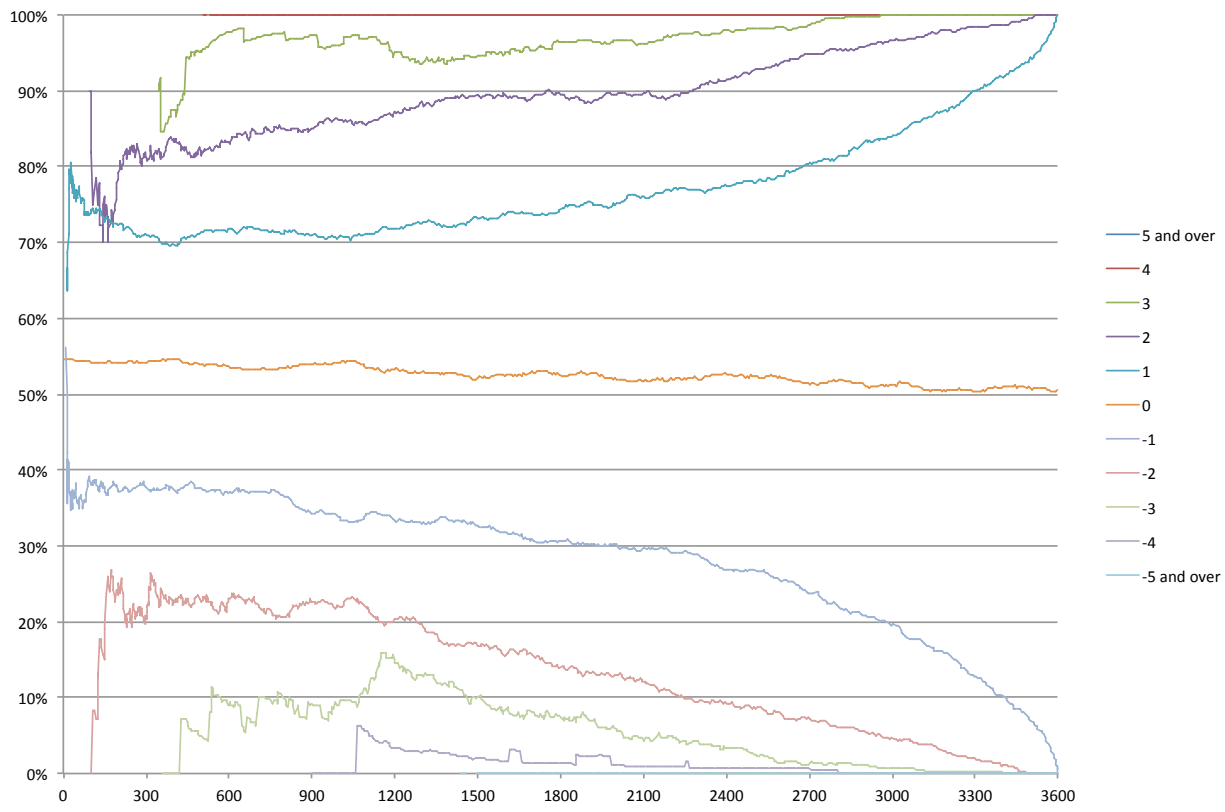


Figure 8: Home team win probabilities based on current score differential

However, the model we are looking to develop will not differentiate between home and away teams. Rather, the model will factor in betting odds at the start of the game as a measure of team strengths. Therefore, the split between home and away teams is unnecessary. Figure 9 combines them to display win probabilities for the leading team. As expected, the results are not extremely different, with the curves following a trend that is similar to the trends in Figure 8.

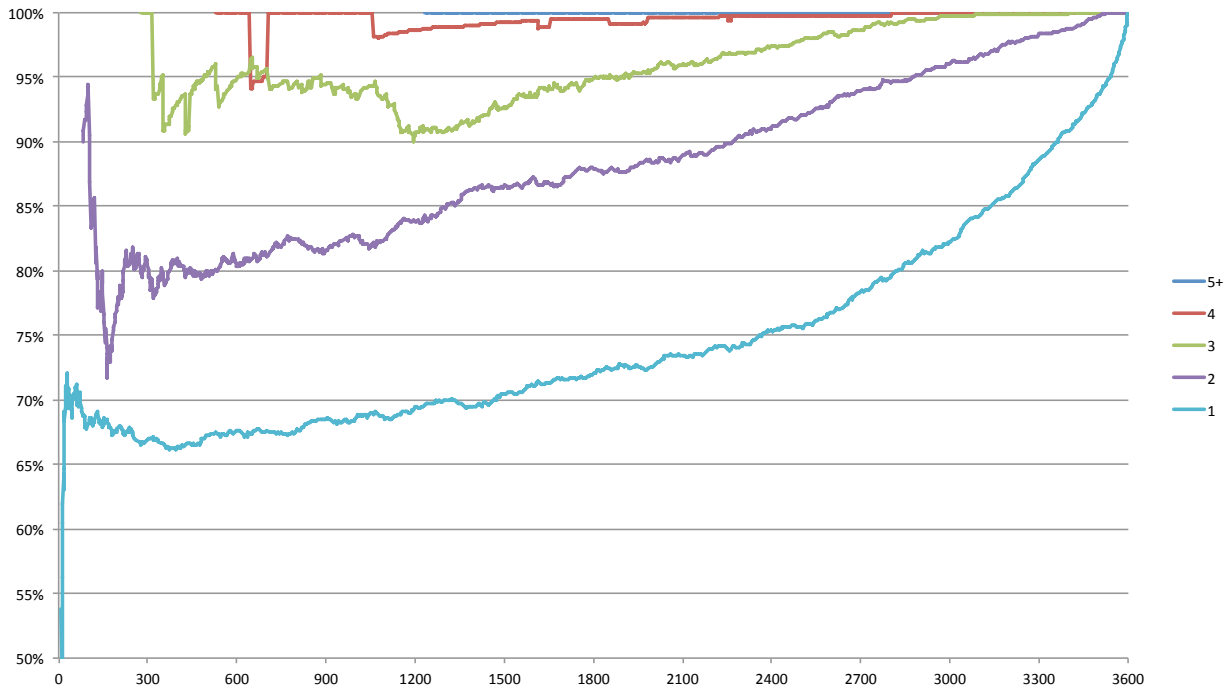


Figure 9: Leading team win probabilities based on current score differential

6. Incorporating Opening Betting-Odds

The next step in the construction of the model was to incorporate Vegas betting odds at the start of the game. This variable would be used as a way to control for the strengths of both teams. There is not one definitive source for betting odds, as different bookkeepers can assign different odds to the same game. Yet given that I needed eight years worth of historical betting odds, my options were slightly limited. I was able to extract the data I needed from the website "covers.com". The website includes betting odds for both teams for every game dating back to the 2008-2009 season. The odds retrieved are in the "American Odds" format. If the odds are positive, they refer to payout one would receive if they placed a $100 bet. For example, if the odds for the Montreal Canadiens to beat the Toronto Maple Leafs are +225, one would receive a $225 payout if they placed a $100 bet. If the odds are negative, they refer to the size of the bet

one needs to place to receive a payout of $100. For example, if the odds for the Montreal Canadiens to beat the Toronto Maple Leafs are -225, one would have to place a bet of $225 to receive a payout of $100.

From these betting odds I derived an implied win probability. The formula depends on whether the odds are positive or negative. Below are the two different formulas.

$$prob = \frac{100}{x_+ + 100} \times 100$$

Figure 10: Implied win probability for positive American Odds

$$prob = \frac{x_-}{x_- + 100} \times 100$$

Figure 11: Implied win probability for negative American Odds

Using the examples above, if the Canadiens had +225 odds, this would result in an implied probability of 30.8%. If on the other hand they had -225 odds, they would have an implied probability of 69.2%. Therefore, for each game since the 2008-2009 season, I not only had the American Odds for both teams, but also their implied win probability. However, for every bet placed, the dealer profits by taking a small cut of the bets. Therefore, the implied probabilities are inflated and add up to a number greater than a 100%. If the dealer took no cut from the bets, then the implied probabilities would add up to 100%. Yet since a cut is taken, the implied probability represents a number higher than the true implied probability. If the probabilities added up to a number less than 100%, then a bettor could make money without risk by placing a bet on both teams, creating an arbitrage opportunity. To account for this, I divided

the implied win probability by the sum of the two implied win probabilities for that game. By construction, those two add up to 100%. There are multiple ways to scale implied win probability, but I judged this to be the best way to go about it. This new probability was used as a team's win probability at the start of the game. For example, on October 22nd, the Canadiens played the Boston Bruins, with betting odds -116 and 103, respectively. This translates to implied probabilities of 53.7% and 48.8%. These odds add up to 102.5%, so I divided both by that value to get new probabilities of 52.4% and 47.6%, respectively.

Given this data, I compiled the win season averages of win probabilities for each team going back to 2008-2009. Though not directly related to the final model, this data represents an interesting analysis of which teams have been the most dominant in the past few years. I first differentiated teams by seasons to see which teams have had the most dominant regular season performances.

Figure 12 represents the top 20 teams in terms of their season-average opening odds. Apart from Detroit and San Jose in 2008-2009, the pack is clustered within close to 2% of each other, showing there is not much variation for dominant teams. An interesting observation to note is the low number of Stanley Cup wins. One would expect that the most dominant teams in the regular season would be much more likely to win the championship, but the relationship seems relatively weak. Since the 2008-2009 season, the average of the season-average opening odds for Stanley Cup champions is 56.3%, roughly the 87th percentile of all recorded seasons.

Team ID	Average Odds
Detroit 2008-2009	63.2%
San Jose 2008-2009	62.5%
Chicago 2009-2010 *	59.9%
Washington 2009-2010	59.9%
Chicago 2014-2015 *	59.7%
Boston 2012-2013	59.4%
Washington 2016-2017	59.3%
San Jose 2013-2014	59.1%
St. Louis 2013-2014	58.9%
Chicago 2013-2014	58.9%
Boston 2011-2012	58.7%
San Jose 2009-2010	58.7%
Vancouver 2010-2011	58.6%
Vancouver 2011-2012	58.4%
Boston 2013-2014	58.2%
Pittsburgh 2012-2013	58.1%
Pittsburgh 2016-2017 *	57.9%
Pittsburgh 2014-2015	57.8%
Boston 2008-2009	57.7%
Los Angeles 2015-2016	57.7%

Figure 12: Top 20 most dominant teams in terms of average opening odds. "*" denotes Stanley Cup champions.

Figure 13 displays the bottom 20 teams of the ranking. Though for the top teams the percentages were clustered, there is a lot more variation for the percentages of the bottom teams. The 2014-2015 Buffalo Sabres were by far the worst team in terms of opening odds. They ended the season with a 23-51-8 record, the worst of the league for that season.

Team ID	Average Odds
Buffalo 2014-2015	31.6%
NY Islanders 2008-2009	36.8%
Buffalo 2013-2014	37.7%
Edmonton 2014-2015	37.9%
Arizona 2016-2017	38.1%
Colorado 2016-2017	39.2%
Edmonton 2010-2011	39.7%
Columbus 2011-2012	39.8%
Arizona 2014-2015	40.2%
Vancouver 2016-2017	40.3%
Edmonton 2009-2010	40.4%
NY Islanders 2010-2011	40.5%
Winnipeg 2008-2009	41.0%
Calgary 2013-2014	41.3%
Tampa Bay 2008-2009	41.4%
Toronto 2015-2016	41.5%
Florida 2013-2014	41.8%
Buffalo 2015-2016	41.9%
Florida 2012-2013	42.0%
Arizona 2015-2016	42.1%

Figure 13: Top 20 least dominant teams in terms of average opening odds.

Finally, Figure 14 compiles the team averages for the past nine seasons, showing which franchises have been the most dominant in recent years. Chicago, San Jose, and Pittsburgh form the top 3. While both Chicago and Pittsburgh have 3 titles each, San Jose is still winless despite constantly performing well in the regular season. Edmonton, Buffalo, and Arizona take the bottom three spots.

Team	Average Odds	Titles
Chicago	56.9%	3
San Jose	56.8%	0
Pittsburgh	56.3%	3
Boston	56.2%	1
Washington	54.9%	0
Detroit	53.6%	0
Los Angeles	53.4%	2
NY Rangers	53.1%	0
St. Louis	52.7%	0
Vancouver	52.4%	0
Anaheim	51.8%	0
Montreal	51.2%	0
Philadelphia	51.1%	0
Nashville	50.2%	0
Minnesota	49.9%	0
Tampa Bay	49.8%	0
Dallas	49.0%	0
New Jersey	48.7%	0
Calgary	47.8%	0
Ottawa	47.6%	0
Carolina	46.9%	0
Columbus	45.7%	0
Winnipeg	46.3%	0
Toronto	46.0%	0
Florida	45.9%	0
Colorado	45.7%	0
NY Islanders	45.7%	0
Arizona	45.7%	0
Buffalo	45.1%	0
Edmonton	43.7%	0

Figure 14: Average of opening win probability odds and number of titles for the past 9 seasons

In the Appendix, I included the season-wide odds for each team dating back to the 2008-2009 season. Returning to the model, I combined the odds data with the original game dataset, so that each game has both team's win probabilities at the start of the game. Incorporating the odds

in the model proved to be a challenge. As discussed earlier, basketball and football models were based on a tradeoff between the score differential and the betting spread. As time goes on, the former gains weight and the latter loses weight in the equation. This tradeoff was something I hoped to replicate, but for hockey games. But the different natures of the models made it impossible to use their method directly. Instead of using percentage odds, the model used the spread from Vegas bookkeepers. The models also attempt to predict the final score and its distribution, while my model calculates the win probability directly from historical data. Given that their methodology was not replicable, I needed to come up with one myself.

My first thought was to replicate Figure 9, but separating teams in three tiers: favored, evenly-matched, and disfavored. Looking at the win probabilities of every game played in the past nine seasons, I used the 66^{th} and 33^{rd} percentiles to separate the lot in three tiers. These percentiles roughly corresponded to opening odds of 55% and 45% respectively. In other words, if the opening odds of a team were above 55%, the team was labeled as "favored." If their opening odds were less than 45%, they were labeled as "disfavored". If they were between the two, then the team was labeled as "evenly-matched." I used the same methodology discussed for the construction of Figure 8 and 9. Instead of coming up for a single line for each goal differential, I came up with three different lines for each differential based on which tier the team was placed in. I expected the lines of the same differential to follow similar trends, with the "favored" line above the other two and the "disfavored" line below the other two.

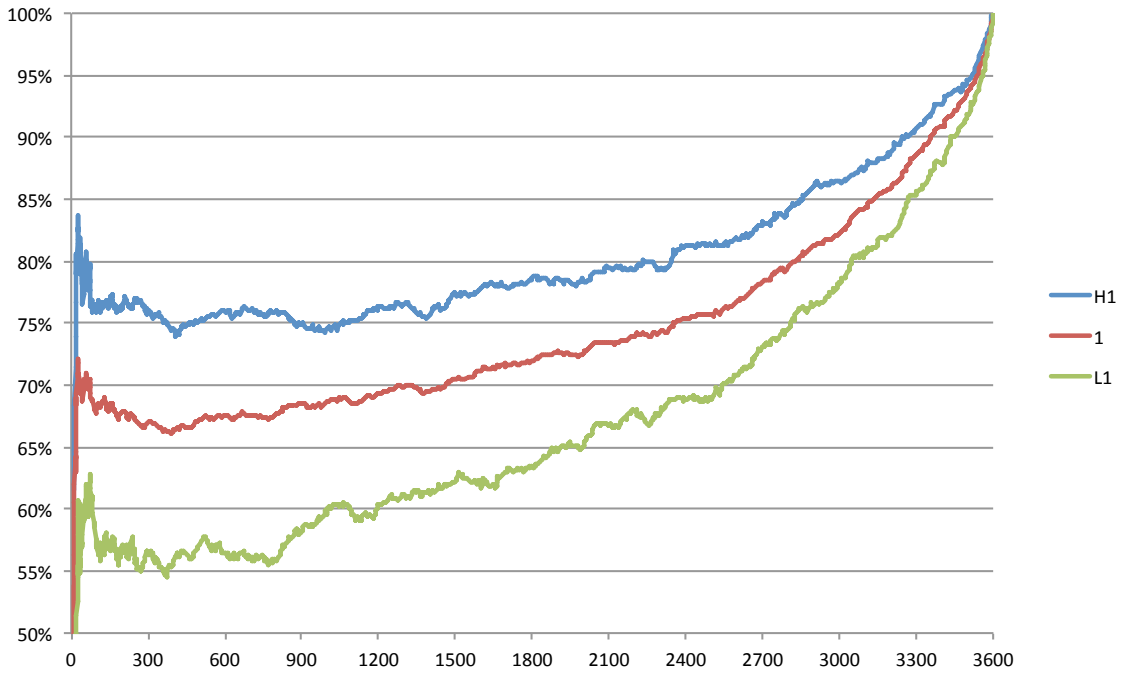


Figure 15: Win probabilities of team leading by one goal. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%

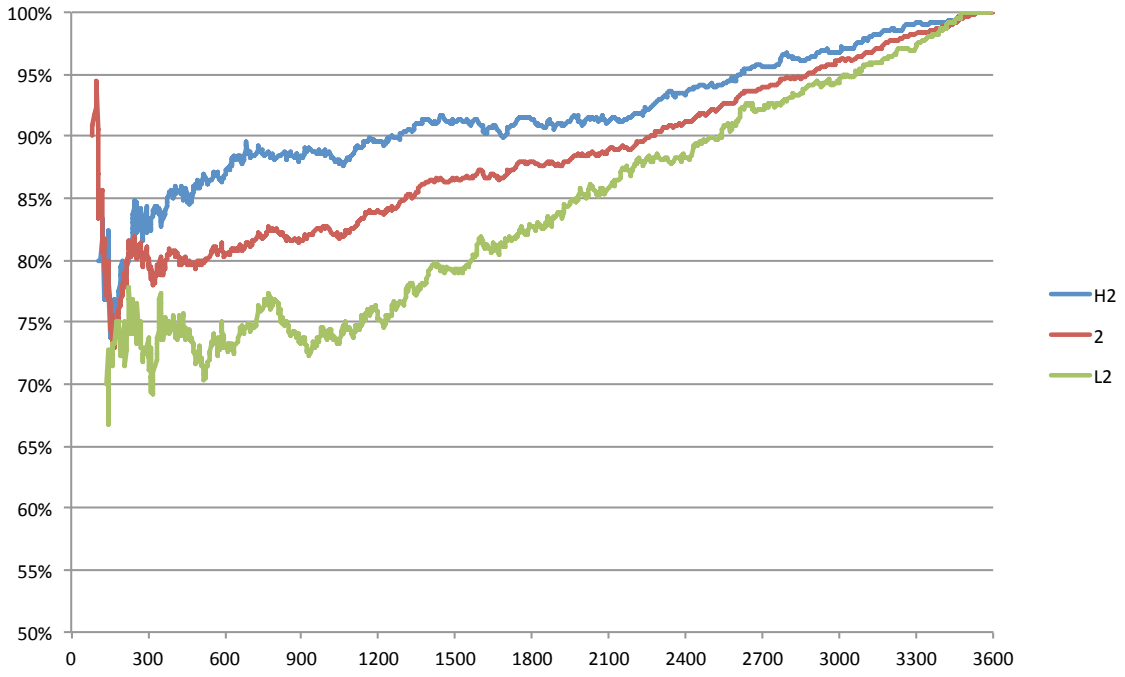


Figure 16: Win probabilities of team leading by two goals. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%

Figures 15 and 16 represent the real time win probabilities for teams up by one and two goals, respectively. Both display expected results. Looking at Figure 15, we see that all three curves follow a similar trajectory: concave up converging to 100%. We can also notice that the spread between the center curves and the other curves decreases as time goes on. We observe similar things with the two-goal differential, except the lines are more clustered towards 100% and the convergence is much more linear.

This method is one way to account for both goal differential and opening odds. It is very simple and intuitive and builds well on previous results. However, there are clear downfalls with the method. The main one is when a team has opening odds close to one of the cut-offs. A slight change in the odds can have a huge difference on the expected win probability, which would be a flaw in the model. Generally, the fact that two teams with very different odds follow the same line represents a flawed and primitive model. One option to remedy the situation is to include more segments, avoiding large gaps between small probabilities. However, the more segments there are, the less data points there are, and the less accurate the estimate it. This was therefore not an effective way to successfully introduce odds data in the model. Instead, I needed to find a way where a small difference in the odds will result in a similarly small change in the win probability.

I decided instead to create a regression-based model. Given a time and a goal differential, the model would have as dependent variable the binary "win" variable (1 if the leading team won the game and 0 if they lost) and would have as independent variable the opening odds. I ran this regression for every goal differential, at every 60-second interval. I also ran separate regressions if the leading team was home or away.

$$win_{(i,t)} = \beta_{1(i,t)} odds + \beta_{0(i,t)}$$

Figure 17: Equation for the regression, conditional on the differential *i* and the time *t*. The "win" variable is a dummy variable, and the "odds" variable is the opening odds minus 50%

For simplicity's sake, I subtracted 50% from the opening odds, so that if two teams were perfectly matched they would both have odds 0. Therefore, the constant term represents the win probability at a specific game-state if two teams are evenly matched. I expect this coefficient to follow a similar trend than the results observed in Figures 6 and 7. The first coefficient represents the weight that a difference from 50% has on the final odds. For example, if a team's opening odds are 60%, then the odds variable would have value 10%, and it would get multiplied by the coefficient β_1. If a team's opening odds are 40%, then the odds variable would be -10%. I expect the size of this coefficient to decrease as time goes on, since the predictive power of the score increases with time. This is analogous to the trend observed in basketball models (see Figure 2) where as time goes on, the weight of the differential increases and the weight of the betting odds decrease.

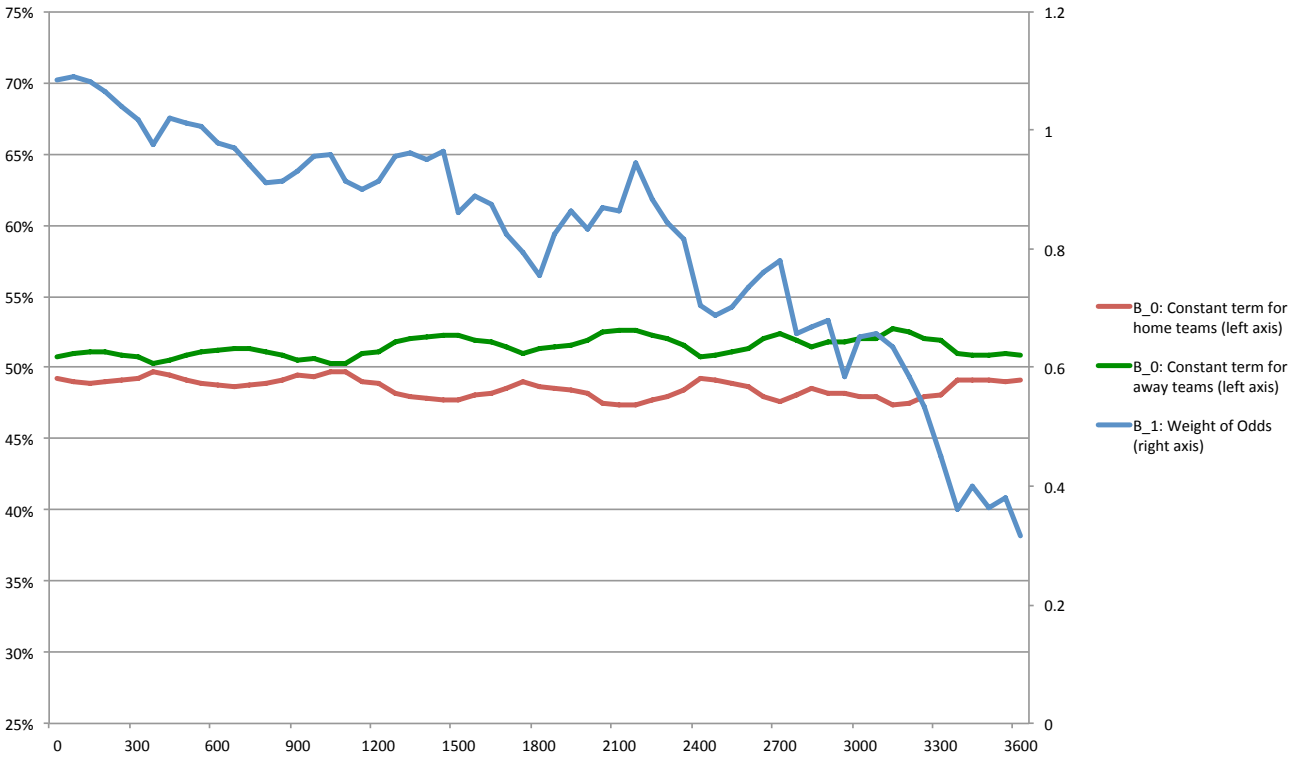


Figure 18: Regression coefficients where the two teams are tied

Figure 18 displays the regression coefficients for the situation where a game is tied. In this graph, each line represent the coefficient of a different regression, ran at 1-minute intervals. Naturally, β_1 will be the same if you calculate the win probability from the perspective of the home team or the away team. The β_1 line follows the expected trend. The weight starts at around 1 and decreases in a concave down fashion, which is what was observed in Everson and Charite's basketball model. I expected the two β_0 lines to be equal at 50%, yet the results are slightly different. The β_0 line for the away team hovers slightly over 50% while the β_0 line for home teams is slightly below 50%. This presents an issue. For example, if two teams both have 50% betting odds, then the model would favor the visiting team. That advantage climbs to 52.6% at minute 52.

There are several reasons why this discrepancy occurs. The first could be that the odds data gathered are biased towards home teams and inflate their odds. Given that I only had one data source available, then it could be the values were slightly off compared to the true market values. The second reason could be that the odds are correct, and this is simply the way to minimize least square residuals. It is possible that given that visiting teams tend to have lower odds, an adjustment is needed in the constant term.

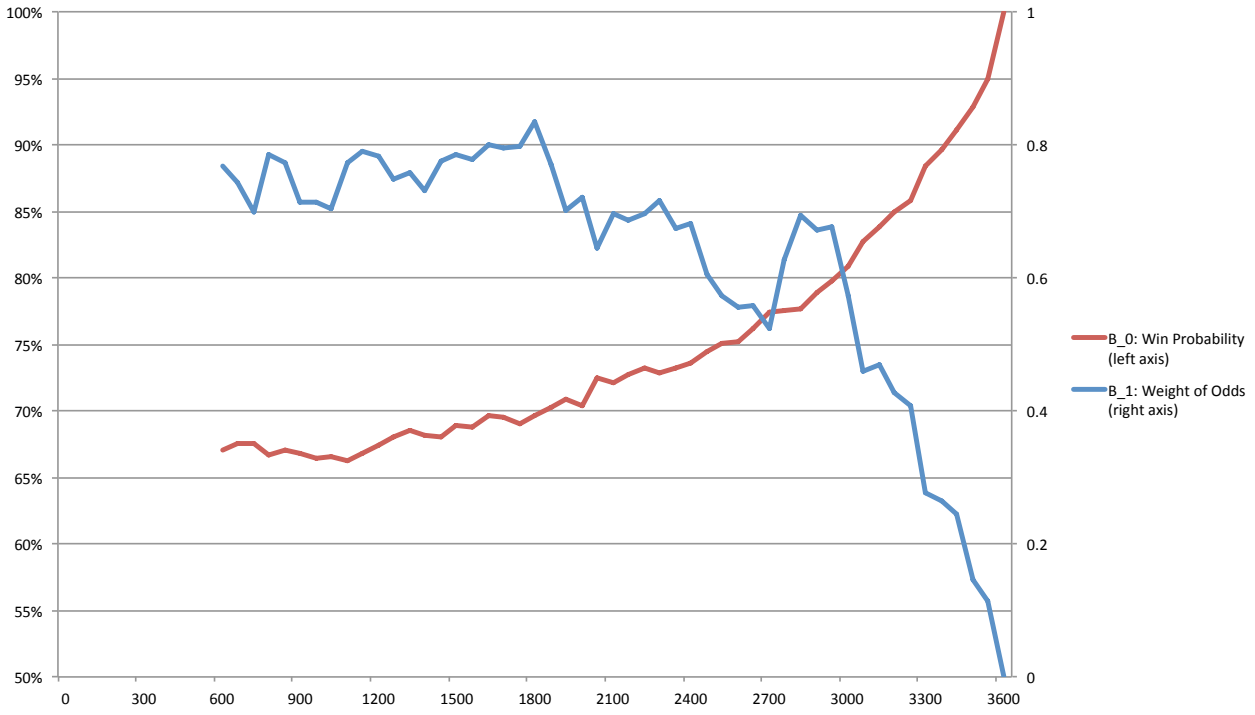


Figure 19: Regression coefficients where the home team is up by one goal

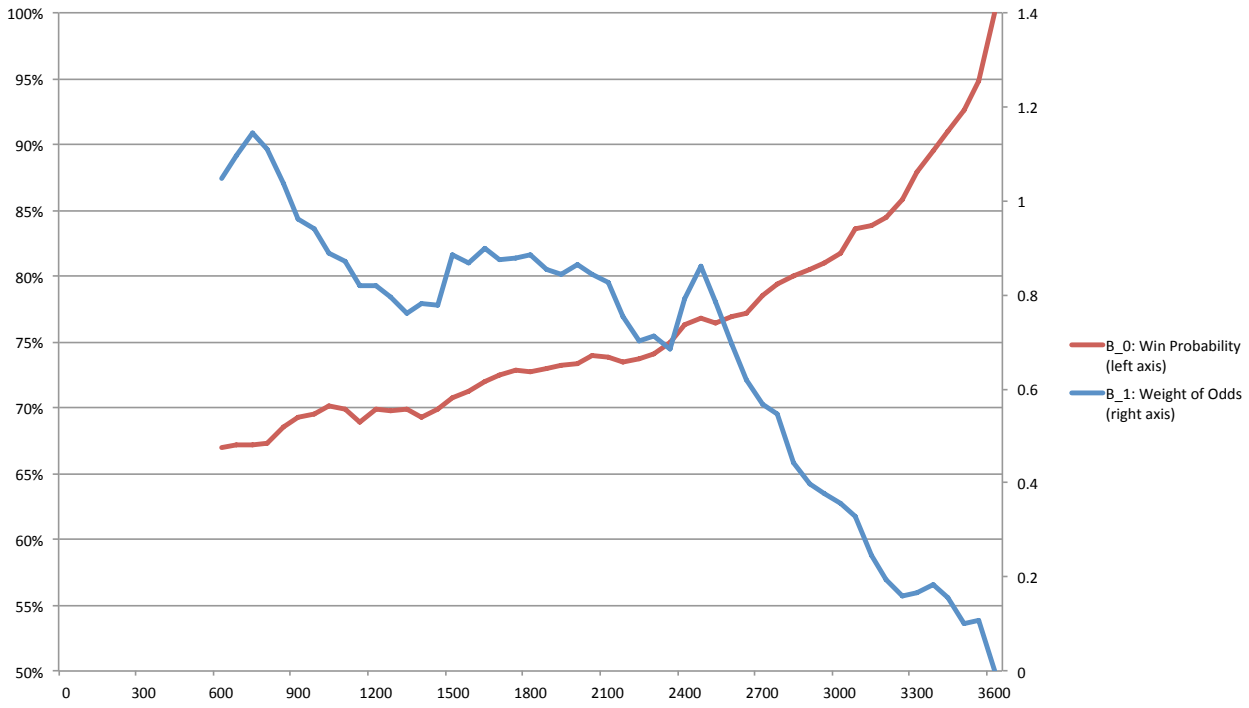


Figure 20: Regression coefficients where the visiting team is up by one goal

Now, let's look at the graphs with one goal differentials. Figure 19 plots the regression values if the home team is up by one goal, and Figure 20 plots the values if the visiting team is up by one goal. Not that for these, as well as all the other differentials, I start the analysis at the 600^{th} second to makes sure I have enough observations for solid results. The lines for the constant terms are quite similar. They follow the same concave up convergence to 100% observed in Figures 8 and 9. However, we can observe that the constant terms for the visiting team is slightly greater than the constant terms for the home teams, with the spread reaching close to 4%. The β_1 curves are slightly different from one graph to the other. For the visiting teams, the curve starts higher, but eventually stabilizes close to 0.8 just like the home team. While the β_1 curve for the home team is more concave down, the curve for the visiting team is closer to linear, especially in the third period.

Looking at all the different differentials (graphs available in the Appendix), there is almost always a discrepancy between the constant terms of the home teams and the away teams. Therefore, to rectify this, I decided to fix the constant terms of both graphs to be the values found in Figure 9. That way, if two teams are equal in strength, then they will have a win probability independent of whether they are the home team or away team. For a tied game, I restricted the constant term to be 50%.

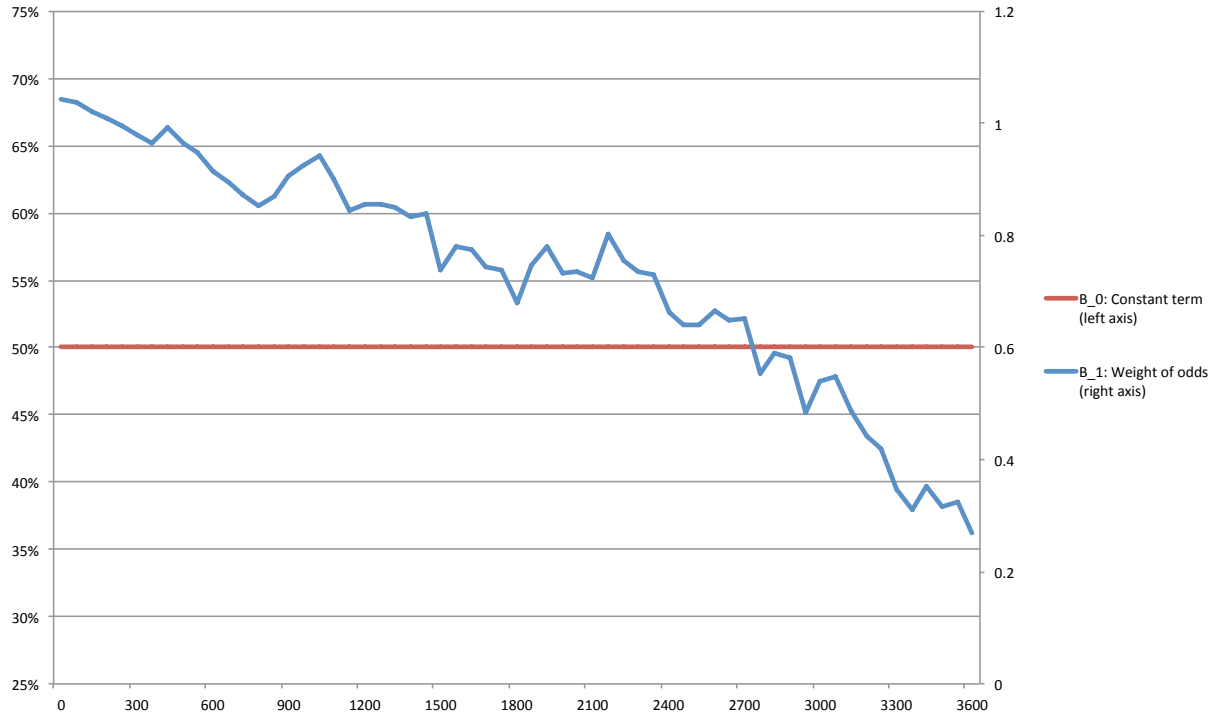


Figure 21: Restricted regression coefficients for a tied game

Figure 21 displays the regression coefficients with the constant term restricted to 0.5. We see that the β_1 follows roughly the same trend as in Figure 18. It is worth noting that the R^2 values for this restricted model will be lower than when not restricted. However, I am willing to make that sacrifice for the model to be more standardized and sensible.

Figure 22 displays the regression coefficients when teams are up by one goal with the constant terms restricted to the historical frequencies displayed in Figure 7. Again, both curves for the weight of the odds are quite similar to the ones calculated in the unrestricted regressions. However, it seems like now the spread between the two curves is smaller, which was the desired result. It is curious why the weight of the visitor odds is so high during the first period, especially compared to the weight of the odds of the home team. After the first period, this spread decreases greatly and the curves are quite close to one another.

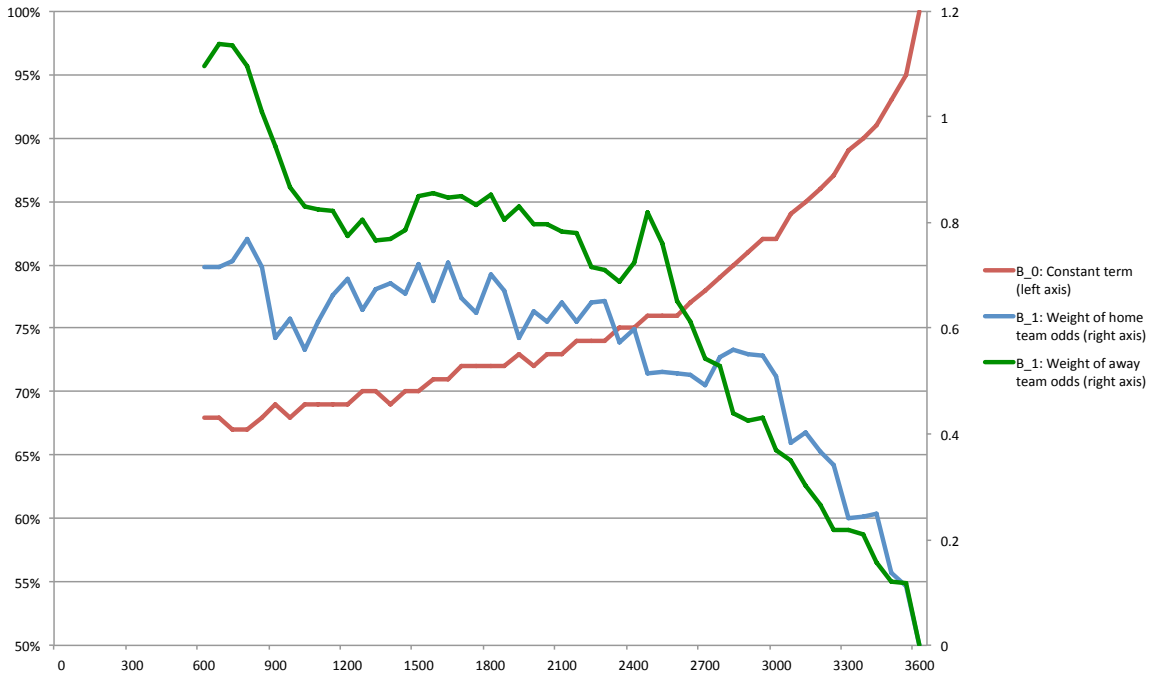


Figure 22: Restricted regression coefficients for a one goal differential

We observe similar results in Figure 23. The spread between the two lines is quite large in the first period, mostly due to the weight of the away team taking on very high values, reaching a weight of 1.2. In the second half of the game however, the spread is close to 0, and the lines follow similar trends as they converge to 0.

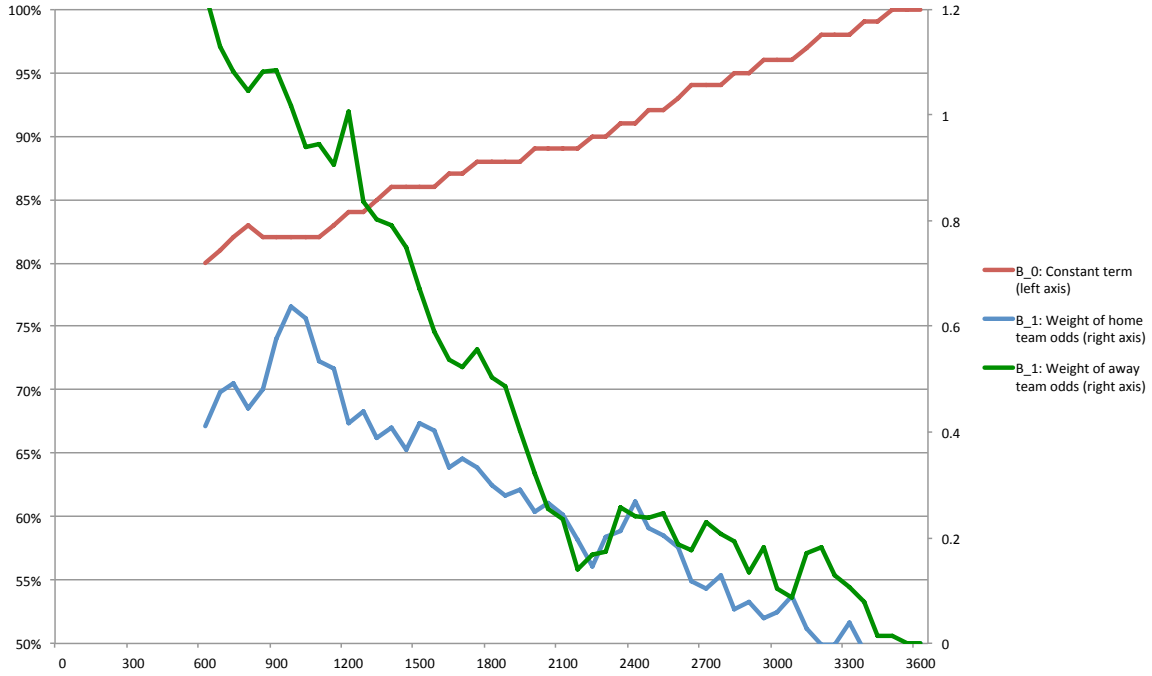


Figure 23: Restricted regression coefficients for a two-goal differential

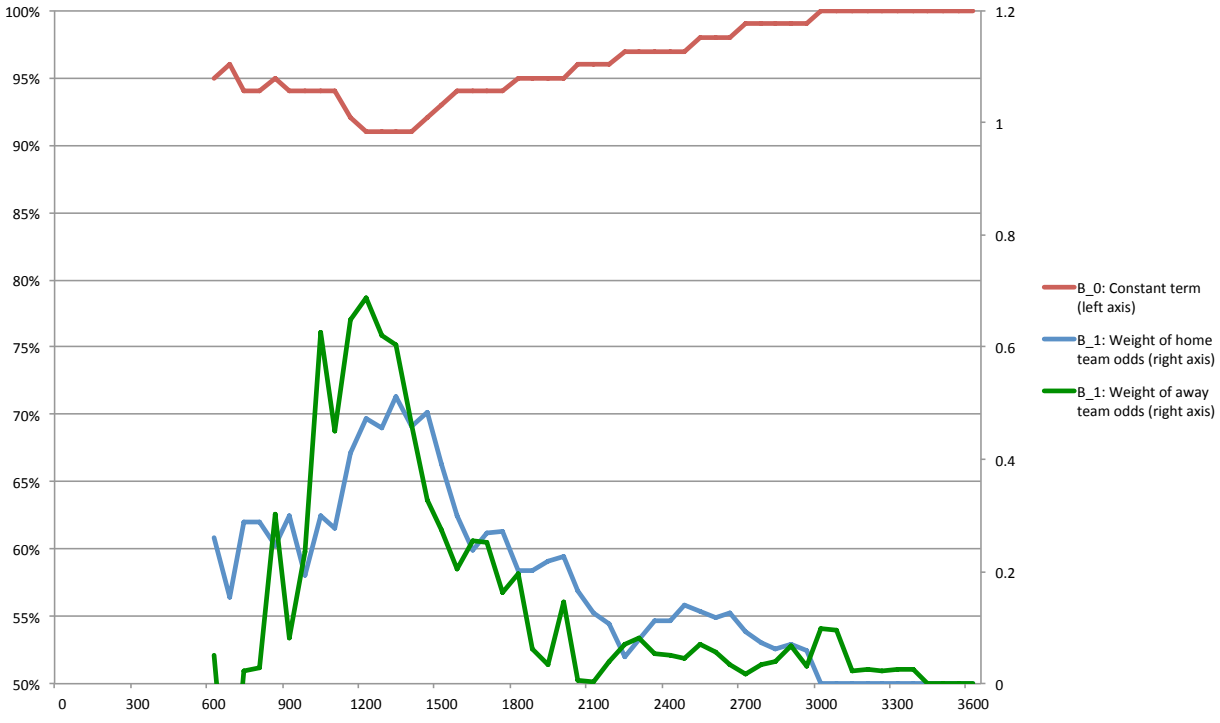


Figure 24: Restricted regression coefficients for a three-goal differential

The coefficients for the constrained regression for a three-goal differential are slightly counter-intuitive. First of all, there is a strange drop in the constant term between minutes 18 and 26. Coinciding with this drop, there is a random large spike in both the home team and the away team β_1 curves. To attempt to rectify this drop, I adjusted the constant term to eliminate the drop. Therefore, for minutes 18 to 26, I fixed the value of the constant to be 94%, coinciding with the constant values for minutes 17 and 27.

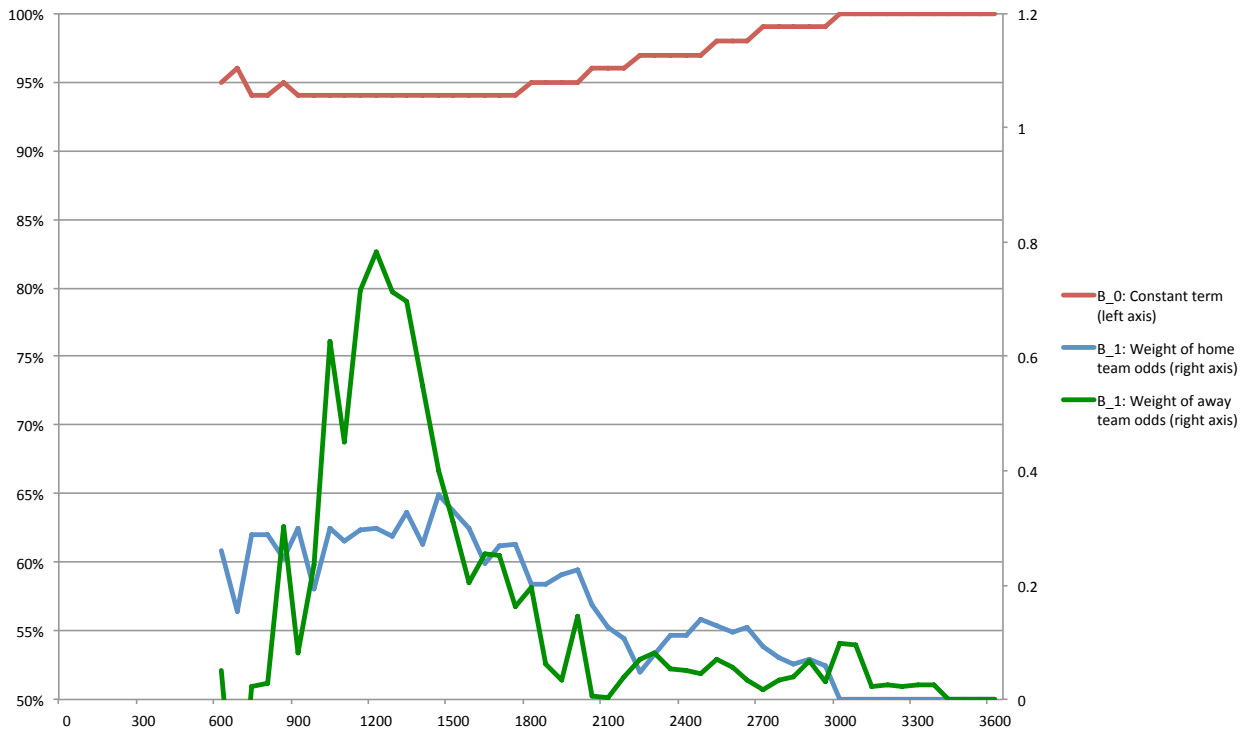


Figure 25: Restricted regression coefficients for a three-goal differential with constant terms adjusted

Figure 25 displays the results of the adjustment discussed above. The adjustment successfully reduced the coefficients for the home team. However, it increased the coefficients for the away team. This is somewhat counter-intuitive, and would only really make sense if on average, away teams that go up 3 goals in that time period are teams with odds less than 50%. I

checked this hypothesis by calculating the mean of the visiting team odds if they were up by 3 during that time period. I found that on average, visiting teams had betting odds of around 47.5% during that time, compared to average odds of around 45% for all visiting teams. On the other hand, some teams that were up by 3 in that time period had odds around 57.5%, compared to average odds of around 55% for all home teams. It is slightly surprising that visiting teams that reach a 3 goal lead on average have betting odds of less than 50%.

Given that this adjustment was not very successful in achieving a better graph, I instead decided to fix the constant term to 91% for values in the first period. Therefore, instead of inflating the constant, I deflated it. Figure 26 shows the resulting curves.

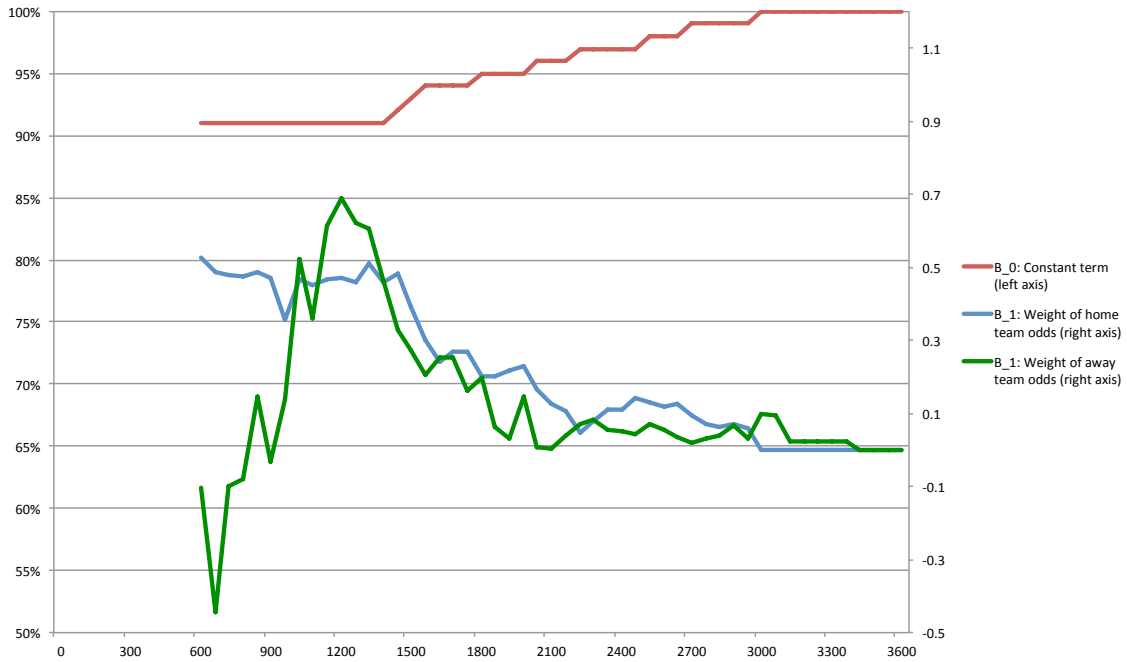


Figure 26: Restricted regression coefficients for a three-goal differential with constant terms adjusted

At first glance, this adjustment seems to stabilize the line for home teams but greatly distort it for visiting teams. However, the curve only dips severely in the first period, and it is

quite rare for teams to get a three-goal lead in the first period. I decided to stick with this adjustment and cap the weight of odds for the visiting team to fix this distortion. This will result in a deflation of the true forecast for a team that is up three goals in the first period, but this in my opinion is the best way to fix the unintuitive dip in the constant term.

For simplicity's sake, I decided to completely eliminate the distinction between home teams and away teams in calculating the weight of the odds. For the majority of game-states, the spread between the weight of the odds for the home teams and visiting teams is usually less than 0.1. Two thirds of teams have odds between 45% and 55%, so their odds values are between -0.05 and 0.05. Therefore, the difference between the forecast with averaging and without would be about 0.1*0.05, which is half of a percentage point. I am willing to sacrifice this small accuracy for a simpler, more usable model. Therefore, for differentials from 1 to 5, I created a new curve that was the linear average between the two curves. The resulting graphs are available in the Appendix. For the three-goal differential graph, I fixed the weight to be 0.5 for the first period to take into account the adjustment I discussed earlier.

7. Smoothing the Curves

Now that I have the curves for every single game-state, I needed to smooth the curves to reduce the jitteriness that resulted from the inherent variation in the data. I decided to use an OLS cubic fit model to smooth the curves. In other words, the regressions would have the desired data to be smoothed out as the dependent variables and would have as independent variables the seconds, the square of the seconds, and the cube of the seconds. I also constrained the regressions so that as time goes to 3600, the constant terms would converge to 1 and the weight of the odds would converge to 0.

$$cons + 3600\ sec + 3600^2 sec2 + 3600^3 sec3 = 1$$

Figure 27: Constraint inputted in STATA for the cubic regressions of the constant terms, where *sec2* and *sec3* are the coefficients for the seconds squared and cubed, respectively

$$cons + 3600\ sec + 3600^2 sec2 + 3600^3 sec3 = 0$$

Figure 28: Constraint inputted in STATA for the cubic regressions of the weight of the odds

Differential	Constant	Sec	Sec2	Sec3
0	0.5	0	0	0
1	0.5463485	2.61E-04	-1.47E-07	3.05E-11
2	0.9313658	-7.29E-05	6.22E-08	-1.02E-11
3	0.9897124	-4.76E-06	6.29E-09	-1.16E-12
4	0.9897124	-4.76E-06	6.29E-09	-1.16E-12
5+	1	0	0	0

Figure 29: Coefficients of the cubic regression on the constant term for each differential

Differential	Constant	Sec	Sec2	Sec3
0	1.11308	-5.33E-04	0.000000337	-7.63E-11
1	1.245982	-7.93E-04	4.30E-07	-8.49E-11
2	1.134884	-4.06E-04	-2.15E-08	1.30E-11
3	0.6834897	-1.24E-04	-1.16E-07	2.72E-11
4	1.007633	-8.57E-04	2.26E-07	-1.84E-11
5+	0	0	0	0

Figure 30: Coefficients of the cubic regression on the weight of the odds for each differential

Figures 29 and 30 display the coefficients of the regressions for each of the cubic fit regressions. I used these coefficients to calculate the constant terms and the weight of the odds for each second in the game. For the weight of the odds, I used the coefficients to calculate the weight in the additional 300 seconds of an overtime period in the event of a tie. I also capped all the values to be between 0 and 1. Therefore, if a predicted value were above 1, it would be adjusted to 1. Similarly, if it were below 0, it would be adjusted to 0. Finally, in the regression analysis, only values after the 600th second were used to calculate the curves for one to three goal differentials, and after the 1200th second for four and five goal differentials. Before that mark, I have the graphs take on the value at second 600 or 1200. For the constant terms, I did the same adjustment for values before the 600th second.

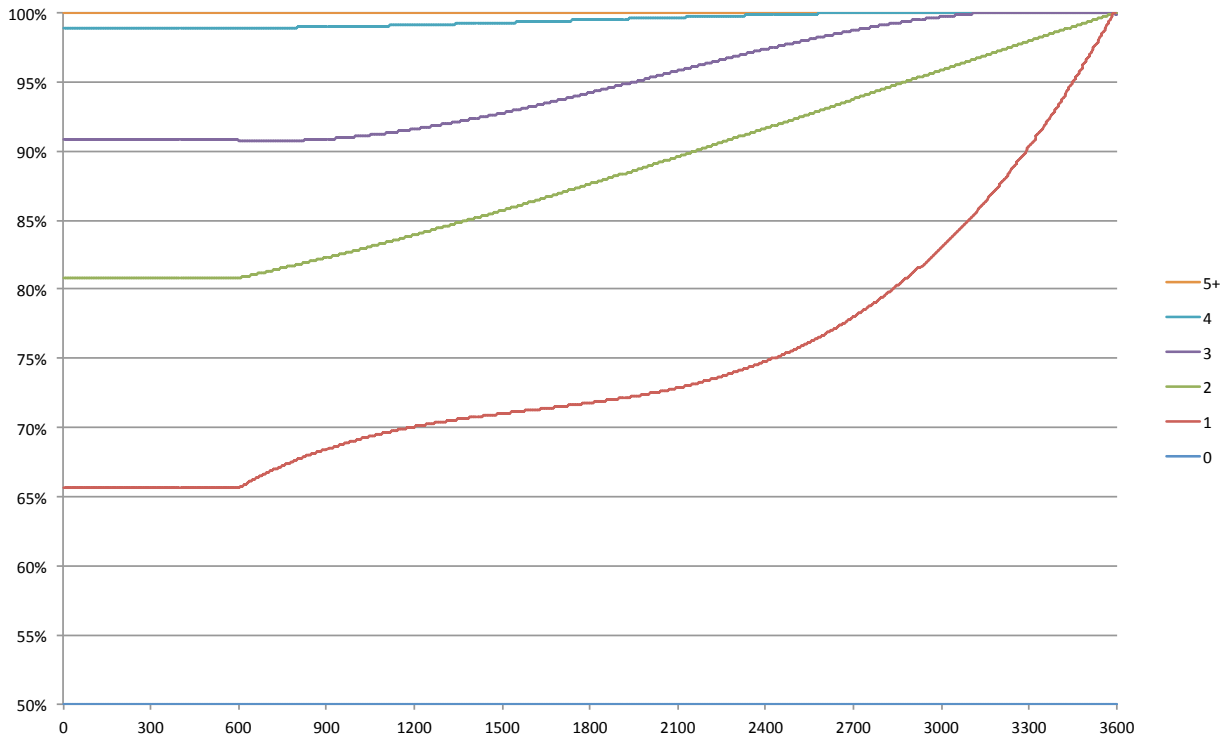


Figure 31: Smoothed curves of the constant terms for each differential

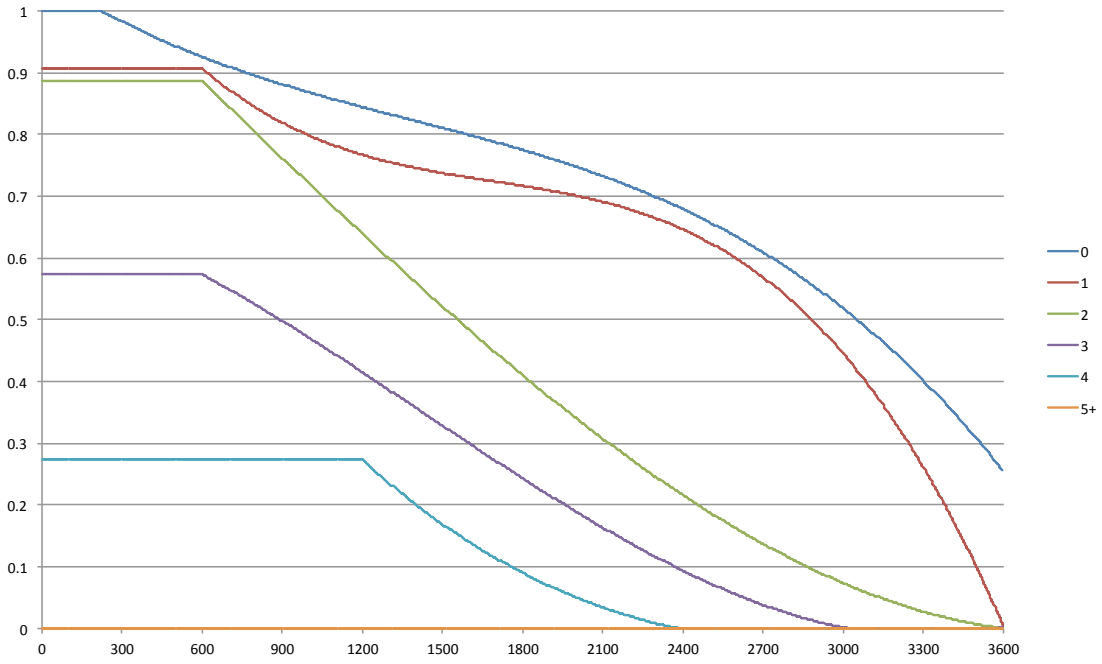


Figure 32: Smoothed curves of the weight of the odds for each differential

Figures 31 and 32 display the final, smoothed curves for the model. Depending on time and differential, the model retrieves a value on both graphs above and uses the odds data to calculate the final forecast. There might be a necessary adjustment to make to that value (i.e. if it is above 100% or below 0%), but these two graphs contain the information necessary to calculate the forecast.

8. Incorporating Penalties

The next step in the construction of the model is to account for power plays. In hockey, if a team is given a penalty, they play one player short for usually 2 minutes. If the other team scores while on power play, the power play ends and the game returns to full strength. If the short-handed team scores, the power play continues.

If a team is on the power play, their chances of scoring increase, so their win probability increases. I decided to replicate Pettigrew's method of using conditional probabilities to calculate the necessary adjustment to the model. To do so, I gathered all the penalties committed in the past 8 seasons as well as all goals scored on a power play by either team in that same time period. My goal was to evaluate each team's probability of scoring a goal before the end of the power play, depending on how much time is left in it. To facilitate the analysis, I eliminated overlapping penalties from my scope as well as penalties that were not 2 minutes long. Therefore, I was left with only single minor penalties that had no other penalties called during their duration. This was a total of 46,194 penalties, or on average just under five penalties per game

From that dataset that included the penalties and the goals scored, I collapsed the data so that it was only expressed in terms of individual power plays. Each line of the data was an individual power play that included the start time, if and when a goal was scored, and which team scored it. In the eventuality that multiple goals were scored on one power play, there were multiple entries for that power play including the different goals scored.

From there, I generated 120 variables denoting each second of the power play. These variables took on a 0 if a goal was not scored in the time left in the power play, a 1 if a power play goal was scored, and nothing if the power play ended (if the power play team already scored a goal before that point). I repeated the same process for shorthanded goals. I was therefore able to produce a second by second forecast of whether or not a power play goal would be scored in the time remaining on the power play, and whether a shorthanded goal would be scored.

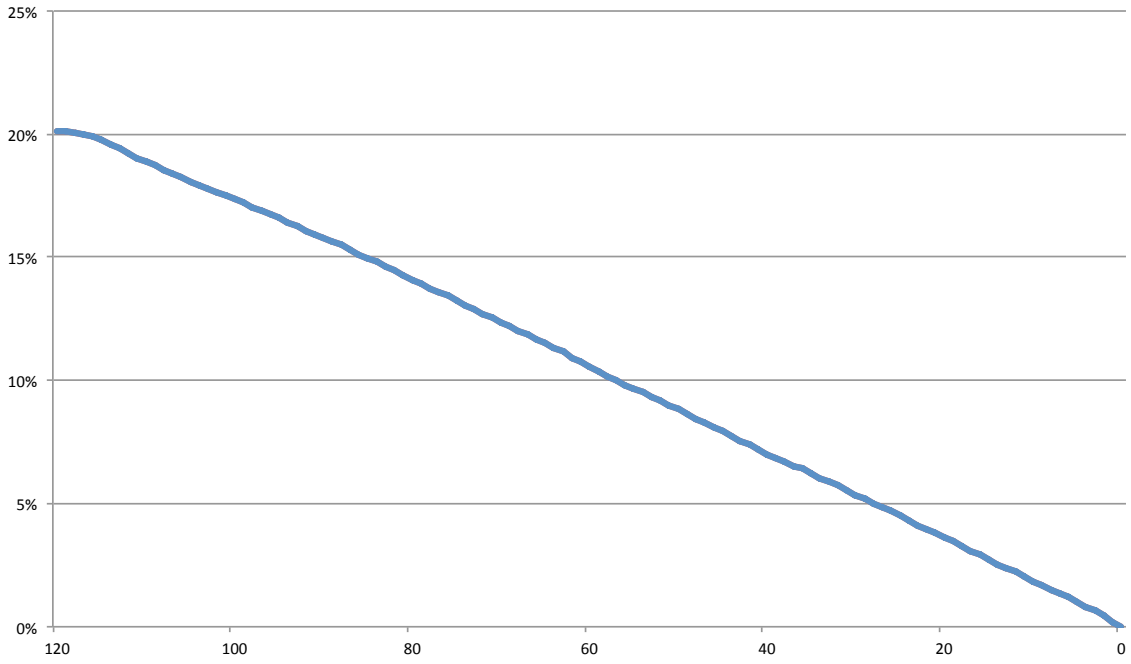


Figure 33: Probability of a power play goal based on how much time has elapsed in the penalty

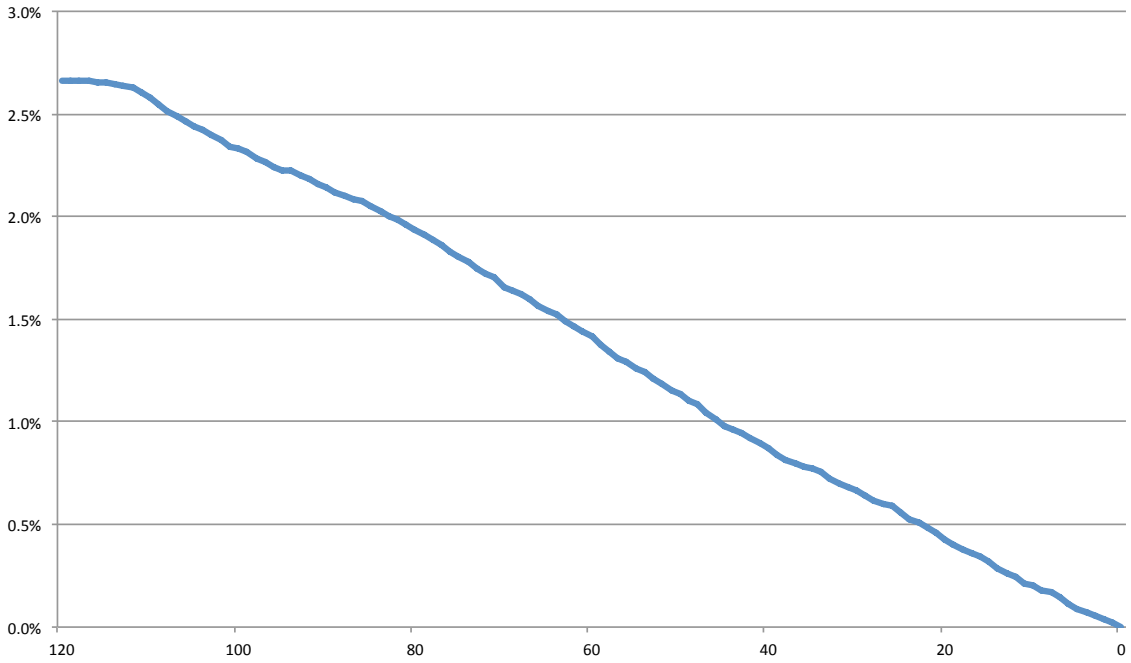


Figure 34: Probability of a shorthanded goal based on how much time has elapsed in the penalty

Figures 33 and 34 display the probability of a goal scored on the power play based on how much time there is left on the power play. For both power play goals and short-handed goals, the trend is very linear apart from the first 5 seconds when goals are rarely scored. We also notice here how unlikely shorthanded goals are. In his model, Pettigrew calculates the probability of the home team winning conditional on a shorthanded goal times the probability of a shorthanded goal. However, that probability is quite small so that it will have only a miniscule effect on the model. For simplicity's sake, I decided to instead use the expected goals scored for my formula. I therefore subtracted the probability of a shorthanded goal from the probability of a power play goal. I also ran a linear fit through that line, giving the result below.

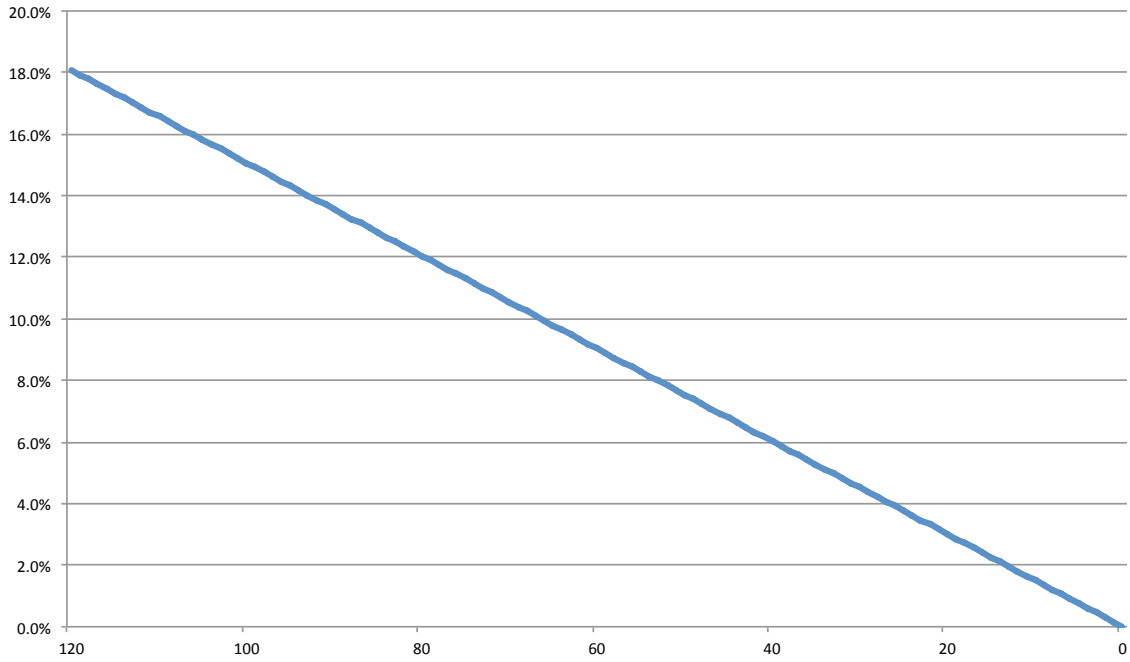


Figure 35: Linear fit of the difference between the probability of a power play goal and a shorthanded goal, constrained to equal 0 at time 0

$$goal = 0.0015066\ sec$$

Figure 36: Equation of the constrained linear fit above, where sec is the time left on the power play

$$P(lead\ team\ win) = P(lead\ team\ win \mid PPG)(0.0015\ sec) + P(lead\ team\ win \mid No\ PPG)(1 - 0.0015\ sec)$$

Figure 37: Win-probability equation conditional on the number of seconds left in a power play

9. Developing a Model for Shootouts

For regular season games, if the game is tied after sixty minutes of play, teams play an additional five minutes of overtime. If the game is still tied after that, a shootout is necessary. In a shootout, each team alternates to send three shooters alone against the goalie. After these three rounds, the team with the most goals wins the game. If there is no winner, an additional round is added until a winner is reached.

In the eventuality of a shootout, one cannot use the model developed above as it is almost completely independent from the previous 65 minutes of play. Therefore, I evaluated shootout data from the past eight seasons to develop a win probability model that solely evaluates shootouts. Just like Pettigrew did in his model, I treated the first three rounds independently, and then collapsed all additional rounds to develop one single estimate for them. For the first three rounds, I calculated historical win probability after each shooter based on the goal differential.

Shooter number	Differential after the shooter from the perspective of the first shooting team				
	-2	**-1**	**0**	**1**	**2**
1			38.0%	72.0%	
2		15.7%	49.3%	84.5%	
3		5.0%	36.9%	78.6%	96.8%
4	0.0%	8.8%	52.7%	91.7%	100.0%
5		0.0%	33.5%	88.6%	100.0%
6		0.0%	50.0%	100.0%	

Figure 38: First shooting team's win probability based on how many shots have been taken and the goal differential

Figure 38 displays the historical win frequencies based on the shooting round and the differential. One way to interpret the table is as follows. After the first shooter, the win probability of his team will be either 38.0% or 72.0%. If the next shooter does not score, simply go down one cell and that is the new win probability. If he does, go the cell one down and one left. Repeat this process for the third shooter, except if he scores go to the cell on down and to the right. Repeat this alternating method until the last shooter goes or a cell containing 0% or 100% is reached.

For example on April 9th 2017, the Carolina Hurricanes and the Philadelphia Flyers played a shootout period to determine the winner. At the beginning of the shootout, both teams had 50% win expectancy. Philadelphia went first and missed, so their win expectancy went to 38%. Carolina scored on their first attempt, so Philadelphia's win expectancy decreased to 15.7%. Philadelphia scored on their second attempt, increasing their win expectancy to 36.9%. Carolina and Philadelphia both missed their next attempts, so Philadelphia's win expectancy went to 52.7% and then down to 33.5%. Finally, Carolina scored on their final attempt, so Philadelphia's win expectancy went to 0%; the game was done, and Carolina won.

If the game is still tied after three rounds, or six shot attempts, for each following round the win probability will depend on whether or not the first shooter scores. Indeed, after the second shooter, the game will either be over or the win expectancy will reset back to 50%. It is therefore only dependent on the outcome of the first shot of the additional round. I therefore calculated the first shooting team's historical win frequency based conditional on what happened in the first shot. If the shooter scores, the win probability of his team is 82.3%; if not, the win probability is 33.3%.

10. Comparison with Pettigrew's Model

With the model complete, I can now compare it to Pettigrew's model for a given game. From his website, I captured screenshots of some of the games he used as examples. The first is between Chicago and Washington, played on October 1st 2013.

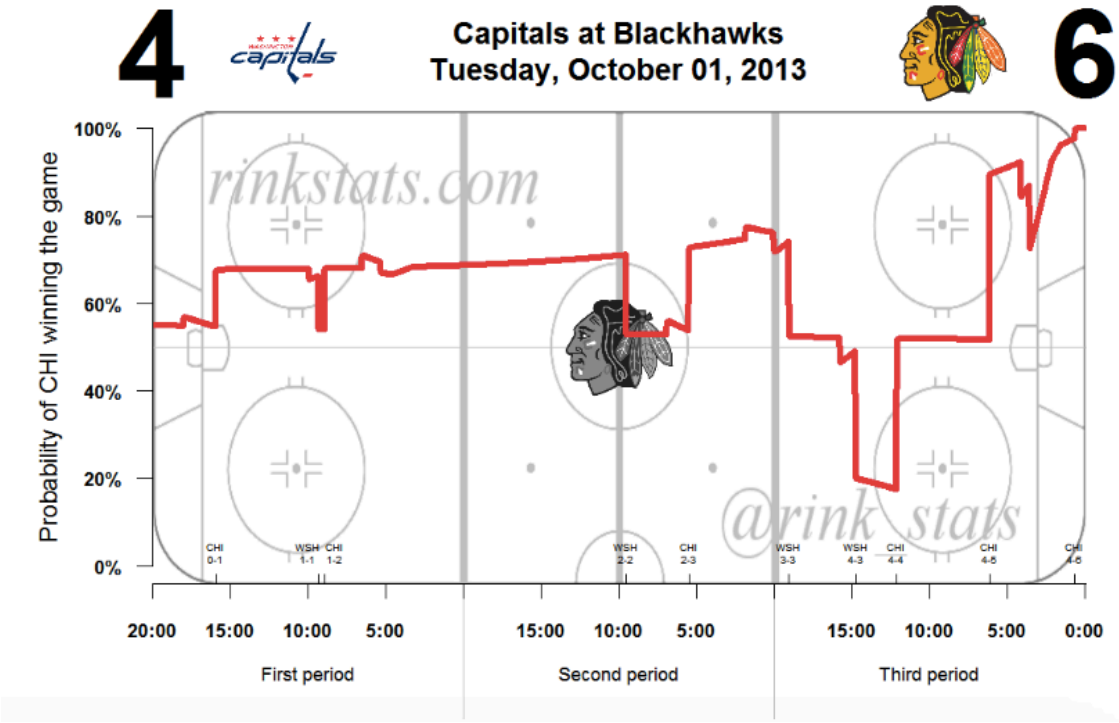


Figure 39: Pettigrew's model applied to a game, taken from Rink Stats (Pettigrew, 2014)

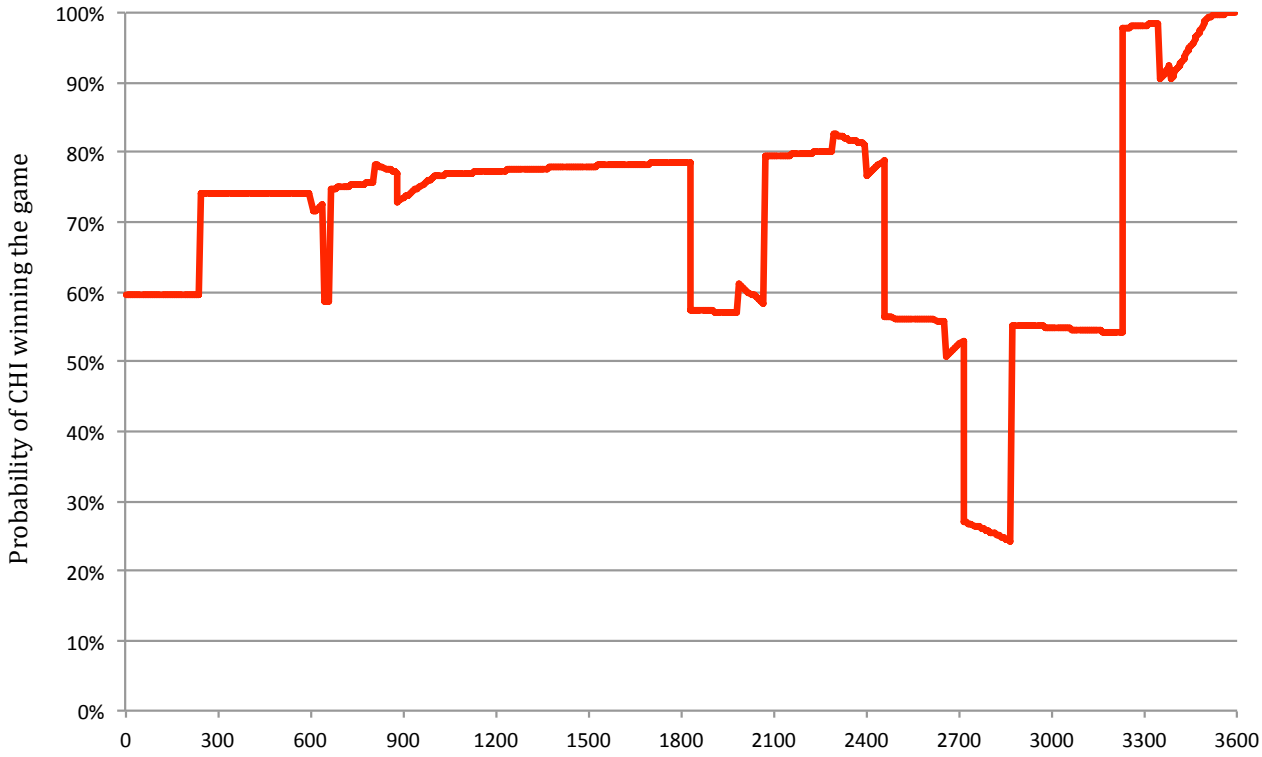


Figure 40: My model applied to the same game as above

Generally the two curves are very similar. It is very easy to identify when goals were scored from the large swings. It is also easy to identify when the power plays occurred from the smaller jumps in the curve. One slight difference occurs towards the end of the game, when Pettigrew's model dips slightly more than mine. This is because a 5-on-3 power play occurred, and my model simply treats it as the start of a new penalty whereas his model recognizes the additional advantage. We notice also that the though the curves follow a similar trend, they aren't quite at the same level. This is because Chicago had pre-game Vegas odds of 59.5%, and my model takes this advantage into account. My model has a different starting point and reaches values that are slightly above Pettigrew's. To demonstrate this difference, in Figure 41 I compared the win probability graph above with the graph if the two teams had been evenly matched.

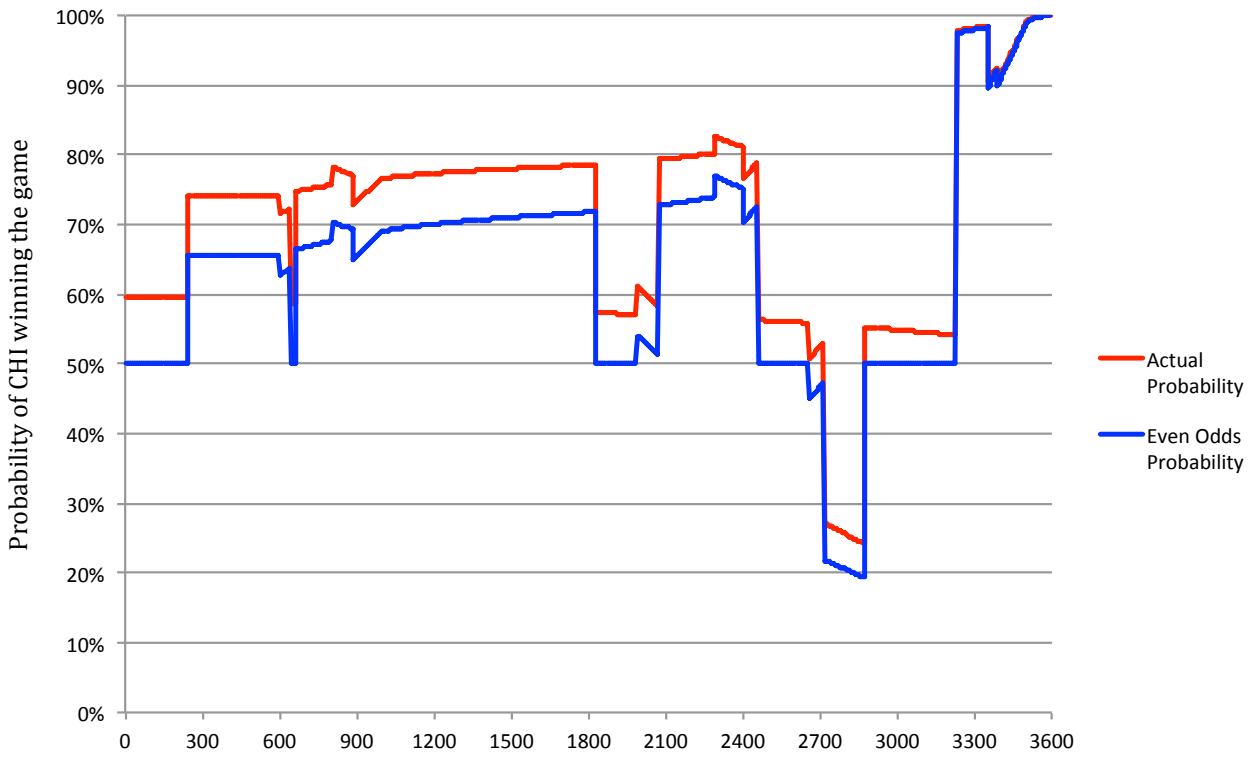


Figure 41: Effect of pre-game odds on the win-probability model

As we can see, the spread between the two curves is non-negligible. Especially in the beginning of the game, when the odds carry a lot of explanatory weight, the spread reaches close to 10%. As the game goes on, the spread becomes smaller as the score differential carries increasingly more weight in the equation.

We repeat the process for another game, this time one that was decided in a shootout period. We will therefore be able to compare my shootout model to Pettigrew's.

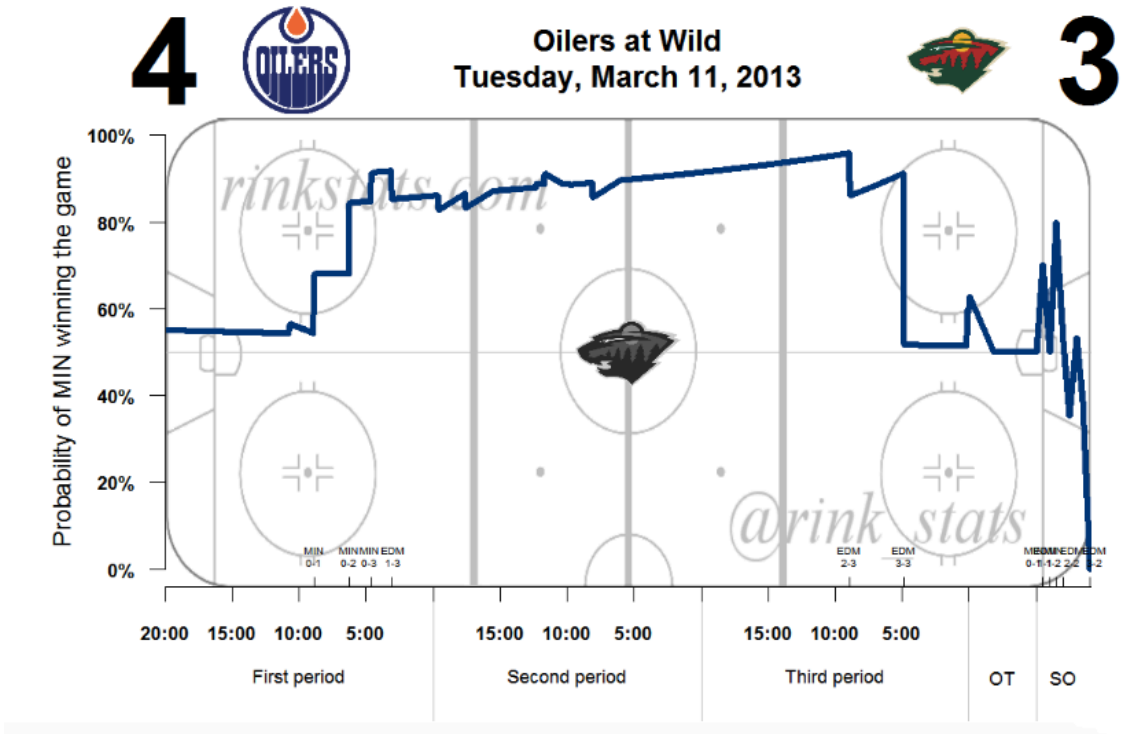


Figure 42: Pettigrew's model applied to a game, taken from Rink Stats (Pettigrew, 2014)

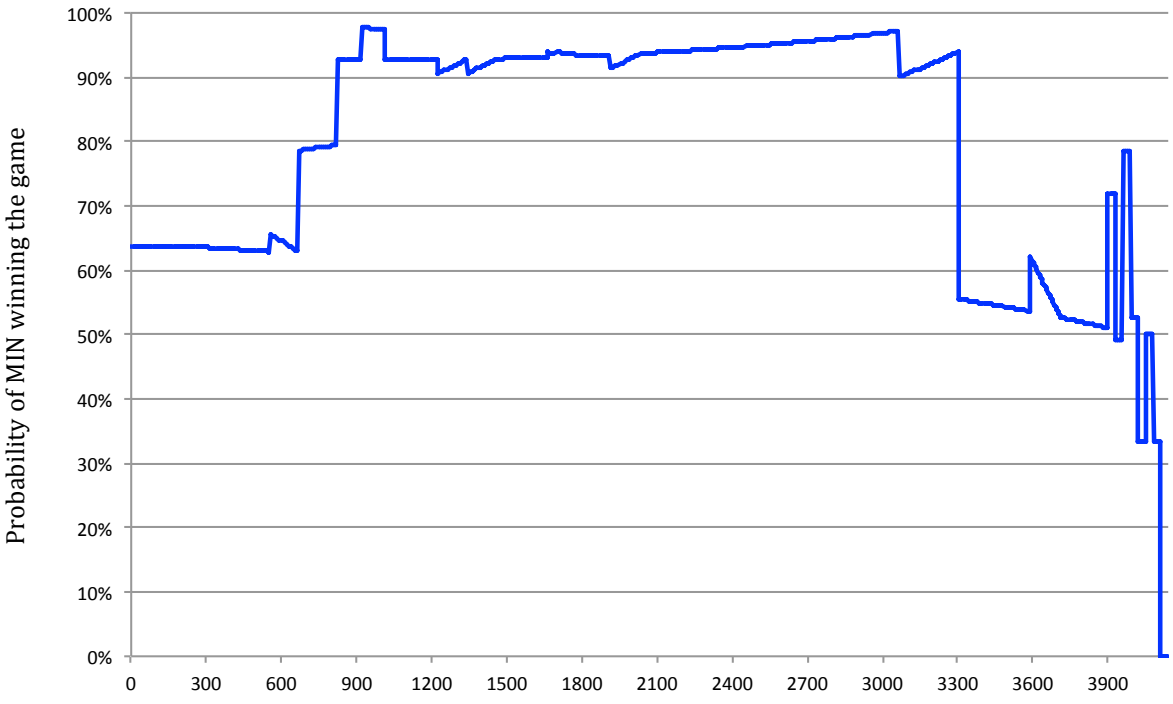


Figure 43: My model applied to the same game as above

As with the previous game, the curves follow very similar trends. The jumps for goals and the smaller jumps for power plays occur at the same time. Again, one major difference between the two curves is the inclusion of pre-game betting odds. Minnesota started the game with a win probability of 63.8%, a clear advantage. We can see that advantage by noticing the spread between Pettigrew's estimate and my estimate. Finally, looking at the shootout period, we see that the two curves are virtually identical. Though it is tough to see the exact values in Pettigrew's model, from simply looking at them we can see how similar they are. I therefore conclude that Pettigrew arrived to very similar shootout probabilities than I did.

11. Conclusion

This thesis focused on building a NHL real-time win probability model from the ground up. I used a game-state method inspired from Pettigrew's work to build the basics of the model, and arrived to very similar results than he did. I was able to develop a novel method of incorporating Vegas odds to the game-state approach by using a multi-regression based method. In the process of building the model, I also pieced together several extensive NHL datasets, including a detailed play-by-play dataset spanning eight seasons and extensive historical odds data.

To improve the model, I would start by increasing the breadth of power plays evaluated to be able to include 5-on-3 penalties. I would also look at shots and time of possession to estimate future goals scored; I imagine this information would be especially useful for tied games. I would also look at the path to the game-state to see identify if a momentum effect exists. Finally, I would develop the model so that it can process information real-time and continuously update itself as the game goes on.

Bibliography

Everson, Phil, and Jimmy Charite. *Mid-Game Precitions for NBA Basketball,* New-England Symposium on Statistics and Sports. 24 Sept, 2013.

"Past Odds Results." https://www.covers.com/pageLoader/pageLoader.aspx?page=/data/nhl/teams/teams.html.

Pettigrew, Stephen. "Win Probabilities Metric, 1.0." RSS, 24 Mar. 2014, rinkstats.com/2014/03/win-probabilities-metric-10.

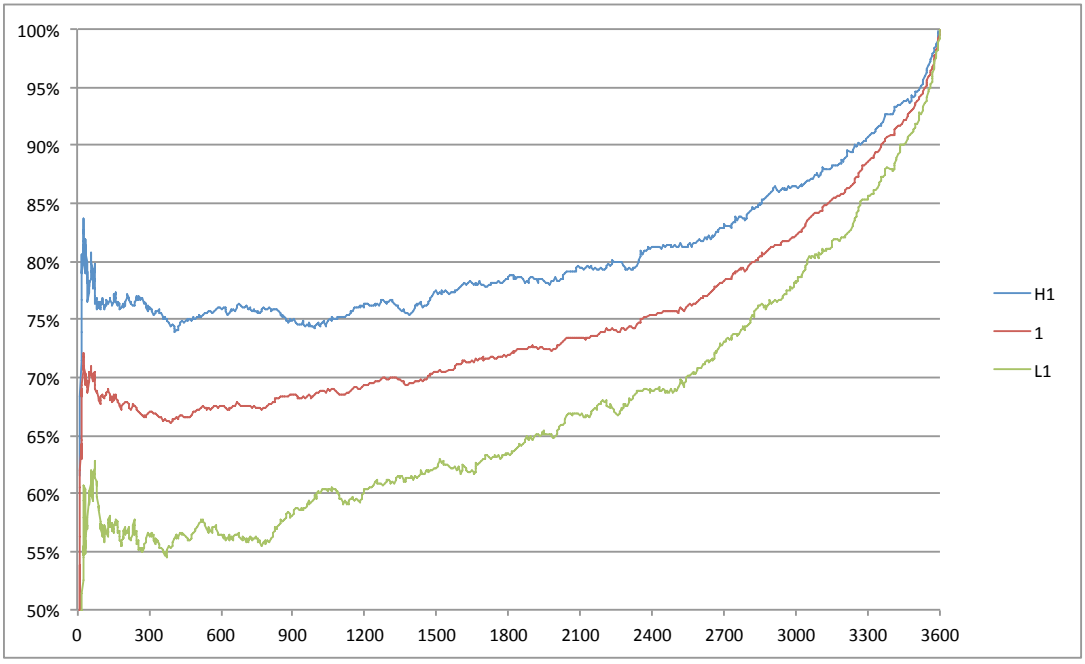
Tango, Tom M., et al. *The Book*: *Playing the Percentages in Baseball*. Potomac Books, 2007.

"Win Probability Model." Pro-Football-Reference.com, www.pro-football-reference.com/about/win_prob.htm.

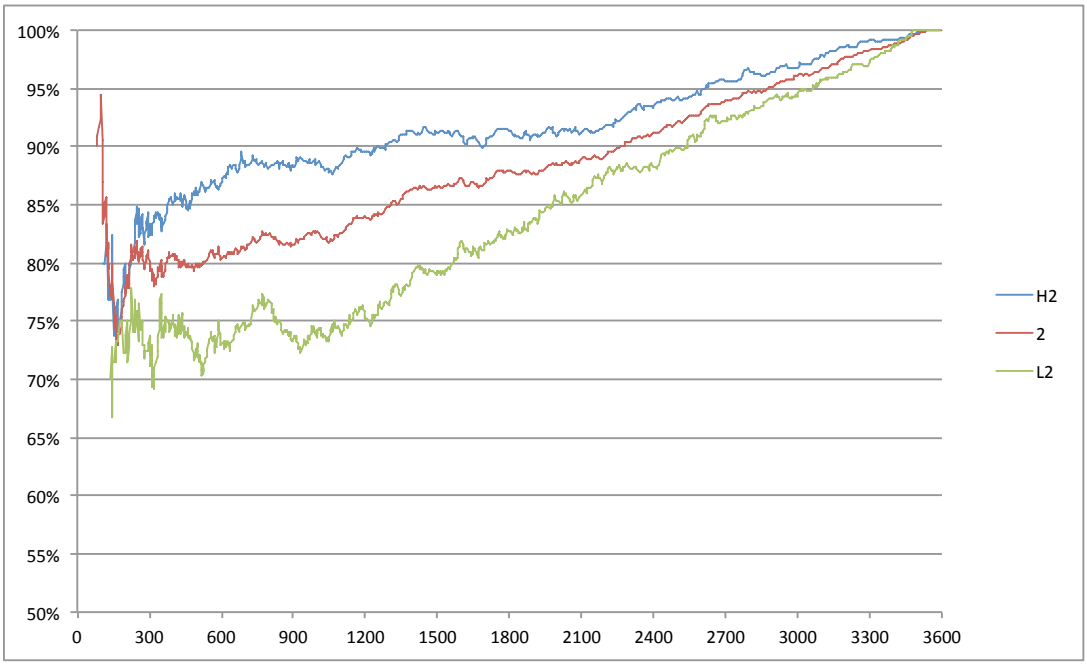
Winston, Wayne L. *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*. Princeton Univ. Press, 2012.

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
Colorado	45.1%	47.5%	45.8%	46.4%	44.3%	51.2%	46.3%	45.2%	39.2%
Detroit	63.2%	55.3%	56.0%	57.1%	50.9%	50.2%	53.9%	51.0%	44.9%
Boston	57.7%	52.9%	55.9%	58.7%	59.4%	58.2%	55.3%	52.9%	55.1%
NY Islanders	36.8%	42.2%	40.5%	44.4%	48.2%	44.3%	54.1%	53.0%	47.5%
Carolina	50.7%	44.9%	49.1%	45.6%	47.5%	45.9%	43.7%	46.3%	48.4%
NY Rangers	52.8%	50.2%	50.6%	54.8%	54.9%	53.5%	54.2%	54.0%	53.3%
Pittsburgh	54.1%	57.2%	53.7%	56.6%	58.1%	57.6%	57.8%	53.9%	57.9%
Toronto	42.6%	45.3%	46.3%	47.2%	47.8%	48.0%	44.4%	41.5%	50.7%
Ottawa	47.8%	50.0%	44.1%	47.3%	48.5%	49.3%	48.1%	46.6%	46.9%
Buffalo	51.8%	52.7%	50.9%	50.0%	45.6%	37.7%	31.6%	41.9%	43.4%
Montreal	54.9%	46.5%	49.8%	47.1%	53.1%	52.4%	53.9%	49.3%	53.7%
Philadelphia	53.5%	54.4%	56.7%	54.1%	48.4%	49.5%	46.9%	47.6%	48.5%
New Jersey	54.1%	54.6%	49.5%	49.3%	49.9%	48.2%	45.1%	44.0%	43.2%
Florida	47.0%	44.0%	44.6%	47.6%	42.0%	41.8%	46.0%	50.8%	49.4%
Washington	56.2%	59.9%	57.2%	52.6%	49.0%	49.4%	53.3%	57.6%	59.3%
St. Louis	43.9%	48.5%	49.7%	54.3%	55.4%	58.9%	56.6%	53.7%	53.1%
Tampa Bay	41.4%	44.5%	50.9%	46.8%	49.8%	51.1%	56.4%	56.0%	51.6%
Calgary	54.1%	52.0%	49.5%	47.4%	43.9%	41.3%	47.0%	45.5%	49.2%
Arizona	44.4%	48.1%	49.2%	48.9%	49.0%	51.0%	40.2%	42.1%	38.1%
Dallas	48.6%	48.5%	49.4%	47.5%	45.5%	48.6%	49.3%	54.9%	48.5%
San Jose	62.5%	58.7%	55.9%	56.6%	53.4%	59.1%	53.7%	54.7%	56.9%
Chicago	55.2%	59.9%	56.5%	54.6%	57.0%	58.9%	59.7%	56.1%	54.0%
Los Angeles	43.4%	50.6%	54.0%	53.0%	55.3%	55.7%	57.2%	57.7%	53.8%
Edmonton	47.7%	40.4%	39.7%	43.7%	45.7%	42.2%	37.9%	43.9%	52.3%
Anaheim	52.4%	48.1%	47.3%	48.6%	52.3%	54.9%	55.2%	55.1%	52.5%
Vancouver	54.0%	54.9%	58.6%	58.4%	57.0%	51.5%	52.3%	45.0%	40.3%
Nashville	46.7%	48.2%	50.1%	51.3%	47.4%	45.9%	54.1%	55.1%	53.2%
Winnipeg	41.0%	46.2%	46.5%	46.0%	47.9%	45.3%	48.9%	47.1%	47.6%
Minnesota	50.1%	47.1%	45.8%	44.0%	50.2%	50.3%	53.0%	53.0%	55.4%
Columbus	46.6%	46.6%	46.6%	39.8%	42.7%	48.2%	44.0%	44.8%	52.2%

Appendix Figure 1: Historical odds for each team for each season. Note that some teams have changed names or cities; their most recent city is the one displayed

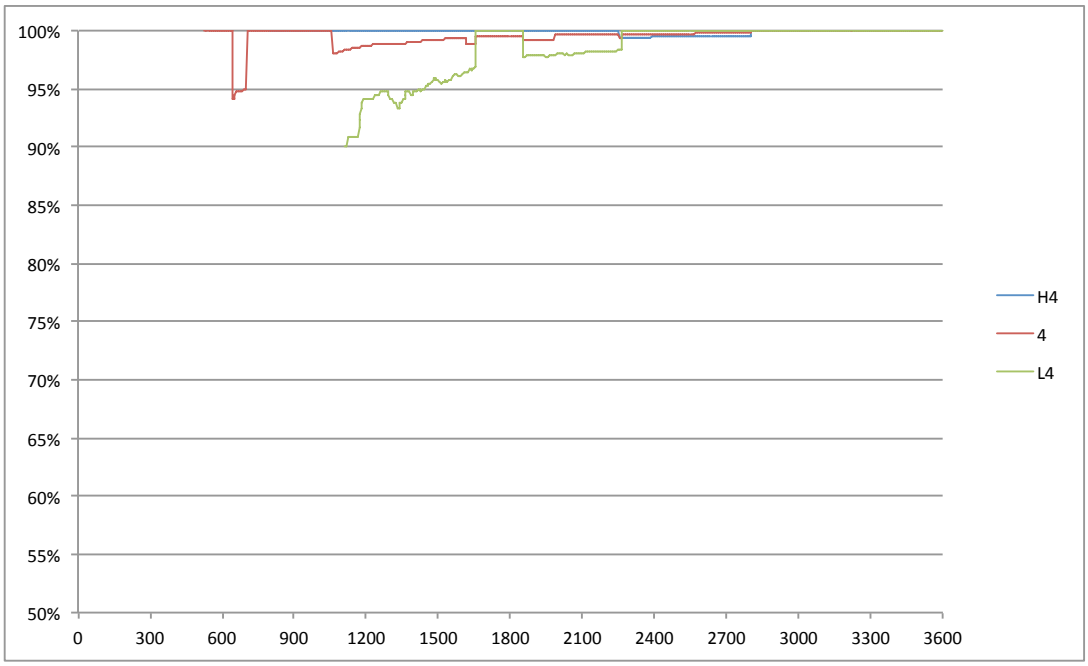


Appendix Figure 2: Win probabilities of team leading by one goal. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%

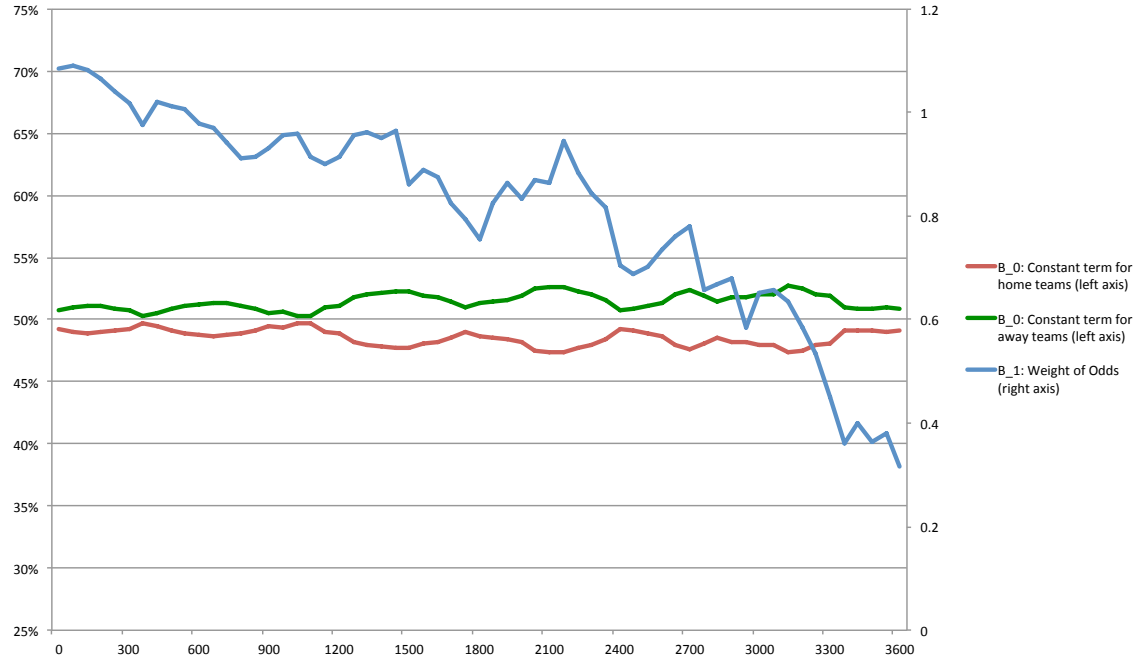


Appendix Figure 3: Win probabilities of team leading by two goals. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%

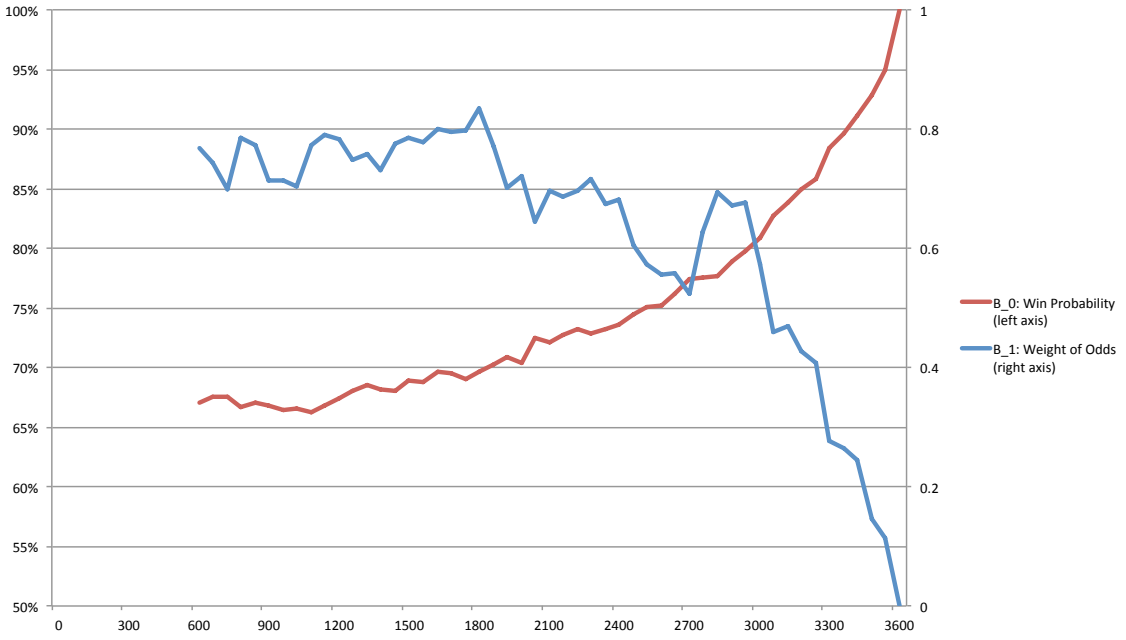
Appendix Figure 4: Win probabilities of team leading by three goals. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%



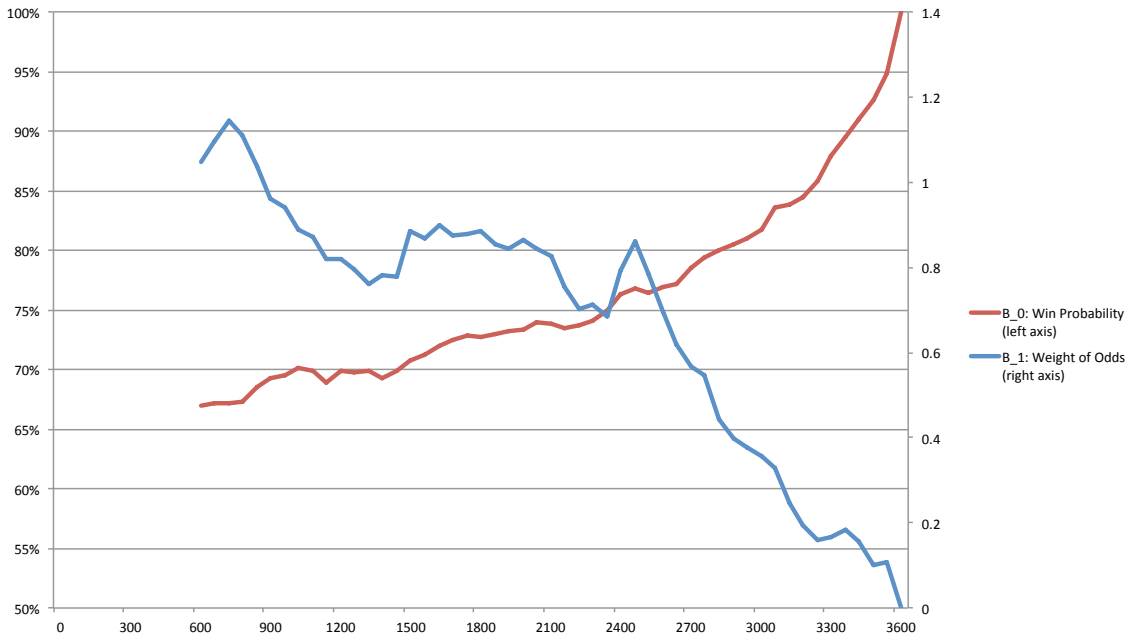
Appendix Figure 5: Win probabilities of team leading by four goals. "H1" refers to teams with win probabilities above 55%, and "L1" refers to teams with win probabilities below 45%



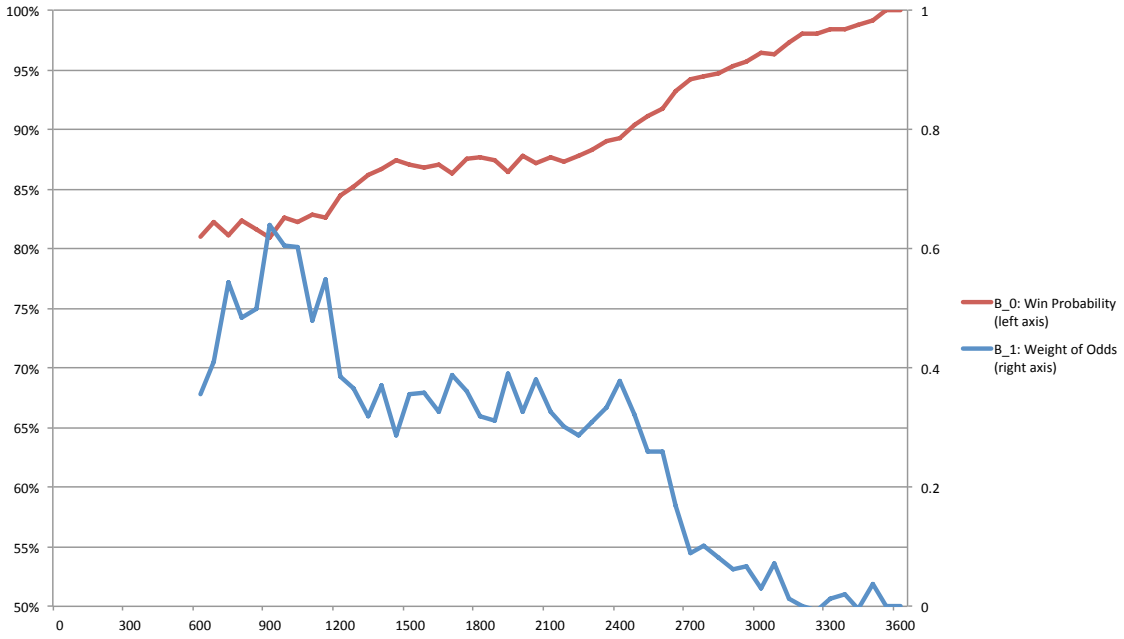
Appendix Figure 7: Regression coefficients where the home team is up by one goal



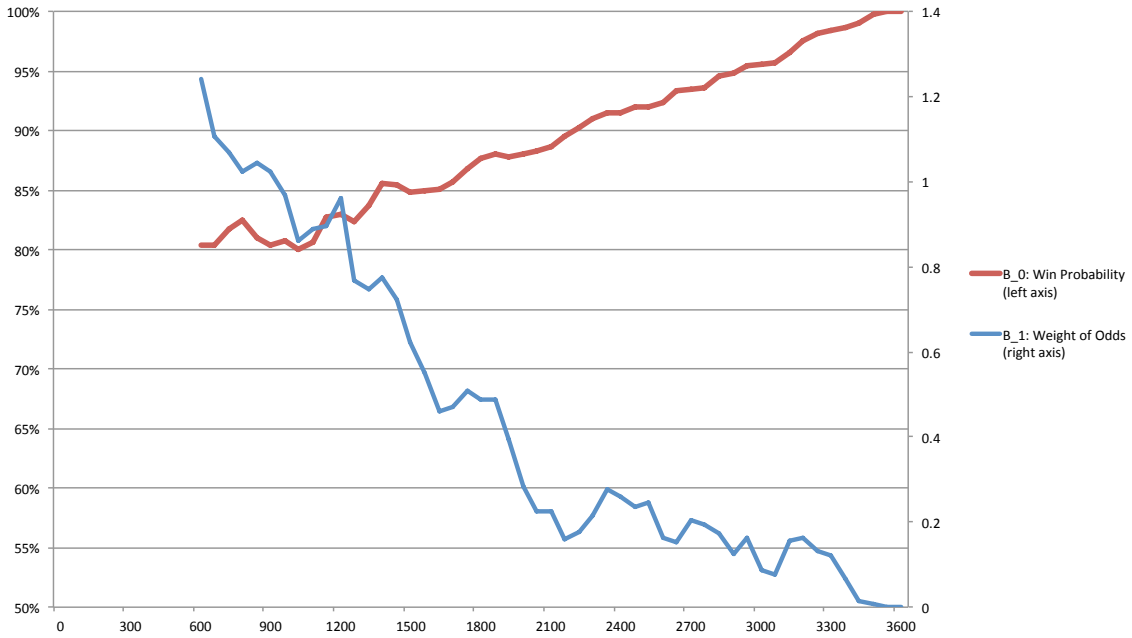
Appendix Figure 6: Regression coefficients where the game is tied



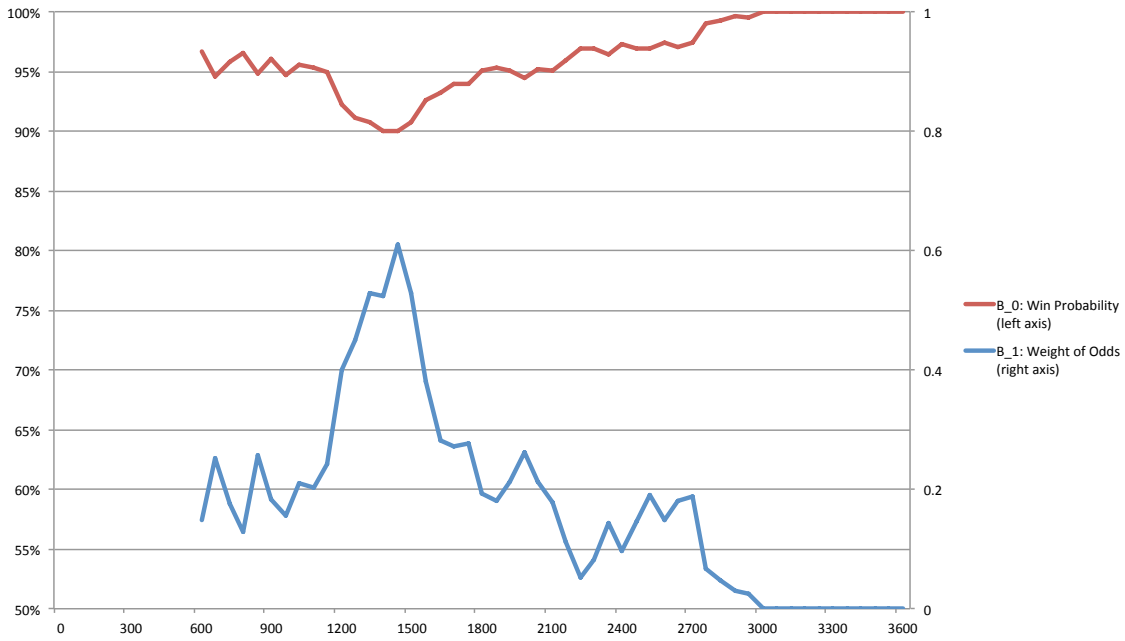
Appendix Figure 8: Regression coefficients where the away team is up by one goal



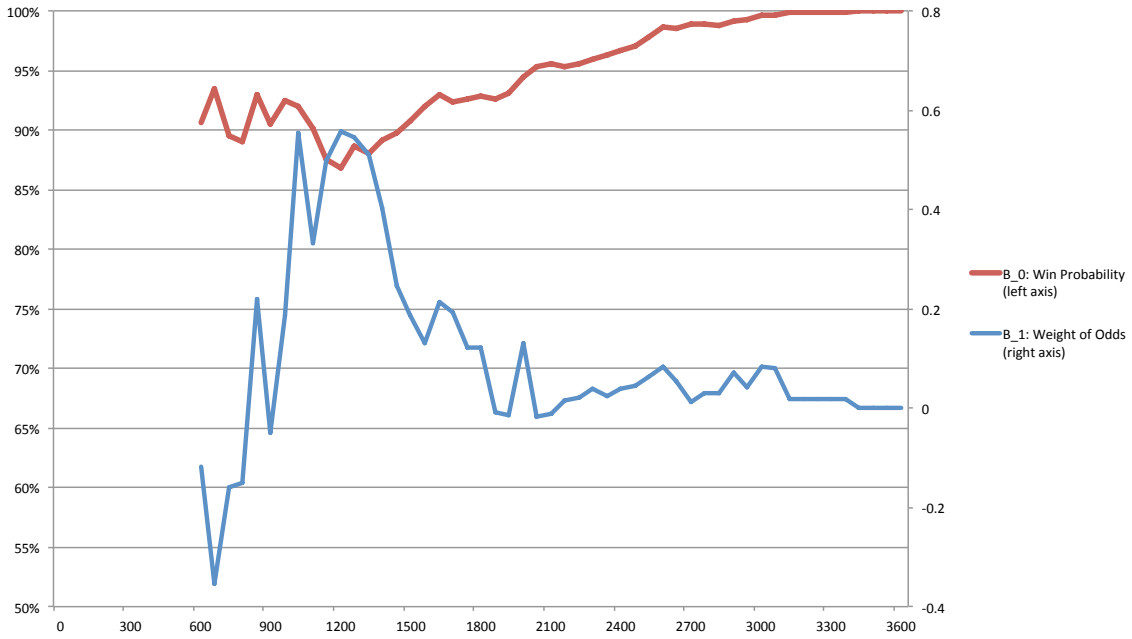
Appendix Figure 9: Regression coefficients where the home team is up by two goals



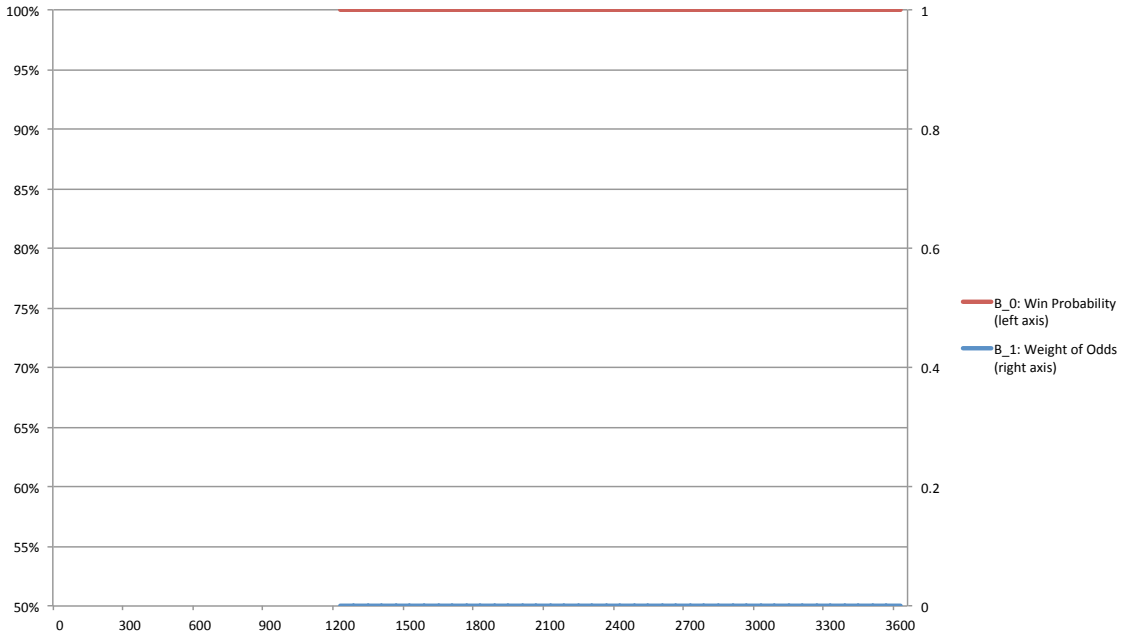
Appendix Figure 10: Regression coefficients where the away team is up by two goals



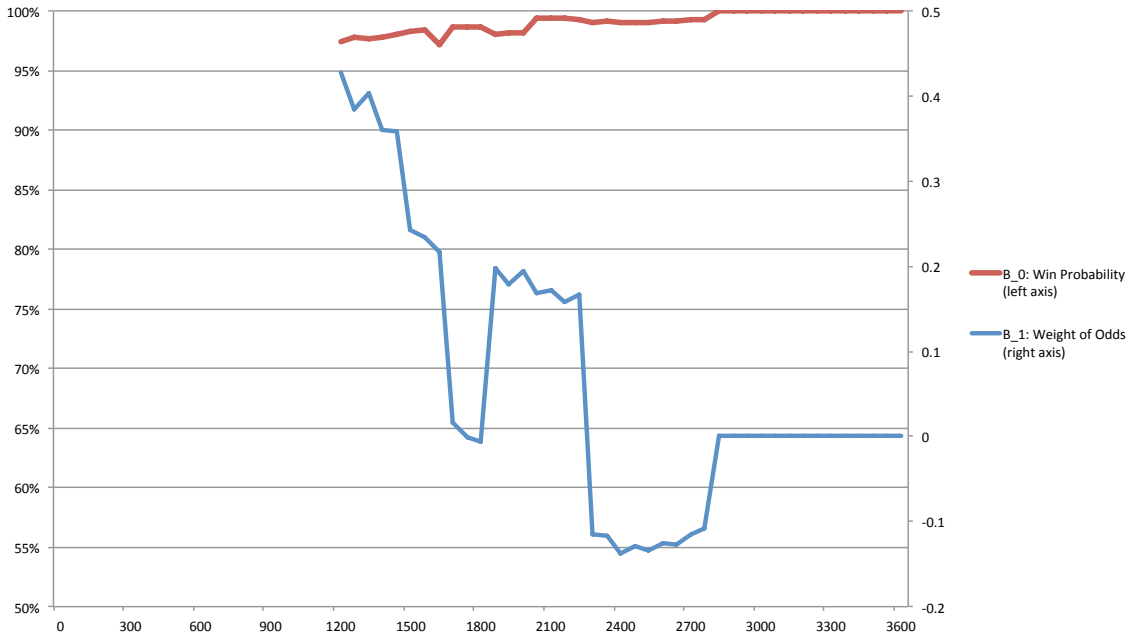
Appendix Figure 11: Regression coefficients where the home team is up by three goals



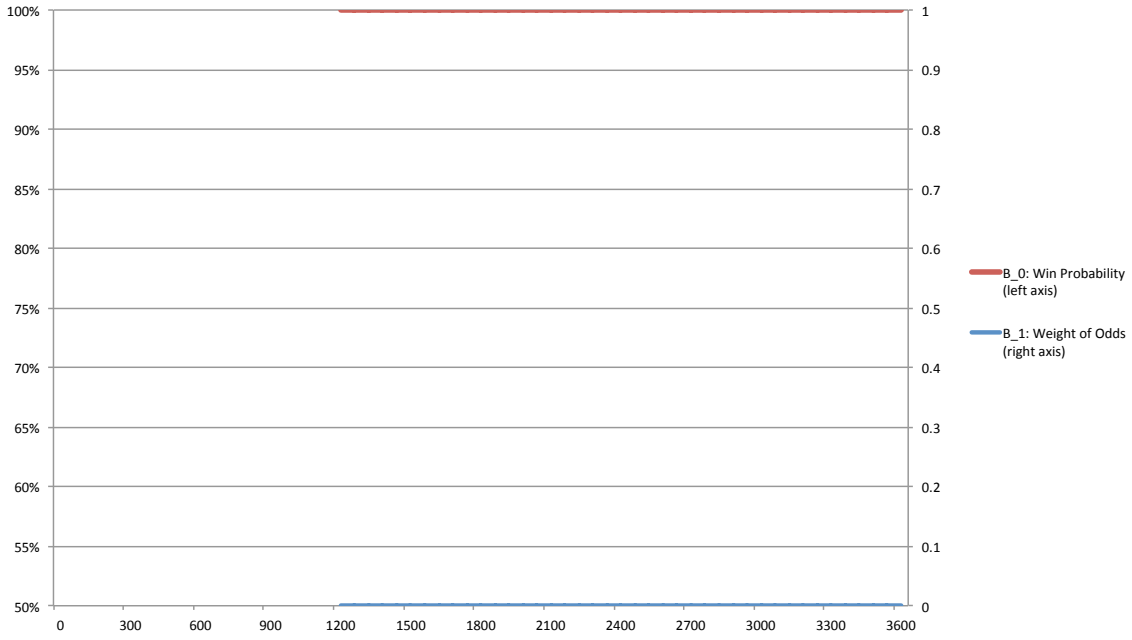
Appendix Figure 12: Regression coefficients where the away team is up by three goals



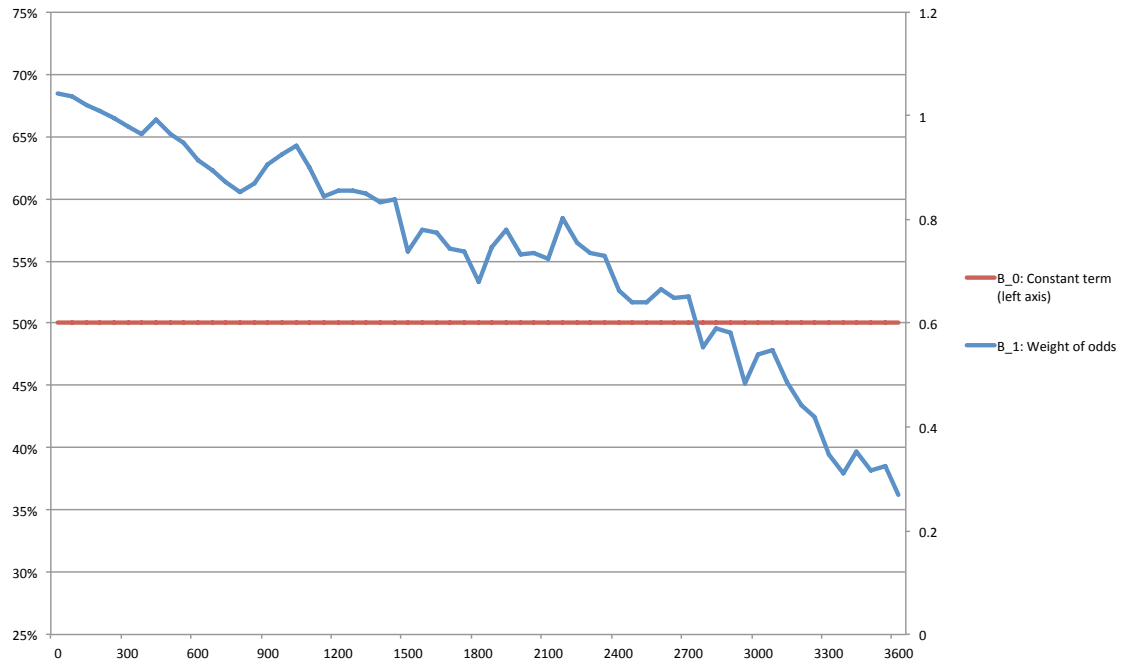
Appendix Figure 13: Regression coefficients where the home team is up by four goals or five or more goals



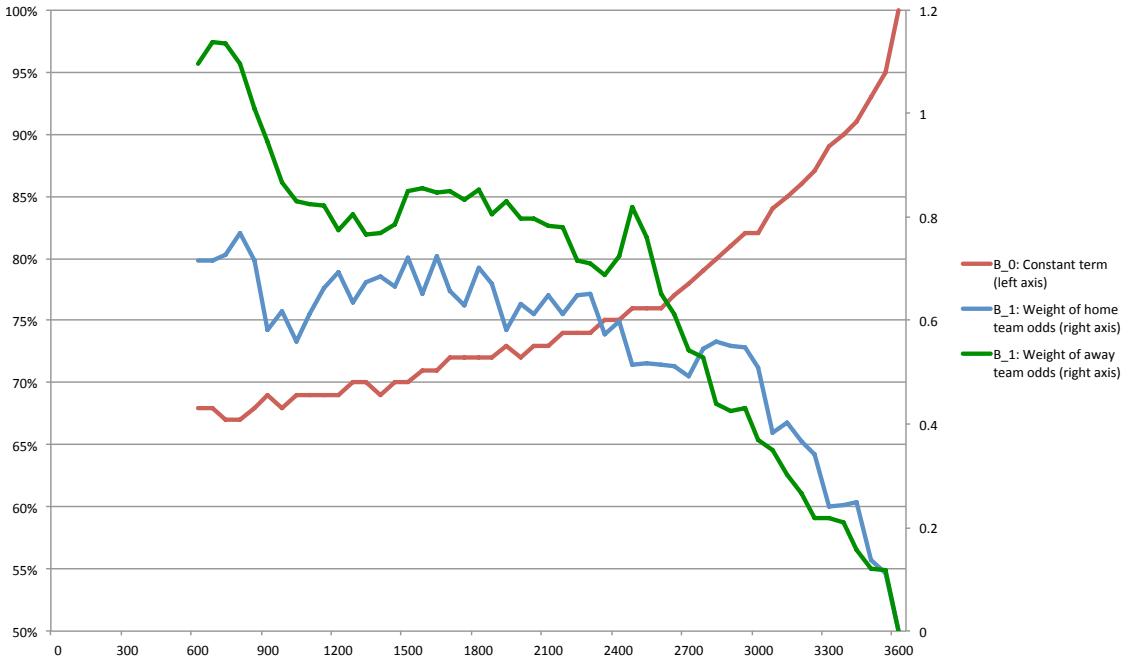
Appendix Figure 14: Regression coefficients where the away team is up by four goals



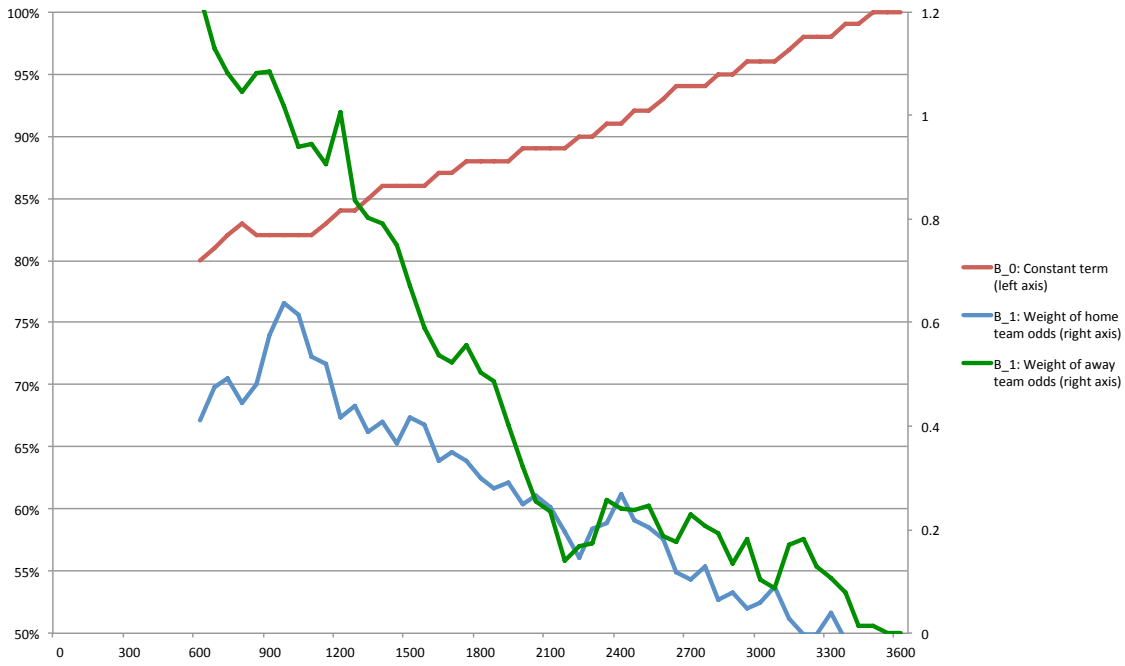
Appendix Figure 15: Regression coefficients where the away team is up by five or more goals



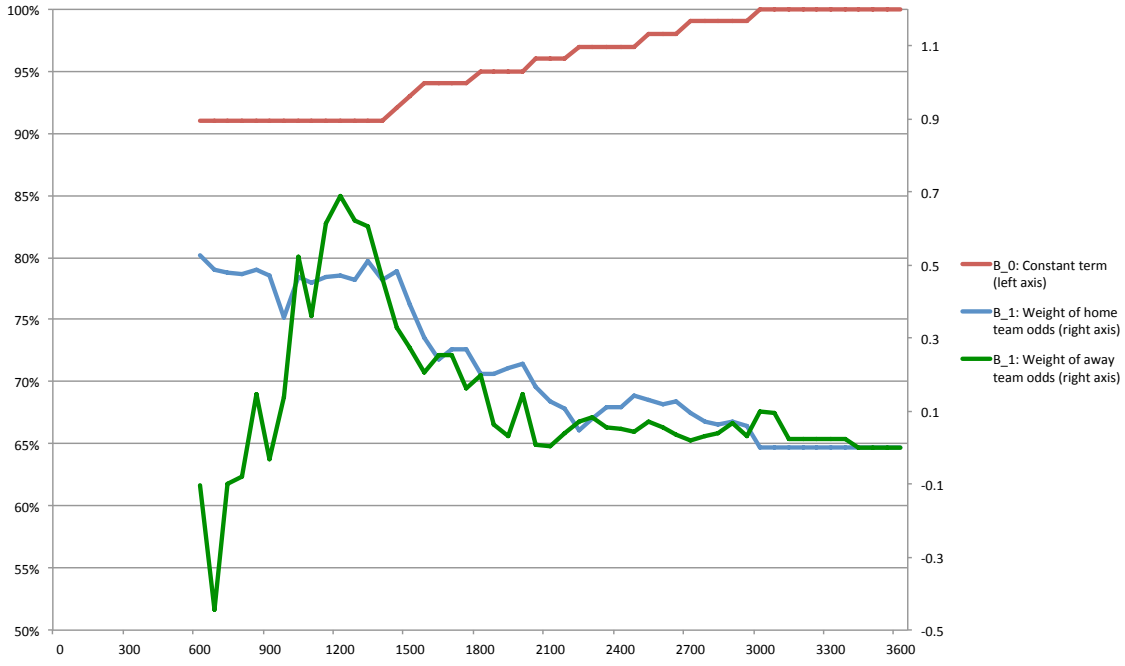
Appendix Figure 16: Restricted regression coefficients for a tied game



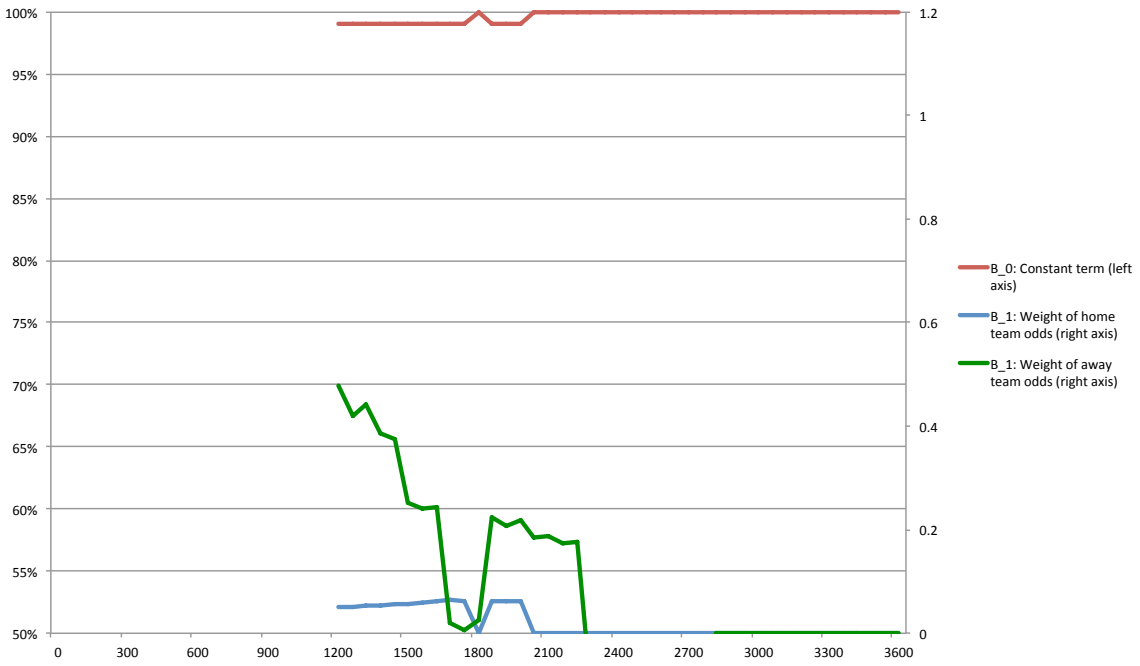
Appendix Figure 17: Restricted regression coefficients for a one goal differential



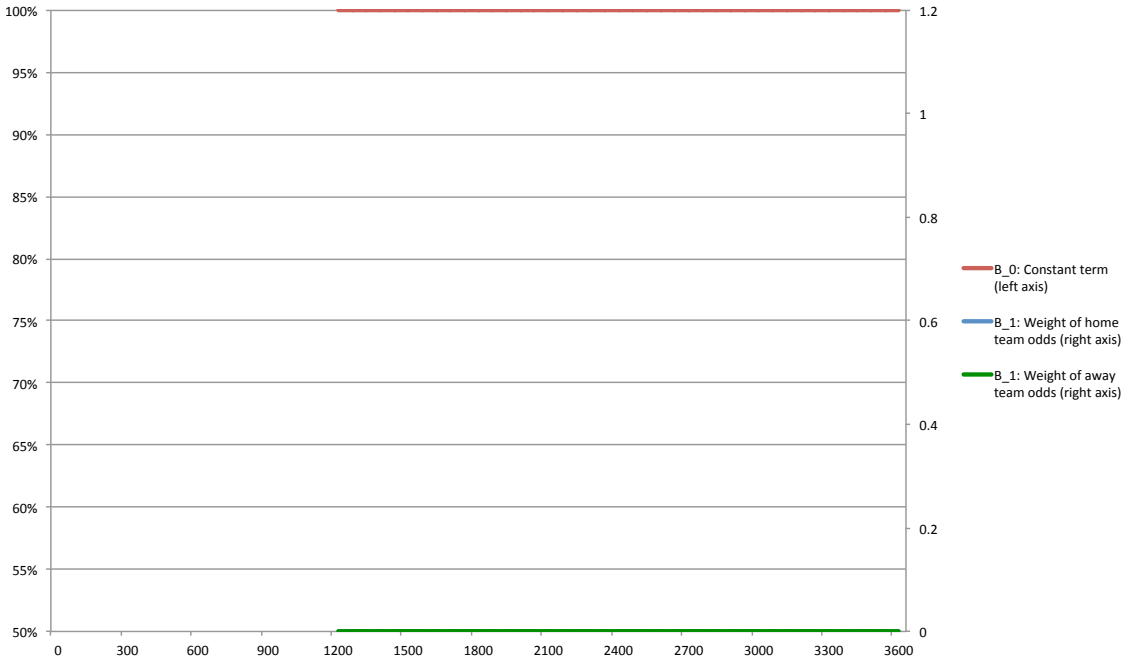
Appendix Figure 18: Restricted regression coefficients for a two goal differential



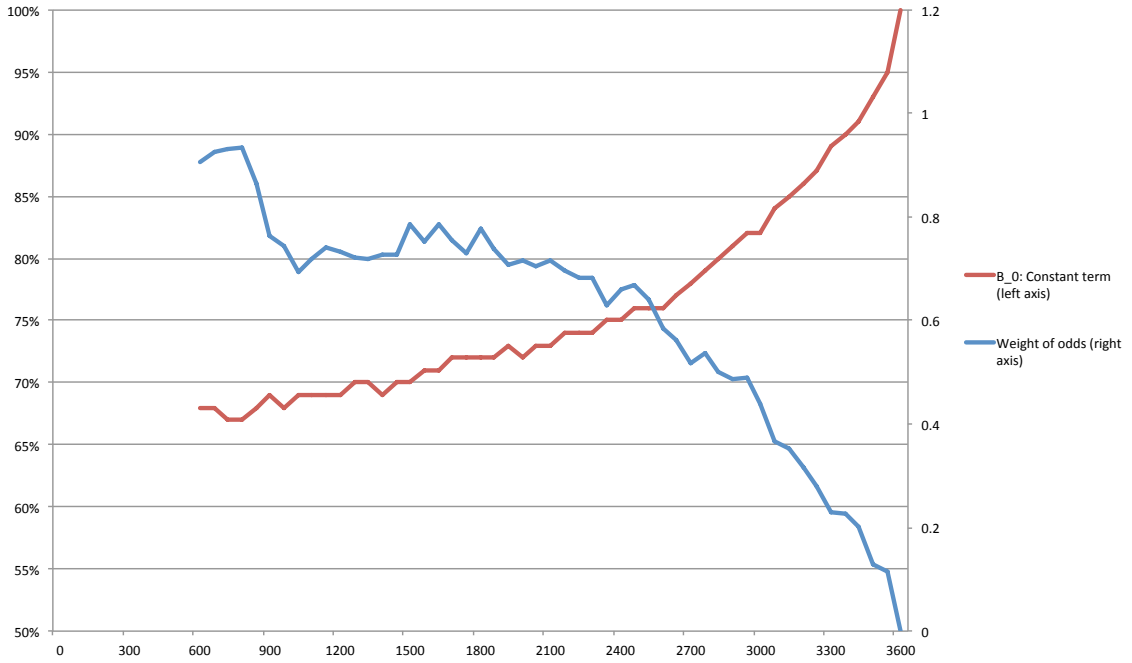
Appendix Figure 19: Restricted regression coefficients for a three goal differential with the constant term adjusted



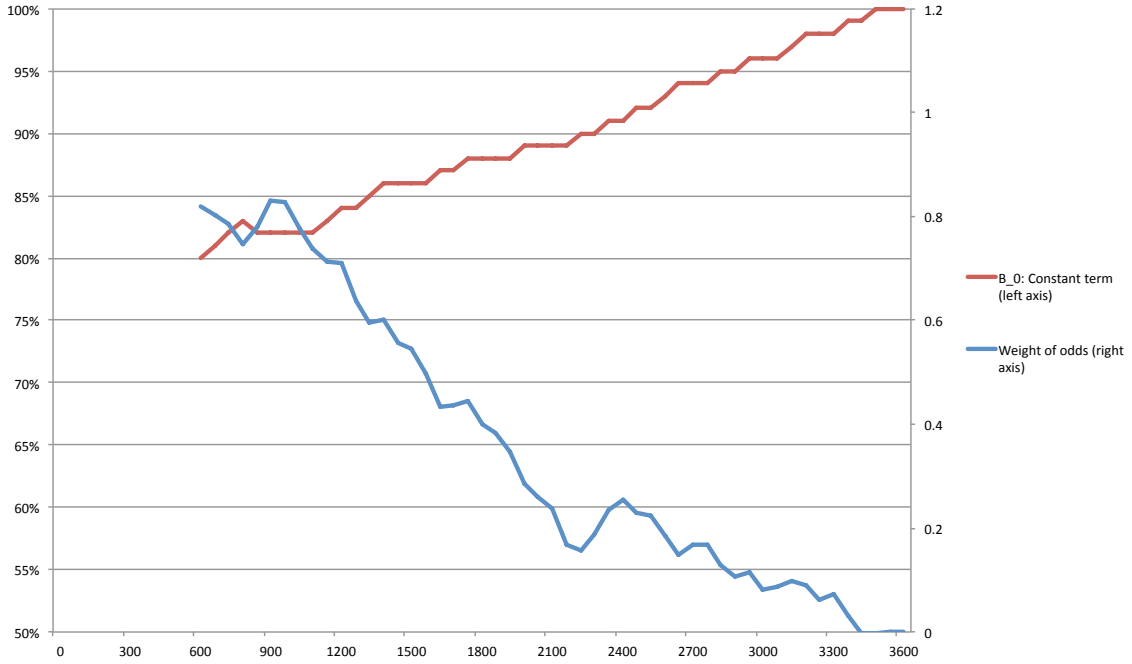
Appendix Figure 20: Restricted regression coefficients for a four goal differential



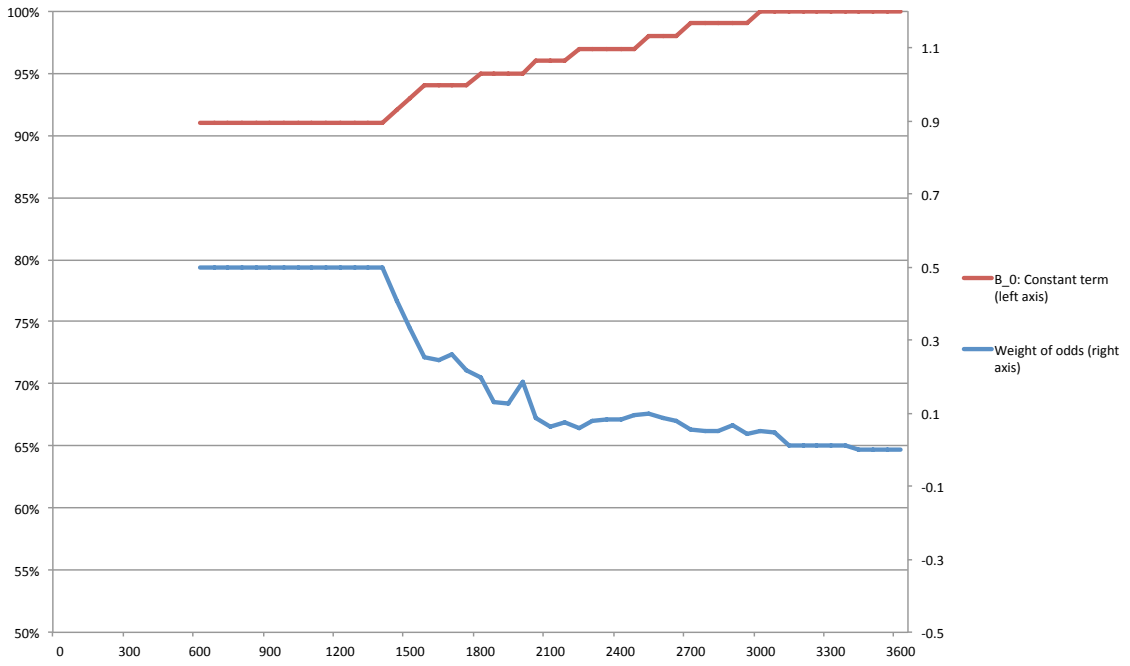
Appendix Figure 21: Restricted regression coefficients for a five goal differential



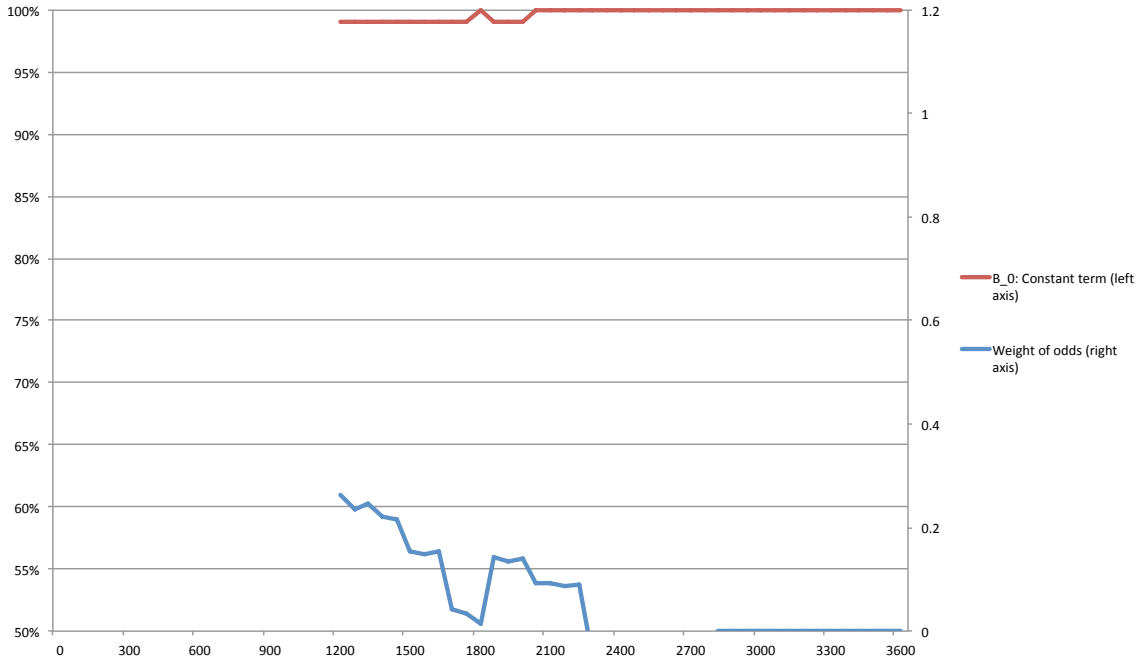
Appendix Figure 22: Restricted regression coefficients for a one goal differential, where the weight of the odds is the average of the home and away coefficients



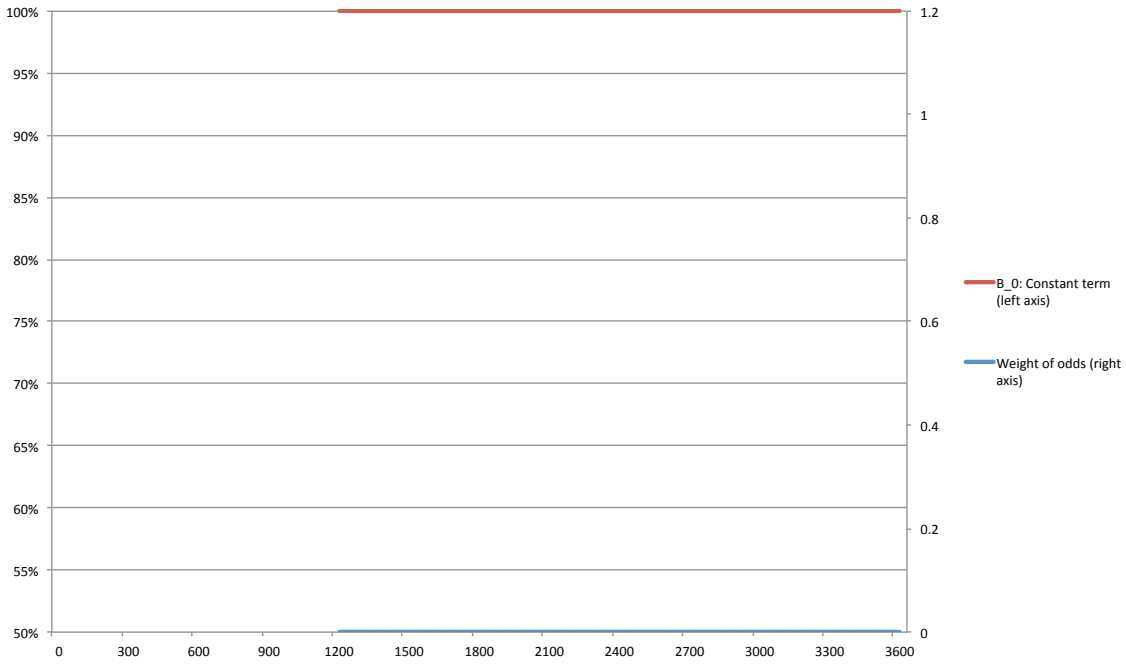
Appendix Figure 23: Restricted regression coefficients for a two goal differential, where the weight of the odds is the average of the home and away coefficients



Appendix Figure 24: Restricted regression coefficients for a three goal differential, where the weight of the odds is the average of the home and away coefficients and both curves were capped before the 23rd minute



Appendix Figure 25: Restricted regression coefficients for a four goal differential, where the weight of the odds is the average of the home and away coefficients



Appendix Figure 26: Restricted regression coefficients for a five goal differential, where the weight of the odds is the average of the home and away coefficients