NHL Win Probability Model

Devin Basley
Earlham College
Richmond, Indiana
dcbasley18@earlham.edu

ABSTRACT

Equal talent and competitiveness in the game of hockey has led to a high level of difficulty in predicting which team wins a game. As new analytical strategies and tools are utilized, the outcomes of hockey games is evolving in how it is predicted. Newer data and information about hockey is becoming more available. There is a growing importance for win probability from front offices around the league to fans of their favorite teams. This paper describes modeling win predictions using logistic regression and real world hockey data from 2008 to 2018.

KEYWORDS

datasets, logistic regression, win probability

ACM Reference Format:
Devin Basley. 2022. NHL Win Probability Model. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

NHL games have often been considered one of the hardest professional sports games to predict. Competitive and equal talent level between teams and "puck luck" makes these games difficult to predict. The impact of "puck luck" such as a goal going off a skate or a shot hitting off the post can significantly change the outcome of a game. These types of impacts can be nearly impossible to account for. My model tries to limit the impact of "puck luck" and predict games accurately. The evolution of hockey analytics has resulted in newer strategies to predict the outcomes of these hockey games. These strategies range from modeling impacts of puck possession vs dumping the pucks deep into the offensive zone to rating the impact of scoring chances and measuring save difficulty. Being able to predict the outcomes of these games accurately is important to the organizations' front offices to best put their teams in positions to be successful. The importance of win probability is the same throughout every sport because it gives the spectator an idea of which team is likely to win and where each team will end up in the standings. For front offices, having an idea of where the team is in the standings matters because it identifies if the team is competing for the playoffs and who their ideal match up is.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Throughout the rest of the paper, we'll discuss the methods and process of the project and the model, the results of the model, and how they compare to other win probability models.

2 RELATED WORK

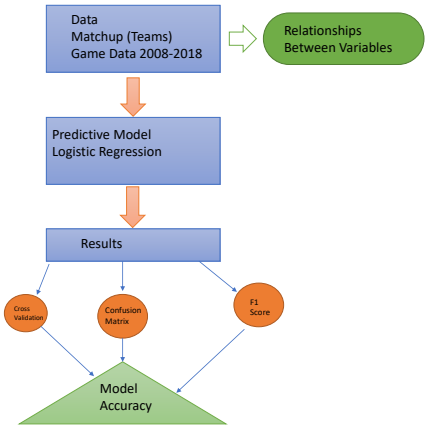
Win probability models are widely used across all sports such as baseball, football and basketball. Some football win probability models use bayesian statistical models, or logistic regression.The biggest difference between football and hockey is the pace of play. Hockey is much more fluid and moves at a faster pace with more unpredictability whereas football is more stagnant and moves at a slower pace. The difference in speed of the game makes football much easier to predict than hockey. Hockey is more similar to basketball in its fluidity and pace of play but the substitution of players in hockey makes it difficult for one player to take over a game whereas in basketball, one player can take over a game and play every minute of the game. Win probability has expanded into much more than just pregame win probabilities.There are now variables such as Win Probability Added in football and baseball that measures how a specific in game play effects the win probability of each team. These techniques have crossed over into hockey in many win probability models. Others use Markov chains, poisson distributions for scoring rates, random forests to predict game outcomes. One of the most popular models is MoneyPuck's win probability model which includes 3 sub models into the main model. Those components are a home team model, away team model and an overtime model. These three models get put together into the large model which used logistic regression with gradient boosting.

3 METHODOLOGY

The flow of my project can be seen in data framework figure. The beginning stages of the project consisted of finding hockey data and cleaning it. The middle stages of my project consisted of exploratory data analysis and model prediction. The final stages of the project is analyzing the model and assessing its accuracy. The exploratory data analysis consisted of looking for relationships between possible predictor variables and the response variable of game outcome. Variables I looked at were blocked shots for teams in the game, faceoff percentage, and the number of powerplay opportunities the team had. From these variables I used a linear regression model to check for relationships between those variables and game outcome. I ran linear regression for the three variables combined and separately but with each model there was not a clear relationship with game outcome. I then used a logistic regression model with the same strategy of the three variables combined and all separately but ultimately I did not include these variables into my main model although the logistic regression performed better than the linear regression. I then moved to looking at expected goals for which how many goals the team is expected to score in the game to be

played. From expected goals for, I took an average for each teams' expected goals for probability over a 10 year period of 2008 to 2017 to see any trends in how expected goal probabilities changed from team to team. After exploratory data analysis, comes the model and I started with a multi linear regression model. To improve upon the multi linear regression model, I changed to a logistic regression model.

3.1 Data Framework



The framework starts with data collection and exploratory data analysis. The data collection and exploratory data analysis steps include looking for relevant variables to help predict game outcome. After finding those variables through data analysis, we used those variables in a logistic regression model on a training data set. Once the model was ran on the training data set, it was then used to predict game outcome on the test data set. From there, we move to model assessment and that was done using a confusion matrix, misclassification rate, and an F1 score.

3.2 Data Collection

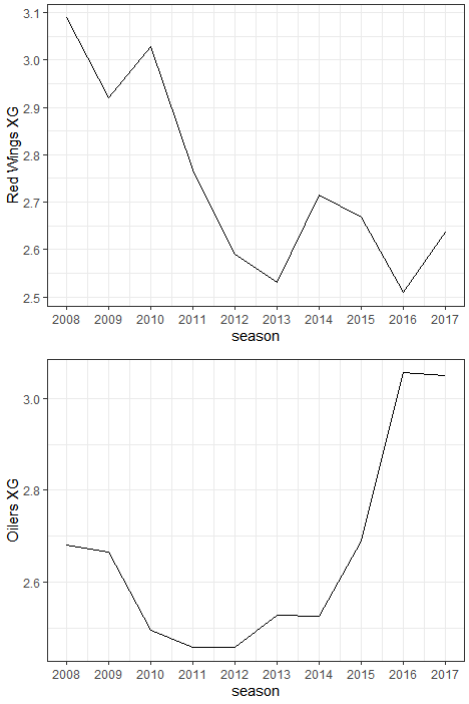
I started out using hockey data from the 2016 season to the 2021 season. The data includes variables such as blocked shots, powerplay opportunities, and faceoff percentage. The data includes 13 smaller data sets containing various game, team, and player information such as who the goal scorers were, the starting goalies, and where the game was played at. Cleaning the data included merging data sets to include possible important variables to predict game outcome.

I then found more data dating further back to 2008 to 2021 because to try and remove some variability from the Covid-19 years in the smaller data set. This data set included variables such as teams' expected goals for, expected time in the offensive zone and each teams' corsi percentage per each game. The data set included 32,534 observations and 112 variables. The data wasn't complete for seasons 2018-2021 so they were removed from the data set. The table below is an example of what the data looked like.

team	season	name	gameId	playerTeam	opposing1	home_or_away	gameDate	position	situation	xGoalsPercentage	corsiPercentage
NYR	2008	NYR	2008020001	NYR	T.B	AWAY	20081004	Team Level	other	0	0
NYR	2008	NYR	2008020001	NYR	T.B	AWAY	20081004	Team Level	all	0.4596	0.6408
NYR	2008	NYR	2008020001	NYR	T.B	AWAY	20081004	Team Level	5on5	0.4857	0.6429
NYR	2008	NYR	2008020001	NYR	T.B	AWAY	20081004	Team Level	4on5	0.0482	0.0909
NYR	2008	NYR	2008020001	NYR	T.B	AWAY	20081004	Team Level	5on4	0.7317	0.9524
NYR	2008	NYR	2008020003	NYR	T.B	HOME	20081005	Team Level	other	0.465	0.625
NYR	2008	NYR	2008020003	NYR	T.B	HOME	20081005	Team Level	all	0.6619	0.6207
NYR	2008	NYR	2008020003	NYR	T.B	HOME	20081005	Team Level	5on5	0.5888	0.6145

3.3 Exploratory Data Analysis

Many variables go into how an NHL team can win a hockey game. Variables I looked at to predict win probability is each team's expected goal probability for each game, their face-off percentage, how many blocked shots they had in the game, the time zone for the game and the time zone from location of the away team, and lastly how many powerplay opportunities the home team had. Each team's expected goals probability had a strong relationship with the outcome of the game.



This figure is an example of the time series for each teams' average expected goals from seasons 2008-2017. This time series is of the Detroit Red Wings and we can see in 2008 and 2009 that their average expected goals was high and they appeared in the Stanley Cup Finals in both of those years. After those years we see a clear decline in their average expected goal probability. This time series was done for all the teams that existed during the 2008-2017 time frame.

3.4 Game Prediction

The initial model for win probability was a multi linear regression model with predictor variables of blocked shots, powerplay opportunities, and faceoff percentage and a response variable of outcome with 0 being a loss and 1 being a win. This model performed poorly. I used logistic regression with the same predictor variables and the same response variable. This model performed better but still not great.

4 RESULTS

To assess the logistic regression model, we used a confusion matrix and calculated the model's misclassification rate, specificity and sensitivity rate, and its F1 score. We assessed the model on

the test data set which was created with random sampling with replacement.

Confusion Matrix

	Predicted Loss	Predicted Win
Actual Loss	3045	1512
Actual Win	1393	2803

From this Confusion Matrix, the misclassification rate was 0.297. The sensitivity rate was 0.649. The specificity rate was 0.686. The F1 score was 0.677. Based on these results we were able to accurately predict a win when the team won at a rate of about 65% of the time and we were able to accurately predict a loss when the team lost at a rate of 69% of the time.

Table 1: Logistic Regression Model Accuracy

Assessment	Rate
F1 Score	67.7%
Specificity Rate	68.6%
Sensitivity Rate	64.9%
Misclassification Rate	29.7%

5 FUTURE WORK

Once NHL tracking data becomes available, more in-depth variables could become useful to predicting game outcomes. Another possible piece of future work would be to include a Win Probability Added component for plays that happen in hockey. Lastly, a piece of future work would be to either add other variables or try a different model to improve upon the accuracy of my current model. A piece of future work that I unfortunately was unable to complete would be to compare my logistic regression model to a random forest model and check which model performed better. Another piece of future work is to take into account whether the team was the home or away team.

6 ACKNOWLEDGEMENTS

I would like to thank Professor Fariba Khoshnasib, Professor Charlie Peck, and the Earlham Data Science Department for their guidance throughout working on this project.

REFERENCES

[1] "Capozzi, Z., 2022. Beyond the Basics: Understanding the Uses of Win Probability. USA Lacrosse Magazine, [online] Available at: <https://www.usalaxmagazine.com/college/women/beyond-the-basics-understanding-the-uses-of-win-probability> [Accessed 17 May 2022]."

[2] "Moneypuck.com. 2022. MoneyPuck.com -About and How it Works. [online] Available at: <https://www.moneypuck.com/about.htm> [Accessed 17 May 2022]."

[3] "Robberechts, P., Haaren, J. and Davis, J., n.d. Who Will Win it? An In-game Win Probability Model for Football. [online] Dtai.cs.kuleuven.be. Available at: <https://dtai.cs.kuleuven.be/events/MLSA19/papers/robberechts$_MLSA19.pdf$ > $[Accessed 17 May 2022]$."

[4] "NHL.com. 2022. Stanley Cup Champions. [online] Available at: <https://www.nhl.com/news/nhl-stanley-cup-champions-winners-complete-list/c-287705398> [Accessed 17 May 2022]."

[5] "Yang, S., 2016. Estimating the Win Probability in a Hockey Game. [ebook] Sudberry, Ontario, Canada. Available at: <https://zone.biblio.laurentian.ca/bitstream/10219/2569/1/Thesis-Shudan%20final%20version.pdf> [Accessed 17 May 2022]."