NHL Win Probability

Devin Basley
dcbasley18@earlham.edu

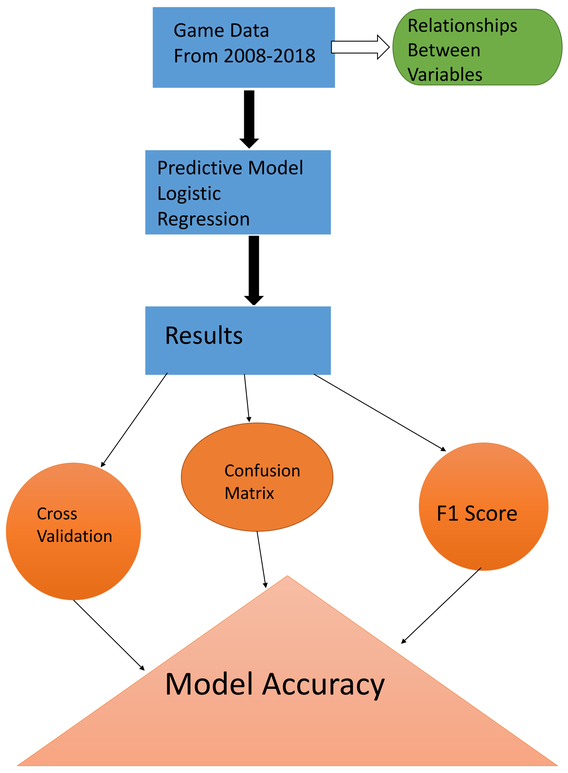
1. MOTIVATION

Predicting NHL hockey games is one of the hardest tasks in sports analytics. Most models can accurately predict the winner of a game 60% of the time. Discovering a way to address variability in "Puck Luck" and talent differences between teams.

2. DATA

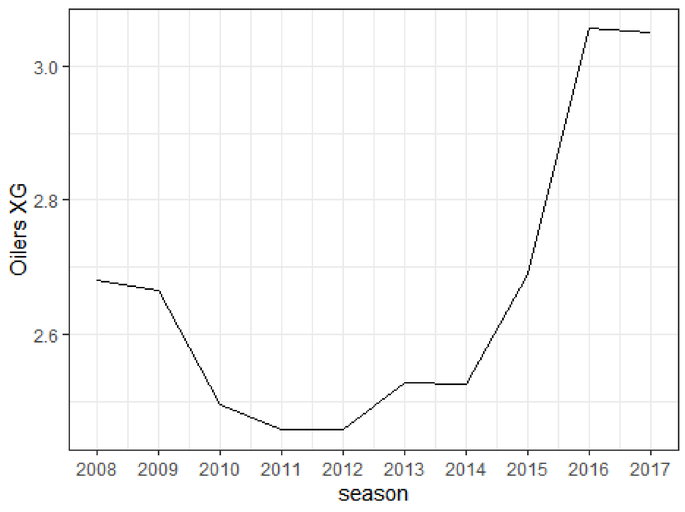
Game Data from 2008-2018.
The Outcome variable is binary with 0 being a loss and 1 being a win. The Expected Goals For and Expected Time in Zone are both continuous variables.

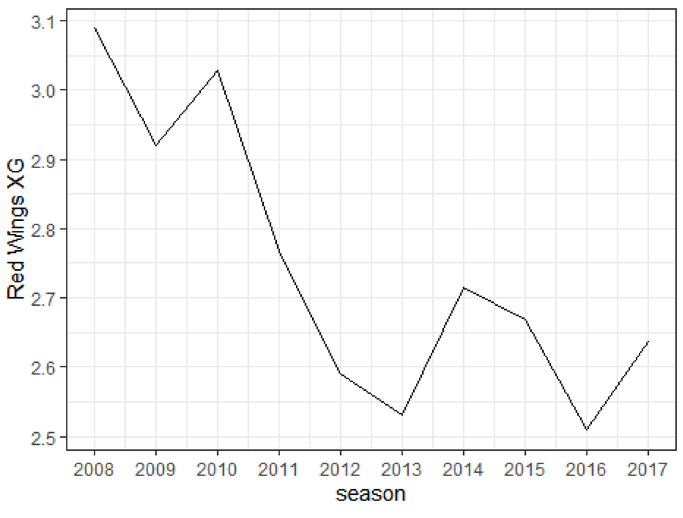
3. PROJECT FRAMEWORK



4. METHODS AND PROCESS

I started with exploratory data analysis to determine any applicable variables to predict win probability. After EDA, applicable variables were Expect Goal Probability, Expected Offensive Zone Time and Corsi Percentage.
To predict win probability, I tested a multiple linear regression with variables Blocked Shots, Powerplay Opportunities, and Faceoff percentage to check if those variables were related to the outcome of the game.
After multiple linear regression, we ran the model using logistic regression for better results.





5. RESULTS

The logistic regression model used Outcome as the response variable with Expected Goals For, Corsi Percentage, and Expected Time In Zone For as the predictor variables.

The Misclassification rate was 29.8%.
The Sensitivity rate was 64.9%.
The Specificity rate was 68.6%.
The F1 Score was 67.7%.

CONFUSION MATRIX RESULTS BELOW:

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

6. CONCLUSION

The misclassification rate is nearly 30% which is high and indicative of poor model performance. The nearly 65% sensitivity rate means we were able to predict a win correctly at a 65% rate. The 68% specificity rate means we correctly predicted a loss at a 68% rate. The F1 Score being nearly 68% is generally low accuracy but relative to predicting hockey games it is a comparable to other win probability models.

7. FUTURE WORK

Possible more in-depth variables once tracking data becomes available.
Use of different models or different variables to improve accuracy such as a Random Forest model.
Include more assessments of model accuracy such as brier scores.
Player Effects on team win probability.

8. 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.