Can Machine Learning Help Predict Goals in an NHL Season?*

Davis Dunkleberger†

April 25, 2023

Abstract

A short summary of what question the project answers, what methods are used, and any policy (or business) implications from the findings.

*Thank you to Professor Ransom for nailing down the topic of research

†Department of Health and Exercise Science, University of Oklahoma. E-mail address: davisdunk@ou.edu

1 Introduction

One of the biggest factors in a season that teams must answer is the product of scoring goals. More goals lead to more wins which can lead to a playoff appearance. The Stanley Cup is the most coveted trophy in hockey and can drive fans and owners crazy. Teams will explore every avenue they can to score more goals, win more games, and win Lord Stanley's Cup.

One of the big trends within the sports world has been the analytics boom brought about in the 2000s. Teams have been exploring new ways to adjust and gain an advantage over the other teams in their league. Major League Baseball adopted Sabremetrics, the National Football League has adopted a more aggressive fourth down strategy, and the National Basketball Association has started taking more 3s and layups all due to analytical reasoning. The National Hockey League has...nothing like that. The lack of hockey analytics is an issue and has lagged the game behind in the public discourse. Creating a framework to make decisions would help advance the sport of hockey in the world of analytics.

With the recent developments of machine learning, seeing if it can accurately predict hockey related statistics and outcomes could help front office decision makers make their decisions on roster construction. Creating an analytical framework to predict the goals of a season is a first step in moving hockey analtyics forward. Front office decision makers are looking for new ways to make decisions. Making their own models to predict things could be an important part of a possible Stanley Cup run.

Machine Learning can be trained to produce a model to classify if shots are a goal or not. This can be used to see if machine learning is a viable process for roster decisions. Testing the tribes of machine learning to see how accurate they are will help front office decisions makers make that

decision easier.

2 Literature Review

1. General topics on analytics movements in sports (Moneyball, 4th down, etc.)
2. Hockey performance analytics movement (Stat Shot)
3. Possible intersection of ML and sports/hockey analytics
4. Possible NHL front office research

3 Data

The primary data source for this research is National Hockey League play by play data from the hockeyR package.

1. Start with 112 columns of data
2. Filter down to shot events(missed shots, shots on net, goals). Excluding blocked shots due to the fact they do not score goals
3. create factors for text variables I'm interested in
4. remove columns that are unnecessary from dataset
5. drop rows that have NA values due to errors caused in empirical research

4 Empirical Methods

While my approach explores a number of different approaches, the primary empirical model can be depicted in the following equation:

$$Goal = \alpha_0 + \alpha_1 SD + \alpha_2 SA + \alpha_3 P + \alpha_4 PSR + \alpha_5 SH + \alpha_6 PP + \varepsilon, \tag{1}$$

where *Goal* is a binary classification for if a shot is a goal or not, and *SD* is the shot distance from the net, while *SA* is the angle of the shot from the middle of the net when facing it, while *P* is the period the game is in, while *PSR* is period seconds remaining, while *SH* and *PP* are dummy variables that denote if the shot was short handed or on the power play, respectively.

1. Set up multiple ML classification models
2. Record accuracy for different ML models
3. Compare accuracy to find best model
4. Denote which one could be used in real application, if any

5 Research Findings

The main results are reported in Table 2.

Summary: Discuss results on if ML is a useful tool for sports analytics and such.

6 Conclusion

Summary: ML is (not) a good tool for this project. Where to go from here and how it impacts sports analytics and hockey specifically. Make clear how to apply things.

Figures and Tables

To be filled in

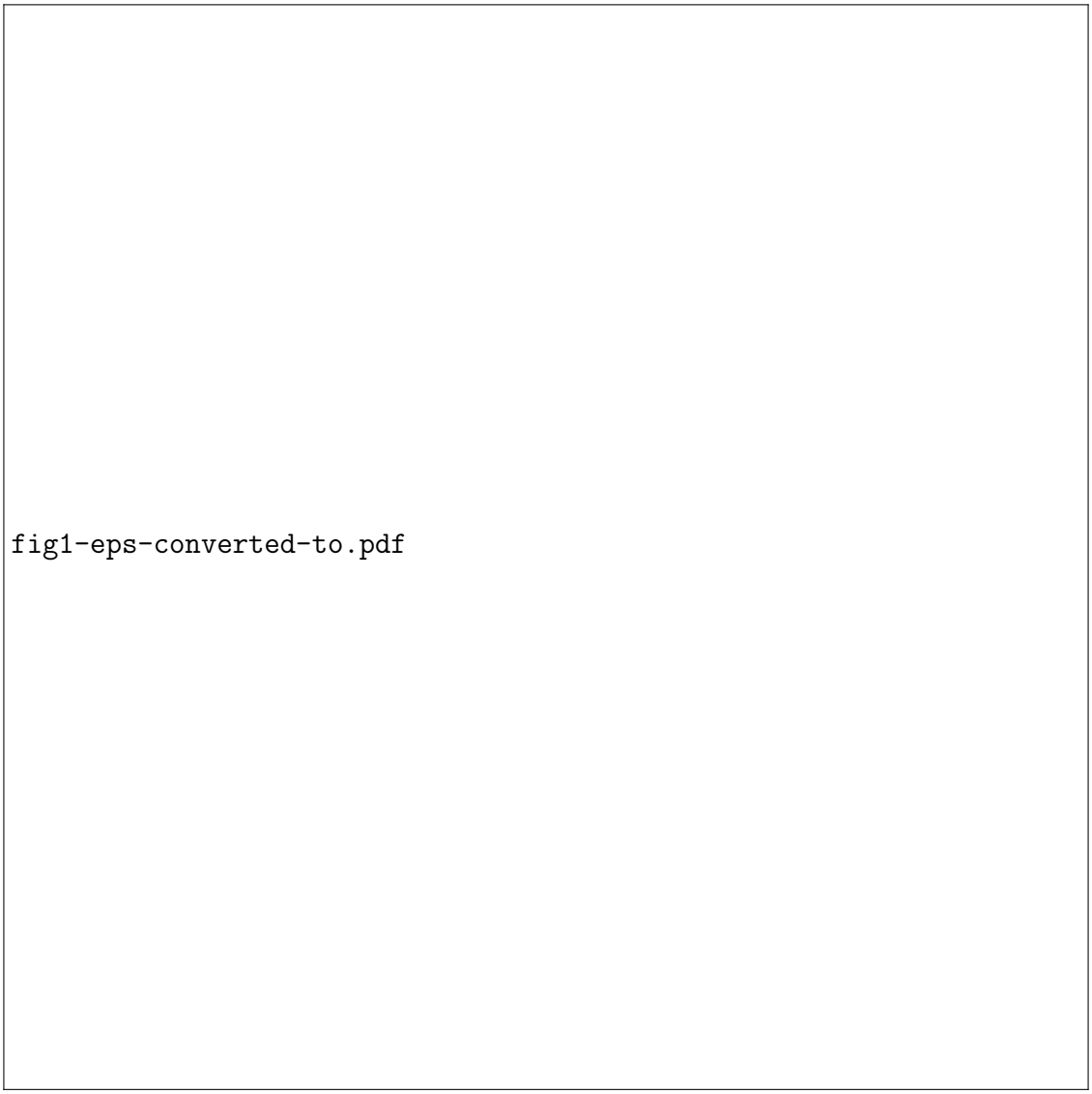

Figure 1: Figure caption goes here

To be filled in

To be filled in

Table 1: Summary Statistics of Variables of Interest

Panel A: Summary Statistics for Variables of Interest

	Mean	Std. Dev.	Min	Max
Outcome variable 1	4.127	1.709	0.000	8.516
Outcome variable 2	1.293	0.648	0.000	0.216
Policy variable	0.685	0.464	0.000	1.000
Control variable 1	0.451	0.497	0.000	1.000
Control variable 2	0.322	0.467	0.000	1.000

Panel B: Sample Means of Outcome Variables for Subgroups

	Group 1	Group 2	Group 3	Group 4
Outcome variable 1	1.782	2.181	3.749	4.127
Outcome variable 2	0.824	0.971	1.215	1.693
N	25,796	75,879	37,157	33,839

Notes: Put any notes about the table here. Sample size for all variables in Panel A is $N = 172,671$.

Table 2: Empirical estimates of parameter of interest

	Few Controls	Many Controls
Variable of interest	-1.977***	-0.536**
	(0.219)	(0.214)
Individual characteristics	✓	✓
Firm characteristics		✓
Location dummies		✓
N	172,671	172,671

Notes: Table notes here. Standard errors in parentheses. ***Significantly different from zero at the 1% level; **Significantly different from zero at the 5% level.