How bad should I win next faceoff?

Big data cup 2021

Stéphane Caron Jean-Philippe Le Cavalier Samuel Perreault

Over the course of a typical hockey game, one could expect that around 50 to 75 faceoffs will be contested. The outcome is often meaningful in determining which team will first dictate the play in the upcoming sequence. This obvious statement is probably enough to suggest that—as the old saying goes—each and every faceoff is important in a hockey game. The main objective of this study is to quantify how much. In other words, in leveraging the data at our disposal, we try to answer the following question:

> **What is at stake when the linesman drops the puck?**

It is reasonable to assume that the answer may vary significantly depending on the context in which a particular faceoff occurs. A non-exhaustive list of important factors may include the current score, the time remaining on the clock, the zone in which the faceoff is taken, and the strength of play (e.g. power play). Many will also argue that—even though there is not a single coach who overlooks the importance of winning faceoffs—some game plans are better to make the most of an offensive zone faceoff win, while others are more effective at salvaging a faceoff loss in the defensive zone.

The first step in answering our question objectively is to define a reliable metric to help us guide our reasoning, and eventually come up with conclusive statements. For the purposes of this study, we will evaluate how the outcome of a faceoff affects the likelihood of scoring or allowing a goal on the ensuing sequence.

Outline of the study. The report is divided in three main parts. In the first part, we analyze the effect of winning faceoffs on a sequence basis, where a sequence is defined as the time between two successive faceoffs. In the second part, we raise the bar higher by performing a time series analysis in which we evaluate how the probability of scoring a goal evolves over time after the faceoff. Finally, in the last section, we perform the two aforementioned analyses on a much larger dataset so as to corroborate our initial findings.

The overall impact of winning faceoffs

The first thing to validate is that winning a faceoff does provide an advantage. A simple way of doing this is to treat each sequence as an observation and fit a logistic regression with a goal indicator as the response variable. By including the faceoff outcome as a covariate, we hope to validate that there is statistical evidence that winning a faceoff has a positive effect. Although reducing the probability of allowing a goal is as important as increasing the probability of scoring a goal, we focus on the latter in this section.

In order to work with relatively homogeneous data, we made the decision to only use the scouting dataset. Indeed, mixing different leagues may introduce noise, as each of them may imply a different style of hockey. Furthermore, the fact that every games in the data involve the Erie Otters allows us to create models from a

Table 1: High-level features of the scouting dataset (Erie Otters perspective)

Min. game date	Max. game date	Games	FO	OZ FO	Goals	OZ Goals
2019-09-20	2020-03-08	40	2441	852	148	63

Table 2: Excerpt of the sequence-by-sequence data

game_date	period	clock_begin	clock_end	FO_win	FO_zone	FO_OZ	GF
2019-09-20	1	20:00	18:57	FALSE	red	FALSE	FALSE
2019-09-20	1	18:57	18:29	FALSE	offense	TRUE	FALSE
2019-09-20	1	18:29	15:27	FALSE	offense	TRUE	FALSE
2019-09-20	1	15:27	14:10	FALSE	offense	TRUE	FALSE
2019-09-20	1	14:10	13:42	FALSE	blue_offense	FALSE	TRUE
2019-09-20	1	13:42	12:41	TRUE	red	FALSE	FALSE

single perspective. It would be somewhat inappropriate to build a model with such asymmetrical data and claim that it represents the reality for the whole league, the OHL in this case.

With that in mind, let us take a look at the data we are working with. A summary of it is provided in Table 1—OZ stands for Offensive Zone—and an excerpt of the training dataset we use for the logistic regression, including created features, is displayed in Table 2.

To get a crude idea of the impact on goal scoring of winning or losing faceoffs, we first fitted a logistic regression with only an intercept and `FO_win` as covariate. We obtained $\beta_{\text{FO\_win}} = 0.974$, which suggests that winning offensive faceoffs does increase the chances of scoring a goal during the sequence that ensues. However, the associated p-value is too high to make any statistically backed conclusion. We nevertheless designed alternative linear predictors using additional covariates to try to capture a clearer signal. Adding the binary `FO_OZ` covariate in the model, we obtained two positive coefficients $\beta_{\text{FO\_win}} = 1.024$ and $\beta_{\text{FO\_OZ}} = 1.063$. Since there is probably an important interaction between the faceoff zone and the outcome of the faceoff, we fitted a last logistic model including the `FO_win` * `FO_OZ` interaction. With this model, the coefficients of interest were $\beta_{\text{FO\_win}} = 0.336$, $\beta_{\text{FO\_OZ}} = 0.237$ and $\beta_{\text{FO\_win+FO\_OZ}} = 1.228$, suggesting that winning a faceoff or having an offensive zone faceoff gives great opportunities, but that the real deal is to win those offensive zone faceoffs. It must be said however that all p-values were well above any reasonable threshold for statistical significance.

We conclude this section with a one-way analysis, shown in Table 3, that groups faceoffs by "situation" (penalty kill, even strength, power play). Although analyzing the impact of winning faceoffs in such specific situations would have been very interesting, Table 3 suggests that this would be inconclusive.

Table 3: Contextual data for faceoff situations

		Goal Against				**Goal For**			
		Defense		Neutral		Neutral		Offense	
	Faceoff	Won	Lost	Won	Lost	Won	Lost	Won	Lost
PK	Sequence	71	129						
	Goal	8	12						
	Rate	11%	9%						
Even	Sequence	292	283	333	378	333	378	245	390
	Goal	19	19	16	22	15	18	13	24
	Rate	7%	7%	5%	6%	5%	5%	5%	6%
PP	Sequence							110	93
	Goal							16	10
	Rate							15%	11%

Breaking down sequences to the second

We now restrict our attention to offensive and defensive zone faceoffs; focusing on goals for in the former case and goals against in the latter. So far, we have gathered *some* evidence that winning these increases the chances of scoring a goal. We know from experience that winning an offensive zone faceoff often leads to a quick shot from the blue line, or more generally to a scoring chance of some sort. Therefore, our intuition tells us that the effect should be most prominent in the very first seconds following the faceoff and that it should become less and less relevant as time goes by. To verify that using the Erie dataset, we analyzed the moment at which the goals following offensive zone faceoffs were scored. Although we do have the exact goal times (down to the second), we preferred to bin them over windows of 5 seconds for now, as it cancels out some of the noise in the data.

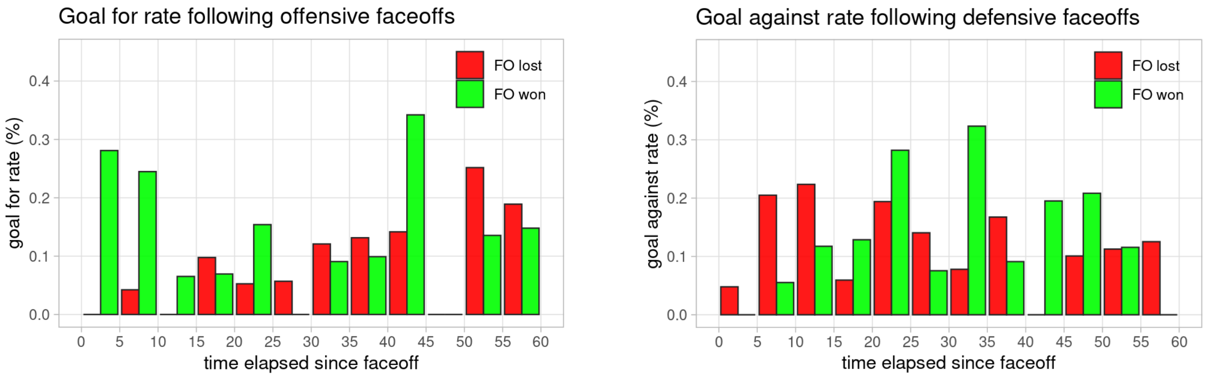

Figure 1: Percentage of times a goal was scored within the time window $[t, t+5)$ following the faceoff, **given that the sequence in question lasted at least** t **seconds**. The left panel is based on goals for Erie following offensive zone faceoffs, while the right panel is based on goals against Erie following defensive zone faceoffs. The data was collected during the 2019-20 season.

Figure 1 is encouraging: it suggests that winning offensive zone faceoffs—or losing defensive ones—leads more often to a *quick goal*, where by *quick goal* we mean within the first five-ten seconds of the sequence. It also suggests that the impact of winning offensive zone faceoffs vanishes after approximately 20 seconds. However, there are some peculiarities to the data, like the fact that the bin $[45, 50)$ of the offensive panel shows no goal while its neighbors are significantly different from zero. Such anomalies are more likely to happen in later bins, as the number of sequences that reach 45 seconds, say, is much less that those reaching 10 seconds. While interpreting Figure 1, it is important to note that it provides conditional probabilities, and therefore that the numbers shown are theoretically invariant to the performance on the faceoffs themselves; although losing more faceoffs, for example, necessarily implies more data for estimating the quantities of interest for this particular situation.

The noise in the bar charts of Figure 1 caused by the lack of data points makes it a bit hard to interpret them. To overcome such difficulty, we constructed a smooth version of these latter using a generalized additive model (*gam*). We relegate the more formal description of the underlying model to the Technical details section at the end of the report. Basically, the goal was to construct a curve that provides the *risk* of a goal being scored at any time t, conditional on the outcome of the faceoff. This *risk* can be interpreted as the probability of a goal being scored within a one-second time frame. To do so, we first expanded the data so that each row of the new data corresponds to a one-second window of time. The resulting curves are shown in Figure 2.

Unsurprisingly, Figure 2 shows nice bumps right after the faceoff when this latter is won by the offensive team. Its continuous nature makes it easier to compare the offensive and defensive performances of Erie. For example, after a defensive zone faceoff lost, they seem to be able to limit their opponents' ability to score. We again note that the advantage provided by winning an offensive zone faceoff seems to vanish approximately 20 seconds into the sequence. It also seems worthy to point out that, for a short period of time around 5 seconds

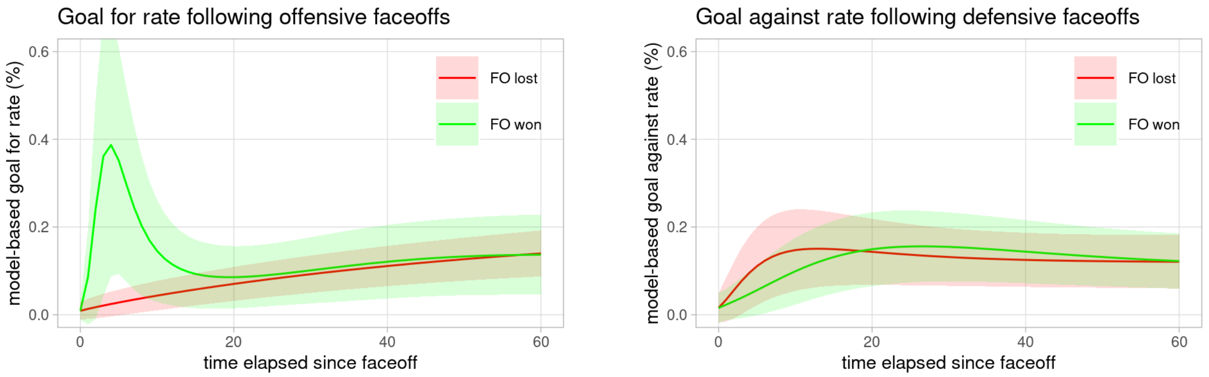

Figure 2: Model-based Erie's goal rates for time windows of the form $[t, t+1)$, **given that the sequence in question lasted at least t seconds.** The left panel covers goals for following offensive zone faceoffs, while the right panel covers goals against following defensive zone faceoffs. The data was collected during the 2019-20 season.

Table 4: High-level features of the NHL dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	Goals	OZ/DZ Goals
2019-10-02	2020-03-12	628	123402	42940	6433	2960

after the faceoff, the 95% confidence intervals for the two curves corresponding to offensive zone faceoffs do not overlap. This contrasts with previous results, where absolutely no statistical significance was found.

Going bigger

In view of our analysis of the Erie dataset, it seems more than reasonable to think that winning offensive faceoffs does increase one's chances of scoring a goal. In order to convince ourselves that our intuition is nevertheless backed by data, we performed a similar analysis using the 2019-20 NHL data.[1] We consider this data to be balanced, as all teams appear in it roughly in the same proportion. Table 4 provides a basic summary of this dataset.

The overall impact of winning a faceoff. As shown in Table 4, the dataset contains 42940 faceoffs that lead to 2960 goals. As we did in the first section of the report, we fitted a logistic regression to this data. We only fitted the more complex model with an intercept, `FO_win` and `FO_OZ` as covariates, an interaction between those two covariates, and the variable `GF` as response. We obtained $\beta_{FO\_win} = 0.138$, $\beta_{FO\_OZ} = 0.559$ and $\beta_{FO\_win+FO\_OZ} = 0.849$. This time, due to the much larger dataset, all the p-values but one, that of β_{FO_win}, were statistically significant. The conclusion reached in the first section is still valid, there is nothing better than winning an offensive faceoff to increase the probability of scoring a goal. Furthermore, we can now say that getting an offensive faceoff—win it or lose it—is more valuable than winning a faceoff anywhere else on the ice.

Breaking down sequences to the second. We again began the more granular analysis by constructing a bar chart showing the proportions of goals for in a given time window, conditional on the outcome of the faceoff. This time, the huge amount of data allowed us to use smaller bins of two seconds. Results are displayed on the left panel of Figure 3. We then fitted our *gam* model to get a smooth version of the bar chart; these results are displayed in the right panel of Figure 3.

From both the bar chart and curve in Figure 3, we can better see the effect of winning offensive zone faceoffs,

[1] We fetched the data using `tidynhl`, an `R` package developed on top of the NHL Stats and Records APIs.

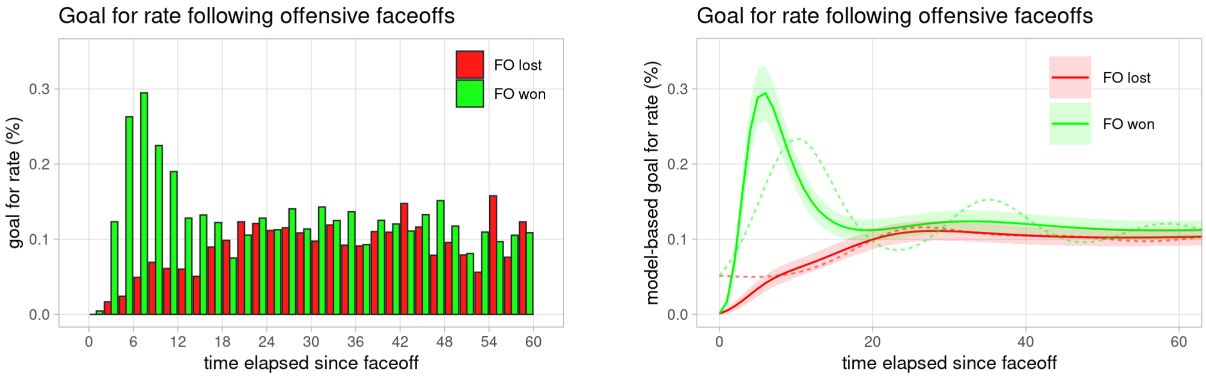

Figure 3: Left panel: Percentage of times a goal was scored within the time window $[t, t+2)$ following offensive zone faceoffs. Right panel: Model-based goal for rates for time windows of the form $[t, t+1)$. All quantities are **conditional on the fact that the sequence in question lasted at least t seconds.** The results are averaged over the data from all NHL teams collected during the 2019-20 season. The dashed line on the right panel shows the results from an alternative model discussed in the Technical details section.

especially in the first 20 seconds. We can also see that the confidence intervals are now much smaller compared to those in Figure 2, and that there is no overlap between them in the first 20 seconds. Beyond that 20 seconds mark, we see that both curves seem to merge together. That suggests that the effect of winning offensive faceoffs does not provide an advantage that lasts forever, as we expected.

As a final analysis, we asked ourselves whether the combined effect of winning (losing) offensive (defensive) zone faceoffs is similar across all teams. To do so, we looked at the team-wise proportion of goals for (against) following won (lost) offensive (defensive) zone faceoffs. Motivated by the shape of the bumps in our *gam* curves, we restricted our attention to the first 20 seconds. For each team, these two proportions were then plotted in an x-y plane; see the left panel of Figure 4. The sweet spot on this graph is located in the bottom right corner, which imply a high proportion of goals for and low proportion of goals against. As one can see, the Buffalo Sabres are among the worst teams according to these metrics and the Boston Bruins seems to be the clear winners. Their associated *gam*-based curves are shown on the right panel of Figure 4. Interestingly, it seems that Buffalo takes little to no advantage of their offensive zone faceoff wins.

Wrap up. It is fair to say that most hockey fans could have guessed that winning offensive zone faceoffs increases the chances of scoring a goal in the sequence that follows. Our objective was to quantify how much, and we believe that the main finding of this study is that offensive zone faceoffs create a 20 seconds window of opportunity that must not be neglected. Studying this effect in more contextualized situations like power plays, penalty kills and overtimes, seem very promising to us, as they might reveal very different patterns.

Technical details

Figure 2 and 3 For convenience, we focus on offensive faceoffs here. Denote $Y_t \in \{0, 1\}$ the random variable indicating whether a goal for occurred in the interval of time $[t, t+1)$, and let $\mathbb{E}(Y_t|x)$ be the expectation of Y_t (i.e., the probability of a goal occurring in that timespan) given that the most recent faceoff was lost $(x = 0)$ or won $(x = 1)$. The curves in Figure 2 were obtained by fitting the generalized additive model (*gam*)

$$g\{\mathbb{E}(Y_t|x)\} = \beta + (1-x)f_0(t^*) + xf_1(t^*), \qquad g(z) = \ln\{z/(1-z)\}, \tag{1}$$

where g is the so-called logit function and, for both $k = 0, 1$, f_k is an unknown (i.e., to be estimated) nonlinear function of $t^* = h(t)$ that approximates $g(Y_t|x = k)$. To fit *gams*, we used the function `gam` of the `R` package

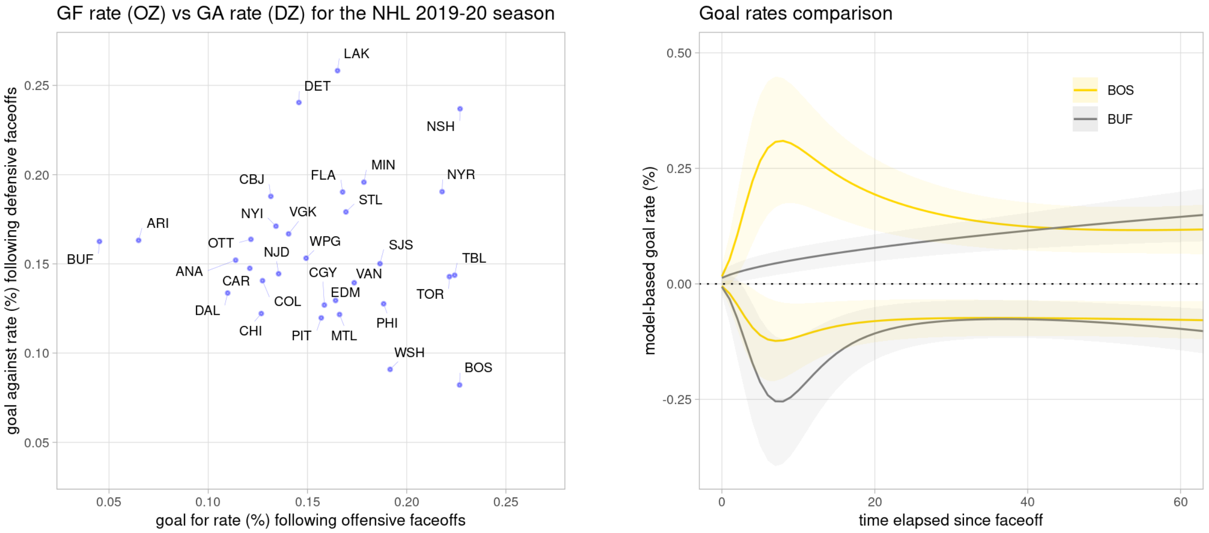

Figure 4: Left panel: The x-axis shows the goal for rate following won offensive zone faceoffs and the y-axis shows the goal against rate following lost defensive zone faceoffs. Right panel: Model-based goal rates for Boston Bruins and the Buffalo Sabres. The upper curves show their performance after won offensive zone faceoffs, while the lower curves show their performance after lost defensive zone faceoffs. The data was collected during the 2019-20 season.

`mgcv`[2]. As explained in the main text, we used $t^* = \log(t+1)$ to generate Figure 2, so as to favor less smoothing near $t = 0$, where we expect a more rapidly changing curve, and more smoothing for larger values of t, for which less data is available. By design, this helps obtaining a (nearly) null risk that a goal occurs at $t = 0$, which is obviously what one would expect. The probabilities reported are obtained by solving this (1) for $\mathbb{E}(Y_{t^*}|x)$ at each value of $t \in \{0, 1, 2, \dots\}$, using the fact that $t^* = h(t)$ for some function h. Two models, one using $t^* = log(t+1)$ and another using $t^* = t$ are shown in Figure 3. The latter model, which was our initial guess, turned out to give a poor fit.

[2] More details. We use the `R` function `s` to estimate the functions f_k, $k = 0, 1$; in doing so, we specify the option `pc=0` to force $f_k(0) = 0$ for both $k = 0, 1$; see our GitHub repository for the exact code. The package `mgcv` implements a Bayesian approach.