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. In other words, in leveraging the data at our disposal, we will try to answer the following question:

What is at stake when the linesman drop the puck?

It is reasonable to assume that the answer may vary significantly depending on the context in which a particular faceoff occurs. One could possibly come with some intuitive factors that may have an impact on the criticalness of winning a given faceoff. A non-exhaustive list may include elements such as the current score, the time remaining on the clock, the zone in which the faceoff is taken, or the strength of play (e.g. power play). Others will argue that — even though there is not a single coach who overlooks the importance of winning faceoffs — some game plans are merely better to make the most of an offensive zone faceoff win while others are more effective to salvage a loss in the defensive zone.

The first step in answering such a 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.

To this end, we made the decision to only focus on the scouting dataset since we wanted to have relatively homogeneous data. Indeed, mixing different leagues may have had introduced noise coming from the different realities among those leagues. Furthermore, since every games in the data highlight the Erie Otters, we will be creating models in the perspective of this particular team. It would be somewhat inappropriate to build a model with such asymetrical data and claim that it represents the reality for the whole OHL.

With these in mind, let's give a first look at the data.

Table 1: High-level features of the scouting dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	Goals	OZ/DZ Goals
2019-09-20	2020-03-08	40	2441	1644	293	122

As shown in Table 1, the dataset contains 40 games of data, with 2441 overall faceoffs. However, only 1644 faceoffs are relevant to our analysis since we restrict ourselves to sequences starting in offensive/defensive zones. The dataset also contains a total of 293 goals scored, but only 122 were actually scored by the team that started a sequence from their offensive zone. Je montrerais les données déjà selon la perspective de Erie.

Organization of the report. The report is divided in three main parts. In the first part, we begin by analyzing the data sequence-by-sequence, that is, with one row per (contiguous) sequence of play. That way, we can first verify our intuition that winning an offensive zone faceoff increases the chances of scoring a goal (on that given sequence). In the second part, once it is confirmed that there is indeed some signal in the data, we perform a more in-depth analysis by looking at the data on a more granular time-scale (second-by-second

at some point). Unfortunately, it then seems irrelevant to further precise the context in which the faceoffs are taken as the data become too sparse. Finally, we investigate whether more robust conclusions could be drawn from a bigger dataset constituted of multiple years of NHL data, which we fetched using `tidynhl`, an R package developed on top of the NHL Open API.

The overall impact of winning a faceoff

At this point, it seems important to verify that winning the faceoff does indeed provide an advantage. For convenience, and because it seems the most interesting case to us, we restrict ourselves to Erie Otters' offensive faceoffs. To verify our intuition, we simply compare proportion of contiguous sequences (beginning with an offensive faceoff) that lead to an Erie Otters goals depending on the outcome of the faceoff. Before we can do that, we have to structure the data so that each row corresponds to a contiguous sequence. Table 2 provides a excerpt of the transformed data, which we refer to as sequence-by-sequence data.

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

game_date	period	clock_begin	clock_end	length_seconds	FO_win	FO_zone	GF	GA
2019-09-20	1	20:00	18:57	63	FALSE	red	FALSE	FALSE
2019-09-20	1	18:57	18:29	28	FALSE	offense	FALSE	FALSE
2019-09-20	1	18:29	15:27	182	FALSE	offense	FALSE	FALSE
2019-09-20	1	15:27	14:10	77	FALSE	offense	FALSE	FALSE
2019-09-20	1	14:10	13:42	28	FALSE	blue_offense	TRUE	FALSE
2019-09-20	1	13:42	12:41	61	TRUE	red	FALSE	FALSE

To get a crude idea of the impact in goal scoring of winning/loosing an offensive faceoff, we fitted a logistic regression with `GF` as response, `FO_win` as covariate and an intercept. In other words, we make use of the logistic function (log of the odds ratio) to model the probability that the Erie Otters score a goal conditional on the value of `FO_win` (either `FALSE` or `TRUE`). We obtain $\beta_{\text{FO\_win}} = 0.171$, which suggests that winning an offensive faceoff does increase the chances to score a goal during the sequence that ensues. However, the associated p-value is 'r round(summary$coefficients[2,][4], 3), and so there does not seem to have enough evidence to confidently make the latter statement.

In Table
reftab:preview, we omitted some columns that bring more context to the faceoff. The most obvious is the information about power play and penalty kill situations. In the table
reftab:features2, we did breakdown the sequences by context and we also added the scoring success rate in each of these situations.

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%

JAI PAS CHECK PANTOUTE. A first conclusion we can draw from table reftab:features2 is that we don't have a lot of goals scored by Erie Otters (63) from sequences that started in their offensive zone. We will see later if it's sufficient to draw significant conclusions. However, we strongly doubt that it will be sufficient for drawing conclusions on a more granular basis, such as power play or penalty killing situations. Another thing we can notice from the table reftab:features2 is the success rate differences between sequences that started with a won or a lost faceoff. From the data we have, we can already see the effect of winning the faceoff on power play situations. For even strength contexts, the impact looks less significant (at least for Erie Otters).

Breaking down the sequences by seconds

So far, we have gathered *some* evidence that winning offensive faceoffs increases the chances of scoring a goal. Our intuition also tells us that this effect should be most prominent in the very first seconds following the faceoff and that, over time, the fact that one has won or lost the faceoff should become less relevant. As a matter of fact, the team that wins an offensive faceoff might get a quick shot from the point or set up a play that leads to an immediate scoring chance. To verify that, let us take a closer look at the times the Erie goals in question were scored; for this analysis, we also look at the goals against Erie following a defensive faceoff. Although we do have the exact goal times (down to the second), we prefer to bin them over windows of 5 seconds for now, as it cancels out some of the noise in the data.

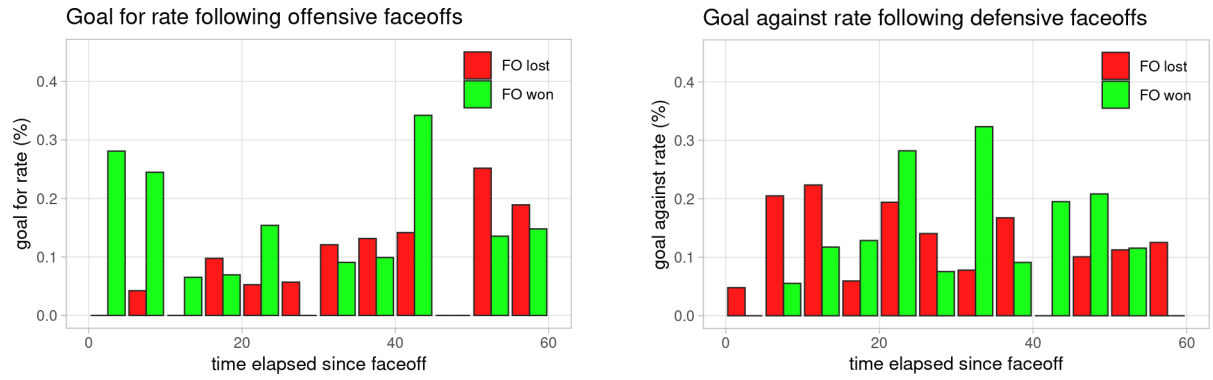

Figure 1: Percentage of times a goal was scored within the timespan $[t, t+5)$ for $t = 0, 5, 10, ..., 175$, **given that the sequence in question lasted at least t seconds**.

Figure 1 is encouraging: it suggests that winning an offensive faceoff (or losing a defensive faceoff) 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 the faceoff in offense (green bars in the left panel, and red bars in the right panel) vanishes after approximately 20 seconds. However, note that no bar actually correspond to more than ZZZ goals, and that 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 reaches 45 seconds, say, is much less that those reaching 10 seconds (XX vs YY in this particular case). While interpreting Figure 1, it's 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 barcharts of Figure 1 caused by the lack of data points makes it a bit hard to interpret them. To overcome such difficulty, we now construct 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, we wish 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 expand the sequence-by-sequence data so that each row of the new data corresponds to a one-second window of time, much like we have done for the barcharts, but with bins of the form $[t, t+1)$ this time. The resulting curves are shown in Figure 4.

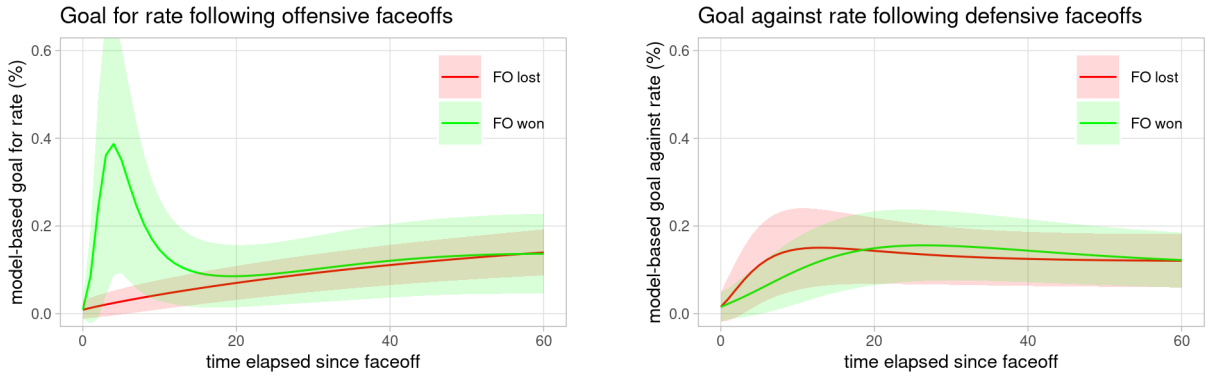

Figure 2: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds. le nom du Y-axis me gosse

Since Figure 4 pretty much conveys the same information as Figure 1, it is not surprising that is shows nice bumps right after the faceoff when this latter is won by the offensive team (Erie in the left panel, and its opponent in the right panel). Its continuous nature makes it easier to compare the offensive and defensive performances of Erie. For example, they seem to be able to limit their opponents' ability to score when they lose the faceoff in defensive zone (right panel, red curve), as their opponents' performance in similar situation (left panel, green curve) shows a much bigger bump, which corresponds to goals for Erie. We again note that when the advantage provided by winning an offensive faceoff seems to vanish around 20 seconds into the sequence. It also seems that when Erie is in defense, winning the faceoff is almost disadvantageous. This is most probably due to the lack of data, and the variability that ensues. In fact, it must be said that the results we have so far are not "significant", in the statistical sense. The pointwise confidence intervals associated to all four curves, also reported in 4, are very large, and so any conclusion drawn from them or from the barcharts in Figure 1 must be taken with a grain of salt. Furthermore, the width of these confidence intervals strongly suggests that performing a similar analysis with a more granular definition of "situation", e.g. distinguishing even strength, powerplay and penalty kill sequences, is probably not a good idea.

Adding more data

In view of our analysis of the Erie dataset, it is reasonable to think that winning offensive faceoffs does provide an advantage, but the evidence so far presented falls just short of legitimate statistical significance. We strongly believe that this is due to a lack of data points caused in part by the restriction to offensive faceoffs. Looking back at the fact that the various probabilities estimated were no greater than $0.5\% = 0.005$, it would have been surprising to obtain statistically significant results with the 1644 offensive faceoffs of the Erie dataset.

In order to convince ourselves that our intuition is nevertheless backed by data, we now perform a similar analysis using the 2019-2020 NHL data.[1] We consider this data to be balanced, as all teams appear in it roughly the same amount of times. Table 4 provides a basic summary of this NHL dataset.

Table 4: High-level features of the NHL dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	Goals	OZ/DZ Goals
2018-10-03	2020-03-12	1378	273148	93823	13996	6382

It seems worthwhile pointing out that we now work with 93823 faceoffs that lead to 6382 goals. As we did in the first section of the report, we first fit a logistic regression with an intercept, `FO_win` as covariate, and the variable `GF` as response. This time, we obtained a $\beta_{FO\_win} = 0.303$ with a p-value of $5.8808827 \times 10^{-31}$. There is thus plenty of evidence that winning an offensive faceoff significantly increases the odds of scoring a goal during the sequences that follow.

We again begin the more granular analysis by constructing a barchart showing the proportion of goals (for) scored in a given time window, conditional on the outcome of the faceoff. The huge amount of data now allow us to use smaller bins of two seconds, that is $[t, t+2)$. The result is displayed on the right panel of Figure 3.

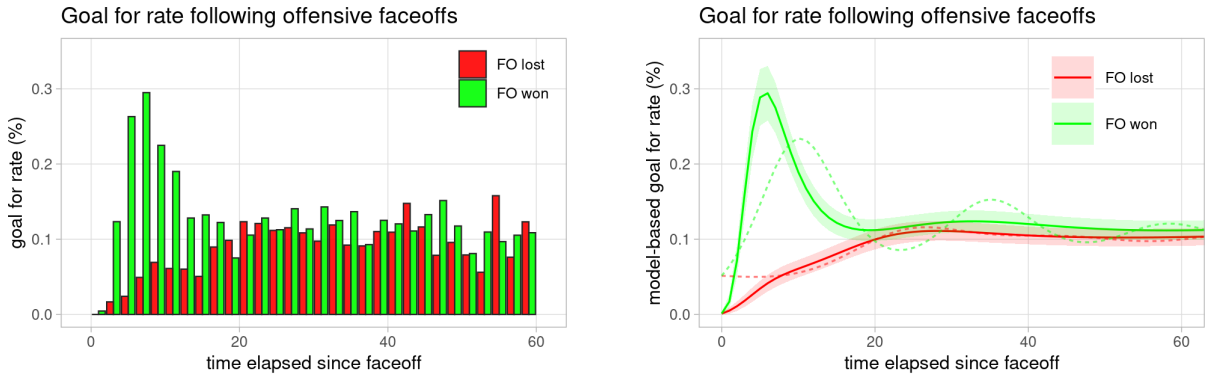

Figure 3: Percentage of times a goal was scored within the timespan $[t, t+2)$ for $t = 0, 2, 4, ..., 60$, **given that the sequence in question lasted at least t seconds**.

From the Figure 3, we can see better the effect of winning an offensive zone faceoff, especially within 20 seconds after a faceoff. Let us construct the smoother version of this bar chart using our gam model (see Figure 3). With this dataset, it is now more clear compared to what we had in the Figure 4.

In the Figure 3, we can see again the bump when winning the faceoff in the moments that follows. When can also see that the confidence intervals are now much more smaller (compare to Figure 4), and do not overlap between each others in the first 20 seconds (approximately). Then, we are able to say that the difference is significant for this time. Beyond that 20 seconds mark, we see that both curves converges and are not significantly different. At some point, we can conclude that winning a faceoff does not provide an advantage that lasts forever.

[1] We fetched the data using `tidynhl`, an `R` package developed on top of the NHL Open API.

Now that we understand the effect of winning offensive faceoffs, we can ask ourselves, is this effect the same among all teams? To compare the different teams against each others, we focused on the so called "bump" that we see after winning (or loosing) a faceoff. To take into account both the offense and the defense, we compare the heights of both bumps (offensive and defensive) for all teams. The offensive bump can be interpreted as the ability to score a goal after a won offensive faceoff while the defensive bump can be interpreted as the ability to contain the opponent after a lost faceoff. In the Figure XX, we can see the offense on the x axis, and the defense on the y axis. In the best scenario, a team would prefer to have a high bump on offense and a small bump on defense (which corresponds to the bottom right corner).

FIGURE XXX

ANALYSE DE LA FIGURE

Let's wrap this up It's fair to say that most of the people watched enough hockey games could have guessed that winning an offensive zone faceoff increases the chances to score a goal in the sequence that follows. However, this analysis allowed us to conclude few interesting things:

- with enough data, we can conclude that the increase is statistically significant
- the effect of winning a faceoff does not lasts forever (about 20 seconds)
- COMPARAISON DES ÉQUIPES (CONCLUSION)

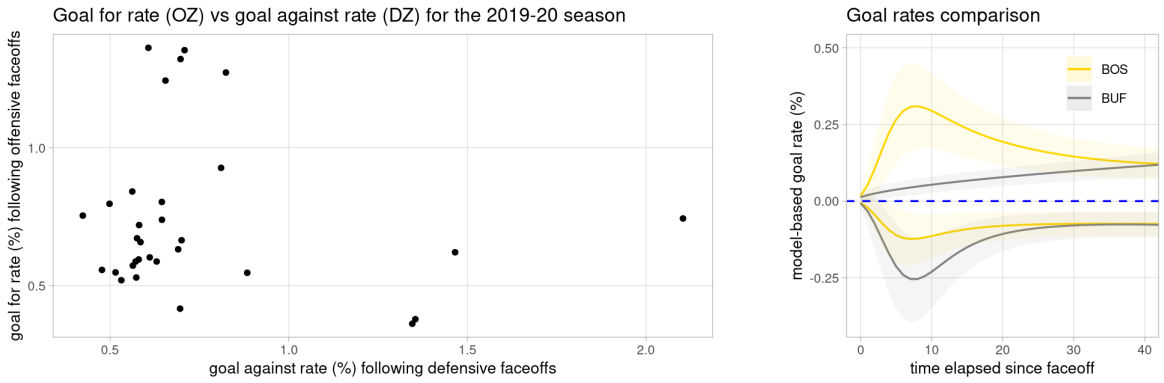

Figure 4: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds. le nom du Y-axis me gosse

We were able to breakdown the faceoffs in different situations, but we think it would also be interesting to explore or push further other situations:

- compare the effect of winning an offensive zone faceoff in power play compare to even strength
- AUTRE OUVERTURE

Technical details

Figure 4 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 4 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 *gam*'s, we use the function `gam` of the `R` package `mgcv`[2]. As explained in the main text, we use $t^* = \log(t+1)$ to generate Figure 4, 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.

[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 GITHUB LINK for the exact code. The package `mgcv` implements a Bayesian approach.