How bad should I win next faceoff?

Big data cup 2021

During the course of a typical hockey game, between 50 and 75 faceoffs take place. The outcome of such events often dictates which team controls the puck at the beginning of the sequence. This, in itself, strongly suggests that (as the old saying goes) "every faceoff is important." However, we also often hear that some faceoffs are "particularly important", perhaps more important than your average faceoff. A coach would say that, for example, just before an offensive faceoff when his/her team is trailing by one goal with less than a minute on the clock. Naturally, this faceoff is 'particularly important' because the team absolutely needs a goal, and a quick one. Loosing the faceoff is certainly not going to help achieve that. In this report, we move away from this definition of "importance", which involved both the time at which the faceoff is taken and the current score. Instead we study the relationship between winning/loosing a faceoff and the chances that a goal occurs. In this sense, we judge the importance of a faceoff by the difference winning or losing it implies on the chances that a goal (for or against, depending on the context) occurs. In other words, we try to answer the question:

What is at stake during an offensive zone faceoffs?

To this end, we decided to use the Erie Otters data of the scounting dataset. We chose this dataset for the simple reason that it involved the Erie Otters in each and every game, thus making it more "balanced' '. Indeed, in the other datasets, the teams involved do not appear the same number of times (or close to), and so this could be a source of complications when it is time to interpret the results. For simplicity, we define the following terms that we use throughout this analysis.

- FO: Faceoff
- OZ: Erie Otters' offensive zone
- DZ: Erie Otters' defensive zone
- FO_win: faceoff won by Erie Otters
- GF: goal gor Erie Otters
- GA: goal against Erie Otters

With these in mind, we are ready to take a first look at the chosen dataset.

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.

What is 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

	Penalty kill			Even strength			Powerplay		
Zone	FOs	Goals	% success	FOs	Goals	% success	FOs	Goals	% success
Defensive									
FO lost	129	12	9.3 %	283	19	6.7 %	7	0	0 %
FO won	71	8	11.3 %	292	19	6.5 %	10	1	10 %
Offensive									
FO lost	10	0	0 %	390	24	6.2 %	93	10	10.8 %
FO won	4	0	0 %	245	13	5.3 %	110	16	14.5 %

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. EXPLAIN WHY To verify that, let us take a closer look at the times the goals in question were scored. Let us also consider the defensive side of the game. 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.

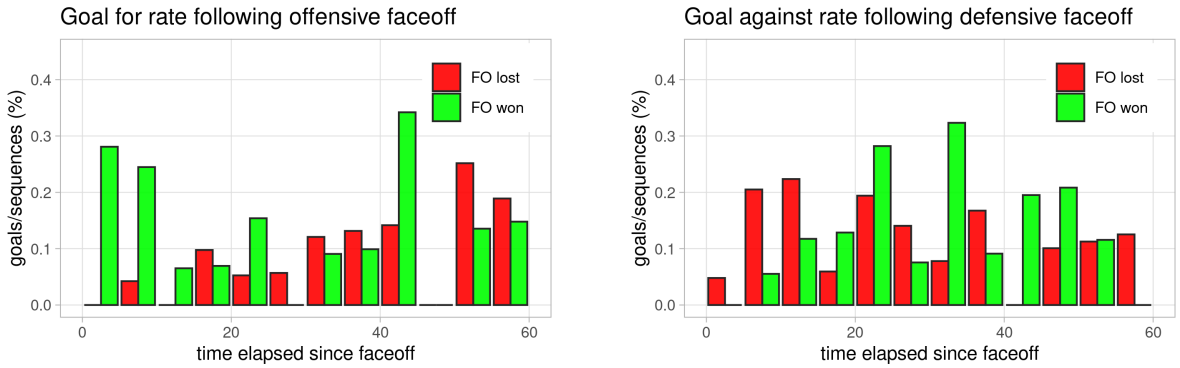

Figure 1: Percentage of times a goal was scored within the timespan $[t, t+5)$ for $t = 0, 5, 10, ..., 175$, **given that the sequence in question lasted at least t seconds**.

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. However, note that no bars 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 bins corresponding to longer durations, as the number of sequences that reaches 45 seconds, say, is much less that those reaching 10 seconds. 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 (alhtough losing more faceoffs, say, means that more data is available for estimating the corresponding case).

The sparseness in the barcharts of Figure 1 caused by the lack of data points makes a bit it hard to interpret them. To overcome such difficulty, we now construct a smooth version of these latter using a generalized additive model (*gam*) and the exact seconds at which the goals were scored; see the **Technical details** section at the end of the report for a more formal description of the underlying model.

The resulting curves, shown in Figure 2, provide not the probability, but the *risk* of a goal being scored at any time t, conditional on the outcome of the faceoff. However, the risk can be interpreted as the probability of a goal being scored within a one-second time frame. Again, we see a nice bump right after the faceoff when it is won by the offensive team. It is also easier to compare the offensive and defensive performances of Erie Otters: they seem to do a better job defensive job than their opponent in the event that the offensive team won the faceoff, but the inverse seems true when the defensive team wins the faceoff. It must be said that the results are not "significant", in the statistical sense. The confidence intervals corresponding to all four curves, also reported in 2, are very large. This strongly suggests that performing this at a more granular definition of "situation", e.g. distinguishing even strengths, powerplay and penalty kill sequences, might not lead to interesting results.

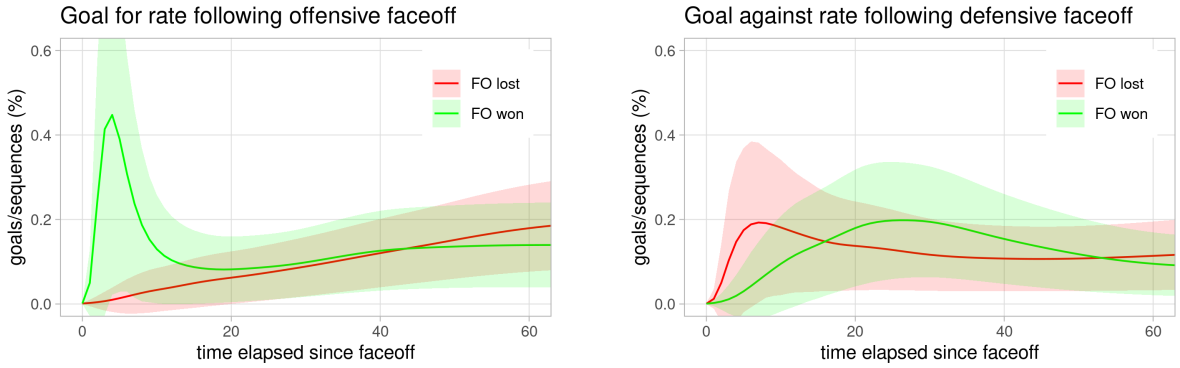

Figure 2: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds. le nom du Y-axis me gosse

Technical details

Figure~(LOESS) Denote $Y_t \in \{0, 1\}$ the random variable indicating whether a goal 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 loess curves in Figure~(LOESS) were obtained by fitting the *generalized additive model* (gam)

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

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 that approximates $g(Y_t|x = k)$.[1] The probabilities reported are obtained by solving this equation for $\mathbb{E}(Y_t|x)$ at each value of $t \in \{0, 1, 2, \ldots\}$.

Figures~(SPLINES-1) and (SPLINES-2). In Figure~(SPLINES-1), we reproduced the analysis involving the model in (1). This time, however, we used splines (DEF) for approximating the functions f_0 and f_1 of the gam (as opposed to the loess method previously used), which allowed us to construct confidence intervals. The more granular results displayed in Figure~(SPLINES-2) were also obtained by means of a splines-based gam. In this case, we fitted a model that included $2 \times ZZZ = ZZZ$ nonlinear functions of t. In addition to the subscripts $k \in \{0, 1\}$ indicating whether the faceoff was lost or won, we use $\ell \in \{1, \ldots, ZZZ\}$ to refer to each of the **ZZZ** situations of interest (NAME THEM). The resulting model is given by

$$g\{\mathbb{E}(Y_t|x, s)\} = \sum_{\ell=1}^{Z} \mathbb{1}(s = \ell) \times \Big\{(1-x)f_{0\ell}(t) + xf_{1\ell}(t)\Big\}, \tag{2}$$

where g is as in (1). Note that (2) could actually be expressed and fit as ZZZ distinct models. However, this formulation makes it clear how to include further covariates that are known to have a similar effect in multiple situations.

Adding more data

Using the Erie Otters data available in the `scouting` dataset, we can see that winning offensive zone faceoffs has an impact, but we can't say it's statistically significant at this point. The reason we suspect is lack of data. To convinced ourselves, we redone the same kind of analysis, but using the last 2 seasons of NHL data. In that case, we used the data of all teams, as it is more equally spread across all teams. In the table 4), we can see some basic informations about that NHL dataset.

[1] To fit this model, as well as all models discussed in this report, we used the `R` package `gam`. Also note that in this particular case, we actually used $t^* = \log(t+1)$ as the time variable, so as to allow less smoothing near $t = 0$, where the observations are more concentrated. We also gave considerably more weight to the observations with timestamp $t = 0$ to force $f_k(0) \approx 0$ $(k = 0, 1)$.

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

Now we have much more faceoffs, and also much more goals as well. In the same way as we did in the first section of this report, we fitted a logistic regression using the variable `FO_win` as our only feature, and the variable `GF` as a target. Using this data, we obtained a $\beta_{FO\_win} = 0.303$ with a p-value of $5.8808827 \times 10^{-31}$. Now that we have more observations, we can conclude that winning an offensive zone faceoff does increase significantly the odds to score a goal during a given sequence.

On sequence-by-sequence basis, we now know that the effect is significant. Considering that, let's recreate the breakdown by seconds as we did in the previous section to see how this effect last over time. As a first step, we can look at the goals scored in time, for which we decided to use bins of 2 seconds here (as we have more data).

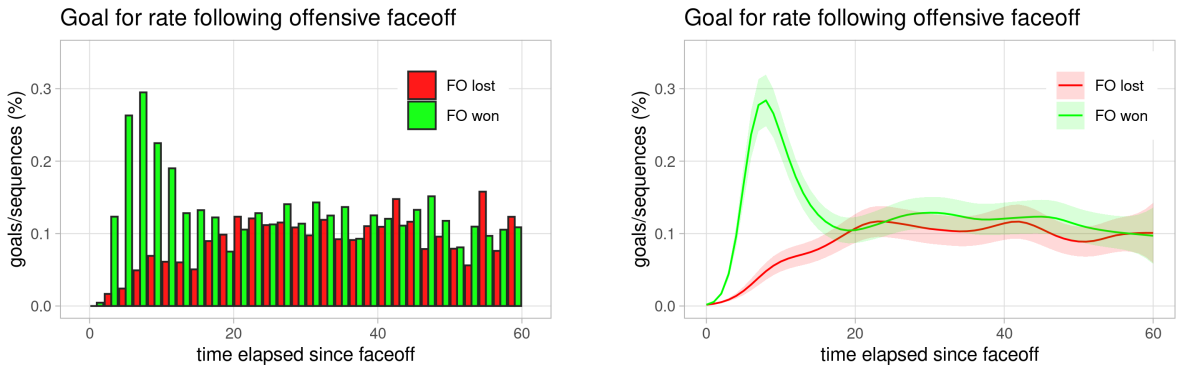

Figure 3: Percentage of times a goal was scored within the timespan $[t, t+2)$ for $t = 0, 2, 4, ..., 60$, **given that the sequence in question lasted at least t seconds**.

From the figure 3, we can already see that the effect of winning an offensive zone faceoff is more obvious within 20 seconds after a faceoff. With this dataset, it is now more clear compared to what we had in the figure 1. Let's construct the smoother version of this bar chart using our gam model (see figure 4).

«««< HEAD

In the figure 4, we can clearly see the spike 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 2), and do not overlap between each others (at least in the first 20 seconds approximately). It also interesting to see that both curves converges 20 secondes and more after the faceoff. At some point, we can conclude that winning a faceoff does not provide an advantage that lasts forever.

AJOUTER UNE AUTRE ANALYSE (OT ou COMPARE TEAMS)

Let's wrap this up It's fair to say that most of the people that knows enough about hockey 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)
- CONCLUSION OT or TEAMS comparisons

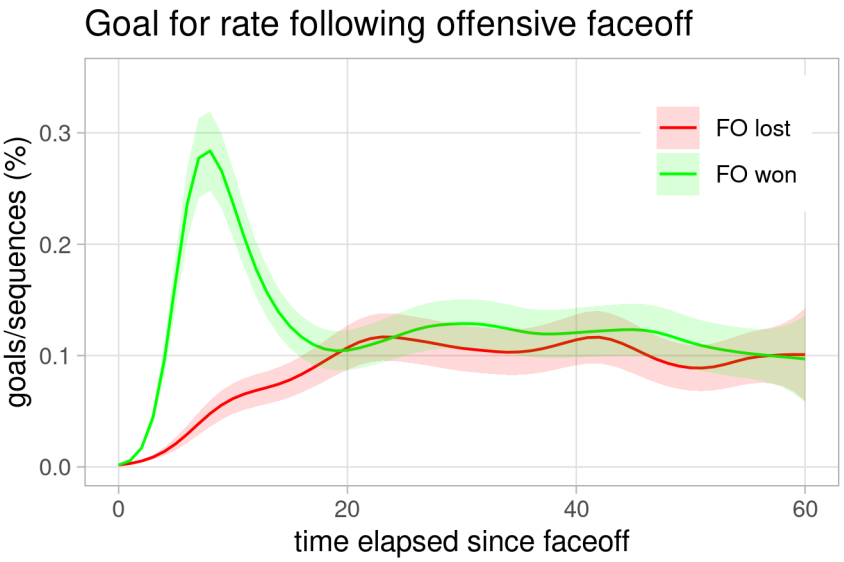

Figure 4: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 60$, given that the sequence in question lasted at least t seconds.

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
- compare different teams together and see how the impact varies between them