How bad should I win next faceoff?

Big data cup 2021

During the course of a hockey game, there are a lot of faceoffs. The outcome of these events usually dictates which team will control the puck at the beginning of a sequence. Thereby, we can say that every faceoff are important. However, many would agree to say that we can attribure even more importance to some faceoffs happening during a game, synonym of key moments. As an example, a faceoff in the offensive zone, trailing by one goal with less than a minute to play would appears to be quite important. In this analysis, we will try to quantify the importance of winning faceoffs. Ultimately, we would like to be able to answer to these kind of questions:

> **To what extend is winning offensive zone faceoffs increases the chances to score a goal? How long after the faceoff we can notice the effect of winning or loosing it? Is winning offensive zone faceoffs even more important in man advantage situations?**

To be honest, we will start by confirming if winning a faceoff in the offensive zone is increasing the chances to score a goal during that sequence. Once that will be confirmed, we will go more in-depth in understanding the importance of those events.

As you may noticed, we targeted offensive zone faceoffs. The reason for that is we think their impact on scoring a goal is more obvious as the team is already close to the opponent net. We could also say we included defensive zone faceoffs, as it only depends from which team point of view you are looking at it. To answer the central question, we decided to use the Erie Otters data made available in the scounting dataset. Our rationale for using this dataset was to use the data of one single team, and not bother having teams that played different amount of games. That way, we think it simplifies the interpretations and our capacity to draw conclusions. In that context, the `scounting` dataset is the one with the most observations. Because we consider one single team in our analysis, the Erie Otters in this case, that means we also consider some events stricly from Erie Otters point of view. For example, we can define these abbrevations that we will reuse throughout this analysis:

- FO: Faceoff
- OZ: Erie Otters Offensive Zone
- DZ: Erie Otters Defensive Zone
- FO_win: Faceoff won by Erie Otters
- GF: Goal For Erie Otters
- GA: Goal Against Erie Otters

In the next sections, we are going to first look at the data sequence by sequence. In this manner, we will be able to conclude if winning an offensive zone faceoff does have a significant impact of scoring a goal on that given sequence. Afterward, we will breakdown our sequences by seconds and see the impact of winning offensive zone faceoffs over time. We will also see if we can draw conclusions from different contexts, such as power play or penalty killing situations. Finally we will see if we can draw more robust conclusions over a bigger dataset, such as using NHL data (API link maybe?).

0.1 Sequence-by-sequence analysis

As a first step, let's have a better understanding of the data we are working with. As mentionned in the introduction, we focused on the `scouting` dataset. In the table 1, we can see basic informations about the this dataset, and what might be relevant to know for our analysis.

Table 1: High-level features of the scouting dataset

Min. game date	Max. game date	Games	FOs	OZ/DZ FOs	GF	GA
2019-09-20	2020-03-08	40	2441	1644	148	145

We have 40 games of data, with 2441 overall faceoffs. However, only 1644 faceoffs will be relevant in our case since we focus on those that happened in the offensive/defensive zones. We have 293 combined scored goals, but only 206 where actually scored from a sequence that started in the offensive/defensive zones.

Talking about sequences, we had to structure the data in a way that we can easily analyze each sequences individually. To illustrate that, we added a preview (see table 2 of how our transformed data looks like on a per-sequence level.

Table 2: Preview of our 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

In the table 2, we omitted to show some additionnal columns that bring contextual informations around the faceoff. The more obvious example of that is power player and penalty killing situations. In the table 3, we added more context and we also try to breakdown the scoring success depending the situation.

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 %

If we take a moment to analyze the table 3, we can see that we don't have a lot of goals scored by Erie Otters for sequences that started in the offensive zone (63 goals) or goals against Erie for sequences that started in the defensive zone (59 goals). Still, that might be sufficient for making conclusions on per-sequence basis. However, we strongly doubt that it will be sufficient for drawing conclusions on a more granular basis.

Another thing we can notice from 3 is the success rate differences between won and lost faceoffs. From the data we have, we can already see the effect of winning the faceoff on man advantage situations. For even strength situations, the impact looks less significant.

To see if the effect of winning the faceoff in the offensive zone is significant, we fitted a logistic regression for which we had the variable `GF` as a target, and the only variable `FO_win` as a feature. In that case, we kept the rows where the faceoffs happened in Erie offensive zone. From the paramtric anova table, we can conclude that model is significant and that winning a faceoff in the offensive zone does increase the chances

to score a goal in that given sequence. We also tested using the `GA` and the defensive zones faceoffs, and the conclusions are the same.

```
## Anova for Parametric Effects
##              Df Sum Sq Mean Sq F value    Pr(>F)
## faceoff_win   2 372.08 186.038   185.6 < 2.2e-16 ***
## Residuals   850 852.00   1.002
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.2 Second-by-second analysis

So far, we have gathered plenty of evidence that winning offensive faceoffs significantly increases your 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. To verify that, let us take a closer look at the exact times the goals in questions were scored.

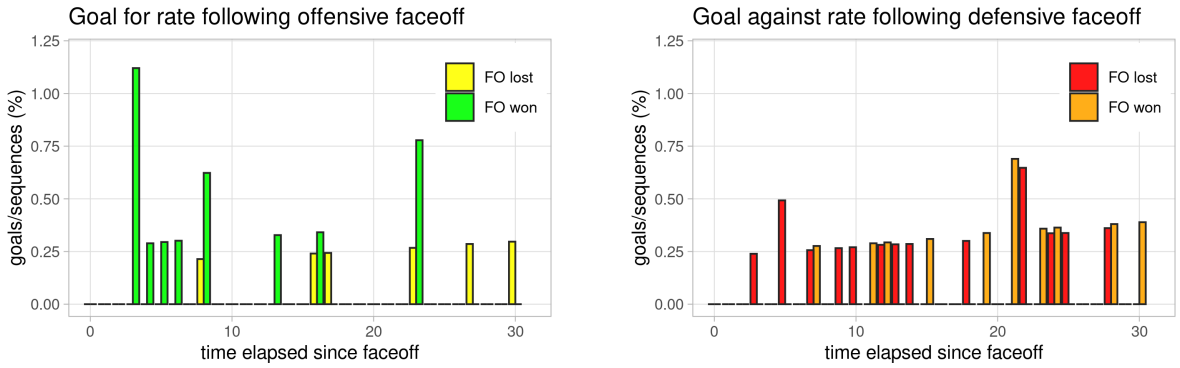

Figure 1: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds.

As shown in Figure 2, and somewhat unsurprisingly, winning the faceoff tends to lead more often to a quick goal. In fact, four goals were scored between $t = 3$ and $t = 4$, which feels a bit odd, but not enough to worry us. It is important to note that the y-axis provides the number of times a goal was scored within the timespan $[t, t+1)$ divided by the number of sequences of duration t or more. All the plotted bars actually corresponds to no more than ZZZ goals, while most bars suggest a zero probability that a goal gets scored within the timespan $[t, t+1)$. Naturally, this is because the number of events is rather limited. Such sparsness makes it hard to communicate the information contained in barcharts like those of Figure 2. To overcome such difficulty, let us construct a smooth version of these latter using loess (*locally estimated scatterplot smoothing*), a well-known method for smoothing scatterplots.

Informal model description: we want curves, an estimate of the probability of a goal occurring during a given time interval (of one second). One curve per situation (like in the bar chart). See Section~(SEC) for a more formal description of the underlying model. **Highlight that one needs to fine tune a smoothing parameter and explain how that was done.** The results are displayed in Figure~(LOESS).

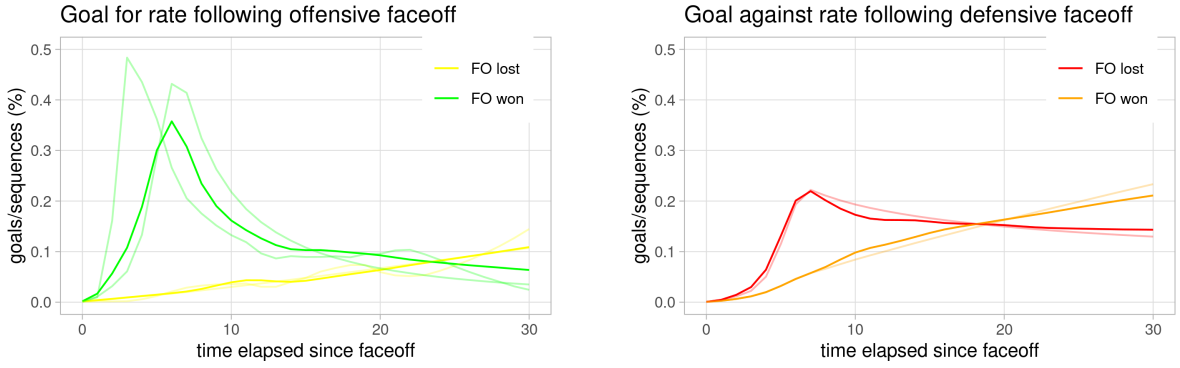

Figure 2: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \ldots, 30$, given that the sequence in question lasted at least t seconds.

1. discuss results in depth — GIVEN THAT THE SEQUENCE WAS LONGER THAN t.

2. SKIP PARAGRAPH. discuss weakness of loess: no confidence intervals for validating like we did previously.
3. discuss that not much data here, so possibly worthless anyways, (although CI from splines are conclusive).
4. discuss interesting questions that we cannot answer, need more data to use a more granular definition of situation.

0.3 The next step: a more granular context

Length: 1 pages

On refait notre analyse mais sur le data de la NHL.

Potentiellement expliquer quelques hypothèses supplémentaires (ex: prend le data de toutes les équipes car environ le même weigth, expliquer différence dans le data si le cas)

0.4 Technical details

0.4.0.1 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 $t \in \{0, 1, 2, \dots\}$.

0.4.0.2 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 $\ell \in \{1, \dots, ZZZ\}$ 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.

[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 $f_k(0) \approx 0$ $(k = 0, 1)$.