

How bad should I win next faceoff?

Big data cup 2021

Stéphane Caron* Jean-Philippe Le Cavalier† Samuel Perreault‡

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 some of these faceoffs are key moments, and thus have even more importance. 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 related to winning faceoffs. In other words, we want to confirm if winning a faceoff in the offensive zone is increasing the chances to score a goal during that sequence. On top of that, we would like to specifically answer this core question

**To what extend is winning offensive zone faceoffs increase the chances to score a goal?
What are the effects on the seconds following the faceoff? Is winning an offensive
zone faceoff even more important in man advantage situations?**

We targeted offensive zone faceoffs, because their impact on scoring a goal seemed more obvious as the team is already close to the opponent net. To answer that question, we decided to use the Erie Otters data made available in the `scouting` dataset. Our rationale for using this dataset was to use data of one single team, which from our perspective, simplifies the interpretation and the conclusions. Moreover, the `scouting` dataset is the one having the most observations for a single team.

In the next sections, we are going to try confirm the impact of winning offensive zone faceoffs on a sequence-by-sequence basis. Afterward, we are going to dive deeper in the effect of winning such faceoffs, by looking at seconds-per-seconds or man advantage situations. In conclusion, we will apply our approach to more data using NHL API (add link).

Sequence-by-sequence analysis

As in all data science projects, we first need to understand the data we are working with. As mentionned in the introduction, we focused on the `scouting` dataset. For simplicity, we will use some abbreviations for few repeated terms in this analysis:

- FO: Faceoff
- OZ: Offensive Zone
- DF: Defensive Zone
- NZ: Neutral Zone
- GF: Goal For Erie Otters
- GA: Goal Against Erie Otters
- Erie: Erie Otters

In the table @ref(tab:features1), we can see basic informations about the `scouting` dataset that appear relevant to our analysis.

*Intact(?) — Corresponding author: EMAIL

†Promu(?)

‡Department of Statistical Sciences, University of Toronto, Toronto, Canada.

Table 1: High-level features of the scouting dataset

Min. game date	Max. game date	Games	FO	Erie won FO	GF	GA
2019-09-20	2020-03-08	40	2441	1103	148	145

In order to tackle our initial problem, we had to structure the data in a way that we can easily analyze each sequences. After doing so, we can now extract additionnal informations about our dataset (see table @ref(tab:features2) and add context such as the position of the faceoff and the)

Table 2: Contextual data for faceoff situations

Zone	Penalty kill			Even strength			Powerplay		
	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 %

From the table @ref(tab:features2), 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 @ref(tab:features2) 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 is not that clear ...

To see if the effect of winning the faceoff in the offensive zone is significant, we fitted a logistic regression. From the parametric anova table, we can conclude that (at least for Erie Otters) the effect is significant:

```
## 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
```

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.

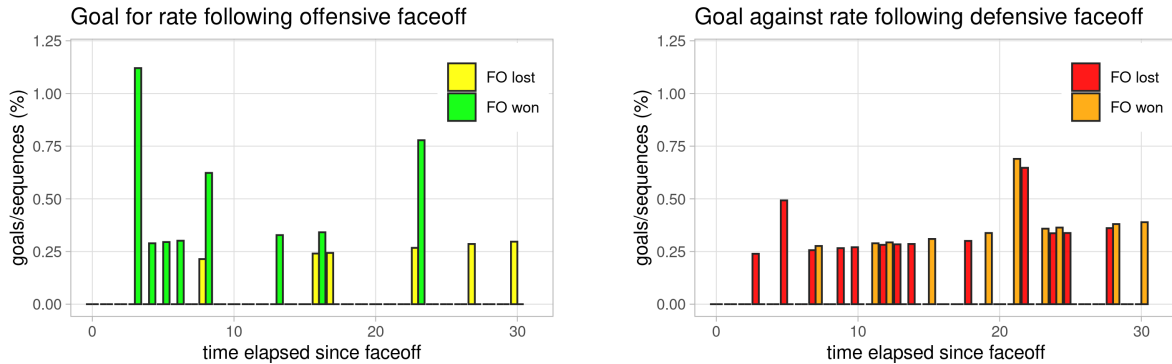


Figure 1: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \dots, 30$, given that the sequence in question lasted at least t seconds.

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 sparseness 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).

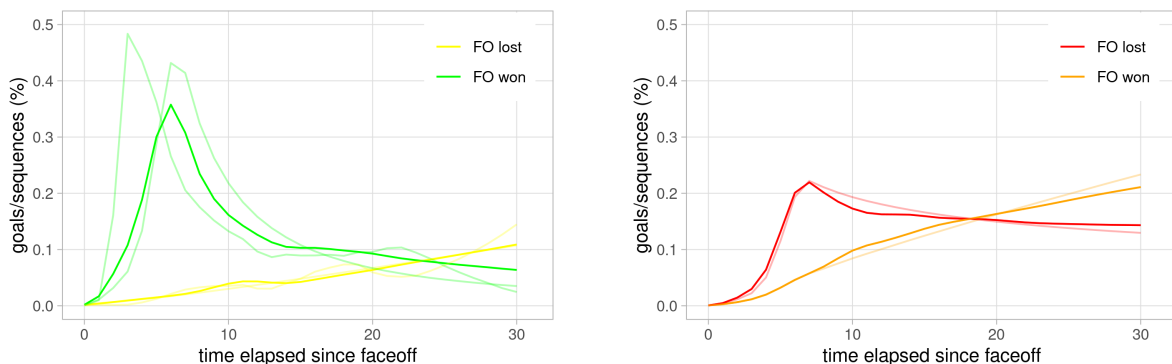


Figure 2: Percentage of times a goal was scored within the timespan $[t, t+1)$ for $t = 0, \dots, 30$, given that the sequence in question lasted at least t seconds.

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.

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 weight, expliquer différence dans le data si le cas)

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)

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

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

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

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

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.

¹To 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 $f_k(0) \approx 0$ ($k = 0, 1$).