Big Data Cup 2021:
VAEP in Hockey to put a value on zone entries

Tim Keller
University of Basel

5.3.2021

Abstract

While looking for a way to evaluate the differences between zone entries I came across the concept of valuing actions by estimating propabilities (VAEP) used in Soccer. By getting a probability of scoring and conceding a goal we will be able to put a value on every event in the dataset including zone entries. This makes it possible to distinguish the different types of zone entries and see the differences in value.

1 Introduction

Because I grew up around Soccer and I had no real relation to Hockey until my early teens, my way to adapt and learn the ways of Hockey always included searching similarities to Soccer. Offside is a concept present in both Hockey and Soccer but the concepts work completely different. While in Soccer the concept occurs multiple times in different places of the same play, the relevance of offside in Hockey is strictly related to the event of zone entries at the blue line to enter the offensive third.
Something that was not at all clear to me for several years is the dump & chase tactic some teams pursue that included playing the puck into the zone with no security of retaining the possession of the puck and risking to set off an attack for the opponent instead of safely carrying the puck into the offensive zone while retaining the possession. While watching more Hockey games I understood some of the advantages that concept has, especially by having the chance to possess the puck with way less distance to the goal like on a carry in. But in the end I still do not see how dump & chase would be more effective than carrying in the puck while entering the offensive zone.
So my goal with my Big Data Cup project was to put a value on these different types of zone entries to see if my feeling about how carried zone entries overall would be more effective and valuable than dumped zone entries or if my assumption is totally wrong and there is much more value in dump & chase than I expected.

2 Models

2.1 non-shot xG Model

Expected goals (short xG) is a concept that is very frequently mentioned in Hockey and in Soccer. The idea is to get a probability of scoring a goal by evaluating a large number of shots depending on things like the distance and angle to the goal and if the information is available the position of the goalkeeper or defenders.
For my model I took the approach to not only assign a xG-value to every shot but actually every single event. The concept is called non-shot xG because every event is seen as a possible place for a shot taken therefore an expected goal value is assigned despite it not necessarily being an attempt to score a goal yet.
The values I included as factors for my model for predicting the probability of a goal being scored in an event are the coordinates, distance to the goal, angle to the goal and the strength state of the teams.
To predict I used a XGboost classifier and a grid search parameter estimator. Because I never before made an expected goals model, I needed something to assist me along the way. Very helpful for me was the tutorial by SciSports lab on how to build own xG model "Tech how-to: Build your own expected-goals model for football" [1]. Their choice in methods seemed very reasonable and that is why I also used them for my model.

2.2 VAEP Model

The basis of the VAEP model is the concept established in the paper "Actions Speak Louder Than Goals: Valuing Player Actions in Soccer" [2] of using the current and two previous events to predict if a goal will be scored or conceded in one of the next 10 events. While implementing the model I oriented myself on the tutorial series by SciSports "Friends of Tracking: Valuing actions in football" [3]. For most of my features I had to twist the data available in the dataset a bit and create some more columns. I turned the game clock into seconds remaining (Something I learned from Nick Wan on his streams on Big Data Cup and Rocket League data), created a column for possession status determining if possession was gained or lost on during the specific event, I calculated the strength states of the team in possession of the puck, changing of a zone while the event and calculated the endpoint coordinates for every action depending of the type of event and the outcome.
This made it possible to include the features of distance to the goal at the beginning and the end of the event, the seconds remaining, goal difference, zone difference, possession status, the difference of the angle to the goal from beginning to the end of the event and the non shot xG calculated in the model I previously explained.
As model I used an xG boost classifier again because it was evaluated as one of the more accurate models in the "Actions Speak Louder Than Goals" paper [2] with 50 estimators and a max depth of 3.

3 Analysis

From the outputs of my model there are clear distinctions between the values of the three different types of zone entries. Carried zone entries are by far the most valuable of the three different types, while the overall sum is far higher than for the other types carried zone entries are also the most frequent types. But by looking at the mean value you can see that with a mean value of 0.008142 in comparison to 0.000643 for dumped zone entries the mean value for carries is more than ten times higher. Surprisingly played zone entries have a negative mean vaep value although played zone entries do not fail as often as dumped entries.
As for analysing the best players at zone entries using the vaep framework there is once again the possibility to use the sum or the mean value provided. For the mean value we need to set a cutoff for the minimum of zone entries required to qualify for the evaluation because otherwise there would be several players with a high ranked mean zone entry vaep value. For that reason we only evaluate players with at least 10 zone entries. The top 10 of the players with the hights vaep sums of zone entries is filled with star players of the American and Canadian Olympic teams, the Boston Pride like Hilary Knight, Sarah Nurse and Samantha Davis. There is also one player of the NWHLs newest franchise the Toronto Six: Mikyla Grant-Mentis is a name that stuck with me from watching the coverage of the game of the Lake-Placid bubble and she is indeed the player with the second highest vaep sum from zone entries. The highest ranked player Blayre Turnbull will be a topic later again but is clearly visible in the top right of Figure 1 showing all the vaep sums.

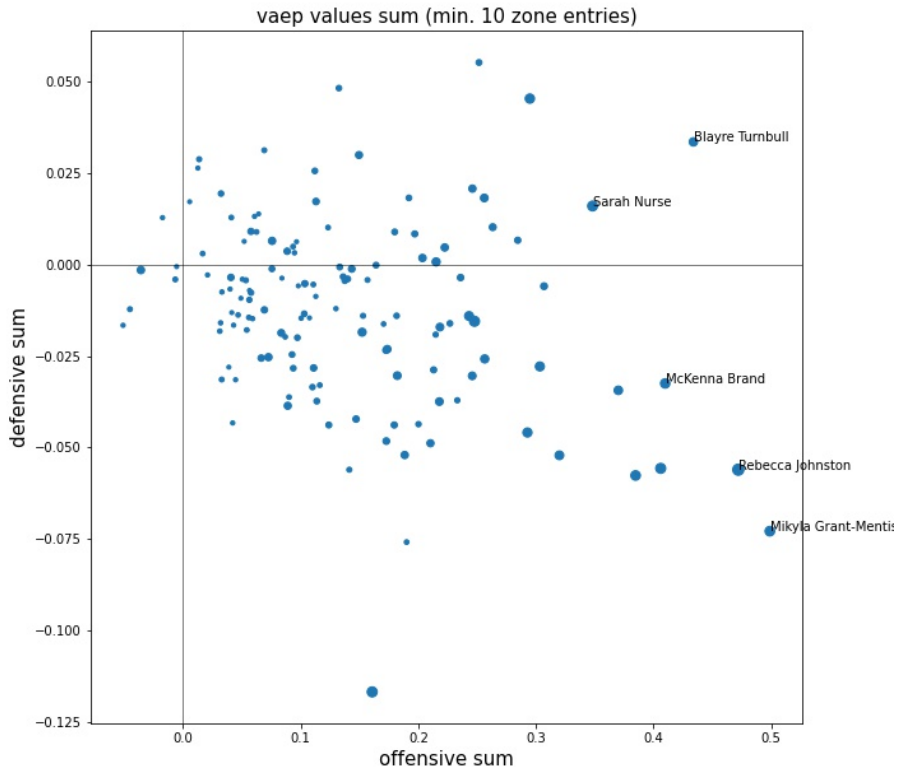

Figure 1: total sum of vaep value from zone entries split in offensive and defensive value

Regarding the mean vaep value of zone entries with a cutoff of minimum 10 zone entries made there is a bigger variety of players from different teams in the top spots like Susanna Tapani of Team Finland or Rachael Smith from the St. Lawrence Saints.

The top players of both the measurements are also the two only players that are in both top ten lists. Tori Sullivan has with a little bit of seperation the highest mean value for her zone entries and with only 21 zone entries still has the 10th highest vaep sum (far less than anyone else above). Blayre Turnbull has the highest sum of zone entries a good amount ahead of Grant-Mentis in the second spot with 14 less attemps. With the second lowest count of zone entries she has also the 6th highest mean value. Sullivan and Turnbull therefore can be seen as the two best players in providing value

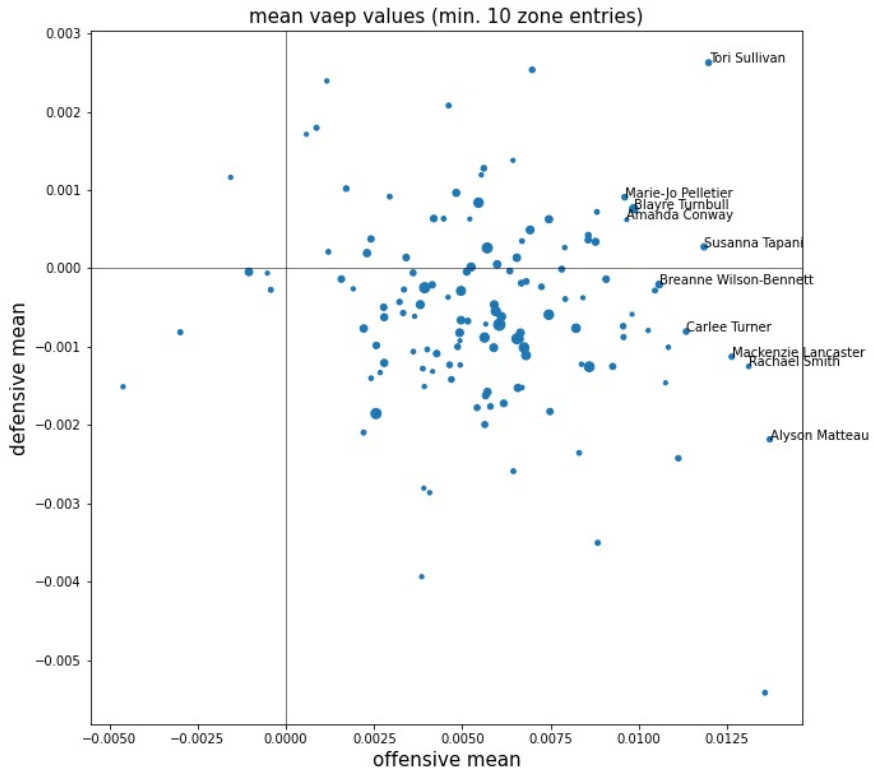

Figure 2: mean of vaep value from a minimum of 10 zone entries split in offensive and defensive value

for their team through zone entries and could be even more valuable if they would handle an even higher number of entries. They both primarily do their zone entries through carries, that was mainly expected because of the differences in vaep value for the different types of zone entries.
To set zone entries into an overall context in the vaep Framework it is important to look how it ranks to other event types in hockey. Like expected a zone entry ins not more valuable than a goal or a Shot but it has the third highest mean value and is fourth in the overall sum with passes getting ahead there. Another reason why it is worth to look into who is providing a lot of value through them and what things work best.

4 Conclusion

I am a bit surprised how big the differences between the different types of zone entries is and how superior a carry in seems to be. The next surprise is that a zone entry via a passing play overall seems to have a negative value and is less effective than a dump in, although puck possession is lost way more often through dump ins. But my original thesis is upheld that carry ins are overall the most effective and valuable way for to enter the offensive zone.
Unfortunately I was not able to do all the analysis I would have loved to do regarding zone entries. It would be very helpful to be able to provide even more information on highly valuable zone entries and their characteristics like the locations where on the blueline the entries are made.

My code for the Big Data Cup 2021 is available on my github repo in the notebook hockey_vaep_final.ipynb: `https://github.com/TK5-Tim/Big-Data-Cup`

5 References

[1] Van Haaren, Jan: `https://bitbucket.org/scisports/ssda-how-to-expected-goals/src/master/`, 2016
[2] Decroos, Tom and Bransen, Lotte and Van Haaren, Jan and Davis, Jesse: Actions Speak Louder than Goals, 2018
[3] Bransen, Lotte and Van Haaren, Jan: `https://github.com/SciSports-Labs/fot-valuing-actions`, 2020