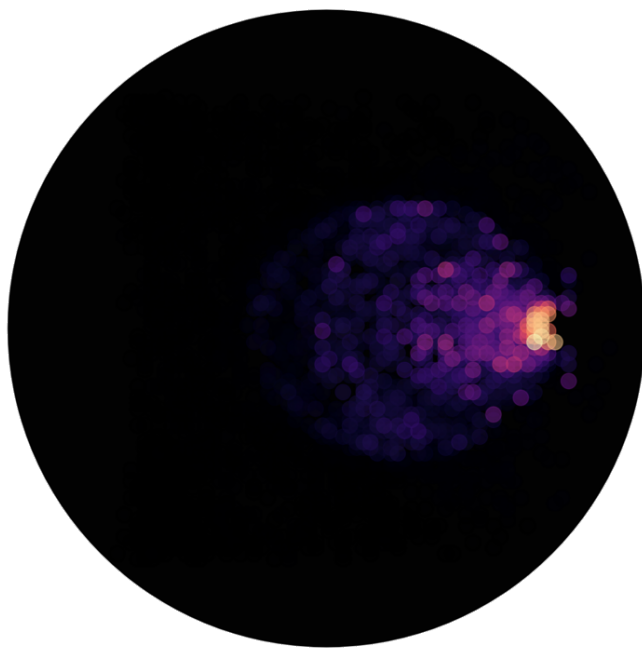


**It's Better to be Good than "Puck-y":
An Analysis of Player Skill Using Expected Goals**

Tej Seth, University of Michigan

Santiago Casanova, Instituto Tecnológico Autónomo de México



I. Introduction

- A. Question: How can we evaluate hockey players based on a set of metrics that are extrapolated exclusively from tracking data?
- B. Thesis: Using knowledge and techniques from across the sports analytics world, we created models to approximate expected goals, win probability, and completion percentage. The data we generated was useful for assigning a composite score to each prospect as well as for clustering prospects together into established categories to recognize which type of hockey player, based on their playstyle, added the most value to their team

II. Data:

- A. A 40 game sample from the Ontario Hockey League team, Erie Otters, made public for the “Big Data Cup 2021”

III. Models

- A. Win Probability Model
 - 1. The win probability model was used to evaluate how prospects would change the course of games by either increasing or decreasing their team’s win probability. The model takes the current period, time left in the period, total time left in the game, skater differential, and score differential to approximate win probability at any given time. There was an observable relationship between win probability and goals (Figure 1), as goals came when the win probability was around 50% for both teams or when the game was close to being decided (<10% or >90%), therefore it also served as an input for the expected goals model.

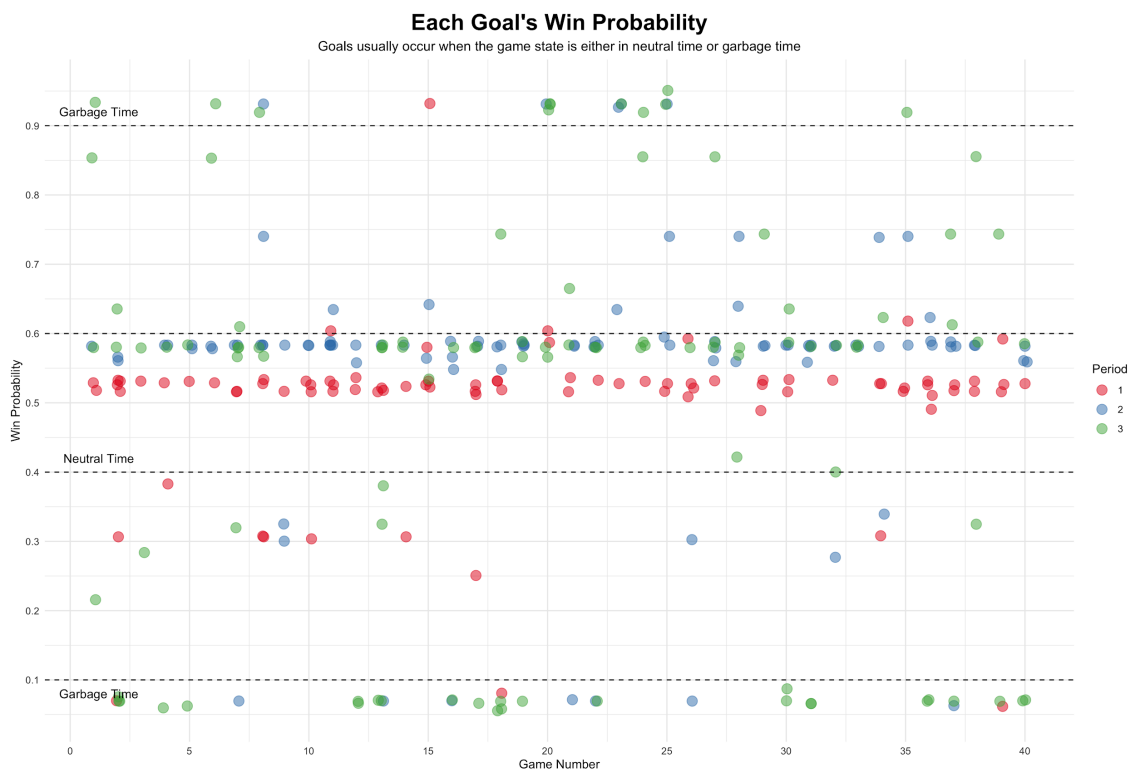


Figure 1

B. Expected Goals

- Using David Sumpter's principles for expected goals in soccer, we developed a logistic regression model that utilizes shot distance, shot angle, and X coordinate of the player as well as win probability, one-timer opportunity, and whether it was a snapshot or wristshot (only two types of shots deemed statistically significant to change xG) to determine the likelihood of any given shot becoming a goal. As the game of hockey revolves around goal-scoring, a precise expected goals model provides a meaningful analysis start point. We used "area under the ROC curve" (AUC) to determine model performance as it works well for imbalanced classification scenarios. Our final AUC on unseen data resulted in 0.7997 which suggests very good/borderline excellent discrimination (rule of thumb deems values ≥ 0.8 excellent (Hosmer & Lemeshow)).

C. Completion Probability

- Taking inspiration from quarterback analytics in football, we created a completion probability model with the added benefit of having positional data of the players involved on any given pass. For this model, we used on-ice skater differential, pass distance, passer/receiver X coordinates, and whether it was a tape-to-tape pass or not

to determine the likelihood of it being completed. Given the fluid nature of Hockey and the lack of defender data, the completion probability model wasn't as precise. Our final AUC on unseen data came in at 0.71 which suggests acceptable discrimination.

IV. Metrics Created

- A. C-Score (Chance score): This metric is essentially the average expected goals per shot. It's a good measure of the average shot quality of a player and it's heavily influenced by a player's positioning.
- B. CG-Score (Chance-Goal Score): Following the "goal is a goal" principle, this metric assigns a score of 1 if the shot resulted in a goal and the expected goal value if the shot wasn't a goal. Similar to C-Score, it measures the quality of the shots taken while also rewarding goals that are scored.

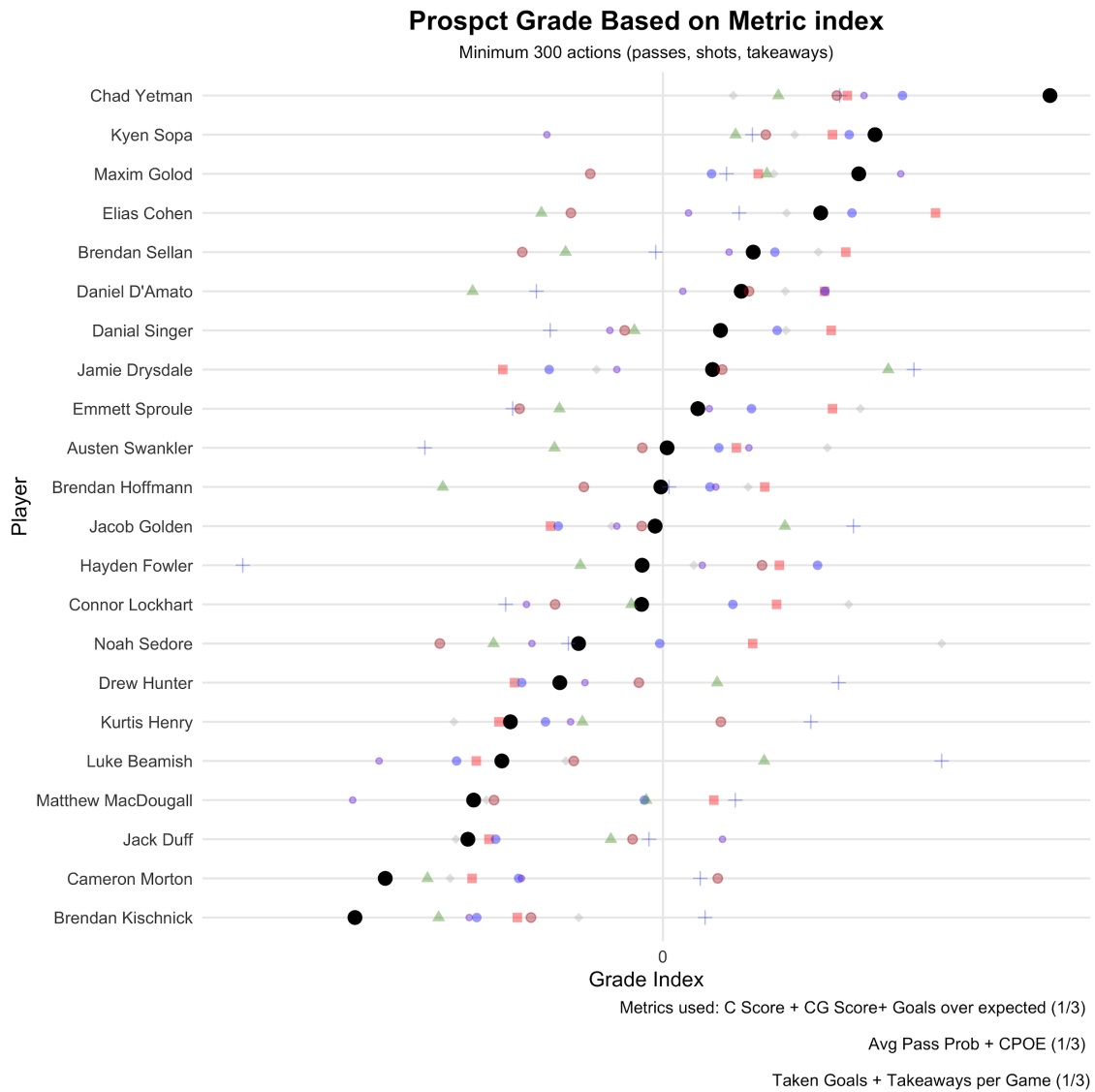
$$CG = \frac{\sum_{i=1}^{no-goal} xG*1000 + \sum_{i=1}^{scores} G*1000}{total\ shots}$$

- C. GoX (Average Goals over eXpected): Average difference between a shot's result and the shot's probability. It rewards risky/hard shots that were converted and punishes easy shots missed. It's a good measure of pure shooting skill.
- D. APP (Average Pass Probability): Average probability of a pass being completed. Rewards players that make "safer" passes rather than risky ones.
- E. CPOE (Completion Percentage Over Expected): Average difference between pass result and pass probability. It measures a player's precision passing skill as players are measured on their ability to complete a myriad of passes from easy to difficult.
- F. TG (Taken Goals): Expected goal value of the opponent at the moment of a takeaway. It's the virtual concept of goals "taken" from the opponent. Rewards defensive plays in dangerous situations.
- G. TA/G (Takeaways per Game): This is an established measure of a player's defensive efforts to take the puck away from the offense.

V. Composite Grade

- A. Using the aforementioned metrics, we created a composite grade that included all phases of the game: scoring, passing and defending (within our limitations). To ensure equal importance within phases and metrics, we standardized all variables and combined them in a weighted average format $([C-Score + CG-Score + GoX]*\frac{1}{3} + [APP + CPOE]*\frac{1}{3} + [TA/G +$

$TG]^{*1/3}$). The final number represents standard deviations from the mean with equal weight (Figure 2).







VI. Clustering

- A. Using the six metrics we created such as Chance Goal Score, as well as already established metrics like Passes Per Game, we were able to cluster every player to play for or against the

Erie Otters into 5 clusters with the attributes that make up each cluster shown below:



B. With each of the players inside of their clusters, we were able to analyze which clusters led to the higher composite grade:

Clusters Ranked by Average Grade					
Cluster	Example Player	Average Grade	Min Grade	Max Grade	Description
2	Maxim Golod 	1.00	-1.20	3.13	Cluster 2's players represent goal scoring: They excel in creating high chances to score, actually scoring goals. Plus they are great at playing defense and average at passing too.
1	Jamie Drysdale 	-0.12	-1.59	0.75	Cluster 1's players prides themselves in their passing ability, having a high volume, and great CPOE. Their biggest flaw is defense.
4	Connor Lockhart 	-0.23	-0.62	0.63	Cluster 4 players take good shots and are elite at scoring. Their defense, however, is average and their passing ability is lackluster
3	Matthew MacDougall 	-1.07	-2.03	-0.13	Cluster 3's players aren't exceptional at anything but they are average at passing and below-average at scoring and getting takeaways on defense.

VII. Conclusion

A. Applying techniques from other areas of the sports analytics world, such as soccer's expected goals methodology and football's CPOE, we were able to create a plethora of models that could be combined to evaluate all areas of a player. We look at each player's defense, passing, and scoring; the three main components of hockey that are measurable with event data. By combining these evaluations into an index, we were able to assign each player a grade that can help scouts support their observations with actionable data on the analytical side of their performance.

VIII. Appendix

A. Sources

1. Clustering: <https://alexcstern.github.io/hoopDown.html>
2. OHL rulebook
<https://cdn.ontariohockeyleague.com/uploads/ohl/2019/10/16162936/2019-2020-OHL-Rule-Book.pdf>
3. David Sumpter xG Methodology as seen on the “Friends of tracking” youtube channel
4. Hosmer & Lemeshow (2013). Applied logistic regression. P.177:

B. Code repository

1. <https://github.com/SCasanova/Big-Data-Cup-2021>