Utilizing Decision Trees with Expected Goals to Understand High-Probability Scoring Chances in Women's Hockey

Nicolas Wispinski (nicolas.wispinski@gmail.com)

Abstract

What features of a shot on net tend to lead to more goals? With data from a sample of the 2018 Women's Olympic Hockey Tournament, the 2021 National Women's Hockey League (NWHL) Isobel Cup, and select NCAA games, an expected goals model is generated from all goals and on-target shots. A logistic regression is modeled on ten features from the dataset with the target vector being successful shots (goals). The probability of the shot leading to a goal is derived from the model for each observation. The top 15% of shot probabilities are classified as high-probability scoring chances. Using the same 10 features as the logistic regression with the new target variable of high-probability scoring chances, a decision tree is generated across the three different levels of data (Olympic, Pro, Amateur) to show the differences between the three groups. Ultimately, distance to the net and angle to the center of the net are the two most important features of a high-probability scoring chance across all three levels of play. The other important features vary based on the level of play. The decision tree also shows thresholds for each of these features to find optimal combinations of events. Insights from this model may be used to guide future offensive strategy irrespective of level of play.

Methodology

The goal of this research is to determine what elements of a shot attempt generate a high-probability scoring chance. What is the relative importance of the features? Where on the ice are high-probability scoring chances generated? Data for this work has been provided by Stathletes as part of the Big Data Cup. The women's hockey data set was used for this research. This dataset was based on Olympic, NWHL, and NCAA games, and each observation represents one event. The observations selected for this research were goals, and on-net shots. Ten features were created from the dataset. These ten features were selected based on assumptions of impact on the likelihood of a goal:

Feature	**Description**
Traffic	Player presence in front of the net for the shot
One-Timer	Whether the shot was taken immediately upon receiving a pass or not
Previous Event	Type of the previous event before the shot (Dummy encoded variable for the five types of events: Pass, Puck Recovery, Shot, Takeaway, Zone Entry)
Home Team	Whether the shot was from the home team or not

Home Advantage	Number of players the home team has on the ice relative to the Away team
Level	The level of play (Olympic, NWHL, NCAA)
Clock Delta	Time since previous observation
Previous Event Distance	Distance from previous event to the shot
Net Distance	Distance from the center of the net to the location of the shot
Net Angle	Angle from the center of the net to the location of the shot

With these features, 1,732 observations were fitted to a logistic regression based on whether the observation resulted in a goal. The output of the logistic regression is the predicted probability of the shot resulting in a goal, based on the 10 features. The observations in the top 15% of predicted probabilities are classified as a high-probability scoring chance. This quantile was chosen to ensure enough labels for the next process, while ensuring the probabilities are still high.

The same observations and features are then used to generate three different decision trees with a maximum depth of three, one for each of Olympic, NWHL, and NCAA games, with the class prediction set to whether or not the shot is classified as a high-probability scoring chance from before. The decision trees rank the features using Gini importance. Each node of the decision tree provides a threshold for the deciding feature, leading to combinations of observations that most often lead to goals.

Results

For each of the three decision trees, the most important feature in creating a high-probability scoring chance is the Euclidian distance to the center of the net from the location of the shot. The next most important feature is the angle from the center of the net to the location of the shot. The importance of the remaining features varies based on the level of play. Most features were deemed unimportant by the decision tree. Feature importance is shown below:

Olympics (n=659)		**NWHL (n=1015)**		**NCAA (n=58)**	
Feature	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**
Net Distance	0.454	Net Distance	0.458	Net Distance	0.507
Net Angle	0.360	Net Angle	0.411	Net Angle	0.308
Shot	0.100	One Timer	0.065	Takeaway	0.085
Clock Delta	0.044	Shot	0.057	Traffic	0.079
One Timer	0.024	Previous Event Distance	0.010	Home Advantage	0.021
Previous Event Distance	0.015				

The features with a lower value of importance highlight the differences between the three levels of play. The inclusion of Clock Delta and One Timer as important features in the Olympics show that the game is played at a faster pace than the other two levels, for example.

Additionally, the output of the decision tree gives threshold values at each node. The thresholds for Net Distance and Net Angle can be combined to create an area representing a high-probability scoring zone:

Olympics
Distance: 11.9 Feet
Angle: 85.7 Degrees
Area: 106 Sq. Feet

NWHL
Distance: 18.4 Feet
Angle: 83.2 Degrees
Area: 123 Sq. Feet

NCAA
Distance: 16.1 Feet
Angle: 93.3 Degrees
Area: 105 Sq. Feet

The above chart shows the physical area in front of the net where the decision tree deems high-probability scoring chances occur most often at each of the three levels of play in the dataset. One possible reason for the NCAA having the smallest scoring zone is the small number of samples (n=58) included in the dataset.

Recommendations

This research shows how and where high-probability scoring chances occur, and how they differ based on the level of play in the Olympics, NWHL, and NCAA. The use of decision trees provides answers to complex questions in an easy to digest format for non-technical audiences. This research can be used by the coaching staff to visualize the importance of shot attempts from the area in front of the net. Aspiring Olympians can use the findings above to adjust their training and playing style accordingly. This research could also be performed on a team-specific basis to prepare defensive strategy for upcoming opponents, as well as finding areas of improvement in the team's own offense. These findings are not limited to women's hockey, or hockey at a high level.