Highway to the Danger Zone

Ben Howell, Carleen Markey, An Nguyen, and Nayan Patel

1. Introduction

During a power play, there are two opposing forces struggling to take control of the game: the penalty kill's attempts to prevent a goal and the power play's drive to get the puck in the back of the net. Ignoring goaltender and shooter ability, these two objectives primarily hinge on defensive team play holding strong to prevent or breaking down to allow high-danger unblocked shots.

In this paper, we attempt to analytically answer the following questions: **in what situations does defensive play break down and in what situations does it successfully prevent shots?** Additionally, what kind of actionable tools can we provide coaches and analysts to apply to their own teams and strategies to both minimize and maximize high-danger shot attempts?

To investigate these inquiries, we drew upon recent trends from the NFL's Big Data Bowl where punt return paths were measured using triangulation and A* search algorithm [5], as well as one author's previous experiences in using minimum spanning trees (MSTs) in star formation research [1]. Building off of these network-based approaches and pre-existing frameworks for xG models [3], we used logistic regression to identify the most impactful variables on power plays and penalty kills for predicting the likelihood of dangerous situations. Using a formulaic model allowed us to integrate it into a `JavaScript`-based webtool (highway-to-the-danger-zone.netlify.app) for ease of use and exploring future research paths.

2. Methodology

The new player tracking data for the 2022 Big Data Cup (2022 BDC) was a double-edged sword: while it provided aspects of women's hockey data never seen before, it also presented brand-new data challenges. In an independent review of individual plays, we found the cleaning and merging process developed by @the_bucketless on Twitter produced an accurate dataset. Thus, we used his cleaned version of the 2022 BDC play-by-play and player tracking data [8] and adapted his provided merge example [9] to produce a dataset of 820 play-by-play events (each event referred to 'game state' hereafter) that occurred during a power play, with listed event details and player locations for each game state.

To get our data ready for the model, we first filtered out any events that the penalty kill was responsible for. This was done because our goal was to understand what makes for good shot suppression for the penalty kill and what helps to get pucks on net for the power play. When the penalty kill team has the puck, the objectives shift to getting the puck out of the zone for the penalty kill, and regaining puck possession for the power play. Measuring the success of these distinctly different objectives is out of the scope of our project. We also filtered out event types that were non-essential to the model such as faceoffs, penalties, and events occurring in the defensive zone (from the power play's point of view). Simply put, we excluded game states that did not involve the warring power play and penalty kill actively preventing the other from trying to achieve their goals of pucks on net and shot suppression, respectively.

With this data, we aimed to keep it as simple as possible, so **we developed a logistic regression model with `scikit-learn` [4] that predicted the probability of a game state being a 'dangerous situation'**, much like how expected goal (xG) models produce probabilities of a goal based on circumstances during a shot. We approach the definition of 'dangerous situations' from a coach's perspective. Tactically, high-danger unblocked shots are created by situations where the power play isolates penalty killers in unfavorable positions to create openings to take an unblocked shot from the 'home plate' area in front of the net (aka a high-danger unblocked shot). These ‘dangerous situations’ are primarily accomplished by well-executed passing and positioning. So, we decided to work backwards and find out how many passes occur before a high-danger unblocked shot to create a quantitative definition of a ‘danger situation’.

To do this, we went through recordings of 13 women’s hockey games from the 2022 Olympics and determined that, on average, 3 unbroken passes occurred between a zone entry or turnover to a high-danger unblocked shot. **We thus defined the game states corresponding to the last three passes of the unbroken pass sequence directly before a high-danger unblocked shot to be 'dangerous situations'** in our data set (marked as 1 for modeling purposes, all other game states marked as 0). If only one or two passes preceded a high-danger unblocked shot after gaining the zone or possession, then only those passes’ game states were marked as dangerous.

As stated before, the new player positioning data for the 2022 BDC opened previously-closed doors in women’s hockey analytics, particularly regarding the measurement of how other players on the ice affect on-ice events. By nature, positional data has player x/y coordinates for players labeled 1-5 on the ice. Unfortunately, this creates inherent problems such as the consistency of denoting 'player 1' vs 'player 5' and so on. To reduce these problems and focus more on the relationships between players rather than their raw positioning, we drew upon previous work that measured network effects in star formation simulations (minimum spanning trees) [1] and punt returns in the Big Data Bowl (triangulation and A* search algorithm) [5].

Minimum spanning trees (MSTs) were chosen as the network measurement method because they identify “a graph consisting of the subset of edges which together connect all connected nodes [without cycles], while minimizing the total sum of weights on the edges” [7] and are deterministic. An example of MST construction from Wikipedia is shown below in Fig. 1.

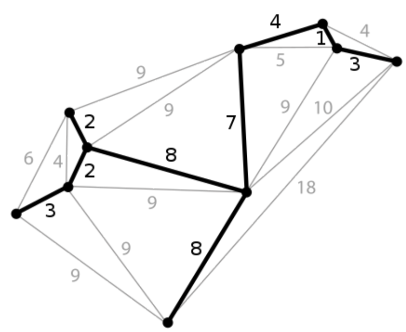

Figure 1: An MST (black) constructed out of a network of distances between each point (gray). [10]

We created three different Euclidean MSTs that, respectively, connect the power play players, the penalty kill players (including the goalie), and all skaters; the edge weights are the Euclidean distances between players. We then developed the following independent variables for modeling, primarily based on these MSTs:

- Average distance between players in the network (for all three MSTs)
- Total summed distance between players in the network (for all three MSTs)
- Average number of connections for each player (for power play MST. This variable for both the all skater and penalty kill MSTs had high variance inflation factors and were dropped)

- Ratio between the average lengths of the power play and penalty kill MSTs (shows if the penalty kill is grouped up in one spot)
- Opponent connection measure from the all skater MST (shows if the majority of the network connections are between teammates or between opponents)
- Puck distance to the attacking net (major factor in shot quality [cite xG models] and therefore impacts if player attempts an unblocked high-danger shot)

Our final model, dubbed **the Highway model**, ended up being a relatively straightforward combination of logistic regression with an L1 penalty for feature selection, class weighting, and `scikit-learn`'s `SelectKBest` with the above variables [4]. The model began with the 10 base variables above, but we added 45 interaction terms by multiplying two base variables together to capture the interactions between variables and what they measured. Using `SelectKBest`, we pared the model down to 45 overall variables to reduce some of the complexity and ensure that the final variable inputs were ones that made a positive impact on the model. However, in the process of nailing down this approach, we ran through a multitude of different failed approaches in an effort to improve the model and address the unbalanced dataset.

We first attempted both over- and under-sampling in our training datasets in attempts to balance our unequal ratio of dangerous situations and non-dangerous situations (80:541). However, over-sampling successes failed to return an improved model and the under-sampled approach left us with too few points in our training dataset. After deciding not to alter the training data, we attempted principal component analysis (PCA) and recursive feature elimination (RFE) on the variables and interaction terms; neither showed an improvement over our final approach and PCA only further decreased the interpretability of our input variables.

New research like the Highway model can be useful in its raw coding form, but **to maximize the functionality of our project for coaches, analysts, and others in hockey operations, we created `JavaScript`-based webtool** (https://highway-to-the-danger-zone.netlify.app/) that allows the user to interactively explore the probability of a spatial configuration of players and puck being a danger state. The webtool is built primarily using `d3.js`, and uses the `interact.js` library to enable dragging functionality. The user can drag dots representing players and the puck around a rink and see how the probability shifts with their movements. The user can also adjust the number of players on each team and show or hide MSTs between players (denoted 'player/skater connections'). If a spatial configuration has at least one base variable with a value outside the values seen in the model dataset, the user will be shown a warning to be skeptical of the displayed probability. Examples of the webtool in action are shown in Fig. 3-4.

3. Discussion

3.1 Highway Model Results and Evaluation

As we developed the Highway model, initial tests returned models that were heavily weighted towards false negatives, essentially indicating pessimism in predicting that something would produce a high-danger-situation. We preferred a model that, if it was going to be wrong, would err towards false positives because there are high-danger situations that don't always result in a high-danger shot or goal, due to outside factors.

Our final model setup produced an accuracy rate of 85%, with a sensitivity of 53% and a specificity of 90%. The visualized ROC curve and confusion matrices are shown in Fig. 2.

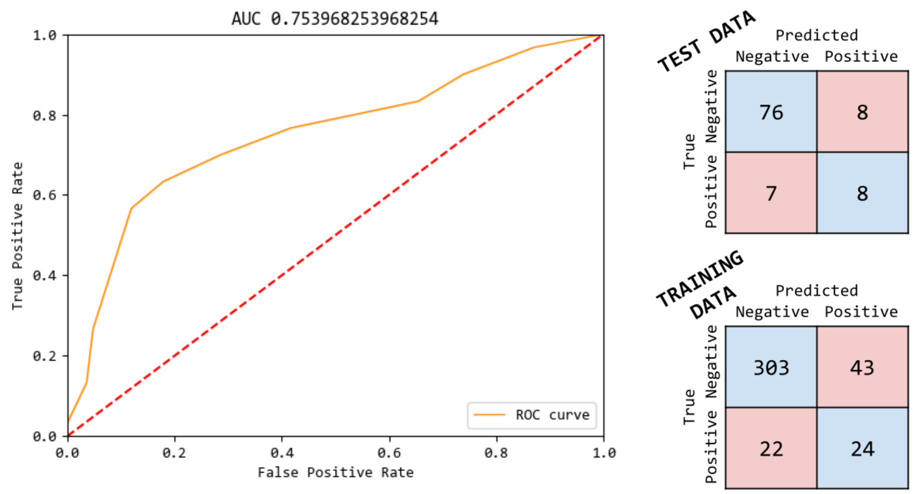

Figure 2: Left: ROC curve and AUC for the test data. Right: Confusion matrices for test and training datasets

Combining these model measures, the model is not necessarily the most robust, but for the purposes of our research and the scope of our project, we viewed it as a success. We believe the combination of a sparse dataset and using a logistic regression may have limited our model's scoring and predictive quality, and would be interested in the future to test further tree-based methods, like random forest or `XGBoost` models, ideally on a larger dataset. However, we stuck with the logistic regression because of its formulaic nature, which allowed us to easily integrate the model into the previously-mentioned webtool for use by a coaching staff.

The other major benefit of using a logistic regression model is its interpretability. For example, we find that the most positively impactful variable on probability of a 'dangerous situation' is the total summed distance in the MST of all skaters. This makes sense, as the more space offensive players have to work with, such as during rushes, the easier it is for them to catch defenders out of position and get a shot off. On the other hand, the most negatively impactful variable was the puck distance to the attacking net, i.e. the farther away the puck gets from the attacking net, the less likely it is for the game state to be a dangerous one. This is also intuitive and links up with the importance of this same variable in xG models. While we would like to do a full examination of the impacts of each variable in this paper, as it would lead to a more granular conceptual understanding of power plays for those researching analytics, we believe this is not where our research can make the most impact.

3.2 Possible Usages Immediately Available to Hockey Operations

Instead, we turn to immediate practical applications for current coaches and analysts. A few years ago, Ryan Stimson put together a book of modern hockey tactics built off of conclusions uncovered from the last decade or so of hockey analytics [6]. Coming full circle, we can now use our webtool to assign empirical probabilities of success to penalty kill and power play tactics that he suggests. Fig. 3-4 show examples of some of these tactics and formations from the book, which have been reproduced with permission, plugged into our webtool.

For this example, shown in Fig. 3, we see the power play using a 1-3-1 setup with PP1 having control of the puck in the corner. Here the penalty kill is set up in a Wedge +1 (or Czech Press), where 3 defenders are tightly controlling the slot/net area with the extra defender (PK1) pressuring the puck. In this situation, our model predicts a probability of this game state being a 'dangerous situation' to be just above 40 percent:

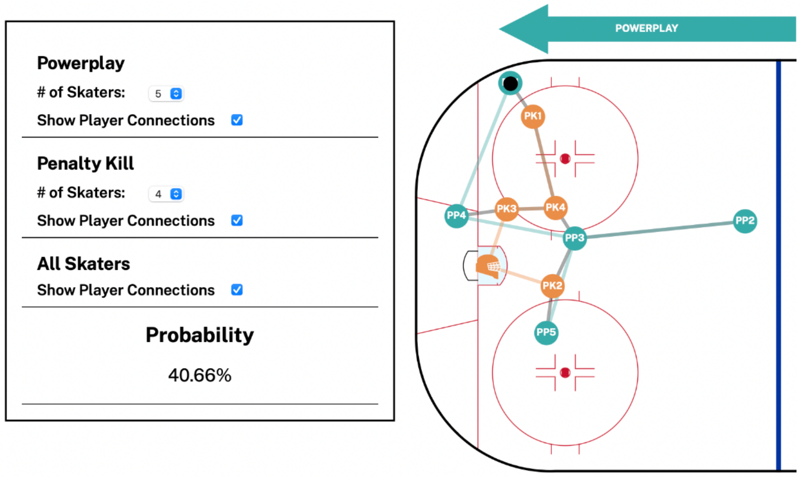

Figure 3: 1-3-1 power play setup with the puck controlled in the corner against a flexible variation of the Czech Press penalty kill setup. Blue dots denote power play skaters and orange dots denote penalty kill players

With this scenario, two possibilities in play occur the most often. The first, shown on the left in Fig. 4, is PP1 cycles the puck up to the point (PP2), who will most likely look to try and bring the puck slightly off the boards. Using our webtool, we are able to see that the most optimal way to decrease the probability of a danger situation would be for PK4, the closest defender to the puck, to switch off and pressure the point, with PK3 rotating in to replace PK4, and PK1 cycling back to cover PP4. This trims the probability down to ~23% (see Fig. 4, left).

The other scenario that could happen is PK1 is successfully able to force a loose puck or engage in a battle in the corner. Shown on the right in Fig. 4 below, the optimal way to play in this situation for the penalty kill would be PK4 to support PK1 along the boards, whether that's in a puck battle, or as an option to chip a pass up the boards. PK2 rotates to fill in PK4's spot on the strong side as you want to overload one side of the ice to create a turnover or be available for a quick breakout in the case of contested possession. This rotation reduces the probability all the way down to just above 10%.

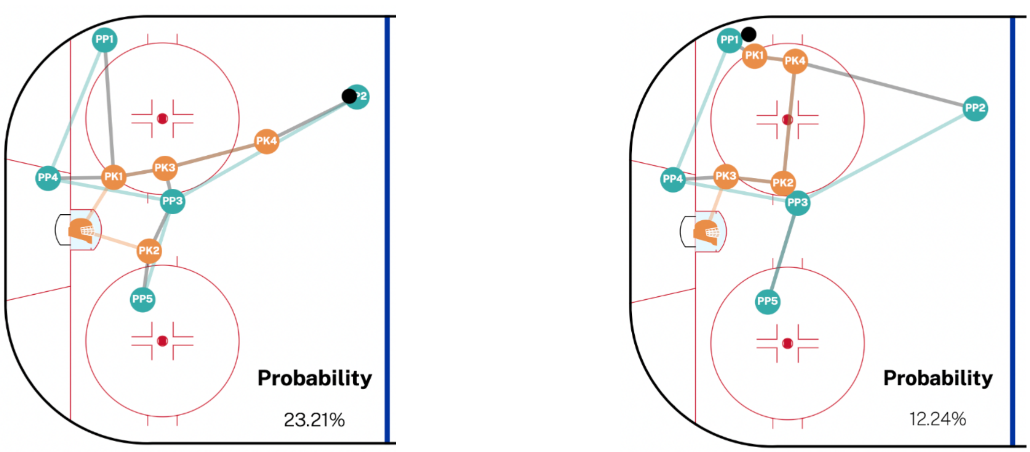

Figure 4: Left: Flexible rotation while keeping gap control if the puck is cycled to the point. Right: If the attacker is pressured off the puck into a corner battle, this would be the optimal rotation for the PK

This process can be done with any set of suggested tactics, allowing coaches and analysts to workshop power play and penalty kill strategies in an easily communicable and objective way.

Therefore, **the current iteration of our project provides coaches and analysts with a well-rounded and developed tool that can immediately integrate into pre-existing game-planning and film review for power plays and penalty kills.** However, there is plenty of room to push our methodology and process further, taking it to the point where it can be used in more of a predictive sense for game-planning purposes.

One question that we would expect from any curious coach, once they've input the game state that they're evaluating, is "okay, so what's the best thing to do now?". By dragging the puck to the other offensive players on the ice, an enterprising analyst can evaluate how changing the puck position (presumably through a pass) would affect the probability of a high-danger state. By looking at the probability across the offensive player positions, one can then determine what the optimal next pass is for a player. This could even be accomplished in-game by using the webtool on a mobile device and arranging all player positions as needed on-screen.

While evaluating the next optimal pass is a relatively straightforward question, another avenue to explore is optimal defender/attacker positioning. If one holds all other player positions constant, one can explore a given player's potential locations to determine what x-y combination would be optimal for that situation. This is possible with the existing webtool. By inputting all player positions for a given frame of game tape, a coach could then move a single player's position until the highest or lowest probability in a reasonable radius of that player's original position is found. By limiting the search radius to a relatively small area of potential points, a coach could offer reasonable feedback in real time to players; it's not feasible to expect a player to be fifteen feet to their right as they come down the ice, but giving them a cue that they should be five feet further right is much more reasonable. This type of automatic search for optimal positioning is a webtool feature we hope to implement in the future. With gap control being a key, fundamental, idea to designing offensive and defensive zone strategies, we feel this application of the webtool could pay dividends with regards to powerful insights alongside ease of use.

In a similar vein of player development and evaluation, this tool and model can also be used to drill down the cause of shot production or lack thereof on penalty kills and power plays. For example, if high-danger shot attempts are being created during on a penalty kill, examining the associated probabilities with the game state and preceding game states for shot attempts can indicate if the issue was due to positioning or tactics, or if it was caused by external factors such as the player's shooting ability or the defending goaltender's ability. On the flip side, if shots aren't being generated on a power play, a similar analysis can be used to discover the cause.

These are far from the only use cases for either the webtool or the Highway model. The immediate possibilities for player tactics, player development, video and live coaching, scouting, delving deeper into the inner workings of power plays, etc. are too numerous to list in this paper.

4. Conclusion

The power play is one of the most dangerous scenarios in hockey, a situation that both teams want to be prepared for. Using the novel player tracking data produced for the 2022 Big Data Cup, we developed a logistic regression model to predict what on-ice situations led to the most, and least, dangerous outcomes, and integrated this model into a webtool for coaches and analysts to investigate on-ice situations of their own design and inputs.

Acknowledgements: Thanks to Ron Yurko, the Stat Sports CMU group, Tyrel Stokes, and Mikael Nahabedian for their feedback on model development and suggestions for possible webtool usages.

References

[1] D. Guszejnov, C. Markey, S. S. R. Offner 1, M. Y. Grudić, C. Faucher-Giguère, A. L. Rosen, and P. F. Hopkins. (2022) Cluster assembly and the origin of mass segregation in the STARFORGE simulations. *Monthly Notices of the Royal Astronomical Society,* under review.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825-2830.

[3] J. and L. Younggren. (2021). A New Expected Goals Model for Predicting Goals in the NHL. evolving-hockey.com/blog/a-new-expected-goals-model-for-predicting-goals-in-the-nhl/

[4] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, CJ Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

[5] R. Ritchie, B. Kumagai, R. Moreau, and E. Cavan. Punt Returns: Using the Math to Find the Path. (2022).
www.kaggle.com/code/robynritchie/punt-returns-using-the-math-to-find-the-path/notebook

[6] R. Stimson. (2018). Tape to Space: Redefining Modern Hockey Tactics.

[7] SciPy Contributors. (n.d.)
docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csgraph.minimum_spanning_tree.html

[8] The Bucketless. (2022). BDC Data. github.com/the-bucketless/bdc/tree/main/data

[9] The Bucketless. (2022). BDC Merge Example.
github.com/the-bucketless/bdc/blob/main/bdc_merge_example.ipynb

[10] Unknown. (2005). commons.wikimedia.org/wiki/File:Minimum_spanning_tree.svg

Appendix - Code

Data and code for the model can be found at https://github.com/nguyenank/bdc22-mst. Code for the webtool can be found at https://github.com/nguyenank/bdc22-mst-website.