A Recurrent Neural Network Model that Quantifies Players' Contributions to Zone Entries and Exits

Christopher Li

Introduction

Most publicly available professional hockey data do not contain pass-level play-by-play information. This limits the extent to which zone entry and exit analyses can be done using public data. Fortunately, the play-by-play data provided for the 2024 Big Data Cup (BDC) does contain information on every pass that is made in a game, which opens the doors to more complex analyses. I use this data from four games between the Canadian and American Women's National Teams to rigorously evaluate players' zone entry and exit abilities.

It is worth noting that the players who are most skilled at zone entries may not necessarily be the players who have the highest volume of zone entries. First off, not all zone entries should be treated equal. For example, some zone entries merely lead to dump-ins and loss of possession, while other zone entries may lead to sustained offensive zone possession and/or shot attempts. My analysis focuses on the ability of players to execute these latter, more valuable zone entries.

Furthermore, merely looking at players' volume of valuable zone entries may not be fully representative of zone entry ability. For example, a player could have a high volume of valuable zone entries predominantly because they play with elite teammates who always set them up for easy carried entries. My goal is to identify players who can not only convert easier zone entry opportunities, but also create their own more difficult zone entry opportunities. To assess players' zone entry prowess, my analysis incorporates the difficulty of each zone entry which is estimated using a recurrent neural network (RNN) model. This allows me to identify players who on average pull off harder entries, which are more valuable to their teams.

Unlike zone entries, zone exits are not explicitly tracked as events in the 2024 BDC data. This leads me to explore a novel way to evaluate players' contributions to defensive zone breakouts and neutral zone play. Ultimately, the goal of defensive zone exits and neutral zone play is to give your team the best chance of entering the offensive zone. Therefore, I define the value of a play to be its contribution towards facilitating an offensive zone entry. I quantify this value as how much a play increases the probability of achieving a valuable zone entry during that possession, which can also be estimated using my RNN model.

Summary of Approach

For this analysis, I define a valuable zone entry as one that leads to the team maintaining puck possession for at least 5 seconds and/or having a shot attempt after the entry. Using the play-by-play data, I develop a model that predicts the probability, after each play in every defensive/neutral zone possession, that a valuable zone entry will result from that possession. The predictions from this model are integral to my downstream player evaluations, regarding both zone entries and exits. For one, I use the probability after the last play before a zone entry as the best estimate of the probability that zone entry is successful. This probability functions as my measure of zone entry difficulty. I also use the outputs of this model to measure how valuable each play is to a breakout/zone exit, based on how much that play increases the probability of having a valuable zone entry (probability of valuable zone entry as of the end of the play minus the probability as of the beginning of the play). This general framework is inspired by StatsBomb's On-Ball Value metric, which measures the value of each event during a soccer match.

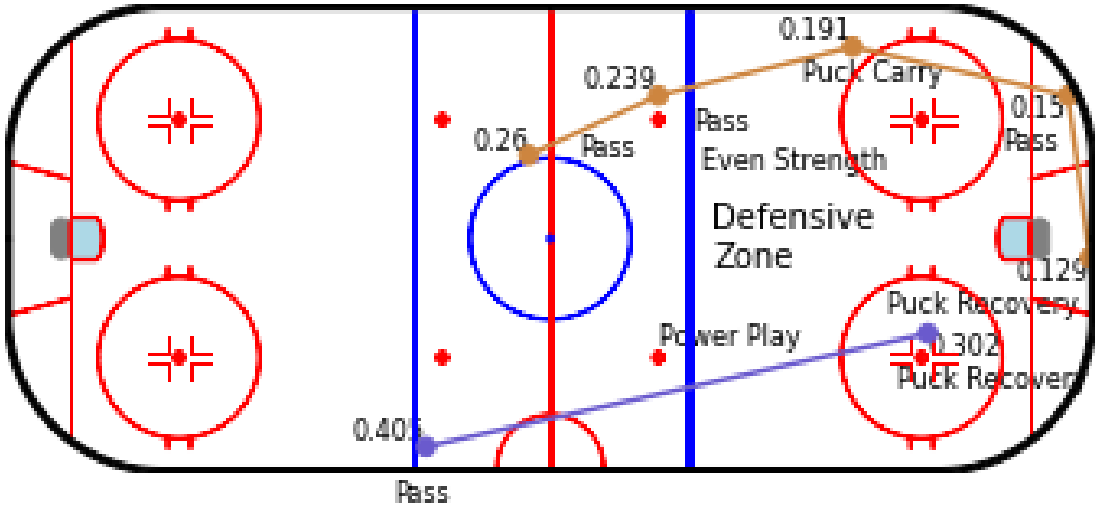
Before training the model, I conduct multiple data wrangling steps which include identifying the set of unique possessions that started in the defensive or neutral zone, flagging which of these possessions led to valuable zone entries, and filtering the data so that only the events before a zone entry were used for the training task. Also, the original play-by-play dataset does not contain data on puck carries, but I feel it is important to incorporate this in my model. For cases when an event's starting coordinate was different from the ending coordinate of the prior event, I use this information to infer when and where a puck carry occurred, as well as which player handled the puck. These puck carries are added to the data as additional events.

The model is then trained on possessions that started in a team's defensive zone or neutral zone, with 70% of the data used to train and 30% of the data used to validate. This data includes possessions that led to a valuable zone entry, possessions that led to a zone entry where neither sustained possession nor a shot attempt were achieved, and possessions where the puck was turned over or dumped in without any zone entry. For this prediction model, I fit a recurrent neural network with one hidden layer (using the Python library Tensorflow) that predicts the probability of having a valuable zone entry after each event in a possession. These predictions are formulated based upon information about all prior events that happened during that possession. This information includes the location, type (Pass, Takeaway, etc.), direction/length (for passes and carries), and situation (5v5, Power Play, or Penalty Kill) of each event. The benefit of using a recurrent neural network architecture is its ability to use sequential data containing sequences of varying lengths to make predictions.

Findings: Model Assessment

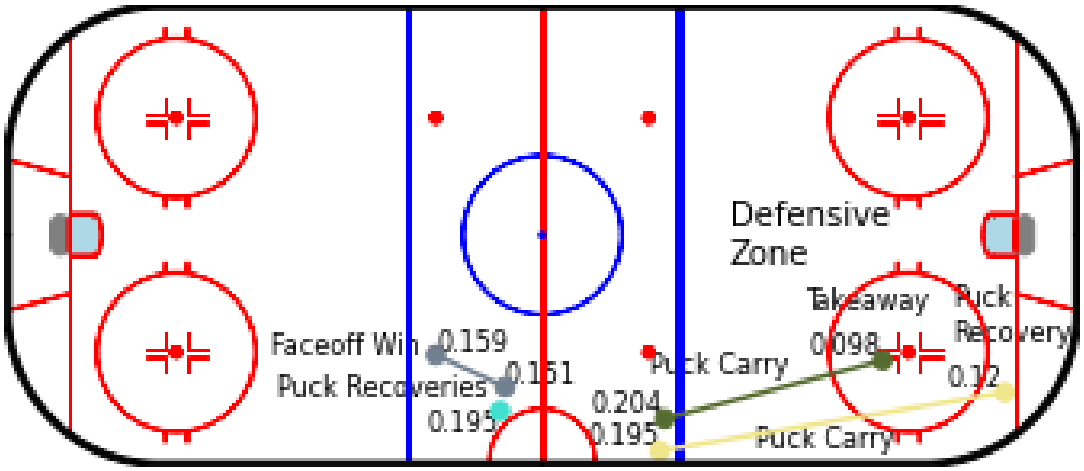
First, I examine the accuracy of my prediction model. It is important that my model produces reliable probability estimates for downstream zone entry and exit analyses. The model achieves an AUC score (area under the ROC curve) of 0.76 on the validation dataset, which confirms it performs well but there is some room for improvement, which could be achieved with a larger training sample of more games.

Next I show some examples to illustrate the outputs of the model and confirm they are consistent with conventional knowledge about what plays are more valuable and how different events impact the probability of zone entries. To start, I present the events that occur in the defensive/neutral zone during two possessions (using the Python library hockey_rink) and examine how the probabilities of valuable zone entry changes.



The possession in orange happens at even strength. It starts with a puck recovery behind the net which is immediately followed by a pass to the corner and then a puck carry up the right side of the defensive zone. Then two consecutive passes are made to get the puck past the red line. In this scenario, monotonically increasing probabilities make sense given the directness and continued forward progress of this breakout. After the final pass that brings the puck near center ice, the model predicts the team has a 26% chance of having a valuable zone entry, based on the prior sequence of plays during this possession. For contrast, I also show a separate possession in purple, which happened during a Power Play.

The probability of having a valuable zone entry as of the time of puck recovery is substantially higher during this Power Play possession (30% vs. 13%), and the probability increases further after one stretch pass. It makes sense that the model's probabilities are much higher during a Power Play, since zone entries are much easier with an extra player. Next I show additional examples of the early stages of four other possessions to highlight the model's ability to account for other nuances.



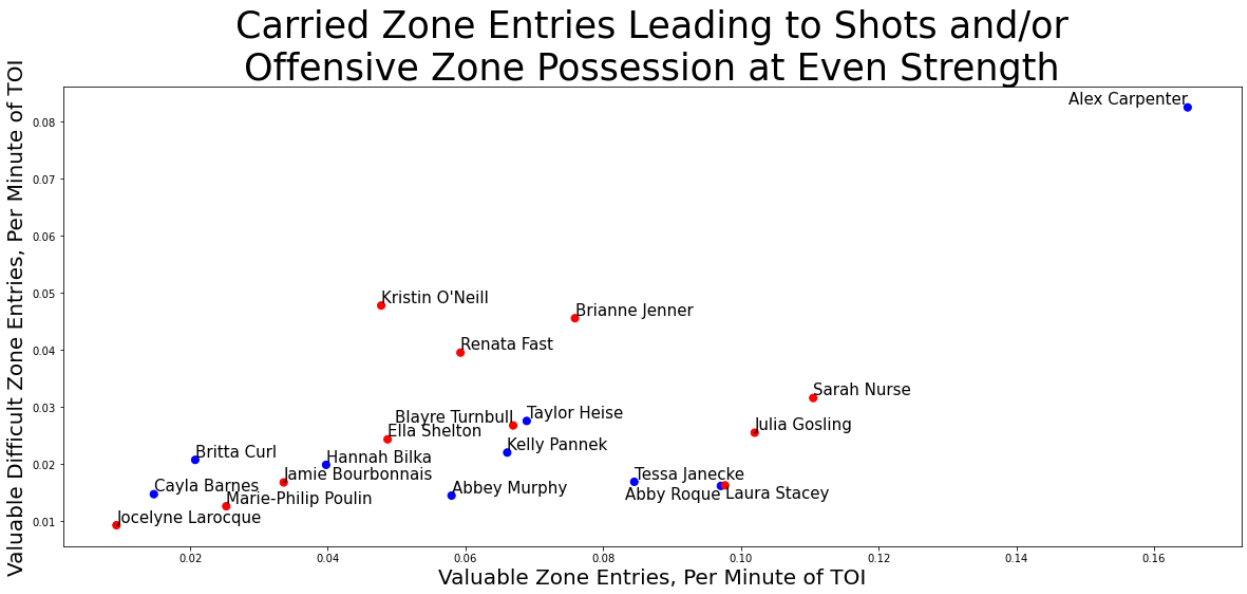
First, I compare the possessions denoted in green and yellow. Possession was gained in the yellow case through a puck recovery while possession was gained in the green case through a takeaway. Despite the puck recovery happening deeper in the defensive zone, its associated probability is higher than the probability after the takeaway. This makes sense because in the scenario of a takeaway, it is likely that the player with the puck is facing more pressure/forecheck from the other team (especially from the opposing team player who lost the puck), which would make the breakout more difficult. In both of these possessions, the puck is carried from the defensive zone up the left side into the neutral zone. The increase in probability from the puck carry after the takeaway is larger than the increase in probability from the puck carry after the puck recovery. A realistic explanation for this is that the puck carry in yellow is on average more valuable because it advances the puck despite probably more defensive pressure following a takeaway. In this way, the model can indirectly account for defensive pressure, which is also shown in the remaining examples.

In the turquoise possession, when the puck is recovered in open play the model estimates a zone entry probability of close to 20%. Meanwhile in the gray possession, when the puck is recovered in a similar spot after a faceoff win, the probability is about 15%. The explanation for this is straightforward. After a faceoff, the defending team likely has all bodies still behind the puck which would make a zone entry harder, while after a puck recovery in open play the defensive team is probably less organized. Now that I have demonstrated the functionality and capability of the prediction model, I proceed to use it for player evaluation.

Findings: Player Evaluation

I evaluate players' carried zone entry ability using two different metrics. The first is a player's number of valuable zone entries across all games included in the data. As a reminder, I define a valuable zone entry as one that leads to a shot attempt and/or maintained puck possession for at least 5 seconds after the entry. The second metric I use is a player's number of zone entries that are both valuable and difficult, where I define a difficult zone entry to be one where at the time the player gains possession of the puck, the probability of a zone entry happening is less than 20%. The difficulty of each zone entry is estimated by my RNN model. I choose the 20% threshold to contain enough difficult entries over the four games of data to be meaningful for analysis, while still maintaining a high level of difficulty. One could also use the average difficulty of a player's valuable zone entries as an alternative measure of difficulty. I choose not to use this since it is more sensitive to outliers, and players should not be punished for executing easier entries.

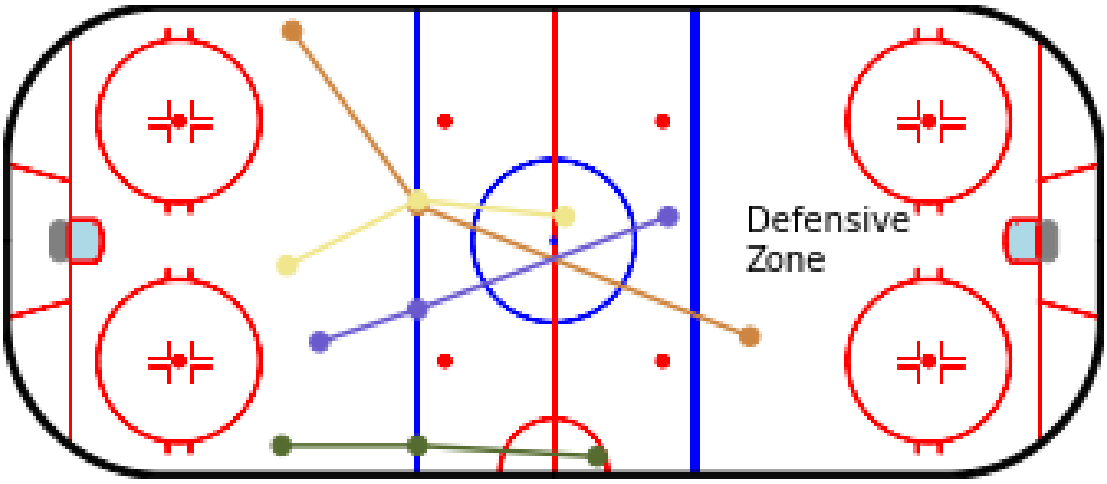
Both metrics are calculated while at even strength and are normalized by minutes played. That way, players who got more ice time and more Power Play time are not given preferential treatment. I chart players on the scatter plot below based on these two metrics, where being higher and farther to the right indicates better performance at zone entries. To ensure the plot is not too crowded, I only plot players who had at least one valuable zone entry and over 30 minutes of total ice time over the four games. Canadian players are represented by red dots and American players are in blue.



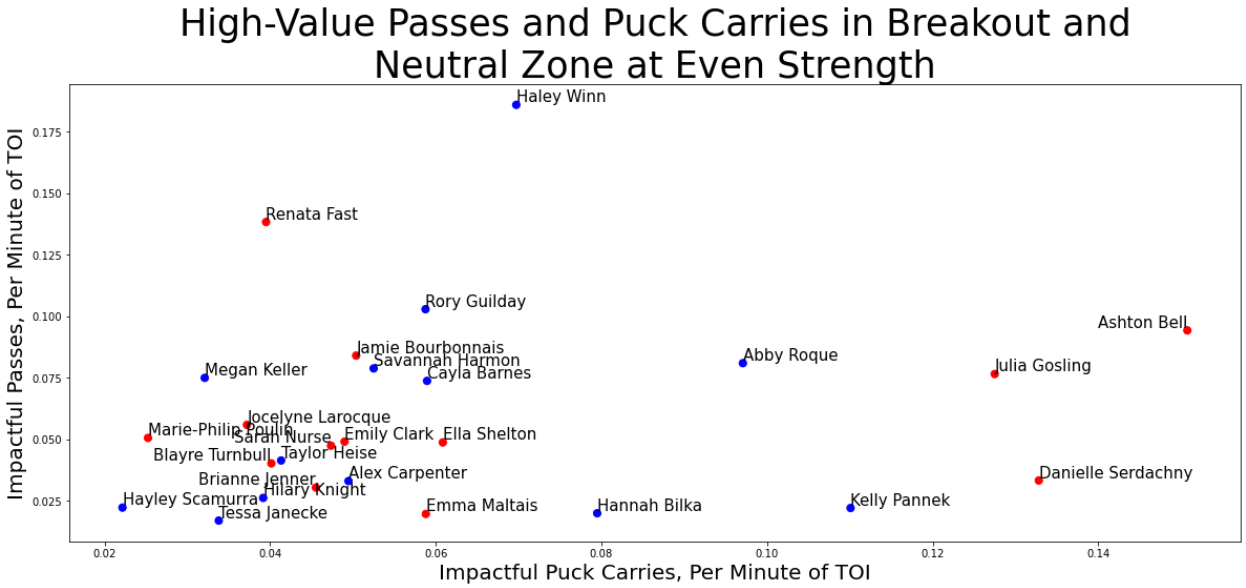
These results demonstrate that identifying the difficulty of zone entries is helpful in understanding which players are the most talented at carried zone entries. For example, some forwards such as Tessa Janecke, Laura Stacey, and Abby Roque perform well at successfully entering the offensive zone. However, these players do not have a high volume of difficult zone entries, which could signify that their zone entry success may be driven by playing with good players or often being positioned near the offensive blue line to receive passes. Other forwards such as O'Neill and Jenner perform well on the difficulty metric despite not having a high volume of entries, and Renata Fast ranks highly on the difficulty metric despite being a defender. These players are likely more skilled at driving their own entry opportunities.

Then we have Alex Carpenter, a clear outlier who has a high volume of both easier and more difficult valuable entries. It would be beneficial for her coaches to design more zone entry plays around her, because of

her outstanding ability to carry the puck into the offensive zone. More generally, when Alex Carpenter is skating forward in the defensive/neutral zone with speed I recommend her teammates feed her the puck. Here is a map of some of Carpenter's more difficult carried offensive zone entries, confirming that Alex Carpenter excels at zone entries when she picks up the puck in her own half of the ice.



Next I analyze players' ability to provide meaningful contributions to zone breakouts and neutral zone plays. As mentioned earlier, I assign the value of a play to be how much that play increases the probability of a valuable zone entry occurring during that possession, using the probabilities outputted by my RNN model. Using this information, I develop two player metrics: impactful puck carries (carries that increase the probability of valuable zone entry by at least 4%) and impactful passes (passes that increase the probability by at least 10%). I determine these thresholds of what constitutes high-value to contain enough instances over the four games of data for meaningful analysis while still maintaining a high standard. Note the threshold that constitutes a high-value pass is much higher than the corresponding high-value puck carry threshold, since passes are more practical and higher-value in a breakout context. If both passing and stick handling are viable options for progressing the puck to a given point, players will usually choose the pass due to efficiency.



This chart highlights players that are strong contributors to zone exits through impactful puck carries and passes. Based on these results, I would recommend that coaches entrust Haley Winn and Renata Fast with more important passes when designing breakout plays and encourage these players to lead zone exits more than their defensive partners. In addition, I would advise coaches to configure plays that give Ashton Bell more open ice in the defensive/neutral zone to carry the puck. I would recommend something similar for Gosling and Serdachny, and I think it is interesting that they perform well on this metric given they are forwards and spend less time in the defensive zone.

Note that all these player evaluations and recommendations should be taken with a grain of salt, given they were only made based on four games of data. Some of these findings may be driven by noise or luck, and ideally a larger sample size should be used to improve the accuracy of my model and validate my conclusions before they influence strategic action.

Conclusion

By training a model to predict the probability of having a valuable zone entry after each play in a possession, I am able to conduct a more in-depth evaluation of players' zone entry and exit talent. This analysis could be leveraged by coaches when determining which players should be the focal point of zone entry and exit plays. Furthermore, coaches could use these types of evaluations on opposing team players to inform defensive strategy. Analytics and management personnel may also be interested in using such a model to make decisions on players to acquire or draft.

While my analysis can infer some information about the level of defensive pressure when estimating zone entry probabilities, it does not capture a complete picture. Since defensive pressure is an important determinant of the difficulty of a zone entry or the value of a breakout pass or puck carry, obtaining access to and incorporating such data would improve the strength of this analysis. It is also important to analyze these metrics in tandem with turnover metrics, since players that often execute difficult zone entries may not be as beneficial to their team if they also frequently turn the puck over while attempting these entries.

Despite these limitations, my model and approach present endless opportunities for further research. On the topic of zone entries and exits, one can dissect the model results further to uncover specific sequences of plays that are effective and underutilized. More generally, my RNN model framework could be applied to offensive zone play-by-play data to estimate the probability of a goal occurring after each play in a possession. This model could facilitate a more nuanced evaluation of players' offensive ability based on how frequently they execute high-value passes or puck carries in the offensive zone.