

**NHL Shot Quality Analysis**

**Project Report**



Garrett Atkinson, Daniel Huang, Ian Pezzella

MSSA 60530 - Human Performance Analytics

Professor. Martin Barron

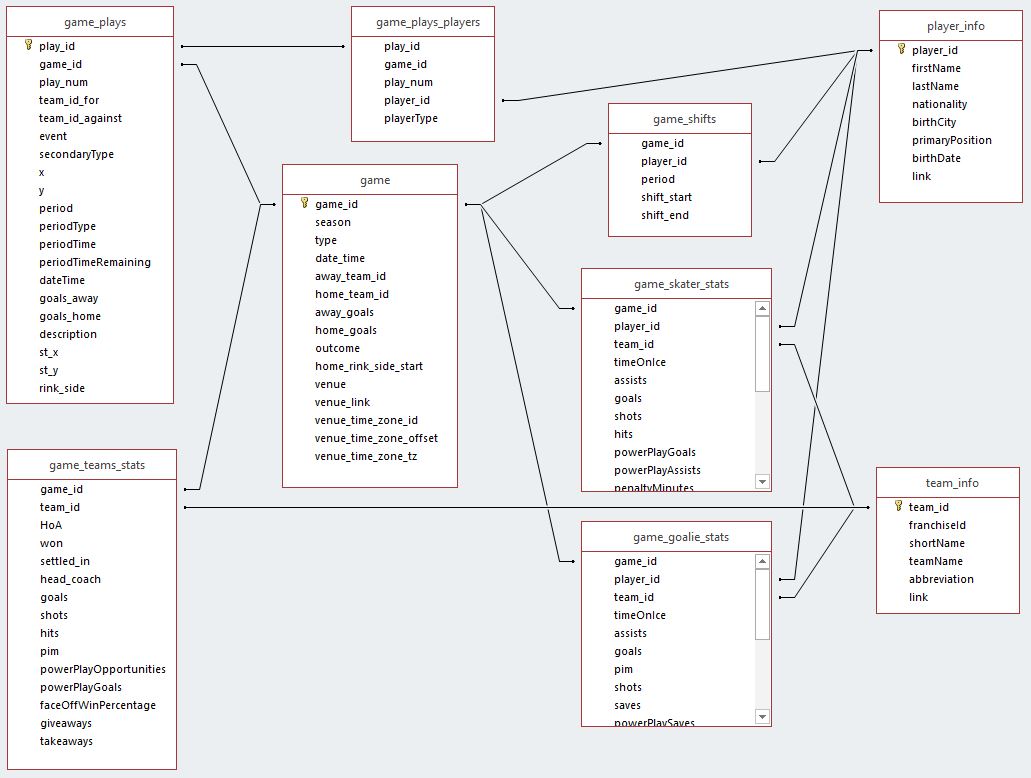
2024 Spring

**Introduction & Related Works**

Since the introduction and success of the Moneyball strategy implemented by Billy Beane and the Oakland Athletics, data analysis has been a hallmark of the sports industry, and its influence has only increased. Twenty years after the initial implementation of Moneyball, we can see sports organizations go well beyond the surface level box-score statistics, and now utilize teams of data analysts and scientists to ensure maximum performance on the field. In both Canada and the United States, hockey has become a beloved and well-followed and scrutinized sport; partially as a result of its popularity, the analytics movement has greatly influenced the sport as well.

Hockey is a simple sport to understand on the surface primarily because of the way outcomes are decided, the team with the most goals wins. In order to better understand scoring and maximize the output of goals scored, people have done numerous studies and analyses on this topic. In August 2023, a Masters of Data Science thesis published by Elliot Barinberg of Ramapo College in New Jersey analyzed player shot quality depending on the situations the player found himself using data from the 2014-2022 NHL seasons and used it to evaluate player quality. The primary result was that Connor McDavid of the Edmonton Oilers was the best player in the world, as he consistently achieved higher shot quality than the average forward in the NHL (Barinberg, 72). The results also highlighted McDavid as a playmaker, as Zach Hyman consistently achieved high shot quality primarily because he shares the same line as McDavid, leading to better scoring opportunities (Barinberg, 79). Another paper published by Gerald Smith of the University of Toronto did a more general analysis of shot quality analysis, using data from the 2014-15 season. Smith concluded that while players like Sidney Crosby and Zach Parise are quality players, they do not produce at a level that commanded their salary during that season (Smith 19). The biggest differentiator between what our group is striving for compared to the previously mentioned studies is that we are going to take a more general approach and attempt to highlight under-the-radar players during those seasons rather than hyper-focusing on specific players like Sidney Crosby and Connor McDavid, and hopefully evaluate skater-on-goalie matchups.

**Data Description**

****

(Graph 1. The relation between different datasets in the database)

The data our group will use comes from Kaggle, and consists of 13 different individual data sets that range from play-by-play data (game\_plays in Graph 1), to descriptive data about the players, to data that reveals whether the player won or lost the game they played in. The play-by-play data contains information from over 5 million plays from games played from the 2016-17 season to 2018-19 season in the National Hockey League and will be the data set that undergoes the most extensive analysis. Examples of play information in the data set include what occurred during the play, the time left in the period during the play’s occurrence, the teams involved in the play, and a general description as to what happened during the play. Other than the game\_plays data, our group joined this play-by-play data with game\_plays\_players and player\_info data because we want to link the shot events to the specific players. We joined these data sets by the primary key: player\_id. Then, we approached the data cleaning by removing the NA values and filtered out events that were either “Goal” or “Shot”. Because the “Shot” event represents a shot made and was saved by the goalie and the “Goal” event represents a shot made and it eventually led to a goal.

In order to prepare our data for modeling, we also created some derived variables from the existing data set. For example, we created 3 variables: 1. Shot Angle, 2. Shot Distance, 3. Goalie saving rate. The shot angle & the shot distance was calculated by the coordinate of the players when they made the shots. These metrics could be useful to evaluate the quality of the shot, especially when a shot has been made in long distance. What’s more the goalie saving rate is a stat to summarize the career saving percentage of the goalie in the “Shot” and “Goal” event. This metric was created as a reference to control for the impact of times when good shots were saved. After that, we split the data into a training data set and a test data set using the 80%/20% rule. To prepare the data for XGBoost modeling, we also utilized the fastDummies package to create dummy variables for the categorical variables in our data set.

**Problem Framing**

Not all shots can lead to goals. When a hockey player takes the shot, there are numerous factors that may impact the chance of score from that shot, such as the types of the shots, the distances to the goal, the time periods of the shots, etc. Our group wants to explore different characteristics of a given shot, and to analyze why some shots result in goals while others don’t. Therefore, the research questions of our project include:

* **What are the main factors that can significantly influence the quality of shots?**
* **Once we identify those main factors influencing the quality of shots, can we construct a predictive model to assess the chance of scoring from a shot?**

In order to answer these questions, our group will analyze the official metrics of Play-by-Play (PBP) NHL data from the 2016-19 seasons. At first, we implemented data visualizations to evaluate the performance of different shot types in NHL. Our group perceived shot type as a crucial factor influencing the quality of a shot. Then, we developed predictive models to predict the chance of a shot leading to a goal using 9 different variables. To approach the predictive models, we selected all the shot events from the cleaned data set, and set the goal variable to be a binary variable containing “0” - not goal or “1” - goal. Then we factorized the remaining variables. In this project, we applied 3 different machine learning methods to our model: Logistic Regression, Random Forest, and XGBoost. When we created the model, we split the data using the 80% training data / 20% validation data rule to avoid over-fitting. The variables we took into account to investigate included:

1. Shot Distance: The distance between the player makes the shot and the goal net
2. Shot Angle: The angle of the player’s shot
3. Skaters Positions: The specific position of the players who makes the shot
4. Shoots Catches: Whether the player shots the puck on left side or right side
5. Skaters Age: The age of the players
6. Period Time Remaining: The time remaining in a period
7. Period: The number of periods of the game
8. Goalie Saving Rate: The career saving rate of the goalie
9. Shot Type:
10. Wrap
11. Tip-in
12. Backhand
13. Snap
14. Wrist
15. Slap
16. Deflected

**Methods**

Due to our problem being classification-based, we decided that using both Random Forest and XGBoost in addition to a logistic regression model would be ideal methods to solve the presented problem. Both situations required splitting the data into training and test data; using the set.seed function and random sampling input, about 20% of the data was used for test data and 80% of the data was used for training data. The split and the samples that were sorted into each category would remain the same throughout the project.

We did the logistic regression model first primarily to serve as a baseline for analysis; “event” (goal or no goal) was the response variable and all other variables were used as predictors. Once that was completed, we noted both the accuracy and balanced accuracy metrics to serve as a benchmark for the remainder of the analysis. The Random Forest Model was done using a simple bagging model. During this process, we decided to weight the events outcomes on a 10:1 scale (10 benefitting goals) due to the “goal” outcome only possessing a value about a tenth as large as the “no goal” outcome; this was primarily done to increase the balanced accuracy of the model. Unfortunately, due to the size of the data and the time-constraints we were presented with, optimally tuning the data would have been far too time-consuming, as the initial bagging model took well over 30 minutes to run on R. Despite these difficulties, we decided to take note of the model’s accuracy and balanced accuracy metrics and include the details about the lack of tuning as an important side-note. After the bagging model was run, we took note of the most important variables according to the model. Different methods were used during the XGBoost process. Prior to any deep analysis with the XGBoost, the test and training data had to be converted to matrices. Once that was completed, we could train the XGBoost model with the training data and get the accuracy and balanced accuracy for it. Afterwards, the XGBoost model underwent an extensive tuning process where the optimal value for max depth, minimum child size, eta, gamma, subsample, and column space were identified and applied to the final model, where the accuracy and balanced accuracy were found for that too and the variables were ranked in terms of importance.

**Results**

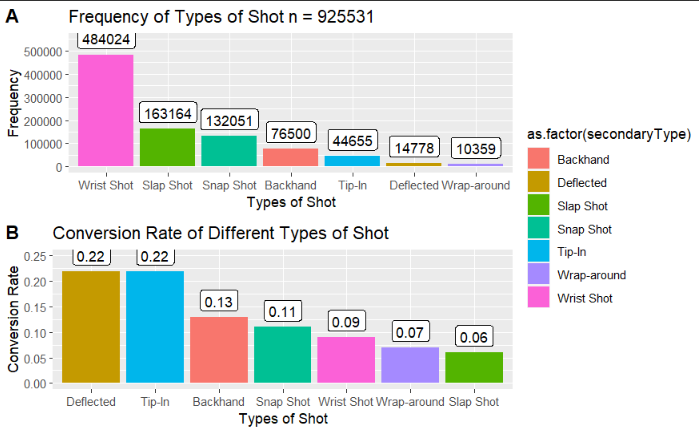
The accuracy for the logistic regression model using “events” as predictors was about 80%, which may initially seem high; however, this is mainly due to the extreme imbalance between the number of goals and normal shots. When looking at the balanced accuracy, it is at a much lower 56%. Due to this low measure for balanced accuracy, this simple model could not be used for deeper analysis and finding the most important variables, so further work was done. The bagging model with weights assigned to goals and non-goals resulted in much higher metrics, which marked a massive increase in model effectiveness. The bagging model possessed metrics of about 94% accuracy and 96% balanced accuracy; both the closeness of the accuracy and balanced accuracy metrics and the high values overall can be explained by the weights, as it allowed each different outcome to be far more equal than they would be in its normal form. When looking at variable importance, the rankings showcased “Period Time Remaining” and “Shot Distance” to be the two most important variables when it came to predicting whether a shot will result in a goal; “Skater Age” was decisively the third most important variable, with “Goalie Saving Rate” and “Shot Angle” ranking as the fourth and fifth most important variables respectively in terms of predicting whether a shot will be a goal or not.

Once the Random Forest analysis was complete and the training and test data were converted into matrices, the tuning process for the XGBoost model could begin. The first parameters tuned were maximum depth and minimum child weight. According to the data frame that shows each different combination of the two parameters and the resulting AUC and error measures, the combination that resulted in the least amount of error was a max depth equaling 15 and a minimum child weight of 1 due to the AUC values being fairly negligible and the error measures being decisively lower than the other combinations. The gamma value was then tuned next, where the data frame shows that a gamma value of 0.00 results in the largest AUC measurement. Column space and subsample were the next two parameters tuned, where the data frame showed the combination of subsample equaling 1.0 and column space equaling 0.6, which possessed the most optimal combination of high AUC and low error. Eta was the final parameter measured, where five models, each consisting of a different eta parameter (0.3, 0.1, 0.05, 0.01, 0.005) were created and then plotted to see which eta value consistently has the lowest error measure. According to the graph, the eta value of 0.1 consistently possessed the lowest error measure, meaning it would result in the lowest error for the model. With all the parameters properly tuned, the accuracy and balanced accuracy for the model was extracted and the variables were ranked in terms of importance. The accuracy metric measured around 81-82%, while the balanced accuracy measured at around 61-62% making it a better model than the logistic regression model. The variable importance graph showed a similar output in terms of the five most important variables, albeit in a different order. “Shot Distance” became the most important variable, with “Shot Angle” and “Period Time Remaining” ranking as second and third respectively; “Goalie Saving Percentage” and “Skater Age” respectively ranking at fourth and fifth.

**Actions**

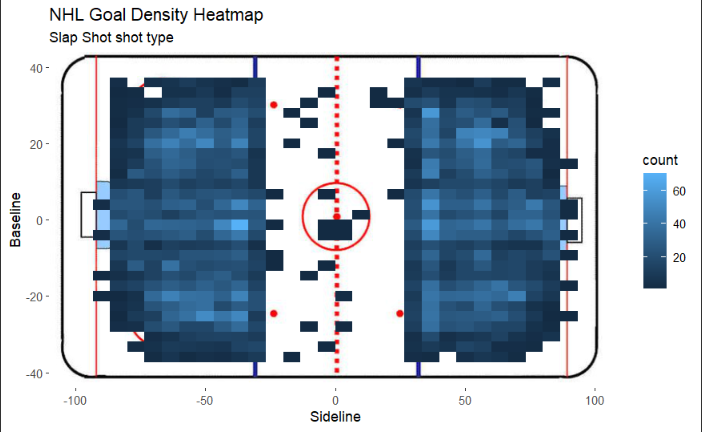
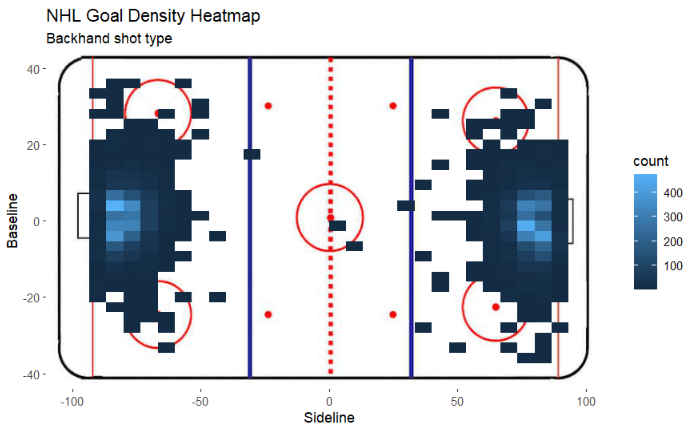
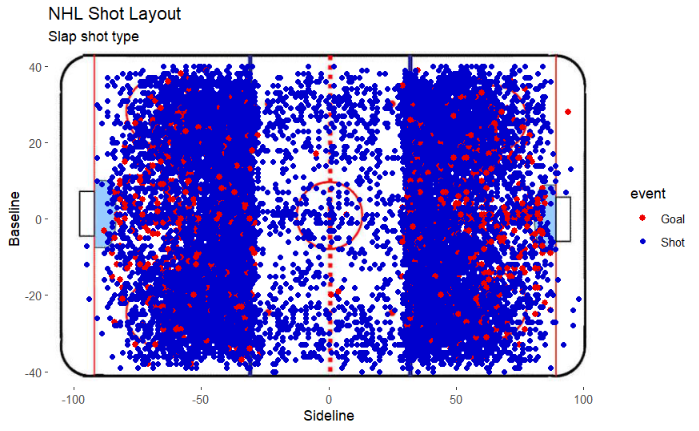
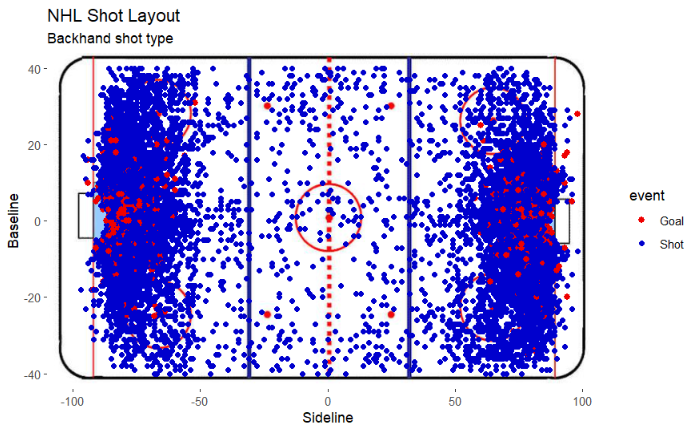
After our analysis, we discovered some important insights to determine the quality of shots, and our group drafted a few strategies for NHL teams and players to enhance their shot performance.

First of all, the visualizations created by our group revealed some important findings regarding the types of shot in NHL. Our group utilized bar charts to illustrate the frequency of each type of shot as well as their corresponding conversion rate. We discovered that the wrist shots are the most common type of shot in all NHL games and the conversation rate for Deflected and Tip-in shots are the highest among all different types. When we combined these two graphs, we found that wrist shots generated the most number of goals. Meanwhile, the snap shot was a great shot strategy with relatively high frequency and high conversion rate. Therefore, the first strategy we proposed is that NHL teams should encourage more Wrist Shot & Snap Shot based on their high frequency & high conversion rates.



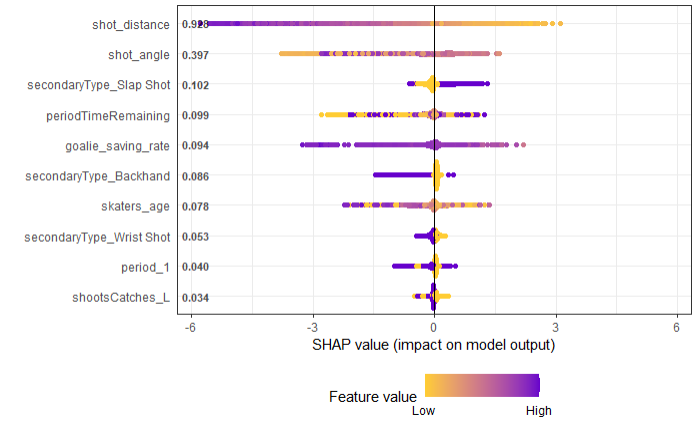
(Graph 2. Shot Types Frequency & Conversion Rate)

What’s more, in order to further analyze the types of shot, our group visualized the location and goal density of each type of shot. With these types of visualizations, our group was able to examine the location of those shots and identify specific features related to the types of shot. Based on our visualizations, our group concluded that the Backhand shots are more effective in short distances and the slap shots are the most effective in long range. Therefore, the second strategy our group recommended is that the players should utilize Backhand shots in short distances and Slap shots in long distances.



(Graph 3. Locations and Goal Density for Slap Shot & Backhand Shot)

Last but not least, our group successfully found out the factors that have the most significant impact on the quality of a shot based on our models. Our group generated a great model using the XGBoost method. We also created a SHAP graph to visualize the impact of different factors on the shot quality. Based on the SHAP graph, we concluded that shot\_distance, shot\_angle, and Slap Shot shot type are the top 3 factors impacting the shot quality. Prominently, the distance of the shot from the goal plays a vital role; shots taken closer to the net have a significantly higher chance of scoring, primarily due to the reduced reaction time available for the goalie and fewer physical barriers obstructing the puck’s path to the net. The angle of the shot is also crucial, as shots taken from sharper angles tend to decrease the size of the visible net to the shooter, whereas shots from more direct angles increase scoring likelihood by offering more target area. Slap shot is an interesting variable, as you utilize more slap shots as your main shot strategies, it is more likely for you to score a goal.



(Graph 4. SHAP Value of Different Variables in The Model)

**Conclusion & Future Work**

In this study, we have successfully harnessed the power of advanced analytics to uncover key factors that influence the probability of shots resulting in goals in NHL games. We discovered that how far the shot is taken from the goal, the angle of the shot, and the type of shot, especially wrist and snap shots, are very important for scoring goals. The XGBoost model we developed was very effective in predicting whether a shot would result in a goal, performing much better than the simpler models we started with.

This research helps to show how data can be used in sports to help coaches and players improve their techniques and strategies. By understanding which types of shots are more likely to score, teams can train more effectively and play smarter. Our charts and models clearly show how different factors, like shot distance and type, work together to influence whether a shot becomes a goal.

For future work, there are several avenues to explore. Firstly, expanding the dataset to include more recent seasons could provide a more comprehensive view of trends and changes in player performance and strategy. Additionally, integrating player fitness and psychological data could offer a deeper understanding of the variables affecting shot outcomes. For example, we could use data generated from wearable devices like catapults to gather more information on athletes’ performance. Finally, further exploration into less common shot types and their situational effectiveness could uncover additional strategies for scoring optimization. These advancements could enhance this project, resulting in more precise and effective strategies to improve shot quality in professional hockey games.

**Bibliography**:

Barinberg, Elliott. “Evaluating How NHL Player Shot Selection Impacts Even-Strength Goal Output Over the Course of a Full Season.” *Ramapo.EDU*, Aug. 2023, www.ramapo.edu/dmc/wp-content/uploads/sites/361/2023/05/MSDS-Byman.pdf.

Smith, Gerald Edward. “A Shot Quality Adjusted Plus-Minus for the NHL.” *Summit Research Repository*, 19 Dec. 2016, summit.sfu.ca/item/16952.