Predicting Goal Scoring in the NHL

A Project submitted in partial fulfillment of the requirements for OR/SYST 568 at George Mason University

by

Muhammad Hassan

Instructor: Dr. Jie Xu

Spring Semester 2022
George Mason University
Fairfax, VA

ABSTRACT

Big Data Analytics is being used by national sports teams to better understand each sport. The analytics is used to evaluate a team's performance, each player, the plays run, and their results. This analysis is being performed by the teams, the leagues in which they play, betting firms, sportscasters, and the sport's fans. This paper performs big data analysis on the National Hockey League (NHL) 2007-2020 seasons and determines the highest percentage of successful shots on goal and those shots that are not as successful. The paper also provides a predictive model by which a shot's success factors are determined..

INTRODUCTION

The National Hockey League (NHL) is a professional sport played by 31 teams in two conferences across Canada and the United States that ends each season by competing to win the Stanley Cup [1].

"In hockey, five players and a goalie per team are on an ice surface and play for a total of 60 minutes. The goal of the game is to put a rubber puck into the opposing team's net using a 1.5 to 2m long stick made of wood or a composite material. The team who scores the most goals in a game is the winner. In the regular season, if a game is tied after 60 minutes, the teams play an extra 5 minutes of sudden death overtime and after that the game is decided by a shootout. In the playoffs, after 60 minutes, additional 20-minute overtime periods are played until a team scores." [2]

This research paper uses a big data analytics approach to understand the scoring characteristics of NHL games. It seeks to identify factors that impact scoring, including factors that could potentially improve the odds of making a goal as well as factors that would prevent shots on goal. The research will examine those discovered factors to identify winning hockey tactics and strategies to making goals.

The research paper's dataset [3] spans the 2007-2020 seasons and records shots taken and their captured attributes during those seasons. The dataset comprises over 1.4 million records and each record has more than 100 variables. This paper takes a data-centered approach to winning hockey tactics and strategies – it will determine what variables play a major role in scoring a goal and which less so. Using the hockey shot data set, the paper will answer the following research questions:

- Places on the ice that scoring shots are most and least likely to come from
- Types of shots are the most and least successful (wrist vs. slap vs. backhand vs. tip-in vs. snapshot)
- Examine if there are significant differences in the scoring patterns in playoff hockey vs. regular season hockey.

Benefits of exploring the NHL shot dataset include: 1) NHL fans will gain a better understanding of the game, 2) hockey coaches can develop strategies with their team given various scenarios on the ice, and 3) sports gamblers can make more informed bets.

.

Related Work

Major league sports are a major business. The NHL, prior to COVID-19, by itself has brought in over $5 billion in annual profit [4]. This leads to high interest in evaluating the sport and providing opportunities to measure NHL performance. Analytical research discovered for this paper spans players, their rankings, their tiers, and team performance.

Naples, Gage and Nussbaum sought to improve the measure of NHL goalie evaluation. Their premise was that goalie performance has historically been measured by "save percentage"

or number of shots on goal blocked divided by the total number of shots on goal. The authors suggest this is a weak measure of goalie performance and statistically verified this suggestion. The authors alternatively considered a range of other variables such as controlling shots per their "Clean" shot formulation, calculating performance against expected goals and by using the "expected goals" metric. Controlling for these variables yielded a better measure of goalie consistency and predictability yielding a correlation value of 0.3123 [4]. The authors highlighted that their results provide clues but are not robust and do not isolate goalie performance from team effects [5].

Pischedda examined NHL match outcomes and researched how machine learning (ML) algorithms may be used to predict those outcomes. Pischedda research was to understand the predictive power of two types of data categories (categorical versus continuous) in the data set used and study the accuracy of the models as applied to betting [6]. His research explored three machine learning techniques: Decision Trees (DT), Multi-Layer Artificial Neural Networks (ANN), and ClusteR. ClusteR is a proprietary software used by a betting company for various sports. ClusteR combines k-nearest neighbor with rule-findings previously analyzed by Weissbock [7]. Pischedda discovered that the categorical variables "Team" and "Location" improved the predictive accuracy of the models – regardless of the machine learning technique used and other factors in the analysis [6]. The research showed that using a team's name and their location increased a team's predicted match result for their next match [6].

Other research examined statistically the rating of NHL forwards and defensemen using their on-ice events. Schuckers and Curro developed a comprehensive measure of player evaluation for NHL forwards and defensemen [7]. Their methodology quantifies the impact of players in preventing and scoring goals by determining the probability that a given play results in a goal. The methodology used considers a range of variables including both shooting and non-shooting events (e.g. turnovers), all players on the ice, and shift starts. The research used a probability approach as NHL games have low scores. Controlling for these variables permit a better measure of NHL forwards and defensemen player effectiveness in preventing and scoring goals [7]. The ranking methodology incorporated shots on goal, shot locations, shot types and a player's movements to provide a rating. Schuckers did not seek to predict a specific shots success – just a player's ranking.

Similarly, Lehmus, Kozlica, Carlsson and Lambrix sought to identify skills that would predict an individual's rank within the NHL by using six machine learning algorithms. The authors discovered that Bayesian classifiers performed the best and had the best sensitivity [8]. They discovered that the models worked best by classifying forwards (versus defensemen or goalies) and identified that game official statistics are dominated by forwards [8]. Like Schuckers, Lehmus did not attempt to predict shot results or what types and locations made a shot more likely to result in a goal.

Dataset

This research uses a dataset containing a record of every shot taken in every NHL game from 2007-2020 seasons. The dataset contains about 1.4 million records and each record has 124 predictor variables that are a combination of numeric, categorical, and ordinal variables. We also have probability variables included in the dataset. These predictor variables will be essential in predicting the shot results.

There are exactly 108 potential useful predictors that wasl be taken into consideration when determining the shot results. Some of the useful predictors are: shot angle, distances, number of

players on the ice, average time since face off and previous shot results. This in turn determines the total number of shots that a team has accumulated over the course of the game. We can also find the total number of shots of the opposing team to determine the team with the most number of shots which will be the winning team. We filtered the dataset to include only shots from 2018 - 2021 seasons. This included a total of 300,406 records.

The dataset was sourced from moneypuck.com directly [2]. Moneypuck is a sports-centered database, especially for NHL. Moneypuck acknowledges that data has been collected from several sources to include the National Hockey League and ESPN. Moneypuck makes no guarantees as to the quality of the data. They state "NHL shot data is known to have issues and biases" [2].

A sample of the predictor variables in the dataset is contained in Table 1. Moneypuck provides a full dictionary of the Moneypuck shot dataset and Appendix A provides a listing of each attribute and its corresponding data type.

Data Preprocessing

The NHL dataset we used for our Predictive Models contains records from 2018 to 2020. The dimensions of the dataset included 124 variables(including the response variable) and 300,405 records/ observations. To begin our preprocessing we handled the missing data, this was done by identifying which predictors had missing data and we determined it was appropriate to delete those records from the dataset based on the variable type and the percentage of missing data. In the adjusted dataset we dropped several probability predictor variables that were not valid for our analysis and other proxy variables. We then encoded the categorical data using dummy variables; predictors such as teamHome, shotTypeBack, shotTypeSnap, shotTypeTip etc. Our next step was converting the IDs to categorical data, having converted the IDs and encoding the categorical data we converted the list to a vector using the "*unlist"* function in R. We then created a subset which was the "goal" variable – the variable of interest.

	shotID	homeTeamCode	awayTeamCode	season	isPlayoffGame	game_id	homeTeamWon	id	time	timeUntilNextEvent	...	xFroze	xRebound	xPlay
0	0	PHI	PIT	2020	0	20001	1	6	16	7	...	0.157823	0.039270	
1	1	PHI	PIT	2020	0	20001	1	9	34	17	...	0.229722	0.033520	
2	2	PHI	PIT	2020	0	20001	1	12	65	0	...	0.342825	0.029958	
3	3	PHI	PIT	2020	0	20001	1	24	171	0	...	0.209750	0.027594	
4	4	PHI	PIT	2020	0	20001	1	27	209	14	...	0.175148	0.033828	

5 rows × 124 columns

Fig. 1 Sample of the Original dataset

	season	isPlayoffGame	homeTeamWon	time	timeSinceLastEvent	period	shotPlayContinuedOutsideZone	shotPlayContinuedInZone	shotGoalieFroze	sh
0	2020	0	1	16	16	1	0	1	0	
1	2020	0	1	34	6	1	1	0	0	
2	2020	0	1	65	2	1	0	0	1	
3	2020	0	1	171	42	1	0	0	1	
4	2020	0	1	209	38	1	0	1	0	

5 rows × 91 columns

Fig. 2 Sample of the cleaned dataset

The next step was filtering out the predictors with low frequencies using the nearZeroVar function. This left our dataset with 98 predictors remaining. We then created a sample of our filtered and unfiltered dataset using 8 predictors, we looked at the distribution of this sample and identified skewness in the distribution. A center and scale transformation on the sample and full data was applied to the sample and full data to address the skewness in the filtered and unfiltered dataset. Next was identifying the correlation among predictors in the original and transformed filtered dataset. The correlation plot below shows how much correlation exists among the 98 predictors. We applied a 0.95 correlation cutoff in the dataset and filtered out the predictors that exceeded this cutoff value from the dataset. Our cleaned up dataset had a dimension of 91 variables (goal subset included) and 300,359 records.

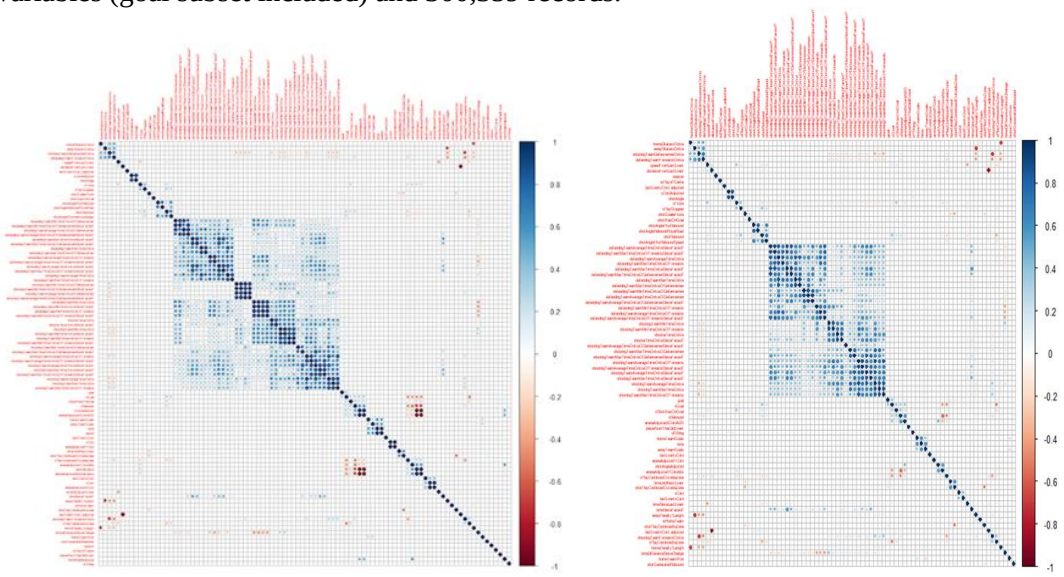

Fig. 3 The images above are the correlation matrix of the predictors.We removed predictors that had pairwise correlations greater than 0.9. Dark blue colors indicate strong positive correlations, dark red is used for strong negative correlations, and white implies no empirical relationship between the predictors. The preprocessing step reduced the number of predictors from 112 to 96.

Data Exploration

The NHL has standardized NHL rink and an example of one is provided by O'Halloran [9]. Figure 1 provides a diagram of a standard NHL rink and the key dimensions found within. There are several key points to make about Figure 1 to inform the reader. There is 89 feet (100 feet – 11 feet) from the center red line to each red line at a goal. When shot positions are captured, shot positions "X" coordinates may range from -100 feet to 100 feet. A shot's "Y" coordinate is measured from the center position of the rink which is 85 feet wide resulting in a range from -42.5 feet to 42.5 feet with the center of the rink being 0 (zero) feet.

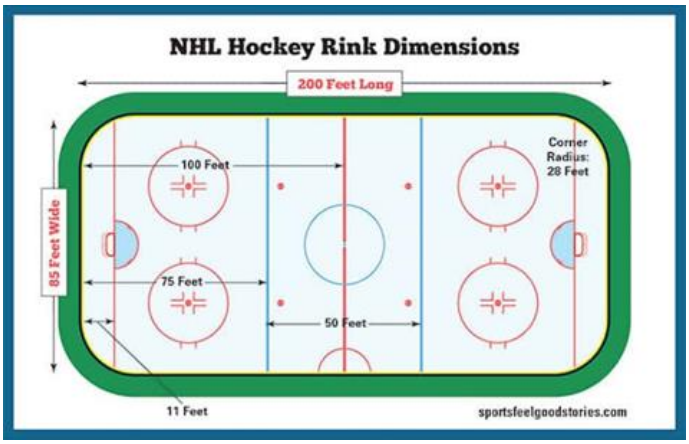

Fig. 4 Standard NHL Rink Dimensions.

Given the dimensionality of the dataset and the extensive preprocessing that was done, we were able to determine several predictors of interest for our analysis. The distribution of the nominal and numerical predictors of interest were analyzed using different exploratory techniques. Some of the variables of interest which we performed exploratory analysis techniques on are: Goals [0,1], Event [Goal, Miss, Shot], Shot Coordinates X [0,100 ft], Shot Coordinates Y [-44,44 ft], Shot Distance [0,100 ft], Shot Angle [-90, 90 degrees], Shot Type [BACK, DEFL, SLAP, SNAP, TIP, WRAP, WRIST], Shooter Handedness [L, R], Home or Away [HOME, AWAY], Season [2018, 2019, 2020], Playoff Game [0,1].

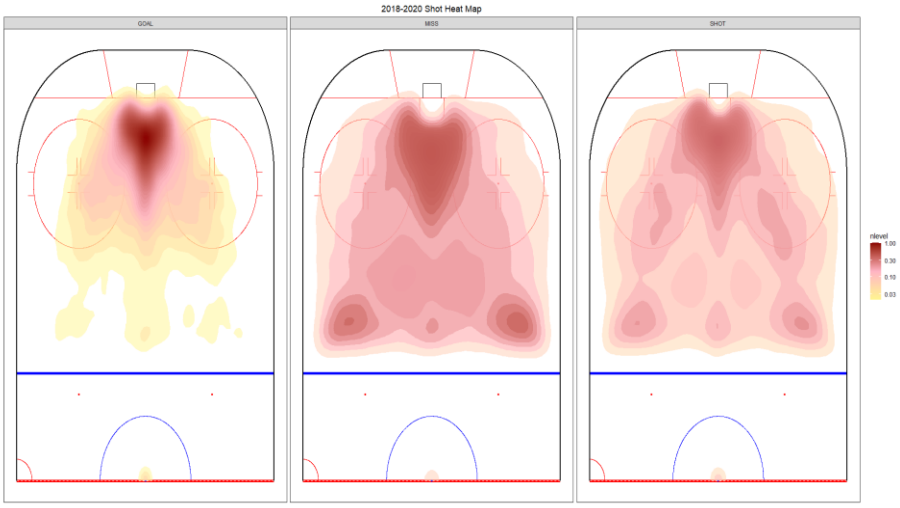

Fig. 5 This is a heat map of all shot attempts by event. Again, shot on goal (SHOT), goal (GOAL), or missed the net (MISS). So we start to get a pretty good idea of what factors contribute to a good vs bad scoring opportunity. From a distance perspective, we confirm that most shots overall and most goals occur close to the net. We also confirm the shot angle to be most successful at 0 degrees and within a range +/- 45 degrees.

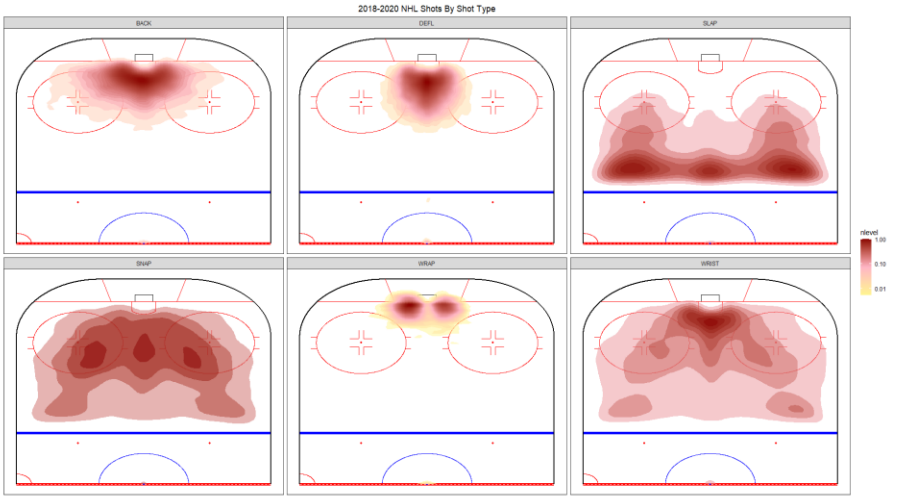

Fig. 6 Another heat map here of the 6 shot types for all shots. The missing data not relevant here, so it was excluded. Also, the tip shot and deflection have been combined as they are essentially the same shot. So we notice some pretty interesting patterns here. 1) wide range of backhand shot attempts. 2) deflections are concentrated right in front of the net, 3) slapshots generally taken a distance from the net in front of the blue line with pockets on concentration at 0 and +/- 45 degrees. 4) Snap shots have a broad range of distance and angle with pockets of concentration in the faceoff circles and straight on about 20 ft or so from the goal. 5) wrap shots focused on either side of the crossfire of the net. 6) and the host volume shot, the wrist shot is another shot taken from almost anywhere, but concentrated close to the net. It's interesting to see overall that each shot type has very distinct hot zones.

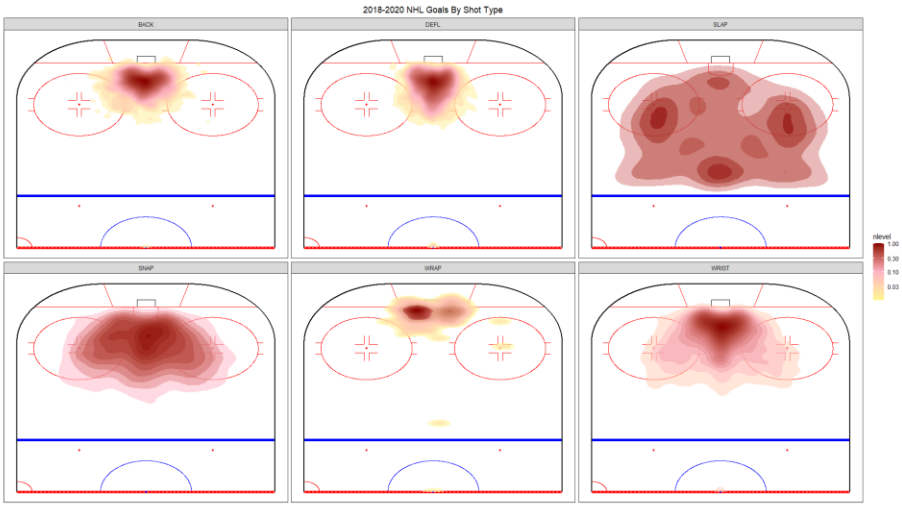

Fig. 7 This is the same chart but filtered for goals only. Here, backhand, snap, wrist, and deflections are pretty similar with a more narrow focus by the net. The slapshot while attempted across the blue line is finding most success in the faceoff circles and in front of the blue line straight on 0 degrees. Wrap shot we see a clear concentration of goal success on the left side of the net. My guess here is that most goalies are right-handed and this shot would end up on the goalie stick side as opposed to the glove side. The stick stick side is generally accepted to be the weaker side for a goalie. This looks to bear out here.

Predictive Models

Once the data was preprocessed we used linear, non-linear, and tree models to classify a two class scoring opportunity problem, a three class event problem and a two class gambling problem. Not all models were computationally possible given our teams' resources. To run more classification methods we tried to run the data from the 2018-2020 season and just the 2020 season. Computing constraints were persistent.

Goal/No Goal Problem

For the two class scoring problem, we were trying to assess whether or not a shot results in a goal. Figure 8 shows the results of the linear classification. With an ROC and AUC greater than .9 we can conclude that the model has significant predictive accuracy. There may be some closely related predictors in this data set given the high degree of accuracy. Figure 9 gives the variable importance plot where shotGoalieFroze has particular importance in classifying the data. ShotGoalieFroze represents a shot where the goalie freezes with the puck within 1 second of the shot. This makes sense as an important variable since every shot with shotGaolieFroze value equal to 1 will not be a goal. For the nonlinear classification models, nonlinear discriminant analysis produced an ROC similar to the logistic regression (Fig. 11). Kappa and accuracy were slightly lower than the logistic regression. For Tree-based models (Fig.12) the bagged tree had similar predictive results as the logistic regression and NDA. Overall the Linear classification model outperformed the Nonlinear and Tree models. This can be measured in terms of computational feasibility, ability to converge to a feasible solution, and providing high predictive ability. Thus, we can infer the data has linear class boundaries based on the outcomes each model has produced. In the Goal/No Goal classification, the five most prolific predictors were shotGoalieFroze, homePenalty1Length, offWing, shotAngleAdjusted, arenaAdjustedYCordsAbs

Event Classification Problem

Given the goal/no revealed the potential predictive power of the resulting shot type, being able to predict the result of a given shot could inform shot selection by identifying the traits of a shot not likely to miss. In this case, for the linear classification methods(Fig 13), partial least squares had the highest AUC, while LDA had the highest accuracy. Naive Bayes was the only nonlinear model we ran for this model and it did not have any predictive accuracy. For our tree models(Fig.15), we found the basic classification tree had similar accuracy results to the LDA and PLS models, with a slightly lower AUC. Just as it was in the Goal/No Goal classification, the Linear models overall outperformed the nonlinear and tree based models in terms of computational feasibility, ability to converge to a feasible solution and providing high predictive ability, we could also infer that the data has linear class boundaries. Based on the variable importance plot we see that the five most prolific predictors were shotPlayContinuedOutsideZone, shotDistance, shotPlayContinuedinZone, arenaAdjustedXCordABS, arenaAdjustedYCordsAbs.

The three predictors that were common in both classification problems are: shotGoalieFroze, arenaAdjustedYCordsAbs, shotAngleAdjusted - with arenaAdjustedYCordsAbs being in the top five of both classification problems. We believe these predictors are the most important in determining whether or not a shot type results in a goal and the result of a given shot and it would be in line with what a hockey expert would deem important.

Gambling Classification Problem

The shot data also held potential relevance for gamblers so we tried to create a model that predicted a winning bet. This involved extensive preprocessing to combine the moneypuck data set with a dataset from sportsbookreviewsonline.com. This data set had point spreads for all regular season games played in the NHL from 2018-2020. We aggregated the shots data so it was grouped by game and the values were either summed or averaged depending on the attribute and its distribution. We then created a key between the game, team and season in each data set and matched the shots data with the gambling data. For the problem we were looking to classify whether or not betting the points spread on the Montreal Canadiens would yield a winning bet. We ran a logistic regression, random forest and supported vector machine to predict the winning bets. When including the potential bet winning data in the model the accuracy of the model was much higher. The sportsbooks make predictions about what will happen in the game and determine how much money a bettor will make from a points spread bet. The earnings amounts being a significant predictor makes sense given the sportsbooks motivation to make accurate predictions. The plus 100 and 200 winnings are more likely to produce losing bets than minus 200. Logistic regression and random forest had the same performance metrics. This accuracy could be potentially useful to a gambler if the data was gathered for all teams in the NHL. Earnings, goal difference and goals were the three most important predictors in the earnings data included model. In the model without earnings data, average goals and avgMaxTimeOnIceDefenceman were the strongest predictors. In hockey, if defenders are on the ice more, they are not committing penalties and thus do not have to play a man down as often as other teams.

Results

Linear Classification											
			Training			Test		Confusion Matrix			
	Models	Tuning Parameter	ROC	Sens	Spec	Accuracy	Kappa		Goal	NoGoal	AUC
1	Logistic Regression		0.996292	0.9779004	0.9870448	0.9853174	0.8935715	Goal	4034	787	0.9956
								NoGoal	95	55155	
2	Linear Discriminant Analysis		0.9959637	0.9997626	0.9732429	0.9741306	0.8278885	Goal	4128	1553	0.9953
								NoGoal	1	54389	
3	Partial Least Squares Discriminant Analysis	ncomp = 10	0.7323	0	0.9999955	0.9313	0	Goal	0	0	0.7294
								NoGoal	4129	559423	
4	glmnet	Alpha = 0.1 Lambda = 0.01	0.996151	0.7457	0.99346	0.9764	0.8006	Goal	3079	366	0.9955
								NoGoal	1050	55576	
5	Sparse logistic regression	Lambda = 0.1 NumVars = 50	0.9959527	0.99976	0.97224	0.9741	0.8279	Goal	4128	1553	0.9953
								NoGoal	1	54389	
6	Nearest Shrunken Centroids	threshold = 6.896552	0.7165885	0	1	0.9313	0	Goal	0	0	0.7119
								NoGoal	4129	559423	

Fig. 8 *Performance of the linear classification models for the goal/no goal problem*. The best optimal linear performance is Logistic Regression with training accuracy of 0.996 and the best linear predictive ability is Logistic Regression with test accuracy of 0.985 and AUC of 0.9956. However, we did receive a warning "glm.fit: algorithm did not converge" As I understand, this warning often occurs when there is perfect separation – the predictor variable is able to perfectly separate the response variable into 0's and 1's.

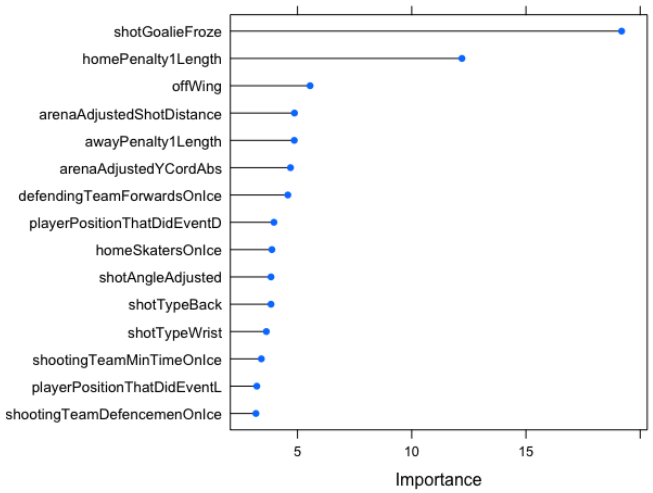

Fig. 9 *Variable Importance Plot of the Linear Model for the Goal/No Goal Classification.*

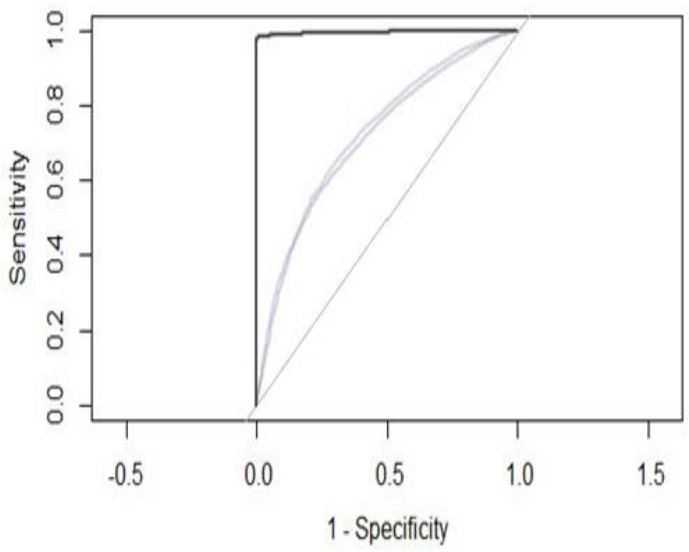

Fig. 10 *ROC curve of the Linear Model for the Goal/No Goal Classification.*

Non-Linear Classification											
			Training			Test		Confusion Matrix			
	Models	Tuning Parameter	ROC	Sens	Spec	Accuracy	Kappa		Goal	NoGoal	AUC
1	Nonlinear Discriminant Analysis	subclasses = 2	0.9960037	0.9997626	0.9732429	0.9741306	0.8278885	Goal NoGoal	4128 1	1553 54389	0.9955
2	Neural Networks	*Attempted, insufficient compute to run*									
3	Flexible Discriminant Analysis	Degree = 1 Nprune = 8	*0.9956881*	*0.9997626*	*0.9732429*	*0.9741306*	*0.8278885*	*Goal* *NoGoal*	*4128* *1*	*1553* *54389*	*0.9949*
4	Support Vector Machines	*Attempted, insufficient compute to run*									
5	K-Nearest Neighbors	*Attempted, insufficient compute to run*									
6	Naive Bayes	fL = 0, adjust = 1, usekernel = TRUE	*0.8865865*	0.001457407	*0.9998688*	*0.931297964*	*0.002151946*	*Goal* *NoGoal*	*5* *4124*	*3* *55939*	*0.8868*

Fig. 11 *Performance of the nonlinear classification models for the goal/no goal problem.* The best optimal linear performance is Mixed Discriminant Analysis with training accuracy of 0.996 and the best linear predictive ability is Mixed Discriminant Analysis with test accuracy of 0.974 and AUC of 0.9955.

Tree Classification											
			Training			Test		Confusion Matrix			
	Models	Tuning Parameter	ROC	Sens	Spec	Accuracy	Kappa		Goal	NoGoal	AUC
1	Basic Classification Tree	cp = 0.0005449261	0.9953442	0.9658972	0.9880864	0.9854506	0.8937628	Goal NoGoal	4003 126	748 55194	0.996
2	Bagged Tree		0.9962032	0.9581926	0.9887134	0.9855837	0.8942474	Goal NoGoal	3985 144	722 55220	0.9956
3	Random Forest	*Attempted, insufficient compute to run*									
4	Boosting	*Attempted, insufficient compute to run*									

Fig. 12 *Performance of the Tree models for the goal/no goal problem.* Only two of the tree models were able to run, the best optimal performance is the Bagged tree with training accuracy of 0.996, and the best linear predictive ability is the Bagged tree with test accuracy of 0.985.

Linear Classification												
			Training			Test		Confusion Matrix				
	Models	Tuning Parameter	Accuracy	Sens	Spec	Accuracy	Kappa		GOAL	MISS	SHOT	AUC
1	Logistic Regression "multinom"	decay = 0.1	0.7181192	0.6750728	0.7508061	0.6978683	0.2793856	GOAL	1016	165	231	
								MISS	23	418	599	0.8304
								SHOT	56	3674	9533	
2	Linear Discriminant Analysis		0.7213356	0.6847001	0.7552985	0.7205218	0.3313506	GOAL	1094	175	233	
								MISS	0	475	376	0.8351
								SHOT	1	3607	9754	
3	Partial Least Squares Discriminant Analysis	ncomp = 28	0.7202348	0.6738465	0.7473932	0.7199491	0.3134515	GOAL	1094	175	233	
								MISS	0	267	177	0.8715
								SHOT	1	3815	9953	
4	glmnet	alpha = 0.1 lambda = 0.01	0.7193599	0.6612759	0.746255	0.7197582	0.3118148	GOAL	1042	140	179	
								MISS	3	356	271	0.8263
								SHOT	50	3761	9913	
5	Sparse logistic regression	NumVars = 1 lambda = 0.1	0.6594239	0.3333333	0.6666667	0.6594337	0	GOAL	0	0	0	
								MISS	0	0	0	0.651
								SHOT	1095	4257	10363	
6	Nearest Shrunken Centroids	threshold = 1.724138	0.6593284	0.334522	0.6668722	0.660006363	0.004307123	GOAL	11	2	5	
								MISS	0	6	3	0.7347
								SHOT	1084	4249	10355	

Fig. 13 *Performance of the Linear classification model for the Event classification problem*.The best optimal linear performance is LDA with training accuracy of 0.721 and the best linear predictive ability is LDA with test accuracy of 0.7205. Best AUC is PLS model of 0.8715.

Non-Linear Classification												
			Training			Test		Confusion Matrix				
	Models	Tuning Parameter	Accuracy	Sens	Spec	Accuracy	Kappa		GOAL	MISS	SHOT	AUC
1	Nonlinear Discriminant Analysis	*Attempted, error. Still working to remediate.*										
2	Neural Networks	*Attempted, insufficient compute to run*										
3	Flexible Discriminant Analysis	*Attempted, error. Still working to remediate.*										
4	Support Vector Machines	*Attempted, insufficient compute to run*										
5	K-Nearest Neighbors	Attempted, insufficient compute to run										
6	Naive Bayes	fL = 0, usekernel = TRUE adjust = 1	0.6592012	0.3440605	0.6695511	0.65943366	0.01480406	GOAL	38	13	32	
								MISS	5	35	41	0.7518
								SHOT	1052	4209	10290	

Fig. 14 *Model Performance of the Non-Linear model for the Events Classification.* We were not successfully able to run many of the nonlinear models, only the Naive Bayes converged, however the performance of the model was not comparable to the Linear classification and Tree-based models.

Tree Classification												
			Training			Test		Confusion Matrix				
	Models	Tuning Parameter	Accuracy	Sens	Spec	Accuracy	Kappa		GOAL	MISS	SHOT	AUC
1	Basic Classification Tree	cp = 0.0005293229	0.7227957	0.6777592	0.7504477	0.7203309	0.3238141	GOAL	1094	175	233	
								MISS	0	385	289	0.8563
								SHOT	1	3697	9841	
2	Bagged Tree		0.7070803	0.6811186	0.7586761	0.7049952	0.3267469	GOAL	1047	138	186	
								MISS	16	853	998	0.8249
								SHOT	32	3266	9179	
3	Random Forest	*Attempted, insufficient compute to run*										
4	Boosting	*Attempted, insufficient compute to run*										

Fig. 15 *Model Performance of the Tree model for the Events classification*. The best optimal tree performance is the basic tree with training accuracy of 0.723, which performed better than all other models, however the linear LDA still had the best linear predictive ability and the linear PLS model had a higher AUC. Overall, of the models run for Events classification, the linear models, particularly LDA and PLS performed best.

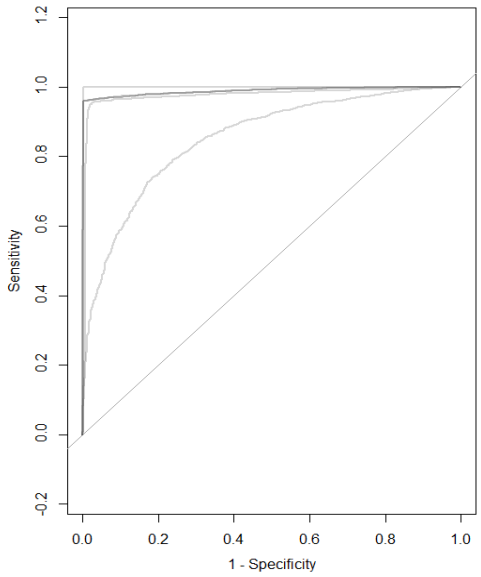

Fig 16 *ROC curve of the Linear Model for the Events Classification*

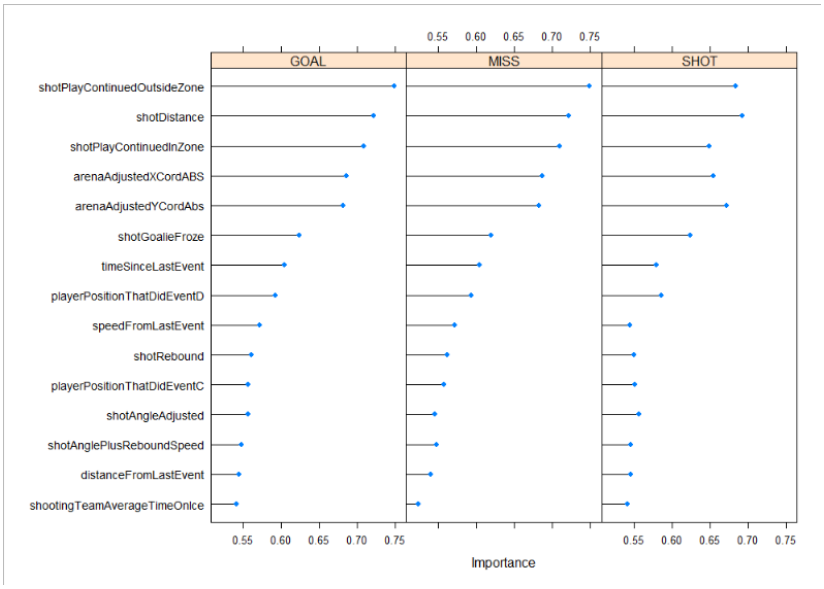

Fig. 17 *Variable Importance Plot of the Linear Model for the Events Classification.*

Model	Accuracy (w/ gambling)	Kappa (w/ gambling)	Accuracy (w/o gambling)	Kappa (w/o gambling)

Logistic Regression	0.8571	0.7144	0.551	0.1061
SVM	0.7551	0.5092	0.7143	0.4274
Random Forest	0.8571	0.7144	0.551	0.1061

Fig. 18 *Performance of Classification Models on the Gambling Classification Problem.*

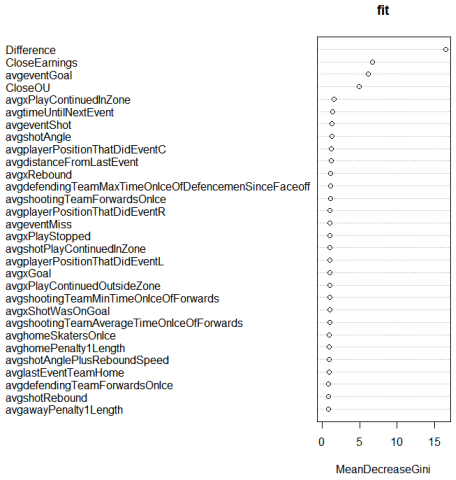

Fig. 19 *Variable Importance Plot of the Gambling Classification Problem..*

Conclusion

Our results demonstrate accurate predictions are possible within the NHL using shot data and can be used when considering strategy. Important predictors like shotPlayContinuedInside the zone and shot rebound suggest that generating rebounds could be the most efficient way to score goals in the NHL. It is also possible to predict gambling results with some level of accuracy. Predicting the outcome of a game is close to 50/50, however when taking into account the winnings a sportsbook will give you for the bet and using a teams aggregated shot data a gambler could use these models to make a reasonably close prediction. Given the high accuracy of the models we must consider the possibility that they are overfit and that the models have some sort of resultant data in the prediction. It is possible with the number of observations, the cross validation made for robust predicting. The strongest predictive methods we implemented were non linear discriminant analysis, linear discriminant analysis, logistic regression, and bagged tree. This suggests that there are linear separation boundaries in the data since the linear and non linear performed with the same level of accuracy. To research further feature

reduction could be performed to reduce the possibility of overfitting. The models could also be tested against data from a variety of years. This could help identify changing trends in the NHL as teams begin to understand their data, like the analytics revolutions in professional baseball and basketball.

REFERENCES

[1] Britannica, T. Editors of Encyclopaedia (2020, December 10). National Hockey League. Encyclopedia Britannica. https://www.britannica.com/topic/National-Hockey-League

[2] Weissbock, J., Viktor, H., Inkpen, D.: Use of performance metrics to forecast success in the national hockey league. European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop (2013)

[3] Tanner, P. (2021, July 8). Player and Team Data. MoneyPuck.com -Download Datasets. Retrieved September 28, 2021, from https://www.moneypuck.com/data.htm.

[4] Gough, Christina. "NHL Revenue by Year." Statista, 2 Feb. 2021, https://www.statista.com/statistics/193468/total-league-revenue-of-the-nhl-since-2006/.

[5] Naples, Marc; Gage, Logan; and Nussbaum, Amy (2018) "Goalie Analytics: Statistical Evaluation of Context-Specific Goalie Performance Measures in the National Hockey League," SMU Data Science Review: Vol. 1 : No. 2 , Article 12. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss2/12

[6] Pischedda, Gianni. (2014). "Predicting NHL Match Outcomes with ML Models. International Journal of Computer Applications," 101. 15-22. 10.5120/17714-8249. Available at: https://www.researchgate.net/publication/284457066_Predicting_NHL_Match_Outcomes_with_ML_Models

[7] Schuckers, M. and Curro, J. (2013) "Total hockey rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events," Proceeding of the 2013 MIT Sloan Sports Analytics Conference, Available at: https://www.statsportsconsulting.com/wp-content/uploads/Schuckers_Curro_MIT_Sloan_THoR.pdf

[8] Lehmus Persson T., Kozlica H., Carlsson N., Lambrix P. (2020) "Prediction of Tiers in the Ranking of Ice Hockey Players." In: Brefeld U., Davis J., Van Haaren J., Zimmermann A. (eds) Machine Learning and Data Mining for Sports Analytics. MLSA 2020. Communications in Computer and Information Science, vol 1324. Springer, Cham. https://doi.org/10.1007/978-3-030-64912-8_8

[9] O'Halloran, Mike. "Hockey Rink Dimensions, Diagrams, Game Time: NHL & International." Sports Feel Good Stories, Sports Feel Good Stories, 20 Oct. 2021, https://www.sportsfeelgoodstories.com/hockey-rink-dimensions-size-diagram/

APPENDIX

Detailed step-by-step instructions on how to run codes

Start by downloading the source data file "shots_2018-2020.csv" to your computer

1. **Scoring Opportunity (2 class): classify whether or not a shot results in a goal**
    a. *Goal / No Goal (2018 - 2020 Season)*
        i. Preprocessing
            1. Open "OR568_Spring2022Project_Preprocessing4_Goal.R"; adjust import to your file path for "shots_2018-2020.csv"
            2. Run full script, the final line of code will write a new csv file of the preprocessed code, "NHL_df3.csv"
            3. Complete, move to the next step.
        ii. Linear Classification
            1. Open "OR568_Spring2022Project_Linear Classification.R"
            2. Run full script, outputs are commented
                a. # 1 Logistic Regression
                b. # 2 Linear Discriminant Analysis
                c. # 3 Partial Least Squares Discriminant Analysis
                d. # 4 glmnet
                e. # 5 Sparse logistic regression
                f. # 6 Nearest Shrunken Centroids
            3. Complete, move to the next step.
        iii. Non-Linear Classification
            1. Open "OR568_Spring2022Project_Nonlinear Classification.R"
            2. Run full script, outputs are commented
                a. # 1 Nonlinear Discriminant Analysis
                b. # 3 Flexible Discriminant Analysis
                c. # 6 Naive Bayes
            3. Complete, move to the next step.
        iv. Tree Classification
            1. Open "OR568_Spring2022Project_Classification Trees.R"
            2. Run full script, outputs are commented
                a. # 1 Basic Classification Tree
                b. # 2 Bagged Tree
            3. Complete, move to the next step.

    b. Goal / No Goal (2020 Season only)
        i. Preprocessing
            1. Open "OR568_Spring2022Project_Preprocessing4_Goal_2020 v3"; adjust import to your file path for "shots_2018-2020.csv"
            2. Run full script, the final line of code will write a new csv file of the preprocessed code, "NHL_df5.csv"
            3. Complete, move to the next step.
        ii. Linear Classification

1. Open “OR568_Spring2022Project_Classification Trees_2020.R”
2. Run full script, outputs are commented
    a. # 1 Logistic Regression
    b. # 2 Linear Discriminant Analysis
    c. # 3 Partial Least Squares Discriminant Analysis
    d. # 4 glmnet
    e. # 5 Sparse logistic regression
    f. # 6 Nearest Shrunken Centroids
3. Complete, move to the next step.

iii. Non-Linear Classification

1. Open “OR568_Spring2022Project_Nonlinear Classification_2020.R”
2. Run full script, outputs are commented
    a. # 1 Nonlinear Discriminant Analysis
    b. # 3 Flexible Discriminant Analysis
    c. # 6 Naive Bayes
3. Complete, move to the next step.

iv. Tree Classification

1. Open “OR568_Spring2022Project_Classification Trees_2020.R”
2. Run full script, outputs are commented
    a. # 1 Basic Classification Tree
    b. # 2 Bagged Tree
3. Complete, move to the next step.

2. Profile of Shot Quality (3 class): classify Events (2020 Season only)

a. SHOT, MISS, GOAL

i. Preprocessing

1. Open “OR568_Spring2022Project_Preprocessing4_Event_2020.R”; adjust import to your file path for “shots_2018-2020.csv”
2. Run full script, the final line of code will write a new csv file of the preprocessed code, "NHL_df_event.csv"
3. Complete, move to the next step.

ii. Linear Classification

1. Open “OR568_Spring2022Project_Classification Trees_EVENT_2020.R”
2. Run full script, outputs are commented
    a. # 1 Logistic Regression
    b. # 2 Linear Discriminant Analysis
    c. # 3 Partial Least Squares Discriminant Analysis
    d. # 4 glmnet
    e. # 5 Sparse logistic regression
    f. # 6 Nearest Shrunken Centroids
3. Complete, move to the next step.

iii. Non-Linear Classification

1. Open “OR568_Spring2022Project_Nonlinear Classification_EVENT_2020.R”
2. Run full script, outputs are commented

a. # 6 Naive Bayes
3. Complete, move to the next step.

iv. Tree Classification
1. Open "OR568_Spring2022Project_Classification Trees_EVENT_2020.R"
2. Run full script, outputs are commented
   a. # 1 Basic Classification Tree
   b. # 2 Bagged Tree
3. Complete, move to the next step.

Attribute	**Data Type**
shotID	Nominal
homeTeamCode	Nominal
awayTeamCode	Nominal
season	Nominal
isPlayoffGame	Nominal
game_id	Nomina
homeTeamWon	Nominal
id	Ordinal
Time	Interval
timeUntilNextEvent	Interval
timeSinceLastEvent	Interval
period	Ordinal
team	Nominal
location	Nominal
event	Nominal
goal	Nominal
shotPlayContinuedOutsideZone	Nominal

shotPlayContinuedInZone	Nominal

Variable	Definition
shotID	Unique id for each shot
homeTeamCode	The home team in the game. For example: TOR, MTL, NYR, etc
awayTeamCode	The away team in the game
season	Season the shot took place in. Example: 2009 for the 2009-2010 season
isPlayoffGame	Set to 1 if a playoff game, otherwise 0
game_id	The NHL Game_id of the game the shot took place in
homeTeamWon	Set to 1 if the home team won the game. Otherwise 0.
id	The event # of the shot in the game
time	Seconds into the game of the shot
timeUntilNextEvent	Time between the shot and the next event that happens in the game after the shot
timeSinceLastEvent	Time between the shot and the event that took place before the shot

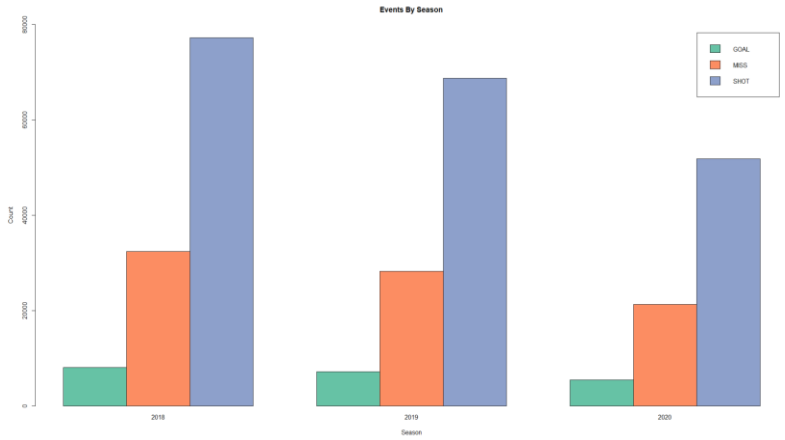

Fig. 20 Chart of Events by Season**:** This is a chart of Events by Season. An event is defined as whether the shot was a shot on goal (SHOT), goal (GOAL), or missed the net (MISS). You'll notice that while the distribution of events does not look to change significantly year-to-year, it does look like the number of shot attempts have decreased over the past couple years. We could speculate this may be due to a rule change or other factors that are beyond the scope of this analysis.

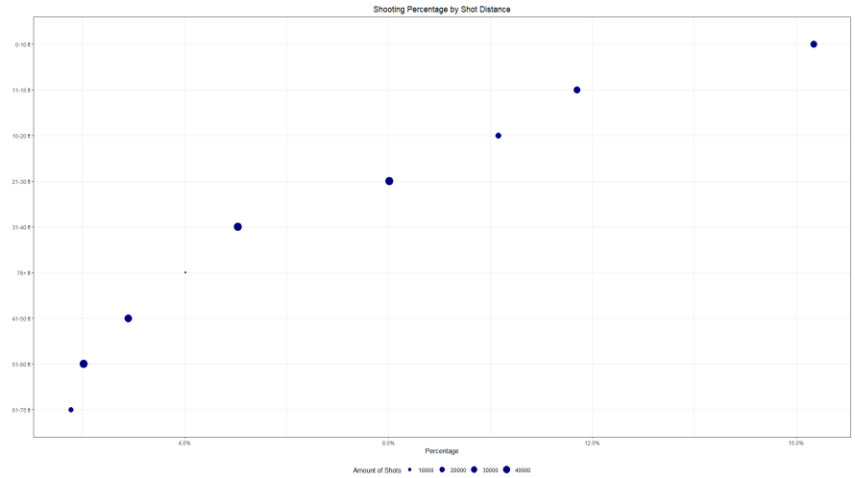

Fig. 21 Shooting Percentage by Shot Distance: This chart gives a visual of shooting percentage by shot distance. You'll notice a non-linear, logarithmic pattern where the shot percentage is clearly higher the closer you are to the goal and lower when further away. Very few shot attempts over 76ft.

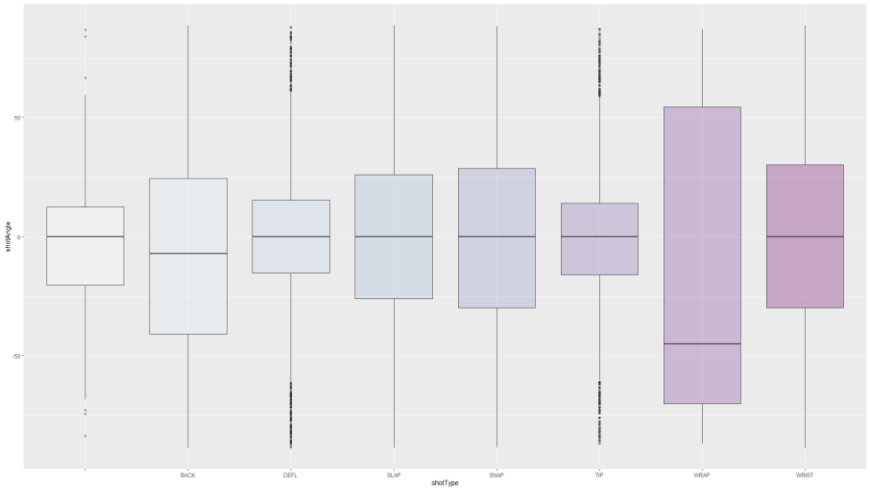

Fig. 22 The Box-plot gives a visual of shot angle by shot type. Most shots mean is at 0 degrees, but we see a few interesting characteristics. 1) a much larger range and skew of shot angle for the wrap shot. We imagine this has to do with the right-handedness of most shooters. The backhand also skews towards a negative shot angle. Deflections and tips (very similar shot types) have the most narrow range around 0 degrees. This makes sense, we suspect a similar chart showing distance would have these shots very close to the goal. There are noticeable outliers.

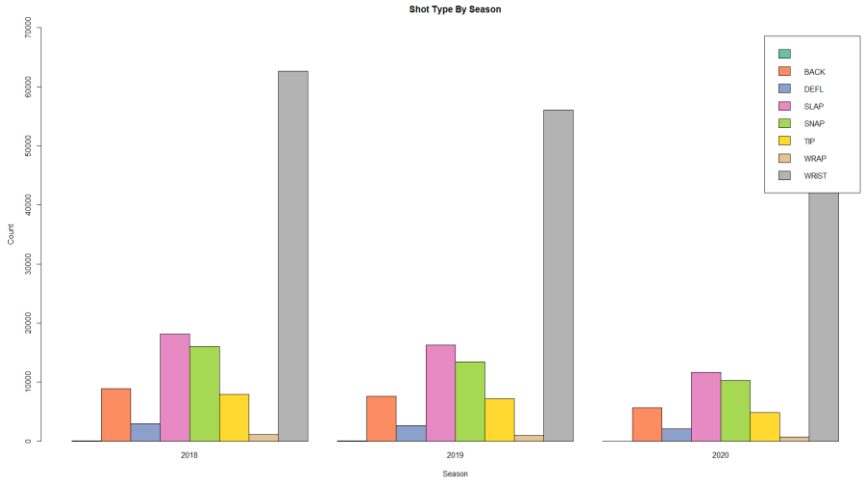

Fig. 23 This chart displays the volume of shots by shot type by season. Clearly the wrist shot is most popular by orders of magnitude, very few wrap shots overall. Distribution across years is pretty static.We included the blanks here to give a visual idea of the percentage of shots that have a defined. Very few shots have an undefined shot type, so we can be confident that each shot type is well represented and the dataset is pretty solid in this regard