BISHOP OR ROOK'D: A CREDIBILITY ESTIMATION FOR NHL GOALTENDER CAREER PERFORMANCE

PROGRESS REPORT

Jeff Faath
School of Computer Science
Georgia Institute of Technology
jfaath3@gatech.edu

Saiem Gilani
School of Industrial & Systems Engineering
Georgia Institute of Technology
sgilani9@gatech.edu

Brian Haley
School of Computer Science
Georgia Institute of Technology
bhaley9@gatech.edu

Michael Parkatti
School of Industrial & Systems Engineering
Georgia Institute of Technology
mparkatti3@gatech.edu

Spring Smith
School of Industrial & Systems Engineering
Georgia Institute of Technology
ssmith494@gatech.edu

1 Introduction and Motivation

As the 4th most popular sports league in America, the National Hockey League generates over $4 billion in revenue each year. Franchises can earn $1.2-1.5 million for every playoff home game, creating incentive for ownership to have successful seasons. With 17 goalies slated to earn over 5 million dollars in 2020, evaluating goaltender efficacy is of utmost importance. Currently, a goaltender's save percentage ($\frac{\text{Goals Allowed}}{\text{Shots on Net Faced}}$) is considered an authoritative measure of a goaltender's ability to stop shots. This measure fails to incorporate the context in which each shot is taken, considering them all to be equally challenging. Our approach will utilize high-dimensional machine learning to estimate probabilities for certain shots to become goals, using factors such as distance, shot type, angle to net, shooter talent, etc. This will provide the necessary context to rate goaltenders accurately and give teams the ability to avoid playing poor goaltenders longer than necessary to properly evaluate their talent level.

2 Problem Description

Each shot a goalie has faced in his career is a Bernoulli trial - a coin flip with a certain probability of being either a goal or not a goal. Each shot has its own unique expectation of becoming a goal (or, its Expected Goals [xG]). We seek to simulate a goaltender's unique set of Bernoulli trials (or shots faced so far in his career) to build an empirical cumulative distribution function (CDF) for the number of goals that the 'average' goaltender would have been expected to allow over that unique set of trials. In comparing the number of goals that the goaltender has actually allowed to this empirical CDF, we can get a sense of how likely that performance was. After each game of a goaltender's career, we can apply a statistical test:

$H_0 =$ The goaltender has average ability to prevent goals: $0.05 <$ Position on CDF < 0.95
$H_{a_1} =$ The goaltender has above average ability to prevent goals: Position on CDF ≤ 0.05
$H_{a_2} =$ The goaltender has below average ability to prevent goals: Position on CDF ≥ 0.95

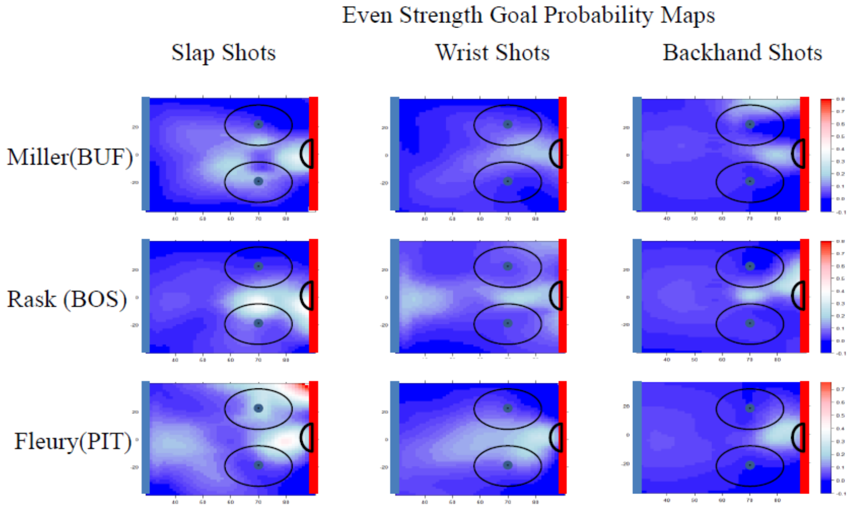
This will result in a rolling test of the goaltender's ability in relation to average expectations over his career. It is expected that the results of this test will provide much earlier insight into the true ability of a goaltender – allowing poor goaltenders to be avoided and good goaltenders to be exploited.

3 Literature Survey

3.1 Hockey Goaltender Analysis

Andrew Thomas mentions goalies can be measured by their save rate using a Poisson distribution. This accounts for the low number of goals, but we alleviate that by looking at shots rather than just goals [15]. Schuckers goes into detail about the old ways of measuring goalies (save proportion) to the newer ones (linear models using shot type, location, and angle from previous shot). None of these models incorporate machine learning or quality of the shooter [11]. Roith and Magel use stepwise analysis to determine important factors for winning a single NHL game and making it to the playoffs. The analysis shows stopping shots is the most important factor. We plan to further analyze the importance of goalies and determine when a team should consider replacement [9]. Schuckers measures shot difficulty and uses average distribution of shots and each goalie's spatially smoothed shot performance map as a basis for goaltender comparison (as in Figure 1 below) [10].

Figure 1: Goal Probability Maps at Even Strength of Different Shot Types for Selected Goalies
Scale is Blue(low) to Red(high)



3.2 General Hockey Analysis

Macdonald uses ridge regression to model how valuable an NHL player is to their team based on their expected contribution to goals/hour. Macdonald focuses only on 5v5 playing stances whereas we plan to consider all playing stances [6]. Schulte, Liu, and Li analyze junior league players with seasonal data to determine success in the NHL for drafting decisions. They use logistic regression to determine whether a player would play in an NHL game, but do not account for how well the player would perform [14]. Shuckers and Curro delve into topics such as home-ice effect and possession changes and what effect they have on probabilities of shots becoming goals. They used these factors to create a rating for forwards and defenseman, whereas we intend to use similar factors for rating goaltenders [12]. Using SPORTLOGiQ's spatio-temporal dataset, team-level pace metrics were developed in even-strength situations across offensive, neutral, and defensive zones by tracking the path of the puck and possession events over a spatial polygon grid and applying Gaussian kernel smoothing baselined to league average. This SPORTLOGiQ spatio-temporal dataset was also used to create player clusterings using an affinity propagation algorithm and activity location data. A Markov model was built to measure their relative scoring impact [16].

3.3 General Sports Analysis

Jamieson analyzed home field advantage in sports across multiple sports under various conditions, showing that home teams have a distinct advantage. For NHL, it was shown that home teams win about 59.5% of the time. This can be considered as a weight in our models [5]. Luke, Daniel, Alexander, and Andrew explain that the location of non-shooters on the basketball court are a significant factor to scoring. We could use non-shooters positions in our model and expand on the position data to categorize offensive strategies and goalie's effectiveness [2].

3.4 General Statistical Analysis

Our model deals with shots on goal where the success rate is generally less than 10%. This leads to a highly imbalanced data set. Haibo et al provided a review of running learning algorithms with imbalanced data, whose insights can be applied to our development [4]. In two papers, Niculescu-Mizil and Caruana investigate methods to create probability estimations for the positive class label in supervised learning classification models, in a general survey across multiple algorithms and specifically for boosted trees. These will help in correcting biases in our expected goal probabilities, regardless of the selected model type [7][8]. Franks et al proposed a "meta-metric" framework to aid evaluating the stability, discrimination and independence of a sport metric. This framework can be applied to our method to ensure the robustness of the model [3].

4 Proposed Method

4.1 Data Collection and Feature Engineering

Using the MoneyPuck shots database and skater/goalie advanced statistics from Hockey-Reference, we were able to join the datasets to create in line shooter and goaltender features using the aggregated statistics from all prior seasons for the players. The SportRadar play-by-play files also contain line-up data which we will incorporate and examine. However, due to the limited amount of line-up data (3 seasons), it may not be appropriate for inclusion.

The data underwent several preprocessing cleaning steps. These steps include removing all empty net shots (as these shots provide no information on goaltender ability) and removing shots greater than 70 feet from the net (to aid in balancing out goals vs. non-goals). Additionally, categorical variables were processed to remove outliers and nonsense entries.

After transforming the shot (x, y) coordinates to a single side of the rink, each shot data point has over 40 features on shooters and 24 features on goalie situational and location performance using each players prior career metrics in those instances, including evaluating special teams (short-handed or power play) and score differential.

4.2 Machine Learning

To determine true goaltender skill, we must first gauge the quality of shots taken in the NHL. The metric traditionally used to assess talent is Save Percentage (SV%). To use SV% alone in assessing talent, one must assume that every shot has an equal likelihood of becoming a goal. Our assertion is that this assumption is invalid.

4.2.1 Innovation #1 - A State of the Art Expected Goals Model

We will build our own Expected Goals (xG) model. Some analytically-inclined blogs have dabbled in creating their own xG models using Logistic Regression or Gradient Boosted Trees. For this project we will build a completely unique supervised machine learning model: the labels will be whether each shot became a goal or not, our features will be the variables created in the data engineering phase of this project, and the outputs will be probabilities of each shot becoming a goal. For more details on our plan for calculating xG, please see Appendix A.

4.3 Statistical Test

The machine learning phase of the project will produce probabilities of each shot going in. The statistical testing phase will use these probabilities to determine whether a goalie is good, bad, or not able to be proven of being either.

4.3.1 Innovation #2 – A statistical test to help determine goalie performance relative to league-average

In this phase, we are going to stochastically simulate each shot the goalie has faced 10,000 times. Essentially, this will result in 10,000 different simulations of his career to-date, assuming league average talent.

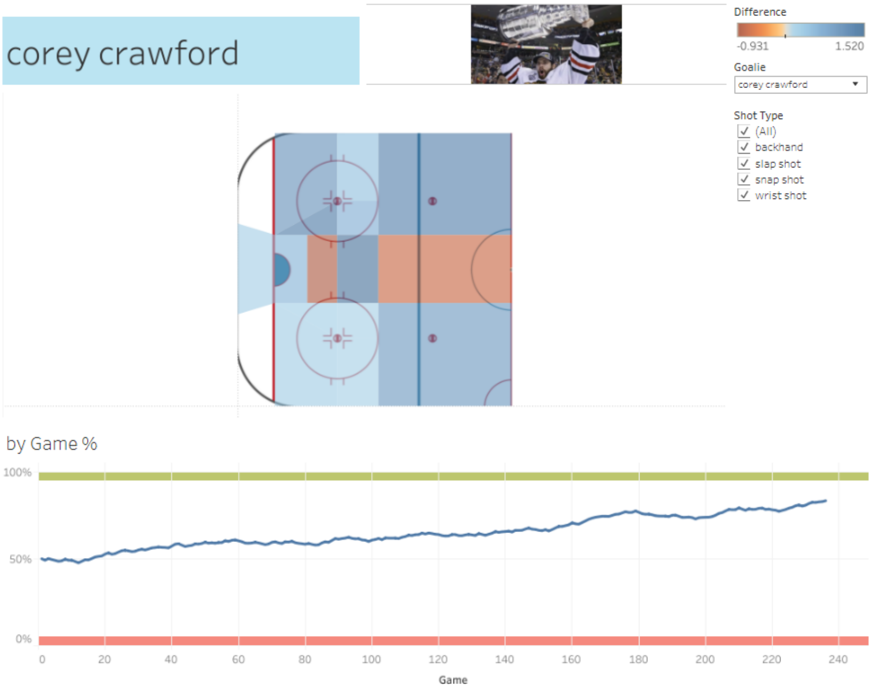
We will then look at the empirical goal distribution of these 10,000 career simulations. At a confidence level of 0.90, we will form a statistical test using the upper 5% of goal results and lower 5% of goal results. If the goalie has let in a goal total that is in the lower 5% of simulated goal results, we will reject the null hypothesis that he has league-average talent, and vice versa. If he's in neither hypothesis range, we'll need to accept the null hypothesis that he has league-average talent.

4.4 Visualization

4.4.1 Innovation #3 - Interactive Visualization of Goalie Performance

We will create an interactive visualization that tells the story of each individual goalie that is not available for consumer consumption today. This visualization will let anybody see how effective a goalie is on multiple levels including shot type, shot location, and season. A user will also be able to see how our model grades each goalie and at what point a goalie can be considered above or below average. This visualization will make the results of the complicated machine learning analysis easy to consume by anybody. Figure 2 (below) is a work in progress that we have created in Tableau. Our next steps will be working to add additional features to the visualization, improving clarity to the user, and making the visualization more attractive.

Figure 2: Visualization in Progress



Goalie
corey crawford ▾

Shot Type
☑ (All)
☑ backhand
☑ slap shot
☑ snap shot
☑ wrist shot

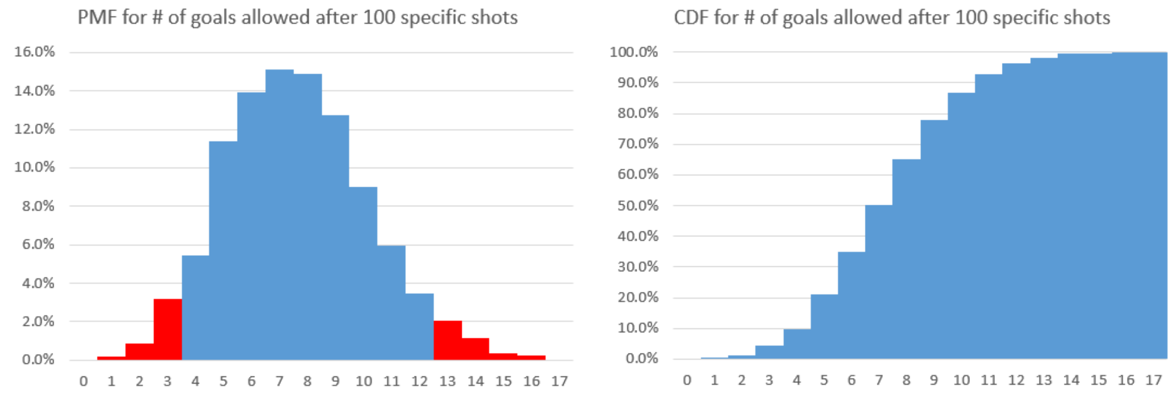
5 Experiments and Results

We conducted a series of tests using a 10,000 observation sample (out of 1.25 million). 15% of this set was held for testing, 85% was used in a 5-fold cross-validation training exercise with hyperparameter tuning using sklearn's `GridSearchCV` function. The best model output by each algorithm was then calibrated using sklearn's `CalibratedClassifierCV` function using a separate sample of 10,000 observations. The brier score loss numbers for each algorithm tested are below:

Algorithm	**Brier Score Loss with Calibration**	**Brier Score Loss without Calibration**
Logistic Regression	0.0624	0.0625
Support Vector (RBF Kernel)	0.0602	0.0600
Random Forests	0.0619	0.0619
Classification Tree	0.0615	0.0625
XGBoost	0.0606	0.0626
K-Nearest Neighbors	0.0611	0.0626
Quadratic Discriminants	0.0617	0.0642

Based on these results, we will explore the best performing methods. For our statistical test, we decided to proceed with a toy example of how the simulations would be run and how conclusions would be drawn. We randomly sampled 100 shots from our dataset – this is a stand-in for the first 100 shots a theoretical goalie would face in his career. We then simulated these 100 shots 1500 times each (less than our planned 10,000 simulations in the full model). From this, we created an empirical probability mass function (PMF) and cumulative distribution function (CDF), as follows in Figure 3.

Figure 3: PMF and CDF for Number of Goals allowed after 100 specific shots



On the PMF, you can see the rejection regions in red – if this theoretical goalie had let in 3 or less goals over these 100 shots, we would accept an alternate hypothesis that he is better than the average NHL goalie. If he had let in 13 or more goals over those 100 shots, he would be demonstrably worse than the average NHL goalie. The average NHL goalie would fall in the middle blue region about 90% of the time. It turns out that 6 actual goals were scored on these 100 shots, so our theoretical goalie would not be able to be proven to be better or worse than average after this sample.

On the right, we have the CDF for the same goalie – where he falls on this curve will drive the 'score' produced for our data visualization.

6 Conclusion

This paper introduces a credibility estimation for NHL goaltender performance. This method uses a machine learning model to evaluate the quality of shots against the goaltender to determine the probability of each shot becoming a goal. Then, statistical testing is used to determine whether a goaltender is good, average, or bad, based on the shots they have faced. This method is evaluated using cross validation to determine the Brier Score Loss with and without calibration. The end result provides an interactive visualization for any goaltender, showing how the goaltender performs on multiple levels including shot type, shot location, and season. This method will allow key stakeholders for the NHL teams to easily determine whether or not their goaltender is performing well compared to other NHL goaltenders.

7 Plan of Activities - New

Phase	Task	Who?	Effort (hrs)	Duration (days)	Goal Start Date	Goal End Date
Admin	Wireframe development	Michael	2			
	Code Organization / Source Control	Jeff	6			
Project Proposal	Doc - Literature Survey	All / Spring edit	15	3	10/8	10/11
	Doc - Plan of Activities	Brian / Spring	2	3	10/8	10/11
	Doc - Technical Problem Definition	Saiem	2	3	10/8	10/11
	Doc - 9 questions	Michael	2	3	10/8	10/11
	Doc - Finalize and transform to Latex doc	Saiem / Spring	4	3	10/8	10/11
	Proposal Presentation Slides	Jeff / Brian	2	3	10/8	10/11
	Video - Creation/Narration	Michael/Brian	2	3	10/8	10/11
Data Collection	Download shot dataset from Moneypuck	Saiem	1	9	10/1	10/10
	Get API keys for data download	All	3	9	10/1	10/10
	Download play-by-play JSON files from the SportRadar API	Saiem	4	11	10/05	10/15
	Download individual player statistics from Hockey Reference (`https://www.hockey-reference.com`)	Saiem	2	3	10/15	10/17
	Download individual Goalie photos from Hockey Reference (`https://www.hockey-reference.com`)	Spring	4	2	11/7	11/9
Data Integration	Join datasets together	Saiem	4	2	10/15	10/17
	QA/Verify cleanliness of data	Brian	4	2	10/17	10/19
Feature Engineering	Performing feature selection and transformation as necessary	Jeff	6	6	10/14	10/24
	Incorporating other advanced metrics from referenced papers	Saiem	6	10	10/14	10/24
ML Modeling	Preprocessing of data	Jeff	6	15	10/20	11/4
	DNN Experimentation	Jeff	6	28	10/20	11/17
	Model Selection/Hyperparameter tuning	Jeff/Michael/Spring	7	28	10/20	11/17
	Calibration of Probability Outputs	All	3	27	10/20	11/17
Statistical Test	Initial experimentation on simulation	Michael	3	8	11/1	11/9
	Full build of simulation model and results	Michael	6	10	11/8	11/17
Progress Report	Construct and submit progress report	All	20	8	11/1	11/10
Visualization Development	Create prototype in Tableau	Brian	12	6	11/4	11/10
	Finalize Tableau Presentation Layer	Spring	6	6	11/11	11/17
	Create poster outline and framework	Michael	5	6	11/11	11/17
Final Report and Poster	Compile all results, figures, comparisons for report, poster and presentations for submission	All	20	14	11/18	11/22

8 Plan of Activities - Old

Phase	Task	Who?	Effort (hrs)	Duration (days)	Goal Start Date	Goal End Date
Planning / Admin	Wireframe development	Michael	2			
Project Proposal	Doc - Literature Survey	All / Spring edit	12	3	10/8	10/11
	Doc - Plan of Activities	Brian / Spring	2	3	10/8	10/11
	Doc - Technical Problem Definition	Saiem	2	3	10/8	10/11
	Doc - 9 questions	Michael	2	3	10/8	10/11
	Slides	Jeff / Saiem	2	3	10/8	10/11
	Video Creation/Narration	Michael/Brian	2	3	10/8	10/11
Data Collection	Download shot dataset from Moneypuck	Saiem	3	9	10/1	10/10
	Get API keys for data download	All	3	9	10/1	10/10
	Download play-by-play JSON files from the SportRadar API for all games in 2015-2018 seasons	Saiem	3	9	10/1	10/10
	Download individual player statistics from Hockey Reference (`https://www.hockey-reference.com`)	Saiem	3	3	10/11	10/14
Data Integration	Join datasets together	Saiem	2	2	10/12	10/14
	QA/Verify cleanliness of data	Brian	2	2	10/12	10/14
Feature Engineering	Performing feature selection and transformation as necessary	Jeff	6	6	10/14	10/20
	Incorporating other advanced metrics from referenced papers	Saiem	4	6	10/14	10/20
ML Modeling	Preprocessing of data	Michael	4	15	10/20	11/4
	Model Selection	Jeff	4	15	10/20	11/4
	Hyperparameter tuning	Spring	3	15	10/20	11/4
	Calibration of Probability Outputs	Michael/All	3	15	10/20	11/4
Progress Report	Construct and submit progress report	All	6	7	11/1	11/8
Visualization Development	Ice Rink Heatmap POC	Brian / Saiem	1	2	10/10	10/11
	Create prototype in Tableau	Brian	5			
	Finalize Tableau Presentation Layer	Spring	5			
	Create poster outline and framework	Michael	5			
Final Report and Poster	Compile all results, figures, comparisons for report, poster and presentations	All	20	14	11/8	11/21
Submit Report		Michael	0	0	11/21	11/21

9 Distribution of Work

All team members contributed equally and are earnestly participating with the project.

10 References

1. Albert, J. (2017). Situational Statistics, Clutch Hitting, and Streakiness. Handbook of Statistical Methods and Analysis in Sports (pp. 89-112). Boca Raton, FL: CRC Press
2. Bornn, L; Cervone, D; Alexander, F; Miller, A. (2017). Studying Basketball through the Lens of Player Tracking Data. Handbook of Statistical Methods and Analysis in Sports (pp. 245-269). Boca Raton, FL: CRC Press
3. Franks, A. M., D'Amour, A., Cervone, D., Bornn, L. (2016). Meta-Analytics: Tools for Understanding the Statistical Properties of Sports Metrics. arXiv:1609.09830v1
4. Haibo H., Garcia E. A. (2009). Learning from Imbalanced Data. IEEE Transactions On Knowledge And Data Engineering, vol. 21, no. 9, pp. 1263-1283
5. Jamieson, J. P. (2010). The Home Field Advantage in Athletics: A Meta-Analysis. Journal of Applied Social Psychology, 40, 7, pp. 1819–1848.
6. Macdonald, Brian. (2012). An Expected Goals Model for Evaluating NHL Teams and Players. MIT Sloan Sports Analytics Conference, 2012. Retrieved from `http://www.sloansportsconference.com/wp-content/uploads/2012/02/NHL-Expected-Goals-Brian-Macdonald.pdf`
7. Niculescu-Mizil, A; Caruana, R. (2005). Obtaining calibrated probabilities from boosting. UAI'05 Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (pp. 413 - 420). Arlington, VA: AUAI Press
8. Niculescu-Mizil, A; Caruana, R. (2005). Predicting Good Probabilities With Supervised Learning. ICML '05 Proceedings of the 22nd international conference on Machine learning (pp. 625 - 632). Bonn, Germany
9. Roith J, Magel R. (2014). An analysis of factors contributing to wins in the National Hockey League. International Journal of Sports Science, 4(3), 84–90. doi:10.5923/j.sports.20140403.02
10. Schuckers, M. (2011). DIGR: A Defense Independent Rating of NHL Goaltenders using Spatially Smoothed Save Percentage Maps. MIT Sloan Sports Analytics Conference. Retrieved from: `http://www.sloansportsconference.com/wp-content/uploads/2011/08/DIGR-A-Defense-Independent-Rating-of-NHL-Goaltenders-using-Spatially-Smoothed-Save-Percentage-Maps.pdf`
11. Schuckers, M. (2017). Statistical Evaluation of Ice Hockey Goaltending. Handbook of Statistical Methods and Analysis in Sports (pp. 307-325). Boca Raton, FL: CRC Press
12. Schuckers, M; Curro, J. (2013). Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. MIT Sloan Sports Analytics Conference, 2013. Retrieved from `http://www.sloansportsconference.com/wp-content/uploads/2013/Total%20Hockey%20Rating%20(THoR)%20A%20comprehensive%20statistical%20rating%20of%20National%20Hockey%20League%20forwards%20and%20defensemen%20based%20upon%20all%20on-ice%20events.pdf`
13. Schulte, O., Zhao, Z., & Javan, M. (2017). Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact. MIT Sloan Sports Analytics Conference. Retrieved from `http://www.sloansportsconference.com/wp-content/uploads/2017/02/1625.pdf`
14. Schulte, O; Liu, Y; Li, C. (2018). Model Trees for Identifying Exceptional Players in the NHL Draft. arXiv:1802.08765v1
15. Thomas, A. (2017). Poisson/Exponential Models for Scoring in Ice hockey. Handbook of Statistical Methods and Analysis in Sports (pp. 271-285). Boca Raton, FL: CRC Press
16. Yu, D., Boucher, C., Bornn, L., & Javan, M. (2019). Playing Fast Not Loose: Evaluating team-level pace of play in ice hockey using spatio-temporal possession data. MIT Sloan Sports Analytics Conference. Retrieved from `http://www.sloansportsconference.com/wp-content/uploads/02/HockeyPace.pdf`

11 Appendix A: Proposed Method for Expected Goals Model

First, we will undertake a series of small cross-validated exploratory models using a multitude of classic machine learning algorithms – this will be useful in understanding which models are best suited to our problem. We will be initially rating our results using the Brier Loss Score – this is simply the mean squared error between the predicted probabilities and the labels.

Next, we will select one type of model (based on its above performance, speed, and interpretability) to move forward with a full model build using all data points and 5-fold cross-validation. The 2017-2018 season will be held out as our final unseen test set (a third-party benchmark is available for that season for us to compare against).