Modeling the probability of an NHL goal for player-placement strategy:

A Naïve (Bayes') approach

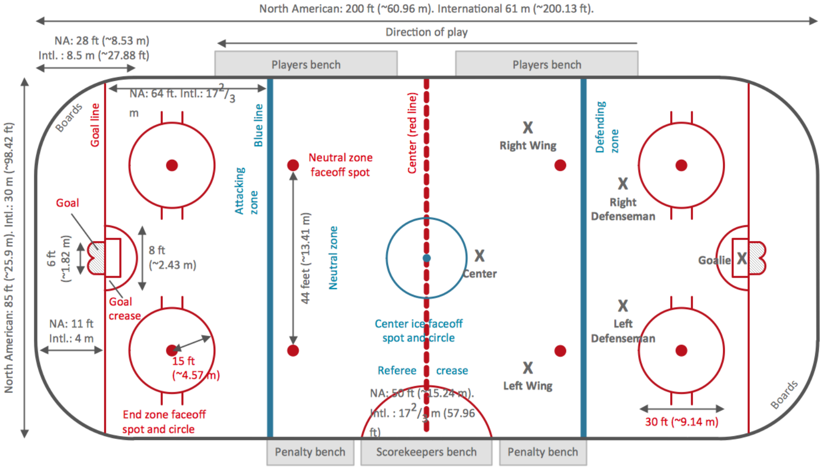
Alex Zajichek

February 20, 2017

Creative Component Presentation

1. Introduction
   - Background
   - Objective
2. The Data
   - `nhlscrapr`
   - Predictors
3. Naïve Bayes' Methodology
   - Empirical Naïve Bayes'
   - Parametric Naïve Bayes'
4. Model Evaluation
5. Results
   - Comparison
   - Implications
   - R shiny application
6. Future work

Background


North American: 200 ft (~60.96 m). International 61 m (~200.13 ft).
Direction of play
NA: 28 ft (~8.53 m)
Intl. : 8.5 m (~27.88 ft)
Players bench
Players bench
NA: 64 ft. Intl.: 17 2/3 m
Boards
Goal line
Blue line
Center (red line)
Defending zone
Right Wing
Neutral zone faceoff spot
Attacking zone
Right Defenseman
Goal
6 ft (~1.82 m)
8 ft (~2.43 m)
44 feet (~13.41 m)
Neutral zone
Center
Goalie
North American: 85 ft (~25.9 m). Intl.: 30 m (~98.42 ft)
Goal crease
NA: 11 ft
Intl.: 4 m
Center ice faceoff spot and circle
Left Defenseman
15 ft (~4.57 m)
Referee crease
Left Wing
NA: 50 ft (~15.24 m).
Intl. : 17 2/3 m (57.96 ft)
End zone faceoff spot and circle
30 ft (~9.14 m)
Boards
Penalty bench
Scorekeepers bench
Penalty bench

Previous work

- Not much work has been done on this specific application

Previous work

- Not much work has been done on this specific application
- A few papers have used logistic regression to model goal probabilities as part of different objectives

Previous work

- Not much work has been done on this specific application
- A few papers have used logistic regression to model goal probabilities as part of different objectives
- Gramacy, Jensen, and Taddy modeled player contribution towards a goal

Objective

Goals:

- Propose a crude but simple alternative to model goal probabilities

Objective

Goals:

- Propose a crude but simple alternative to model goal probabilities
- Compare model performance to logistic regression

Objective

Goals:

- Propose a crude but simple alternative to model goal probabilities
- Compare model performance to logistic regression
- Create R shiny application to explore results

Objective

Goals:

- Propose a crude but simple alternative to model goal probabilities
- Compare model performance to logistic regression
- Create R shiny application to explore results

Possible Implications:

- Understand shot characteristics more likely to lead to a goal

Objective

Goals:

- Propose a crude but simple alternative to model goal probabilities
- Compare model performance to logistic regression
- Create R shiny application to explore results

Possible Implications:

- Understand shot characteristics more likely to lead to a goal
- Put players in favorable (or unfavorable) situations on the ice

The Data: nhlscrapr

- R package giving web-scraping abilities to download NHL play-by-play data

The Data: nhlscrapr

- R package giving web-scraping abilities to download NHL play-by-play data
- Observation example:

season	gcode	refdate	event	period	seconds	etype
20092010	20001	2830	1	1	0	FAC

a1	a2	...	...	...
9 BRENDAN MORRISON	21 BROOKS LAICH	...	...	...

Predictors

- Considered 579181 shots taken from 2007-2015 within 65 minutes of gameplay (2002 - 2006 didn't contain shot coordinates)

Predictors

- Considered 579181 shots taken from 2007-2015 within 65 minutes of gameplay (2002 - 2006 didn't contain shot coordinates)
- Predictors used: *angle, catch, distance, game type, height, home, manpower, minute, position, shot side, type, weight*

Naïve Bayes' Methodology

For a given shot taken during an NHL game, let

$$Y_i = \begin{cases} 1 & \text{for a goal} \\ 0 & \text{for a save} \end{cases} \tag{1}$$

and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{i12})$ be the 1 x 12 predictor vector for the i^{th} shot taken, where $j = 1, ..., 5$ for continuous predictors, and $j = 6, ..., 12$ for categorical.

Naïve Bayes' Methodology

Conditional densities and probability mass functions:

	Continuous	Categorical	Marginal probabilities
Goal	$f_j(x_{ij} \vert Y_i = 1)$	$P_j(X_{ij} = x_{ij} \vert Y_i = 1)$	$P(Y_i = 1)$
Save	$f_j(x_{ij} \vert Y_i = 0)$	$P_j(X_{ij} = x_{ij} \vert Y_i = 0)$	$P(Y_i = 0)$

Naïve Bayes' Methodology

For the i^{th} shot, if we let

$$G_i = P(Y_i = 1) \times \prod_{j=1}^{5} f_j(x_{ij}|Y_i = 1) \times \prod_{j=6}^{12} P_j(X_{ij} = x_{ij}|Y_i = 1) \tag{2}$$

$$S_i = P(Y_i = 0) \times \prod_{j=1}^{5} f_j(x_{ij}|Y_i = 0) \times \prod_{j=6}^{12} P_j(X_{ij} = x_{ij}|Y_i = 0) \tag{3}$$

$$\begin{aligned} P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) &= \frac{P(Y_i = 1, \mathbf{X}_i = \mathbf{x}_i)}{P(\mathbf{X}_i = \mathbf{x}_i)} \\ &= \frac{P(Y_i = 1, \mathbf{X}_i = \mathbf{x}_i)}{P(Y_i = 1, \mathbf{X}_i = \mathbf{x}_i) + P(Y_i = 0, \mathbf{X}_i = \mathbf{x}_i)} \\ &= \frac{P(Y_i = 1) \times P(\mathbf{X}_i = \mathbf{x}_i|Y_i = 1)}{P(Y_i = 1) \times P(\mathbf{X}_i = \mathbf{x}_i|Y_i = 1) + P(Y_i = 0) \times P(\mathbf{X}_i = \mathbf{x}_i|Y_i = 0)} \\ \text{naïve assumption} \rightarrow &= \frac{G_i}{G_i + S_i} \end{aligned} \tag{4}$$

Empirical Naïve Bayes' (ENB)

- Assumed no parametric form to predictors

Empirical Naïve Bayes' (ENB)

- Assumed no parametric form to predictors
- Used R's `density` and `approxfun` functions to obtain density estimates of continuous predictors

Empirical Naïve Bayes' (ENB)

- Assumed no parametric form to predictors
- Used R's `density` and `approxfun` functions to obtain density estimates of continuous predictors
- Categorical probabilities were calculated as the proportion of observations belong to a given level

Empirical Naïve Bayes' (ENB)

- Assumed no parametric form to predictors
- Used R's `density` and `approxfun` functions to obtain density estimates of continuous predictors
- Categorical probabilities were calculated as the proportion of observations belong to a given level
- Evaluated equation (4) to obtain predicted probabilities

Parametric Naïve Bayes' (PNB)

- Examined empirical densities to determine common parametric model to fit to each predictor

Parametric Naïve Bayes' (PNB)

- Examined empirical densities to determine common parametric model to fit to each predictor
- In the spirit of the naïve approach, some approximations were crude, but chosen for simplicity

Parametric Naïve Bayes' (PNB)

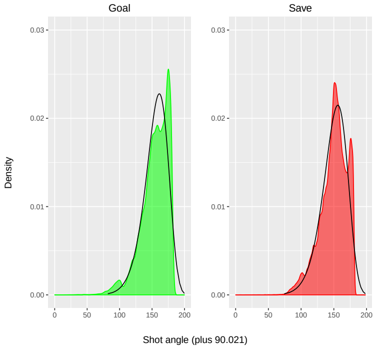
- Examined empirical densities to determine common parametric model to fit to each predictor
- In the spirit of the naïve approach, some approximations were crude, but chosen for simplicity
- Parameters were estimated by maximum likelihood once a model was chosen

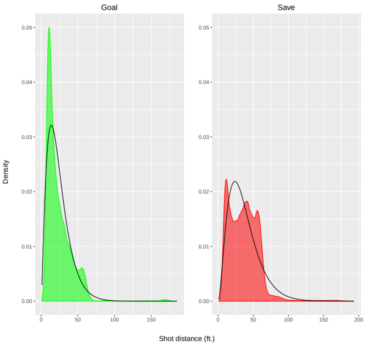
Parametric Naïve Bayes' (PNB)

Predictor	Parametric distribution
Angle	Weibull
Distance	Gamma
Height	Normal
Minute	Weighted Uniform
Weight	Normal
Catch	Binomial
Game type	Binomial
Home	Binomial
Manpower	Multinomial
Position	Binomial
Shot side	Binomial
Type	Multinomial

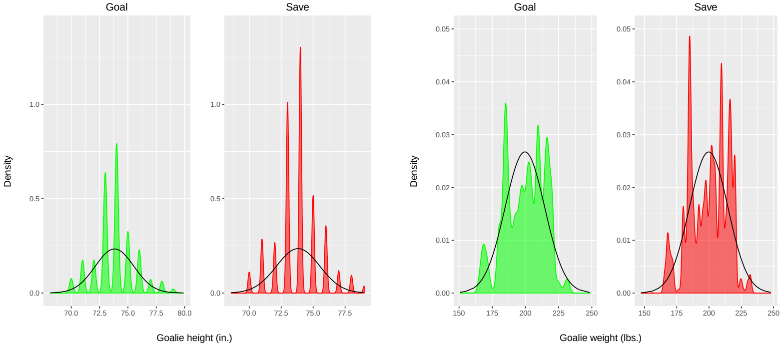
- For categorical predictors, ML estimates are just the sample proportions, so no difference occurred between ENB and PNB

Angle and Distance

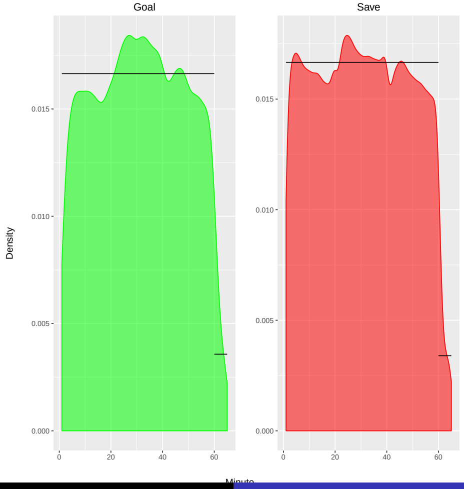




Height and Weight



Minute



Model Evaluation

If $\hat{p}_i$ is the predicted probability, then for a given classification threshold, $t \in [0, 1]$, we can define a classification as

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq t \\ 0 & \text{if } \hat{p}_i < t \end{cases} \tag{5}$$

Model Evaluation

If y_i is the observed outcome of the i^{th} shot and n is the total number of shots taken, then

Model Evaluation

If y_i is the observed outcome of the i^{th} shot and n is the total number of shots taken, then

$$error\ rate \rightarrow ER = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)}{n} \tag{6}$$

$$false\ positive\ rate \rightarrow FPR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)(1 - y_i)}{n - \sum_{i=1}^{n} y_i} \tag{7}$$

$$false\ negative\ rate \rightarrow FNR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i) y_i}{\sum_{i=1}^{n} y_i} \tag{8}$$

where

$$\mathbb{1}(y_i \neq \hat{y}_i) = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{if } y_i = \hat{y}_i \end{cases} \tag{9}$$

Model Evaluation

If y_i is the observed outcome of the i^{th} shot and n is the total number of shots taken, then

$$error\ rate \rightarrow ER = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)}{n} \tag{6}$$

$$false\ positive\ rate \rightarrow FPR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)(1 - y_i)}{n - \sum_{i=1}^{n} y_i} \tag{7}$$

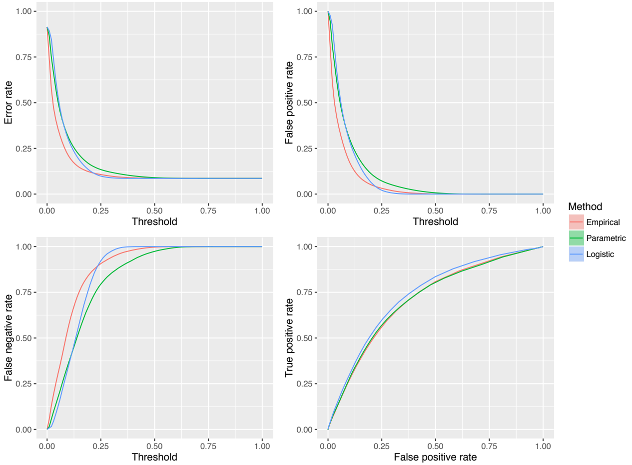
$$false\ negative\ rate \rightarrow FNR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i) y_i}{\sum_{i=1}^{n} y_i} \tag{8}$$

where

$$\mathbb{1}(y_i \neq \hat{y}_i) = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{if } y_i = \hat{y}_i \end{cases} \tag{9}$$

*10-fold CV was carried out to obtain accurate estimates of each of the three error measures at a set of thresholds {0,0.01,0.02,...,0.99,1}.

Model comparison



Model comparison

		Error rate			False postive rate		
Model	Optimal Threshold	ENB	Logistic	PNB	ENB	Logistic	PNB
ENB	.0566	.3340	.4940	.4717	.3339	.5263	.4975
Logistic	.0932	.2229	.3169	.3271	.1940	.3169	.3259
PNB	.0913	.2271	.3231	.3325	.1994	.3245	.3325

False negative rate			ROC curve
ENB	Logistic	PNB	AUC
.3340	.1503	.1974	.7117
.5312	.3170	.3394	.7351
.5213	.3080	.3325	.7122

Implications

- ENB better for identifying where players should *not* shoot from (maximum *true negative rate*)

Implications

- ENB better for identifying where players should *not* shoot from (maximum *true negative rate*)
- PNB and logistic regression better for identifying where players *should* shoot from (maximum *true positive rate*)

Implications

- ENB better for identifying where players should *not* shoot from (maximum *true negative rate*)
- PNB and logistic regression better for identifying where players *should* shoot from (maximum *true positive rate*)
- Use combination of methods depending on strategic approach (offense/defense)

R shiny application

https://alexzajichek.shinyapps.io/nhlshiny/

Future Work

- Build more complex model by taking into account individual skill, and team skill

Future Work

- Build more complex model by taking into account individual skill, and team skill
- Broaden the scope of the analysis to account for shoot-outs and all of overtime

References

Ekstrm, C. (2016). MESS: Miscellaneous Esoteric Statistical Scripts. R package version 0.4-3.

Grothendieck, G. (2014). sqldf: Perform SQL Selects on R Data Frames. R package version 0.4-10.

Hastie, T., James, G., Tibshirani, R., and Witten, D. (2014). An Introduction to Statistical Learning with Applications in R. Springer, New York.

Jensen, S. (2013). Measuring player contributions in hockey. Chance, 26(3):34-38.

References

Jurafsky, D. (2011). Text classification and naive bayes'. University Lecture.

Murphy, K. P. (2006). Binomial and multinomial distributions. The University of British Columbia Lecture.

R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Schuckers, M. E. (2016). Draft by numbers: Using data and analytics to improve national hockey league (nhl) player selection. MIT Sloan: Sports Analytics Conference, 1559.

References

Thomas, A. and Ventura, S. L. (2014). nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R. R package version 1.8.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S: fitdist. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. (2007). Reshaping data with the reshape package. Journal of Statistical Software, 21(12):1?20.

Wickham, H., James, D. A., and Falcon, S. (2014). RSQLite: SQLite Interface for R. R package version 1.0.0.