# Modeling the probability of an NHL goal for player placement strategy: A Naïve (Bayes') approach

Alex Zajichek

February 10, 2017

**Abstract**

This paper proposes a naïve approach to model the probability of an NHL goal being scored based on shot attributes, goalie and shooter information, and game situation to give insight into player placement strategy by identifying favorable and unfavorable shots and locations on the ice. Naïve Bayes' models were built to obtain probabilities in two formulations: estimation from empirical densities, and estimation from parametric models fit to the data. Every shot taken in the NHL from 2007-2015 within 65 minutes of game-play was used in model fitting. Previous works have proposed logistic regression models for similar objectives, so it was of interest to be implemented here for comparison purposes. For each model, 50 replications of 10-fold cross-validation were carried out to get estimates of the overall error rate, false positive rates, and false negative rates at a set of thresholds representative of $[0, 1]$. The Empirical Naïve Bayes' (ENB) model had the smallest false positive and overall error rate, but largest false negative rate out of the three. There were negligible differences in the error measures among the Parametric Naïve Bayes' (PNB) and logistic regression models, as well as the *area under the curve* (AUC) of the *receiver operating characteristic* (ROC) curves for all three models. The strategic implications of the results suggest the ENB model be preferable for identifying shots/locations where shots are more likely to be saved, whereas the PNB and logistic regression models are better suited for identifying shots/locations where shots are more likely to be goals. An R shiny web application was also developed which implemented the ENB model. It allows the user to choose various input variables to calculate predicted probabilities of goals being scored, which are displayed on a heat map overlaying an ice rink.

## 1 Introduction

In recent advances of technology and computing power, data-driven decision making has become very prevalent in the professional sports industry. In particular, many National Hockey League (NHL) teams have begun using data analytics for everything from in-game decision making to player acquisition (Schuckers, 2016). In such a fast-paced sport, in-game events tend to be difficult to model due to their random nature. This paper proposes a Naïve Bayes' approach to model the probability of a goal being scored in an NHL game given various information about the shot that was taken, the goalie and shooter, and game situation. Being able to accurately quantify this phenomenon could potentially allow teams to have better understanding of the characteristics of a shot that may be more

probable to lead to a goal or lead to a save. This information could then be utilized to develop player-placement strategy for specific in-game situations in a given matchup. Previous work has proposed logistic regression as a method for modeling this probability, in which the main objective was to evaluate player contribution towards a goal being scored (Jensen, 2013). The Naïve Bayes' approach serves as an alternative to the logistic regression model and will be compared and contrasted. Various packages and functions in the R statistical language will be used throughout.

# 2 Data

The data used in the analysis was obtained from NHL.com via the `nhlscrapr` package (Thomas and Ventura, 2014). Play by play game data was downloaded for every NHL game from 2002-2015. Each row of the original data gave information about a given event occurring during a particular game at a particular time, displaying the event that occurred, names of the players on the ice, names of the players involved in the event, etc. It was then stored in a local SQLite database for ease of access by the `RSQLite` package (Wickham et al., 2014).

The original data was far from being in the form needed to carry out the analysis. A large amount of time was spent manipulating the data to be in the correct form using the `sqldf` and `reshape2` packages (Grothendieck, 2014; Wickham, 2007). This involved identifying shots taken during penalties and with pulled goalies based on the distribution of players on the ice for each team at a given time, calculating shot angles based on the coordinates of the shot and the distance it was taken from, and manually inputting demographic information for the goalie that a given shot was taken on. Due to data and model restrictions, all observations in the years 2002-2006, all shots taken after 65 minutes of gameplay, and all shots that weren't considered a goal or a save were omitted from the analysis.

## 2.1 Predictors

After the data was explored, the following predictors were used to model the probability of a goal being scored on a given shot: *angle, catch, distance, game type, height, home, manpower, minute, position, shot side, type, weight.* These variables contain demographic information about the goalie, attributes of the shot taken, as well as game situation. Refer to Table 2.1 in the appendix for detailed descriptions of each variable.

# 3 Naïve Bayes' Methodology

The Naïve Bayes' framework has been around for a while. It uses a direct application of Bayes' theorem to obtain conditional probabilities, but instead of finding the joint-conditional distribution of random variables, it allows a (naïve) assumption of conditional independence. In recent years, it has been used extensively in machine learning applications for things such as text classification, among others (Jurafsky, 2011).

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. If we denote $x_{ij}$ as the $j^{th}$ predictor in the vector, then, as noted in Table 2.1, $j = 1, 2, ..., 5$ are the indices for continuous

predictors, and $j = 6, 7, ..., 12$ for categorical predictors. Table 3.1 gives general notation for the distributions of continuous and categorical predictors, as well as the marginal probabilities of a shot being scored or saved. Out of the 579181 shots considered in the original data, 8.6% were goals, leaving 91.4% being saves. We will allow this notation to represent the densities and probability mass functions for both modeling frameworks that will be introduced in Sections 3.1 and 3.2.

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

Table 3.1: The first column represents notation for the density of the $j^{th}$ continuous predictor ($j = 1, ..., 5$), the second represents the notation for probability mass function of the $j^{th}$categorical predictor ($j = 6, ..., 12$), and the third gives the marginal probabilities for each outcome.

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

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

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

where $G_i$ contains the terms for a goal, and $S_i$ contains the terms for a save, then the probability of a goal being scored given information about the shot can be obtained:

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

As noted, equation (4) uses an assumption of conditional independence of the predictor variables given the outcome of the shot. This allows us to find the product of each univariate distribution evaluated at the observed outcome rather than finding the joint-conditional distributions. Note that for continuous variables, because of the normalization term, we only need to evaluate the density functions at the observed outcome to carry out the derivation. From this point on we will refer to $\hat{p}_i$ as the *estimate* of equation (4) for the $i^{th}$ shot from the data.

## 3.1 Empirical Naïve Bayes'

The Empirical Naïve Bayes' (ENB) model was developed by assuming no parametric form for the observed data, using only the empirical densities of each predictor variable for estimation.

**(i)**

For the continuous predictors, the `density` function was used to obtain the empirical densities. The `approxfun` function was then applied to these objects to approximate a smooth density function, allowing for new inputs to be evaluated (R Core Team, 2016). For example, to find the conditional empirical densities of the *minute* variable, the following code was executed:

```
g_min = approxfun(density(g$Minute)$x,density(g$Minute)$y,yleft = .00000001, yright = .00000001) #Y=1
s_min = approxfun(density(s$Minute)$x,density(s$Minute)$y,yleft = .00000001, yright = .00000001) #Y=0
```

where `g_min(60)` would represent an estimate of $f_4(60|Y_i = 1)$, the conditional density of the *minute* variable for shots that were goals evaluated at 60.

**(ii)**

The conditional empirical densities for categorical predictors were found by finding the proportion of observations belonging to each level. For example, the conditional densities of the *home* variable were obtained by the following:

```
g_home = table(g$home)/length(g$home) #Y=1
s_home = table(s$home)/length(s$home) #Y=0
```

where `g_home` then consists of the proportion home and away games for goals scored, which are estimates of the values $P_8(X_{i8} = 1|Y_i = 1)$ and $P_8(X_{i8} = 1|Y_i = 0)$, respectively.

The process displayed in (i) was carried out for all continuous predictor variables and (ii) was carried out for all categorical predictor variables. With these terms, the probability of a shot becoming a goal could be estimated using equation (4).

## 3.2 Parametric Naïve Bayes'

In contrast to the ENB model, a Parametric Naïve Bayes' (PNB) model was also developed. In this setting, the conditional empirical densities displayed in Figure 3.1 were analyzed to determine an appropriate well-known parametric distribution to be fit to the data via *maximum likelihood*. Table 3.2.1 gives the chosen distribution for each variable. In the same spirit of the naïve approach, some of these approximations were crude, but chosen for the sake of simplicity.

| 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 |

Table 3.2.1: Shows the parametric distribution fit to each predictor variable for the PNB model. Maximum likelihood estimation was used to estimate parameters of each distribution.

For continuous variables, once a sufficient distribution was chosen for a particular variable, the `fitdistr` function was used to estimate parameters (Venables and Ripley, 2002). For example, the *distance* variable, which describes how far from the net the shot was taken from, was chosen to have a gamma distribution. The following code was executed to obtain the estimates:

```
m_g_dist = fitdistr(g$distance, 'gamma')$estimate #Y=1
s_g_dist = fitdistr(s$distance, 'gamma')$estimate #Y=0
```

where `m_g_dist` then contains the maximum likelihood estimates for the shape ($\alpha$) and scale ($\beta$) parameters of the gamma distribution for the *distance* variable for shots that became goals. Referring to the notation in Table 3.1, $X_{i2}|Y_i = 1 \sim Gamma(\alpha, \beta)$, leading to $f_2(x_{i2}|Y_i = 1)$ being estimated by the density of a gamma distribution. Except for the *minute* variable, this process was carried out for all continuous predictors.

In the PNB framework, the *minute* variable was chosen to have a *weighted uniform* conditional distribution. Let $r_1$ be the proportion of goals scored and $r_0$ be the proportion of saves in an NHL game during *regulation*. Similarly, let $o_1$ be the proportion of goals scored and $o_0$ be the proportion of saves in an NHL game within five minutes of *overtime* (again, all other shots were excluded from the analysis). Thus, $r_1 + o_1 = 1$ and $r_0 + o_0 = 1$. Therefore,

$$f_4(x_{i4}|Y_i = 1) = \begin{cases} \frac{r_1}{59} & 1 \leq x_{i4} \leq 60 \\ \frac{o_1}{5} & 60 < x_{i4} \leq 65 \end{cases} \tag{5}$$

is the estimated conditional density of *minute* for goals scored, and

$$f_4(x_{i4}|Y_i = 0) = \begin{cases} \frac{r_0}{59} & 1 \leq x_{i4} \leq 60 \\ \frac{o_0}{5} & 60 < x_{i4} \leq 65 \end{cases} \tag{6}$$

is the estimated conditional density of *minute* for saves. We can observe that the effect of the *minute* variable on predicted probabilities from equation (4) is only dependent on the values of $r_0, o_0, r_1, o_1$.

In addition, it did not make practical sense to assume a uniform density for all shots taken, but it did seem to be appropriate given when game play was occurring: *regulation* or *overtime*.

The conditional parametric distributions fit to categorical variables resulted in the same estimated conditional probabilities as in Section 3.1. Each categorical variable was determined either binomial or multinomial, leading to maximum likelihood estimates simply being the sample proportions belonging to each level for a particular variable (Murphy, 2006). In effect, there were no differences among the probability masses of categorical predictors in the ENB and PNB models. Figure 3.2 displays bar plots of each categorical predictor separated by goals and saves.

### 3.3 Logistic Regression

As previously mentioned, logistic regression models have been proposed to examine player contribution towards the probability of a goal being scored (Jensen, 2013). Logistic regression was also implemented here to compare performance to the ENB and PNB models. Specifically, if $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{12})^T$ is a 12 x 1 parameter vector and we assume

$$log\left(\frac{P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)}{P(Y_i = 0|\mathbf{X}_i = \mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{12} x_{i12} = \beta_0 + \mathbf{x}_i\boldsymbol{\beta} \tag{7}$$

implying

$$P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i\boldsymbol{\beta}}} \tag{8}$$

then the optimal value of $\boldsymbol{\beta}$ is found by maximizing

$$\begin{aligned} L(\boldsymbol{\beta}|data) &= \prod_{i=1}^{n} P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)^{y_i} P(Y_i = 0|\mathbf{X}_i = \mathbf{x}_i)^{(1-y_i)} \\ &= \prod_{i=1}^{n} \left(\frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i\boldsymbol{\beta}}}\right)^{y_i} \left(1 - \frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i\boldsymbol{\beta}}}\right)^{(1-y_i)} \end{aligned} \tag{9}$$

with respect to $\boldsymbol{\beta}$, which is the *likelihood* function (Hastie et al., 2014). The `glm` function was used to estimate parameters (R Core Team, 2016).

## 4 Model Evaluation

To compare and evaluate the performance of the three models, three measures of classification error were considered. For a given classification threshold, $t \in [0, 1]$, we can define a predicted classification as

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq t \\ 0 & \text{if } \hat{p}_i < t \end{cases} \tag{10}$$

If $y_i$ is the observed outcome of the $i^{th}$ shot as defined in equation (1), and $n$ is the total number of shots taken, then

$$error\ rate \rightarrow ER = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)}{n} \tag{11}$$

$$false\ positive\ rate \rightarrow FPR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i)(1 - y_i)}{n - \sum_{i=1}^{n} y_i} \tag{12}$$

$$false\ negative\ rate \rightarrow FNR = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i \neq \hat{y}_i) y_i}{\sum_{i=1}^{n} y_i} \tag{13}$$

where

$$\mathbb{1}(y_i \neq \hat{y}_i) = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{if } y_i = \hat{y}_i \end{cases} \tag{14}$$

It was then of interest to obtain accurate estimates of each of the three error measures at a collection of thresholds representative of $t$. Namely, the chosen set of thresholds was {0,0.01,0.02,...,0.99,1}.

## 4.1 Cross-validation

To get an accurate estimate of the error measures, *10-fold cross validation* (10-fold CV) was implemented by randomly partitioning the row indices of the data into ten non-overlapping sets, fitting the model on the rows corresponding to the indices of the union of nine subsets, then predicting the outcome of the response corresponding to the remaining subset (Hastie et al., 2014). This process was then repeated for all ten subsets, allowing each observation to be predicted without being used in the model fitting. The average error measure of all ten subsets was then obtained.

Specifically, if $P_k$ (k=1,2,...,10) contains the indices for the $k^{th}$ partition, $P_k \cap P_m = \emptyset$ for $k \neq m$, and $ER_k, FPR_k, FNR_k$ are the three error measures for the $k^{th}$ partition, respectively, then the following loop can be carried out:

for k in 1:10
  calculate $ER_k$
  calculate $FPR_k$
  calculate $FNR_k$

then,

$$ER_{cv} = \frac{\sum_{k=1}^{10} ER_k}{10} \tag{15}$$

$$FPR_{cv} = \frac{\sum_{k=1}^{10} FPR_k}{10} \tag{16}$$

$$FNR_{cv} = \frac{\sum_{k=1}^{10} FNR_k}{10} \tag{17}$$

This process was carried out for 50 replications in order to assess the stability of the cross-validation error estimates. The resulting standard error estimates of the replications were extremely negligible, implying very similar error measures across iterations.

## 4.2 Comparison

Figure 4.1 shows four different plots. The estimated $ER$, $FPR$, and $FNR$ at each threshold level for each model are shown in the upper-left, upper-right, and lower-left, respectively. The remaining plot contains the *receiver operating characteristics* ($ROC$) curves for each of the models (Hastie et al., 2014). The ROC curve is a popular evaluation method for binary classification models which plots the $FPR$ v.s. $1 - FNR$ (known as the *true positive rate*) for thresholds in $[0, 1]$. The $ROC$ curve of an ideal classifier would "hug" the upper-left corner of the plot, giving maximum area under the curve. The right-most column of Table 4.1 displays the *area under the ROC curve* ($AUC$) for each of the three models (Ekstrm, 2016). We observe very similar estimates of $AUC$, with the logistic regression model having the highest by roughly 2 percentage points. This indicates a very similar overall classification ability in the three classifiers, giving a small advantage to the logistic regression model according to $AUC$.

There is not a notable difference among the PNB and logistic regression models according to the $ER$, $FPR$, and $FNR$. The ENB model, however, strays away from the other two. Figure 4.1 shows a relatively large advantage for using the ENB model if an objective is to minimize the $FPR$, $ER$, or both. A consequence though is that the $FNR$ is increased.

So what threshold should be used for classification? Because the occurrence of a goal being scored in an NHL game is small relative to the number of shots taken, a traditional threshold of 0.5 is clearly a poor choice. The $FNR$ is near 1.0 in all three models at this threshold. One way to choose a better classification threshold is to find the *optimal threshold*, which can be defined as "the threshold at which the *threshold v.s. FPR*, and *threshold v.s. FNR* curves intersect." Besides the AUC, Table 4.1 also displays the *optimal threshold* for each of the three models. At each of these thresholds, the $ER$, $FPR$, and $FNR$ were calculated for each model. This allowed comparison of the model performances at reasonable threshold levels. We can observe that at each reported threshold, the $ER$ and $FPR$ of the ENB model are significantly lower than those of the other two. Again, the consequence of this performance is a poor $FNR$ for the ENB model relative to the PNB and logistic regression models (Hastie et al., 2014).

| | | Error rate | | | False postive rate | | | False negative rate | | | ROC curve |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Optimal Threshold | ENB | Logistic | PNB | ENB | Logistic | PNB | ENB | Logistic | PNB | AUC |
| ENB | .0566 | .3340 | .4940 | .4717 | .3339 | .5263 | .4975 | .3340 | .1503 | .1974 | .7117 |
| Logistic | .0932 | .2229 | .3169 | .3271 | .1940 | .3169 | .3259 | .5312 | .3170 | .3394 | .7351 |
| PNB | .0913 | .2271 | .3231 | .3325 | .1994 | .3245 | .3325 | .5213 | .3080 | .3325 | .7122 |

Table 4.1: At the *optimal threshold* of each model, the overall error rate, false positive rate, and false negative rate were calculated for all three models. The rightmost column gives the AUC of the ROC curves shown in Figure 4.1.

# 5 Discussion

Sports in general are difficult to quantify due to the fast-pace and randomness of events occurring. Specifically, goals being scored in hockey are at times unpredictable due to deflections, bouncing pucks, tip-ins, and more, making even statistically sound strategies vulnerable to error. The goal of the ENB and PNB models were to induce probabilities on shots taken in order to strategically identify and place players in situations where the chance of scoring is the largest (or smallest). But because of the low rate of scoring in hockey relative to the number of shots taken, the resulting probabilities

are small leaving the models susceptible to large error rates. Possible implications of the results for a team's strategy of player placement, supported in Table 4.1 and Figure 4.1, can be explained from two perspectives:

**Offense**

For a given match-up, a team may be interested in avoiding shots in the *offensive zone* in which goals would not be scored. This corresponds to minimizing the $FPR$, in turn maximizing the *true negative rate*. The ENB model is the one that does this out of the three. In context, a false positive error would be classifying a shot as a goal in which the goalie ends up saving. By minimizing this error, players would be taking less shots that would end up being saved, allowing them to focus on more favorable locations. The $FNR$, which would be classifying a shot as a save in which ends up being a goal, may not be as important to diminish from an offensive perspective, because the hockey team's objective is to score goals. If they end up taking a shot that is not predicted to be a goal, and it is, they will still be satisfied with the outcome. On the contrary, if a team's objective is to strategically place players in locations where a goal *will* be scored, they would want the $FNR$ to be minimized, which implies maximizing the *true positive rate*. Because the PNB and logistic regression models gave similar error measures, either of these would be a better option for this objective. In conclusion, offensively, the ENB model is best for identifying locations/shots that players should not take, where the PNB and logistic models are better for identifying locations/shots that players should take. In any single model, there will always be a trade off in error measures, so using multiple models in a player placement strategy would give the best outcome.

**Defense**

On the other end of the ice, a team may want to develop player placement strategy for the *defensive zone*. Similar to the offensive perspective, each model will have benefits and drawbacks depending on the objective of the strategy. By minimizing the $FPR$ from a defensive perspective, a team would be able to better identify shots/locations that their opponents would be less likely to score from. A team could then put in place strategies to force the opponent to take unfavorable shots from unfavorable locations. Again, to achieve this objective, the ENB model would do the best job out of the three. The minimization of the $FNR$ would allow a team to better identify shots/locations that their opponent *is* likely to score from, giving the defense the ability to try to force the opponent out of favorable shot locations or the goalie the opportunity to prepare for certain types of shots. The PNB or logistic regression model would be preferable for this objective. Like the offensive strategy, each model alone has drawbacks, so using multiple models in the defensive perspective would also give the better outcome.

# 6 Web application via R shiny

A web application was built in R shiny to further investigate and visualize the ENB model, which was chosen for computational advantages. It allows the user to visualize predicted probabilities on a heat map which overlays part of an ice rink. The color scale used remains static as inputs change allowing the user to see relatively small changes in predicted probabilities. Figure 6.1 shows a partial screenshot of the application. The *distance*, *angle* and *shot side* variables are built into the heat map in order to get probability estimates at each coordinate in the plane. The left-hand panel displays the

categorical predictor variables with drop-down menus to select specific levels for prediction, as well as a slider bar to select the minute of the game. In addition, the user also has the option to leave a variable blank which then omits it from prediction. This feature gives the capability to see how subsets of predictor variables change the predicted probabilities on top of the level-specific changes.

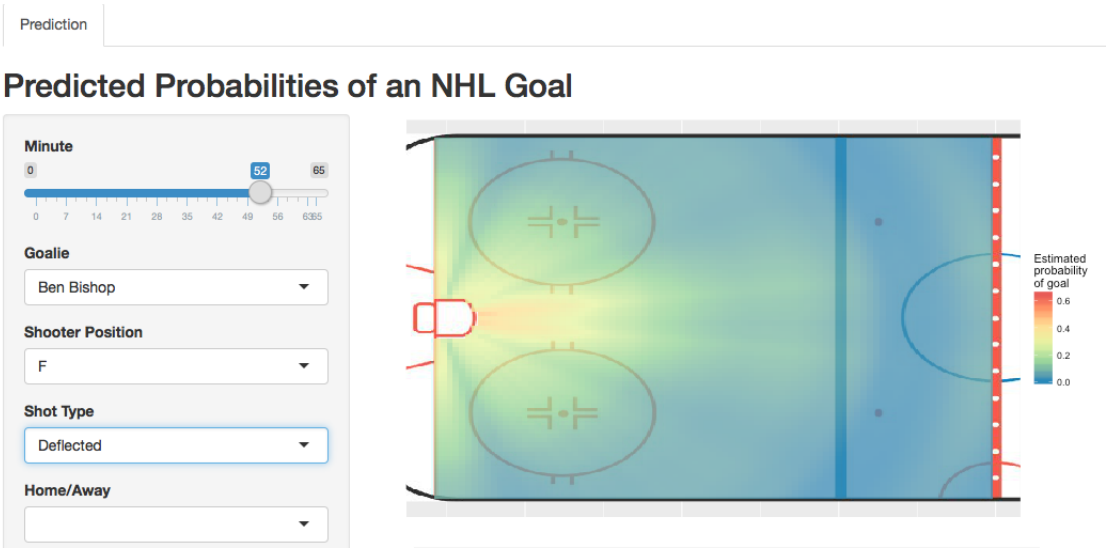


Figure 6.1: Partial screenshot of an R shiny web application implementing the ENB model. The application allows the user to choose levels of various predictor variables and displays the resulting probabilities on a heat map that overlays an ice rink.

The goalie names available for selection are most of the goalies in the 2016-2017 NHL season according to NHL.com. In the modeling framework, a specific goalie was defined by the combination of their height, weight, and catching hand. Consequently, a specific combination may lead to multiple goalies, which is a one of the limitations of this framework.

# 7 Future work

A restriction of the chosen predictors is that they do not account for specific player or team ability. Future work may consider adding on to the chosen set of predictors by building more complex models that incorporate specific player and goalie skill sets, measures of team performance, and match-up specific attributes. This would allow predictions to be tailored towards a specific game plan by utilizing the differences that individual players and teams have.

The scope of this analysis is also somewhat limited. It does not account for shots missing the net, shots being taken in playoff games after the first five minutes of overtime, or shoot outs. These may be attractive cases to consider modeling because implications of goals/saves in overtime or a shoot out are far greater than a goal, say, in the middle of the first period.

# Appendix

.

| $j$ | Variable | Description |
|---|---|---|
| 1 | *angle* | The angle at which the shot was taken (degrees). |
| 2 | *distance* | The distance the shot was taken from (feet). |
| 3 | *height* | The height of the goalie in which the shot was taken on (inches). |
| 4 | *minute* | The minute of the game in which the shot was taken (0,65]. |
| 5 | *weight* | The weight of the goalie in which the shot was taken on (pounds) |
| 6 | *catch* | The catching hand of the goalie (left, right) |
| 7 | *game type* | Indicating a regular season or playoff game (regular, playoff) |
| 8 | *home* | Indicating if the goalie in which the shot was taken on is on the home team (home, away). |
| 9 | *manpower* | Manpower on the ice due to penalties or game situation from the goalie's perspective (even, short, power play, pulled). |
| 10 | *position* | The position of the player who shot the puck (forward, defense) |
| 11 | *shot side* | From the goalie's perspective, the side of the ice in which the puck was shot from (left, right). |
| 12 | *type* | Type of shot taken by the shooter (backhand, deflected, slap, snap, tip, wrap, wrist) |

Table 2.1: Descriptions of the variables used in the predictive models

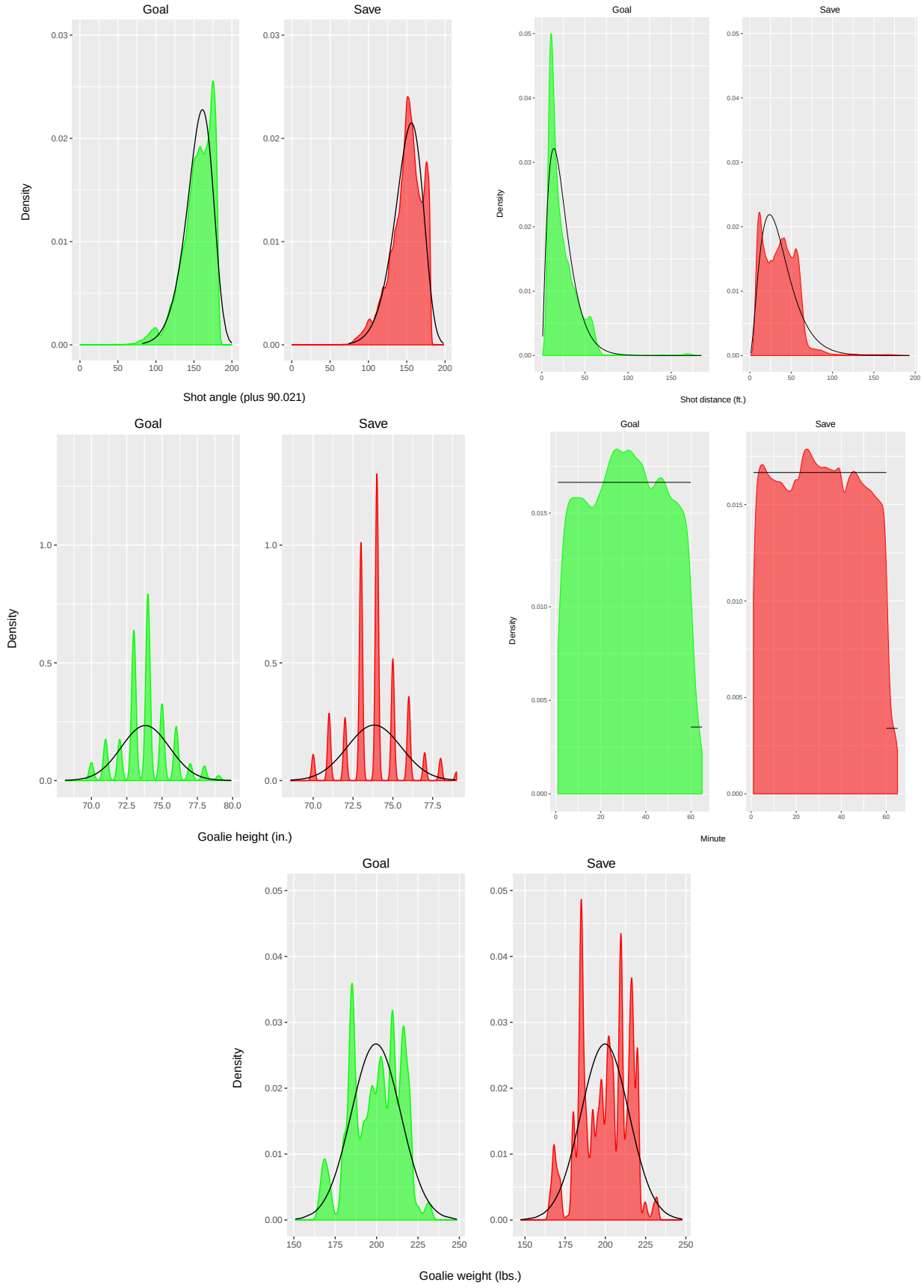


Figure 3.1: Empirical density of each continuous predictor variable with an overlay of the parametric distribution fit via maximum likelihood.

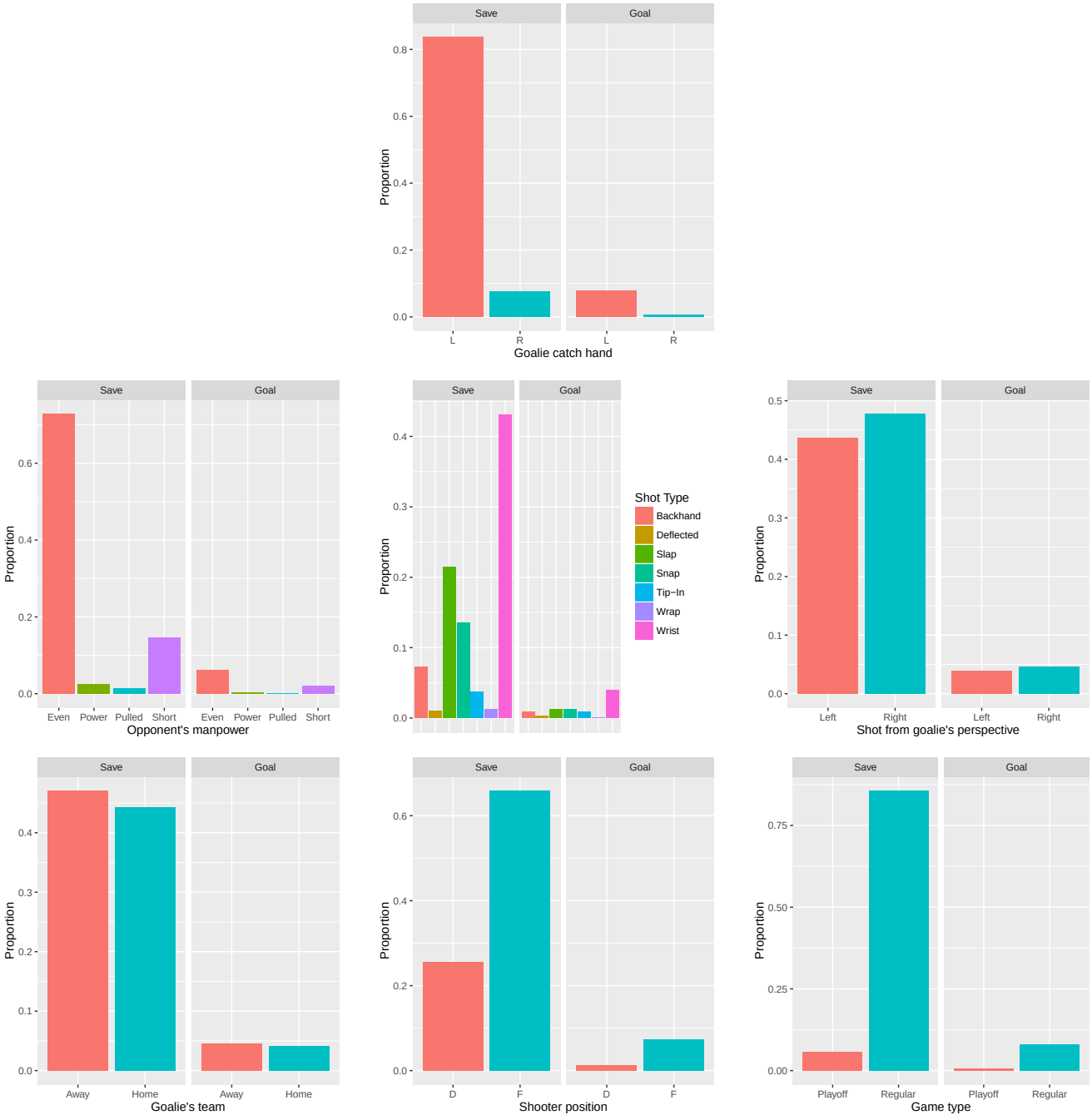


Figure 3.2: Bar plots for each categorical predictor separated by goals and saves.

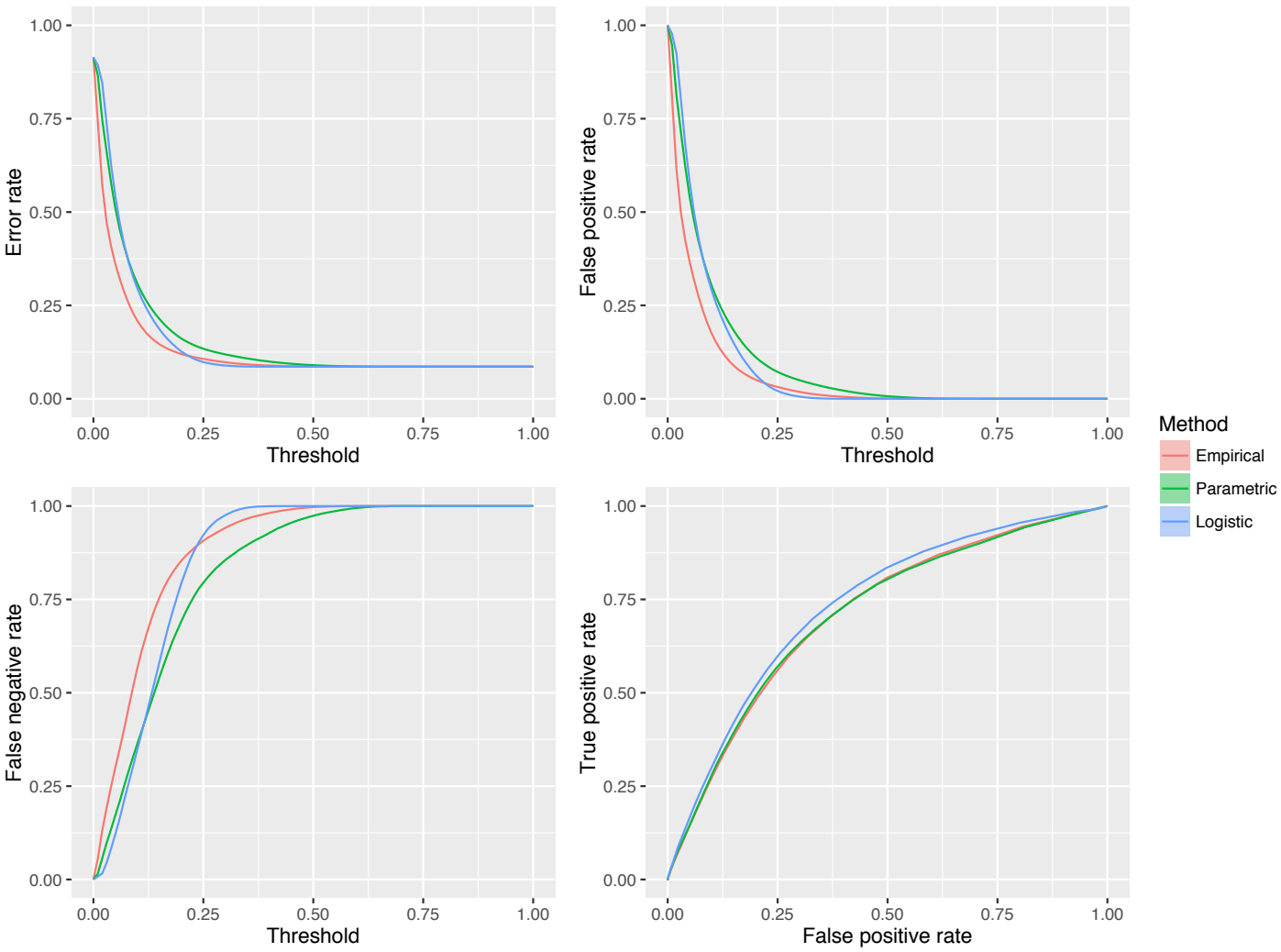


Figure 4.1: Upper left: Test error rates; Upper right: False positive rates; Lower left: False negative rates; Lower right: ROC curves; For each method, each type of error was calculated by averaging 50 replications of 10-fold cross-validation at classification thresholds in [0,1]. Estimated standard errors from the replications were negligible. The ENB model is red, PNB model is green, and logistic regression model is blue.

# 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.

Jurafsky, D. (2011). Text classification and naïve 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.

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.