Analysis of NHL Goalies in 2022

EXECUTIVE SUMMARY

The role of a goalie in ice hockey might appear straightforward, centered around one core objective: preventing the puck from entering the net. However, assessing goalie performance extends far beyond the simple tally of blocked shots. Numerous intricate factors contribute to each goal scored in the sport. Consequently, sophisticated statistical metrics have been developed to comprehensively analyze a goalie's effectiveness.

Recognized as the most pivotal player on the ice, a goalie's performance holds significant importance in every ice hockey game ever played. Understanding the multitude of variables impacting this performance is crucial to the success of a professional NHL team. Thus, this report aims to identify and evaluate the primary factors that influence goals-against. Using these factors, the report will also discover the answers to three important research questions. Using Principal Component Regression, how does the overall data look when predicting goals against? What is the most significant factor in determining an upper-escalante goalie in the NHL? How does the in-game situation affect the amount of goals scored?

Before predictive analysis, the validity of the recorded statistics must be examined. The dataset used can be found at [MoneyPuck.com -Download Data](https://moneypuck.com/data.htm), and was downloaded into R-Studio for analysis. This raw data contains many variables that include team statistics and individual goalie statistics. It was then manipulated in order to create a concise and clear dataset which gave the season total statistics for each goaltender. After deletion of the pre-calculated expected goals and subjective metrics, a PCR model was created to check the validity of the recorded statistics when predicting goals. Since we see an immediate drop of RSME after just one Principal Component(FIGURE 6). This means the fit can be explained by correlated phenomena and the variables used are extremely effective at predicting goals when combined into Principal Components. PC1 has just about the lowest amount of cross validation error among all 20 Components calculated and explains 94.29% of the variance in our response variable goals(FIGURE 7). This is due to high correlation among components and means that further analysis must be done in order to accurately predict goals against.

The most important factor in predicting goals in the NHL during the 2022 season was the number of times that the play continued in the offensive zone after a rebound(FIGURE 2), meaning the goalie did not freeze1 the play or the defending team did not clear their zone after the initial shot. This makes statistical sense and can be attributed to extremely important in-play factors such as defensive fatigue, offensive puck control, and positioning. The more time and possession a team has in the offensive zone, the more likely they are to score and this is displayed in all the Linear Models. An upper-escalante goaltender is one who is able to deal with these challenges and makes saves during an extended offensive possession by the opposing team.

Another massive factor when determining if a play results in a goal is the rebounds a goalie gives up. It’s not so much just about how many rebounds a goaltender allows, but the quality of those rebounds. Goalies are trained from a young age to control their rebounds either into the netting above the glass or into their own body in order to receive a stoppage in play. However, it is quite apparent in the data that when the play continues after a shot, there is a huge uptick in the amount of goals scored. What was surprising is that the model shows significance in predicting goals when the play continues outside of the attacking zone after a rebound(FIGURE 1). This is due to the nature of the game’s most scoring-abundant situation, the power-play2. The power-play is also the only time in the game when a stoppage will not be called if the penalized team ices3 the puck. This rule completely changes the style of the game as the puck is continuously being moved from one end of the ice to the other. Despite all this movement, the result of the play is much more likely to end in a goal because the penalized team is playing with 4 men instead of 5. Although the dataset used combined every type of situational goal, the models created give us the information needed to decipher just how important power-plays are in the NHL. We can also determine from this that the highest likelihood of scoring-situation in an NHL hockey game occurs while a team is on the powerplay and controls the puck after a rebound.

After analyzing the data, there are many interesting insights one can take away. The most shocking in my opinion is that rebounds are far more important than originally suspected. While I was aware of the importance of rebound control. I did not take into account the location or possession of the puck after the fact. It is clear now that NHL teams must focus on recruiting goaltenders who have excellent rebound control. Proper rebound control provides a stoppage of play and when the opposing team has an extended offensive possession, the goalie’s team is allowed time to rest, re-group, and recover. In such a fast paced game, a rested team is necessary to prevent goals against.

DATA & APPROACH

The main goal in this analysis was to take a deep dive into the in-game metrics recorded in the National Hockey League in order to predict the amount of goals scored on a certain goalie. A data containing NHL data for the 2022 goalies from moneypuck.com was downloaded and analyzed in R Studio. I consolidated the data by creating a new dataset containing every goalie’s season totals for each statistic.(The original data was broken up by in-play situations such as power-plays and penalty kills) Once this was done, a training dataset was created from it containing 70% of the observations and the remaining 30% left as a testing dataset. Then, regression techniques were applied to the season totals training data in order to predict the amount of goals scored on the goalies. Regression techniques such as linear, pairwise, subset selection, Principal Component, Partial Least Squares and Boosting were used.

DETAILED FINDINGS

LINEAR MODELS

-Deletion of the pre-calculated expected goal metrics and goal total metrics was done for the linear models. This was for the purpose of creating our own expected goal models and not using pre calculated prediction metrics by moneypuck.com.

–This includes variables like expected goals, lowDangergoals, and xHighDanger goals

-In the Linear models, variables that were unnecessary were not used, such as penalties and penalty minutes. These would be very interesting predictors if not for the fact that it calculated penalties for each goalie instead of against. Goalie penalties are extremely rare, so the observations were removed.

-A normal Linear model was constructed first, followed by a reduced model containing the significant factors. It is clear that the quality of a rebound is the most important factor when predicting goals. All three predictors have to do with the play after a rebound occurs. If the puck was not frozen, or collected by the defending team, there is a much higher likelihood of the play resulting in a goal.

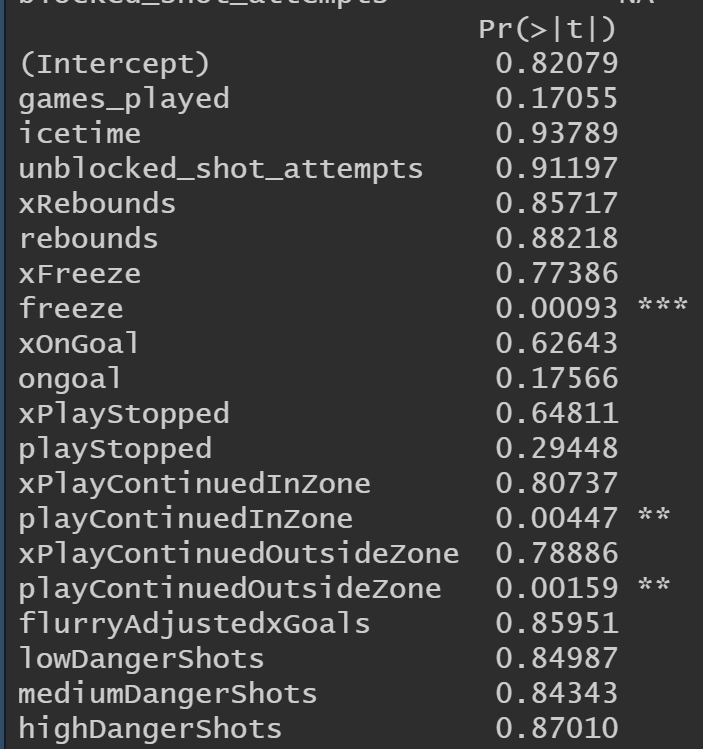
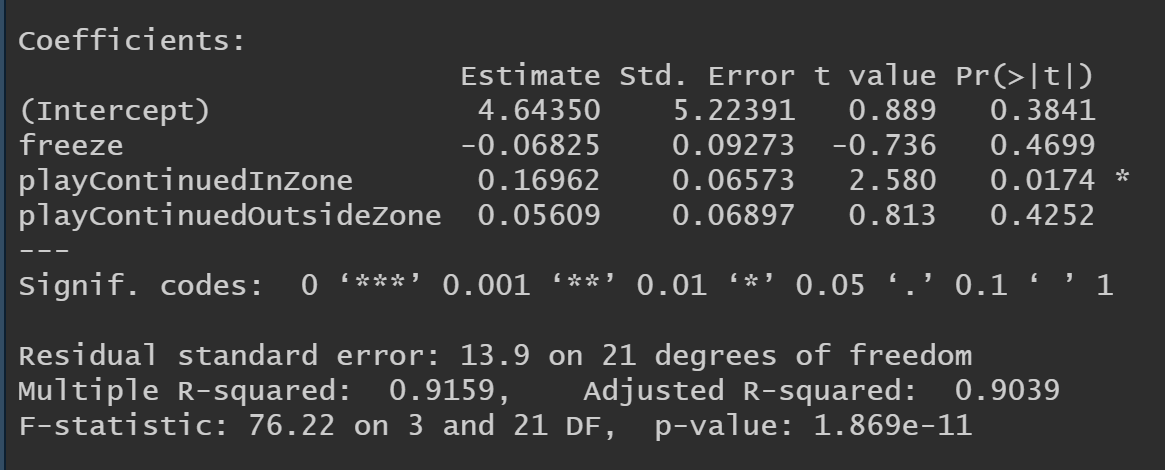


FIGURE 1 FIGURE 2

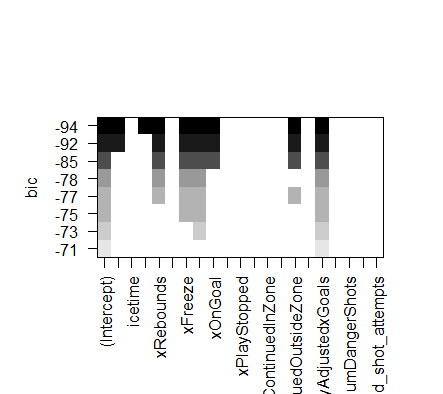
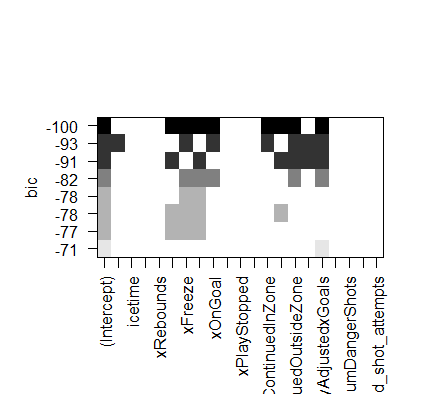
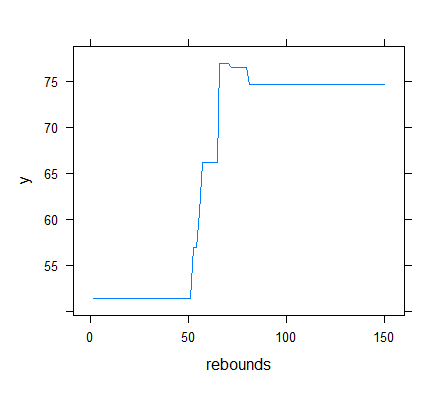


FIGURE 3 FIGURE 4

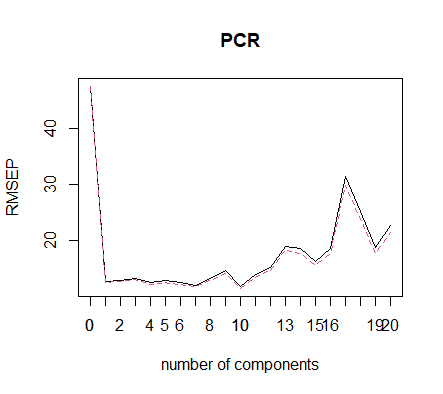
-We also see evidence of significance between these same three factors in the regression by subset selection (FIGURE 3) and Stepwise Regression (FIGURE 4)

There is also evidence in Figures 3 & 4 of significance within the rebound variable. This significance is backed by a boosting model, which displays Rebounds as the primary predictor (FIGURE 5)

FIGURE 5

Predicted Goals for Figure 5 is at zero until 50 rebounds because the goalies recorded all had at least that many.

PRINCIPAL COMPONENT REGRESSION



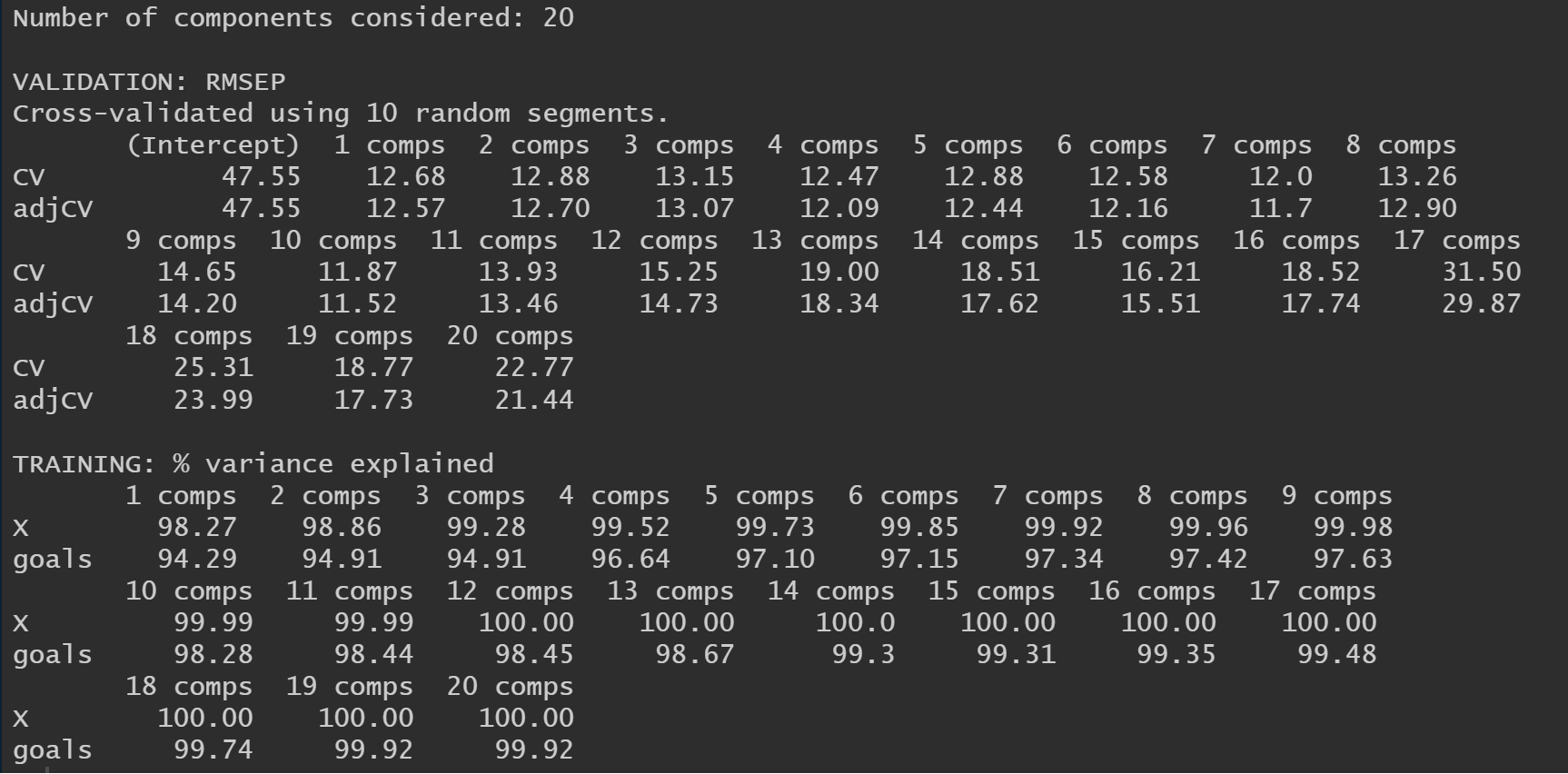
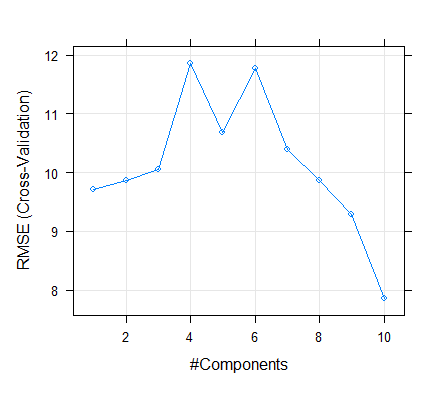


FIGURE 6 FIGURE 7

Using PCR we were able to determine that there may be some correlation within the dataset. With only one Principal Component being able to explain so much variation, it is clear that the components calculated were accurate and that our PCR is explaining as much variance within the dataset as possible. It was able to reduce dimensionality extremely quickly with only one Component.

PARTIAL LEAST SQUARES 

It is also interesting to note that a Partial Least Squares model looks a bit different compared to the PCR. I believe this is due to the bias reduction property that comes with PLS. Since we see such a quick explanation of variability in the first Principal Component, we can assume there exists some bias among variables. PLS looks to eliminate this bias which is why we can see a drop in RMSE after about 8 components.

VALIDITY & RELIABILITY ASSESSMENT

In the future, I believe the NHL will start focusing on scouting and recruiting goalies with outstanding rebound control. Now that the game is evolving and becoming faster every year. The ability to slow the game down and create a stoppage is increasingly valuable. Now that there are advanced metrics focused around rebounds and possession after the rebound, I hope to see NHL teams using these statistics in order to recruit the best goaltender. In such an intense and quick game, being able to control the pace of play is the most important factor for goals-against.

APPENDIX

Icing- During even strength play if a defending player shoots the puck from behind the middle red line, and it reaches the other side of the ice without another player touching the puck, a stoppage in play will be called and a face off in that defender's zone will occur.

Freeze- When a goalie gathers possession of the puck (other than it being on his stick). A stoppage of play will commence.

Power Play- When the opposing team is penalized. The other team will have a 1 to 2 man advantage called a powerplay

NHL- National Hockey League

R-CODE:

goalies <- read.csv("C:/Users/nicki/Downloads/goalies (1).csv")

season\_totals <- goalies[goalies$situation == 'all', ]

season\_totals <- season\_totals[, !(names(season\_totals) %in% c("season", "position","situation"))]

#Test 30% and Training 70%

train <- sample(1:nrow(season\_totals), 25)

goalies.training <- season\_totals[train, ]

goalies.testing <- season\_totals[-train, ]

y\_train <- goalies.training$goals

y\_test <- goalies.testing$goals

library(caret)

library(glmnet)

library(pls)

library(dplyr)

library(MASS)

library(class)

one\_hot\_encoding <- dummyVars(goals ~ ., data = goalies.training)

x\_train <- predict(one\_hot\_encoding, goalies.training)

x\_test <- predict(one\_hot\_encoding, goalies.testing)

#sapply(lapply(goalies.training, unique), length)

##LINEAR

lin\_model <- lm(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training)

summary(lin\_model)

plot(lin\_model)

lin\_model2 <- lm(goals ~ freeze + playContinuedInZone +playContinuedOutsideZone ,

data = goalies.training)

summary(lin\_model2)

##GLM

glm\_model <- glm(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals, data = goalies.training)

summary(glm\_model)

#Same exact results as linear model

##subset selection

library(leaps)

goalies\_subset = regsubsets(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals, data = goalies.training)

reg\_summary = summary(goalies\_subset)

par(mfrow = c(2,2))

plot(reg\_summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")

plot(reg\_summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l")

adj\_r2\_max = which.max(reg\_summary$adjr2)

points(adj\_r2\_max, reg\_summary$adjr2[adj\_r2\_max], col ="red", cex = 2, pch = 20)

plot(reg\_summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")

cp\_min = which.min(reg\_summary$cp) # 10

points(cp\_min, reg\_summary$cp[cp\_min], col = "red", cex = 2, pch = 20)

plot(reg\_summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")

bic\_min = which.min(reg\_summary$bic) # 6

points(bic\_min, reg\_summary$bic[bic\_min], col = "red", cex = 2, pch = 20)

coef(goalies\_subset, 8)

summary(goalies\_subset)

plot(goalies\_subset)

##pairwise

goalies\_pair = regsubsets(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training, method = "forward")

plot(goalies\_pair)

##ridge

ridge\_fit <- train(x = x\_train, y = y\_train,

method = 'glmnet',

trControl = trainControl(method = 'cv', number = 10),

tuneGrid = expand.grid(alpha = 0,

lambda = seq(0, 10e2, length.out = 20)))

ridge\_fit

coef(ridge\_fit$finalModel, ridge\_fit$bestTune$lambda)

##Lasso

lasso\_fit <- train(x = x\_train, y = y\_train,

method = 'glmnet',

trControl = trainControl(method = 'cv', number = 10),

tuneGrid = expand.grid(alpha = 1,

lambda = seq(0.0001, 1, length.out = 50)))

lasso\_fit

coef(lasso\_fit$finalModel, lasso\_fit$bestTune$lambda)

##Partial least squares

pls\_model <- train(x = x\_train, y = y\_train,

method = 'pls',

trControl = trainControl(method = 'cv', number = 10),

tuneGrid = expand.grid(ncomp = 1:10))

pls\_model

plot(pls\_model)

##Trees

library(tree)

Goalie.tree = tree(goals~.,data = goalies.training)

summary(Goalie.tree)

Goalie.tree

par(mfrow = c(1,1))

plot(Goalie.tree)

text(Goalie.tree)

#no need for pruning

##Bagging

library(randomForest)

bag\_goal = randomForest(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training,mtry = 10, importance = TRUE)

plot(bag\_goal, xlim=c(0,150))

axis(side=1, at=seq(0, 150, by=25))

plot(bag\_goal, xlim=c(0,55))

axis(side=1, at=seq(0, 55, by=5))

##random forest

rf.goalie <- randomForest(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training, mtry=3, importance=TRUE)

plot(rf.goalie, xlim=c(0,40))

axis(side=1, at=seq(0, 40, by=5))

##Bosting

library(gbm)

boost.goalie <- gbm(goals ~ .-xRebounds -games\_played -icetime -playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals -unblocked\_shot\_attempts,

data = goalies.training,

distribution = "gaussian", bag.fraction = 0.85)

plot(boost.goalie)

#knn

set.seed(1)

train.X = cbind(goalies.training$icetime)

test.X = cbind(goalies.testing$icetime)

train.Y = cbind(goalies.training$goals)

knn.pred = knn(train.X, test.X, train.Y, k=1)

table(knn.pred, goalies.testing$goals)

#Logisitic Regression

exp(coef(glm\_model))

#LDA

lda.fit1 = lda(goals ~ .-playerId -name -games\_played -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training)

lda.fit1

##QDA

qda.fit = qda(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies)

qda.fit

##dataset too small for QDA use LDA instead.

#PCR

pcr.fit <- pcr(goals ~ .-playerId -name -team -penalties -penalityMinutes

-xGoals -lowDangerGoals -highDangerGoals -mediumDangerGoals

-lowDangerxGoals -highDangerxGoals -mediumDangerxGoals,

data = goalies.training, scale = TRUE,

validation = "CV")

plot(pcr.fit, main = 'PCR')

summary(pcr.fit)

validationplot(pcr.fit, main = 'PCR')

axis(side=1, at=seq(0, 20, by=1))