Hockey Team Success: A Sum of Its Parts

Molly McNamara

Springboard Foundations of Data Science Capstone Project

Hockey and the Potential Impact of Data Science

The National Hockey League (NHL) is considered to be one of the four major professional sports leagues in the United States (though ranked behind baseball, football and basketball). The NHL operates between the United States and Canada (where hockey is by far the most popular sport). The economics of hockey are complicated by the shared border and exchange rates, but team success during the regular season and the playoffs ultimately drives financial success for the team and the league. Teams see an impact of improved performance and win totals on season ticket sales and sponsorship revenue.

Establishing the right balance of player skills and team needs to make a successful run at the playoffs is the challenge of every NHL general manager (GM). While in-game coaching decisions are left to the coaching staff, the GM is responsible for acquiring and dismissing players and coaches. A skilled GM evaluates the existing team members and identifies gaps or needs in order to find and obtain talent that will support and improve the team as a whole. Better analysis of player and league data can help guide a GM in improved decision making when acquiring team talent, hopefully resulting in a more successful team.

The Data

The project dataset is a combination of NHL player and team statistics for years 2000-2011 and NHL draft data from the late 1980s through 2010. Variables include player names, ages and birthdates, countries of origin, positions played, scoring and defensive statistics by year, junior level and NHL teams and leagues played in, coaching records, and more. The information was gathered from several sources and in several separate files and formats.

Data Wrangling

Player/team statistical data was available in CSV format, separated into files by topic; draft data required download, compilation in a TXT file and conversion of the entire dataset to CSV.

Five datasets were read into R: Master (basic demographic information on all NHL players from 1908 to 2011), Scoring (scoring statistics for all NHL players from 1908 to 2011), Teams (team statistics and playoff outcomes from 1908 to 2011), Goalies (defensive statistics for all NHL goaltenders from 1908 to 2011), and Draft (player draft information from 1979 to 2010). The date range for Draft data was determined by identifying the earliest drafted player still playing in the period of interest for this project (the project will

evaluate game data from 2000-2011 - a player drafted in 1979 was still playing in 2000) and the latest possible drafted player who would play in this time period (2010). Note that there is no data for 2004 due to a league-wide labor dispute (lockout) wherein the entire season was cancelled and no games were played.

The R packages used for data wrangling were as follows:

```
library(dplyr)
library(readr)
library(tidyr)
library(countrycode)
```

Columns for all 5 datasets were renamed to understand the variables more easily. In the Draft dataset, the Player column was split into two columns, separating the player name and ID and AmateurTeam/League was split into team and league. The Master, Scoring, Goalies and Draft datasets were then joined/merged by Player_ID and Year into one dataset (AllData).

Unnecessary or duplicative variables were removed. Multiple fields were converted to factors or integers as appropriate. Birth and death month/day/year columns were joined into a single date column for each event. The datasets were subset to the years 2000 to 2011.

The datasets were evaluated with the summary() command to assess the presence of any obvious outliers. NA values in both sets were identified. In the case of certain variables, such as death statistics, NA is appropriate as most of the players in the subset timeframe had not yet died. Scoring statistics that displayed NA (for example, a player did not score any goals in a given year) were replaced with a 0.

New variables were created for team playoff success, birth regions, draft round, and additional scoring and goaltending statistics. Finally, the cleaned datasets were merged by year and team and the new cleaned data file was written and saved as CSV file.

Final Dataset Composition

The variables in the final dataset are as follows:

- Team ID
- Conference and Division Team
- Season end rank and Playoff result
- Team made playoffs (0 or 1 – generated from original data)
- Team Games, Wins, Losses, Ties, Overtime Losses, Points, Shootout Wins and Shootout Losses
- Team Goals For and Goals Against
- Team Penalty Minutes and Bench Minors
- Year (Season)
- Player ID
- First Name and Last Name

- Height and Weight
- Shooting Hand
- First and Last NHL season
- Years of Experience (calculated from original data)
- Position Played
- Birth Country, Birth State/Province, Birth City
- Birthdate and Deathdate
- Birth Region (determined from original data)
- Stint with team
- Games Played
- Goals, Assists and Points
- Penalty Minutes
- Plus Minus Rating
- Powerplay Goals and Assists
- Shorthanded Goals and Assists
- Gamewinning Goals and Gametying Goals
- Shots
- Goals and Shots Per Game (calculated from original data)
- Percent of Team Goals Scored (calculated from original data)
- Percent of Team Games Played (calculated from original data)
- Draft Pick
- Draft Round (calculated from original data)
- Draft Year
- Draft Team
- Draft Age
- Amateur Team and League
- Goalie Minutes Played
- Goalie Wins and Losses
- Goalie Shutouts
- Goalie Goals Against Goalie Shots Against
- Goalie Save Percentage (calculated from original data)

Limitations

Several limitations are readily apparent in the dataset prior to analysis: (a) the lack of player statistics prior to draft which might be especially predictive of future performance; (b) lack of draft combine results which is kept confidential by the NHL; and (c) Time On Ice (TOI), a well-known statistic that tracks how much players are used in the game, was not available.

Preliminary Exploration

The purpose of the initial data exploration was to screen for any obvious trends in the data and evaluate the relationships between player demographics and characteristics and teams making the playoffs.

Player birth countries were mapped to a world map to evaluate the many origins of players in this dataset.

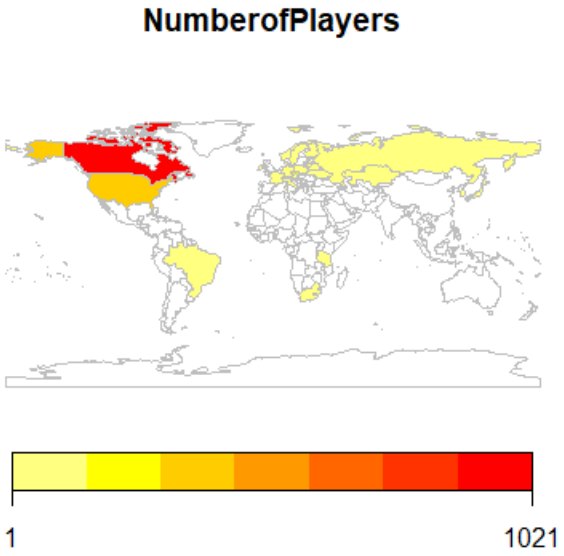

Figure 1: Player Countries of Origin

While it appears drafted players originated from many different countries, the NHL is based in the US and Canada; it might be useful to see regionally where the most players come from.

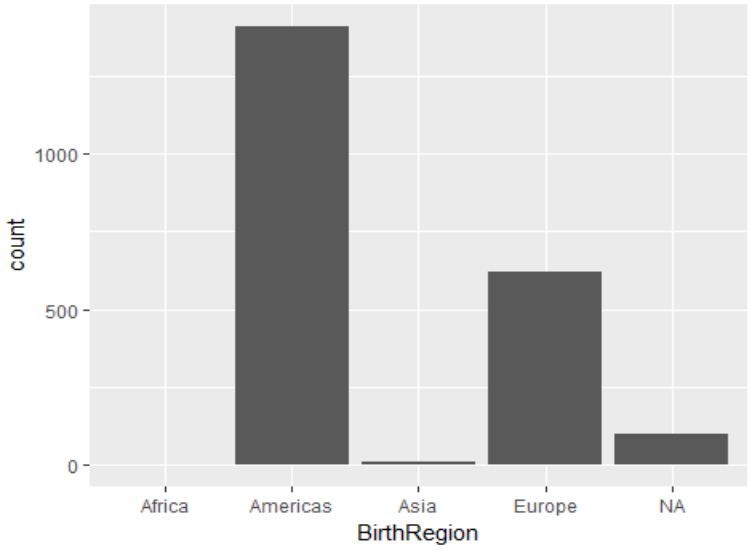

Figure 2: Player Region of Origin

The Americas and Europe generate the vast majority of players drafted in the NHL.

Draft statistics were initially evaluated by viewing a distribution of the draft rounds of the players in this dataset.

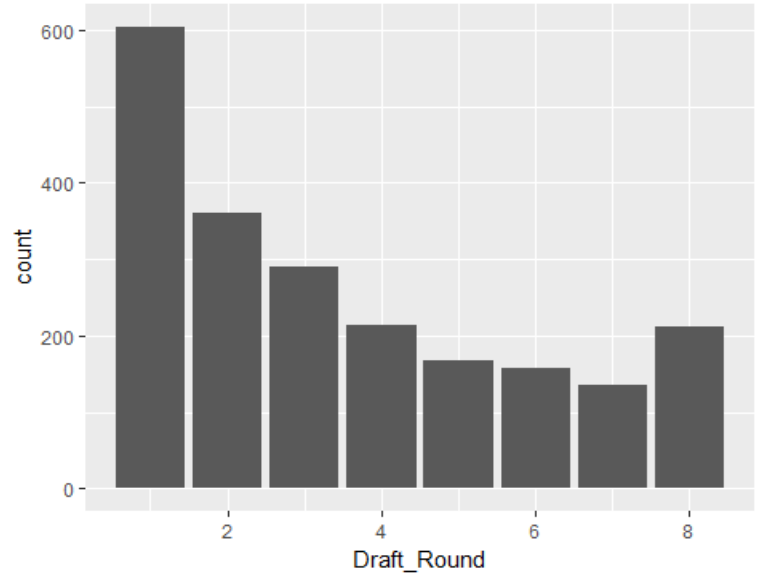

Figure 3: Distribution of Draft Round

The draft round in the dataset appears to be skewed towards earlier rounds. A comparison between draft round and active years (up to 2012) in the NHL supports the idea that

players from earlier draft rounds may have somewhat longer NHL careers.

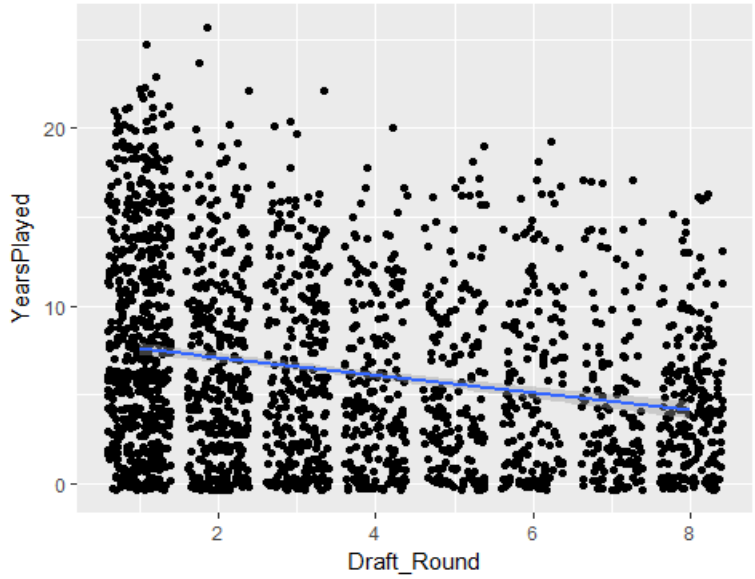

Figure 4: Draft Round and Years of Experience

Another way to evaluate relationship between player draft position and team playoff success might be to see if higher-drafted players score more goals per game. It does look like there is a slight trend for more goals per game from higher-drafted players but it's not clear if this impacts making the playoffs.

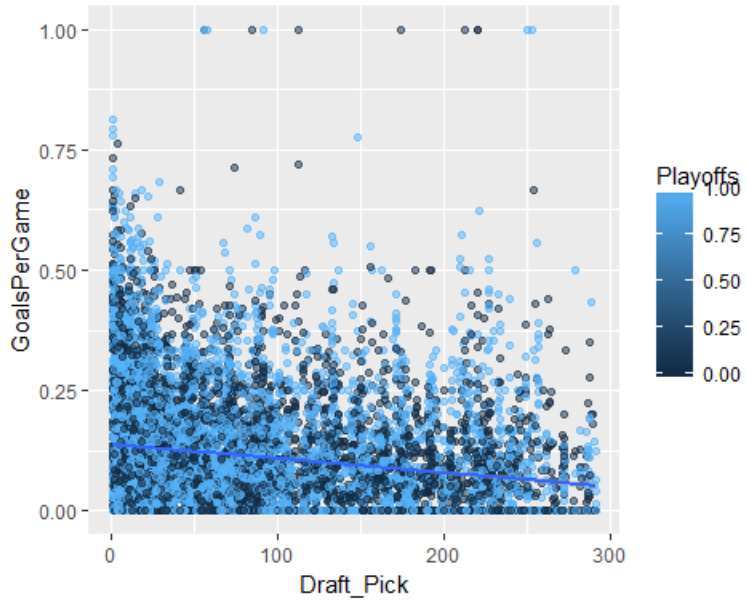

Figure 5: Goals per Game by Draft Position

It should be expected that teams who score more goals in a year and allows less goals in a year are more likely to go to the playoffs.

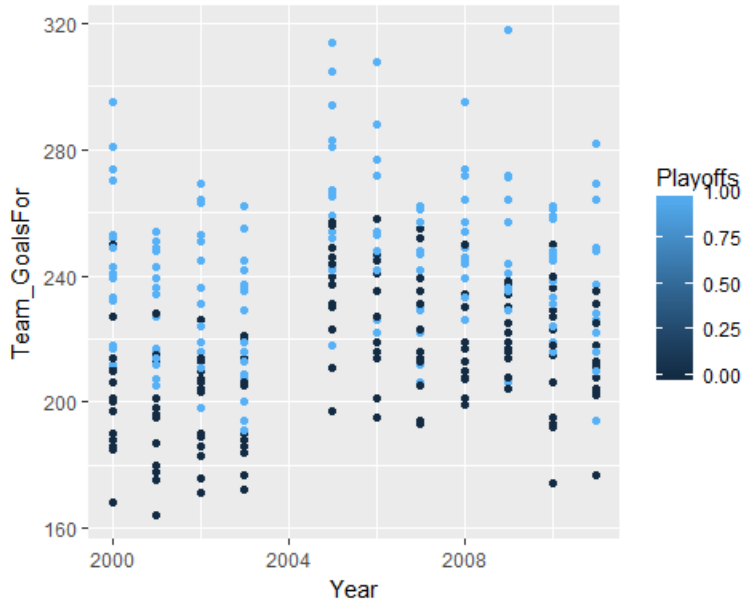

Figure 6: Team Goal Totals by Year

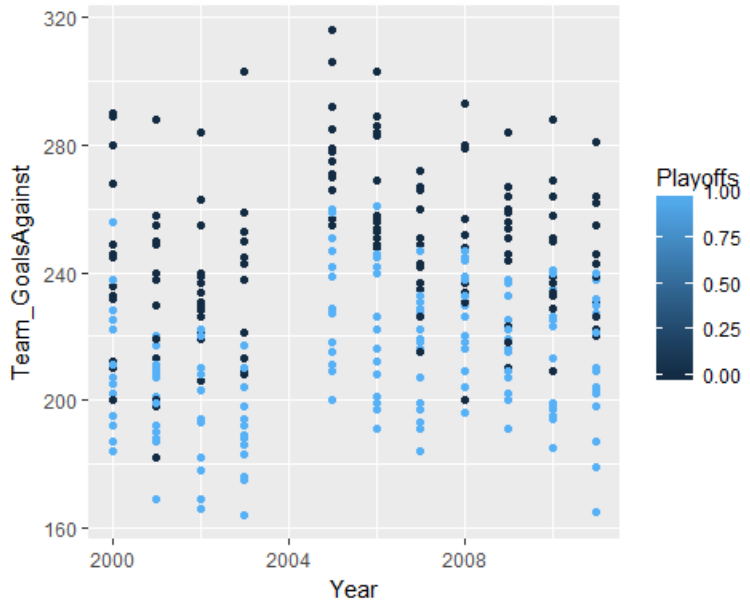

Figure 7: Team Goals Allowed by Year

This hypothesis appears to be true for the highest goal-scoring teams and lowest goal-allowing teams year-to-year.

Another interesting area to evaluate is defensive statistics related to goaltenders. A goaltender's save percentage (the percent of shots they save from becoming goals) and

shutouts (games where they don't allow the other team to score at all) should influence their team's win rate and perhaps making the playoffs.

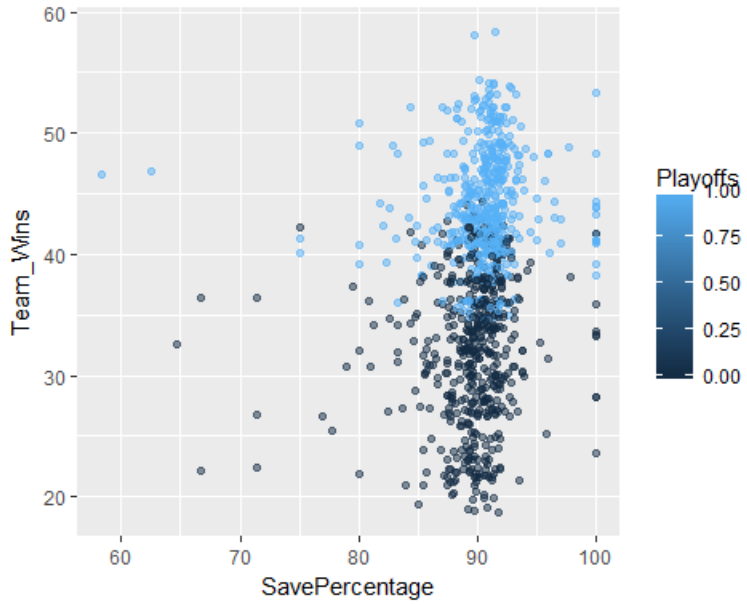

Figure 8: Team Wins by Goalie Save Percentage

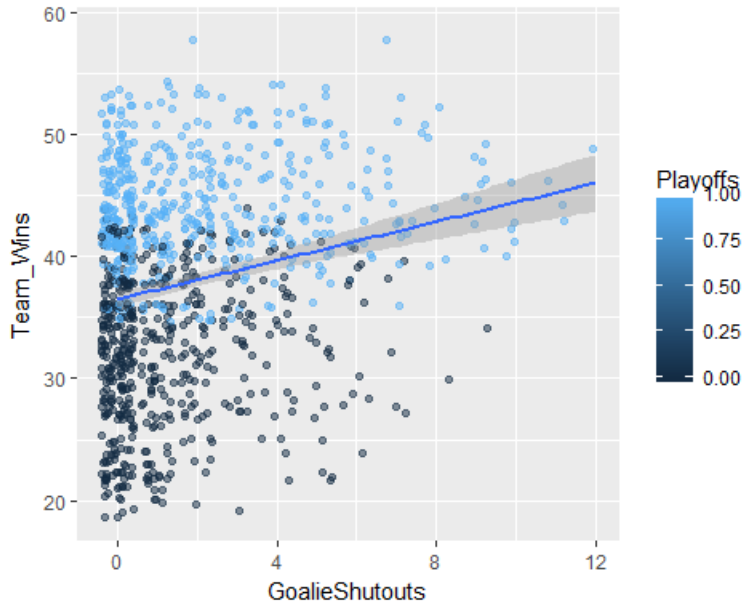

Figure 9: Team Wins by Goalie Shutouts

The relationship is not entirely clear from these graphs, though more shutouts seem to indicate more team wins, but the influence of this feature may be more apparent when the data is modeled.

Modeling Approach

In examining the data in more detail over the course of the wrangling and exploring exercises, several simple truths are clear: teams that win more games, score more goals and allow less goals by their opposition go to the playoffs. However, as teams are built and restructured over years, the question still exists of what player characteristics contribute most strongly to success. The main goal of the project remains the primary focus from the project proposal - to determine what factors contribute to regular season success that is realized in a playoff berth.

- The project approach at this point is to:
  Build the best predictive model of team success (defined as making the playoffs) from the player and draft characteristics.
- Define underrated and overrated players, determined by draft position and scoring history, and evaluate if player performance is predictable at the time of the drafy by amateur team, amateur league, position played or other factors.

Machine Learning Approach to Capstone Project

The main question for this project was what player demographic/draft/play characteristics are most predictive of team success (making the playoffs). The secondary question was whether player characteristics at the time they are drafted into the league, using a subset of underperforming and overperforming players, can predict their future performance. As this is labeled data, it was approached as a supervised learning problem.

The player features (independent variables) used to predict team success were Shooting Hand, Years of Experience, Position Played, Games Played, Goals, Assists, Shots and Points, Penalty Minutes, Plus-Minus Rating, Goals and Shots Per Game, Percent of Team Goals Scored, Percent of Team Games Played, Draft Pick, Draft Round, and Draft Age.

The player features (independent variables) that may be used to predict whether a player will over or under deliver in terms of performance post-draft were Height and Weight, Position Played, Draft Year, Draft Team, Draft Age, and Amateur League.

The primary question was evaluated using several different types of algorithms to build a model that allows for less than linear relationships between the variables; model performance was assessed to determine what type of algorithm produces the best model. The secondary question was evaluated using logistic regression to determine how the variables may impact the final outcome of player performance.

Repeated k-fold cross-validation was used with the caret package to estimate model accuracy for the various models predicting team performance. The logistic regression model for player performance was evaluated with a chi-squared test using ANOVA.

Data Analysis

Modeling Predictors of Team Success

A number of R packages were utilized for the modeling and analysis.

```
library(mlbench)
library(caret)
library(readr)
library(rpart)
library(dplyr)
library(C50)
library(plyr)
library(ipred)
library(e1071)
library(ROCR)
library(randomForest)
```

The trainControl function was set for all models to be run with three separate 10-fold cross-validations as the resampling scheme.

The formula to be used in the modeling was defined.

```
formula <- Playoffs ~ Shooting_Hand + YearsExperience + Goals + Assists + Pen
alty_Minutes + Shots + GoalsPerGame + ShotsPerGame + PointsPerGame + PercentG
oals + PercentGames + Draft_Pick + Draft_Round + Draft_Age + SavePercentage +
GoalieShutouts + GoalieMinutes
```

The caret package in R was used to quickly run a number of different types of models on the same data. The data was evaluated using logistic regression, CART, C5.0, bagged CART, random forest and stochastic gradient boosting models.

The resamples function was used to aggregate the results from the models and method accuracy was compared across models.

```
## Models: logistic, cart, c50, bagging, rf, gbm
## Number of resamples: 30
##
## Accuracy
##               Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## logistic 0.5194805 0.5648295 0.5898166 0.6008131 0.6306604 0.7473684    0
## cart     0.5536398 0.5697538 0.5747126 0.5751777 0.5778894 0.5967433    0
## c50      0.5354406 0.6027299 0.6340996 0.6259022 0.6594828 0.6791188    0
## bagging  0.5900383 0.6290709 0.6397129 0.6384139 0.6532661 0.6676245    0
## rf       0.6206897 0.6429598 0.6553379 0.6540559 0.6637931 0.6829502    0
## gbm      0.6159004 0.6388889 0.6452845 0.6443529 0.6526220 0.6695402    0
```

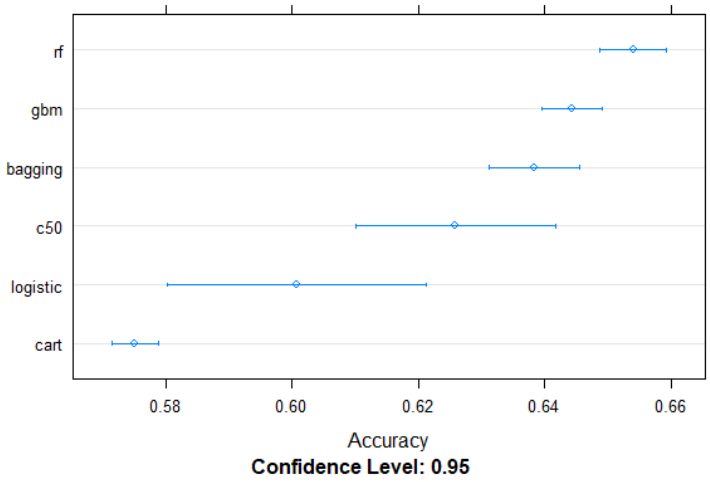

Figure 10: Accuracy by Model

The random forest model showed the highest accuracy amongst the models. The various features that contribute to the model were evaluated to see their influence on the modeling.

```
## rf variable importance
##
##                 Overall
## PercentGoals    100.000
##Draft_Pick       78.445
##Penalty_Minutes  54.639
##ShotsPerGame     51.273
##PercentGames     48.771
##PointsPerGame    47.765
##YearsExperience  47.019
##GoalsPerGame     41.597
##Shots            39.379
##Goals            28.424
##Assists          24.523
##Draft_Age        14.312
##SavePercentage    9.744
##Draft_Round       7.891
##Shooting_HandR    6.225
##GoalieMinutes     6.099
##GoalieShutouts    0.000
```

A new model was generated using the variables of greatest importance from the first random forest model.

```
formula2 <- Playoffs ~ PercentGoals + Draft_Pick + Penalty_Minutes + ShotsPer
Game + PercentGames + PointsPerGame + YearsExperience + GoalsPerGame + Shots
+ Goals + Assists
```

```
fit.rf2 <- randomForest(formula2, data=dataset, trControl="oob", na.action =
na.roughfix, preProcess=c("center", "scale"))
```

The mean decrease gini (a measure of how each variable contributes to the model) was plotted to determine the most important features, the AUC was calculated, and a ROC curve generated to assess the results of the model.

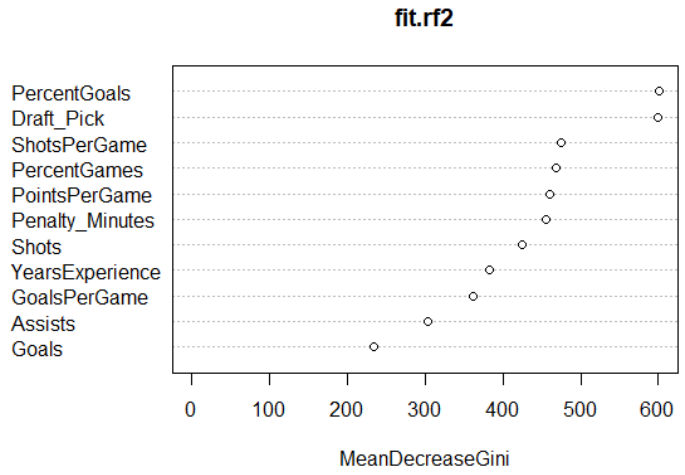

Figure 11: Mean Decrease Gini for Selected Features

```
## auc
## [1] 0.6924906
```

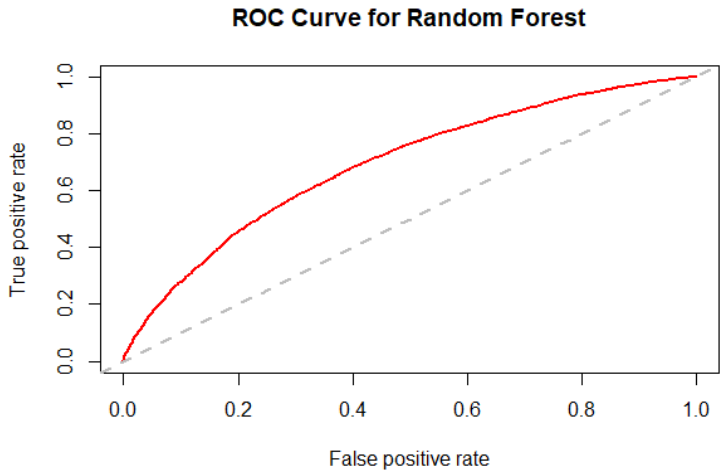

Figure 12: ROC Curve for Final Random Forest Model

Modeling Predictors of Future Player Performance

A secondary analysis aim was to define underrated and overrated players, determined by draft position and scoring statistics, and evaluate if player performance at the time of the draft is predictable by amateur team, amateur league, nationality, position played or other factors. A logistic regression model with three 10-fold cross validations was used to assess the features which may differentiate future player performance.

```
formula2 <- OverUnder ~ Height + Weight + Position_Played + Draft_Team + Birt
hRegion + Draft_Age + AmateurLeague
```

The results of the model show an average accuracy rate across the cross-validated models of 93% and the 20 most important variables (of 71 options).

```
Generalized Linear Model
2325 samples
   7 predictor
   2 classes: '0', '1'
Pre-processing: centered (85), scaled (85)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2718, 2718, 2718, 2717, 2719, 2718, ...
Resampling results:
  Accuracy   Kappa
  0.9333757  0.4589219

glm variable importance
only 20 most important variables shown (out of 71)

                               Overall
Draft_Age                       100.00
Position_PlayedD                 60.70
Height                           59.51
AmateurLeagueRussia              45.04
Position_PlayedL                 39.60
`Draft_TeamHartford Whalers`     33.90
`Draft_TeamPhiladelphia Flyers`  31.64
`Draft_TeamBoston Bruins`        29.53
`Draft_TeamNew York Rangers`     28.81
`Draft_TeamDallas Stars`         27.63
`Draft_TeamWashington Capitals`  27.23
`Draft_TeamCalgary Flames`       26.84
`Draft_TeamNew York Islanders`   25.01
`Draft_TeamSan Jose Sharks`      23.91
`Draft_TeamVancouver Canucks`    22.91
AmateurLeagueUSHL                21.60
`AmateurLeagueH-East`            21.27
`Draft_TeamNew Jersey Devils`    19.66
`Draft_TeamEdmonton Oilers`      19.44
`AmateurLeagueHigh-MA`           19.14
```

Conclusions

Using a random forest algorithm, it was possible to build a model that predicted team regular season success (making the playoffs) with near 70% accuracy. The features which supported the prediction were:

- Scoring Statistics (Goals, Assists and Shots)
- Draft Pick Position
- Penalty Minutes
- Goals, Shots and Points Per Game (rate of scoring)
- Years of Experience
- % of Team's Goals Scored by a Player and % of Team's Games Played by a Player

The secondary project question was how well one could predict a high draft pick underperforming or a low draft pick overperforming in their future career. The strongest predictor variables in a logistic regression model were Draft Age, Height, Position Played, certain Amateur Leagues, and some of the drafting teams. This model predicted the outcome with 93% accuracy.

Both analyses would likely benefit from more data, especially pre-draft, that is, the player statistics from their amateur league years of play. Ideally, this analysis could be enhanced if amateur data could be compiled and included and the entire dataset expanded to more years of play.

Client Recommendations

Based on the results of this analysis, player selection advice to NHL GMs would be as follows:

- When looking to increase a team's chances of making the playoffs, where possible look for an experienced player selected in a higher draft position who is able to play a high percentage of the team's games (perhaps indicative of a lower injury rate), and who has a strong scoring history and/or high shot rate
- Consider players who take penalty minutes (this may be an indicator of willingness to take risks to make plays).
- Player position, draft age, height, and coming from certain amateur leagues may give an indication as to how successful a lower or higher drafted player will perform longer-term. If the chance arises to take a player that meets the criteria (for example a tall Russian League forward) in a lower draft round, that saves the team money and earlier-round draft picks. A GM who can score a more successful player in a later draft round does his team a major service!

References

Player and team statistics were obtained from: The Hockey Database(link) by Doug Reynolds, via Open Source Sports.
Draft data was obtained from: Sports Reference LLC. "NHL Entry And Amateur Draft History." Hockey-Reference.com - Hockey Statistics and History(link).

All prior project reports, original and final data files, and R code can be found in the project Github repository(link).