Who Will Make the Playoffs?: Predicting the 2024 NHL Standings

Peter D. DePaul III

03-15-2024

Abstract

This reports presents a comprehensive analysis of hockey statistics with the goal of predicting the final National Hockey League (NHL) standings utilizing statistics from the latter part of the regular season (10-15 games left). This project meticulously examines the relationship between various team performance indicators and their standings at the end of the season. The development of the predictive model is a robust process with 6 models tested — including multivariate linear regression, support vector machines, and random forests — with the results revealing insightful variables for prediction. The findings of this model has the ability to help team management understand their season-end goal, whether it's making a playoff push or giving up and planning for next year. This offers a tool for any analyst or fan who are eagerly anticipating the playoffs.

Contents

1 Introduction

When it comes to sports it's hard to argue that there is anything more important than the playoffs. Perhaps the only thing more important is being one of the teams that qualifies for the playoffs tournament. In the National Hockey League (NHL), the playoffs have an interesting format compared to most leagues. The final NHL playoffs spots are usually separated by 2-3 standings points, and the last 10-15 games of the season usually significantly impact the final standings. I believe there's a way to predict the final points totals for each of the 32 NHL teams based upon their team statistics and standings with about 10-15 games left per season.

1.1 Background

The NHL has a long history with the earliest version of the Stanley Cup Playoff taking place in 1917. The early NHL was tumultuous and often teams wouldn't survive a few years. The modern league is generally considered as beginning in 1942-1943 with the "Original Six" teams. Today the NHL has expanded to over 32 teams, and follows a 16 team playoff format. The 16 team playoff format has existed since the 1979-1980 season when there were 21 teams in the NHL. Today the playoffs follow a format where there are 8 teams each from the Eastern Conference and Western Conference. Within these conferences there are 2 divisions so the 3 best teams from each of these divisions qualify for the playoffs. Additionally, there are 2 "Wild Card" teams for each conference that make up the final playoff spots (Records 2024).

1.2 Variable Overview

Below is the subset of variables I utilized from the larger data sets to build the model for my final points predictions. There are 14 total variables in the data set including the predicted variable `endSeasonPoints`. The 13 predictor variables were all used in the modeling process and the breakdown is 11 numeric, and 2 categorical.

Table 1: Table of Variables

Variables	Description	Type
clinchIndicator	Indicating if the team clinched a playoff spot	Categorical
gamesPlayed	Total Number of Games Played	Numeric
goalDifferential	Goal Differential of Team	Numeric
l10GoalDifferential	Goal Differential over Last ten Games	Numeric
l10Losses	Losses over Last ten Games	Numeric
l10OtLosses	Overtime Losses over Last ten Games	Numeric
l10Points	Standings Points over Last ten Games	Numeric
losses	Losses	Numeric
otLosses	Losses in Overtime	Numeric
points	Standings Points	Numeric
shootoutLosses	Shootout Losses	Numeric
wildcardSequence	Wildcard Ranking	Numeric
endSeasonPoints	Predicted Points	Numeric
decade	Decade of Play	Categorical

2 Exploratory Data Analysis

2.1 Does Goal Differential Matter for Predicting Points?

For understanding a team's points at the end of a season, it's important to understand how well that team did from a scoring perspective. Luckily in the NHL there's an all encompassing scoring statistic, "Goal Differential". Goal differential is defined as Goal Differential = Goals For − Goals Against. It's a simple metric to understand, negative usually indicates poor play, 0 indicates a team is playing about even on both sides of the puck, and a positive one indicates great play from a team (Bourne 2022).

Table 2: Regression Model Coefficients

	Estimate	Std. Error	p-value
(Intercept)	88.593	0.213	0
goalDifferential	0.338	0.005	0

From the Table 2 above we can see that both the intercept (`end_season_points`) and the predictor (`goalDifferential`) have p-values that are 0, which indicates both are significant. This implies that when `goalDifferential` is equal to 0 that the `end_season_points` are significantly different from 0. This also implies that `goalDifferential` has a significant impact on the prediction of `end_season_points`.

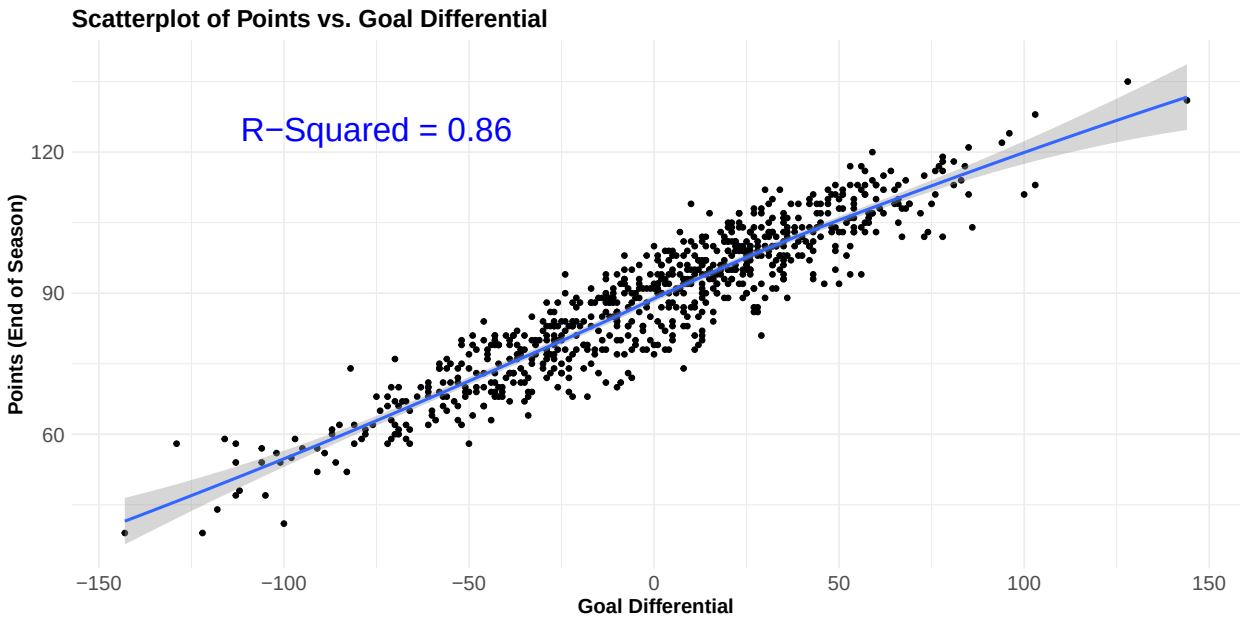

Figure 1: Scatterplot of Points vs Goal Differential

From the fitted linear regression line we're able to see that the $R^2 = 0.86$ indicates there is a strong positive linear relationship between Goal Differential and end of season Points [See Figure 1]. Teams with higher (more positive) goal differentials often finish higher in the league standings. With lower (more negative) team goal differentials teams more often than not historically have finished lower in the league standings (Bourne 2022).

2.2 How does Goal Differential impact the Playoffs?

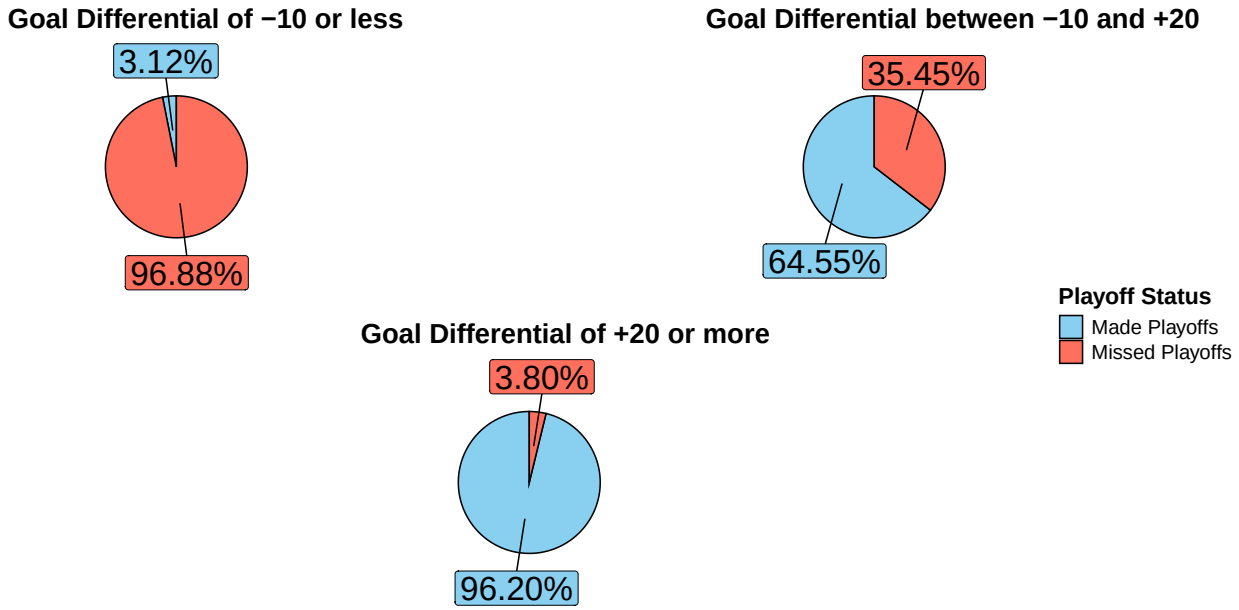

Figure 2: Playoff Status by Goal Differential (GD)

There is a stark importance in the impact of a team's goal differential on their playoff status [see Figure 2]. Teams with a goal differential of less than −10 are essentially guaranteed to miss the playoffs. Teams with a goal differential between -10 and +20 make the playoffs about 2/3 of the time. Finally we are able to see that teams with a goal differential greater than +20 are essentially guaranteed to clinch a playoff spot.

2.3 How much better are Playoff Teams?

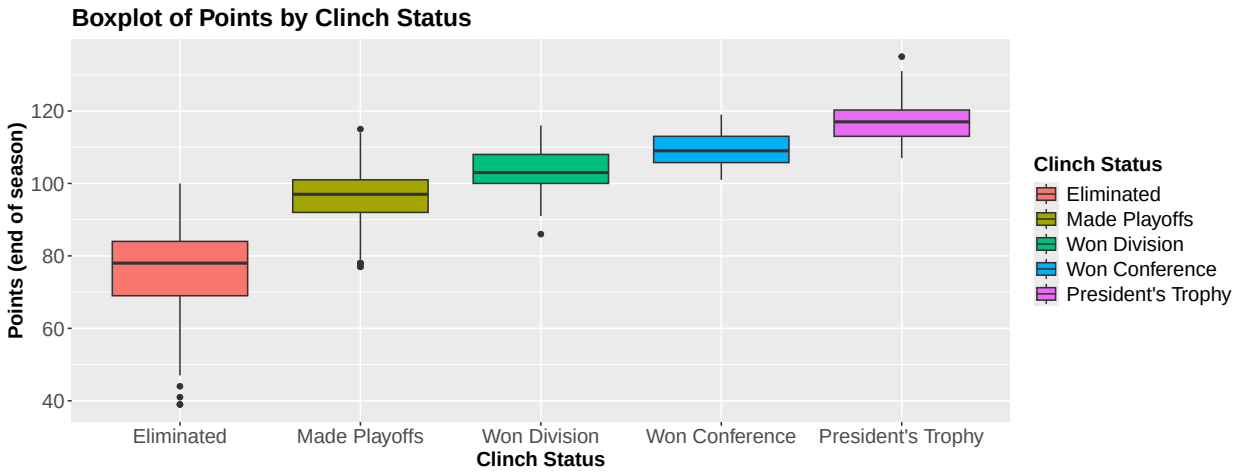

Figure 3: Boxplot of Team Points by Clinch Status

Team points will increase as one looks left to right as the President's Trophy is the highest honor, since it means that team led the entire NHL in points at the end of the season (NHL.com 2023). Most interesting is observing the outliers who made the playoffs, and the range of those who were eliminated from the playoffs. Teams who make the playoffs are usually in at least the 100 points area. However it is observed that some teams score upwards of 100 points and still get eliminated from the playoffs [See Figure 3].

Table 3: Table of Playoff Outliers

Clinch Status	Team	Season	Points
Made Playoffs	Toronto Maple Leafs	20212022	115
Won Division	Carolina Hurricanes	19981999	86
Made Playoffs	Ottawa Senators	19961997	77
Made Playoffs	Montreal Canadiens	19961997	77

The visible outliers who made the playoffs are interesting cases [See Table 3]. The 2021-2022 Toronto Maple Leafs only Made the Playoffs because they were in the same division as the Florida Panthers who won the President's Trophy. Meanwhile the 1996-1997 Ottawa Senators and Montreal Canadiens benefited from playing in an exceptionally weak conference, which resulted in them making the playoffs. Finally, the 1998-1999 Carolina Hurricanes were the best team in the worst division in hockey and were able to come away as champions of the Southeast Division.

2.4 How has the NHL Dynamic Changed?

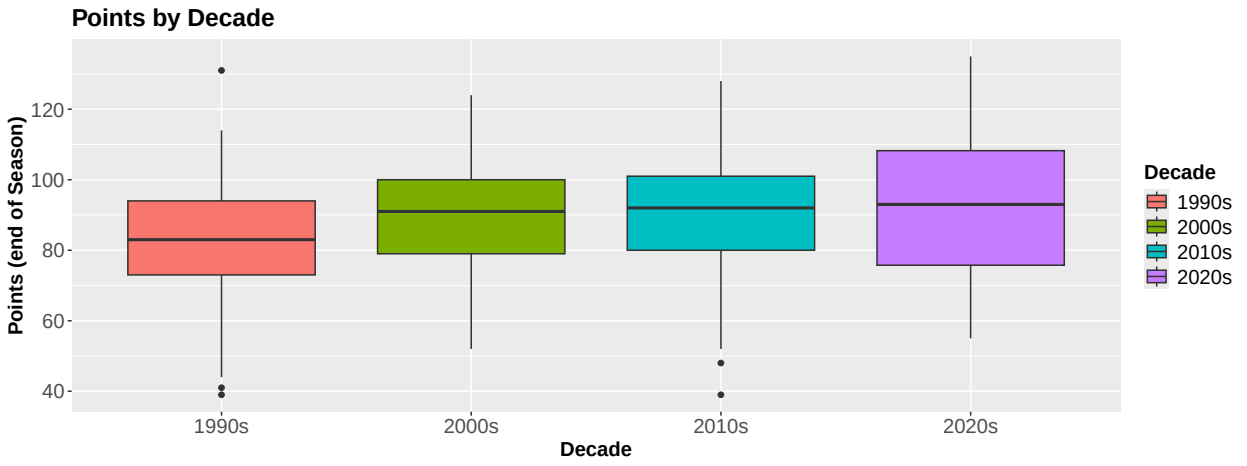

Figure 4: Boxplot of Points by Decade

The reason the difference in decades interests me is because as the NHL has grown over the years there has become a movement towards higher overall point totals in the league. The boxplot confirms this idea as we can see since the 1990s there has been an overall increase in team's points at the end of the season [See Figure 4]. This likely has to do with the more even level of competition in the NHL. In modern standards players condition year-round, and higher levels of play are consistently expected compared to older eras (Bouthillier 2008).

3 The Model

By defining a team's success by their points at the end of the season, I aim to create a predictive regression model which predicts team points at the end of season. It will utilize specific near-end season data variables to make the prediction of the end of season points.

3.1 Model Data

Data was collected using the NHL API through Python (Zmalski 2024). I scraped team statistics and standings for regular season games from 1979-1980 through 2023-2024. There are two distinct sets of data, the data from NHL seasons through 65-70 games in the season (near-end season data). We then had complementary end season statistics for each season with the exception of 2022-2023.

3.2 Feature Engineering

The near-end season data was the data used to train our predictive model. I made the decision to add two features to the near-end season data.

The following variables were added:

- `end_season_points`: The team's points at the end of the season, from the end of season data.
- `decade`: The decade in which the team's season took place.

3.3 Feature Selection

Following the feature engineering process, we employ the Boruta feature selection algorithm to verify the importance of our predictive features. The Boruta algorithm utilizes a Random Forest method for confirming the importance of variables. The algorithm confirmed that all the variables were significant predictors of a team's points at the end of the season [See Table 1].

3.4 Model Creation

I constructed the models using `tidymodels` workflows (Kuhn and Wickham 2020). The step-by-step model creation process is as follows:

1. I utilized the near-end season data for seasons from 1995-1996 to 2022-2023 as the training data for the model.
2. To further ensure the best model possible, hyper parameter tuning was utilized for all models except the linear regression model. Tuning finds the ideal hyper parameters for the models based on the training data.
3. In the pre-processing phase of the model recipe, the workflow checks to ensure there are no variables with zero variance (all the same values), add dummy variables for categorical variables which not already encoded, and finally scale and center all numeric variables. These choices minimize the variance caused by outliers in the predictor variables.
4. I perform the hyper parameter tuning process by constructing a grid of 1500 unique values (depending upon each model's specific parameters) then re-fit the model for each set of parameters. Each fitted model is cross-validated for performance, and at the end the best model is chosen by the cross-validation model with the lowest root mean square error (rmse).

For the model selection process I tested out 6 different models. These models included:

- Linear Regression
- K-Nearest Neighbors (KNN)
- Polynomial Support Vector Machine (SVM, `kernlab` (Karatzoglou, Smola, and Hornik 2023))
- Random Forest (`ranger` (Wright and Ziegler 2017))
- Gradient Boosted Tree (`xboost` (Chen et al. 2024))
- Multi-layer Perceptron (`nnet`(Venables and Ripley 2002))

One final note is that once the models were tuned I was interested in observing if the outliers within the model were affecting the predictive performance. This is plausible since the training data set only consisted of 740 observations. I decided to examine the models with the same hyper parameters, but compare their performance with and without outlier observations in the training data.

4 Results

4.1 Cross-Validation Metrics

For choosing the best model I decided to focus on two specific regression variable metrics, the Root Mean Square Error (RMSE) and the R-squared Coefficient of Determination. For RMSE a lower value indicates a better performing model. A higher R-squared value indicates a better model fit when comparing models with the same variables. While reading the following table it is important to keep in mind that the SVM model was unable to make a prediction for the 2023-2024 San Jose Sharks. I'm not quite sure why this happened as all other models did not encounter a problem.

4.1.1 Models by RMSE

Table 4: Table of Models by RMSE

with Outliers				no Outliers			
Model	Metric	Mean	Std. Error	Model	Metric	Mean	Std. Error
SVM	RMSE	3.832	0.092	SVM	RMSE	3.758	0.067
Boosted Tree	RMSE	4.029	0.106	Boosted Tree	RMSE	3.908	0.073
MLP	RMSE	4.073	0.086	Rand. Forest	RMSE	3.969	0.095
Rand. Forest	RMSE	4.180	0.137	MLP	RMSE	4.047	0.076
KNN	RMSE	5.290	0.096	KNN	RMSE	5.235	0.141

We are able to see that for both the model with outliers and the model without outliers the SVM model and Boosted Tree model performed the best [See Table 4]. It's also clear to see that the worst performing model in both cases was the K-Nearest Neighbors model, so that is excluded from selection. I further excluded the SVM model from my selection due to the `NaN` model generated for the San Jose Sharks prediction. The MLP and Random Forest models performed well but the MLP performed marginally better with it's rather low standard error. In terms of the RMSE metric, the Boosted Tree model was the best performing model.

4.1.2 Models by R-Squared

Table 5: Table of Models by R-Squared

with Outliers				no Outliers			
Model	Metric	Mean	Std. Error	Model	Metric	Mean	Std. Error
SVM	R-squared	0.938	0.003	SVM	R-squared	0.939	0.003
Boosted Tree	R-squared	0.933	0.004	Boosted Tree	R-squared	0.937	0.003
MLP	R-squared	0.932	0.003	Rand. Forest	R-squared	0.935	0.003
Rand. Forest	R-squared	0.928	0.006	MLP	R-squared	0.932	0.004
KNN	R-squared	0.892	0.006	KNN	R-squared	0.894	0.010

When evaluating the R-squared metrics of the models with outliers and no outliers, the results are exactly the same as with the RMSE. SVM and Boosted Tree were the two best models followed by the MLP then the Random Forest model [See Table 5]. The K-Nearest Neighbors model was again the worst performing model. The SVM again was excluded due to the `NaN` prediction. From R-squared, again the best performing model was the Boosted Tree model. The Boosted Tree model created using the `xgboost` package was the best performing model overall. This was the model whose predictions were utilized for the final results. Furthermore I specifically used the model trained on the data without outliers. The models without outliers had a noticeable performance improvement.

4.2 Model Parameters

Parameter Description	R Documentation	Value
Number of Randomly Selected Predictors	mtry	11
Number of trees	trees	127
Minimal Node Size	min_n	6
Tree Depth	tree_depth	6
Learning Rate	learn_rate	≈0.0434
Minimum Loss Reduction	loss_reduction	≈0.1236
Sample Size	sample_size	1
Number of Iterations Before Stopping	stop_iter	5

Table 6: Boosted Tree Tuning Parameters

The model parameters were chosen using tidymodels tuning to hyper-parameter tune for all the parameters of the model with the exceptions of `sample_size` and `stop_iter` which I arbitrarily chose for the model [See Table 6]. I utilized a 1500 unique row random tuning grid to train 1500 boosted tree models. Additionally, cross-validation was performed on each model after it was fit using its random hyper-parameters. At the end of the process, I collected metrics and decided the best fit based upon the model with the lowest RMSE. That best fit model is represented by the hyper-parameters above

4.3 Model Predictions

Table 7: Table of Final Standings Predictions

Playoff Teams			Non-Playoff Teams		
Clinch Status	Team	Points	Clinch Status	Team	Points
President's Trophy	Florida Panthers	112	Eliminated	NY Islanders	91
Won Conference	Vancouver Canucks	112	Eliminated	Minnesota Wild	90
Made Playoffs	Boston Bruins	110	Eliminated	Washington Capitals	89
Won Division	NY Rangers	110	Eliminated	St. Louis Blues	88
Won Division	Winnipeg Jets	110	Eliminated	Seattle Kraken	88
Made Playoffs	Colorado Avalanche	108	Eliminated	Pittsburgh Penguins	88
Made Playoffs	Dallas Stars	108	Eliminated	Calgary Flames	87
Made Playoffs	Carolina Hurricanes	108	Eliminated	Buffalo Sabres	86
Made Playoffs	Toronto Maple Leafs	106	Eliminated	New Jersey Devils	85
Made Playoffs	Edmonton Oilers	106	Eliminated	Ottawa Senators	75
Made Playoffs	Los Angeles Kings	101	Eliminated	Arizona Coyotes	74
Made Playoffs	Vegas Golden Knights	98	Eliminated	Montreal Canadiens	71
Made Playoffs	Tampa Bay Lightning	97	Eliminated	Columbus Blue Jackets	66
Made Playoffs	Nashville Predators	96	Eliminated	Anaheim Ducks	59
Made Playoffs	Philadelphia Flyers	95	Eliminated	San Jose Sharks	52
Made Playoffs	Detroit Red Wings	93	Eliminated	Chicago Blackhawks	51

My predictions put 93 points as the cutoff for playoff contention this year [See Table 7]. As we can see there are several teams on the verge of this mark. The separating factors between the teams fighting for playoff spots is only 2-4 points, which is 1-2 wins.

4.4 Variable Importance

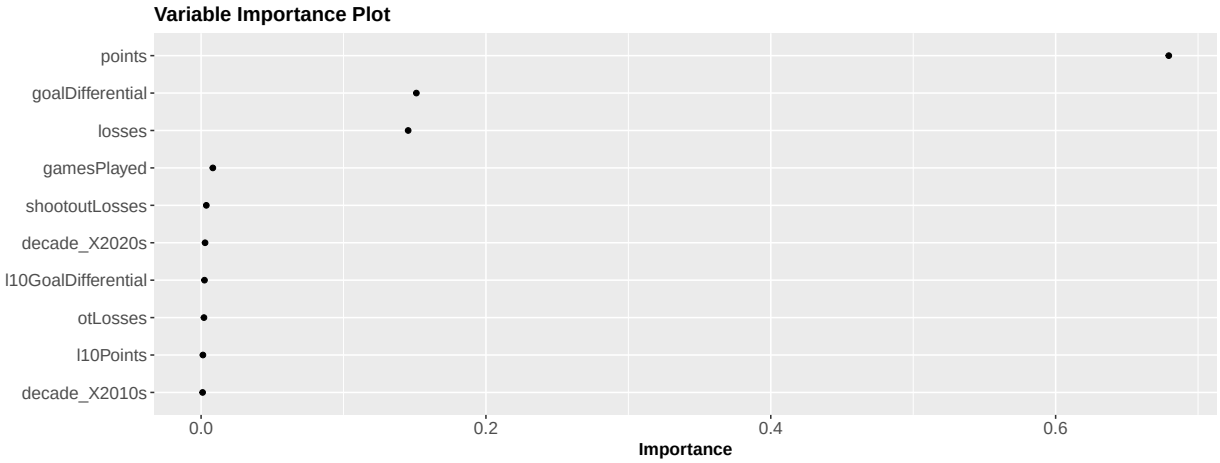

Figure 5: Variable Importance Plot

For our predictive model the most important variables were points at the mid-point `points`, `goalDifferential`, and a team's total losses `losses` [See Figure 5]. The goal differential is the best descriptive variable for evaluating a team's overall success, and predicting a team's success in the future.

5 Conclusion

5.1 Goal Differential's Impact

There is abundant information provided above to suggest that goal differential is the best predictive statistic we can use to evaluate a team's success. To put it simply teams with highly positive goal differentials win a lot and lose a little, the opposite holds true as well. The important part of goal differential is understanding it scales the game down a level, and does not let a team hide its flaws.

5.2 Playoff Team's Performance

During the regular season team's which miss the playoffs tend to have poor goal differentials, and struggle to produce at a high level. The data analysis has supported that the team's who barely miss the playoffs are within 2-4 points of a playoff spot. This is the difference of 2 wins in a season, and it helps with understanding how important every game is throughout the NHL season.

5.3 Model Results

Our model predicted `end_season_points` with an RMSE = 3.91 which I consider very high performance based upon the available data. There was less than 1000 observations of training data, and yet it's still able to make a rather effective model. There is room for improvement with the model, but I think the model is at a great point. It keeps the model simple, and looks to simple team counting stats. The advanced stats are not always the answer, and hockey is still not the most data rich sport.

The meaningful outcome of this model is there's more to determine a team's performance than `points` and `losses`. I was able to find that there are a few important variables for predicting a team's success including `goalDifferential`, `shootoutLosses`, and `decade`.

5.4 Possible Improvements

After coming up with the idea for this project I decided I wanted to keep the variables as simple as possible, and restricted to base statistics. It would likely be beneficial to have additional information such as line strength, and performance when at an advantage or disadvantage. However I had faith this model would perform to a high level, and my expectations are exceeded at this point so I am satisfied with the results.

5.5 Looking Ahead

I hope to compare the results of my models to the true results at the end of the NHL season. Additionally, once the end of season comes I plan to develop a model which aims to predict the Stanley Cup champion based upon teams data at the end of the regular season. Overall the model will look to predict playoff performance among all NHL teams who make the playoffs. This will have to wait about a month as the NHL continue to play regular season games.

Bibliography

Bourne, Justin. 2022. "What Goal Differential Can Tell Us about NHL Standings and Playoff Hopes." https://www.sportsnet.ca/nhl/article/what-goal-differential-can-tell-us-about-nhl-standings-and-playoff-hopes/.

Bouthillier, Chris. 2008. "The Evolution of Hockey: How the NHL Has Grown over the Years." https://bleacherreport.com/articles/46151-the-evolution-of-hockey-how-the-nhl-has-grown-overthe-years.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2024. *Xgboost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=xgboost.

Karatzoglou, Alexandros, Alex Smola, and Kurt Hornik. 2023. *Kernlab: Kernel-Based Machine Learning Lab*. https://CRAN.R-project.org/package=kernlab.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. https://www.tidymodels.org.

NHL.com. 2023. "What Goal Differential Can Tell Us about NHL Standings and Playoff Hopes." https://www.sportsnet.ca/nhl/article/what-goal-differential-can-tell-us-about-nhl-standings-and-playoff-hopes/.

Records, NHL. 2024. "All-Time Playoff Formats." https://records.nhl.com/history/playoff-formats.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Zmalski. 2024. "NHL-API-Reference." https://github.com/Zmalski/NHL-API-Reference.