Data Mining and Machine Learning Final Project

Daniil Deych, Alex Mykietyn, Cameron Wheatley

2023-04-24

What team-relative regular season statistics can tell us about NHL playoff team performance?

I. Introduction and Background

As team and player statistics have garnered a more prominent role in sports, attempts at predicting performance has become a go to area for Las Vegas bookies and sports fans in general. Getting into any debate about how "good" your team will be will undoubtedly incur some version of statistical performance. Be it your gut or a fancy statistical model, those numbers fuel the desire to predict the future. We are no different.

With this project we are hoping to build a predictive model which NHL team will win a given 7 game playoff series using the teams' regular season statistics. Using the difference between teams' statistics in we calculate the home team's advantage (or disadvantage) in each category.

Historically, the primary Machine Learning tool used in predicting sports performance has been neural networks (Weissbock el al., 2013), here we are looking to test a variety of other Machine Learning models.The end goal of all these methods being to find which model will provide the highest percentage of correct winners on the testing set.

A typical regular season hockey game consists of three 20-min periods. During the period, each team puts out five players and a goalie onto ice, and they attempt to win the game by putting a puck into the net of the other team. If by the end of the three periods the game is tied, the game goes into a 5 minute "sudden death" over time, after which, if still tied, the game goes into a shootout. In the shootout, each team takes turns by sending out one player to score a penalty shot on the opposing team's goalie. After 3 attempts, the team with most penalty shots scored wins the match and the losing team gets attributed an Overtime Loss (OTL). These overtime rules were instituted part way through our data set in 2006, previously a tie would be assigned if no one scored in the overtime period. If during the game, either team commits a foul, the other team receives a Power Play for a pre-determined amount of time (typically 2 mins), during which the fouling player gets sent off for that amount of time and the other team plays with an extra player on the ice, this situation often results in a goal.

II. Data

NHL's website has a copious amounts of data going as far back as 1917, but the game has changed dramatically since then, so we had to be discerning about which years to look at for our data. We decided to choose the most recent 20 seasons that played all the 82 mandated games, which makes the earliest season we accounted for to be 1998.

As a side note, we excluded several shortened seasons that took place between 1998 and now. COVID pandemic shortened the 2020-2021 and 2019-2020 season, while player strikes lead to lockouts for the 2012-13 and 2004-05 season.

To create our data set, we used the raw data from NHL's website and created our unique data set. For each year, we isolated all the playoff match ups that year, and created a separate row for each match up, designated by Home/Away team, the rest of the row lists the regular season difference-statistics between Home/Away teams of the given matchup. (See Appendix below for more details). It is this difference in regular season stats that will fuel our predictive models.

The variables of interest attempt to describe team performance in various situations. Many variables track break down the wins, losses and win percentage of teams based on the margin of victory in regular season games. We also have statistics relating to performance by period and the propensity of a team to come from behind or blow leads. Considering the number of goals scored on Power Plays, we have statistics measuring team performance in these situations as well as the frequency they take penalties and draw penalties from the other team.

III. Method

The main challenge with our approach is our relatively small data subset. With only 20 seasons and 300 playoff series against more than 60 variables, we do not have sufficient rows of data to work with the data set directly. To account for that we elected to use step wise selection and lasso approach to reduce the number of variables. In both cases we are concerned about the degrees of freedom in our model and so we limit variable selection to a maximum of 10 variables given that we will only be training on 200 observations.

We will select variables using a lasso regression and a stepwise selection process. We will then plug the selected variables into probit, ordered probit, and random forests models. Given that some of our statistics are highly correlated, we will also use Principle Component Analysis to summarize the data set and plug this into a probit model.Given the nature of our small data set we attempt to give ourselves less variance in our testing set we hold out (100 observations) from our training data. To attempt to avoid over fitting our models we test potential parameter values on resamples of the training set and select the model that performs best on these training set resamples . We then test the accuracy for each model by calculating the absolute improvement and lift against our null model that Home team (which is equivalent to being a higher seed) always wins.

Step-wise Selection

Step-wise selection variables are - face-off win percentage (FOW%), penalty kill percentage (Net.PK), percentage of games won byc 2 goals (win..2goal.game), number of penalties drawng against the other team (Pen.Drawn.60), percentage of games won by more than 3 goals (win..3.goal.game), goals against in second period (GA.in.p2), percentage of games won while leading in period 2 (win..lead.2p), percentage of games won while leading in period 1 (win..lead.1p)

Lasso selection

Lasso selected variables are total goal differential (total_goal_differential), goals against in period 2 (GA.in.P2), percentage of games after scoring first (W..SF), net power play kill percentage (Net.PK.), percentage of games won while leading in period 2 (win..lead.2p), percentage of games won by more than 3 goals (win..3.goal.game), number of penalties drawn against the other team (Pen.Drawn.60), shots per game differential (shots_differential), regulation wins (RW).

Using lasso selected variables

Probit Model

```
##    yhat
```

```
## y    0  1
##   0 11 27
##   1  9 53
```

```
## [1] "Probit Model Accuracy"
```

```
## [1] 0.64
```

```
## [1] "Null Model Accuracy"
```

```
## [1] 0.62
```

Ordered Probit Model

```
##     Ordered_Probit_Lasso_Pred
##      -4 -3 -2 -1  1  2  3  4
##   -4  0  0  0  2  0  3  0  0
##   -3  0  0  0  6  0  3  0  0
##   -2  0  0  0 10  0  7  0  0
##   -1  0  0  0  2  0  5  0  0
##   1   0  0  0  3  0  9  2  0
##   2   0  0  0  4  0 16  3  0
##   3   0  0  0  6  0  7  2  0
##   4   0  0  0  2  0  7  1  0
```

```
## [1] "Ordered Probit Model Accuracy"
```

```
## [1] 0.2
```

```
## [1] "Null Model Accuracy"
```

```
## [1] 0.23
```

Logit Model

```
##    yhat
## y    0  1
##   0 11 27
##   1  9 53
```

```
## [1] "LogitModel Accuracy"
```

```
## [1] 0.64
```

```
## [1] "Null Model Accuracy"
```

```
## [1] 0.62
```

Random Forest

```
##    yhat
## y    0  1
##   0 12 26
##   1  9 53
```

```
##    yhat
## y    0  1
##   0 12 26
##   1  9 53
```

```
## [1] "Forest Model Accuracy"
```

```
## [1] 0.65
```

```
## [1] "Null Model Accuracy"
```

```
## [1] 0.62
```

Step-wise Selection

Probit Model

```
##    yhat
## y    0  1
##   0 16 22
##   1 10 52
```

```
## [1] "Probit Model Accuracy"
```

```
## [1] 0.68
```

```
## [1] "Null Model Accuracy"
```

```
## [1] 0.62
```

Ordered Probit Model

```
##     Ordered_Probit_Step_Pred
##       -4 -3 -2 -1  1  2  3  4
##   -4  0  0  0  4  0  1  0  0
##   -3  0  0  0  4  0  5  0  0
##   -2  1  0  0  8  0  7  1  0
##   -1  0  0  0  5  0  2  0  0
##   1   0  0  0  5  0  7  2  0
##   2   0  0  0  4  0 11  8  0
##   3   0  0  0  5  0  6  4  0
##   4   0  0  0  4  0  5  1  0
```

```
## [1] "Ordered Probit Model Accuracy"

## [1] 0.2

## [1] "Null Model Accuracy"

## [1] 0.23
```

Logit Model

```
##    yhat
## y    0  1
##   0 16 22
##   1 10 52

## [1] "Logit Model Accuracy"

## [1] 0.68

## [1] "Null Model Accuracy"

## [1] 0.62
```

Random Forest

```
##    yhat
## y    0  1
##   0 13 25
##   1  8 54

## [1] "Forest Model Accuracy"

## [1] 0.67

## [1] "Null Model Accuracy"

## [1] 0.62
```

PCA

```
##    yhat
## y    0  1
##   0 14 24
##   1 11 51

## [1] "PCA Model Accuracy"

## [1] 0.65

## [1] "Null Model Accuracy"

## [1] 0.62
```

IV. Results

PCA:

PCA Absolute Improvement = 3%, Lift = 1.05

Lasso selected variables

Probit Model: Absolute Improvement = 2%, Lift = 1.03

Ordered Probit Model: Absolute Improvement = -3%, Lift = 0.87

Logit Model: Absolute Improvement = 6%, Lift = 1.09

Random Forest: Absolute Improvement = 3%, Lift = 1.05

Step-wise selected variables

Probit Model: Absolute Improvement = 6%, Lift = 1.10

Ordered Probit Model: Absolute Improvement = -3%, Lift = 0.87

Logit Model: Absolute Improvement = 6%, Lift = 1.10

Random Forest: Absolute Improvement = 5%, Lift = 1.08

V. Conclusion

In the end none of our models showed consistent improvement over the base model that predicts the higher seed (Home team) to be the winner of the match up. The improvement and lift vary significantly depending on the seed used and are not reliable. Upon discussion we came up with several potential reasons for that.

Firstly, our data set is not expansive enough. With 60 variables and only 300 observation that is not rigorous enough to run good train/test splits. The accuracy estimates for both our model and the null model vary wildly depending on the selection of observations in the testing split making it difficult to get a comparison of the true improvement and lift of our model, the above results are merely one sample. We attempted to account for over fitting in the training set by resampling the training set and selecting parameters that perfomred well on all bootstrapped samples, but that technique did not show much improvement either.

Another possible explanation is that our model was not able to provide a good indication of who will win a playoff series because our data set did not include some of the potential confounding variables. Since NHL data set that we used simply calculates the season average statistics, we don't account for how teams are performing at the end of the year which may change significantly.

Another issue with not emphasizing the end of year performance is that it is very common practice for teams that are playoff bound is to improve their roster by bringing on high quality players later in the season. The improvement in performance form these players is not likely to show up on the team's season long statistics. To account for that a good statistic to add would have been a salary that is added in mid to late season trades. The assumption there is that a players performance should be highly correlated with their salary. The nature of the NHL's hard salary cap means that we could also calculate the percentage of the total salary cap occupied by an incoming player, this would control for salary differences over time.

The next confounder that we could not account for is the health of the team. If an important player on the team was out for most of the season due to a serious injury that would dramatically reduce their regular season stats, which our models would predict to reflect in their playoff runs. The reverse of that scenario would be true as well, an effect of high impact player that going down right before the beginning of the playoffs would not show up in any of our models.

To summarize, step-wise selection seemed to perform better across all of our models, but the variance of our attempts to measure model accuracy puts these results in question. More data would and needed to properly run these algorithms, specifically enough data that the proportion of home wins in the test set does not vary so wildly.