ECO395K

Data Mining and Machin Learning

Professor James Scott

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

By Alex Mykietyn, Cameron Whetley and Daniil Deych

1. 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 for NHL playoff team against an opponent, using the regular season relative statistics between the two teams as predictive variables for the model. Using the difference between corresponding stats for the head-to-head playoff teams instead of simple raw stats, allows us to discount for many potential confounders that our small (comparatively) data set could not account for.

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. Comparing supervised learning tools such as linear regression, probit and ordered probit models, lasso and step-wise selection and unsupervised learning approaches - random forest, boosted tree model and a Principle Component Analysis. The end goal being, 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).

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. Power Play efficiency on defense or offense both are very relevant hockey statistics and are likely to play a crucial role in determining season performance.

1. 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 season, lockout seasons of 2012-13 and 2004-05 did not have games at all.

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.

1. Method

The main challenge with our approach is our relatively small data subset. With only 20 seasons (and thus only 20 Stanley Cup winners) and 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 stepwise selection and lasso approach to reduce the number of variables to 8 and then 5.

Once we have identified our list of best predictive variables, we then use those variables to build our list of models: linear, probit, ordered probit, random forest, boosted trees model and principal component analysis. We then test 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.

Finally, once we identify our best model, we will use it to make the predictions for ongoing Stanley Cup playoffs.

1. Results
2. 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. 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, as especially with a smaller test set the variance in the data is bound to be captured in the model. We attempted to account for that using as k-fold cross validation, but that technique did not show much improvement either.

Another possible explanation is that 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, and the team performance is more relevant to how the team ends the season, our data does not account for that kind of heterogeneity.

A very common practice for teams that are playoff bound is to improve their roster by bringing on high quality players later in the season, whose performance 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 share of the salary cap that is being used by the team, as the team goes into the playoffs. The assumption there being that teams with no salary space under the cap are likely to have the better players than the ones that do.

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,

References

Weissbock, J., Viktor, H., & Inkpen, D. (2013). Use of Performance Metrics to Forecast Success in the National Hockey League.