Patrick Jordan

Amir Jafari

Data Science Capstone

December 11, 2018

Analysis of the NHL Data API: What Does it Take to be a Championship Team?

I decided to conduct my capstone project on the National Hockey League's rich data, and more specifically I wanted to see if I could pull out from statistical analysis what exactly separates a 'great' team from a championship team. All cards on the table I am a huge Washington Capitals fan, I was immediately enamored with the team from the first NHL game I attended shortly after arriving to the Washington DC metro area in 2006. That was the second half of Alexander Ovechkin's rookie year, and the building was electric. Since then the Capitals have won the President's Trophy, given to the NHL team with the best record in the regular season, three times, back to back from the 2015-2017 seasons. However, winning this trophy does not guarantee success in the playoffs, in fact since Ovechkin has come into the league only two teams have gone on to win the Stanley Cup, the leagues iconic championship trophy, after notching the best record in the regular season and one of those teams was the Chicago Blackhawks in the 2012-2013 shortened lockout season. The point is the Capitals arguably have been one of the best teams of the last decade, and its teams from 2015-2017 were arguable the best of the bunch on paper and statistically in the regular season, but they could not finish the season as champions, they could not even get past the second round of the playoffs.

So, at the end of the 2016-2017 campaign while hindered by the NHL's salary cap, the team let go or traded away a number of their top veteran players. They additionally lost one of their top defensive prospects during an expansion draft to the Las Vegas Golden Knights. Most

of the league had written them off as having a rebuilding year. The Capitals were forced to fill in positions with a number of young talents from their minor league system, they finished first in their division but were not in the running for the President's Trophy. Yet this team made it through the playoffs and won the championship! So, the focus of my project was to look across the league at the team level statistics to identify a delta from teams past.

The project was done using python, I wanted to challenge myself to get better in this coding language, and while I would have preferred to conduct this in R, I learned so much more doing it in python. I initially explored the NHL's Data API [1]. It is incredibly rich with statistical data going back to the inception of the league in 1917. The API, however is not documented by the league, but luckily for me there were a number of fans who went through a lot of trouble figuring out what the endpoints were, without these resources I would not have been able to start exploring the API [2,3]. Once I was able to get into the API, and get a data payload returned I realized that like most modern data sources, the NHL returned their data in a JSON payload. This would be great if I had been well versed in how to appropriately unpack said JSON, and if it was a one-layer JSON I could have covered that as well, however this is a heavily nested structure and I could not figure out how to flatten the files in order to use them, that will be a learning task for another day. I was fortunate to find a python package that would scrape and return the play-by-play data for me by season and package it nicely into a csv file [4]. This tool was a life-saver; however, it did take a long time to download the data. One season took on average four hours to complete and each file contained about 650k rows of data for every game in the season to include playoffs.

The play-by-play endpoint is a live feed, and will give you information as its recorded, down to the x,y coordinate on the ice for each play. The information is well structured and has a

schema, but most of the data is textual. So, the next task was to transform it into statistical information so that I could conduct analysis on it. Once compete I created visualizations of the data by season which is included in the data folder on GitHub. Here is the rub, of the teams that won the President's trophy, and the 2018 team that won the cup, aside from a slightly higher shooting percentage, and obvious goals for/against differential nothing really stood out. The 2018 did have higher per game blocks for in the playoffs which certainly helped, but the goalie still faced on average the same amount of shots he did the year before. The scope of this project did not allow me to dive into individual player statistics and I believe looking at that would yield better results, but from the team aspect I could not pinpoint the statistic that mattered most, aside from the obvious goals for and against. As one of the local sports anchors said after the Capitals closed out the Pittsburgh series, "The young guys on this team are too dumb to know that they are supposed to lose".

From there I created two Machine learning models, first I removed all stats derived from goals scored aside from the shooting percentage. And then created a naïve Bayesian model that performed 67 and 69% accuracy, and a Logistic Regression leveraging Grid search with five-fold cross validation that performed between 98 and 100% accuracy at classifying a win from the statistical data of the game. Unfortunately, I did not build the models such that they could forecast a seasons results given part of the seasons performance, but I would like to explore that at a later time.

Works Cited

1. National Hockey League, ***NHL Data API***, **https://statsapi.web.nhl.com/api/v1**
2. Sidwar Kevin, ***The Undocumented NHL Stats API,*** https://www.kevinsidwar.com/iot/2017/7/1/the-undocumented-nhl-stats-api
3. Hynes Drew, ***NHL Stats API Documentation,*** https://gitlab.com/dword4/nhlapi/blob/master/stats-api.md
4. Shomer Harry, ***Hockey-Scraper,*** https://github.com/HarryShomer/Hockey-Scraper