Josh Croasdale

**02/16/2024**

# CAPSTONE PROJECT - SPRINT 0

- **The Problem area: What is your area of interest? Within it, what challenges or opportunities could your project address?**

The area of interest for my capstone project is the NHL (National Hockey League) and the statistical analysis of player performance. More specifically, I would like to create a model that predicts future player performance, based on past performance indicators. At the very core, my goal is to predict two common counting stats: goals and assists, for a given player based on historical data. I recognize that this sort of predictive analysis can get very complex. The most basic (and inaccurate) form of this process would be to take the sum of each of these two stats for a given player, and divide them by their number of games played to get an average over their career. In reality, there are so many factors that may influence these counting stats and there are so many metrics that can and are used to make predictions. A challenge for me will be to determine an appropriate and practical number of factors to consider in my analysis.

To give an idea of the potential complexity of this problem, here is a non-exhaustive list of some factors that may influence how many goals or assists a given NHL player will produce:

- Past performance
- Position
- Ice time

- Power play ice time
- Quality of team
- Quality of linemates
- Quality of opponents
- Shots on goal
- Shot attempts
- Age
- Experience

These factors can vary from season to season and even from game to game. The best predictive models are able to incorporate as many factors as possible and weight them accordingly. There are so many more factors that may be considered, depending on how technical or abstract you wanted to get with your modeling. .

One big area of opportunity that my project could target is in sports betting. Sports betting is big business. If bettors had reliable predictive models at their disposal, they may be more confident in making certain bets in their attempt to beat the betting market. On the flipside, the market setters use predictive modeling in order to set the very betting lines that bettors try to beat.

Another potential link my capstone project would have (and more near and dear to my heart) is to fantasy hockey. In this realm you essentially try to pick the best players in order to build a “fantasy team” of players in an effort to outperform other people who have also picked teams. Accurate predictive models may be able to help people pick the best players.

- **The User: Who experiences these problems? How would they benefit from the outcomes of your project?**

I have touched on this in the previous discussion. See above. Many people try to predict these metrics (goals, assists and much more) on a daily basis and would stand to benefit from more accurate predictive models (beating the betting markets, beating your friends in fantasy).

- **The Big Idea: How can machine learning bring solutions to these areas? Research how other people have approached the problem previously. Refer to the "Intro to Capstone" slides on synapse for an overview of different machine learning approaches.**

I may use numerical prediction in my machine learning approach. Target variables: goals, assists. Takehiro Matsuzawa published a paper[1] where he uses many methods, including LinearRegressions, K-NearestNeighborhoodRegressions, RandomForestRegression, and Neural Network Regression. His goal was to predict how a given player in the NHL would perform in their next 5 games, based on data from their previous 10 games. Hockey-statistics.com provides insights into player game contribution with a Deep Learning Modeling application[2].

- **The Impact: What societal or business value do you anticipate your project to add? If possible, try to quantify the scale of the problem (in dollars, in CO2, in time spent, ...)**

Effective predictive modeling has the potential to be highly influential in the world of sports betting. The sports betting market is predicted to reach $USD 182 Billion by 2030[3]. Hockey in particular is a more difficult sport to predict due to the inherent randomness that exists in the game (high speeds, many players, small and fast moving puck, influence of goaltending). My proposed project will start small but will hopefully be built upon in the future.

- **The Data: Identify several possible datasets in this subject area and describe them at a high level. Include references. If you struggle to find more than one or two datasets, this might mean a Data Science approach to the problem will be challenging. Check in with your Educator.**

There are many sources of NHL game data:

- https://moneypuck.com/ provides extensive player level and team level data ranging from the current NHL season all the way back to the 2008-2009 season. They also provide a data dictionary that explains many of the columns and metrics in the dataset. At this stage in my process, this is all of the data that I anticipate needing for this project. The data will need to be combined into one dataset, as currently each NHL season is contained in a separate .csv file. I plan to only include players in my dataset who have met a certain threshold of minimum games played. Data for “minor leaguers” and fringe players is unnecessary. This becomes a subjective matter to decide what the minimum games played cutoff will be. More research is needed. Another challenge for this project will be what to do with rookie players. They have no historical NHL data to draw from, but yet many of them are crucial players in the league today and should be included in any up-to-date analysis.

- https://www.hockey-reference.com/leagues/NHL_2024_skaters.html provides extensive raw data. Although I need to figure out how to convert raw text into a csv file
- https://gitlab.com/dword4/nhlapi has NHL API Documentation.

- **The Alternative: In a few sentences, summarize a problem in an alternative subject area that also interests you.**

I have a personal history in powerlifting and would be interested in exploring the sport from an analytical perspective. A few ideas:

- What impact do different parameters have on success in the sport (age, sex, height, weight, nationality, ethnicity)?
- Can we predict future world record totals (lifting total = squat + benchpress + deadlift) based on historical trends?

https://www.openpowerlifting.org/ provides a wealth of data

# References


1. Using Machine Learning to Predict Future Points in the NHL, TakehiroMatsuzawa. https://dash.harvard.edu/bitstream/handle/1/37366468/MATSUZAWA-SENIORTHESIS-2017.pdf?sequence=1
2. https://hockey-statistics.com/2021/08/25/deep-learning-modeling-of-hockey-game-contribution/
3. https://www.grandviewresearch.com/industry-analysis/sports-betting-market-report
4. https://moneypuck.com/data.htm