Joshua Croasdale
Spring 2024

Capstone Project

Predicting NHL Player Performance Using Linear Regression and Other ML models

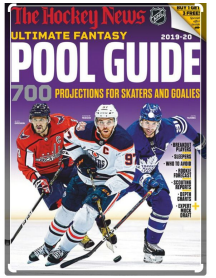
Overview

- The NHL is an entertainment product
  - People become highly invested in the outcomes
    - Financially invested (betting, jobs)
    - Personally invested (passion)
- Big business in Canada and the US
  - Ticket sales, merchandise sales, endorsements
  - Online **Sports gambling** industry is worth 30 billion USD in North America in 2024![1]
- Fantasy hockey → Ever-growing industry
  - Not as economically impactful as sports betting
    - More of a passion-driven endeavour
    - But there is an industry built around fantasy sports
  - **More focused on season-long outcomes** as opposed to individual games

1. https://www.statista.com/outlook/dmo/eservices/online-gambling/north-america

Hypothetical Scenario

- I am an aspiring data scientist and statistician that has been hired by a new sports news company to flesh out the ‘Fantasy Hockey’ section of their website
- First order of business: **Create a model that will predict (hopefully well) the season-long statistical outcomes for any given NHL player for the upcoming season**



- The company charges users a fee for access to the model
- User can input the specific parameters of their own fantasy league and receive detailed projections for every player
  - Use these projections to help them win their fantasy leagues!

Narrowing It Down…

- For the scope of this analysis I am going to focus on predicting two important stats:
  - **Goals** → Goal is awarded to the last skater to touch the puck before it crosses the opposing team's goal line
  - **Assists** → awarded to up to two players who touched the puck immediately before the goal scorer put it across the goal line
- This is a multiple linear regression problem
  - Target variable is continuous
- In the future…
  - Classification → game-by-game predictions
    - Will a given player score a goal in their next game?
    - Will a given team win their next game?

The Dataset

- All player data was provided by:

  https://www.hockey-reference.com/leagues/NHL_2024_skaters.html

- NHL player data for each season from 2005-06 to 2023-24 (19 seasons) was exported into a .csv format and saved into individual .csv files using a text editor
  - All files stored in one folder
- All .csv files were compiled into one dataframe with python and processed from there
  - I added a ‘Year’ column to every individual dataset
- Do I have enough data?
  - Each row of my dataframe is an entire season’s worth of data for one player!

Next Steps…

- Drop features that have high covariance and run analysis again

- Use a three-year weighted average of most recent three seasons, with the remainder of career averaged out and weighted accordingly
  - Can this be done? YES
  - How will this work with test vs train data?

Obstacles and Limitations

- Inherent randomness of ice hockey
- Every player is unique!
- Quality of team
- Quality of linemates
- Quality of competition
- Assists
  - Primary vs Secondary assists
- Aging curves of players
- What to do with rookies?
- What to do with ‘tweeners’?