

Predicting NHL Goal Scoring



Capstone Sprint 1

Taylor Gallivan

20 Oct. 2023
































"When I started [in the NHL in 2015], it was, 'Oh, if they have one person in analytics, they're so innovative.' Now if you don't have more than one person, you're behind."

- *Alexandra Mandrycky, Assistant GM, Seattle Kraken, 2022*

As of Oct. 2022:

- ➔ 23 teams w/ 3 or more analytical hires (72% of league)
- ➔ 9 teams w/ 5 or more analytical hires (28% of league)
- ➔ Most? Maple Leafs with 8 ... but still no Cup

'Analytical' hires in NHL front offices

 Jeff Solomon	Vice-President of Hockey Operations	 Zac Urbach	Hockey Analyst	 Toran Singleton	Analytics	 Benjamin Morgan	Statistical Analyst
 Ryan Lichtenfels	Director of Team Services	 Josh Weissbach	Hockey Analytics Consultant	 Sul Tarnambene	Analytics	 Ken Burkler	Data Science Programmer
 Gabriela Switaj	Team Services Analyst & Amateur Scout	 Cori Lawrence	Hockey Analytics Consultant	 Alex Martynov	Analytics	 Cheryl Metcalf	Special Assistant to the General Manager
 Matt Poni	Director of Analytics	 Mark Janick	Assistant General Manager	 Sam Fortner	Analytics	 Rob Petrasavage	Senior Analyst, Hockey Research and Development
 Chase Watson	Full Stack and Data Engineer	 Steve Orsley	Director of Hockey Strategy/Scouting and Development	 Kelly Keogh	Analytics	 Judy Cohen	Analyst, Hockey Research and Development
 Leo Stempniak	Hockey Data Strategist	 Alex LePore	Hockey Analytics Coordinator/Professional Scout	 Jon Sullivan	Vice President of Hockey Strategy & Data Management	 Bruce Peter	Analyst, Hockey Research and Development
 Chase Glasberg	Assistant Data Analyst	 Ben Lines	Hockey Analytics Coordinator/Amateur Scout	 Richard Dry	Director of Sports Technology	 Jill Reimer	Analyst, Hockey Research & Development
 Jeremy Rogalski	Director of Hockey Analytics	 Bryan Campbell	Director of Statistical Analysis and Hockey Administration	 Michael Dillon	Performance Analyst	 Wesley Waldner	Senior Development, Hockey Research and Development
 Joshua Puhkamp-Hart	Hockey Operations Data Scientist	 Dan Kosinski	Hockey Operations Data Analyst	 Daniel Hovasse	Software Developer/Data Analyst	 Jed Ong	Developer, Hockey Research and Development
 Campbell Weaver	Data Engineer, Hockey Operations	 Justin Mahe	Manager of Hockey Analytics	 Kara Stephan	Full Stack Engineer	 Andrew Law	Developer, Hockey Research and Development
 Seth Reimer	Data Engineer	 Shaun Mahe	Video Coach & Hockey Operations Coordinator	 Tim Patsyoun	Director of Hockey Operations	 Adam Fox	Senior Analyst, Hockey Analytics
 Dan Panfili	Developer	 Sunny Mehta	Vice President of Hockey Strategy & Intelligence	 Elias Collette	Analytics Consultant	 Ryan Blech	Video Analyst, Hockey Analytics
 Jason Nightingale	Assistant Director of Amateur Scouting	 Tom Bark	Assistant to the General Manager	 Ian Anderson	Director of Hockey Analytics	 Miss Hooken	Junior Analyst
 Jason Karmann	Associate General Manager	 Ryan Kruse	Vice President of Research and Development	 Jacob Hurtburt	Lead Developer	 Rachel Doernie	Analyst, Hockey Analytics
 Sam Ventura	Vice President of Hockey Strategy and Research	 Josh Vulliamy	Senior Analyst	 Tim Minton	Director of Hockey Information/Video	 Misha Doroslov	Director of Hockey Operations
 Domonic Galambin Jr.	Data Scientist	 Hayden Speck	Analyst	 Cole Anderson	Lead Data Scientist	 Tom Poraszko	Hockey Operations Analyst
 Matthew Barlowe	Data Engineer	 Jason Lewis	Video Technician	 Matthew Karlner	Analyst	 Dustin Walsh	Hockey Operations Analyst
 Chris Snow	Assistant General Manager	 Rosie Yu	Software Engineer for Research and Development	 Kathryn Yates	Hockey Analyst	 Tim Barnes	Director of Hockey Analytics
 David Johnson	Database Architect and Analyst	 Matt Sells	Vice President of Hockey Strategy	 Nick Clynne	Senior Data Scientist, Hockey and Business Operations	 H.T. Lenz	Manager of Hockey Analytics
 Michael Cheron	Quantitative Analyst	 Christopher Boucher	Director of Hockey Analytics	 Katarina Wu	Data Scientist	 Adam Koneff	Research & Data Coordinator
 Connor Rankin	Video Analyst	 John Sandgwick	Vice President of Hockey Operations and Legal Affairs	 Andy Sautler	Hockey Operations Analyst	 Jordy Finnigan	Data & Video Coordinator
 Eric Tubley	Assistant General Manager	 Mario LeBlanc	Video Coach	 Doug Wilson Jr.	Director of Scouting	 Matt Prebotaine	Video and Analytics Coach
 Kevin Kan	Senior Developer, Hockey Operations	 Adam Douglas	Sports Science and Performance Director	 Charlie Townsend	Hockey Analyst/Assistant to the NHL Coaching Staff	 Neil Platon	Vice President, Statistical Data
 Matt Waller	Data Engineer	 Matt Hamann	Hockey Operations Analytics Coordinator	 Alexandra Mandrycky	Director of Hockey Strategy & Research	 Tim Seppa	SQL Developer & Stats Analyst
 Margaret Conniff	Data Scientist	 Dalton Linkus	Research and Data Development Engineer	 Dori Chu	Quantitative Analyst	 Christopher Baker	SQL Developer & Stats Analyst
 Arik Parnas	Director of Analytics	 Tyler Dellow	Vice President of Hockey Analytics	 Nemita Nandakumar	Senior Quantitative Analyst	 Michael Roobis	SQL Developer & Stats Analyst
 Dawson Springles	Associate Director of Analytics & Lead Data Scientist	 Matt Cano	Director of Hockey Analytics	 Eric Mathison	Hockey Operations Data Developer	 Connor Reed	SQL Developer & Stats Analyst
 David Wood	Hockey Analyst	 Jared Lumsford	Data Scientist	 John Macisoulis	Hockey Operations Data Engineer	 Julianne Falvo	SQL Developer & Stats Analyst
 Kyle Davidson	Vice President of Hockey Strategy & Analytics	 Jon Fung	Software Developer	 Tim Ohaai	Head Video Analyst	 Christopher Remondelli	SQL Developer & Stats Analyst
 Mary Dettartolo	Coordinator, Research & Development/Hockey Analytics	 Craig Lewis	Software Developer	 Ryan Miller	Director of Hockey Operations	 Samuel Wood	Stats Author & Research Analyst
 Andrew Conlis	Hockey Analytics/Video Analyst	 Austin Wallace	Data Engineer	 Mike Penhman	Hockey Data Analyst	 Nicholas Carr	Stats Author & Research Analyst
 Josh Flynn	Assistant General Manager	 Frank Gardner	Analytics	 Michael Peterson	Director of Hockey Analytics		

Source: The Athletic, Oct. 2022

Source: The Athletic, Oct. 2022

Project Overview

- The sports analytics industry has been growing consistently since the early 2000's → first widely adopted in the MLB, but now a prominent feature in every major North American league (MLB, NBA, NFL, NHL, MLS)
- Have been a hockey fan as long as I can remember: fond memories attending games, cheering on Team Canada, watching Hockey Night in Canada
- With the bevy of NHL data available, merging my hockey fandom with data science was a perfect fit for a capstone project

Problem Question:

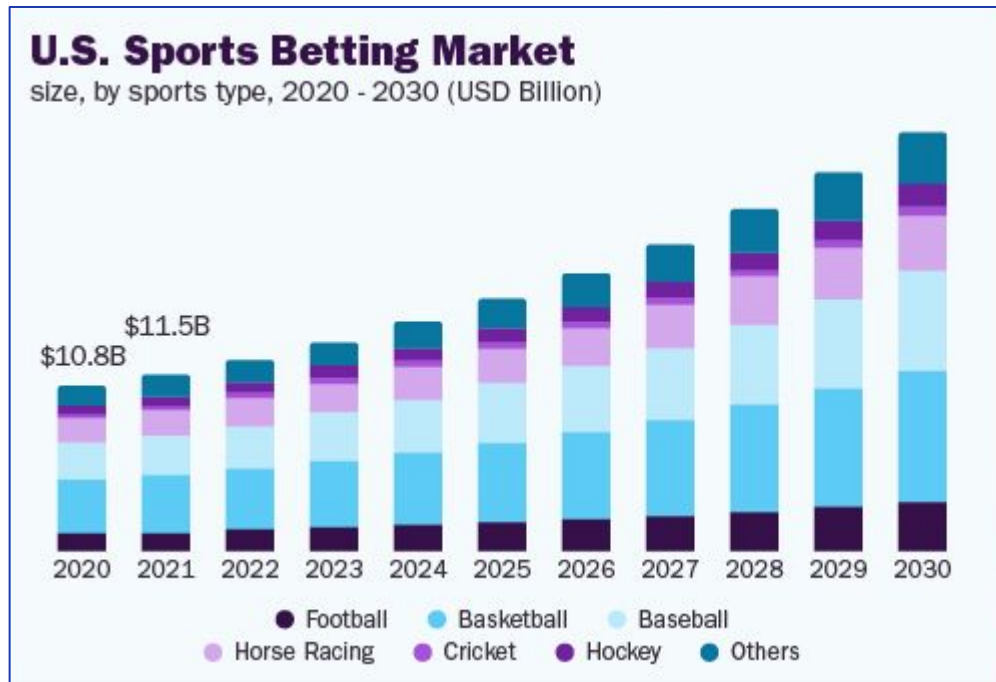
Can an ML model (or models) be trained to accurately predict a player's goal output for a season, based on their individual characteristics (a mix of statistical and categorical features)

Project Vision

Though not well documented, the NHL makes large portions of their API (Application Programming Interface) open to public access:

1. Retrieve statistical data directly from the NHL API, for all seasons between 1990-1991 and 2022-2023
2. Perform EDA on the cleaned dataset to identify preliminary trends within predictive variables → currently 47 features in dataset
3. Fit and test models for accuracy - current candidates include:
 - a. Generalized Linear Model (GLM)
 - b. Gaussian Process Regression (GPR)
 - c. Support Vector Machine (SVM) Regression
4. Target variable: total goals scored (in an 82 game season)

Potential Impact



Source: GrandView Research

Sports betting is big business:

→ USD 83.6 billion market value in 2022 (global)

→ ~10% CAGR expected growth

Pipedream: sell my model to a Vegas sportsbook & retire at 35

Realistic Outcome: use it for my fantasy hockey draft and still finish third place

The Data...

30,000 player ID's:

- Returns a DataFrame with season-by-season statistics (goals, shots, shooting percentage, etc.)
- Irrelevant positions excluded (goalies) as well as fringe players (less than 3 seasons or 200 total games played)
- Concerns: ability to account for external/untracked factors like contract status, injury history, player conditioning

```
# Call 1: 8477246 - 8482246

main_df_test = pd.DataFrame()
base_url = 'https://statsapi.web.nhl.com/api/v1/people/'
rangel = range(8477246, 8482246)

for num in rangel:
    people_url = f'{base_url}{num}'
    response = requests.get(people_url)

    if response.status_code != 404:
        suggestions = json.loads(response.content)['people']
        player = (pd.json_normalize(suggestions))

        if player['primaryPosition.code'][0] != 'G':
            stats_url = f'{base_url}{num}/stats/?stats=yearByYear'
            response = requests.get(stats_url)
            content = json.loads(response.content)['stats']
            splits = content[0]['splits']

            df_splits = (pd.json_normalize(splits, sep = "-" )
                        ).query("league_name == 'National Hockey League'")

            if df_splits.shape[0] >= 3:
                df_splits['player_id'] = player['id'][0]
                df_splits['first_name'] = player['firstName'][0]
                df_splits['last_name'] = player['lastName'][0]
                df_splits['position_code'] = player['primaryPosition.code'][0]
                df_splits['stat_games'] = df_splits['stat_games'].astype(int)
                total_games = df_splits.groupby(['player_id', 'first_name', 'last_name', 'position_code'])['stat_games'].sum().reset_index()
                filtered_total_games = total_games[total_games['stat_games'] > 200]

                if not filtered_total_games.empty:
                    df_splits['season_start_yr'] = [x[0:4] for x in df_splits['season']]
                    df_splits['season_start_dt'] = [datetime.strptime(x + '0930', "%Y%m%d") for x in df_splits['season_start_yr']]
                    df_splits['season_end'] = [x[4:8] for x in df_splits['season']]

                    df_splits['weight'] = player['weight'][0]
                    df_splits['height'] = player['height'][0]
                    df_splits['shot_dir'] = player['shootsCatches'][0]
                    df_splits['birth_date'] = pd.to_datetime(player['birthDate'][0])
                    df_splits['age'] = (np.floor((df_splits['season_start_dt'] - df_splits['birth_date']) / np.timedelta64(1, 'Y')) )
                    df_splits['age'] = df_splits['age'].astype(int)
                    df_splits['position_name'] = player['primaryPosition.name'][0]
                    df_splits['position_type'] = player['primaryPosition.type'][0]
                    df_splits['birth_country'] = player['birthCountry'][0]
                    df_splits['nationality'] = player['nationality'][0]

                    main_df_test = pd.concat([main_df_test, df_splits], sort=False).reset_index(drop=True)
                else:
                    pass
            else:
                pass
        else:
            pass
    else:
        pass
```

Next Steps

1. Finish compiling data set
 - a. Check for duplicates, remove unnecessary columns
 - b. Convert 'time on ice' values from string to datetime data types
 - c. Make franchise names consistent
 - d. Adjust scoring totals for era
 - e. Adjust scoring for shortened seasons
 - f. Feature selection
 - i. Evaluate applicability of aggregated values (for instance, 3-yr weighted averages vs. most recent season's stats)
 - ii. Evaluate the difficulty/practicality of including additional external variables