Predicting NHL Goal Scoring

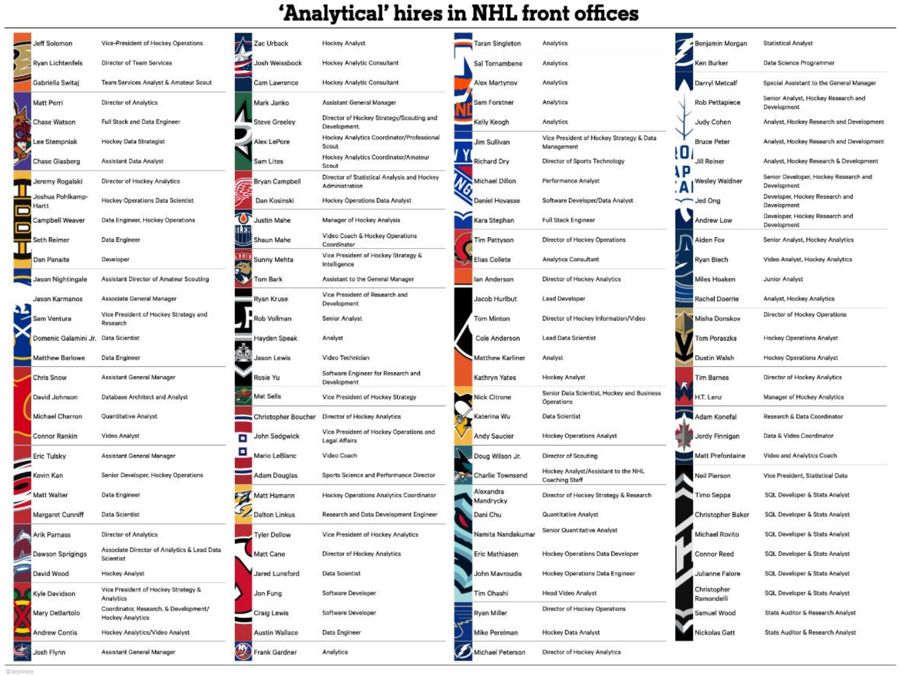
Capstone Sprint 1

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*



'Analytical' hires in NHL front offices

Jeff Solomon	Vice-President of Hockey Operations
Ryan Lichtenfels	Director of Team Services
Gabriella Switaj	Team Services Analyst & Amateur Scout
Matt Perri	Director of Analytics
Chase Watson	Full Stack and Data Engineer
Lee Stempniak	Hockey Data Strategist
Chase Glasberg	Assistant Data Analyst
Jeremy Rogalski	Director of Hockey Analytics
Joshua Pohlkamp-Hartt	Hockey Operations Data Scientist
Campbell Weaver	Data Engineer, Hockey Operations
Seth Reimer	Data Engineer
Dan Panaite	Developer
Jason Nightingale	Assistant Director of Amateur Scouting
Jason Karmanos	Associate General Manager
Sam Ventura	Vice President of Hockey Strategy and Research
Domenic Galamini Jr.	Data Scientist
Matthew Barlowe	Data Engineer
Chris Snow	Assistant General Manager
David Johnson	Database Architect and Analyst
Michael Charron	Quantitative Analyst
Connor Rankin	Video Analyst
Eric Tulsky	Assistant General Manager
Kevin Kan	Senior Developer, Hockey Operations
Matt Walter	Data Engineer
Margaret Cunniff	Data Scientist
Arik Parnass	Director of Analytics
Dawson Sprigings	Associate Director of Analytics & Lead Data Scientist
David Wood	Hockey Analyst
Kyle Davidson	Vice President of Hockey Strategy & Analytics
Mary DeBartolo	Coordinator, Research, & Development/ Hockey Analytics
Andrew Contis	Hockey Analytics/Video Analyst
Josh Flynn	Assistant General Manager
Zac Urback	Hockey Analyst
Josh Weissbock	Hockey Analytic Consultant
Cam Lawrence	Hockey Analytic Consultant
Mark Janko	Assistant General Manager
Steve Greeley	Director of Hockey Strategy/Scouting and Development.
Alex LePore	Hockey Analytics Coordinator/Professional Scout
Sam Lites	Hockey Analytics Coordinator/Amateur Scout
Bryan Campbell	Director of Statistical Analysis and Hockey Administration
Dan Kosinski	Hockey Operations Data Analyst
Justin Mahe	Manager of Hockey Analysis
Shaun Mahe	Video Coach & Hockey Operations Coordinator
Sunny Mehta	Vice President of Hockey Strategy & Intelligence
Tom Bark	Assistant to the General Manager
Ryan Kruse	Vice President of Research and Development
Rob Vollman	Senior Analyst
Hayden Speak	Analyst
Jason Lewis	Video Technician
Rosie Yu	Software Engineer for Research and Development
Mat Sells	Vice President of Hockey Strategy
Christopher Boucher	Director of Hockey Analytics
John Sedgwick	Vice President of Hockey Operations and Legal Affairs
Mario LeBlanc	Video Coach
Adam Douglas	Sports Science and Performance Director
Matt Hamann	Hockey Operations Analytics Coordinator
Dalton Linkus	Research and Data Development Engineer
Tyler Dellow	Vice President of Hockey Analytics
Matt Cane	Director of Hockey Analytics
Jared Lunsford	Data Scientist
Jon Fung	Software Developer
Craig Lewis	Software Developer
Austin Wallace	Data Engineer
Frank Gardner	Analytics
Taran Singleton	Analytics
Sal Tornambene	Analytics
Alex Martynov	Analytics
Sam Forstner	Analytics
Kelly Keogh	Analytics
Jim Sullivan	Vice President of Hockey Strategy & Data Management
Richard Dry	Director of Sports Technology
Michael Dillon	Performance Analyst
Daniel Hovasse	Software Developer/Data Analyst
Kara Stephan	Full Stack Engineer
Tim Pattyson	Director of Hockey Operations
Elias Collete	Analytics Consultant
Ian Anderson	Director of Hockey Analytics
Jacob Hurlbut	Lead Developer
Tom Minton	Director of Hockey Information/Video
Cole Anderson	Lead Data Scientist
Matthew Karliner	Analyst
Kathryn Yates	Hockey Analyst
Nick Citrone	Senior Data Scientist, Hockey and Business Operations
Katerina Wu	Data Scientist
Andy Saucier	Hockey Operations Analyst
Doug Wilson Jr.	Director of Scouting
Charlie Townsend	Hockey Analyst/Assistant to the NHL Coaching Staff
Alexandra Mandrycky	Director of Hockey Strategy & Research
Dani Chu	Quantitative Analyst
Namita Nandakumar	Senior Quantitative Analyst
Eric Mathiasen	Hockey Operations Data Developer
John Mavroudis	Hockey Operations Data Engineer
Tim Ohashi	Head Video Analyst
Ryan Miller	Director of Hockey Operations
Mike Perelman	Hockey Data Analyst
Michael Peterson	Director of Hockey Analytics
Benjamin Morgan	Statistical Analyst
Ken Burker	Data Science Programmer
Darryl Metcalf	Special Assistant to the General Manager
Rob Pettapiece	Senior Analyst, Hockey Research and Development
Judy Cohen	Analyst, Hockey Research and Development
Bruce Peter	Analyst, Hockey Research and Development
Jill Reiner	Analyst, Hockey Research & Development
Wesley Waldner	Senior Developer, Hockey Research and Development
Jed Ong	Developer, Hockey Research and Development
Andrew Low	Developer, Hockey Research and Development
Aiden Fox	Senior Analyst, Hockey Analytics
Ryan Biech	Video Analyst, Hockey Analytics
Miles Hoaken	Junior Analyst
Rachel Doerrie	Analyst, Hockey Analytics
Misha Donskov	Director of Hockey Operations
Tom Poraszka	Hockey Operations Analyst
Dustin Walsh	Hockey Operations Analyst
Tim Barnes	Director of Hockey Analytics
H.T. Lenz	Manager of Hockey Analytics
Adam Konefal	Research & Data Coordinator
Jordy Finnigan	Data & Video Coordinator
Matt Prefontaine	Video and Analytics Coach
Neil Pierson	Vice President, Statistical Data
Timo Seppa	SQL Developer & Stats Analyst
Christopher Baker	SQL Developer & Stats Analyst
Michael Rovito	SQL Developer & Stats Analyst
Connor Reed	SQL Developer & Stats Analyst
Julianne Falore	SQL Developer & Stats Analyst
Christopher Ramondelli	SQL Developer & Stats Analyst
Samuel Wood	Stats Auditor & Research Analyst
Nickolas Gatt	Stats Auditor & Research Analyst

Source: The Athletic, Oct. 2022

As of Oct. 2022:

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

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)
- With the bevy of NHL data available, merging my hockey fandom with data science was a perfect fit

Problem Statement:

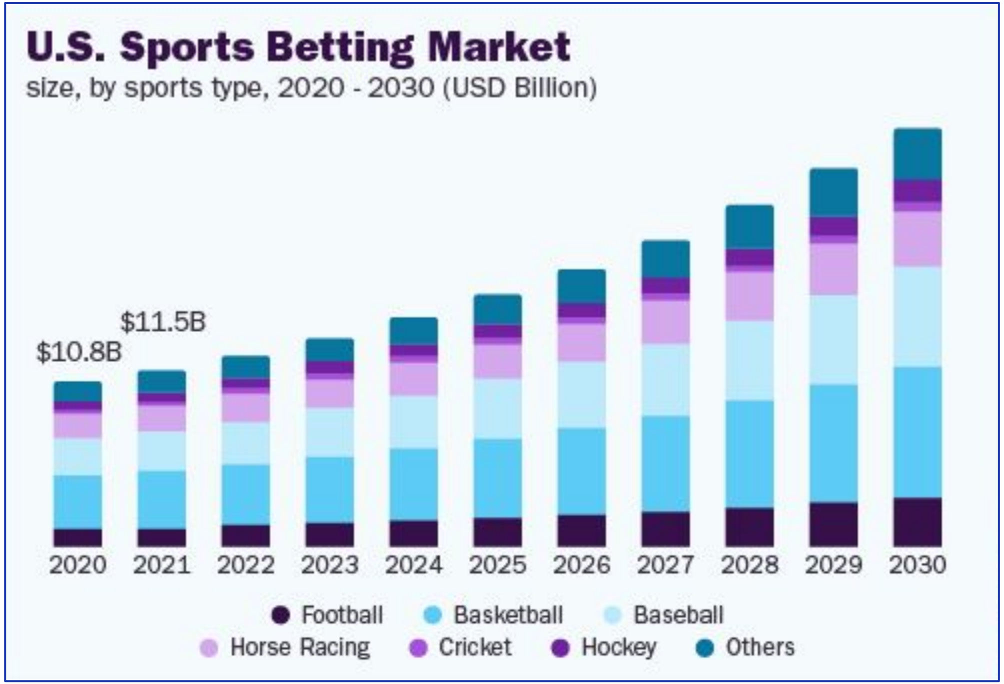
Using season-by-season player statistics, fit an ML model that can effectively predict next season's goal total for a given player

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 (dataset currently has 47 columns)
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

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

```python
# Call 1:  8477246 - 8482246

main_df_test = pd.DataFrame()
base_url = 'https://statsapi.web.nhl.com/api/v1/people/'
range1 = range(8477246, 8482246)

for num in range1:
    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. Remove unnecessary columns, check for duplicates
    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. Evaluate applicability of aggregated values (for instance, 3-yr weighted averages vs. most recent season’s stats)
    g. Evaluate the difficulty/practicality of including additional external variables