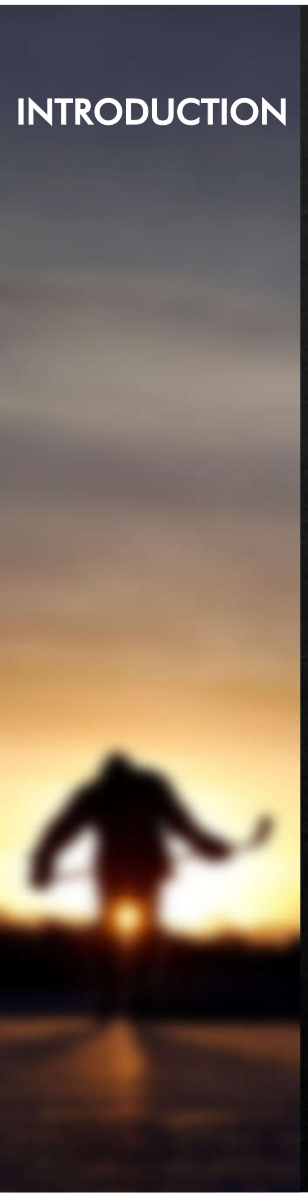
Scouting with AI

Tom Callegari

UNIVERSITY of TORONTO
SCHOOL of CONTINUING STUDIES
Data Analytics Bootcamp

2019



INTRODUCTION

*"It's not a **science**, it's definitely an **art**."*

Kevin Cheveldayoff, Winnipeg Jets GM

*"The hardest part is the **judging** truly doesn't come for **years** to come. The big thing about the draft is it's all **based** on **projections**. It's the **ability** to **project** and show some **blind faith** on players that show **some** **attributes** that you think can be developed. "*

What is the NHL Draft?

What players are eligible?

Why is it important?

Why scout with data and AI?

PROBLEM BREAKDOWN

How do the CHL and NHL relate?

15 16 17 18 19 20 21 22 23 24 25

Age

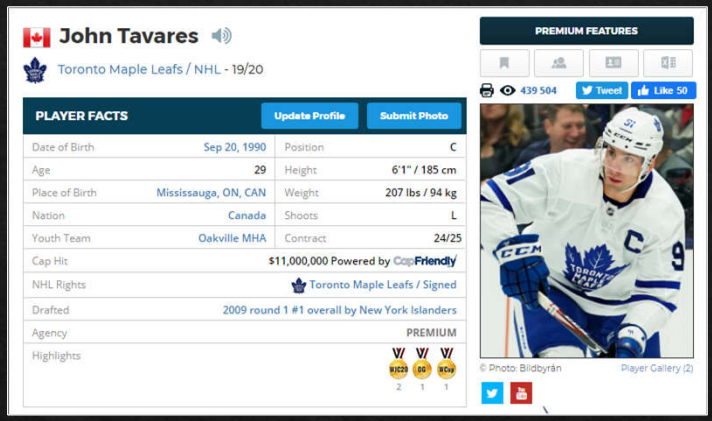
What data is available?

To approximate a function for projecting player potential to some point in future time, enough of the "right" data must be collected.

What data can be used?

How should the data be organized?

Meta



John Tavares

Toronto Maple Leafs / NHL - 19/20

PREMIUM FEATURES

439 504 Tweet Like 50

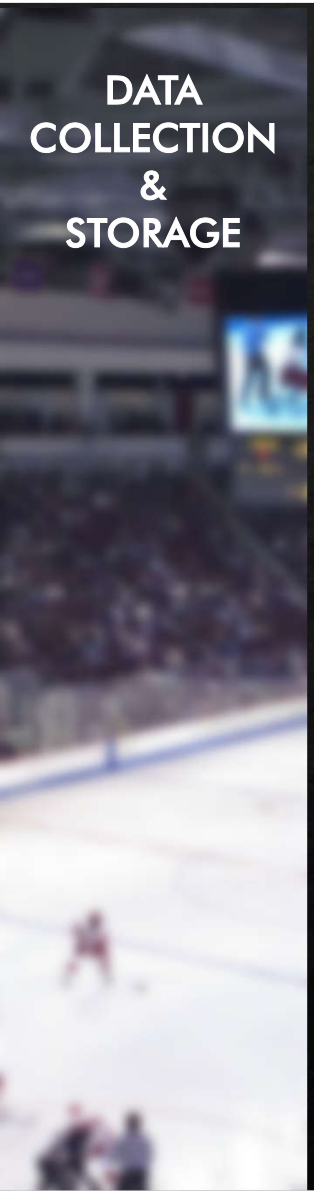
PLAYER FACTS	Update Profile	Submit Photo		
Date of Birth	Sep 20, 1990	Position	C
Age	29	Height	6'1" / 185 cm
Place of Birth	Mississauga, ON, CAN	Weight	207 lbs / 94 kg
Nation	Canada	Shoots	L
Youth Team	Oakville MHA	Contract	24/25
Cap Hit	$11,000,000 Powered by CapFriendly		
NHL Rights	Toronto Maple Leafs / Signed		
Drafted	2009 round 1 #1 overall by New York Islanders		
Agency	PREMIUM		
Highlights	WJC20 2 · WC 1 · WCup 1		

© Photo: Bildbyrån Player Gallery (2)

Player Statistics

PLAYER STATISTICS

S	TEAM	LEAGUE	GP	G	A	TP	PIM	+/-
2003-04	Toronto Marlboros Bantam AAA	GTBHL	90	95	92	187	76	
2004-05	Toronto Marlboros Mn Mdgt AAA "A"	GTMMHL	72	91	67	158	62	
	Milton Icehawks	OPJHL	20	13	15	28	10	
2005-06	Oshawa Generals	OHL	65	45	32	77	72	-13
	Canada Ontario U17	WHC-17	5	6	8	13	8	
	Canada U18	WJC-18	7	2	3	5	4	4
2006-07	Oshawa Generals	OHL	67	72	62	134	60	25
2007-08	Oshawa Generals "A"	OHL	59	40	78	118	69	22
	Canada U20	WJC-20	7	4	1	5	2	0
2008-09	Oshawa Generals "C"	OHL	32	26	28	54	32	5
	London Knights	OHL	24	32	18	50	22	5
	Canada U20 "A"	WJC-20	6	8	7	15	0	7
2009-10	New York Islanders	NHL	82	24	30	54	22	-15



DATA COLLECTION & STORAGE

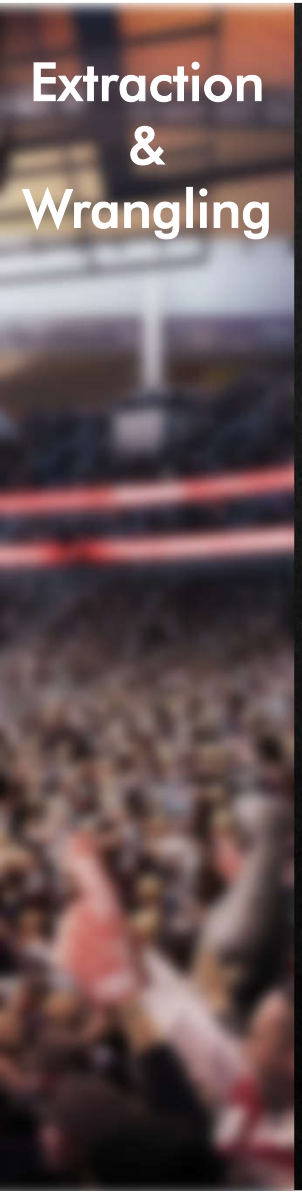
Web Scraping

Beautiful Soup

elite hockey prospects

Player Profiles URL's who played in the CHL from 2005-06 to 2017-2018

6054 Player Profiles Pages



Extraction & Wrangling

Pull from MongoDB

PyMongo

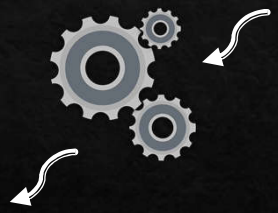
Pandas

Pandas

By Player Season

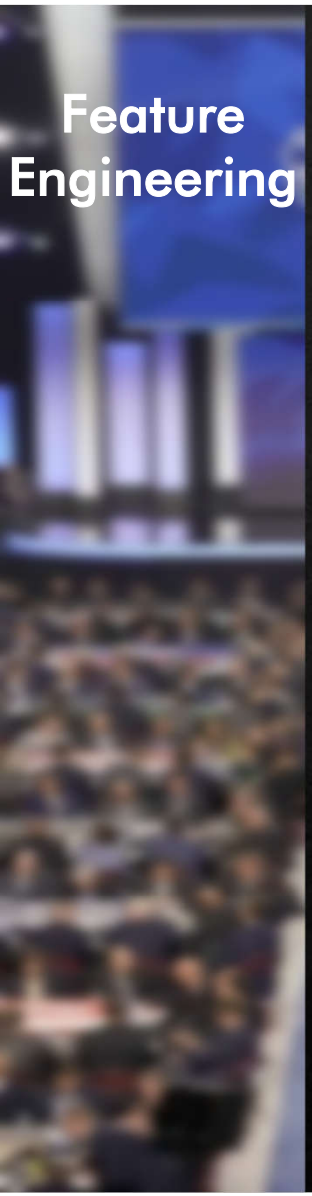
ep_id	games_played	goals	assists	plus_minus	penalty_min	league	team	age
11101	72	16	11	3	89	WHL	Seattle Thunderbirds	18
11101	26	8	8	0	14	WHL	Tri-City Americans	19
40288	0	0	0	0	0	NBPEIMHL	Cornwall Thunder Midget AAA	16
40288	1	0	0	0	0	MJAHL	Charlottetown Abbies	17
40288	61	6	4	-11	26	QMJHL	Prince Edward Island Rocket	18
40288	30	13	15	0	46	MJAHL	Charlottetown Abbies	19
40288	17	0	3	-1	4	QMJHL	Prince Edward Island Rocket	19
40288	41	20	31	0	52	MJAHL	Charlottetown Abbies	20
40288	29	19	20	0	64	MJAHL	Summerside Western Capitals	21
106101	33	7	17	0	50	NSMMHL	Valley Wildcats	16

Reshape



ep_id	games_played_18	goals_18	assists_18	penalty_min_18	plus_minus_18	team_18	league_18
11101	72	16	11	89	3	Seattle Thunderbirds	WHL
40288	61	6	4	26	-11	Prince Edward Island Rocket	QMJHL
106101	58	13	28	28	-2	Bridgewater Lumberjacks	MJAHL

One Row per Player



Feature Engineering

Features

Response $y = \text{league\_25}$ → Refactored → NHL | Lower | Not Active

Variables $X_i = \{ \text{age\_14} \rightarrow \text{age\_17} \}$ → Continuous

Sentiment Analysis – Scout Text

"Tavares has excellent offensive instincts and has a fantastic ability to find open space in his team's offensive zone. His shot is already on a high level in NHL standards and his stickhandling is phenomenal. Tavares isn't the best skater but he is always explosive after entering the opposing team's zone."

→ NLTK

Natural Language Processing (NLP) Variables

↓

subset

↓

647 rows x 33 columns

Model Training

Python Packages Used

TensorFlow

Keras

scikit learn

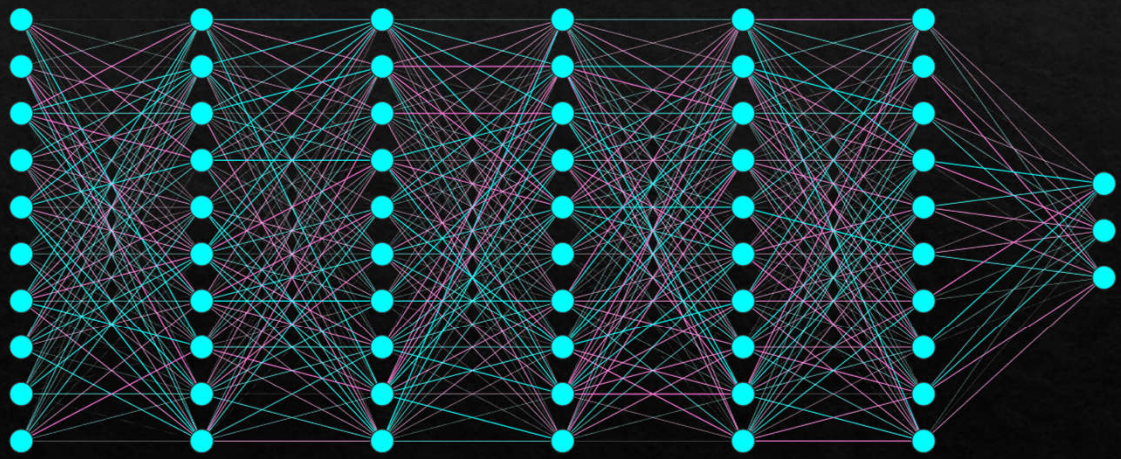
FCFF-NN Architecture

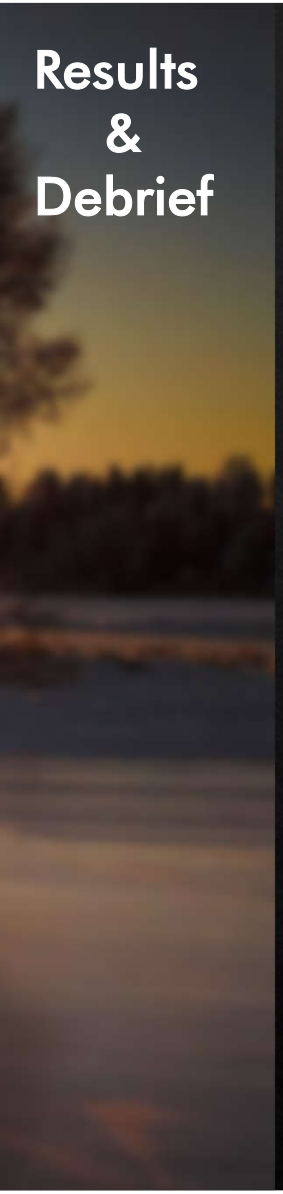
Input layer: 100 units

5 Hidden layers: 1000 units

Output layer: 3 units

70 Epochs





Results & Debrief

Results

	Precision	Recall	F1-Score	Support
Lower	0.47	0.53	0.49	48
NHL	0.3	0.22	0.26	0.76
Not Active	0.75	0.77	0.76	95
Micro Avg	0.62	0.62	0.62	162
Macro Avg	0.51	0.51	0.5	162
Weighted	0.61	0.62	0.61	162

Debrief

- Cleaner NLP (Names, Misspelling, Hockey Intuition)
- Use Scrapy and not BeautifulSoup
- Scrape other sites to fill in blank data (hopefully)
- Find more scouting blurbs

Future Considerations

- Data completeness
- Psychological data (Family history, Education, Personality Profiles, etc.)
  - There is quite a bit of noise in the model, some of which could conceivably be explained with quantified mental/psychological data or some useful proxy.