**Predicting Hockey Player’s Future Performance Based on Previous Games Played by Using Linear Logistic Analysis**

by

Hongyu(Nelson) Dai

Khaled Gabc

Wanjin Yoo

Peter Chepelev

University of Victoria

March 2017

UNIVERSITY OF VICTORIA

ABSTRACT

Predicting Hockey Player’s Future Performance

Based on Previous Games Played by Using Linear Regression Data Mining of Performance of Hockey Players

by Hongyu(Nelson) Dai

Khaled Gabc

Wanjin Yoo

Peter Chepelev

Instructor: Alex Thomo | Department of Computer Science

Because of the Vancouver Stanley Cup riot, 2011 was a sad year for most Vancouverites, especially for the die-hard Canuck fans. If the Vancouver Canucks had figured out which players they should have acquired and which position to play those players, they would have had a better chance to win over the Boston Bruins in game seven and the sad story would not have happened. With the massive changes in technology of data mining, almost everything can be predicted by analyzing all the data we have. By tracking players’ data, we can easily conclude how they are doing. Coaching staff, fans as well as sports analysts are trying to take advantage of the abundance of data to satisfy their different analytical needs.

This report analyzes each player’s statistics and predict their future performance, for this task the most suitable and fundamental mechanism should be linear regression which is also used in many professional and experimental applications. Data were scraped from eliteprospects.com and stored in a json file. The predictions of the performance of the players we have tested on latest games based on the model has shown to be satisfactory.

**TABLE OF CONTENTS**

[**§ Introduction**](#_l9dkbs8f4vme) **2**

[**§ Data Collection**](#_v7zd2u3eokj2) **3**

[**§ Data Cleaning and preprocessing**](#_u1nlbph9kltw) **4**

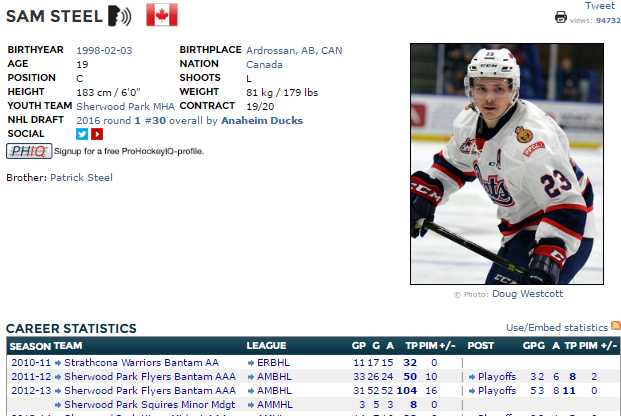
[**§ Data Mining**](#_8c0r595liyrk) **5**

[**Website Prediction**](#_n3uw9z9qyviy) **7**

[**§ Discussion of Results**](#_5zgmmx7femri) **8**

# § Introduction

Forecasting results of sporting events and the performance of the athletes represents a natural machine learning applications with massive scale. Professional sports contain a lot of convenient data and the player's performance is often the result of the random distribution, and provide the data to predict the appeal. Data holders may wish to influence the outcome of calculation, because they can directly benefit from the certain result, or due to privacy issues and hide their input. In this case, it is necessary to model a functional utility, and design an appropriate incentive mechanism for personal use. Hockey is particularly well suited for this kind of applications because the performance of hockey players in different positions can be measured using the same statistics. In addition, the application of machine learning algorithms to predict the performance of individual athletes also has a natural extension of the predicted game results. By predicting the performance of each player's score and adding them, it is possible to determine whether the individual's performance can use an accurate loss classifier. In this project, we started using a linear regression (compared with the NaïveBayes and SVM implementations) to develop a model to predict how would the hockey players do on the opponent in the future games.



**Example of scraped web page**

# § Data Collection

We decided to scrape information from eliteprospects, a website that contains a database of thousands of hockey players and various statistics such as games played, points scored, height, date of birth and many others.

Firstly we need to create a list of all players that played between 1995 and 2016, and where there position is forward. Using a visual crawler tool called import.io we are able to extract urls in this table and press button ‘Next’ to extract players from other pages too. The result of this is that we were able to generate a Excel file named [ playersUrls.csv ] that contains player;s names, and their page’s URL.

Secondly using Scrapy, a python library we needed to scrape information from individual pages of players such as score made, games played, height, and DOB. Usings the list generated in step 1, and XPATH code to accurately choose what parts we need to extract, we were able to make a JSON file that contains all this information.

A code snippet that contains the logic behind extraction from each player’s page is below:

def parse(self, response):

yield {

'name': response.xpath("//span[@id = 'fontHeader']/text()").extract\_first(),

'birth':response.xpath("//table[2]//table//tr[1]//td[2]//a/text()").extract\_first(),

'height':response.xpath("//table[2]//table//tr[4]//td[2]/text()").re(r'^\d\S\*')[0],

'season': response.xpath("//table[@class='tableborder'][1]//td[@style][1]//text()").re("\d{4}-\d{2}|^\xa0$"),

'Leage':response.xpath("//td[3][@style][not(@id)][not(@class)]//font[1][not(@id)][not(@style)]/text()").extract(),

'GP':response.xpath("//td[4][@style][not(@id)][not(@class)]//font[1][not(@id)][not(@style)]/text()").extract(),

'TP':response.xpath("//td[7][@style][not(@id)][not(@class)]//font[1][not(@id)][not(@style)]/text()").extract()

}

A single JSON entry of a player (http://www.eliteprospects.com/player.php?player=195952) in the JSON file looks like this:

{

"Leage": [ "PCBHL", " Midget", " WHL", " WHC-17", " WHL", " WHL", " WHL", " John Reid Bantam", " U16 Cup", " Jr Super Series" ],

"name": "TY RONNING",

"GP": ["72", "12", "56", "6", "24", "67", "66", "6", "-", "2"],

"season": ["2011-12", "2012-13", "2013-14", "\u00a0", "2014-15", "2015-16", "2016-17", "2011-12", "\u00a0", "2016-17"],

"TP": ["153", "25", "20", "4", "2", "59", "53", "11", "-", "1"],

"height": "175",

"birth": "1997-10-20"

}

# § Data Cleaning and preprocessing

**First Step**

To extract useable data from the JSON file that was created from the web scraping process, it was loaded into a python script. The JSON file had many unusable data in it, so these had to be removed. For example, there were some seasons that the player object did not play in, and they were removed. There were blank spaces in the season field in the data due to the notation of the website, where any blank spaces in seasons field refer to the season nearest above it that has a value in the season field. Therefore, all blank spaces had to be replaced with a value. Furthermore, the website sometimes uses dashes when describing a value of zero for games played or points. These had to be replaced with the number zero. After this basic cleaning, the data had to be put into a usable format. For each player, target seasons where the player’s age at the start of the season ranged from 16 to 20 were set. Then, these target seasons were used to merge all data for each season together. The player’s NHL performance was then appended to the player record. Finally, the list of player records was output to a CSV file. See figure below for an example of output.

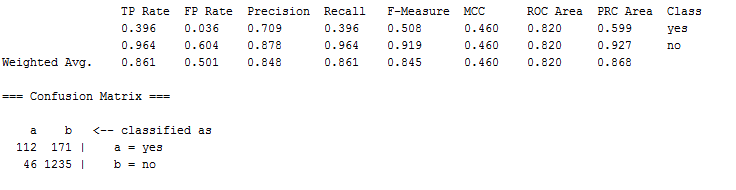


**Improvement**

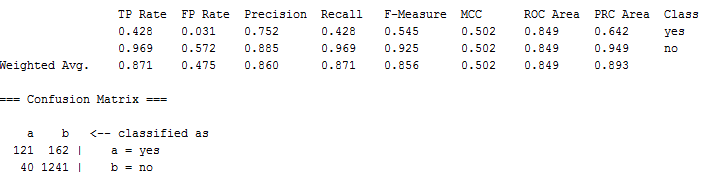
A derived attribute, improvement was used to describe whether the player is improving during their junior career. A file was created, ImprovementGenerator.py that opens a CSV file containing lists of performance by year for each player. We go through this list and remove the seasons that they didn’t play and then create a linear regression line from the remaining values. This program outputs the slope of the line, this refers to the improvement of a player during their junior career.

# § Data Mining

After experimentation, using the average of all of the seasons in which a player played was chosen to describe their performance. (see figures below)

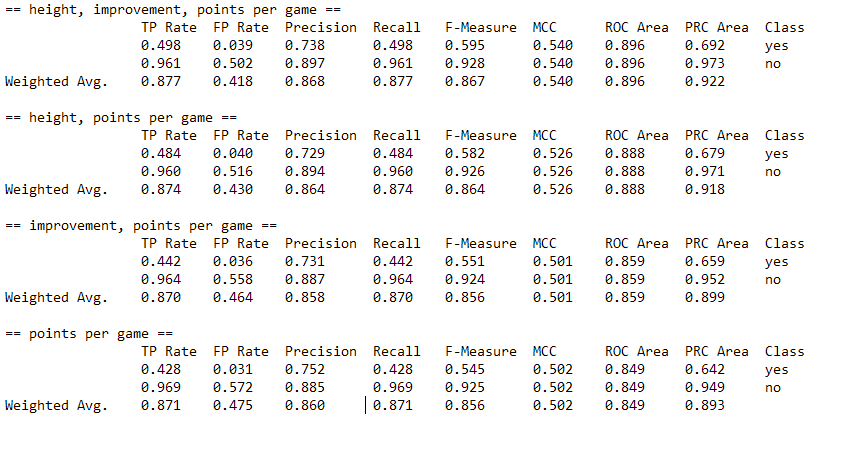


**Accuracy of multiple variable data mining**

****

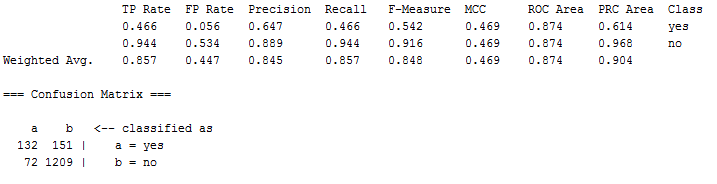
**Accuracy of single variable data mining**

Whether or not the player played at least one game in the NHL is the class attribute. It was shown through further experimentation that also including height and improvement gave better results in the classification. (see figure below for statistics)

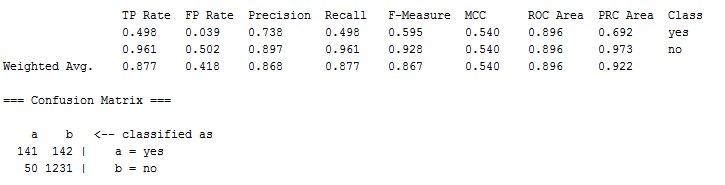


**comparison of different attributes included**

The data was put into WEKA, and naive bayes and logistic regression algorithms were tested.(see figures below)



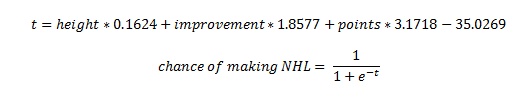
**Accuracy of Naive Bayes classifier**



**Accuracy of Logistic regression classifier**

Logistic regression was chosen as the better algorithm and the learned equation for deciding whether the player will make it into the NHL is shown in figure below.

**equation for finding chance of making NHL**



# § Website Prediction

Using the formula generated from the Data Mining process, and using player’s statistics generated by the by CSVtoJSON.py, we built a simple website that you can use to predict if a player will is likely to be successful in NHL or not. The website is simply taking a player name, and it substitute player’s attributes in the formula to generate the prediction.

Refer to index.html in ‘website’ folder to view the website, and look at player\_example.txt to get sample player names.

**CSVtoJSON.py**

Takes a CSV file that contains name, height, improvement, and score of each player, and It converts it JSON format so that It can be used by the website.

A sample player data looks like the following:

"BRETT MILLER": {  
 "points": "0.211413034",  
 "height": "191",  
 "improvement": "-0.11468314"  
 },

Code snippet to show how javascript computes the percentage.

function calc() {  
var selectedName = document.getElementById("nameInput").value.toUpperCase();

var res = (players[selectedName].height \* 0.1624) +

(players[selectedName].improvement \* 1.8577) +

(players[selectedName].points \* 3.1718 ) - 35.0269 ;

var percentage = 1 / ( 1 + Math.pow ( 2.718281828, -res ) );  
  
document.getElementById("result").innerHTML= Math.round(percentage\*100) +'%';  
}

# § Discussion of Results

0ur algorithm is designed to predict players’ performance in NHL based on their previous games(our training data) . We normalized our data from 0 to 100 scale because it is easier to interpret.If a player scored high, it means the player is more likely to be successful in the future games. In this way we can find those players who has smaller central scale but better overall statistical in performance. On the other hand, we can also find players who are receiving too

much playing time because of his luck in one game.

Overall, the logistic regression model is able to predict player’s future performance within 87% of accuracy. This accuracy could be improved in future studies by taking into account the strength of the team that the player is on and taking into account how large a role the player played in the team’s offensive production. Including shots and plus/minus may also improve the analysis. The running time of the algorithm was insignificant, as the size of the dataset is not very large.

**REFERENCES**

https://en.wikipedia.org/wiki/2011\_Vancouver\_Stanley\_Cup\_riot(Vancouver Stanley Cup riot news)