Final Capstone Report

Justin Anto
June 25 2023

Introduction

My task with this project is to find who the best two way forwards are in the NHL and define what makes them effective through statistical analysis and machine learning.

There are 3 position categories in ice hockey; forwards, defense and goaltenders. Forwards are tasked with producing most of the offense for their team; goals, assists on goals etc. Defensemen are tasked with stopping opposition players from attempting to score and goaltenders tend the goal area. They use specialized equipment to block any shots of the puck from the other team's players.

This project will focus on the forwards and a type of forward called "two way" forwards. These are players that excel at producing offense for their team and stifling their opposition on defense. This is an incredibly valuable type of player to any team. Having a player that can do both at an elite level is undervalued, because it's hard to define what exactly makes each one of them great.

Being able to identify who these players are with statistics and machine learning models will give ice hockey teams an edge in finding and bringing these players to their teams. Additionally, being able to define why a two way player is effective means you can bring in other players with complementary skills. This will aid in shoring up any weaknesses that a team might have.

Additionally the NHL has implemented what is called a "salary cap". This means that every team has a maximum amount of dollars they can spend on player contracts per year. Being able to identify elite two way forwards who may go unnoticed gives teams an edge in getting these players signed to their roster- at a bargain. Getting more effective players for less money is crucial for success in today's NHL.

Background on the subject matter

The idea of an elite two way forward isn't new. The NHL awards a trophy for this every year. It's called the Frank J Selke trophy for the best defensive forward. Why not just look at this list? This list isn't exhaustive and it's voted on by journalists around North America. They can't possibly watch every player in every game. So, we'll turn to statistics and player tracking for help.

Statistics and the tracking of them isn't new in hockey. The advent of what are called micro-statistics and advanced statistics *is* new. Micro-statistics are individual events that happen during the game. It could be; a player skating with the puck or another player disrupting a pass of the puck from two other players. Ice hockey is an incredibly fast sport to play and watch. These events can go unnoticed, but there is a growing ability to track them. These micro-statistics let us peer into the game at a granular level and discover new levels

of detail. There has also been an explosion of what are called advanced statistics; aggregations of micro-statistics that give us a higher level view of player behavior on the ice.

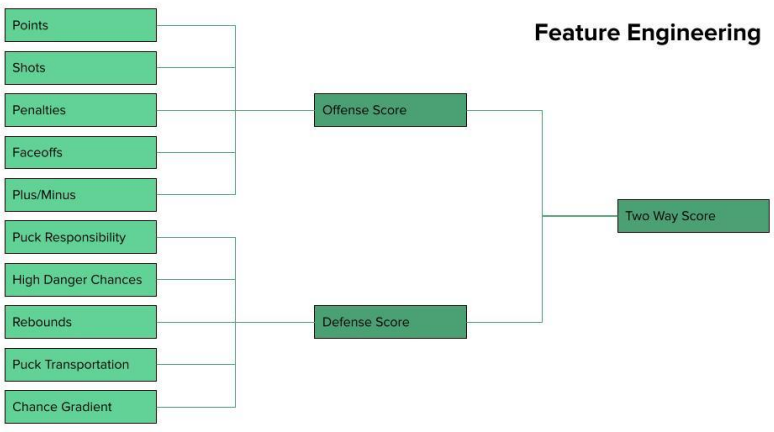
Summary Of Cleaning And Preprocessing

To accomplish this task, I've downloaded datasets from a website called "Money Puck". They are a web based service that provides data, visuals and analysis for journalists and individuals. Their datasets include statistics, advanced statistics and micro-statistics. Ranging from 2008 - 2022, they have made all of their datasets available and free to download in CSV format.

Since this data is used by MoneyPuck (I used the "skaters" datasets) to produce all of their analysis and visuals, it was pre-cleaned. Other than a summary EDA, there wasn't any need for cleaning. There was a lot of pre-processing to do in the form of feature engineering and dimensionality reduction. Each dataset has around 4500 rows, and 150 columns. However, not all of these columns and rows are necessary or useful for defining two way players and their performance. In researching for this project, I decided that the best course of action was to create my own advanced statistics from the statistics and micro-statistics provided in the dataset. This would let me measure an individual's offensive and defensive performance separately, and then let me aggregate all of that into a catch-all metric that I eventually called "two way score".

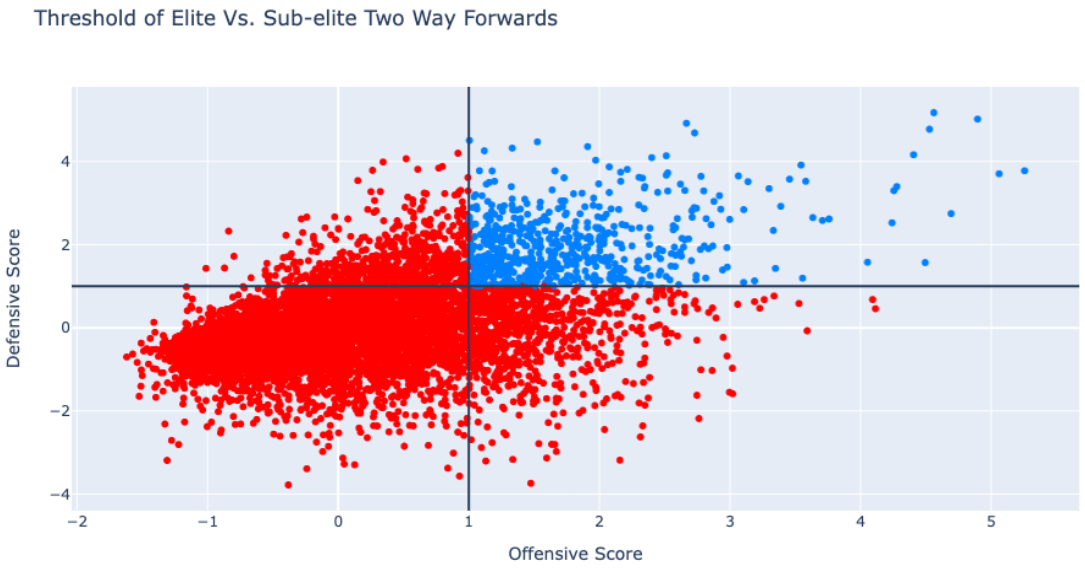
To create the offensive advanced statistic ("offense score") I drew influence from an ice hockey journalist named Dom Luszczyszyn. He works for a sports news outlet called The Athletic. He borrowed from basketball and baseball to create a metric to evaluate a player's performance in a single game called "GameScore". This featured many of the ideas I had about what my offensive score should look like. With some adapting, I was able to come up with a grouping of engineered features. These features measured and scored a player's offensive performance over an entire season's worth of data. An article written by Dom Luszczyszyn about GameScore is available here.

For the defensive metric, there was less information out there. So, with some research and relying on my own knowledge of hockey I was able to produce a metric ("Defensive Score") that measured a player's defensive impact over a season's worth of hockey. A flow chart of my feature engineering looks like this:



This wasn't the solution I imagined it to be. In further looking into the data, I found that a player might have a high offense score, low defense score and come out with a great two way score. This isn't the type of player I'm looking for. To solve this, a threshold for each of offensive and defensive scores was created. The idea being that an elite two way forward is someone who exels at both offense and defense.

This threshold produces a visual like this:



This produced exactly the effect I wanted and provided binary value of whether a forward was considered an elite two way player or not.

Insights, Modelling and Results

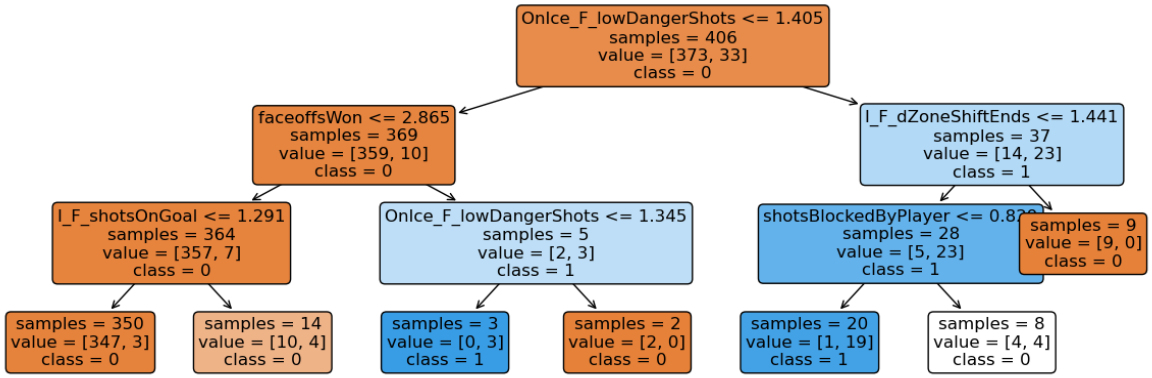
Following this, I moved on to machine learning models. I began with Logistic Regression to divine which of these features and statistics were most indicative of two way excellence. My assumption being that goals, assists and the quantity of what are called "high danger chances" either for or against the player would be most indicative (click here for an article about high danger chances) of two way performance.

The model found something quite different. It determined over several iterations that the quantity of "low quality chances" for and against a player, as well as faceoffs won or lost were most indicative of two way performance (an article on faceoffs can be found here). Current thoughts in ice hockey circles are that high danger chances are more important than low danger chances. By looking at the percentages, you're more likely to score on a high danger chance than a low one. I believe this model keyed in on the *quantity* of shots. The saying goes "quantity has a quality all its own", and this - at least in terms of shots in hockey - seems to confirm that. The faceoffs conversation is a healthy-ish debate in hockey circles. I believe they're very important to a player's success, I just didn't think they were the second most important thing.

I reduced the dimensionality of my dataset by dropping any stats that the model wasn't sure were indicative of success. Following this step, I ran a KNN machine learning model to see if

my player classification threshold could be replicated. The model performed well, with high accuracy. The accuracy was suspiciously high. Usually KNN models require some tuning and optimization, but this gave me top tier accuracy right away. Since the features being evaluated were created from each other (using engineered features to create engineered features), this was expected. I used the KNN model as a confirmation, in concert with the winners and nominees for the Frank J Selke award. This confirmation that I had things right was great to have.

My final model was what's called a Decision Tree. This model can show how it goes about splitting the players in the dataset into the elite and sub elite threshold I set.



While it can look a bit cluttered, it clearly shows that the low danger shots and faceoffs were some of the first splits made, showing what was happening in my logistic regression models was accurate and replicable!

Findings and Conclusions

As a summary, I acquired data, engineered features and ran machine learning models to identify two way players and what about them makes them so effective. I was ecstatic to be able to identify the elite two way forwards of the league. Especially those who have not won or been nominated for the Frank J Selke award. These players are elite, but don't get the recognition they deserve. This is an opportunity for a team to scoop up a great player at a bargain. My machine learning models helped illuminate my incorrect assumptions about what exatly determines whether a two way forward is elite or not.

In the future, I would actually like to add more data from different sources. I think some of my engineered features don't discern individual impact as well as I would like. This means adding more micro-statistics. I'd also like to revisit this project as I grow my skills. I've learned a lot by completing this project and I look forward to learning more!