The Evolution of Scoring in the NHL

*Note: Sub-titles are not captured in Xplore and should not be used

1st Friedfertig
Department of Computer Science
College of Charleston
Charleston, United States
friedfertigc@g.cofc.edu

2nd Charlie Fuller
Department of Computer Science
College of Charleston
Charleston, United States
fullercf@g.cofc.edu

***Abstract*—This document is a final project for our Data Mining (CSCI 334) class. We compared statistics of the top 10 National Hockey League (NHL) point scorers in the following seasons: 1998-1999 (the first season that all the data we need was all collected), 2008-2009, and 2018-2019. We are going to use these statistics to find patterns in based on the year and develop a thesis regarding the evolution of scoring. We also want to determine what variables, for example new technology, have influenced these statistics over the years. Our data is scraped directly from quanthockey.com and is used in order to test and train our model.**

***Index Terms*—Data Mining, Machine Learning, Classification, Web Scraping, Sports Analytics, Decision Making, Hockey**

I. Introduction

There aren't many better things to celebrate than a goal by one of your favorite sports teams. Whether it is a beautiful individual effort or an unbelievable team play, the excitement is exhilarating. While that feeling is different in all sports, to us, there is nothing more excited than a goal in hockey.

Gordie Howe has 801 career goals and retired in 1970-1971, Wayne Gretzky has 894 career goals and retired in 1999, Alexander Ovechkin has 780 goals and he is yet to be slowed down! Times have changed and different skills are necessary to put the puck in the net, but how have variables that go into scoring evolved over the last thirty years? That is what we strive to find out.

II. Background

Going into the project, we as a group decided we wanted to tackle a topic that was interesting to both of us. As a partnership, we immediately recognized that we both have a heavy interest in sports, specifically sports analytics. One of us being a member of the College of Charleston Men's Hockey team, it was pretty easy to put together an idea. Originally, we wanted to look at the data from one year but had issues determining a goal in mind. However, with the help of our professor, Navid Hashemi, he helped us originate the idea of comparing data from 3 years, ten years apart each, over a 30 year span to see patterns in variables that contributed to scoring. With that, we were ready to get started!

III. Context

Hockey is one of the most complicated sports of all time. While we both carry great amounts of enthusiasm, we understand that there is a large group of people who cannot even fathom the complexity of the sport. There are a couple concepts and definitions that are useful prior to reading further in our project:

- Points: Total Number of Goals plus the Total Number of Assists
- Plus/Minus: A player's plus/minus is determined by taking the number of goals that he is on the ice for his team and subtracting the number of goals that he is on the ice for the other team.
- Time On Ice (TOI): The average amount of time a player is on the ice during a game.

Overall, there has been a major shift in the evolution of scoring over the last thirty years and we couldn't help but want to analyze it.

A. 1998-1999

In a day where hockey was known strictly for violence and absolute wild behavior, 1998-1999 was a great season for hockey filled with a lot of excitement. A main focus on how the data will be received is the addition of technological advancement in the sport. 20 years ago hockey was a very different sport because of how much this technology has changed. Specifically back then goalies were equip with smaller and yet clunkier pads meaning more of the net would be left vulnerable and the goalie would be slower at getting to a fast moving puck.

B. 2008-2009

Goal tending technology in the early 2000s had a significant increase which lead to players having to adapt to a more well guarded net which would explain the decrease in overall goals during this time period.

C. 2018-2019

After the rapid advancement in goalie technology began to slow and players were able to adapt to the new style of play to match with the newer technology they were able to pick up where they left off which can be shown in the data through more total goal scores.

IV. Previous Works

We looked towards Kaggle to see if there were any similar previous projects that utilized a similar idea. There are many different projects on the NHL and its goalscorers; however, we did not come across one that was necessarily similar to this.

V. Data Collection/Data Sets

We collected data online from quanthockey.com, a database tool that allows access to NHL statistics in every way imaginable. They have stats from 19 distinct leagues located all over the globe ranging from the NHL to the American Hockey League (AHL) even data on The Olympics. The database includes Hockey data that has been recorded for decades and formatted here on their website. This website contains all the data we need from the years 1998-present. We plan to use all statistics. In doing so, we can determine what scorers were more efficient in each year and try to differentiate each statistic for each year. Below is the link to the data that we plan to use for our research. QuantHockey's database consisted of all sorts of statistics that were useful in our project.

- Rank
- Name
- Team
- Age
- Position
- Games Played
- Goals
- Assists
- Points
- Plus/Minus
- TOI (in minutes)
- Shots

VI. Data (Pre)Processing

We extracted the top 50 players from each selected season (1998-99), (2008-09), (2018-19). After that we stripped the bottom 40 players and were left with three tables of the top 10 players from each respective season. We had to remove about ten variables which we did not feel were relevant to the evolution of scoring. We did not think that the machine learning models would be accurate if it included these variables.

VII. Determining Important Attributes

A. Age

B. Goals and Assists

We wanted to see if players were getting more points from goals or assists throughout time.

1) Shots: We wanted to see if new goaltender technology could possibly make total shots less relevant for scorers.

VIII. ML Models

We had considered a couple different ML Models to use throughout this project. We did our work in RapidMiner Studio and initially used a Deep Learning Model, a Random Forest, and a Decision Tree; however, we ended up using the Deep Learning Model as we thought it was trained to the highest level. We trained a deep learning machine model to predict the number of points based on these variables:

- Age
- GP
- Goals
- Assists
- Points
- Plus/Minus
- TOI (in minutes)
- Shots

We deemed these variables to be the most important to consider throughout our project and found our training results to be quite successful. Our RapidMiner was set up simply:

1) Steps: First, we retrieved the data for each year.

Second, we used the "Set Role" function and set points as a label.

Third, we used the "Split Data" function and set enumerations at 0.7 and 0.3 to split the training model from the real model.

Fourth, we connected the "Split Data" to both a "Deep Learning" operator and an "Apply Method" operator sending the higher percentage through a deep learning model.

Fifth, we added a performance operator testing for root mean squared error (RMSE) and absolute Error.

A. 1998-1999 Machine Learning Training

- MSE: 4.0028253
- RMSE: 2.0007062
- R-squared: 0.8636033
- Mean Residual Deviance: 4.0028253
- Mean Absolute Error: 1.5830487
- Root Mean Squared Log Error: 0.019582905
  Scoring History (Training Session 1 — Training Session 2)
- Duration: 0.067 — 0.089 (seconds)
- Training Speed: 1166 — 2692 (obs/sec)
- Iterations: 1 — 10
- Samples: 7 — 70
- Training RMSE: 8.36660 — 2.00071

As shown above for the (1998-99) data it can be seen that for both training sets the duration was very low along with the training speeds. The first iteration ran once and the second ran 10 times with respective samples of 7 and 70 percent. It also shows that the RMSE and other errors were substantially low.

B. 2008-2009 Machine Learning Training

- MSE: 15.965915
- RMSE: 3.995737
- R-squared: 0.71195513

- Mean Residual Deviance: 15.965915
- Mean Absolute Error: 3.3836544
- Root Mean Squared Log Error: 0.039516453
  Scoring History (Training Session 1 — Training Session 2)
- Duration: 0.142 — 0.156 (seconds)
- Training Speed: 1000 — 3500 (obs/sec)
- Iterations: 1 — 10
- Samples: 7 — 70
- Training RMSE: 11.28452 — 3.99574

As shown above for the (2008-09) data it can be seen that for both training sets the duration was very low along with the training speeds. The first iteration ran once and the second ran 10 times with respective samples of 7 and 70 percent. It also shows that the RMSE and other errors were substantially low.

C. 2018-2019 Machine Learning Training

- MSE: 2.1733336
- RMSE: 1.474223
- R-squared: 0.94952923
- Mean Residual Deviance: 2.1733336
- Mean Absolute Error: 0.7965133
- Root Mean Squared Log Error: 0.012882786
  Scoring History (Training Session 1 — Training Session 2)
- Duration: 0.041 — 0.054 (seconds)
- Training Speed: 2333 — 4666 (obs/sec)
- Iterations: 1 — 10
- Samples: 7 — 70
- Training RMSE: 7.21448 — 1.47422

As shown above for the (2018-19) data it can be seen that for both training sets the duration was very low along with the training speeds. The first iteration ran once and the second ran 10 times with respective samples of 7 and 70 percent. It also shows that the RMSE and other errors were substantially low.

IX. Deep Learning Points Prediction

We trained our model to predict the total number of points based on Average Time on Ice/Game, Total Shots, and Plus/Minus.

A. 1998-1999 Predicted Total

Our totals were very accurate and predicted high scoring and high amounts of shots confirming our correlations.

B. 2008-2009 Predicted Total

Our totals predicted low scoring based on low shots and overall bad job keeping up with important variables.

C. 2018-2019 Predicted Total

Similar to 1998-1999, the model predicted correctly as it predicted high scoring/shots and overall a lot of points.

D. Overall Results

We feel as though the machine model was trained successfully as these numbers for the most part are sensible and accurate.

X. Other Approaches

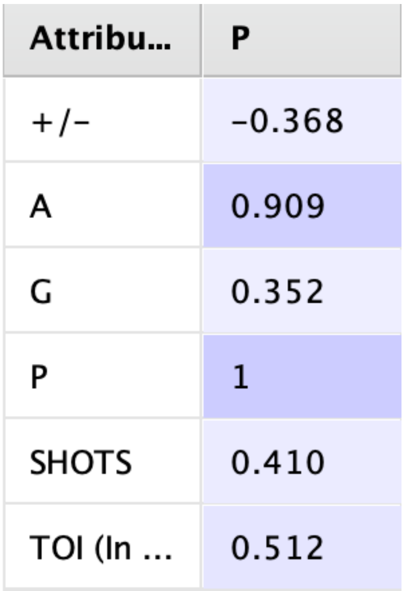
While our Machine Learning Model was very successful in its training, we also wanted to use other ways to manipulate and analyze the data. We completed K-means Clustering in RapidMiner for each year.

A. Correlation Through Rapid Miner

We used an auto-model method in RapidMiner in order to determine the correlation of points to the top 10 from each year to the following variables:

- Goals
- Assists
- Points
- Plus/Minus
- TOI (in minutes)
- Shots

We found the following clusters in relation to total points.



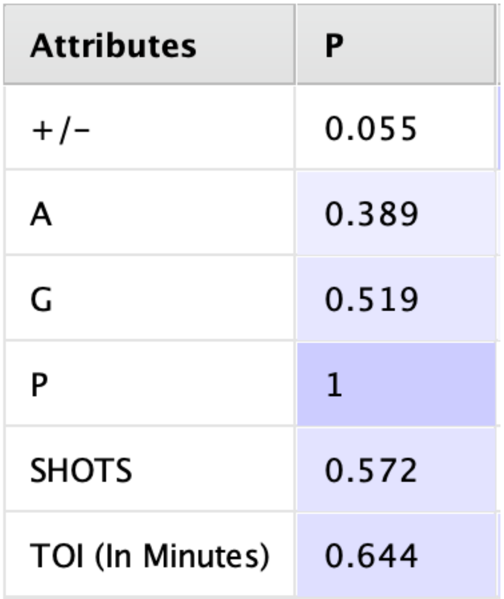
Attribu...	P
+/-	-0.368
A	0.909
G	0.352
P	1
SHOTS	0.410
TOI (In ...	0.512

Fig. 1. ML Model Predictions based on 1998-1999 Data

B. Clustering Results

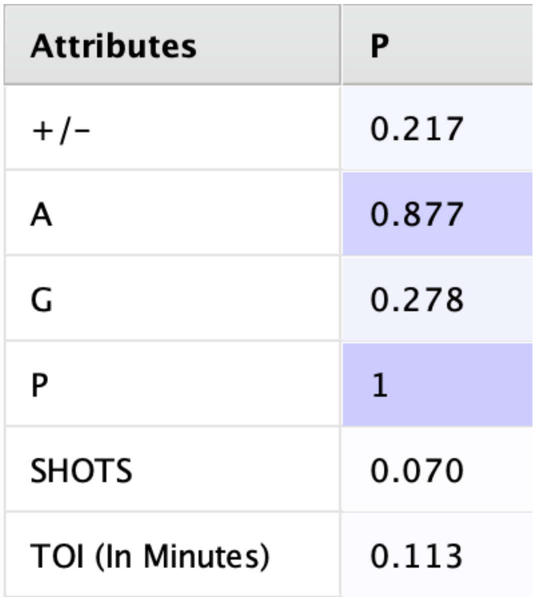
We were able to come up with a few conclusions based on our clusters.

1) Plus/Minus: In 1998-1999, plus/minus has a negative correlation with Points. We take this as a factor that goalie technology was overall much worse back then and there were a lot more high-scoring games. This being the case, no matter how many goals top goal scorers are out there for their team, the pace of scoring meant that they can be out for just as many goals against. Plus/Minus increased to a positive correlation by the time we got to 2018-2019. We assume this is because as they finally learned to score against this new goalie technology, that they also learned how to defend more against opponents.



Attributes	P
+/-	0.055
A	0.389
G	0.519
P	1
SHOTS	0.572
TOI (In Minutes)	0.644

Fig. 2. ML Model Predictions based on 2008-2009 Data



Attributes	P
+/-	0.217
A	0.877
G	0.278
P	1
SHOTS	0.070
TOI (In Minutes)	0.113

based on 2018-2019 Data

2) Goals and Assists: In 1998 and 2018, Assists were a dominant force in scoring. We see a slight decline in 2008-2009 which we see as a reason there wasn't as much scoring as we associate high amount of points with higher assists than goals. This is also proven as the year with the least amount of points has a higher correlation with goals versus assists.

3) Shots: Shots have consistently been a driving force for points in the NHL. However, in 2018-2019 we feel that scorers are scoring more on excellent passing and teamwork plays therefore bringing that total down.

XI. Discussions

We wanted to determine what made the star players stand out from the rest and see what (if any) variables gained or lost importance over two decades. We were extremely impressed with how accurate the deep learning machines were at predicting points. RapidMiner was a great software to use as it was fluid, easy to implement, and easy to interpret.

XII. Conclusion

NHL scoring has evolved in many ways; however the same statistics are consistent throughout for the most part. The top players have become less valuable to their team as a whole as it has evolved to more of a team game. Overall, we can conclude that NHL teams should focus on drafting play makers who can make other people around them better versus pure goal scorers, a common trend that has failed team's drafts over the last few years.

A. Authors and Affiliations

Chase Friedfertig, Charles Fuller

Acknowledgment

We'd like to thank quanthockey for providing us with the data. We'd also like to thank our teacher, Navid Hashemi for giving us the idea for the project as we had a lot of fun working on it. Lastly, we'd like to thank our Business Analytics professor, Olga Biedova for showing us RapidMiner and giving us a great way to use Machine Learning Models.

References

Check out our Project Repository on Github at https://github.com/cfriedfertig/The-Evolution-of-Scoring-in-the-NHLthe-evolution-of-scoring-in-the-nhl

References

[1] quanthockey.com, April 2022.