**Applied Data Science and Machine Learning in Sports Statistics**

Richard Sauve

Spring 2023

In Partial Fulfillment of

Data 4395-Senior Project

Department of Mathematics and Statistics

|  |  |
| --- | --- |
| Faculty Advisor: |  |
| Dr. Katherine Shoemaker: |  |
| Committee Member: |  |
| Dr. J. Patrick King: |  |
| Committee Member: |  |
| Dr. Gideon Gogovi: |  |
| Department Chair: |  |
| Dr. Ryan Pepper |  |

Table of Contents

Contents

[Abstract 3](#_Toc134527371)

[Data 3](#_Toc134527372)

[NHL 3](#_Toc134527373)

[Data Scrapping 4](#_Toc134527374)

[FiveThirtyEight 4](#_Toc134527375)

[Two Agents 4](#_Toc134527376)

[Sports Bettor 5](#_Toc134527377)

[Elo Systems 5](#_Toc134527378)

[History 5](#_Toc134527379)

[Sigmoid 6](#_Toc134527380)

[Margin of Victory 7](#_Toc134527381)

[Autocorrelation 8](#_Toc134527382)

[Shift Equation 9](#_Toc134527383)

[Marginal Process 9](#_Toc134527384)

[Optimizations 10](#_Toc134527385)

[ROC and AUC 13](#_Toc134527386)

[Final Comparison 14](#_Toc134527387)

[Team Analyst 16](#_Toc134527388)

[Model Choice 17](#_Toc134527389)

[Random Forest 18](#_Toc134527390)

[Feature Importance 19](#_Toc134527391)

[Hyper-Parameters 22](#_Toc134527392)

[Further Analysis 23](#_Toc134527393)

[Betting Application 23](#_Toc134527394)

[Further Team Analytics 23](#_Toc134527395)

[Conclusion 23](#_Toc134527396)

[References 25](#_Toc134527397)

[Data 25](#_Toc134527398)

[Resouces 25](#_Toc134527399)

# Abstract

Sports is a multi-billion-dollar revenue global industry with data ranging from historical aggregate data to modern day play possession times in seconds. This report aims to apply Data Science and Machine Learning tools to sports using the National Hockey League (NHL) data. The data is sourced from online resources but is all original NHL owned data.

The project is split into two hypothetical data agents within the sports industry. The first is the sports bettor whose goal is to raise margins or stack the odds as much in their favor. The second agent is the team analyst designed to use data to improve the team’s overall performance within the organization (think Moneyball).

The goal is to provide insight into real-world problem solving and to explore the limits of data science and machine learning.

# Data

Data is crucial to the central concept of this project. Throughout this process data integrity is maintained by limiting the amount of manipulation and curation of the collected data. Credible official sources are used with all data available on the connected GitHub repository.

## NHL

Data sources for this project include official National Hockey League data. Data collection by the NHL is but remains blocked behind the NHL website by limiting practical scraping of data. To get around this issue, the NHL licensed the availability of its data to online data repository. This project used hockey-reference for its ease of access using data scraping tools in python.

## Data Scrapping

Data scraping functions were developed in python using the libraries beautifulsoup and cloudscraper. The primary use of these functions was to extracts the timeline and results of games. This provides the ability for elo system to generate probabilities based off observed game outcomes. The second use is modified to extract all game statistics for a given team to be entered into the machine learning models.

## FiveThirtyEight

FiveThirtyEight maintains an open policy pertaining to data and methods they used to generate their probabilities. They advertise their elo system structure and parameters used to scale the elo values. Their dataset also contains historical probability generations and features dating back to the inaugural season in 1917. For the purposes of this project, only the data for the 2022-2023 season is used as a comparison for the final optimized elo system.

# Two Agents

Sports data is vast containing information ranging from historical metrics from players to modern day play time in seconds. Collection and availability of this data is at all-time highs, meaning that all agencies involved within sports possess this data. This means betting agencies and bettors, opposing sports teams, fans, and anybody who feels like looking at it. To better understand motivation and use of this data, sports statistics will be separated into two separator agents, the sports bettor and the team analyst.

The sports bettor contains any observer of sports in the realm of betting, this means agencies looking to set betting lines, news sites looking to report on the probability of the outcomes, and individuals looking to make money. The goal of this theoretical sports bettor is to maximize their gains. This is done by achieving small marginal gains over time to result in compounding profits. This data to all presents a significant problem as it becomes near impossible to achieve an advantage as all competitors possess the same data. This means other methods are explored, in particular elo systems. They are designed track live performances and give the best estimate of the probability of one of the teams winning. This ultimately will be shown to be the best method.

The second split is those within the sports organizations whose goal is trying to increase their performance during the season. This second agent is called the team analyst. The team analyst has strong potential for machine learning as many models are able to generate value from an analyst’s perspective. Due to the goal of the team analyst, model performance and reliability are not the most important traits. Interpretability is the most important factor as this agent is looking to seek out weaknesses and strengths within the teams and in the competition.

# Sports Bettor

Data science for the Sports Bettor provides numerous applications but is restricted in that the betting market involves generated probabilities and money-lines before the event happens. This means that machine learning techniques are not industry standard. Turning to alternative sources, such as FiveThirtyEight, elo systems are ideal as they provide probabilistic outputs and rely on statistical principles.

## Elo Systems

### History

To give insight, elo systems were created by Hungarian-American Physics Professor Arpad Elo (1903-1992), the namesake of the design. Elo was a chess master and the best player in Milwaukee during his time. Elo designed this system for chess, which is a zero-sum game between two competitors. These characteristics themselves show why this system is so prevalent today in competition. Elo believed that the performance of a team was based off its average performance over time along with the strength of the competitors. If this is translated to statistical terms, it becomes the mean performance in a given set.

### Sigmoid

A basic elo system generates probabilities based off differences between two competitor’s ratings or elo values. It uses the sigmoid function as a base due to its possible output range of 0-1. To explore this method, FiveThirtyEight is a news organization that utilizes statistical analysis in their reporting. FiveThirtyEight provides advertised values for their elo system.

EloDiff – Elo Value Difference: This given by taking the home team’s elo rating and subtracting the away team’s

HTA - Home Team Advantage: FiveThirtyEight adds this metric to break ties and account for momentum when playing in the home arena. The recommended value is 50, meaning the home team receives an added 50 elo points to their rating. When the teams play in neutral settings, the HTA value is 0.

Chart, line chart

Description automatically generated With equivalent elo values, it is understood that the probability of either team winning is 0.5 and as the elo value difference grows so too will the probability difference.

### Margin of Victory

After the games are completed, the observed team win and lose can be used to calculate the shift in elo values. The first step is to factor in the margin of victory, meaning but how much a team wins. Logarithmic decay is used to emphasize that the greater the margin of the win the less significant the win plays in the change of elo values. Larger margin of victories will still add more but sports contain momentum factors that can spiral teams into larger wins/lose. This is calculated from the perspective of the victor.

MV – Margin of Victory: This metric is the difference in goals between the two competing teams. The margin of victory uses a natural log function to demonstrate a decay in increasing margins of victory. This is due to momentum factors taking place in sports with teams feeling defeated and losing more, and teams that are winning carrying the momentum to score more goals.

E. G. Home Team Goals: 4 Away Team Goals: 2 MV = 2

MVM – Margin of Victory Multiplier: This is one factor in the shift in elo ratings from resulting games. It contains two metrics analyzed during the optimization process. MVM\_A is the scaling factor and MVM\_B is the base or intercept. These metrics adjust the MV.

### Autocorrelation

FiveThirtyEight adds an autocorrelation factor; this is to account for the tracking of past and present results correlating with each other. In the NHL this is found to be small so FiveThirtyEight uses a tiny factor. The scaling is so tiny that it rarely contributes to the overall elo value shift. When teams are even this autocorrelation will contribute little to the elo value shift but will add little adjustments to the shift with teams of large differences to bring them closer to the average over time. This autocorrelation factor is multiplied into the Margin of Victory Multiplier.

AC – Autocorrelation

ACF – Autocorrelation Factor: The advertised FiveThirtyEight value is 2.05

ACM – Autocorrelation Multiplier: This is the scaling factor of the autocorrelation, with and advertised value of 0.001

### Shift Equation

The most important equation is the shift equation. This combines the results together and multiplies them by the shift factor K. The shift equation is only needed to be calculated once and the resulting shift value is added to the winning team and subtracted from the losing team.

K – Shift Factor: FiveThirtyEight has an advertised value of 6

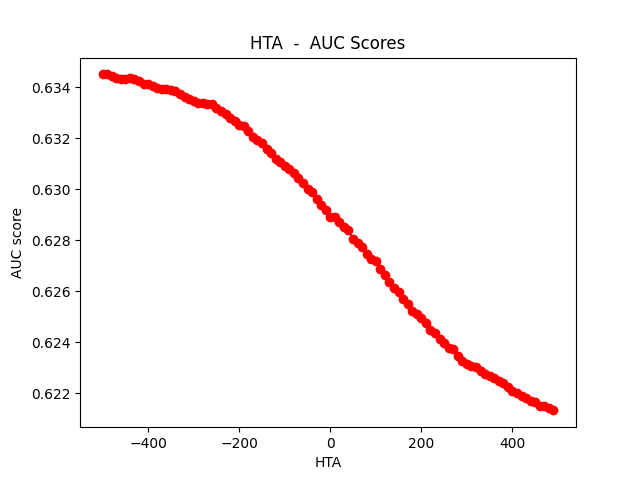
TeamWin: This is essentially the observed probability of the team winning, meaning that if the team won then it is 1 and if they lost then its value is 0

TeamProb: This value is the probability of the team winning, given by the sigmoid function

## Marginal Process

The betting process is zero-sum, and these probabilities and money-lines are very finely tuned. Overtime it is extremely hard to gain money as averaging out will either lead to zero gains or even loses. This means any marginal gain is considered an advantage against the betting market. Optimizations and improvements face two major obstacles to being able to access any advantage or gains. The first is that all betting agents are using elo systems and similar methods to generate their probabilities and money lines as they are designed to track live ratings. The second limitation, an elo system has two characteristics in the same way a vector does, the magnitude and the direction. If the top team’s elo value is 600 or 6000 with the rest of the teams being scaled then fundamentally our elo system is the same, this is simply changing the “magnitude” of the elo system. The means to make optimizations, there needs to be shifts in the direction of the elo system itself.

## Optimizations

 The optimization process was inspired by the gridsearch methods for machine learning. Custom functions were developed but resulted in a computational requirement that was not practical. Instead, individual optimization were performed, by searching a range for the individual metrics while keeping the remaining metrics constant. When choosing the optimal value for the parameters, the value producing the highest AUC score was chosen.

The HTA value when expanding forms the sigmoid curve due to it being in the input position in the function. This shows that FiveThirtyEight adds the HTA metric to the sigmoid function as a principle rather than because of model performance. For the optimizations, the new model value will be 0 to consider no bias. This is also a principle choice as the optimization of this metric does not produce consistent results.

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated The shift factor optimization revealed that its likely FiveThirtyEight chose 6 as a base value and built the rest of the structure based on this fact. The optimal value of 6.45 was chosen due to the right-side peaks being smoother than the left-side peaks.

MVM\_ A optimal value choice was 0.89

Chart, scatter chart

Description automatically generated

MVM\_B optimal value choice was 0.9

A picture containing shape

Description automatically generated

The ACF seems to play no role in the improvement of the model.

Chart, line chart

Description automatically generated

The ACM value plays no role in the improvement of the model.

Final Optimizations:

|  |  |  |
| --- | --- | --- |
|  | Advertised | Optimizations |
| HTA | **50** | **0** |
| K | **6.0** | **6.45** |
| MVM\_A | **0.6686** | **0.89** |
| MVM\_B | **0.8048** | **0.9** |
| ACF | **2.05** | **2.05** |
| ACM | **0.001** | **0.001** |

## ROC and AUC

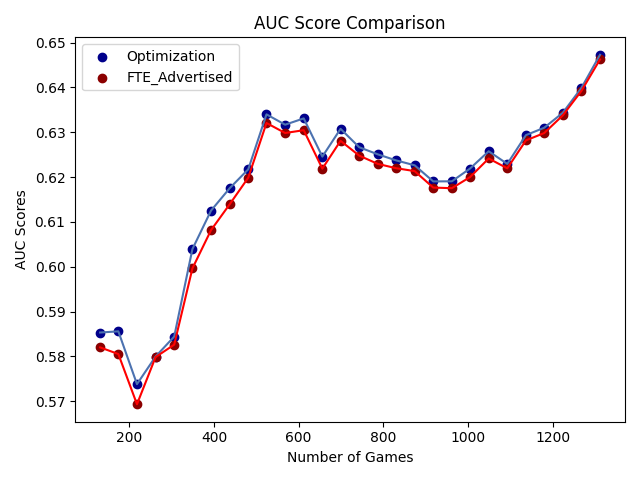
To access and potential gains compared to the FiveThirtyEight system, standard optimization techniques will be applied, tracking small movements in model metrics and comparing the Area Under the Curve (AUC) value on a Receiver Operating Characteristic (ROC) curve. The ROC curve is a graph of the True Positive (TP) values vs the False Positive (FP) values at various thresholds. The closer the curve it to the top left, the better the Chart, line chart

Description automatically generatedperformance of the model.

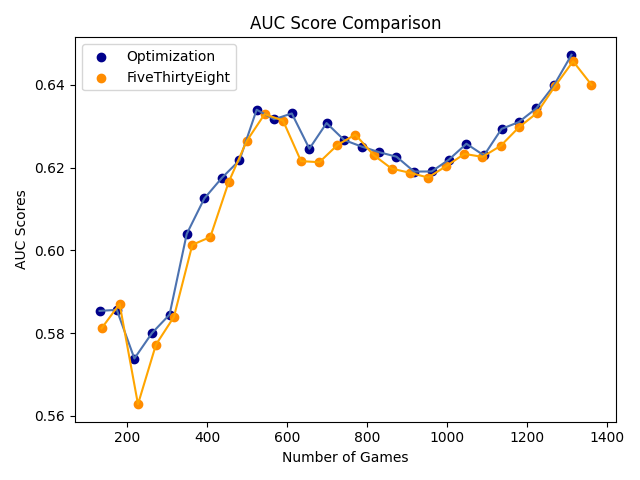
The blue line represents random guessing with an AUC value of 0.5, and the orange line is FiveThirtyEight’s model performance with an AUC value of 0.64.

## Final Comparison

The nature of the NHL data is a time series, and to better grasp the performance of the elo systems, the AUC values were compared based off of the values over time.



This is the comparison between the advertised values and the optimizations. Our optimizations gave us small marginal gains at the start of the season and during the normalization period when elos readjust from the previous year. The small gain seems to close over time as elo systems are meant to drift towards the mean. This is also a period in the season in which teams begin to win more to get into the playoffs and lose more to get a higher seed in the draft.



This is the comparison to FiveThirtyEight’s performance; it shows clearly that they make their own adjustments and do not follow explicitly the advertised system. This was discovered during the process and originally motivated optimizations. Just as when compared to the advertised values, the two drift closer towards as the season moves towards the end.

# Team Analyst

Machine learning possesses a stronger opportunity for the team analyst. Machine learning models often contain partial ability to interpret the weight that features contribute to the predictions. Interpretability is the key component to the team analysts as they are attempting to bring practical information to the organization. There are some models that excel greater than others in the ability to interpret model dynamics, but the model still needs to fit well with the data being used.

## Model Choice

The game stats collected:

|  |  |
| --- | --- |
| *Abbreviation* | *Description* |
| S | Shots |
| PIM | Penalty in Minutes |
| *PPG* | Power Play Goals |
| *PPO* | Power Play Opportunities |
| *SHG* | Short Handed Goals |
| *SA* | Shots Against |
| *PIMA* | Penalty in Minutes Against |
| *PPGA* | Power Play Goals Against |
| *SHGA* | Short Handed Goals Against |
| CF | Corsi For, at even strength (Shots + Blocks + Misses) |
| CA | Corsi Against, at even strength (Shots + Blocks + Misses) |
| *CF%* | Corsi For Percent, at even strength, (CF / (CF + CA)) |
| *FF* | Fenwick For, at even strength (Shots + Misses) |
| *FA* | Fenwick Against, at even strength (Shots + Misses) |
| *FF%* | Fenwick For Percent, at even strength (FF / (FF + FA)) |
| *FOW* | Faceoff Wins |
| *FOL* | Faceoff Losses |
| *FO%* | Faceoff Win Percentage |
| *oZS%* | Offensive Zone Start Percent, (oZS / (oZS + defensive ZS)) |
| *PDO* | Shooting Percent + Save Percent |

CF, CA, FF, FA, FOW, FOL are all dropped due to the percentage metrics measuring these values already. PDO is dropped since it includes the goals in the match and the models caught this aspect and were all relying significantly on the metric. Before being removed all the models were performing with accuracies above 0.90.

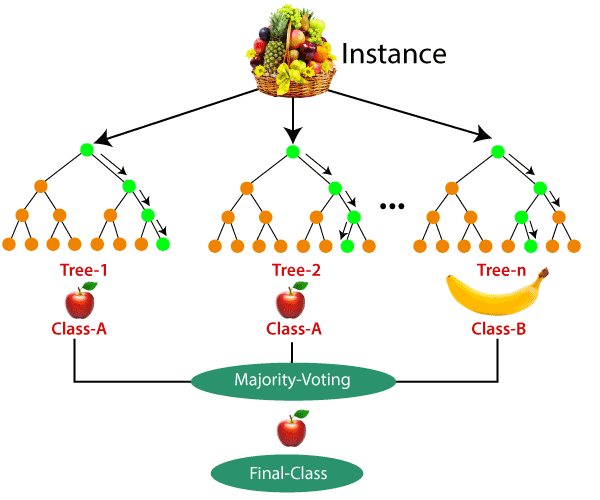
The means to choose a model will be a pipeline analyzing the accuracy, precision, and recall values of the model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | Recall |
| Logistic Regression | **0.552** | **0.609** | **0.778** |
| Support Vector Machines | **0.586** | **0.667** | **0.667** |
| Decision Tree | **0.552** | **0.647** | **0.611** |
| Random Forest | **0.655** | **0.667** | **0.889** |
| K-Nearest Neighbors | **0.621** | **0.652** | **0.833** |

The pipeline values point towards random forest being the best model to fit the data.

## Random Forest

Random forest is the best candidate likely due to hockey containing lots of randomness from all the components associated with the sport. Random forest’s random generation of decisions is likely able to form better to the numerous features and their nature. The use of a majority consensus allows the most optimum was to predict outcomes as well as determine the level of importance of the features.



This is a simplified structure to how the consensus works. For the binary classifier, if a majority of the trees determine the class to be a win, then it gets classed as a win.

Qr code

Description automatically generated

This is an individual showing the

branching off into leaves and other branches based on the values. This tree shows SA, oZS%, and S.

## Feature Importance

Chart

Description automatically generated Feature importance allows the team analyst to extract the desired information from the models. The feature importance is weighted by and order by the feature contributing most to the binary classification.

Chart, funnel chart

Description automatically generated

Comparing these two feature importance graphs derives insight for the team analyst. The red graph shows a stronger focus on FO% compared to blue whose most important feature is oZS%. This allows teams to identify areas of focus when playing against teams, and weaknesses to work on part way through the year. Another metric to bring attention to is PIMA, this shows that these factors have different significance for these teams.

Another form of checking model performance is the confusion matrix. This is a chart that compares a classification model’s predictions to the real values. Since the team analyst is not focused on developing the model but rather interpretability, these charts show more the overfitting of the model to the training data. This means that confusion matrixes can be used to show a winning and losing team. For these matrices, 1 is a win and 0 is a lose.

Graphical user interface, application

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

The TOR is overfitted to predicting wins showing that TOR is a strong team with lots of wins. The CGY chart shows the model predicting lots of losses, showing that CGY struggled in the first part of the season. The LAK chart is more well rounded as this was a middling team finished the season in the middle of the standings.

### Hyper-Parameters

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, line chart

Description automatically generated As already mentioned, hyper-parameter tuning is not necessary as this is to increase the reliability of the model where the analyst is attempting to use the model to extract information. Tuning the hyper-parameters would fit the model to better predict future games but would sacrifice the weights of the features on the already played games. The process shows that the lowest value for the maximum leaf nodes is best.

After adjusting the maximum nodes, this increases the reliability of the model.

# Further Analysis

## Betting Application

After developing elo systems and optimization, the next steps involve incremental process of inclusion of more information into the elo system. One possible big component is the injury of players as losing players that generate goals and chances significantly hinders the performance of the team. Another type of improvement would involve changing the starting elos of the team, this project used FiveThirtyEight’s elos, but the findings revealed an adjustment period that can be better accounted for through transferring elos with players or re-adjusting based off salaries or average performance of players. All these methods allow the sports bettor to bettor tips odds in their favor when betting against money-lines or generating them.

## Further Team Analytics

The potential for the team analyst is endless due to the ability for data science tools to breakdown data and statistics. The next step could be doing historical analytics, comparing divisions and multiple teams, playstyles, coaches, any possible way in which sports are traditionally understood. This also allows for non-normal ways to re-interpret sports that might not have been previously known.

# Conclusion

Data science applications are vast in sports statistics. For instance, the use of data visualization provides an easier way to understand the plethora of data within the field for this project. Sports statistics was broken down into two separate agents showing how separate strengths in data science on both sides of the industry. The sports bettor was able to use data science tools to better understand elo systems and the overall statistics involved in the betting industry. Optimization principles from machine learning accomplished an advantage over the originally designed FiveThirtyEight system. The team analyst used machine learning to gain access information latent within overwhelming data. Random forest’s modeling methods fit well to the randomness of sports and can be used to aid the analyst.

# References

All graphs and codes can be found available at <https://github.com/RichardSauve/Senior-Project>

## Data

*2022-23 NHL schedule and results*. Hockey. (n.d.). https://www.hockey-reference.com/leagues/NHL\_2023\_games.html

Ryanabest. (2021, October 7). *How our NHL predictions work*. FiveThirtyEight. https://fivethirtyeight.com/methodology/how-our-nhl-predictions-work/

Fivethirtyeight. (n.d.). *Data/NHL-forecasts at master · fivethirtyeight/data*. GitHub. https://github.com/fivethirtyeight/data/tree/master/nhl-forecasts

*2022-23 Toronto maple leafs team gamelog*. Hockey. (n.d.). https://www.hockey-reference.com/teams/TOR/2023\_gamelog.html

## Resouces

Langville, A. N., & Meyer, C. D. (2012). Chapter 5: Elo’s System. In *Who’s #1?: The science of rating and ranking*. essay, Princeton University Press.

Lutz, M. (2020). *Learning python*. O’Reilly.

Fenner, M. E. (2022). *Machine learning with python for everyone / Ji Qi Xue Xi Python ban / make E. fenna zhu*. Ji xie gong ye chu ban she.