Puck Predictions: Unraveling the NHL Game Forecasting Riddle

Jason Vasquez Dylan Skinner Jeff Hansen
Benjamin McMullin

March 28, 2024

Abstract

The goal of this project is simple: predict the outcomes of NHL games from any given state. As simple as the problem statement is, however, the solution is not so straightforward. To solve this problem, we will use a variety of machine learning techniques, including logistic regression, XGBoost, and ARIMA models. Additionally, we utilize a form of MCMC to simulate the outcomes of games from any given state. Our hypothesis is that we will be able to successfully predict the outcomes of NHL games with a high degree of accuracy using these tools.

1 Problem Statement and Motivation

In the world of sports analytics, predicting the outcomes of games is a common and challenging problem, with live win predictions adding an extra layer of complexity. For most sports, there are a plethora of widely accepted—yet hidden—predictive models and methods that are used to predict games. In addition to this, most sports have easily accessible statistics and graphics that give current win probabilities for any live game.

Hockey, however, is a different story. While there are some methods used to predict the outcome of National Hockey League (NHL) games, these models typically belong to sport books and their nuances are not publicly disclosed. Additionally, hockey analytics is not as developed as it is in other sports, such as basketball or baseball. This lack of model transparency and public interest in hockey analytics makes predicting the outcomes of NHL games a very underdeveloped and challenging problem. Previous attempts and research into predicting NHL games has relied on methods such as

decision trees and artificial neural networks [2] (from 2014), naïve bayes and support vector machines [3] (from 2013), and Monte Carlo simulations [4] (from 2014).

In addition to model research, some research has also gone into developing new features that can be used to better predict the game outcomes. The two biggest engineered classes of features are the Corsi and Fenwick[1] metrics (both around 2007).

Our project seeks a similar outcome to the research mentioned above: predict the outcomes of NHL games. Not only this, but we seek to provide live, acturate win probabilities for any given game state. Despite the simplicity of the problem statement, as mentioned, the solution is not so straightforward. The NHL provides fast-paced games with many events occuring in quick succession. Our goal is to use this abundance of data and new approaches to build upon previous research.

Our motivation for this project exists strictly as fans of the sport and as data scientists. Our model is not intended to be used for gambling or any other nefarious purposes—any use of this model for such purposes is a misuse of our work.

2 Data

Our data came from the hockeyR Github repository[1]. This repository contains an abundance of data about every NHL game that has occured since the 2010-11 season. This data includes information about the events that transpire in a game (hits, shots, goals, etc.), which teams are playing, who is on the ice, and the final score of the game. The data is stored in a series of `.csv.gz` files, allowing for easy access and manipulation.

Each game in a season is given a unique identifier (`game_id`), which is constant across all events in a game. Every event that occurs in a game will be stored in the `event_type` column. There are 17 unique event types, including things such as game start, faceoff, shot, hit, and goal. Most of these event types are not relevant to our analysis, so we remove them from the dataset. After removing the unnecessary events, we are left with nine events: blocked shot, faceoff, giveaway, goal, hit, missed shot, penalty, shot, and takeaway. These events are attributed to the team and player that

[1]These metrics were created by sports bloggers Tim Barnes and Mark Fenwick, respectively. We were unable to locate the original blog posts talking about these metrics, but a good article to learn more about the math can be found here `https://thehockeywriters.com/corsi-fenwick-stats-what-are-they/`.

performs the event. We only take into consideration the team that performs the event and discard the player information.

The data also contains information about when the event occured. This appears in a variaty of formats, but we only use the `game_time_remaining` column. `game_time_remaining` starts at 3600 (60 minutes) and counts down to 0. If the game goes into extra time, i.e., it is tied after 60 minutes, `game_time_remaining` will be a negative value.

We found that our data did not contain any missing values that was not easily explainable. For example, if a game is starting, there will be no events for penalties, which will result in a `NaN` value in the penalties column. Additionally, any data that was confusing or not easily explainable (for example the home team having 7 players on the ice and the away team having 5), was manually verified by watching a clip of the game where the event occured to make sure the event was recorded correctly. We did not find any incorrectly recorded events, so we did not remove any strange events from out dataset.

3 Methods

3.1 Bayesian Network

We first used a Bayesian Network to establish a benchmark for probability using several key features.

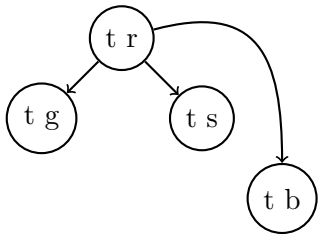
Bayesian networks are probabilistic graphical models that represent probabilistic relationships among a set of variables using a directed acyclic graph (DAG). In a Bayesian network, nodes represent random variables, and directed edges between nodes represent probabilistic dependencies between the variables. Each node in the graph is associated with a conditional probability distribution that quantifies the probability of that variable given its parent variables in the graph.

For our purposes, we predefined the structure of the network, and used the data to calculate the conditional probabilities for each node. We then used the network to calculate the probability of a team winning given the current state of the game.

The computational complexity of Bayesian Network inference is high, with exact inference being an NP-hard problem TODO CITE. Using the python package pgmpy, we originally tried to fit a network with all 26 of our features, but our computational resources failed to fit this network. Then, to get a baseline for our future predictions, we simply fitted the model with the base features of home goals, away goals, home shots, away shots, home

blocked shots, and away blocked shots. These features were chosen as priors because of our opinion that they are the most important to the game, based upon our knowledge of hockey.

(Basic graph to test)



3.2 t-SNE, UMAP, and PCA

3.3 Regression and XGBoost

3.4 MCMC Game Simulation

To simulate hockey games, we setup a Markov Chain where the states are a tuple of three consecutive events that occurred. For example, if the home team won a faceoff, the home team lost the puck, and then the away team shot the puck and missed, the markov chain would look like figure 1. The probabilities of transitioning from one triple-state to another triple-state is calculated by:

$$\begin{aligned} P(s_{t+1} = (B, C, D) \mid s_t = (A, B, C)) &= P(D \mid (A, B, C)) \\ &= \frac{\{\text{\# of times (A,B,C,D) happend}\}}{\{\text{\# of times (A,B,C) happened}\}} \end{aligned}$$

Where A, B, C, D are events that can occur in a game, and the tuple (A,B,C,D) represents that "A then B then C then D" happend right after eachother in a game.

To simulate a game, we performed this monte carlo algorithm:

	curr_state	next_event	probability
39970	FACEOFF_HOME,GIVEAWAY_HOME,MISSED_SHOT_AWAY	SHOT_AWAY	0.1853207
39969	FACEOFF_HOME,GIVEAWAY_HOME,MISSED_SHOT_AWAY	BLOCKED_SHOT_AWAY	0.1138666
...	...	...	...
39981	FACEOFF_HOME,GIVEAWAY_HOME,MISSED_SHOT_AWAY	PENALTY_AWAY	0.0167210
39976	FACEOFF_HOME,GIVEAWAY_HOME,MISSED_SHOT_AWAY	GOAL_HOME	0.0085793

Figure 1: Sample Markov Chain for the current state (home team won faceoff, home team lost puck, away team missed shot) with the probabilities of the next event sorted from highest to lowest

4 Results

5 Analysis

6 Ethical Considerations

Predicting win percentages or outcomes in hockey games, like any sport, raises several ethical considerations. Here are some key points we want to address:

- **Gambling and Addiction**: Our win percentages and predictions might be used by those who wish to gamble, which could lead to addiction and financial harm, especially if undue trust in placed in these methods. Any publication of these methods or predictions would be accompanied by promoting healthy and responsible gambling practices.
- **Fairness and Integrity of the Game**: Sometimes, coaches and players becoming aware of their chance of winning can affect how the game is played, potentially harming the integrity of the sport. We must be careful to not provide an unfair advantage to any team or player. Inaccurate predictions could lead a team to believe that the game is out of reach when it isn't, and we want to avoid that.

Overall, our predictions, like any, should not be considered declarative for gambling or performance purposes, but are rather an interesting exposition into the complexity of sporting events.

7 Conclusions

References

[1] Dan Morse. hockeyR-data: A collection of hockey datasets for use with the hockeyR package. `https://github.com/danmorse314/hockeyR-data`, 2024. Accessed: March 2024.

[2] Gianni Pischedda. Predicting nhl match outcomes with ml models. *International Journal of Computer Applications*, 101:15–22, 09 2014.

[3] Jason Weissbock, Herna Viktor, and Diana Inkpen. Use of performance metrics to forecast success in the national hockey league. In *European Conference on Machine Learning: Sports Analytics and Machine Learning Workshop*, 2013.

[4] Josh Weissbock. Forecasting success in the national hockey league using in-game statistics and textual data. 2014.