

# Puck Predictions: Unraveling the NHL Game Forecasting Riddle

Jason Vasquez      Dylan Skinner      Jeff Hansen  
Benjamin McMullin

March 27, 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

## 2 Data

Our data came from the hockeyR Github repository[1]. This repository contains an abundance of data about every NHL game that has occurred 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 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 occurred. This appears in a variety 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 occurred 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 our 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.

### 3.2 t-SNE, UMAP, and PCA

### 3.3 Regression and XGBoost

### 3.4 MCMC Game Simulation

## 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 is 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.