NHL Game Prediction Modeling

BrainStation Initial Capstone Presentation

Mark Shumka

October 20, 2023

Opportunity

The outcome of National Hockey League games is notoriously difficult to predict relative to other sports

The insights that can come from an accurate prediction model would be of value to multiple stakeholders, including

1. Team management – owners, general managers, coaches, and others running professional teams
2. Hockey journalists/websites – people who create content for hardcore and casual hockey fans
3. Bettors – people who gamble on hockey, and the sportsbook owners who facilitate the process

Potential Impact

- Visibility into the underlying drivers of wins would help team management prioritize player acquisition based on relative performance on key metrics. Application to team construction could increase the likelihood of success, which leads to increased revenue (ticket sales, merchandise).
- A model that is even slightly more accurate than the betting odds is a huge advantage to people betting on hockey.

As a measure of success, model performance will be evaluated on the ROI it would provide if used to bet on actual games

How can Data Science address the problem?

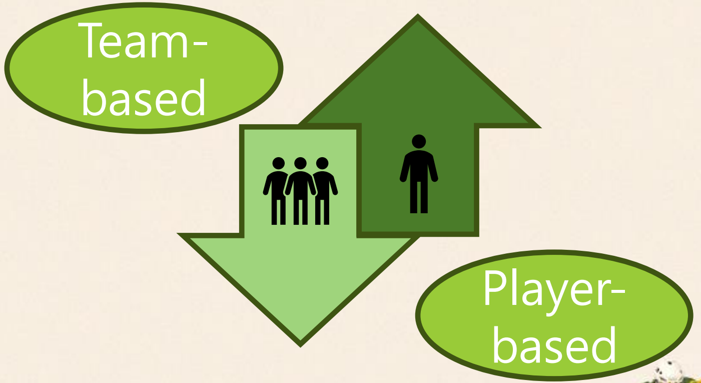
Understand patterns and trends from historical data to develop a model to predict the outcome of future games

Basic Stats vs. Advanced Stats

BASIC STATS
Shots
Goals
Hits
Save Percentage
Power Play Percentage
Faceoffs

ADVANCED STATS
Corsi
Fenwick
xGoals
High Danger Chances
PDO
Flurry Adjusted Expected Goals

Top-Down vs. Bottom-Up



Team-based

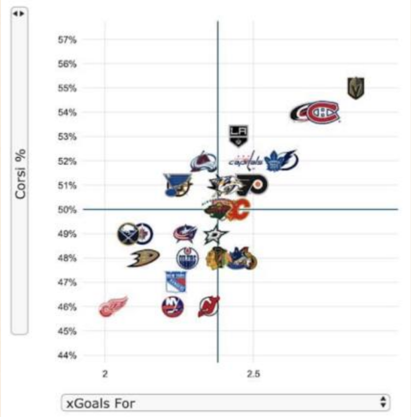
Player-based

The goal of the analysis is to identify and quantify the stats that have the biggest impact on the outcome of games

The Dataset

Two Datasets in one

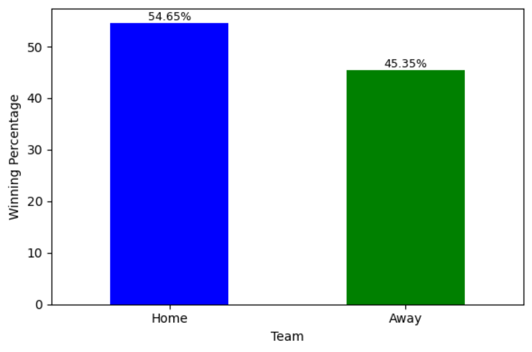
55k columns
16 rows



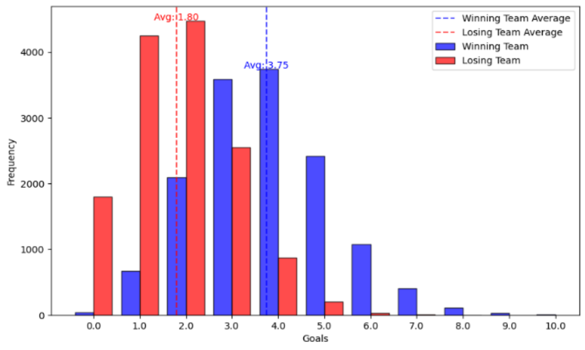
190k columns
111 rows

- 28k columns, 110 rows
- Data from every regular season game from 2008-2020
- Team-based stats
- Dependent Variable = Game outcome (Win/Loss)
- Independent Variables = Basic Stats & Advanced Stats (and their components)

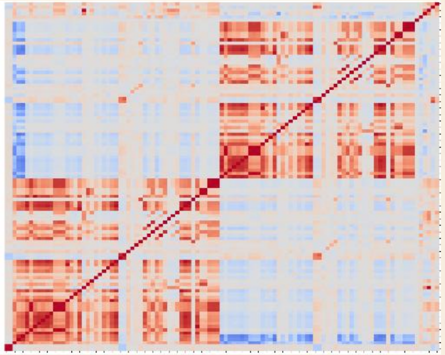
EDA and Initial Findings



Home teams win ~55% of games

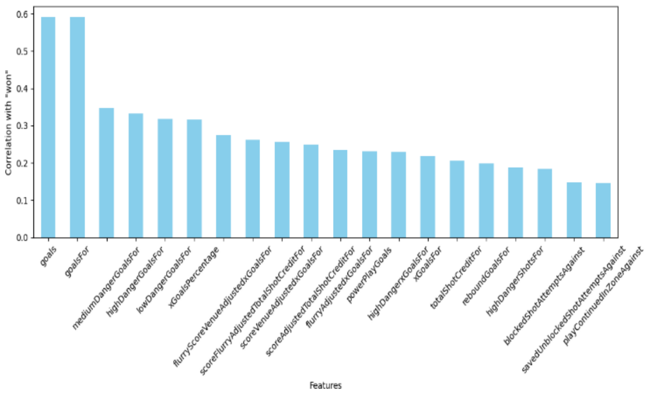


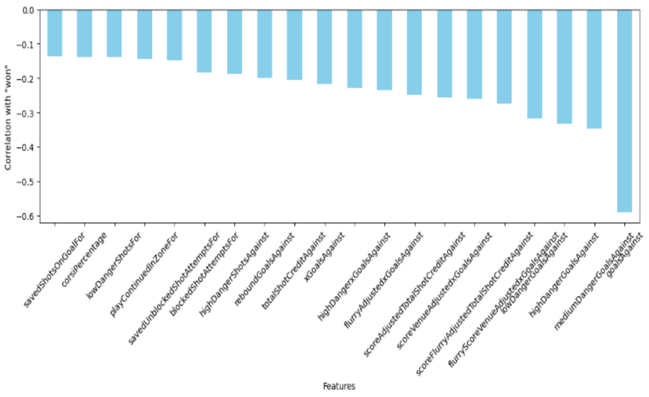
Teams score an average of 3.75 goals when they win, 1.8 when they lose



Correlation heatmap of 100+ variables!

Top 20 variables most and least correlated with wins





Open Questions and Next Steps

Condense the dataset

- Similarity of variables
- Multicollinearity
- Biggest drivers of wins
- Biggest predictors of wins

Expand the dataset

- Missing data
- Goalie stats
- Time, distance traveled between games
- Calculated metrics from raw data (e.g., save%)
- Salary cap hit
- Time series (i.e., trailing n games)

Modelling

- Logistic regression + machine learning
- Phase I – Identify key variables and develop a reactively accurate model
- Phase II – Develop forward-looking model with defined inputs that predict outcomes