Regular Season NHL Game Prediction Model

Nolan Whittaker

CMPT 353 - Computational Data Science

Simon Fraser University

1. Introduction

After the Canuck's heroic 2023 season and playoff run, I found myself reintroduced to the amazing sport of hockey. Having watched all 82 of their games this past 2024 season, I began to wonder, "Could I predict who will win?". This is no simple task given the complexity and unpredictability of the sport. There is no way to foresee a star player having an off night, an untimely penalty late in the game, a goaltender delivering a career performance, or a key player being ejected. However, I believe it's possible to make a data-driven prediction that, while not perfect, can provide meaningful insights into the likely outcome of a game. The goal for my project is to develop a model that could predict the 2024 NHL season with a 61% accuracy.

2. Identifying Key Factors

To predict the winner of any given NHL game, we need to determine the key factors that define a successful team. In my opinion, a winning team has success in 3 elements:

- Ability to win close games
- Consistent Offensive Performance
- Reliable goaltending

The ability to win close games and maintain consistent offensive production is crucial in the NHL. Teams that excel in these areas gains a significant edge and usually finish high in the standings. Similarly, reliable goaltending is also essential for success. For example, the 2024 Colorado Avalanche initially relied on goaltender Georgiev, who posted a save percentage of .874 and the team struggled at the beginning of the season. Later, the team replaced him with Mackenzie Blackwood, who recorded a save percentage of .913 over more games, eventually helping the Avalanche climb the standings to make the playoffs. Without good goaltending, no team in the NHL including the Avalanche would be able to consistently win games.

Basic NHL statistics can help us measure these 3 elements. Offensive performance can be measured using metrics such as goals per game, high danger chances, and shots per game. The ability to win games can be quantified using metrics such as win percentage, current win streak, and the total number of wins. Reliable goaltending can be captured through statistics like save percentage. Aside from the regularly tracked statistical data, we need to consider several contextual factors that may influence a team's chances against a particular opponent. These factors include head-to-head record, back-to-back games, and whether the team is playing at home or away. All these scenarios have been known to impact a team's chances of winning, as we will show later.

3. Obtaining Relevant Data

For this analysis, we used Moneypuck.com to download game-level statistical data covering the 2008 to 2024 season. This dataset contained over 218,000 rows representing different scenarios (all, 5v5, or 5v4). I filtered this dataset to only include full game stats and added a column indicating which team won that game. However, the data did not specify which team won in a shootout, leaving those games recorded as ties. Due to licensing restrictions of some websites, I manually created shootout_data.csv, which contains the results of each shootout and the winning team on the corresponding date. Combining these datasets provided the necessary information to calculate each team's statistics, and determine which team won each game from 2008 to 2024.

In addition to game-level statistics, I also parsed the NHL API to gather standings data for the end of each season. This allowed me to compile a CSV file containing each team's final rankings, and total points. This dataset enabled me to perform statistical analyses, comparing individual team stats with their leaderboard positions, and to explore correlations between specific performance metrics and overall season success.

4. Data Preparation

To prepare the data for machine learning models, I calculated the season average for these statistics which included, shots on goal against, shots on goal for, goals for, goals against, save percentage, and win percentage. Additionally, I tracked contextual factors such as whether the team was playing on a back-to-back, the number of wins in the last five games, and the head-to-head record between the two teams.

5. Statistical Analysis

To support the variables I used to predict the winning team, I performed linear regression analyses comparing each team's total shots for and against, high-danger shots for, and goalie save percentage in comparison to their final season ranking.

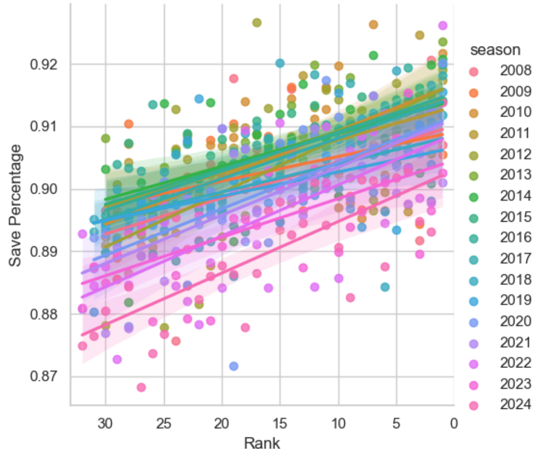

Figure 1 - Save Percentage

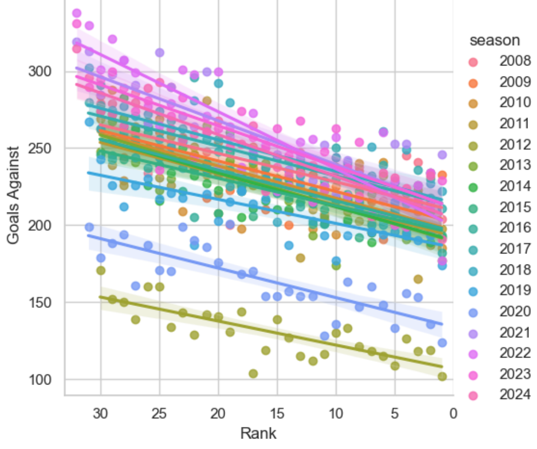

Figure 2 – Goals Against

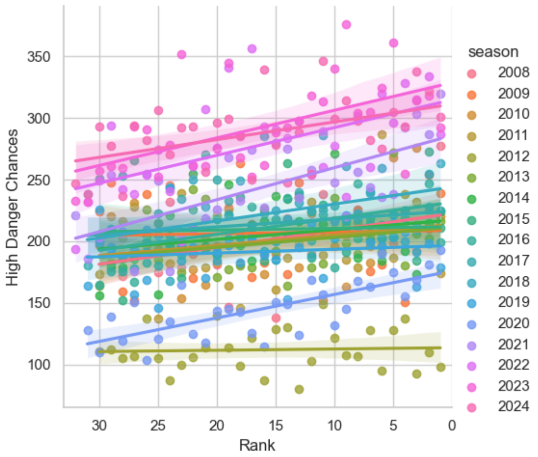

Figure 3 – High Danger Chances

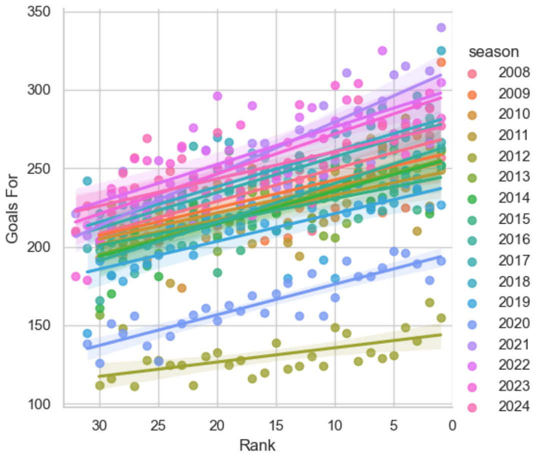

Figure 4 – Goals For

The correlations between team statistics and season rank are as follows:

- Goals For vs Rank: r = -0.43
- Goals Against vs Rank: r = 0.49
- High Danger Shots For vs Rank**:** r = -0.17
- Save Percentage vs Rank: r = -0.56.

These results show that goals against, and save percentage have the strongest linear relationships with team rank, and that all the columns beside high danger shots show significance for predicting successful teams. In addition to this, I similarly analyzed the impact of back-to-back games and home versus away status in relation to winning.

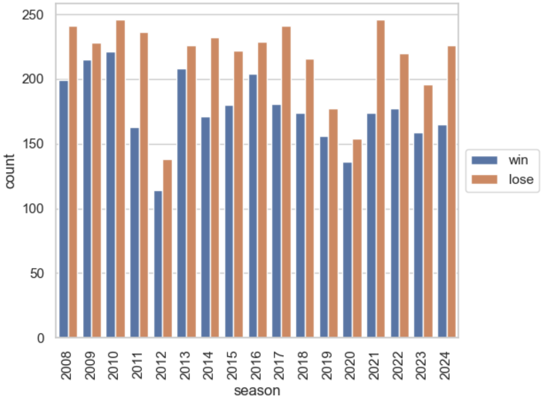

Figure 5 – Back to Back Game Results

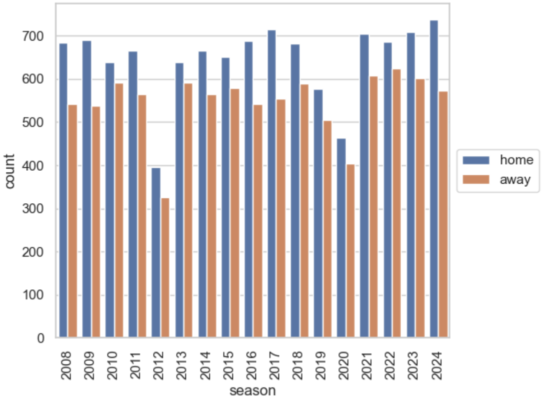

Figure 6– Home and Away Wins

Figure 5. and Figure 6. show that teams playing a back to back are more likely to lose, and teams tend to perform better when playing at home, reflecting the well documented home-ice advantage in the NHL and the fatigue of playing back to back.

6. Machine Learning Results

The machine learning models were trained on averaged team statistics over the season, along with contextual features such as head to head record, back to back, and home or away game. We utilized Random Forest Classifier and Gradient Boosting because predicting NHL games is complex and it requires utilizing multiple features for an accurate prediction. The Random Forest introduces randomness in the tree construction which helps our prediction capture subtle patterns in the games of a season, and reduces the chances of overfitting. Similarly, Gradient Boosting was used with the hope to better capture complex interactions between features by sequentially improving on the errors of previous trees**.**

7. Results

Random Forest Result

Depth min_leaf	3	4	5	6	7	8
5	0.5922	0.5899	0.5830	0.5792	0.5754	0.5739
7	0.5929	0.5884	0.5876	0.5774	0.5792	0.5800
10	0.5929	0.5960	0.5792	0.5708	0.5701	0.5731

Gradient Boosting Result

Depth min_leaf	3	4	5	6	7	8
5	0.5670	0.5670	0.5647	0.5373	0.5434	0.5282
7	0.5701	0.5647	0.5647	0.5541	0.5320	0.5396
10	0.5739	0.5525	0.5571	0.5373	0.5449	0.5602

The results are quite surprising. On average, we notice that the Random Forest outperforms Gradient Boosting across most configurations, with the best Random Forest achieving an accuracy of 0.5929, in comparison to 0.5739 for Gradient Boosting. Additionally, we observe that as the depth of the trees increases, the accuracy of both models generally decreases. This suggests that the models begin to overfit even at relatively low depths which is unexpected. Another notable observation is the effect of minimum leaf size. We notice that moderate leaf sizes tend to yield better performances. In addition to predicting the 2024 NHL season, I also applied the same optimal models to all other seasons to check whether the 2024 prediction was an anomaly. On average, the model achieved around 56% accuracy across all seasons. A particularly notable result was the 2021 season, where the model reached 61% accuracy. This higher performance can likely be attributed to the shortened season caused by the COVID-19 pandemic, which may have reduced variability and made game outcomes slightly easier to predict.

8. Limitations

Although we had a significant amount of data, we were still missing crucial elements that can heavily influence NHL game outcomes. Some of the unavailable statistics include power play percentage and penalty kill percentage. Additionally, other contextual factors such as the absence of star players like Quinn Hughes for the Canucks or Connor McDavid for the Edmonton Oilers could have a major impact on game results and, if included, might further improve the predictive power of the model.

9. Conclusion and Future

Although I failed to reach my desired 61% accuracy for the 2024 season, I am happy that we managed to achieve a 59% accuracy. This level of accuracy still performs better than randomly guessing, and can provide a solid hint towards which team is statistically more likely to win. The results showed that predicting highly complex NHL games is no easy feat, and even small differences in team performance, goaltending, or situational factors like home/away and back-to-back games can significantly affect outcomes. This conclusion leaves significant room for improvement that could be addressed by incorporating additional features, such as player injuries, starting goalie matchups, and in-season roster changes. By incorporating more features and data, future predictions could achieve higher accuracy and can help provide deeper insights into the features that allow teams to perform successfully in the NHL.

10. Project Experience

- Extracted, Transformed, and loaded 16 years of NHL data, performing grouping and aggregation to prepare it for statistical analysis and machine learning prediction.
- Engineered key features including team performance metrics, goaltending statistics, and situational contexts to enhance model predictive power.
- Predicted the 2024 NHL season with 59% accuracy using a Random Forest classifier and evaluated model performance across prior seasons to validate robustness.
- Developed key visualizations using Seaborn to highlight the most important factors influencing game outcomes, including offensive metrics, defensive metrics, and goaltender performance.

11. Credit

The 2008–2024 NHL data used in this project was downloaded from Moneypuck.com. All credit to Moneypuck for providing the dataset.