NHL Analytics

Using Machine Learning to Predict NHL Game Outcomes

Sawyer Jacobson

6/5/2021

Introduction

In all competitive professional sports, the ultimate goal is for one team or individual to win the game, match, or race, depending on the sport. For the mainstream team sports in the United States, whether or not a team makes the playoffs in each league is determined by their overall record and, depending on how the league is structured, which division they play in. The most popular professional sports leagues in the United States, such as the MLB, NFL, and NBA, have seen a lot of work in predictive analytics at various levels over the years. While most analytics can be useful to each individual organization to help them utilize data to field the best, most competitive team by optimizing on field performance, another interesting, and profitable, application of predictive sports analytics is in the growing industry of sports betting. Sports betting in the United States was illegal for close to 50 years between the 1970s until 2018[1] when the US Supreme Court ruled in favor of New Jersey to allow legalization of sports betting. The ban was initially put in place because of the resulting corruption that found its way into matches via athletes. The most notable examples being point shaving (purposefully missing shots) and even match throwing. The two main types of bets that can be made are *Moneyline* and *Spread* bets. A Moneyline bet is a traditional bet on who will win the game while a Spread bet is a bet on the point differential between each team at the end of the game. Both types of bets are great candidates for machine learning, but the focus here will be on predicting the winner of the National Hockey League (NHL) games with the potential application to Moneyline betting.

At first thought, picking the winner of a game or race may seem somewhat trivial: simply pick the team with better stats or record. In some cases, this might be a pretty safe bet, more so for some sports than others or if there is such an extreme difference in skill/performance between the teams. The latter of which is less likely in the NHL due to a salary cap (maximum dollar figure each team can spend on players) being in place across the league causing teams to

[1] https://www.cnbc.com/2018/05/14/us-supreme-court-rules-for-new-jersey-in-states-fight-to-legalize-sports-betting.html

be more or less similar in overall talent. However, there are many other factors that can affect the outcome of a game. Star players will have hot streaks or slumps that may have a great impact on the outcome of a game. A team on a losing streak may have a sudden burst of energy and turn their season around. Home field advantage plays a role as well, again more so in some sports than others. Lastly, there is the aspect of luck. In sports such as football and baseball, chance plays a noticeable role, but the outcome is more so impacted by the skill, performance, and confidence put forth by the teams. It's arguable that random chance plays a larger role in hockey than these other sports, affectionately referred to by some as "puck luck". Hockey is a very fast paced game played on an 85x200 foot enclosed sheet of ice using a roughly 3-inch diameter rubber puck with the goal of one of the 5 players on each team shooting the puck into the net past the opposing teams goaltender. Factors such as home ice advantage can be seen in historical data, and referee bias during games plays a role as well. Due to the speed, something as simple as an unlucky bounce off of the enclosing boards, a deflection in front of the net, or a turnover because of rough ice has the potential to almost instantaneously change the momentum and thus outcome of the game. These factors are impossible to capture in data and thus impose a theoretical limit[2] of around 62% to the accuracy of which a hockey game outcome can be predicted. Since this is a binary problem, the baseline for correctly predicting the winner could be set at 50%. This project will use an ensemble machine learning approach using traditional hockey statistics as features along with feature engineering to assess how well game outcomes can be predicted over the past decade based on overall accuracy.

Dataset Description/Preparation

Compared to other mainstream professional sports, the NHL has not seen as much work or research into analytics and thus does not have as much easily accessed data. A significant portion of this project went towards collecting these data and streamlining a process for future use. All of the data was scraped from the NHL's statsAPI[3]. The API stores the data in JSON format, and, while there is a large amount of data contained there, the overall interface is somewhat clunky with no official documentation. The only documentation that exists is created by fans such as Drew Hynes[4].

---

[2] Joshua Weissbock, Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data, 2014

[3] https://statsapi.web.nhl.com/api/v1/expands

[4] https://gitlab.com/dword4/nhlapi

The data used in this project consisted of game outcomes, in-game team statistics, and in-game individual player statistics for all regular season games collected for the 2010/2011 through the 2020/2021 seasons. Additionally, goalie and referee game data for the same time frame were collected for future use and not yet utilized in this current project. There are roughly 1,250 games each season with some variation over the years due to lockouts, expansion teams, and shortened seasons in 2019/2020 and 2020/2021 due to COVID-19. In total, data for 12,592 games were collected.

The game results data consists of one observation per game with a unique identifier for the game and one for each team, an indicator for whether the home team won, the final score for each team, and the venue where the game took place. The in-game team statistics consists of two observations per game, one for the home team and one for the away team, and contains identifiers for both teams, the game, an indicator for whether the team was at home, and traditional hockey statistics such as goals scored, penalty minutes, shots on goal, and more along with several special teams' statistics. Finally, the player data contains one observation for each player for each team for each game given that the player participated in the game. Like the other datasets, there are unique identifiers for each player, team, and game along with traditional statistics similar to what is present in the team data.

Since the player and team level data only contained data associated with each individual game, these data needed to be wrangled into a more desirable form. While there may be come correlation to player or team statistics across seasons, each season was addressed separately. For both players and teams, stats that add up over the course of a season, such as goals, hits, etc., were updated to be cumulative totals to be representative of the player or teams' performance through each game. Other statistics, such as time on ice for players and powerplay percentage for teams, were summed and divided by the number of games played at that point to calculate a running average.

While no single player statistic can personify how a team is playing, the concept of the "hot hand" may be an effective indicator of how a team is performing offensively. This is the idea that an athlete continually successful at a task, such as scoring, will have a greater chance of accomplishing said task in the future. We will express the "hot hand" concept as a player in a particular season on a *point streak* which is defined as the player having at least one goal or assist in 2 or more consecutive games with the value being the total number of games in that

streak. This feature is utilized in the game level data by counting the number of players on point streaks a team has after each game.

Similar features were created for the team level data in the form of winning and losing streaks defined as number of consecutive games won or lost respectively. This falls under the related idea that a team wining consecutive games will have gained a boost in confidence leading to an increased chance of winning the next game. Similarly, a team on a lengthy losing streak may lose some confidence necessary to perform and break out of that rut. Losing streaks may also lead to staff changes in an attempt to turn things around, but that level of granularity is beyond the current scope of this project.

At this point in our data preparation, an observation of team statistics for game N includes all game stats up through and including game N. We want to use only past data for predictions so in each season, each team had their stats from game N shifted up to game N+1 to be used as features for the next game on their schedule. Consequentially, since this require data from past games in a season, the first game of each season for each team will be dropped.

Figure 1: Number of games won and lost by home team for the 2010/2011 to 2020/2021 seasons.

Lastly, the modeling goal is to predict the game outcome. This project will approach this target by framing the prediction to be whether or not the home team will win. The team game statistics data has 2 observations per game, one for home and one for away. Define the home

team data as V_{Home} and the away team data as V_{Away}. The final feature vector for game N+1 with k games in a season will then be defined as

$$V_{Home-Away,N+1} = V_{Home,N} - V_{Away,N}, \text{ for } N \in [2, \ldots k],$$

the difference between the home and away team statistics through game N. Therefore, positive values will indicate higher values for the home team while negative values indicate higher values for the away team. As seen in Figure 1, in every season since 2010/2011, the home team of a game has won more often than not. Some seasons see an almost equal proportion of games won by the home team as lost, but nevertheless, the trend shows some evidence towards the notion of home field advantage. With the modeling data setup as previously described, the idea is that some remnants of this phenomenon will be learned from the data to assist with predictions.

In Figure 2, we see the distribution of the difference in total goals scored by the home and away teams in games for the feature vector defined above for all seasons colored by the winning team. The overall distribution is approximately normal. On the right half of the plot, we see that, as expected, the home team having scored more goals led to more wins. However, the converse is less common. To the left of 0, there are a few instances where the away team having scored more goals translated into more wins, but much less often and trending towards coin flip odds.

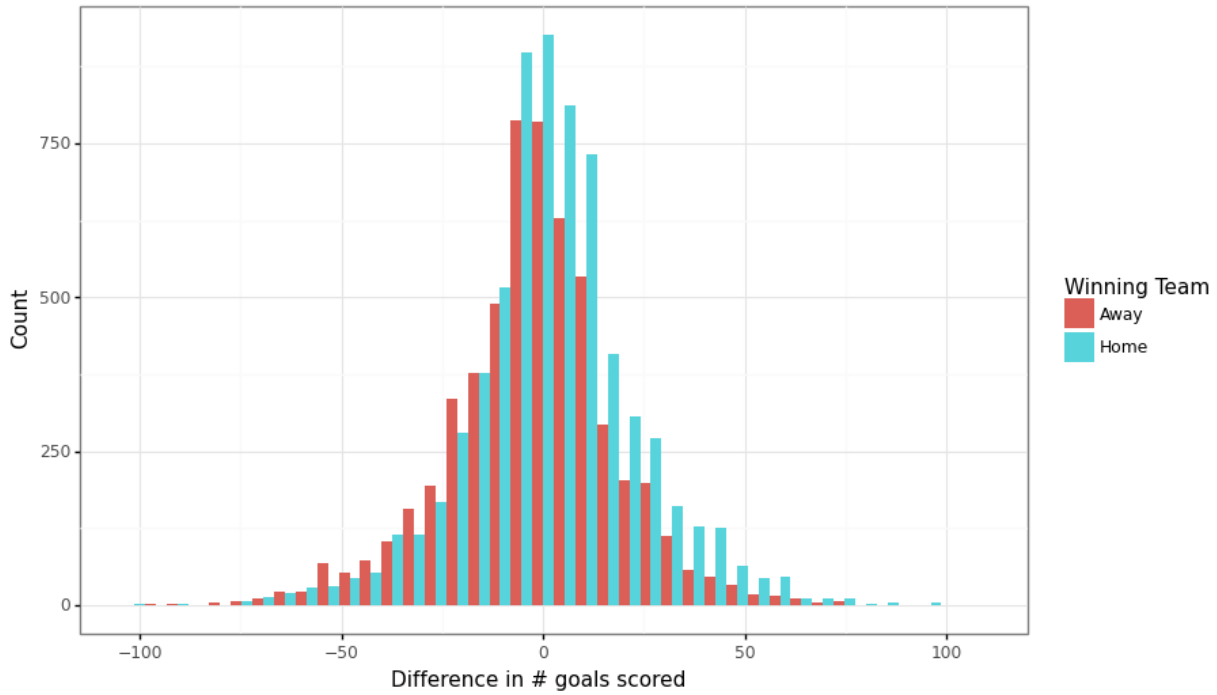


Figure 2: Distribution of in difference in goals scored between home and away teams for all seasons.

Methods

With the data processing performed as described above, I will now go over the modeling process. The data to be used in modeling include the traditional cumulative in-game team statistics as well as the extra features created above that signify hot, winning, or losing streaks. This gives a total of 15 features used as input variables. The metric used to assess performance will be overall accuracy since, in sports betting, there is no benefit to minimizing false positive or false negative rates since a bet is only won or lost. There is no consolation prize.

The style of play in the NHL has changed significantly over the last 2 decades. A league that used to be very physical and valued grit as near equivalent to skill has since shifted to a faster, more skill-based game. While it is expected that model performance will change between seasons due to random chance outcomes, I believe different variables may be identified as having more importance to winning games throughout this time span. Starting with the 2012/2013 through the 2020/2021 season, each season will be used separately as a test set and have all game outcomes from that season predicted. All games prior to that test season will be used as training data with 5-fold Cross Validation (CV) used to optimize the models over a group of prespecified hyperparameters. Each of the seasons described above will have their own models. To be clear, the 2015/2016 season will be used as an example. All data prior to this season (2010/2011, …, 2014/2015) will be used as the training data in combination with CV to optimize the models. All games in 2015/2016 season are predicted with the performance of each algorithm scored and recorded.

Altogether, six classification machine learning algorithms were fit and optimized for prediction accuracy to then predict on the test data. These models include Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), both Linear and Non-Linear Support Vector Machines (SVM), and a Gradient Boosting Classifier. The mix of tree based, distance based, and linear classifiers provide ample variety to be able to learn trends in these data. Additionally, two Voting Classifiers were fit using the training data optimized models usually a hard and soft voting rule. The hard voting rule uses the predicted outcome from a majority rule voting, and the soft voting rule uses the argmax of the predicted probabilities. To reduce the bias present in the soft rule voting classifier, models with perfect training accuracy were excluded from the ensemble. Feature importance from the RF model for the 2012/2013 and 2020/2021 seasons was also obtained.

Results

Model predictions were evaluated on their accuracy on test set predictions across all specified seasons. Figures 3 and 4 below depict the accuracy for both the training and the test datasets respectively. The dashed line in each subplot represents the baseline 50% accuracy or

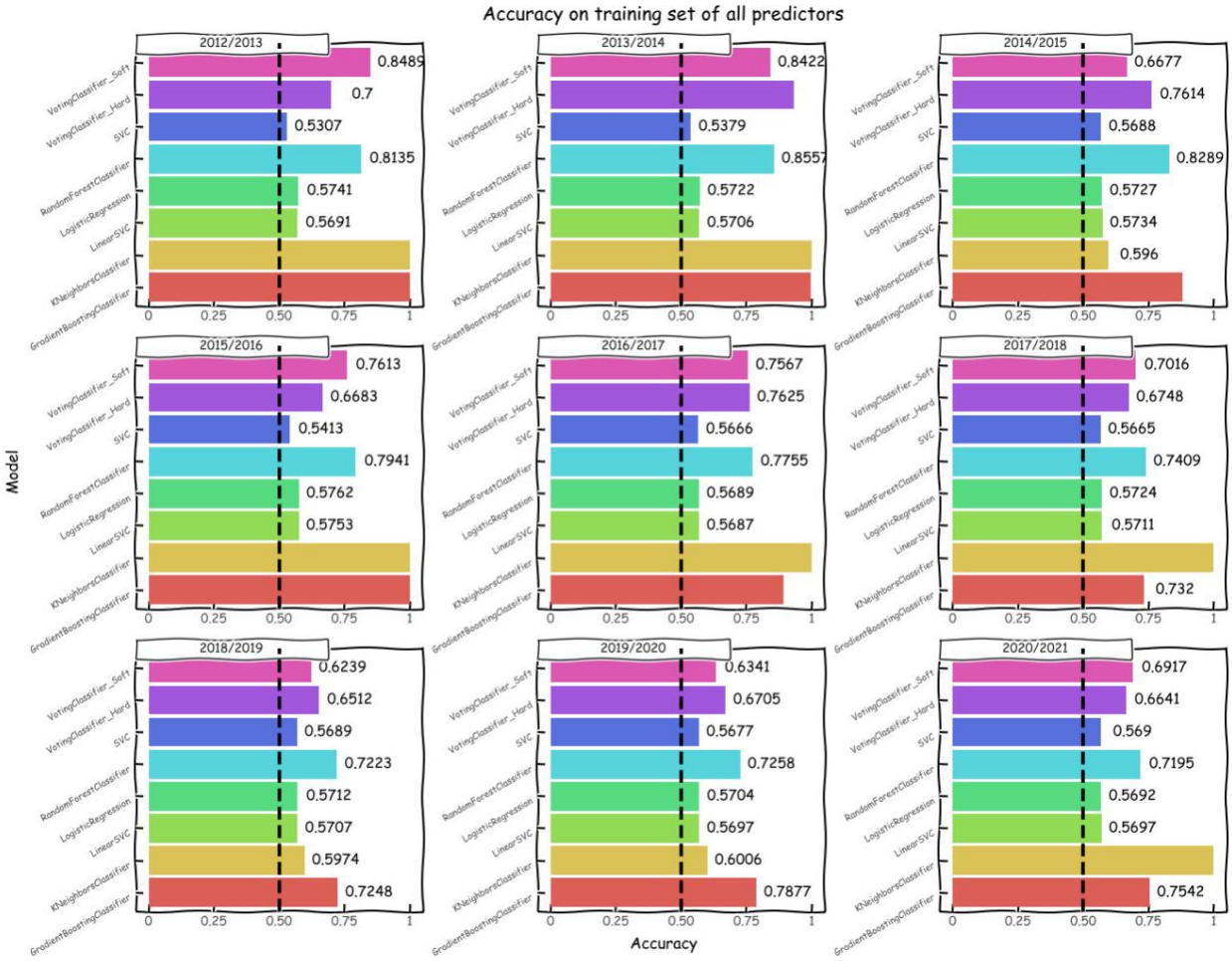


Figure 3: Accuracy for each model on the training data used for each season. Text numbers represent the overall model accuracy

the equivalent of predicting the winner via a coin toss. In Figure 3, we see a lot of variation in model performance on the training data. When performance is compared to the test data in Figure 4, models such as GBM and KNN were largely prone to overfitting (near perfect training accuracy) across all training data for each season. Random Forest was also consistently overfit but to a lesser extent. The ensemble classifiers also saw slight overfitting.

Looking closely at Figure 4, there was a fair amount of variation in the model prediction accuracies across each season. All models were able to beat the baseline 50% accuracy in all

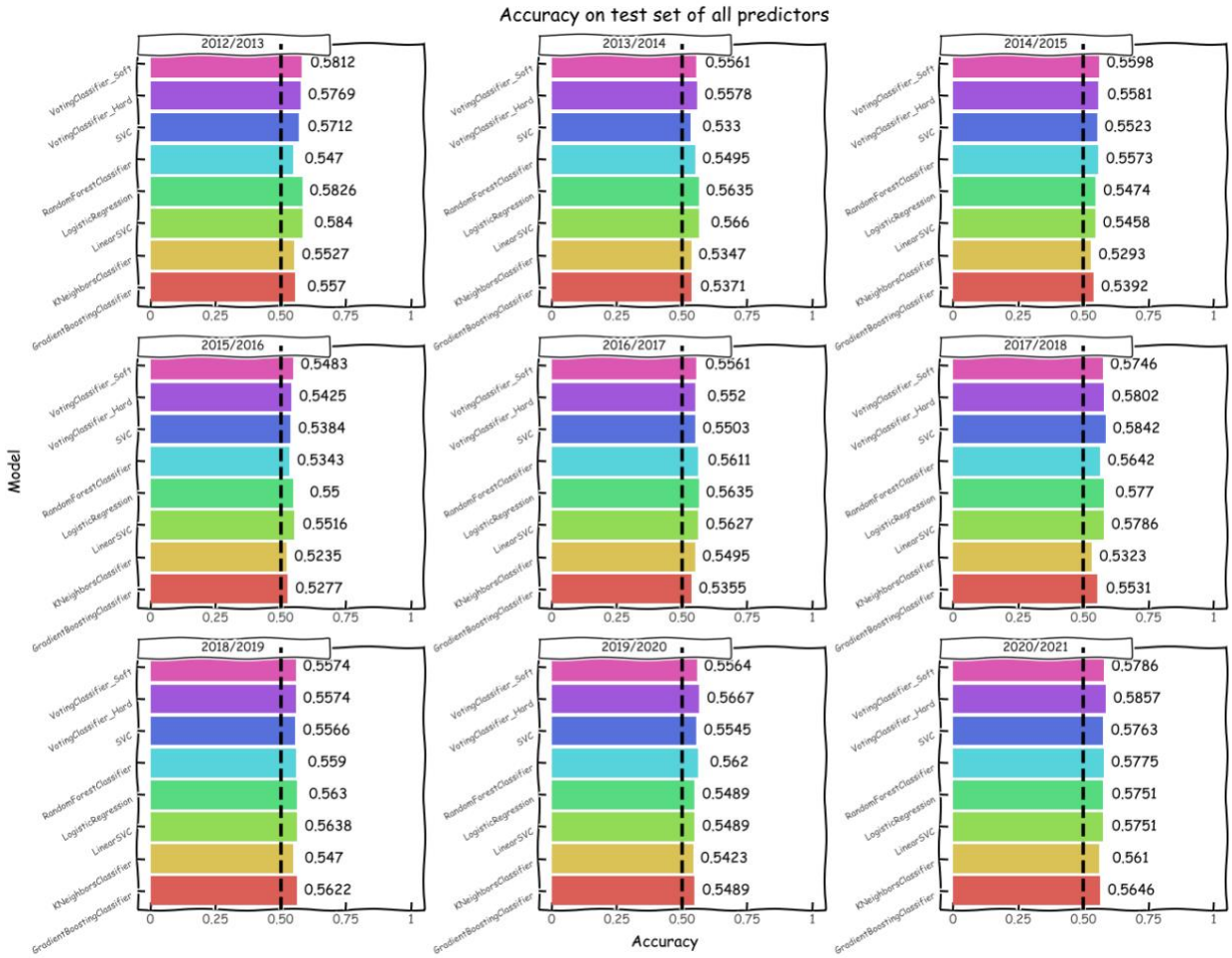
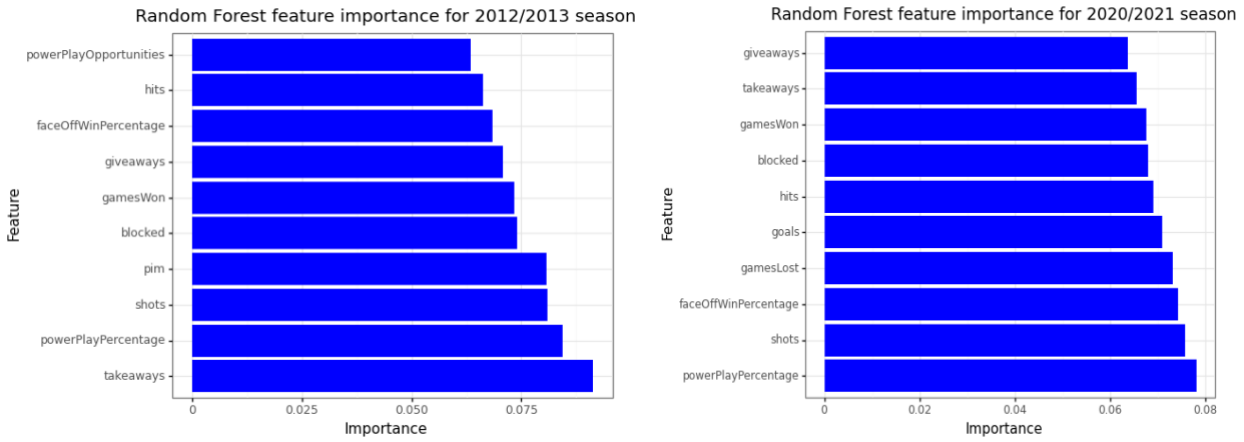


Figure 4: Test set accuracy for each model for each season. Text numbers represent the overall model accuracy.

seasons. The most accurately predicted seasons were in 2012/2013, 2017/2018, and 2020/2021. The highest model accuracy was the hard rule Voting Classifier in the 2020/2021 season with an accuracy of 58.57% followed by the Non-Linear SVM from the 2017/2018 season with 58.42% accuracy. The simple Logistic Regression model performed very similarly to the Linear SVM. No single model was able to consistently outperform the others on the test data. Additionally, more available training data (each increasing year had roughly an additional 1,200 games of training data) did not directly translate into better model performance.

Feature importance for the top 10 features of the Random Forest models for the 2012/2013 and 2020/2021 seasons can be seen in Figures 5a and 5b. The highest importance variable for the 2012/2013 season was the difference in number of takeaways while the highest importance variable for the 2020/2021 season was power play percentage.



Figures 5a and 5b: Top 10 Random Forest feature for each model.

Discussion

Simply put, predicting the outcome of an NHL game is hard. Through the analysis above, the max prediction accuracy we obtained through using largely traditional game statistics was 58.57%, well below the theoretical 62% upper limit mentioned earlier. This analysis gives more evidence that this limit does truly exist as this threshold was not crossed over any season analyzed. While all models were able to beat the 50% baseline accuracy for the test data, this threshold was often not beaten by very much. Additional training data did not have a noticeable impact on performance, evidenced by performance not improving as the seasons progressed and more data were available. In many cases, simple linear classifiers such as Logistic Regression and Linear SVM were able to outperform more complex ensemble models like Random Forest and GBM. In Figures 5a and 5b, we were able to see the feature importance for the Random Forest models in 2 separate seasons. The differing top features in each season give evidence to a slight change in the style of play as well as signifying what may be important factors in determining the outcome of an NHL game.

Ensemble models are a popular method of prediction under the theory that the average prediction of many weak predictors can equate to a strong predictor. Random Forest is a simpler example of this, although in this case it was prone to overfitting. The ensemble Voting Classifiers were able to often outperform individual models, including garnering the highest accuracy of all models in tested on these data. Neural Networks, popular models for complex problems, are often top performers in sports analytics, but they were not able to be implemented in this project due to technical difficulties.

The low prediction accuracies seen in all models gives more evidence that random chance (luck) plays a huge factor in the outcome of a hockey game give that the overall skill level between teams should be roughly equal due to the NHL's salary cap. A lucky bounce or goal can completely change the momentum of a game in favor of either team. Outside factors such as illnesses, injuries, and mental health/confidence issues are all difficult to quantify but can play a significant role in a team's performance.

In future work, more advanced hockey statistical features should be included. Most of the features used in this project were traditional hockey statistics, but in recent years, more advanced measures of team performance have been developed and would likely increase model performance. Using play-by-play data may also be beneficial to predicting a game outcome by updating probabilities based on certain events, but I believe that would be useful under a different context.

Overall, despite the disappointing accuracies, machine learning can certainly be useful in the context of sports betting. This project focused specifically on predicting all games of a season. However, a sports better does not, and mostly likely will not, place a bet on each and every game. When two equivalent teams play each other, the outcome is almost always uncertain. That falls more under the umbrella of gambling. Using machine learning models to predict probabilities would be the most effective implementation in betting. Focusing on games that have a high (home team wins) or low (away team wins) probabilities, and ignoring the more uncertain probabilities, would maximize the potential for profits. A further implementation of this would be to tune the models on these probabilities and implement a cost function to denote the behavior of winning or losing a bet. While it is likely though that such a dynamic game will never see overly accurate predicted outcomes, there is much room for improvement using data science and machine learning.