NHL Hockey Player Impact Significance To Outcome of NHL Match-Ups

Abstract

Sports betting is culturally ingrained into society. Sports betting is primarily done on professional sports such as NFL football, MLB baseball, and NHL hockey. These odds are calculated by bookmakers, bookies for short, and are determined from mathematical analysis of data from databases or datasets. However, these statistics calculated by the bookies, still lead to much uncertainty on the prediction. In this project a modelling method is proposed to predict NHL sports match-ups based on rigorous data from the year 2000 up until the year 2020. It will focus on hockey players' impact on match-up results and also include timezone difference analysis, to offset potential player fatigue with jet-lag. During this analysis, supervised learning using classification techniques will be used such as logistic regression and random forest tree techniques. It will then determine the odds for NHL sport team match-ups with a comparative analysis between logistic regression and random forest trees. This project has been able to yield a 74% accuracy of NHL match-ups using the random forest trees technique.

1. Description of Applied Problem

NHL hockey betting does not happen as frequent as other professional sports. In fact MLB games are bet on five times more than NHL games (Author, 2022). This could be the case where NHL games are harder to predict than other sports. According to many analysts, lack of information on current lineups and injuries leaves betters guessing (Whyno, 2021). There is lack of data points as compared to other professional sports, such as player tracking and puck tracking (Cotsonika, 2022).

There are several bookmakers that do analysis from databases, some are more accurate than others however. Most odds that come from bookmakers are odds compiled from outsourced specialists. Specialist companies compile odds for in-play or pregame markets which means that some sportsbooks may have near-identical odds to rival betting sites (Young, 2020). Since these odds rival betting sites, it is safe to assume eventually these calculated odds would become standardized. However, until then some of these sports' match-up odds can be calculated from just

each teams' latest match-ups, which can give calculated odds, but the odds are not necessarily an in-depth analysis, which is why bidders should choose their bookmakers wisely.

When it comes to predicting NHL matchups, the accuracy of the prediction is much lower than other major sports match-up accuracy, primarily due to how hard it is to predict an NHL game. From the University of Ottawa, a team of analysts were only able to predict NHL match-ups with an accuracy of 59% (Weissbock et al., 2013). If the odds of matchup prediction could increase in accuracy, perhaps NHL hockey sports betting could also increase too. However, as of 2022, the NHL league is putting forth puck tracking and player tacking measures to improve statistical analysis, which as a result will help improve NHL matchup predictions, but unfortunately won't be in this analysis (Cotsonika, 2022).

2. Description of Available Data

The available data in the dataset is vast and has a compilation of twenty years of data that includes every game outcome, all player statistics, and all team data from the years 2000 to 2020 (Ellis, 2020). It also includes in-depth stats, such as which player scored on which team on a specific day against a specific goalie. There are over five million plays recorded in this dataset and over twenty-six thousand game outcomes recorder. There is a lot of information that can be drawn from each individual play that can complete a full in-depth analysis of team match-up outcomes. However in this analysis, not all the data mentioned is going to be included. The key datasets that are going to be used, will focus upon the game team stats, which includes results of every game since 2000 to 2020. With it, it also includes the accumulated statistics of shots, blocked shots, face off percentage, etc for each game, for each team within a matchup as well as the game result. The game predictions include both regular season outcomes and playoff outcomes. To include player impact to the analysis, game skater stats will also be included. This entails goals, assists, penalty minutes, shots, blocked shots, takeaways, and giveaways for each player playing in each game. There is also just general stats of each game played. This includes venue timezones and when each game has been played, as well

as which season the NHL game took place in. This data is sufficient enough to help predict game outcomes.

3. Analysis and Visualization Techniques

3.1. Preprocessing phase

To have high quality results, preprocessing the data and cleaning the dataset should be applied in case of inaccuracy from human or computer error.

In this analysis, three primary datasets are going to be used. One dataset which outlines the match outcomes and match basic statistics such as team shots, team blocks, team hits, etc. A dataset that focuses on individual player statistics per game, and the last dataset that focuses on general game information such as venue timezone and the time of the game. All other data will be excluded.

The data structure for the classifier is going to analysis each game played in the NHL since 2000 to 2020. Each game analyzed will include the home and away team statistics. All columns analyzed per game are:

- away_settled_in (REG OR OT)
- away_head_coach (COACH NAME)
- home_head_coach (COACH NAME)
- type (R OR P)
- season (SEASON YEAR)
- time (24hr military time)
- game_id (Number)
- away_team_id (Number)
- away_shots (Number)
- away_hits (Number)
- away_pim (Number)
- away_powerPlayOpportunities (Number)
- away_faceOffWinPercentage (Number)
- away_giveaways (Number)
- away_takeaways (Number)
- away_blocked (Number)
- home_team_id (Number)
- home_shots (Number)
- home_hits (Number)
- home_pim (Number)
- home_powerPlayOpportunities (Number)
- home_giveaways (Number)
- home_takeaways (Number)
- home_blocked (Number)
- away_team_weight (Number)
- home_team_weight (Number)
- timechange (Number)

Majority of the data is numeric while the rest is categorical.

There were unfortunately a lot of anomalies, and 13659 games had to be dropped. There was game data included in the dataset with improper team ids or some games had to be continued and the stats or just null. There is also missing data values such as hits and blocked shots. Some of the statistics such as face off win percentage were not tracked until the year 2010. For complete accuracy, it was necessary to drop all of these games with missing or incorrect values. A KNN imputer could potentially fill in these gaps of missing information, but accuracy is key. Faceoff wins determine puck possession and puck possesion is key to success (habseyesontheprize). So all games from before the year 2010 were dropped.

The data for each game were tupled into two rows for both home stats and away stats respectfully. Joining both the tupled data (home team stats and away team stats) into one row was necessary with a prefix of home or away on data columns as seen in the columns analyzed above.

Adding customized features to help enhance accuracy seemed necessary. Added features to this analysis was to create a team weighted score. Add the time the game was played at and the timezone difference between venue and away team to correlate possible team performance.

To add the team weight score, every single player was grouped by their player id to average out their stats per game. So every single player would have their shots per game average, their hits per game average, their blocked shots per game average, etc. Then, from these averaged out stats, a scoring system was put into place to score how valuable that player is in general. The weighted factors to determine the players value was as follows (their average multiplied by a weighted factor):

- pointsPerGame * 1.5
- shotsPerGame * 1
- shotBlocksPerGame * 1.3
- puckTakeAwaysPerGame * 5
- puckGiveAwaysPerGame * -5
- plusMinusPerGame * 1.5
- penaltyMinutesPerGame * -0.01
- hitsPerGame * 1

Then for each game played, the data was then grouped on each player onto their

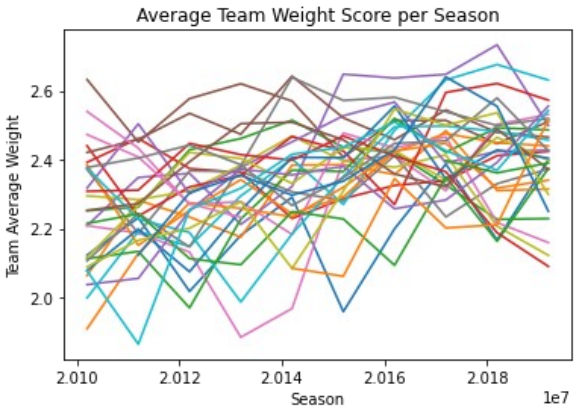
designated team and then onto the game id and an average team weight score was calculated by taking the mean of all the players weighted value grouped on their respective team. This gives an accurate weighted value per game for each away team and home team.

The time data within the dataset itself needed to be adjust for daylight savings time, so if the venue had a DT timezone, it was necessary to correct it. Then to add the timezone offset, it was necessary to compare the timezone of the away team and the home team (venue). The time offset difference should help determine possible team performance differences under time conditions that aren't typical for the respective team.

3.2. Analysis phase

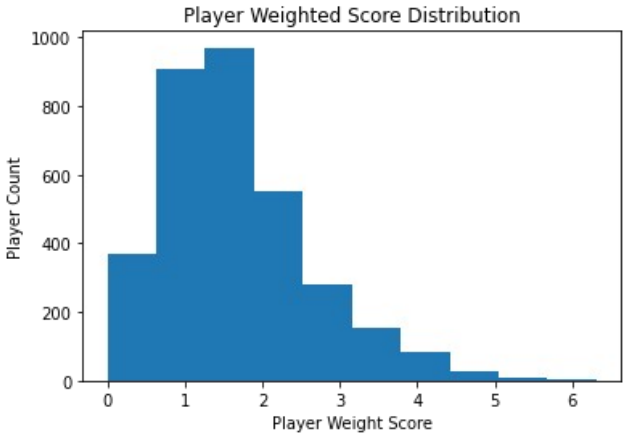
The goal of the project is to predict a game match-up accurately while also proving that certain players do impact match-up results. Certain NHL players are considered game changers and it can be identified through team weighted averages. Since about half of the team game data has been dropped due to insufficient data, the years 2010 – 2020 of every team have their team weight scores calculated in Figure 1.

Figure 1.



This data is interesting as there seems to be a general increase in weighted scores over the 10 year period of time. This implies that the weighted scores set to the players in this analysis may be a favourable assessment for what team management looks for in players and builds their team in favour of these desired attributes that the players have been weighted on. On this conclusion, the team weights may be an effective measuring feature.

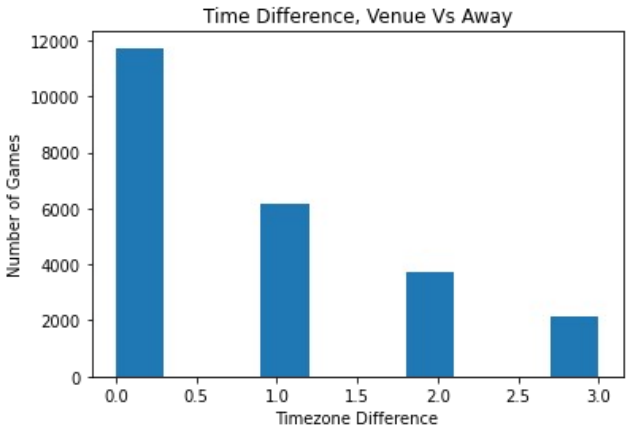
Figure 2.



In Figure 2. we can see the general distribution of player weights, where it is evident that superstars do exist and have a major contributing factor to team performance. In the top five players assessed with a weighted score, it consisted of major NHL AllStars, with Alex Ovechkin being the biggest impact player, following was Pavel Bure and Mario Lemieux who have both been inducted into the hockey hall of fame. With these All-Star and hockey hall of fame players being within the top weighted scores ranks, it is fair to assume that the weighted score analysis may be a key feature to add to the classifier.

The NHL games that happen between teams occur over the whole north American continent. There are timezone differences between away games and home games occasionally. This could impact team performance, as NHL players can experience jet-lag or fatigue. In Figure 3., we can see how many games occur in different time zones.

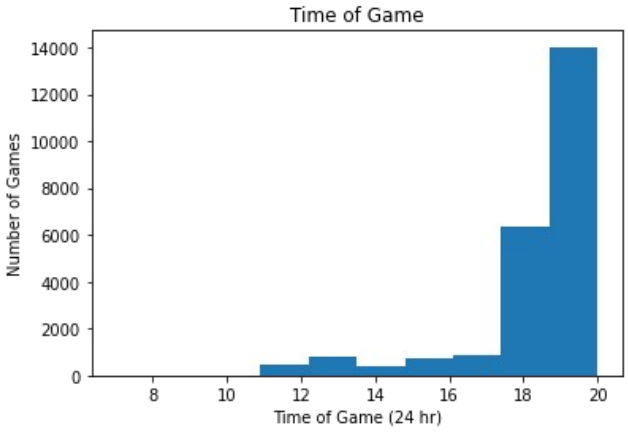
Figure 3.



Out of the total twenty-three thousand games being measured, a little bit more than two thousand games have been played that have a three hour timezone difference between the away team and home team venue. Which is about nine percent of the games. Fifteen percent of NHL games recorded have a two hour difference, with twenty five percent of the games having a one hour difference. In total, just over fifty percent of all NHL games in this dataset have been played with a one hour timezone difference or greater. That is

a pretty significant factor to take into account.

Figure 4.



In Figure 4. it shows the amount of games taken place at their specific GMT time. Teams being accustomed to specific game times could offset their performance. With game times altering, with the timezone differing between home and away teams, it could significantly impact how teams perform. Knowing when the game time starts, could help make the timezone difference feature more significant.

These three added features should help contribute to help predict NHL team match-ups more accurately.

3.3 Visualization phase

This analysis included two classifiers for comparison. It used logistic regression and random forest trees.

With the logistic regression, the columns of numeric data were standardized and the categories were hot encoded.

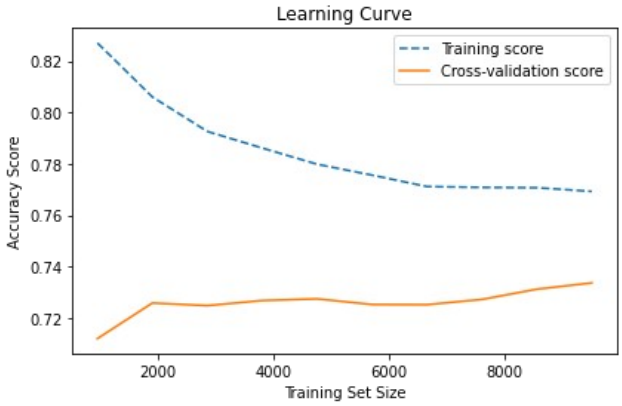
The logistic regression classifier has used the holdout evaluation technique, with the data being split between two categories. The test data and training data. The test data size is equalling thirty percent of all data in the dataset, and seventy percent of the data, or the rest of the data is for training.

The logistic regression classifier used the newton-cg solver for the gradient decent and then also used the L2 regularization to try to avoid over-fitting and penalize coefficients.

The results for the logistic regression yielded good results as compared to the 59% accuracy score benchmark performed by the University of Ottawa with a 74% accuracy. As seen in Figure 5., it can be

seen where the training score and cross-validation score is plotted over time. This is also known as the learning curve, where it gives a good idea of whether the classifier is either over-fitted or under-fitted. The gap between the training score and cross validation score near the end indicates the lines haven't converged, which then considers this neither over-fitted or under-fitted, but right at the "sweet" spot. It can be argued that more training data would be needed to help increase cross-validation score, but there is no more data to train. An approach of using the sag solver with L1 penalization was used to try to reduce some coefficients so it may fit better, but ended up being nearly the same results with a little less effectiveness resulting with a 69% accuracy. Looks like it is at a good spot to make NHL matchup predictions.

Figure 5.

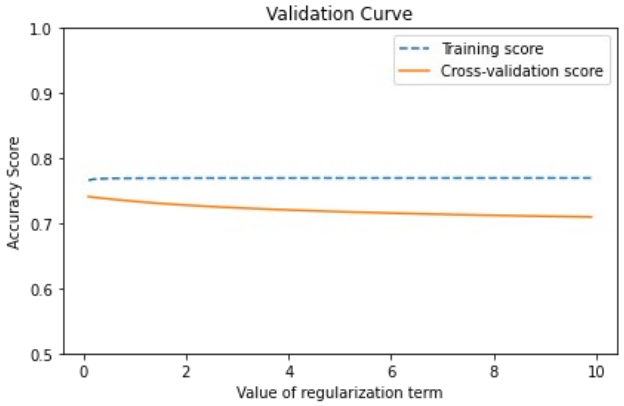


The confusion matrix of the logistic regression classifier also looks as so:

TP= 4837	**FP**= 1603
FN= 1939	**TN**= 3521

With these results, several statistics can be calculated. The confusion matrix determines that the accuracy score is 70%, recall score is 71.4%, precision score is 75.1%, sensitivity score is 71.4% and specificity score is 68.7%. Fairly good.

Figure 6.



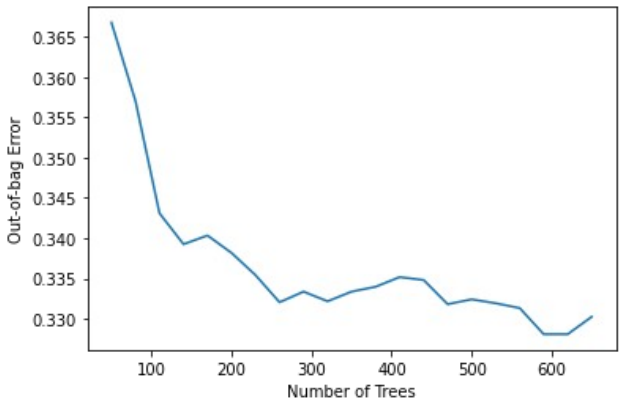
In Figure 6. we can see the validation curve which is another measure of over-fitting and under-fitting. Since the lines are almost parallel, we can see that there are no hyper-parameters of the model making it less accurate. The log loss on this model is quite high, with a log loss of 10.196. This does make sense as it confirms how hard it is to

predict an NHL game. Not all features are necessarily relevant for a team victory. Do to the uncertainty of NHL games, the model is still well configured.

By curiosity, a random forest tree classifier on the same dataset has been tried. This was done to just see if a different classifier can yield better results. The data has been treated the same, with standardizing the numeric values and hot encoding the categories. The data was also split 30% for test data and 70% for training data.

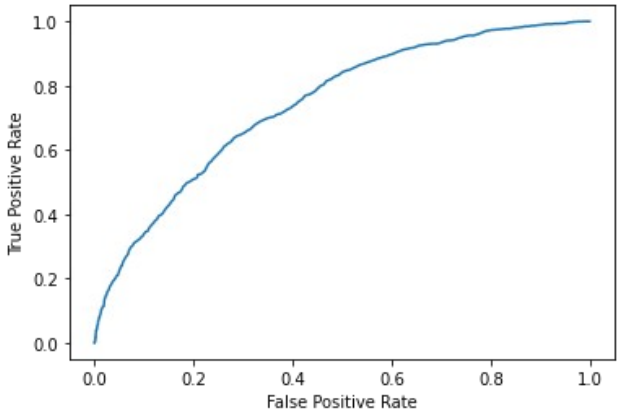
The approach of the random forest tree classifier will start with fifty tress and then add thirty trees iteratively, until there are six-hundred-fifty trees. While doing so, error tracking of the out of bag error will occur. This approach will help find where the error rate will bend to get a good estimate of when the model is fairly trained and will give good training results.

Figure 7.



This random forest tree was able to also achieve a 74% accuracy rate. But doesn't look like it can improve much more as seen in Figure 7. Considering logistic regression classifier managed the same AUC score of 74%, the random forest tree classifier at this point is assumed to be well trained.

Figure 8.

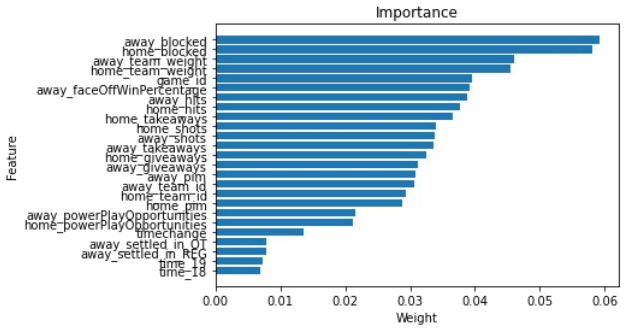


The AUROC score in Figure 8. isn't the greatest to determine separability, but does imply again that hockey is a little hard to

predict. Overall not too bad when comparing to the benchmark of University of Ottawa.

In Figure 9., it was good to know if the added features impacted the decision making process in the random forest classifier. It turns out the team weights for both the home and away teams were very significant factors, ranking in third and fourth as the most important features. The timezone feature ranked twenty-first in importance whereas the time of game feature didn't fall to far behind the timezone. Surprising the timezone difference isn't too much of a significant impact.

Figure 9.



The random forest tree classifier did outperform the logistic regression just slightly yielding an AUC of 74.3% whereas the logistic regression yielded a 74.1%. Considering the team weighted feature was classified as an important feature shown by the random tree classifier, proves that player impact does matter to the team outcome during the matchup. The random tree classifier also beat the benchmark done by the University of Ottawa by nearly 15%. Overall, these models are well trained and ready to predict NHL matchups.

4. References

Ellis, M. (2020, December 11). NHL Game Data. Kaggle. Retrieved May 27, 2022, from https://www.kaggle.com/datasets/martinellis/nhl-game-data

Weissbock, J., Viktor, H., & Inkpen, D. (2013). Use of Performance Metrics to Forecast Success in the National Hockey League. CEUR Workshop Proceedings. Retrieved May 27, 2022, from http://ceurws.org/Vol-1969/paper-06.pdf

Young, J. (2020, October 30). How do bookmakers create sports odds? Bookies.com. Retrieved May 27, 2022, from https://bookies.com/guides/howbookmakers-create-odds

Author, G. (2022, March 22). How popular is NHL Betting in 2022? The Win Column. Retrieved August 6, 2022, from https://thewincolumn.ca/2022/03/21/how-popular-is-nhl-betting-in-2022/#:~:text=Plenty%20of%20people%20bet%20on,than%20the%20average%20NHL%20game.

Boyle, C. (2013, April 4). Why possession matters: A visual guide to fenwick. Eyes On The Prize. Retrieved August 6, 2022, from https://www.habseyesontheprize.com/2013/4/4/4178716/why-possession-matters-a-visual-guide-to-fenwick

Cotsonika, N. J. (2022, March 9). NHL Puck, player tracking honored at Sloan Sports Analytics Conference. NHL.com. Retrieved August 6, 2022, from https://www.nhl.com/news/nhl-puck-player-tracking-honored-at-sloan-sports-analytics-conference/c-331649276

Whyno, S. (2021, May 12). As sports betting expands, hockey's secrecy makes wagering on NHL a gamble | CBC sports. CBCnews. Retrieved August 6, 2022, from https://www.cbc.ca/sports/hockey/nhl/nhl-betting-secrecy-1.6024319