CAPSTONE FINAL REPORT

A Individualized Approach to NHL Betting Analytics

Samuel Wilson 10160657

DATA 501 | W 2020 | University of Calgary

1 Introduction

1.1 Goal

Sports betting has been around since the time of the Ancient Greeks, where spectators would place wagers amongst themselves on the outcome of Olympic events in hopes of profit. Although much time has passed, the goal of the sports better is all the same: How can I effectively guarantee profit? Through the use of purposeful feature selection – in a method I refer to as 'strength of ratio' – and the implementation of logistic regression my project aims to answer this question. More specifically, through the lens of an NHL team. In addition, this paper serves to identify which particular attributes have the greatest impact on the outcome of a game and ultimately, assess the logistic regressions ability to predict the future from past events. To do so, we will be looking solely at place-line bets, which is simply a bet made on whether a team wins or loses, obviously only paying out if you guess correctly. Our measure of success will be if the model can accurately predict >50% of the games. Whether you are motivated simply by fiscal reasons, or if you are simply interested in NHL team analytics, this work serves to benefit a multitude of people. After all, who doesn't like money?

1.2 Previous Work

The difficulty with predicting the outcome of NHL games stems from a variety of factors. One such challenge is that players are constantly entering and exiting the league, effectively changing our population and its size with each season. But perhaps the most complicated of the challenges arises specifically within the team aspect itself in the form of confounding variables that are very hard to attribute for; such as luck, psychological well-being, refereeing, etc. It is much easier to predict the outcome of a single individual than that of a team, comprised of several individual efforts. Therefore, increasing the variability of the team's success. All of these factors of course, hinder the accuracy of the model, not only by the longevity of its application, but also bit its ability to predict trends from years past. Given the complexity of the sports betting world, and the multitude of sports and events that can be bet on, there are a plethora of models that could be derived. One such individual was Tuan Doan Nguyen, a data scientist at Quora Inc. He developed a Machine Learning (ML) – free based algorithm to predict the Premier League (soccer) results using a simple Poisson process and his model was able to accurately predict the outcome of 64.1% of the matches tested. Another individual, Jordan Bailey, on the other hand, used a ML-based algorithm that used previous box score statistics to predict the OVER/UNDER on future basketball games. He developed one OVER model and one UNDER model which was able to correctly predict 59.09% of and 54.16% of matches tested, respectively. Hockey on the other hand has very little ML based work, and this is probably attributed to the abundance of factors specific to hockey that other, more heavily researched sports, don't share. These factors include, but are not limited to shorter shifts, higher speeds, and the allowance of greater contact and fighting. The most notable work being from the University of Science and Technology, Beijing, where Wei, Rozann Whitaker and Thomas Saaty utilized a hybrid method that uses both data and judgements by the 'experts' to make predictions. Through the implementation of a support vector machine and an ANP network, this allowed them to predict the outcome of 89 postseason NHL games with an

accuracy of 77.5%. These approaches, although impressive, have inherent assumptions that seem problematic. Nguyen's Poisson assumption makes a bold claim about the nature of the game of soccer, implying that one minute of play has no effect on the following minutes of play. Bailey's model, on the other hand, succeeded in beating the average (>50% accuracy) only after the addition of a confidence threshold, which in itself was only 62%, adding yet another layer of uncertainty. Whereas Gu, Whitaker and Saaty rely heavily on the opinions of people, making the assumption that said experts are right the majority of the time – which we have seen in real world application, is not the case. Where my method differs is this: my individualized approach to predictive statistics, my proposed 'strength of ratio' formula, and accessibility of data. All of which I will explain further in the following section.

1.3 Approach

Past approaches try to assess the potential success of a single team based on the league as a whole and make predictions on all games accordingly. Although, this method is valid, depending on our goals, it may not necessarily be the best approach. We can think of the success of a team as a measure of the percentage of games they win in a season. Now, let Team A and Team B both have a winning percentage of 0.500, meaning they win half the games they play. This would mean both teams are equally successful, but this does not necessarily mean their methods of success are similar. For example, say Team A wins every game 1-0 and Team B wins every game 10-9. The point differential is the same, but Team A's wins will be more highly correlated with defense/goaltending, and Team B's wins will be more highly correlated with offense. Thus, their methods of success differ drastically. My approach aims to overcome this hurdle by analyzing just one team at a time, and what variables have the greatest correspondence to the outcome of their games specifically. After all, our goal is simply to be profitable. One does not need to know and bet on the outcome of every single NHL game to make money. For example, would you rather bet on 3 games a night with a 60% accuracy, or allocate said money to one bet with a 90% accuracy? I know what I would choose. My hope is by limiting our regression to one team, we can increase the accuracy of our model, and effectively strengthen our betting strategy by increasing our confidence in bets placed. Ultimately, by doing so, one can hope to be more profitable while maintaining a higher level of confidence. The Calgary Flames are everyone's favorite hockey team, so it seemed only fitting that they be the subject of my testing for the remainder of the experiment.

2 Methodology

The following was achieved in Excel and Python, with the emphasis of time spent in Python using the packages pandas, math, numpy, seaborn, matplotlib, scipy, and sklearn.

2.1 Data Wrangling

The first step was to acquire the necessary data for analysis. Thankfully, Hockey-Reference.com has a massive database of statistics for all NHL teams dating back to 1917. For the sake of this report, we look only at the last 5 years for the aforementioned reasons related to reducing bias, variability, and inaccuracy. I was able to download the box scores and related team statistics for every Flames game since the beginning of the 2014 season, for a total of 410 games with 28 corresponding features, in the form of .txt files directly into Excel. Of those 28 features, 18 corresponded to the flames individually, 6 to their opponent, and the remaining 4 shared between them. Here, as you can see in Figure 1, some small changes were made to the original header descriptions in order for pandas to later be able to use and differentiate between the columns. Once in python, pandas was used to add the power play percentage (PP%) and penalty kill percentage (PK%) variables, which is defined as 'PPG' divided by 'PPO' and 'PPGA' divided by 'PPOA', respectively.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1										Team	Team	Team	Team	Team		Opponent	Opponent	Opponent	Opponent
2	GP	Date		Opponent	GF	GA				S	PIM	PPG	PPO	SHG		S	PIM	PPG	PPO
3	1	2018-10-03	@	Vancouver Ca	2	5	L			35	7	0	7	0		23	19	0	
4	2	2018-10-06		Vancouver Ca	7	4	W			37	12	3	6	1		20	12	2	
5	3	2018-10-09	@	Nashville Pre	3	0	W			27	8	2	4	0		43	8	0	
6	4	2018-10-11	@	St. Louis Blu	3	5	L			34	14	0	4	0		32	10	2	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	GP	Date	HomeAdv	Opponent	GF	GA	Outcome	S	PIM	PPG	PPO	SHG	SA	PIMA	PPGA	PPOA	SHGA	CF	CA
2	1	2018-10-03	0	Vancouver Ca	2	5	L	35	7	0	7	0	23	19	0	1	0	58	2
3	2	2018-10-06	1	Vancouver Ca	7	4	W	37	12	3	6	1	20	12	2	6	1	52	3
4	3	2018-10-09	0	Nashville Pre	3	0	W	27	8	2	4	0	43	8	0	4	0	36	5
5	4	2018-10-11	0	St. Louis Blu	3	5	L	34	14	0	4	0	32	10	2	6	0	47	3

Figure 1: Screenshot of dataset in Excel before and after 'cleaning'; used to show change of variable names and addition of others.

2.2 Data Analysis

The logistic regression, although less demanding than the linear regression, still has some essential assumptions that must be met. Those being (i) that the data is free of missing values, (ii) that the dependant variable is binary, (iii) independence of predictors, (iv) little no to multicollinearity between independent variables and (v) has a large sample size. The choice to use logistic versus linear was based on the categorical nature of our dependant 'Outcome' variable, the robustness of the model to handle outliers and different data types, and its ability to bypass assumptions such as normality or homoscedasticity. The next logical step was then to confirm these assumptions through statistical analysis. Starting with the first assumption, all columns, with the exception of a few in 'PP%' and 'PK%', were found to be nonempty. The empty columns were a result of my definition of the variables, which gave us a division error when either team did not have a power play opportunity in said game. To rectify this, the mean of each column was calculated, and the missing values were then replaced with this new value. Given the large dataset size, this seemed more appropriate than abolishing the rows all together. Continuing on, as seen in Figure 2 and Figure 4, assumption (ii) was verified easily via a count plot and assumption (iii) was verified via scatter plot, respectively. As for assumption (iv), the Spearman correlation coefficient was calculated for every feature in the dataset.

Spearman was used instead of Pearson to overcome the assumption of normality and because of the ordinal nature of our dependent variable. Taking *r* = 0.50 as the threshold for 'too much' multicollinearity, only features who fell below this mark were to be used in the model. In this way, as shown in Figure 3, almost all values fell into this category, and as long as we didn't use a combination that rose above this value, our assumption is met. The last assumption was verified after our selection of variables, as the required sample size changes accordingly with the number of predictors included in the model. The rule of thumb being, your dataset size must exceed the value given by the number of independent variables with at least 10 entries divide by the least probable outcome. For example, if the least probable outcome is 0.5, and you have 5 independent variables, then your dataset size must exceed: (10*5/0.5) = 100 entries. I go into more detail on this in the next section.

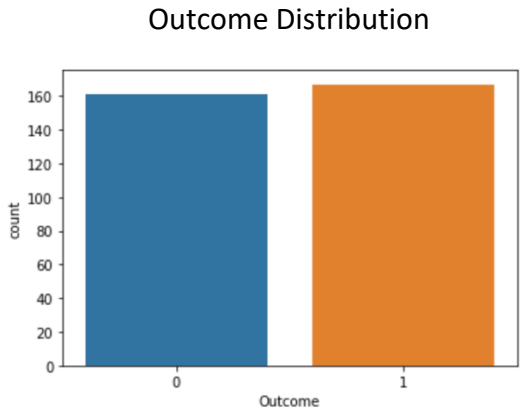

Figure 2: Count plot of the distribution of the variable 'Outcome'; used to confirm that our dependant variable is binary, taking on only two values.

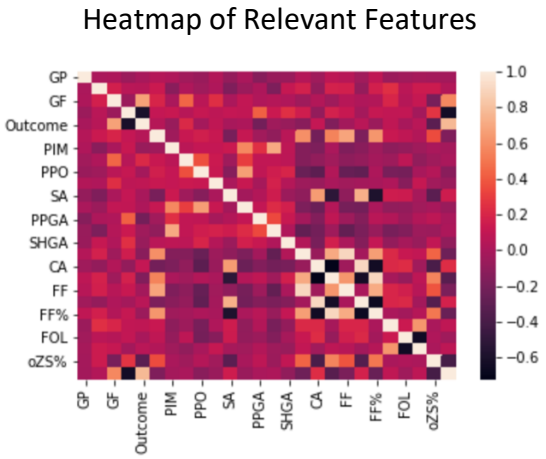

Figure 3: Heatmap of spearman correlation coefficients between all 28 features of interest; used to check for correlation to the variable Outcome and to check multicollinearity.

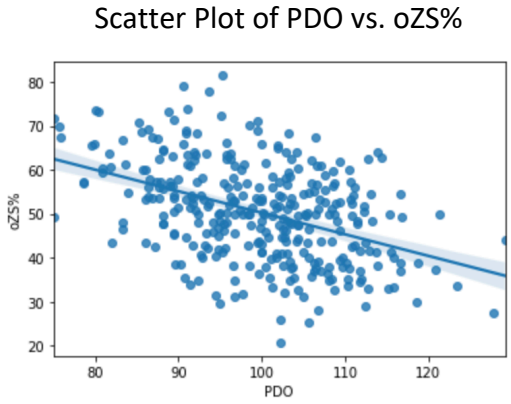

Figure 3: Example of one of the scatter plots created, PDO vs. oZs%; used to assess the independence of the predictor variables.

2.3 Ratio of Strength

The first method implemented was to simply choose independent variables based on their relative spearman correlation to the dependent variable: 'Outcome'. Setting a threshold at r=0.2, anything that fell below this was considered to not have a significant effect on the outcome of a hockey game, leaving us with 8 variables to choose from. At this point, our last and final assumption (v) is met, as with a maximum of 8 variables, we will need a maximum of 160 entries, which we are well above. As for variable choice, at first, it seemed intuitive to use the variables with the highest spearman coefficient, but after several model compositions, the accuracy of the model was still undesirable and I began considering alternative approaches – quite literally, in the form of trial and error. It was this experimentation with parameters that led me to the serendipitous derivation of the method I have dubbed 'ratio of strength'. For all independent variables included in the model, the relationship is defined as follows:

$$\rho_{ROS} = \frac{Spearman\ Coefficient\ to\ Dependent\ Variable}{\sum Spearman\ Coefficent\ to\ Independent\ Variables}$$

with larger numbers corresponding to more powerful predictors. By this method, the top 5 variables were chosen to best represent the model. The exclusion of the other three was a personal choice based on prior knowledge of the game and its aforementioned confounding factors.

2.4 Logistic Regression

Once the independent variables had been chosen, sklearn was used to split the dataset, train it and eventually build the resultant model. The logistic model was trained on the first 229 games and tested on the last 99. In the next section, we will dive deeper into the success and accuracy of the model.

3 Results

Using our 'ratio of strength', data analysis and logistic regression, I predicted the outcomes of 99 Calgary Flames NHL games with an accuracy of 97.97%, which is extremely high in comparison to past attempts aforementioned in section 1. Of the 99 games, 44 games were won and predicted correctly, 53 games were lost and predicted correctly, and the remaining 2 games were won and predicted incorrectly. The results of the confusion matrix and its resulting accuracy score can be seen below in figures 5 and 6.

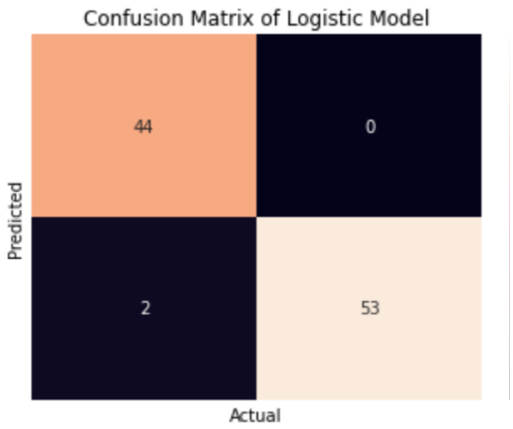

Figure 5: Plot of confusion matrix with results of logistic regression; used to show the accuracy of the model and the counts on each result.

```
accuracy_score(y_test, predictions)
```

```
0.9797979797979798
```

Figure 6: Screenshot of accuracy score from in-code; used to further verify the results of Figure 4.

Interestingly enough, the independent variables that proved for the best result (PPG, PDO, GF, oZS%, PPGA) were not indicative of my earlier assumptions prior to experimentation. Most surprisingly so was PDO, oZS% and PPG. Based on previous knowledge of the game, I had assumed these stats would play a much smaller role in the derivation of this logistic model; further reaffirming my earlier criticism of judgment-based models. Based on our earlier definition of success – accurately predicting more than 50% of the games – I would conclude that the experiment was a success, surpassing this benchmark by a whopping 47.97%. Thus, according to these results, in comparison with the standard league wide analyzations, the individualized team approach appears to be superior in determining the outcome of an NHL hockey game. In terms of profit, with a 97.97% accuracy, it seems near impossible not to be profitable adopting such a method.

4 Discussion

Overall, this result is very promising, especially considering the degree of accuracy by which we have surpassed both our success threshold, and the accuracy levels of past approaches. For this reason, it is hard to suggest an alternative method. With that being said, Hockey-Reference.com supplies a vast majority of stats, both individual and team based. An interesting alternative approach may be to draw from a different pool of statistical features, different than our own, and see if the accuracy of the model could be even further strengthened. Moreover, the algorithm we used select our independent variables, to my knowledge, has not been used before. In this way, further research would have to be done to ensure the validity of the algorithm's application to other fields of interest. It may be of interest to use this method on other teams within the league and see how their variables of interest matchup against ours within the same pool of 28 features we began with. Are they the same, different, or do they

share commonalities between them? As for those, who's knowledge in statistical theory is greater than mine: I challenge them to prove or disprove the strength and utility of my algorithm.

5 Conclusion

I had sought out to develop a model that could accurately depict the outcome of NHL hockey games and effectively ensure profit when placing place-line bets. I did so through the individual analyzation of the Calgary Flames team, and the purposeful selection of independent variables using the 'ratio of strength' method outlined in section 2. We defined a successful model as one that could accurately predict said outcome more than 50% of the time. Through the logistic regression of 410 games dating back to 2014, this criterion was exceeded drastically by the model, using 5 predictors from our dataset and reporting an accuracy score of 97.97%. Through the course of this project I both developed existing coding skills, as well as picked up many new ones. In terms of technical skills, I learned how to use pandas to analyze and manipulate datasets, use seaborn to plot and visualize results, use sklearn to train a logistic model, as well as how to interpret the results of confusion matrices and plots. Not to mention, expanded my knowledge of both supervised and unsupervised machine learning, learnt how to better interpret and understand scientific papers, and (hopefully) how to write a better paper. This is clearly beneficial to myself and my education, but my real hope is that this may provide insight for my fellow peers into how they select their independent variables and how my individualized method can provide perspective on how we analyze data.

6 Future Work

As briefly discussed in section 4here are a variety of ways future research and development could be made. I have outlined them as follows:

(i) Short term – Apply this same method on other predictive variables available on Hockey-Reference.com or Corsica.com and see how the accuracy of the models compare.

(ii) Medium term – Apply this same method to other teams within the NHL league and see how the predictive power of the 'ratio of strength' holds up.

(iii) Long term – Delve deeper into the statistical theory and legitimacy of my 'ratio of strength' method in order to assess the robustness of its application.

References

E-Books

Machine Learning with MATLAB. The MathWorks, Inc., 2020.

Journal Articles

Jordan Bailey, **Applying Data Science to Sports Betting**. *Towards Data Science*, 2018.

Tuan Nguyen, **'Making big bucks' with a data driven sports betting strategy**. *Medium,* 2019.

Wei Gu, Thomas L. Whitaker, Rozann Whitaker, **Expert System for Ice Hockey Game Prediction: Data mining with Human Judgment**, *World Scientific*, 2016.

Websites

Towards Data Science – https://towardsdatascience.com

Medium – https://medium.com

Hockey Reference – https://www.hockey-reference.com/leagues/NHL_2020.html

Corey MS – https://coreyms.com

Corsica – https://corsicahockey.com

Edureka – https://edureka.co