Capstone Project 1 Final Report

Predicting NHL game winners with machine learning

Kevin Hill

Introduction

The National Hockey League (NHL) is often considered the smallest of the "Big 4" major North American sports, also consisting of the National League Football (NFL), Major League Baseball (MLB) and National Basketball Association (NBA) [1]. The NHL claims the lowest revenue and smallest fan base of those sports, but that has never deterred me from being a huge fan. Perhaps the smallest status has kept sports analytics from becoming a larger part of the NHL. On the other hand, it may be the speed and dynamic action that make it difficult to apply analytics to the sport. In any case, analytics is a relatively small aspect of NHL hockey, but is definitely becoming more popular and important.

For those who are not fans of the sport, I shall present a very brief summary of the sport. Hockey is played (in the NHL, on ice) by teams, who are allowed 5 skaters and 1 goaltender per team to be active at any given time. Each player uses a stick with a short bend at the end, shaped like the letter "L", and the object is to put the puck, a small rubber disk, into the opposing team's goal. The team scoring more goals when regulation time expires is the winner. A game is 60 minutes long, divided into 3 periods of 20 minutes each, with a short overtime if the teams are tied at the end of the game. When a player is penalized, that player is not allowed to participate for a short time, and that player's team is short a player as a result. This situation is called a power play for the team with the larger number of skaters, and a penalty kill for the penalized team. I enjoy the fast pace and fluid action of a hockey game and attempting to apply data science to the game is what inspired me to this capstone project.

The goal of my project is to predict the winner of an NHL game, specifically by attempting to recreate the results of work done by sports analysts at the University of Ottawa [2]. They reported an accuracy in their prediction model of roughly 60%. A secondary goal is to determine which statistics have the most influence over which team wins a game. I approached this project with the intention to find ways for a team to improve its chance of winning a game. Thus, the primary clients I envisioned when planning this project include any NHL team, as well as other hockey teams outside the NHL, such as minor league teams linked to the NHL, for example the American Hockey League (AHL), and national hockey teams outside North America. Since gambling often involves choosing the winner of major league sports events, this information could also benefit gamblers betting on games.

Data Wrangling

The dataset I am working with was provided by the original researchers. In their published document, they noted their data would be available if they were contacted by email. The data consists of basic statistics pertaining to a hockey team, such as total number of goals scored and team shooting percentage, as well as derived statistics, such as Fenwick close, which attempts to measure puck possession via shots, all defined in Figure 1.

Fenwick Close	Used to estimate puck possession, derived from team shots for versus team shots against
GF	Total team goals scored for
GA	Total team goals scored against
GIDiff	Difference between goals for and goals against
PP%	Percent of power plays team scored during
PK%	Percent of opponent penalties killed off
sh%	Total team successful shooting percentage
sv%	Total team save percentage
PDO	Used to measure "luck", derived from shoot % and save %
win streak	Number of games team has won consecutively
standing	Team ranking in their conference
5-5 F/A	Ratio of goals scored for versus goals scored against when both teams have 5 skaters

Figure 1. Definition of statistics used

The data set also contains statistics gathered after each game, presumably aggregated into the totals, and betting lines on each game. The original authors didn't use any of those statistics, and I also removed them from the data during wrangling.

My original plan was to build a baseline model using regularized logistic regression with hyper-parameter tuning. It is a logical model to use, because this problem is a binary classification problem, and the logistic regression model is designed to address that type of problem. Logistic regression attempts to construct an equation of sorts, in which each variable is assigned a weight, or importance, toward the outcome. In this study, the outcome is either the home team wins or the home team loses. Binary classification refers to a problem that can be observed by organizing data into classes, and binary means there are only two possible classes, in this case home team wins or home team losses. So the goal of this project is to classify unknown games as one of those two outcomes. The data was organized in a manner

that was convenient for their model but was not as convenient for mine. For their study, each game was comprised of two blocks of statistics, one for the home team and one for the away team, and I preferred the data for each game to be merged into a single block containing all the game information. Thus, the bulk of my data wrangling consisted of renaming features to distinguish home and away values, then rearranging the data to facilitate working with my model. Apart from contacting the original analysts, the data is also available through various websites, including NHL.com and Corsica.hockey, but the data from the authors has some preliminary calculations already done, saving me some time for this project.

Exploratory Data Analysis

To begin my preliminary exploration, I examined the summary of the statistics to look for any immediate potential issues. I found the away team power play maximum value and penalty kill minimum value appeared unusual. The high power play number was almost 80%, while the low penalty kill was under 30%. If the data started at the beginning of the season, those numbers would not be strange, but close to one third of the way in, they seemed a little out of place. Therefore, I created some violin plots, showing density of data points, with the home team data on the left and the away team data on the right, to see if visually anything would clarify the apparent anomaly. The plots for away power play and penalty kill did in fact seem to verify something possibly incorrect. Another plot, home team win streak, also looked like it might need further investigation. The plots shown on the top, in figure 2, were created from the initial data import. The spikes in the top right, middle left, and bottom middle plots were what I was inspecting. After doing some research, I found that in one of the recorded games, the values for the away team's penalty situation appeared to have been switched. The away team in that game had their power play rating switched with their penalty kill rating. The bottom middle plot ended up being correct but needed a closer look. It was due to a single team having an extended winning streak. The corrected plots are shown on the bottom, in figure 3. The only other apparent difference is in the PDO plot. The home teams show a little indent under the mean that the away teams don't seem to show. Perhaps home teams exhibit slightly less "bad luck" than away teams. Other than those situations, visually the data appears very similar for both home and away teams.

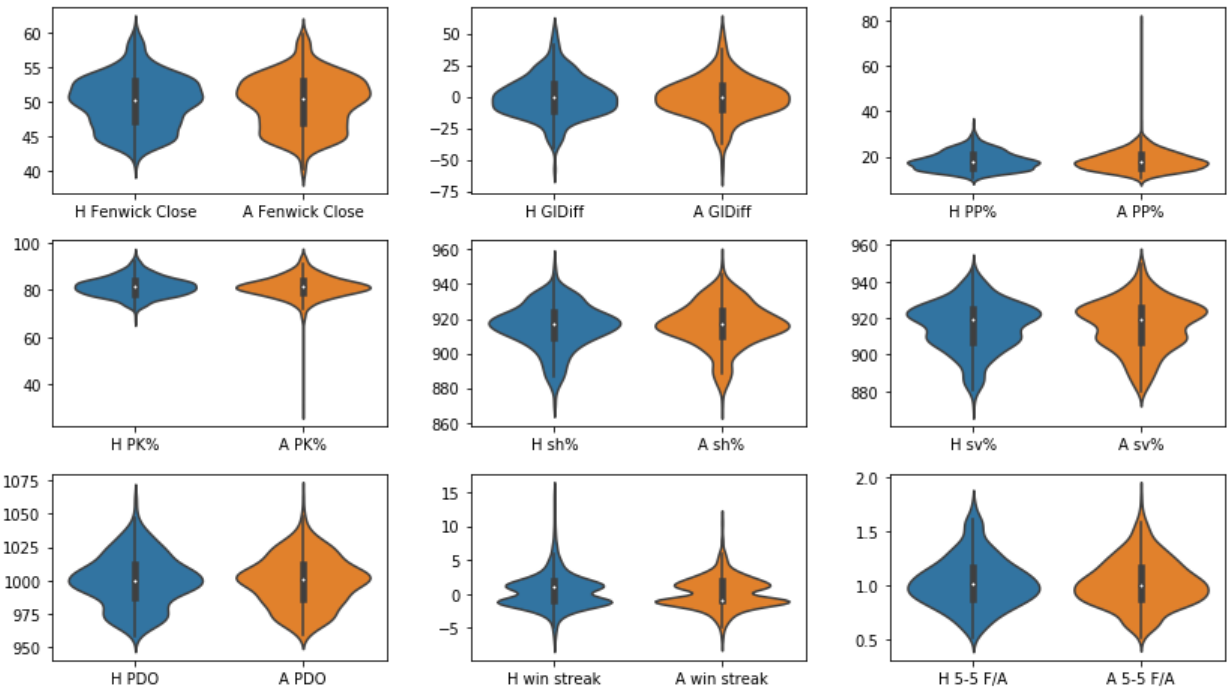

Figure 2. Violin plots of original data

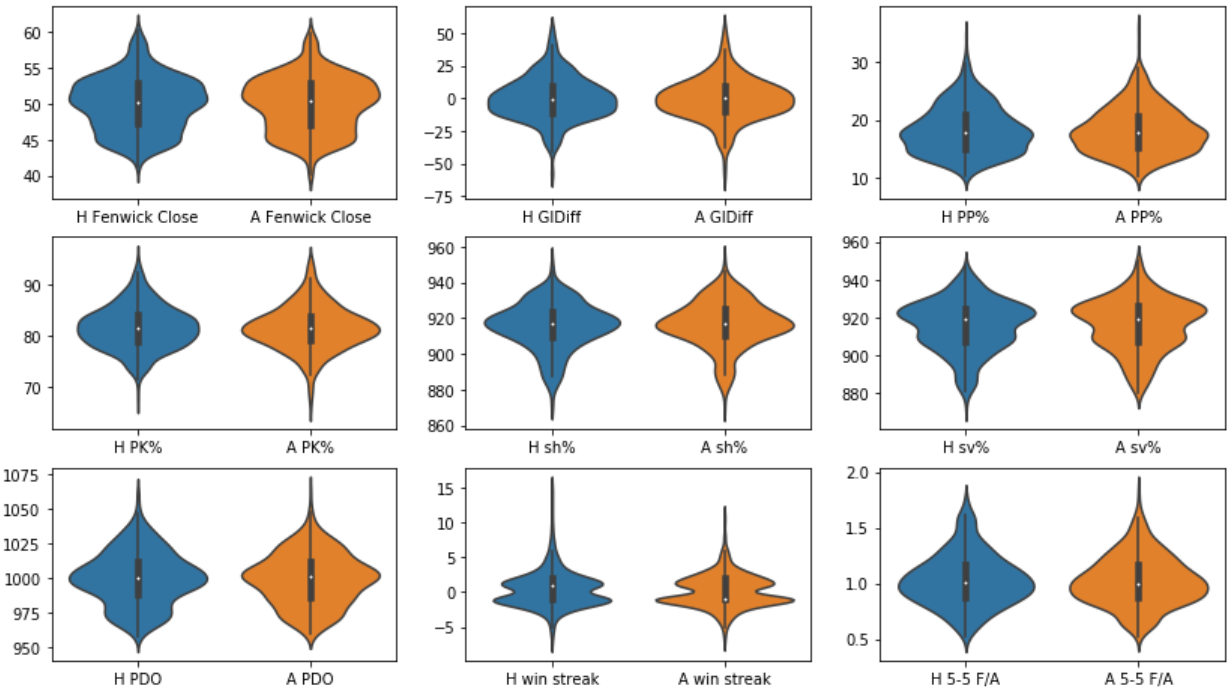

Figure 3. Violins plots of corrected data

Next, I created several scatter matrix plots of the data. Scatter plots show the data for two different variables plotted in a grid format. Again, the purpose is to look for trends in the data by pairing different values together. The matrix format takes several variables and creates scatter plots of them all in one visualization for convenience. Also shown in the matrix is a histogram where the same variable is plotted with itself, showing a bar plot of data values for that statistic. I created matrices for the home and away teams, shown in figure 4 for the home teams and figure 5 for the away teams. All the plots show home team wins colored orange and away team wins colored blue. The last scatter matrices I created, figure 6 and figure 7, show the data for both teams. I was hoping to see some correlation between home and away team data, but I didn't find anything obvious. In scatter plots, data that is correlated makes a shape more like a line than a blob. Lines that slope up to the right mean an increase in one variable also tends to increase the other variable, while lines that slope down to the right mean an increase in one variable tends to decrease the other variable. For example, it's slightly vague, but in figures 4 and 5, in the 4^{th} column, 2^{nd} row, the plots appear to slope upward to the right, instead of form shapeless blobs. The apparent slope means both home and away teams that killed penalties well tended to score more overall. The opposite is also apparently true, teams that didn't kill penalties well tended to be scored on more overall.

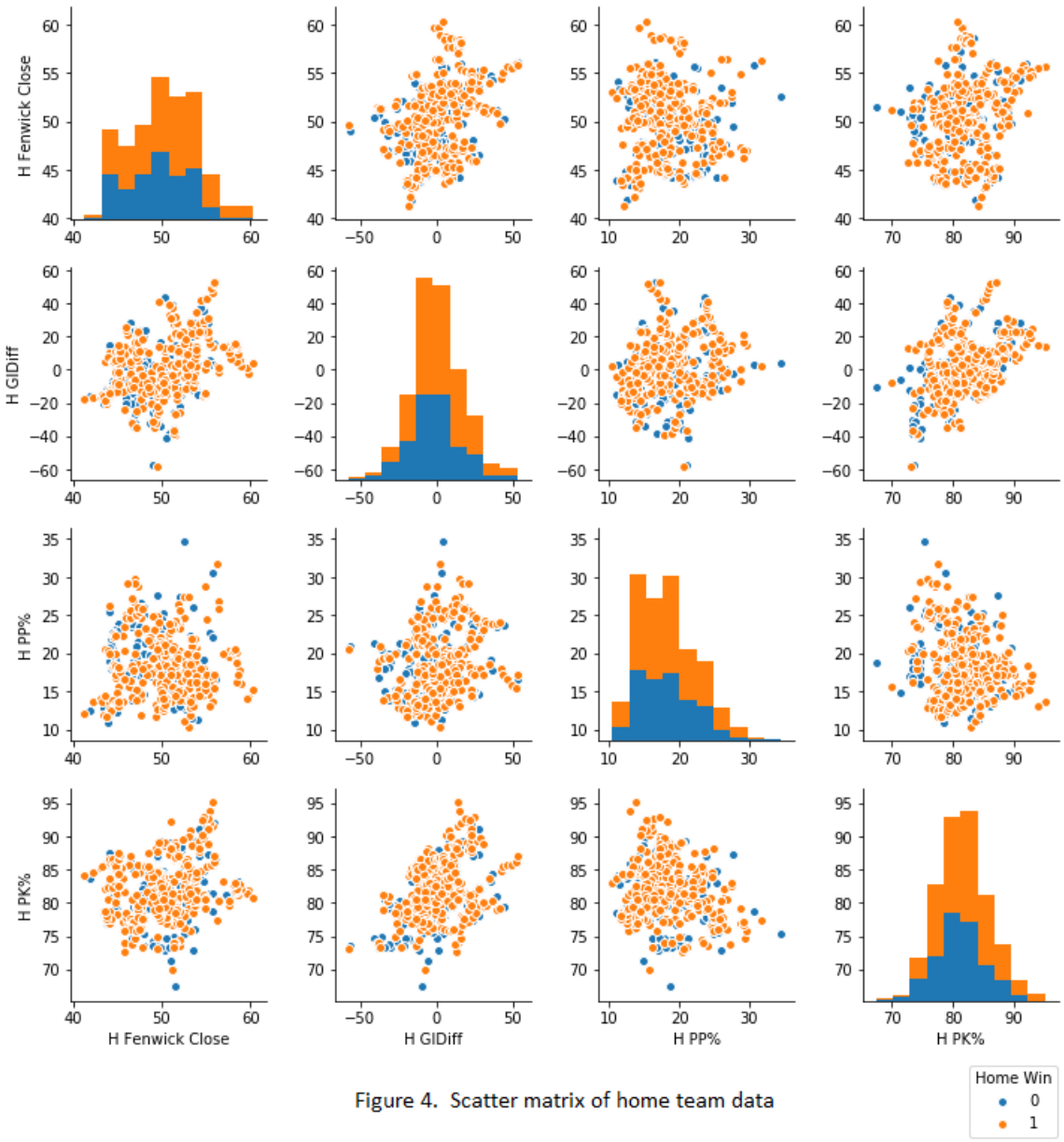

Figure 4. Scatter matrix of home team data

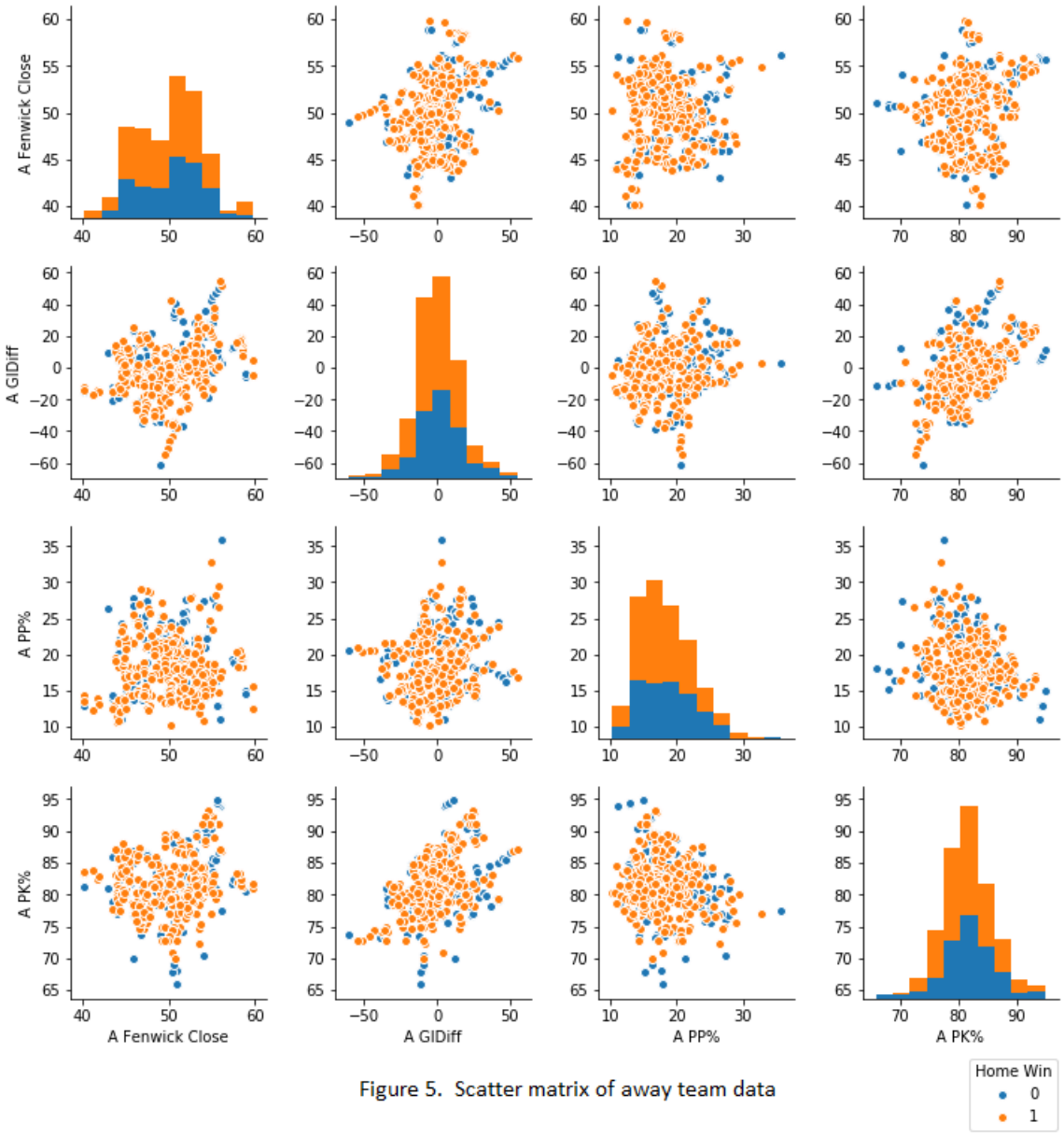

Figure 5. Scatter matrix of away team data

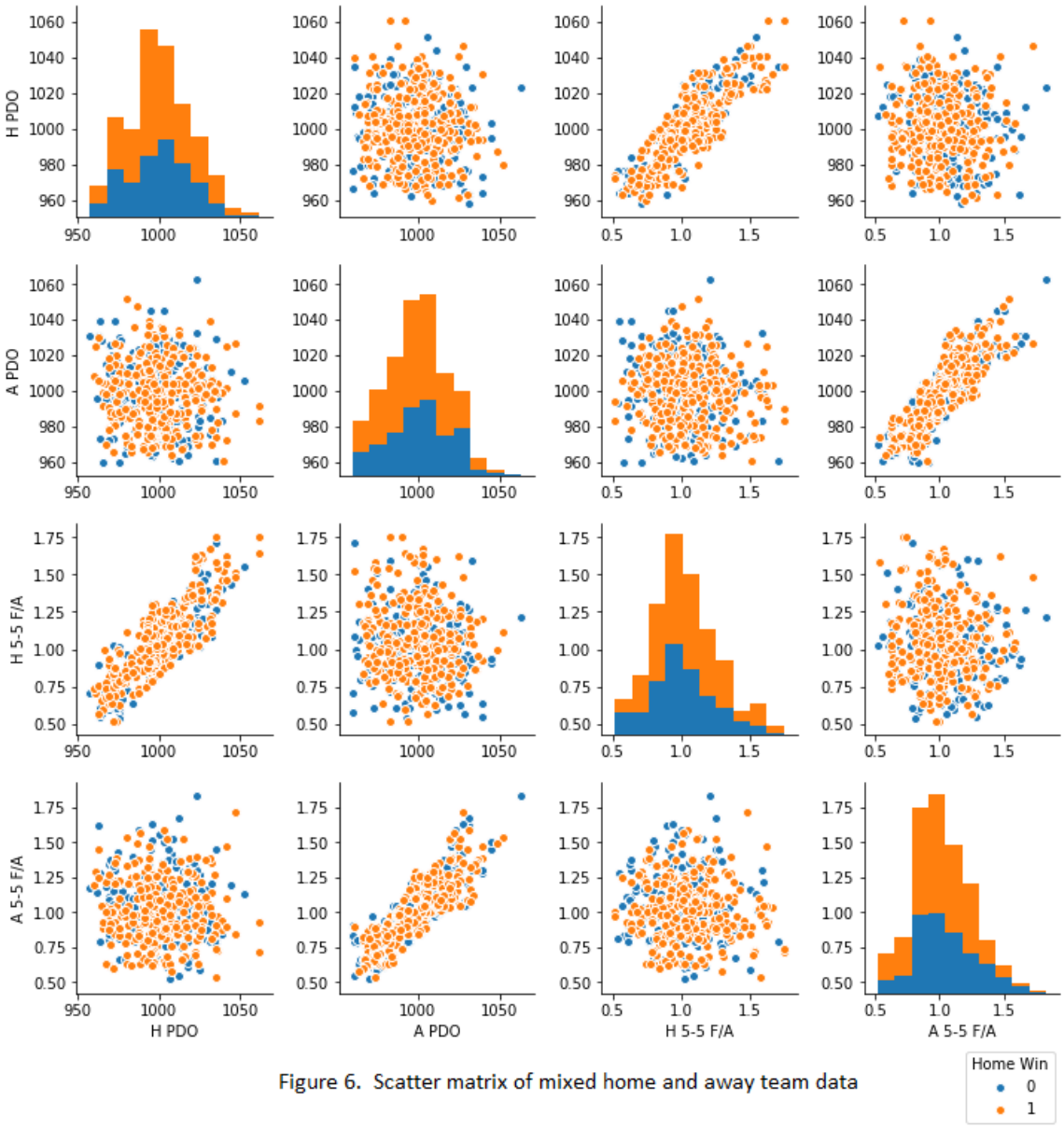

Figure 6. Scatter matrix of mixed home and away team data

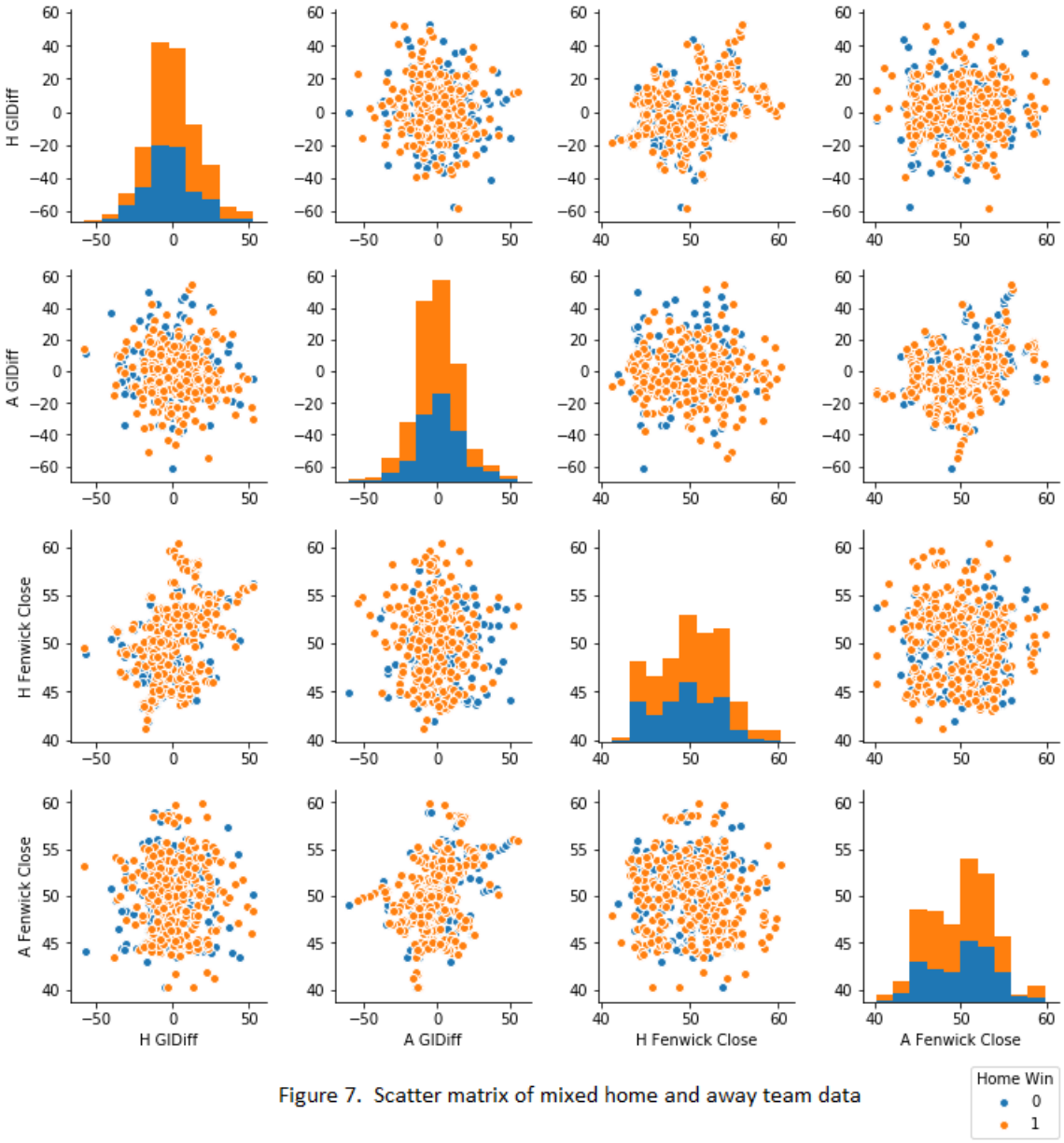

Figure 7. Scatter matrix of mixed home and away team data

My final exploratory step was to create histogram plots of the data that appeared to be correlated. The three best candidate values were goal difference, full strength scoring and PDO or luck. These are shown in figure 8. The top plots show a more gradual downslope on the positive side, suggesting the home teams tended to score slightly more than the away teams. The middle plots show the home teams with a larger plateau on the positive side and a slightly bigger positive tail, again suggesting the home teams tended to score more at full strength that the away teams. Finally, the bottom plots seem to show the peak of PDO for home teams to the left of 1000, while the peak of PDO for away teams to the right of 1000. 1000 is "average luck," so perhaps away teams tended to be luckier than home teams.

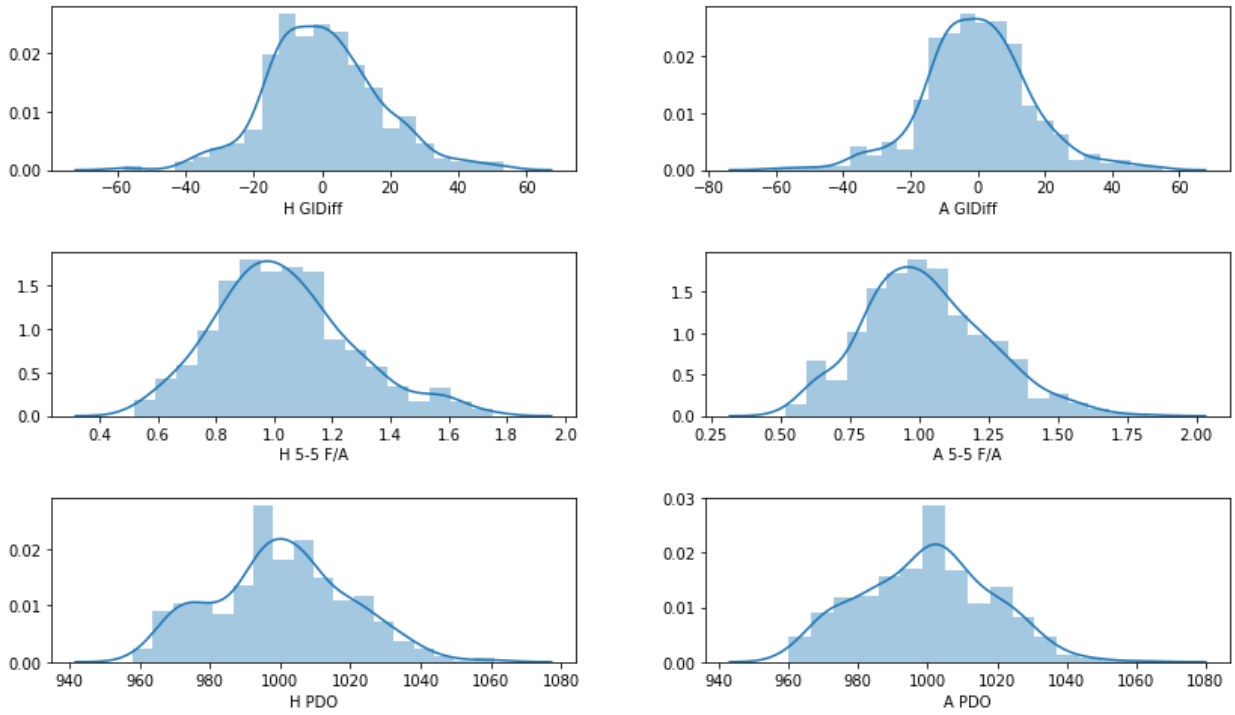

Figure 8. Histograms of correlated data

Inferential Statistics

According to the data, during the season I'm analyzing, 2013-2014, the home team won slightly less than 60% (0.586) of their games. I thought, in a perfect world with no home advantage, that number should be closer to 50%, so I created a hypothesis test for that purpose. My stated null hypothesis was, the mean win rate for home teams was 50%, while the alternative hypothesis was, the mean win rate for home teams was greater than 50%. I ran 10,000 random test seasons with a mean 50% home win rate and calculated the p-value based on how many of those seasons mean home win rate was greater than the observed mean home win rate. I also used Python's SciPy module to calculate a one-sample t-test for comparison. The observed mean home win rate appears to be plausible and not necessarily random chance, because I calculated a p-value of 0.00008, and generally a value less than 0.05 shows significant statistical evidence to reject the hypothesis. Shown in figure 9, is the histogram of my random test seasons compared to the observed value. Based on this graph, it is obvious the observed mean home win rate is not even close to the hypothetical 50%.

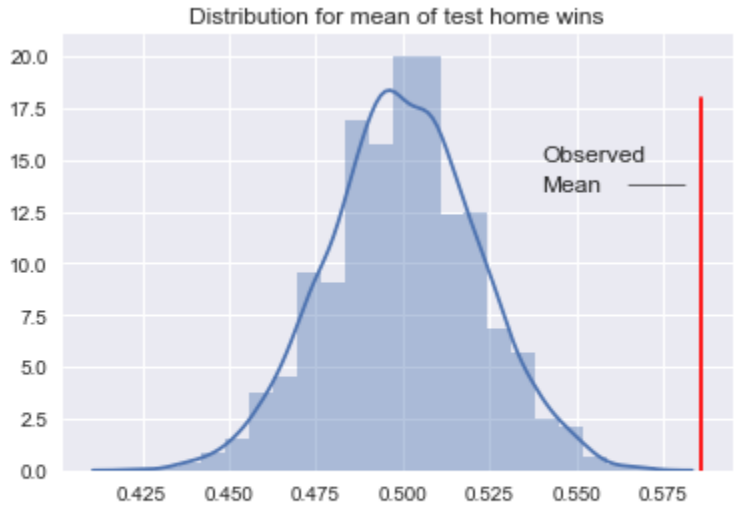

Figure 9. Histogram of test mean win rate

Home teams that season had a positive mean goal difference, meaning they tended to score more than they were scored on, while the away teams had a negative mean goal difference. I tested a hypothesis that the difference between these values may have been smaller than the data shows. My stated null hypothesis was, the mean goal difference for home teams was 0, while the alternative hypothesis was, the mean goal difference for home teams was greater than or less than 0. For this test, I again simply used the SciPy module, but this time used a two-sample t-test, because I couldn't be sure the mean goal difference was

positive or negative. I did not find enough statistical evidence to reject that, because the p-value I found from that test was 0.6759, much greater than 0.05.

Finally, I ran some correlation tests to determine which of the recorded statistics have the greatest potential impact on the outcome of the game and found quite different results for the home and away teams. I used the Spearman test for correlation, because I did not know if the data was normally distributed, which is a requirement to use the Pearson test for correlation. The values for correlation range from -1 to 1. A value near -1 means highly negatively correlated or increasing one value should greatly decrease the related value. A value near 1 means highly positively correlated or increasing one value should greatly increase the related value as well. A value near 0 means uncorrelated or increasing one value should have little or no effect on the related value. The values I calculated are shown in figure 10. As you can see, none of the values are very high, so correlation is relatively low. However, we can compare the values to each other. The home teams apparently benefitted most from controlling the puck when both teams have all 5 skaters, known as full strength, concentrating on their penalty killing, and scoring more than the other team when both teams are at full strength, in that order. The away teams apparently benefitted most from full strength scoring, scoring during their power plays, and PDO, again attempting to simulate luck, also in that order. Also, notice, as expected, the values for the away teams are negative, meaning increasing those values has a negative effect on the home team winning. The opposite is true for the home teams, except for PDO, which suggests the home teams may actually be hurt a little by luck.

```
Correlation of home goal difference to home team win is 0.108594861032
Correlation of home full strength goal ratio to home team win is 0.0769764024523
Correlation of home PDO or luck to home team win is -0.00313187820171
Correlation of away goal difference to home team win is -0.100108562433
Correlation of away full strength goal ratio to home team win is -0.0665165539607
Correlation of away PDO or luck to home team win is -0.0606635871359

Correlation of home fenwick close to home team win is 0.107594912024
Correlation of home power play % to home team win is 0.00660495470368
Correlation of home penalty kill % to home team win is 0.0890216906726
Correlation of away fenwick close to home team win is -0.0352729093513
Correlation of away power play % to home team win is -0.0715503978477
Correlation of away penalty kill % to home team win is -0.0516689223862
```

Figure 10. Spearman correlation coefficients

Baseline and Advanced Model Construction

I began my baseline model by building a logistic regression classifier using a train-test split of the default 75%-25% on the full season data. In what is known as machine learning, this method divides, or splits, the data into two different sets, one composed of 75% of the data which is used to train the model, and the other composed of the remaining 25% which is used to test the accuracy of the model on data it hasn't been exposed to. One potential problem with the split of data was the imbalance of win rate between the home and away teams. The home teams won close to 60% of the games, but a random split may not accurately reflect that. In order to minimize the imbalance of the home teams winning so much more than the away teams, I used an optional parameter in the split to maintain the same ratio of home wins in the training and testing data, known as stratifying. Part of training the model is called hyperparameter tuning, which involves adjusting several parameters of the model to try to maximize the accuracy. I calculated the best accuracy from L1 and L2 regularization but found from the classification report that both the precision and recall were very low when predicting a home team loss. The classification reports for these initial models are shown in figure 11. In the report, precision refers to the number of home team wins the model predicted that ended up being actual home team wins. Recall refers to the amount of home team wins the model predicted out of the total number of actual home team wins. Both of the scores range from 0 to 1, with a score closer to 1 being better. F1-score is simply a score calculated from both of those values, and support is the number of games in each set of data, training or test data, where 1 was an actual home team win and 0 a home team loss. I also found many estimated coefficients, again for both L1 and L2 penalties, were zero or extremely small. The estimated coefficients are the model's suggested weight for each variable, so those results were especially surprising and confusing. I was unsure how the model could predict anything if the features contributed little to nothing to the regression equation.

```
[Final Training Classification Report Best L1 Regularization:]
             precision    recall  f1-score   support

          0       0.00      0.00      0.00       160
          1       0.59      1.00      0.74       227

avg / total       0.34      0.59      0.43       387

[Final Test Classification Report Best L1 Regularization:]
             precision    recall  f1-score   support

          0       0.00      0.00      0.00        54
          1       0.58      1.00      0.74        76

avg / total       0.34      0.58      0.43       130

[Final Training Classification Report Best L2 Regularization:]
             precision    recall  f1-score   support

          0       0.00      0.00      0.00       160
          1       0.59      1.00      0.74       227

avg / total       0.34      0.59      0.43       387

[Final Test Classification Report Best L2 Regularization:]
             precision    recall  f1-score   support

          0       0.00      0.00      0.00        54
          1       0.58      1.00      0.74        76

avg / total       0.34      0.58      0.43       130
```

Figure 11. Initial classification report

Since logistic regression seemed relatively inaccurate, I repeated the calculations with random forest classifiers, setting estimators for square root, log, and no maximum of features, as well as number of trees, for hyperparameter tuning. Random forest classifiers begin with what is known as a decision tree. The tree is created by that model examining each variable and branching at different values toward the outcome, again in this case, the home team wins or the home teams loses. A completely optimized tree yields the same results every time, so tends to work extremely well during training, but relatively poorly on unknown data. To compensate, decisions can be given in random order, forcing the tree to reach different outcomes. The forest refers to the training of many of these decision trees and taking the most common outcome as the predicted one. These all produced better scores than the logistic regression classifiers but were again still low on scores predicting home losses, with precision and recall scores around 0.4, as shown in figure 12. I've only shown the results on the test data

for the random forest classifier, because that classifier tends to score very well on training data naturally. Next, I ran calculations for all previous classifiers when the data was split 70-30 and 80-20, all with similar results. The authors had also gathered incomplete data for several games prior to when they started their calculations. I imputed the means of the missing data and trained all the classifiers with the extra data, and still had similar outcomes, as shown in figure 13.

```
[Final Test Classification Report Best sqrt Random Forest:]
             precision    recall  f1-score   support

          0       0.47      0.44      0.46        54
          1       0.62      0.64      0.63        76

avg / total       0.56      0.56      0.56       130

[Final Test Classification Report Best log2 Random Forest:]
             precision    recall  f1-score   support

          0       0.45      0.37      0.41        54
          1       0.60      0.68      0.64        76

avg / total       0.54      0.55      0.54       130

[Final Test Classification Report Best None Random Forest:]
             precision    recall  f1-score   support

          0       0.39      0.33      0.36        54
          1       0.57      0.63      0.60        76

avg / total       0.50      0.51      0.50       130
```

Figure 12. Initial random forest classification report

```
[Final Training Classification Report Best L1 Regularization with extra data:]
             precision    recall  f1-score   support

          0       0.60      0.39      0.47       182
          1       0.64      0.81      0.71       244

avg / total       0.62      0.63      0.61       426

[Final Test Classification Report Best L1 Regularization with extra data:]
             precision    recall  f1-score   support

          0       0.48      0.38      0.42        61
          1       0.60      0.69      0.64        81

avg / total       0.55      0.56      0.55       142

[Final Training Classification Report Best L2 Regularization with extra data:]
             precision    recall  f1-score   support

          0       0.61      0.38      0.47       182
          1       0.64      0.82      0.72       244

avg / total       0.63      0.63      0.61       426

[Final Test Classification Report Best L2 Regularization with extra data:]
             precision    recall  f1-score   support

          0       0.46      0.34      0.39        61
          1       0.58      0.69      0.63        81

avg / total       0.53      0.54      0.53       142
```

Figure 13. Initial classification report using extra data

My final attempt to generate better predictors was to resample the data, because the home team did win almost 60% of the games, generating an imbalanced classification feature. Resampling refers to altering the data to try to correct the imbalance. Over-sampling is the process of adding to the less represented data, in this case the home team losses, and under-sampling is the process of subtracting from the more represented data, in this case the home team wins. I chose to do my resampling before the train-test splitting, rather than split then resample, to try to minimize the effect of creating or subtracting data from the existing imbalance. I first tried over-sampling, using naïve random, SMOTE and ADASYN methods, which are all different ways to randomly create new data, followed by random under-sampling, and again trained all the classifiers with the new data set. This finally achieved results similar to the original authors. All of these methods scored precision and recall scores between roughly 0.55 and 0.6. The best results were from the random forest classifier with square root

maximum features on random naïve over-sampled data, as shown in figure 14, again showing only the results on the test data. Figure 15 shows the estimated coefficients from that classifier. In the random forest classifier, there are no negative coefficients, so the larger numbers show variables with higher suggested importance. To finish, figure 16 shows the ROC curve, figure 17 shows the precision recall curve, and figure 18 shows the confusion matrix for the best scoring classifier.

[Final Test Classification Report Best sqrt Random Forest naive over-resampled data:]

	precision	recall	f1-score	support
0	0.66	0.71	0.68	76
1	0.69	0.63	0.66	76
avg / total	0.67	0.67	0.67	152

Figure 14. Best achieved classification report

Estimated Coefficients for sqrt trees:

	estimated coefficients	features
4	0.054654	H PP%
5	0.054109	H PK%
17	0.049989	A PK%
16	0.049949	A PP%
1	0.047618	H GF
12	0.047444	A Fenwick Close
0	0.047004	H Fenwick Close
14	0.046232	A GA
7	0.046017	H sv%
13	0.043993	A GF
11	0.043963	H 5-5 F/A
2	0.043710	H GA
6	0.043403	H sh%
8	0.043311	H PDO
20	0.040617	A PDO
18	0.039530	A sh%
3	0.039377	H GlDiff
19	0.037648	A sv%
23	0.035523	A 5-5 F/A
22	0.034108	A standing
15	0.031969	A GlDiff
10	0.030622	H standing
9	0.027084	H win streak
21	0.022127	A win streak

Figure 15. Estimated coefficients from best classifier

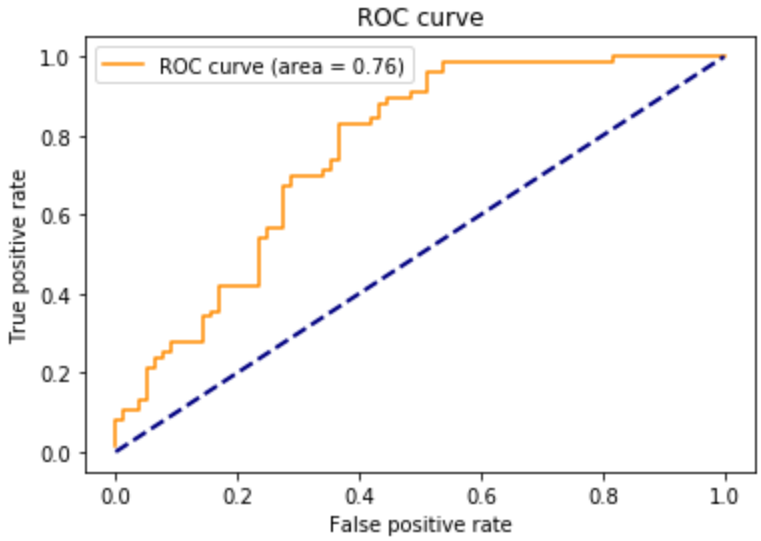

Figure 16. ROC curve for best found classifier

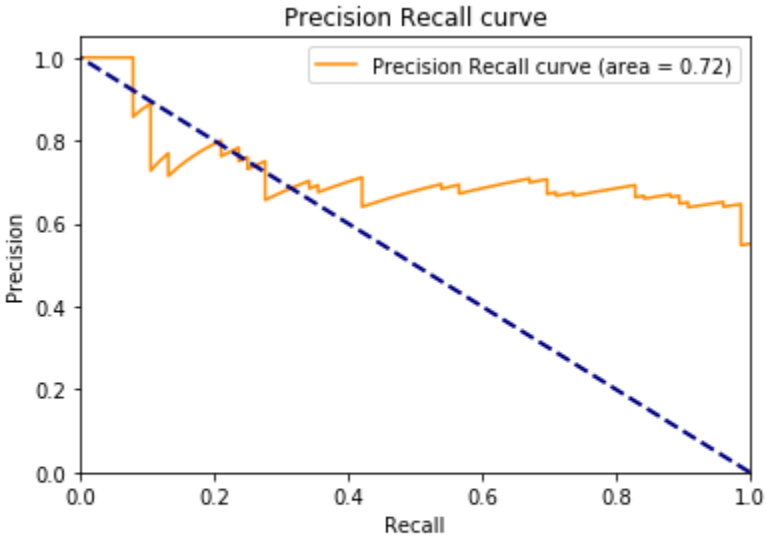

Figure 17. Precision recall curve for best found classifier

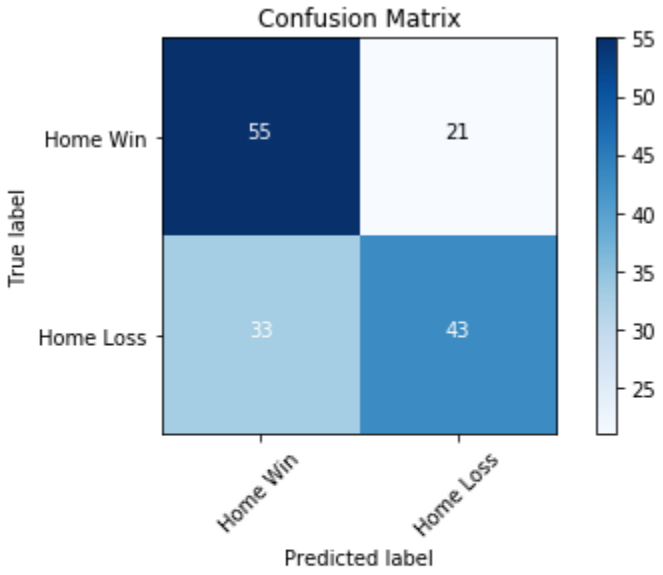

Figure 18. Confusion matrix for best found classifier

Conclusions

My final results are close to those of the authors of the original study, with my best accuracy just over 65%, but among the better classifiers, average accuracy around 60%. It has been suggested that luck is responsible for 38% of a team's winning, and both the original study and my own seem to agree with that [3]. According to the confusion matrix in figure 15, the best model did fairly well predicting home wins and slightly less well predicting home losses. Its biggest weakness appears to be false positives, where it predicted home wins that were actually home losses. This could lead to circumstances where the home team expects, or predicts, a win but ends up with a loss, which would actually exaggerate the existing problem with the model.

Recommendations to Clients

What I can report to NHL teams are the statistics I've calculated to be most influential on the outcome of the game. Figures 10 and 15 can be referred to. According to the classifier I tested with the best results, the most important measurements for the home and away teams were special teams, meaning both power play and penalty killing, with the home teams having slightly more influence than the away teams. According to all the well scoring classifiers, home teams benefit most from Fenwick close and penalty killing, while for the away teams, mostly power play is important. Other statistics for teams to pay attention to include, for the home teams, Fenwick close and penalty killing, and to a lesser degree, power play. Otherwise, overall standing and win streak were important to some classifiers. For the away teams, what mattered most were fairly equally power play, Fenwick close and penalty killing. Also, PDO, overall standing, win streak, and shooting percentage were all represented a little. So, both teams seem to need to worry about puck possession and special teams, but the away teams seem to need to worry about a few more things as well.

Future Work

There is only data from a single NHL season in these projects, so studying more seasons may provide additional information. Studying the effect of other existing variables, such as hits and zone starts, may help the model. Another possibility would be trimming less impactful variables from the calculation, because I used all the collected data in my model. It is also possible that statistics that are not currently measured may be more useful in predicting the winner. Either of those possible additions are beyond the scope of this project but would certainly be interesting future projects.

References

1. Denson, Ben: The NHL is dead. The Cornell Daily Sun. http://cornellsun.com/2016/02/09/denson-the-nhl-is-dead/ (2016)

2. Weissbock, J., Viktor, H., and Inkpen, D.: Use of Performance Metrics to Forecast Success in the National Hockey League. University of Ottawa, Ottawa, Canada (2013)

2. username Hawerchuk: Luck in the NHL Standings. https://www.arcticicehockey.com/2010/11/22/1826590/luck-in-the-nhl-standings (2010)