Capstone Project 1
Predicting NHL game winners with machine learning

Kevin Hill

Statement of Purpose

- The goal of my capstone project 1 is to predict the winner of a National Hockey League (NHL) game, and to determine which statistics have the most influence over which team wins a game.
- The primary clients I envisioned when planning this project are mainly NHL teams, and other non-NHL hockey teams.
- This project was inspired by an article I read from sports analysts at the University of Ottawa, who tried this prediction already. I am also a huge hockey fan and wanted to join my passion for the sport with my growing knowledge of data analysis.

Problem Formulation and Data Origin

- The main problem to solve is: Does the home team win a given hockey game?
- An additional problem to solve is: Which hockey statistics have the most influence on whether or not the home team wins?
- The data was collected by the authors of the article, and was obtained from them via email with their permission. It contains various statistics ranging from basic hockey game and team stats to analytic stats derived from those basic ones. The same data can be obtained online from such websites as NHL.com or Corsica.hockey.

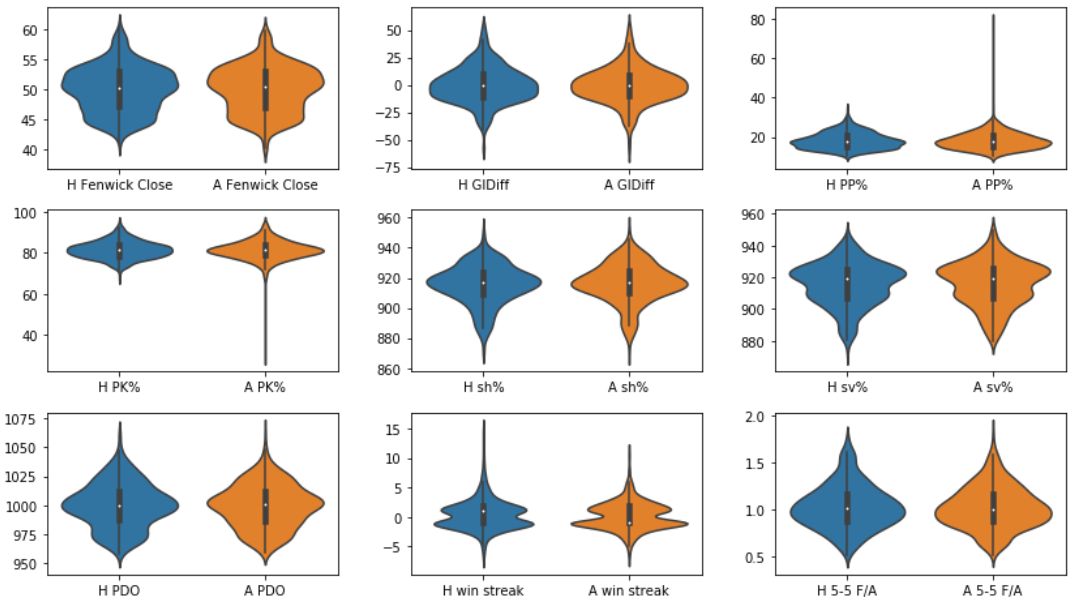
Data Wrangling

- The data was already processed by the authors for their study. However, their model worked with the data for a game in two blocks of statistics, one for the home team and one for the away team. My planned model needed one block of statistics containing all the data for a game, so most of my data wrangling involved rearranging the data to fit my model.
- There was some data the authors had collected that was incomplete that they didn't use, so I imputed the means of the missing data for possible inclusion.

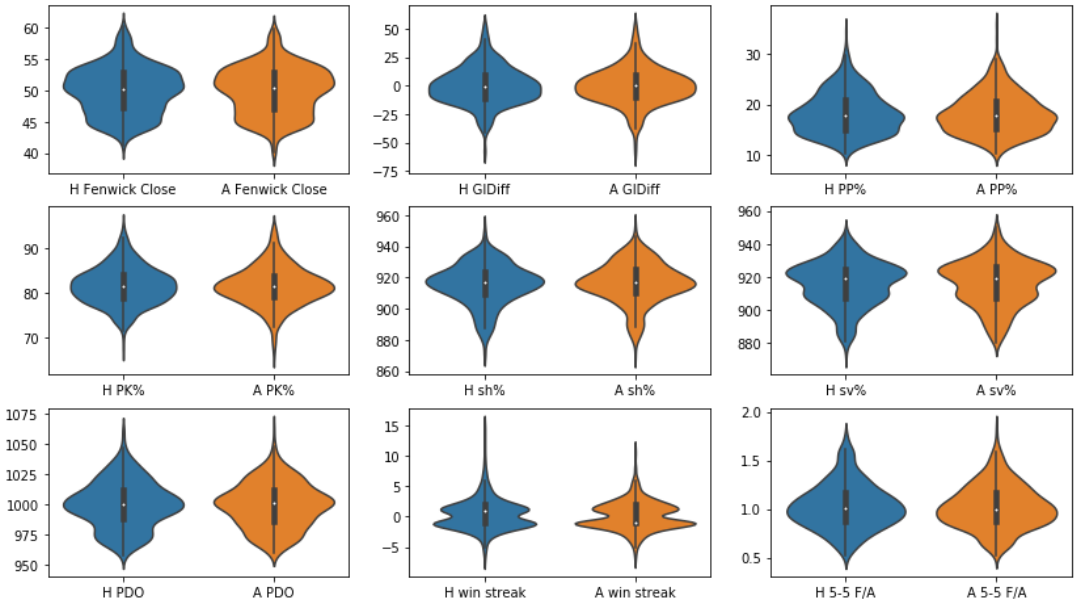
Exploratory Analysis

- I created some violin plots, showing density of data points, with the home team data on the left, in blue, and the away team data on the right, in orange.
- Three of the plots needed some investigation.
- Two appeared to have data switched, which I corrected, while the third ended up being correct.
- Afterward, no large distinction appeared between home and away teams.
- The original plots from the imported data and the plots with the corrected data appear on the next slide.

Imported data

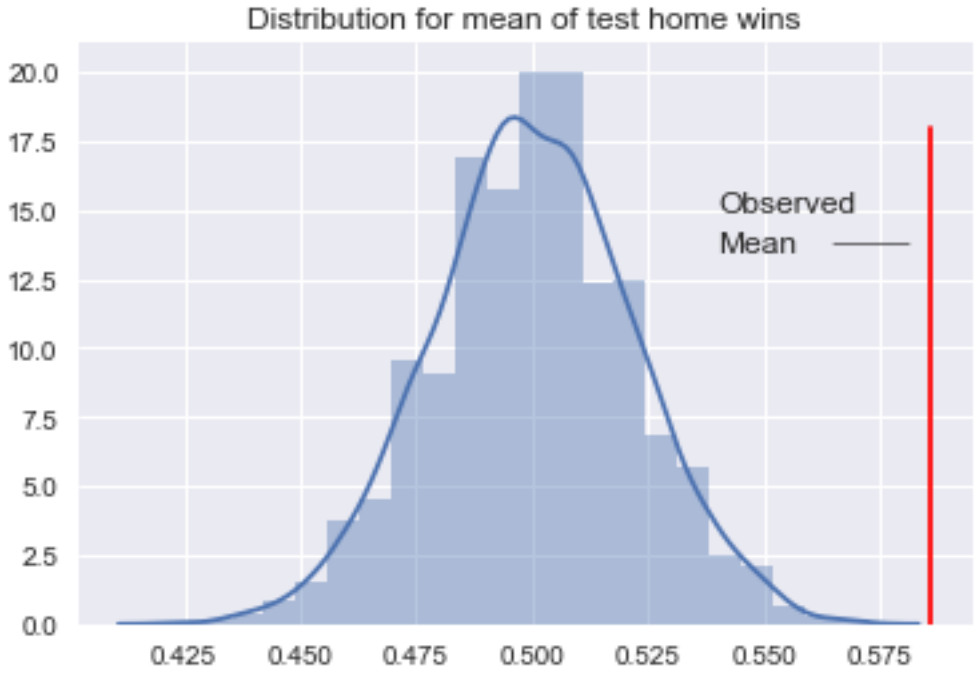


Corrected data



Inferential Statistics

- According to the data, the home team won almost 60% of the games. I tested a null hypothesis that the home team won 50% of the games, with the alternative hypothesis being the home team won more than 50% of the games. I obtained a p-value of 0.00008 in that test, which is less than 0.05, so there was significant statistical evidence to reject the null hypothesis.



- I also 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.
- The home teams 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 benefitted most from full strength scoring, scoring during their power plays, and a stat that attempts to show luck, known as PDO, also in that order.

In-Depth Analysis

- Since the problem involves two possible outcomes, home team wins or home team loses, I first chose to build a logistic regression model to solve the binary classification problem. Using hyperparameter tuning, I would select the best performing classifier.
- However, poor recall scores and very low estimated coefficients made me choose another classifier, the random forest. Again, hyperparameter tuning would provide me the best model. Results were better, but still low.
- The next slide shows the classification reports for the best of both classifiers.

[Final Test Classification Report Best Logistic Regression:]

	precision	recall	f1-score	support
0	0.46	0.30	0.36	54
1	0.60	0.75	0.67	76
avg / total	0.54	0.56	0.54	130

[Final Test Classification Report Best Random Forest:]

	precision	recall	f1-score	support
0	0.44	0.37	0.40	54
1	0.60	0.67	0.63	76
avg / total	0.54	0.55	0.54	130

- Precision measures how many home team win predictions (line 1) were actual home team wins.
- Recall measures how many home team wins were predicted (line 1) out of the total actual home team wins.
- F-1 score is an overall score using both precision and recall.
- All 3 scores range from 0 to 1, with closer to 1 being better.
- Support is the total number of games predicted.

Final Results

- Since the home teams won so many more games than the away teams, the data is called imbalanced, which may be why the scores are low.
- The best results came when the data was resampled, meaning data was either created pseudo-randomly or ignored to create a more balanced set of home team wins.

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

	precision	recall	f1-score	support
0	0.62	0.59	0.60	76
1	0.61	0.63	0.62	76
avg / total	0.61	0.61	0.61	152

- These was very close to the results the original authors achieved.
- With the balanced data, the scores are higher overall.
- This is the best model I can build to make the best possible prediction on unknown data. However, even this model is only a little better than flipping a coin.

Recommendations

- Even if a winning prediction is vague, I can suggest which statistics had the most influence on the models.
- All the models show for home teams:
  - Fenwick close and penalty killing are very important
  - Power play, overall standing and win streak are somewhat important
- And for away teams:
  - Power play, Fenwick close and penalty killing are very important
  - PDO, overall standing, win streak, and shooting percentage are somewhat important

- However, the best scoring model shows these for home teams:
  - Fenwick close and penalty killing
- And these for away teams:
  - Only power play
- Combined with the earlier analysis shows home teams might rely on penalty killing and full strength puck control, but away teams might rely on power play, full strength puck control, and a little luck.

Future Considerations

- There is only data from a single NHL season in these projects, so studying more seasons may provide additional information.
- It is also possible that statistics that are not currently measured may be more useful in predicting the winner.