Table of Contents

1. Problem Description

The initial problem I am attempting to tackle was to classify shots from NHL players as a goal or a missed shot using different information about the shot attempt. Working with the data, I determined it may be better suited to use to see if a shot attempt would be successful based off of selected features in the dataset. Therefore, the objective of this analysis is to determine which features of a National Hockey League game can best objectively classify which shot will result in a goal versus which will result in a no-goal. I was able to find an appropriate dataset online containing the factors I needed.

2. Description of Data

The dataset used in this project is seven National Hockey League seasons worth of shot attempts (including the playoffs). This includes 665,000+ shot attempts matched with such data as shot type, shot location, team shooting, player shooting, side of the rink and goal/no-goal. The NHL does not make access to their data very easy so I initially though I would have to scrape what I needed through their API. I was lucky enough through my exploration to find someone who had used NHL shot data for a different analysis and that data was used here (source referenced in this document). This should be plenty of records to use machine learning to determine whether or not a given shot should be successful or to classify “good” or “bad” shots.

3. Method Description

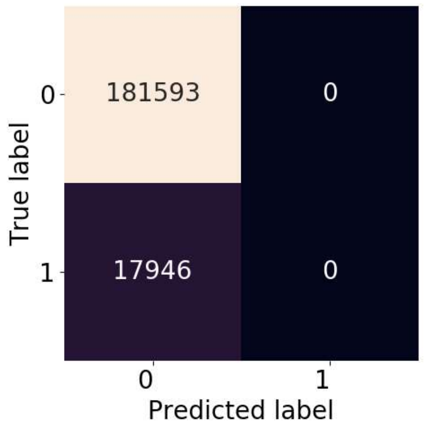
I decided to use a decision tree algorithm for classification of the shots which means I chose the gini function to evaluate the performance and characteristics of each node and to use a confusion matrix to test the accuracy of my predictions. I also used Principal Component Analysis to reduce the feature options by choosing the ones with the largest coefficient magnitude. Initially, when testing and training my algorithm I was receiving fairly normal prediction accuracy rates with no abnormal statistics throwing off my analysis. However, as I continued to check the accuracy of my code, some changes needed to be made which I will cover in the “Experimental Setup section.

4. Experimental Setup

I used several python libraries to facilitate my analysis including: sklearn, pandas, numpy, matplotlib, and seaborn. The first step in the analysis was to do some preliminary exploration of the data to determine if any fields need to bet altered or if nulls need to be addressed. The key takeaways of this were that I knew I wanted to look at shot location as the first feature to possibly build upon and that there were some string variables that would need to be converted via a dictionary to numerical values (i.e. shot type).

5. Results

The Principal Component Analysis led me to 4 features I wanted to test: X Coordinates, Y Coordinates, Shot Type, and TeamFor_ID. After the initial round of experimentation, it became clear that something was odd about the way it was configured. Looking at the image below, the results would appear to overfit on true negatives and struggle to classify true positives.



I attempted a couple of things to resolve this overfitting issue but to no success. I normalized the shot data to try to simplify the algorithm to only be on one side of the rink. I also tried removing the coordinates completely to no success. Essentially the algorithm was very good a saying a shot from behind the net for instance would not be successful but couldn't tell for certain in front of the net if a shot would go in or not. I was able to get the algorithm to correctly class ~100 true positives but was not able to get any higher than that.

6. Summary and Conclusions

In conclusion, this data was much more complex than I originally suspected. There were more factors than I was prepared to try to summarize for my algorithm. I would definitely attempt to pull or find defensive information to better prepare my decision tree to handle shots in front of the net as to better successfully identify predicted positive results. When I initially chose this as my project, I though the location would really be more of a successful indicator of goals in the NHL. I was disappointed with the results but feel like with more time I really could have squeezed some information out of this dataset.

7. References

1. Data Set: https://www.kaggle.com/martinellis/nhl-game-data , 2018 , Martin Ellis
2. NHL Data Source, Gamecenter: https://www.nhl.com/gamecenter/bos-vs-stl/2019/06/03/2018030414#game=2018030414,game_state=final , 2011-2018
3. Inspiration: https://pdfs.semanticscholar.org/c0bc/e3faf2ff533292362059c19ab8608889e71e.pdf , 2014, Gianni Pischedda, International Journal of Computer Applications
4. Sklearn research: https://scikit-learn.org/stable/modules

8. Computer Appendix

Code Contribution: 40 -15/ 40+21 = 41%