CLASSIFYING NHL SHOT ATTEMPTS

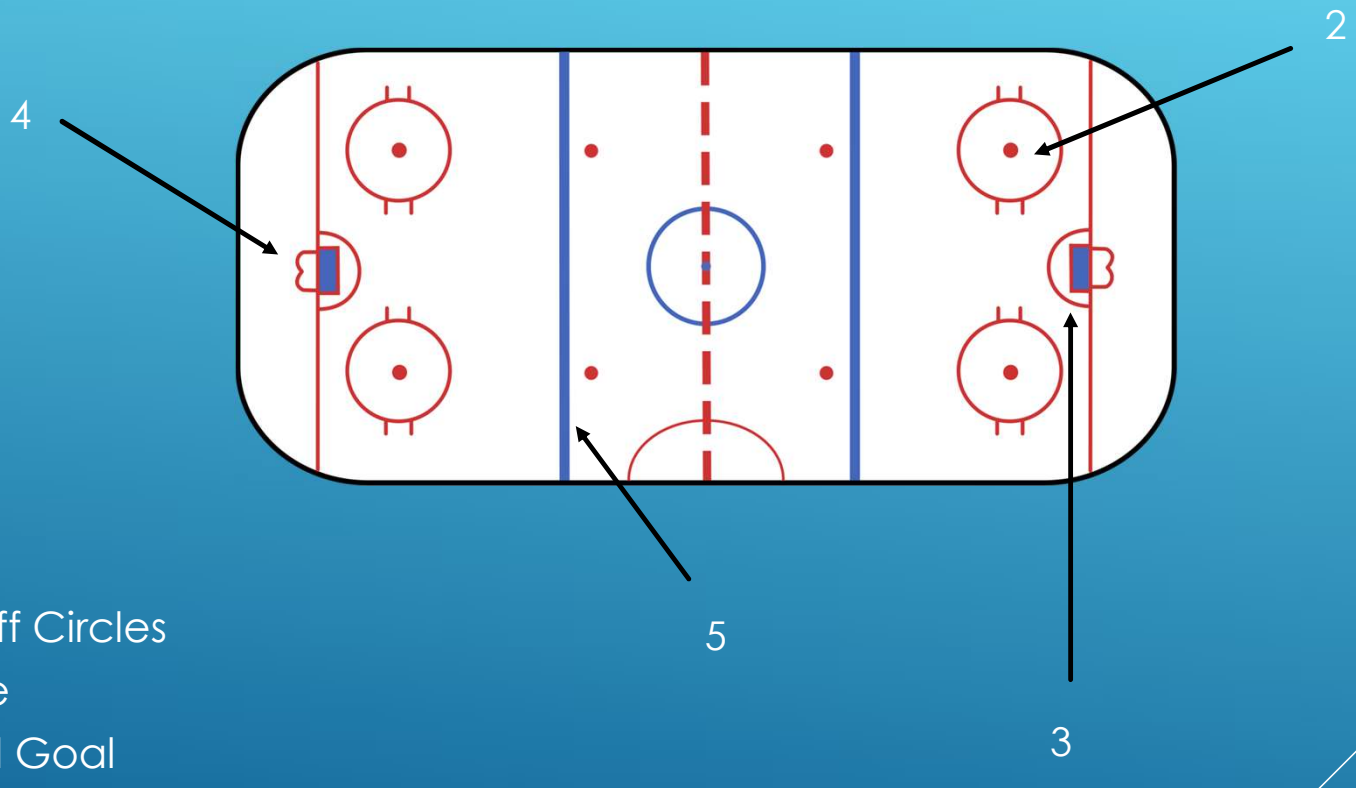
INTRODUCTION

- Problem: What factors most influence whether a shot in an ice hockey game is a success (goal) or failure (no-goal).
- What does an NHL season look like? What kind of information could we gather about games or shots?

NATIONAL HOCKEY LEAGUE

- 31 teams in the league
- Each team plays 82 games in a season + post-season playoffs
- Over the last 7 seasons teams have averaged between 25 – 40 shots per game but over the same time span only average between 2.5 – 3.1 goals per game

A LOOK INTO THE GAME

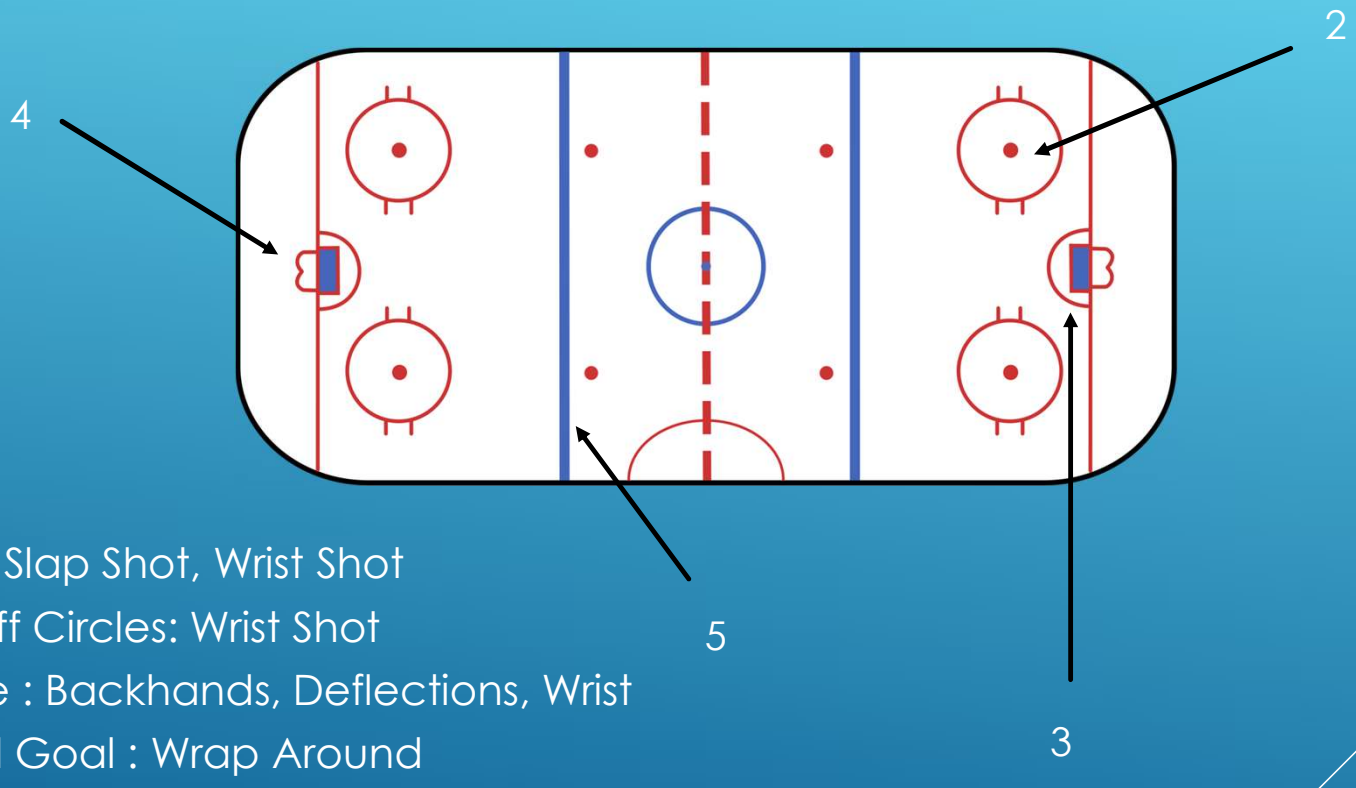


1. Zones
2. Faceoff Circles
3. Crease
4. Behind Goal
5. Blue Line

RELATING TO THE DATA

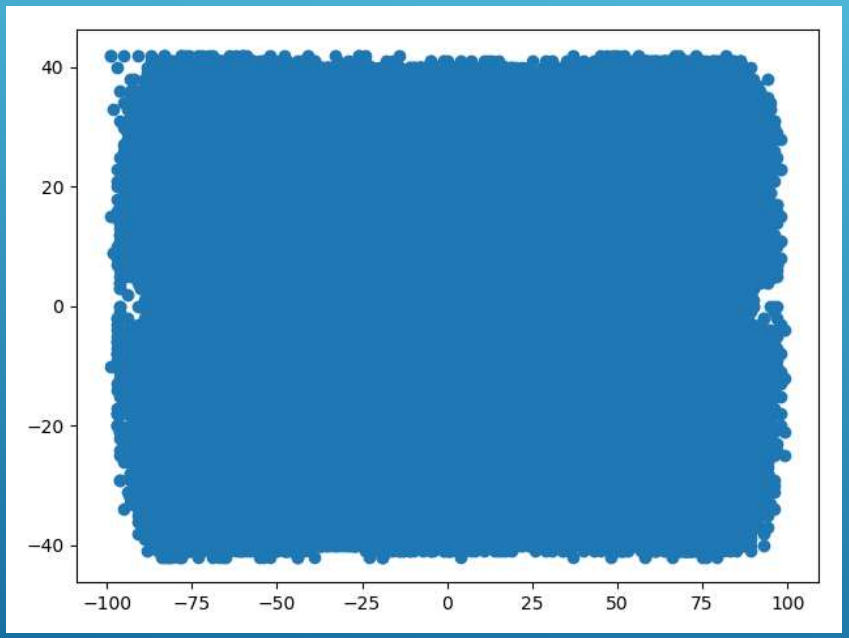
- Dataset used in this analysis consists of 660,000+ shots with ~60,000 of them resulting in goals
- Shot types: Backhand, Deflection, Slap Shot, Snap Shot, Tip-In, Wrap-Around, Wrist Shot
- Additional possible features: Shot coordinates (X,Y), rink side, players, teams, period type (regular / overtime), time remaining
- Need to choose features and use the result of our shot (goal: 1, no goal:0) as the target

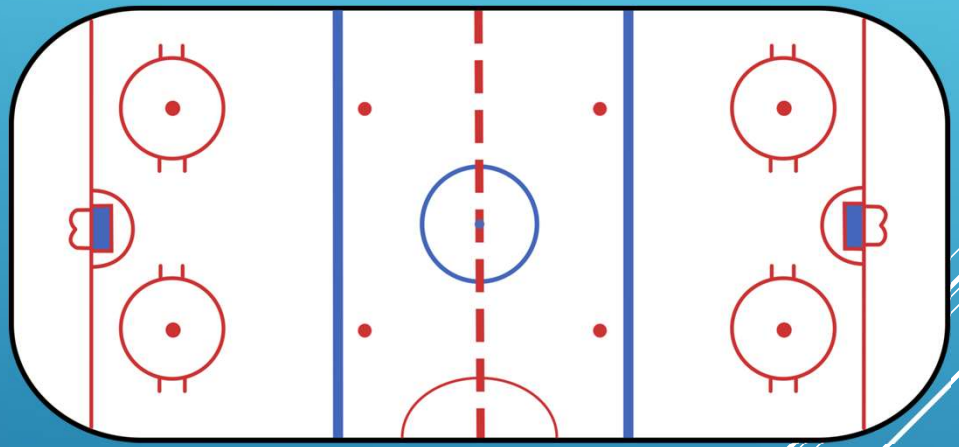
WHERE DO THE SHOTS OCCUR?



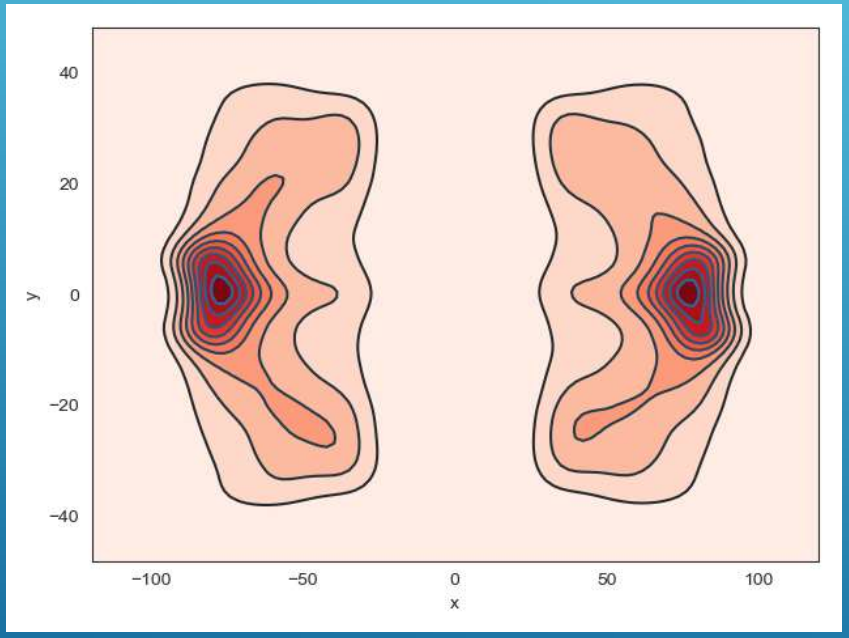
1. Zones: Slap Shot, Wrist Shot
2. Faceoff Circles: Wrist Shot
3. Crease : Backhands, Deflections, Wrist
4. Behind Goal : Wrap Around
5. Blue Line: Slap Shot

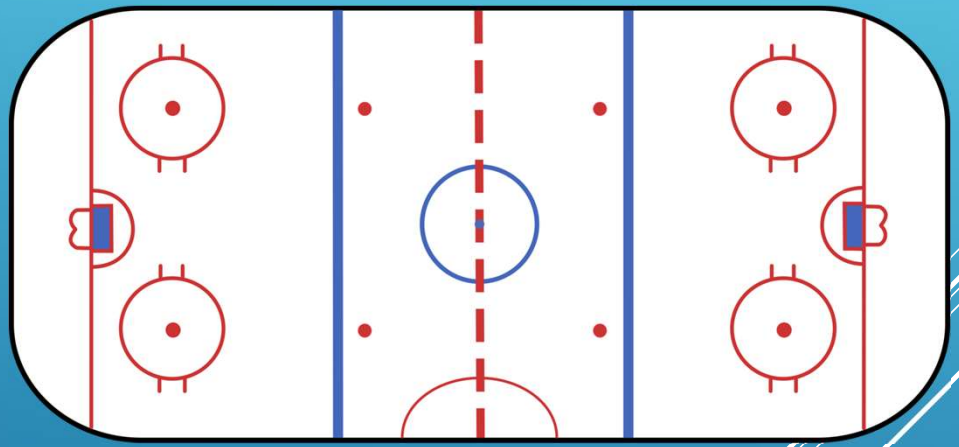
A LOOK AT THE SHOTS



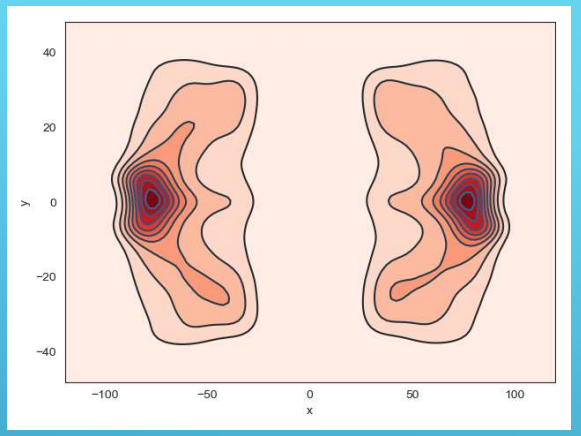


A LOOK AT THE SHOTS





FEATURE SELECTION



- Just from visualizing the shots locations it's apparent that the coordinates are probably of interest to the analysis
- Perfomed Principal Component Analysis (PCA) to reduce the dimension selection based on variance or magnitude of their coefficient
- The 4 features chosen to focus on were X-Y Coordinates, Shot Type, and TeamFor_ID based on the results of PCA

ALGORITHM CHOSEN

- Just from visualizing the shots locations it's apparent that the coordinates are probably of interest to the analysis
- Perfomed Principal Component Analysis (PCA) to reduce the dimension selection based on explained variance
- The 4 features chosen to focus on were X-Y Coordinates, Shot Type, and TeamFor_ID based on the results of PCA
- Decision Tree Classifier was chosen based on number of features for a first attempt

FIRST ATTEMPT

- On the first attempt of classifying whether a shot is a goal or no-goal based on the features selected
- Tried using a decision tree classifer from sklearn with no parameters tweaked and there was a result of 0.84 accuracy.
- Not terrible, but could be better

SECOND ATTEMPT

- On the second attempt at classifying whether or not a shot was a goal, the decision tree split qualifier was set to consider information gain as a factor for decision making.
- The result here pushed the classifying accuracy up to 91%

RESULTS

- The validation of these results was not as conclusive as I would have liked them to be
- A confusion matrix showed overfitting on true negatives and struggled to correctly classify true positives
  - Too similar results for features we have and not enough data for "x factors" i.e shots can go in or not go in from pretty much every location on the ice, goalie, etc….
  - Limiting shots to 1 side of the ice had no effect
- Even though shot location had a high magnitude coefficient as compared to others

True	0	178577	3016
Label	1	17814	132
		0	1
		Predicted	Label

SUMMARY

- While location may be a strong feature to classify whether or not a shot is successful, it alone is not enough to train an algorithm
- Would have much more thoroughly investigated during preprocessing if possible

HOW WOULD I CONTINUE THIS RESEARCH

- Combine results with another method of classification for better resulsts
- Normalize the shot data more throroughly
- Go further in-depth analyzing shot selection for specific players/teams
- Try to find goalie stop data for specific shot id's (couldn't find 1:1)