NHL Capstone

Nathan Wodarz

September 2021

Outline

- Problem Statement
  - Background
  - Problem
- Data Wrangling
  - Data Sources
  - Data Gathering
- Exploratory Analysis
  - Initial Findings
- Modeling
  - Classification
  - Probability
  - Feature Importance
- Future Directions
  - Home Goals
  - Feature Engineering
- Conclusion
  - Summary

Problem Statement

Problem Statement

Background

- **National Hockey League** (NHL): professional ice hockey league
- 32 teams in the United States and Canada
- Season Format
  - 82 regular season games per team
  - Playoffs involve 16 teams
    - Four playoff rounds
    - Each round is best-of-seven format

Problem Statement
Background

- Object of game: score more goals by shooting puck into net

Problem Statement

Problem

- Average shot attempts per game: over 100
- Average goals per game: around 6
- **Can we determine if a shot attempt will result in a goal?**

Data Wrangling

Data Wrangling
Data Sources

- NHL Live Feed
  - JSON format
  - Updated during and after game

Data Wrangling
Data Sources

- NHL Live Feed


```
}, {
  "players" : [ {
    "player" : {
      "id" : 8476856,
      "fullName" : "Matt Dumba",
      "link" : "/api/v1/people/8476856"
    },
    "playerType" : "Shooter"
  }, {
    "player" : {
      "id" : 8469608,
      "fullName" : "Mike Smith",
      "link" : "/api/v1/people/8469608"
    },
    "playerType" : "Goalie"
  } ],
  "result" : {
    "event" : "Shot",
    "eventCode" : "MIN525",
    "eventTypeId" : "SHOT",
    "description" : "Matt Dumba Wrist Shot saved by Mike Smith",
    "secondaryType" : "Wrist Shot"
```

Data Wrangling
Data Sources

- NHL Play-by-play
  - HTML format
  - Posted after game, not updated

Data Wrangling

Data Sources

- NHL Play-by-play

239	3	SH	10:05 9:55	SHOT	MIN ONGOAL - #24 DUMBA, Wrist, Def. Zone, 172 ft.	77 C	93 C	79 L	7 D	27 D	41 G	14 C	26 L	24 D	25 D	32 G

Data Wrangling
Data Gathering

- Shot attempt data
  - Game meta-information
    - Game location
    - Teams playing
    - Regular season vs. playoffs

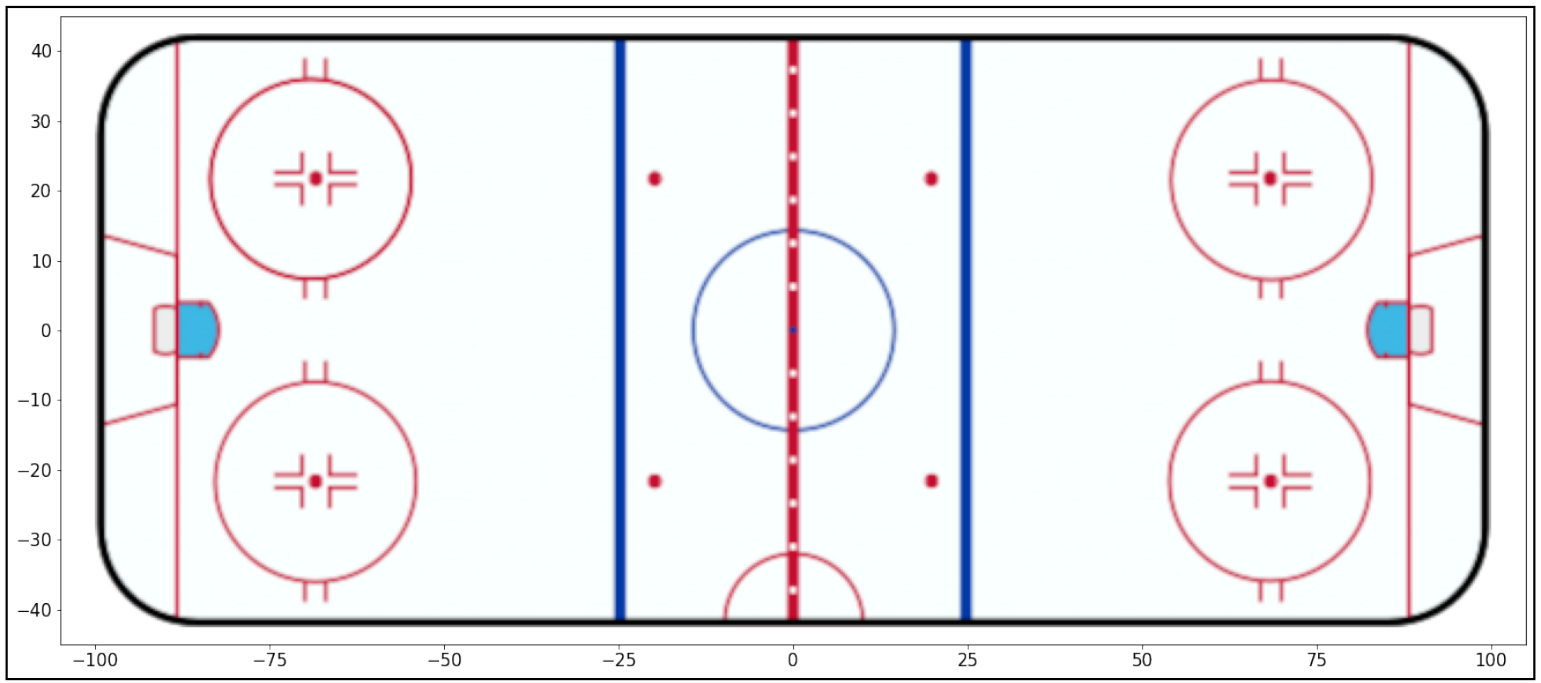
Data Wrangling
Data Gathering

- Shot attempt data
  - Game state
    - Current score
    - Time elapsed/remaining
    - **Number of players on ice**

Data Wrangling

Data Gathering

- Shot attempt data
  - Shot information
    - Location of shot attempt



Data Wrangling
Data Gathering

- Shot attempt data
  - Shot information
    - Type of shot

Data Wrangling
Data Gathering

- Shot attempt data
  - Shot information
    - Result of shot - Goal, Save, Miss, Block

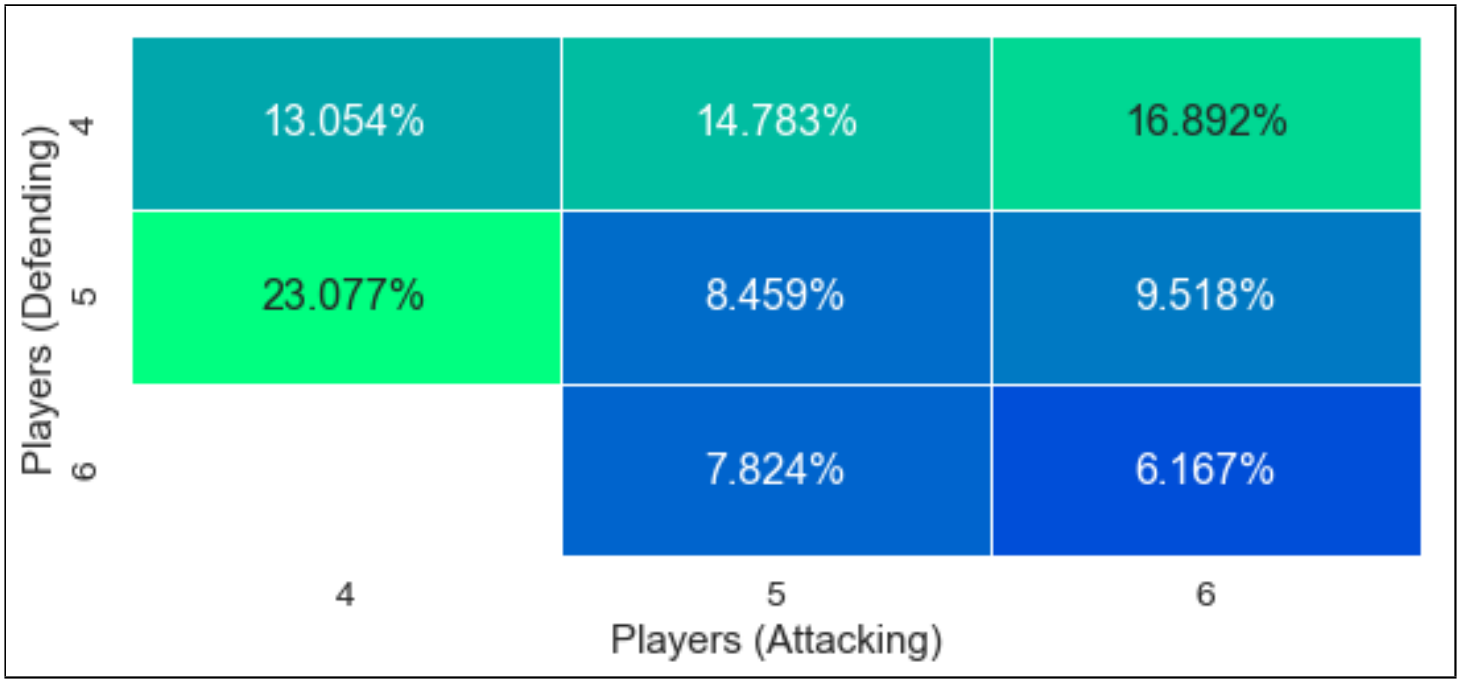
Exploratory Analysis

Exploratory Analysis
Initial Findings

- 1,358 games
- 157,363 shot attempts
- 118,280 non-blocked attempts

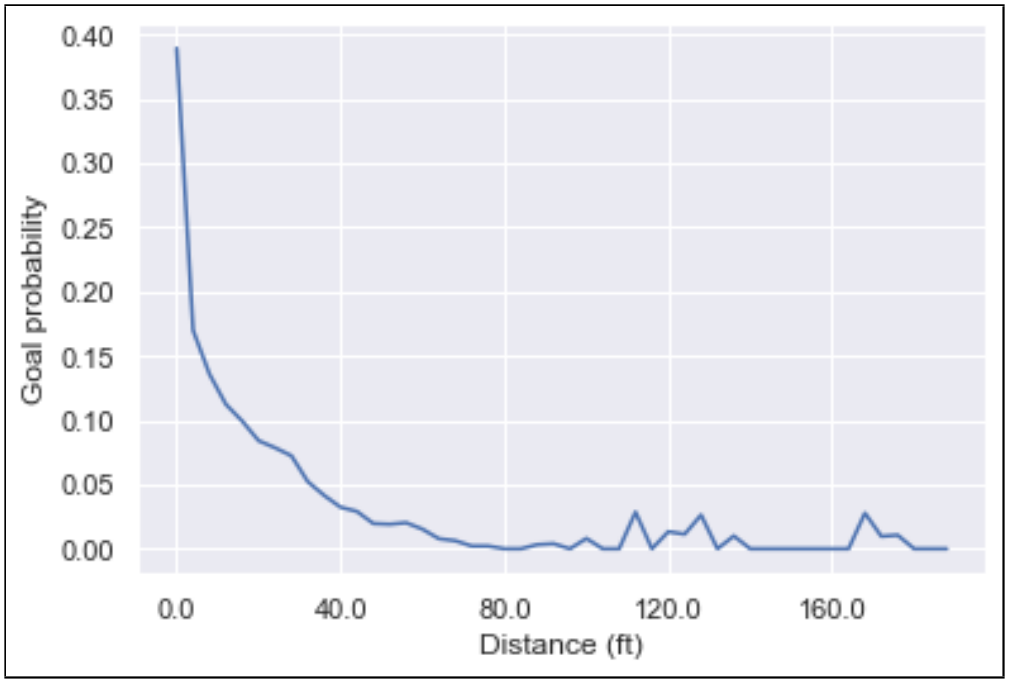
Exploratory Analysis
Initial Findings

- Number of players on ice is important



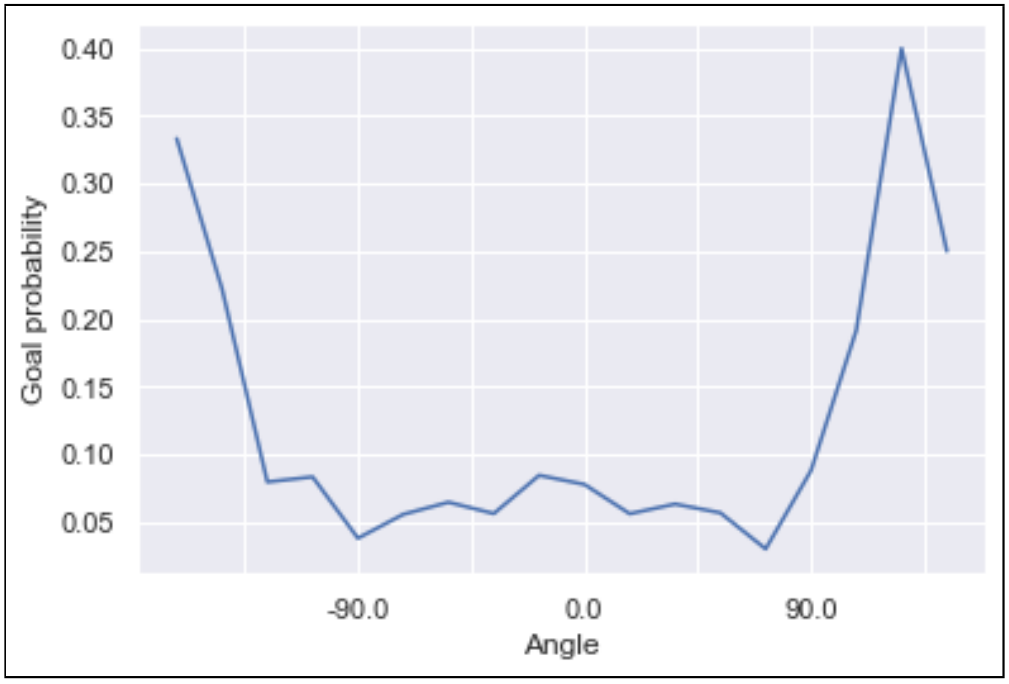
Exploratory Analysis
Initial Findings

- Shot distance is important

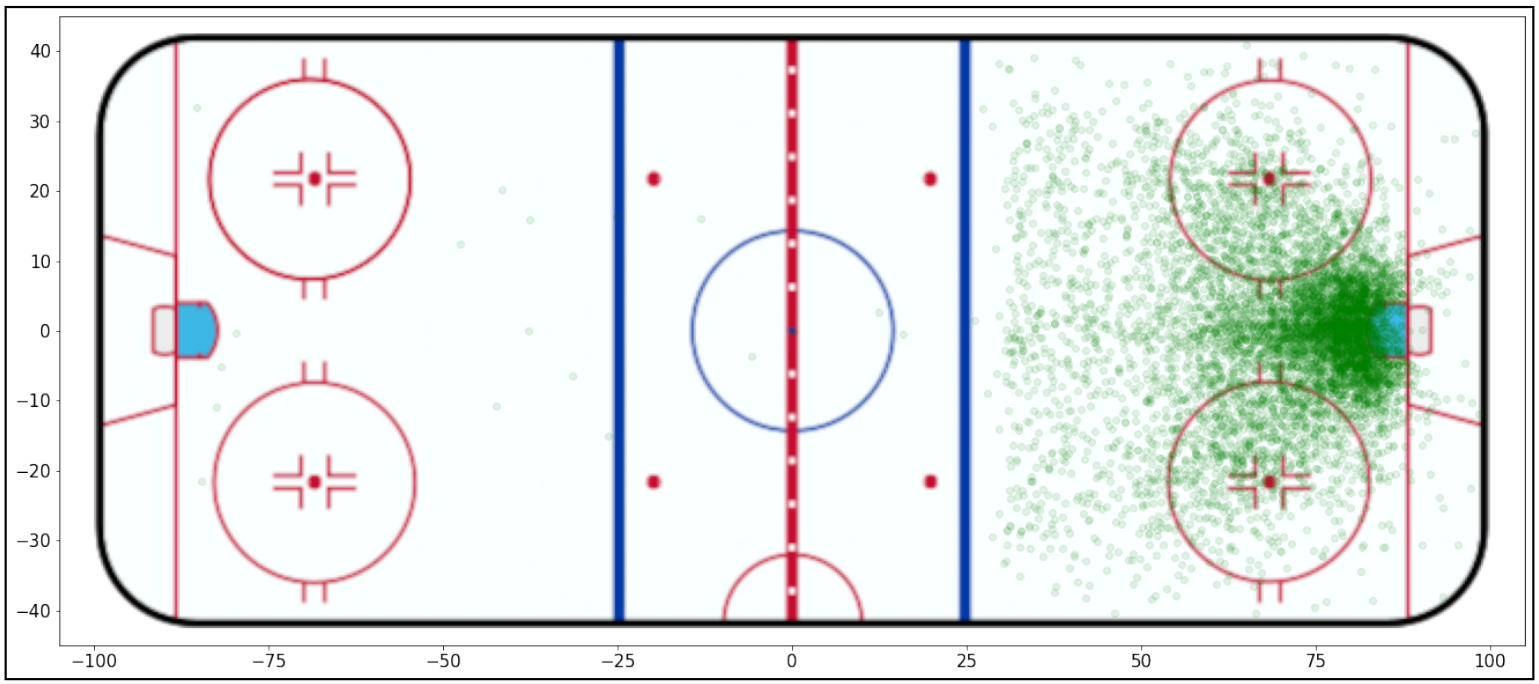


Exploratory Analysis
Initial Findings

- Shot angle is important



Exploratory Analysis
Initial Findings



Modeling

Modeling
Classification

- Initially attempted to classify as goal/no goal
- Classifiers (scikit-learn, imblearn), metric F_1
  - Logistic Regression
  - Random Forests
  - Support Vector Classifiers
  - Gradient Boosting
  - Others (Bagging, Naive Bayes, Ensembles, Neural Networks)
- Sampling
  - Random undersampling/oversampling
  - Synthetic Minority Over-sampling Technique (SMOTE)
  - Adaptive Synthetic Sampling (ADASYN)
  - Tomek Links

Modeling
Classification

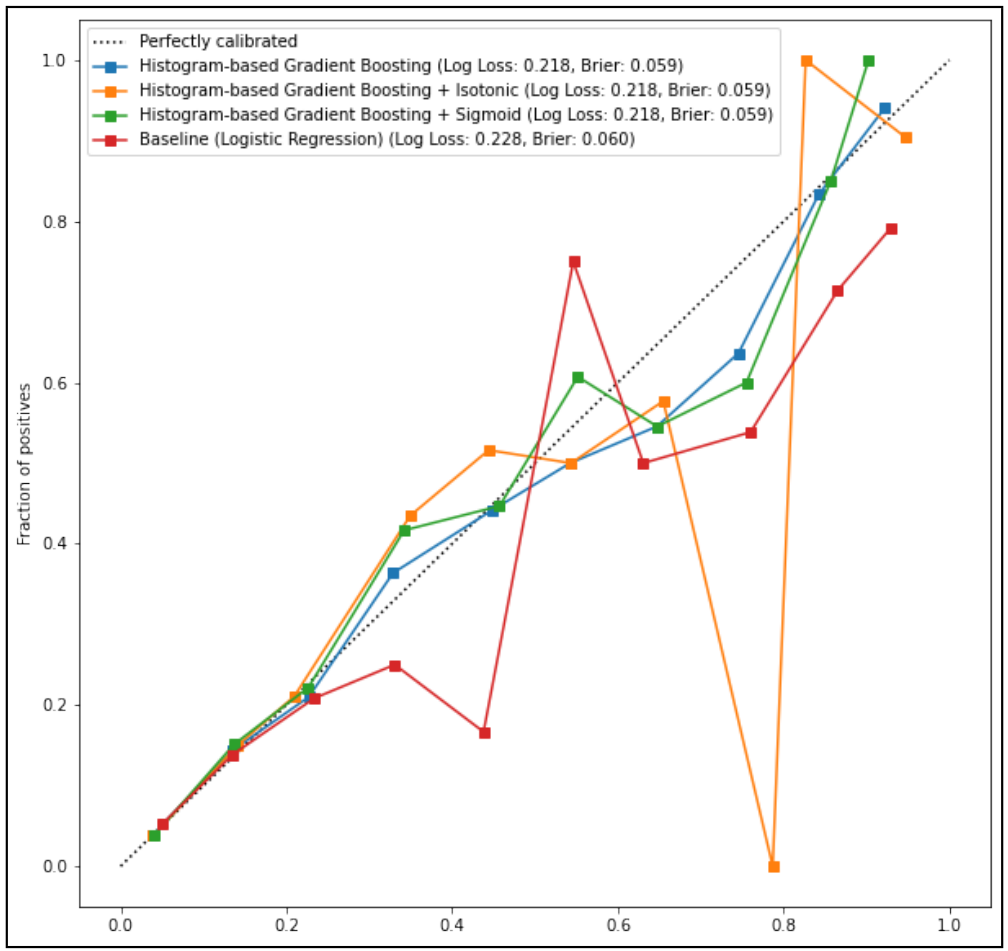
- Results were disappointing
- Best result with Voting Ensemble, ($F_1 = 0.231$)
  - Included Gradient Boost, Logistic Regression, and EasyEnsemble
- Likely issue: No clear boundary between goals and no goals

Modeling
Probability

- Second attempt: Probability metrics
- Metric: log loss
- Classifiers examined with and without probability calibration

Modeling
Probability

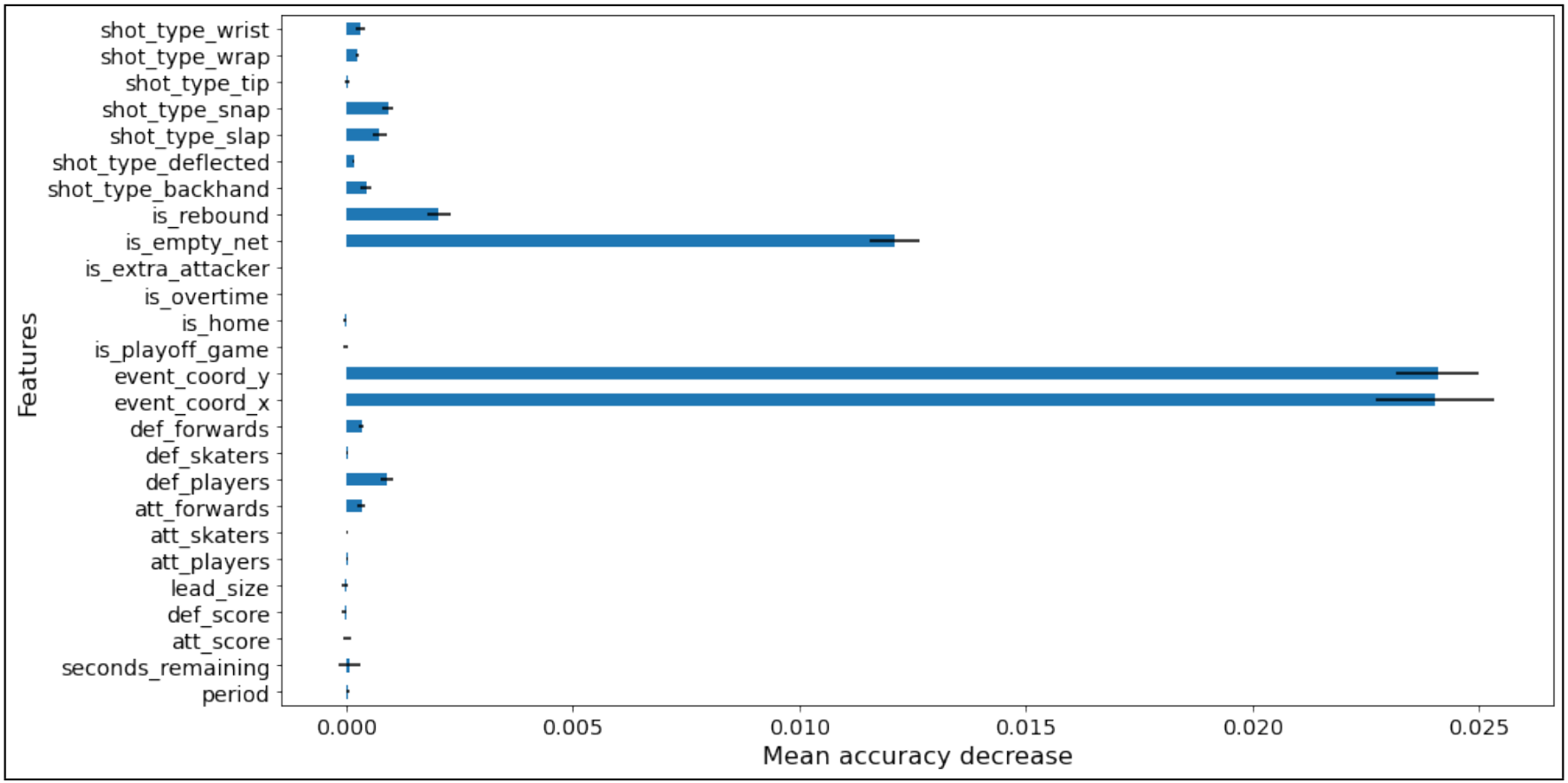
- Best performance: Histogram-Based Gradient Boost, uncalibrated



Modeling

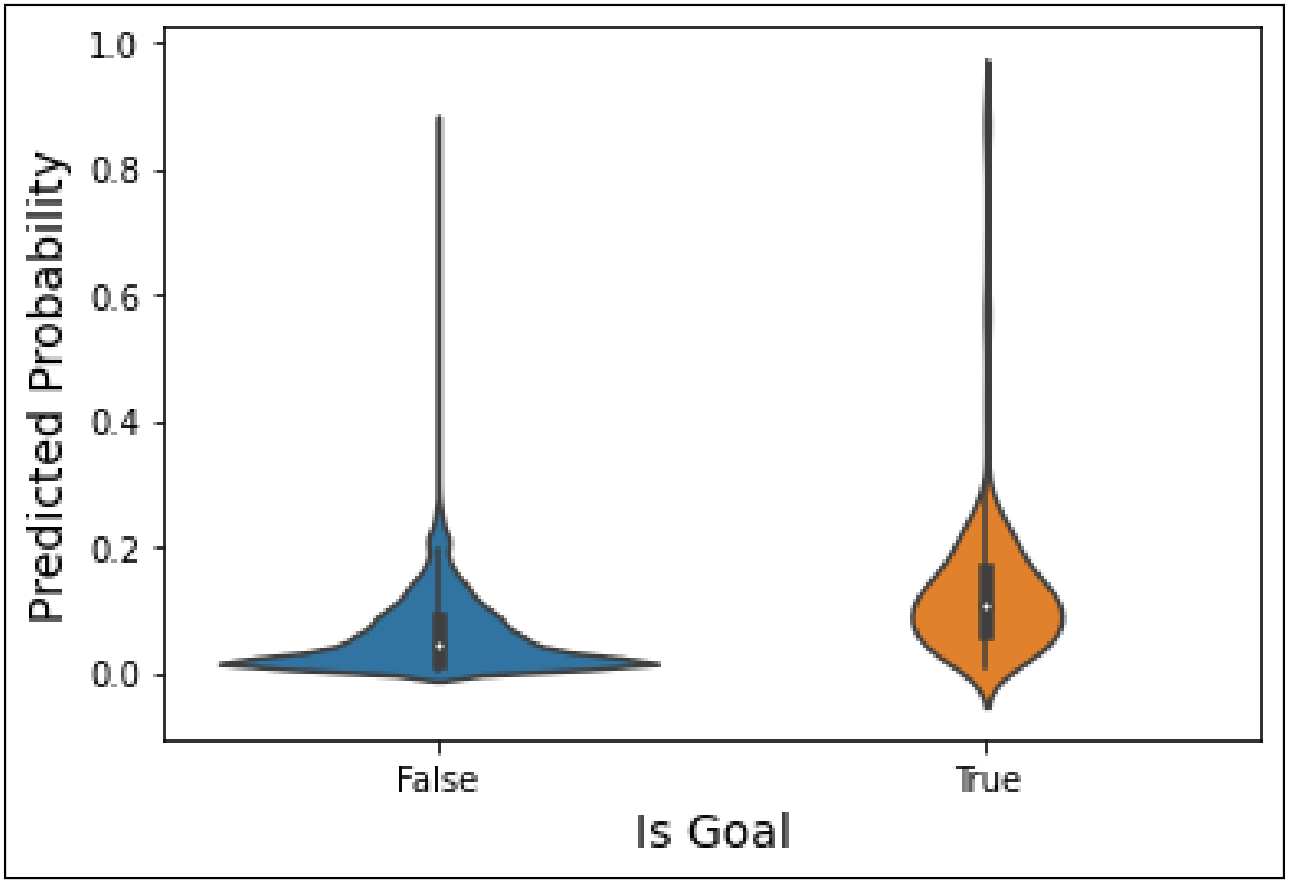
Feature Importance

- Location most important, followed by empty net



Modeling
Feature Importance

- Distribution of no-goals is skewed



Future Directions

Future Directions

Home Goals

- Goals more likely for home team (6.698% to 6.625%)
- Statistically significant difference ($p = 0.019$)
- Why is this the case?
  - Home crowd support
  - Familiarity with arena
  - Rules
- Pandemic allows for testing - games played without crowd and/or at neutral sites

Future Directions

Feature Engineering

- Multiple models, corresponding to game state
- Binning shot positions

Conclusion

Conclusion

Summary

- Probability metrics provided better results
- Best performance from uncalibrated Histogram-Based Gradient Boost
- Splitting models may provide better future performance

Conclusion

Thanks!

- Nathan Wodarz
- Email: nmwodarz@gmail.com
- LinkedIn: https://www.linkedin.com/in/nathanwodarz/
- Github: https://github.com/nmwodarz

- Project Report