NHL Capstone Report

Nathan Wodarz

Abstract: Write your abstract first and then write it again last. Your abstract should grab your reader's attention and also tell them what the paper is about in plain English. Avoid using jargon, undefined terms, or acronyms in your abstract; if you must, define them. Write your abstract before your paper in order to structure your thinking and then rewrite it when you are finished with your paper to make sure it is tight, clean, and accurately portrays what the paper is about. Include only what is absolutely necessary. Tell the reader what you set out to do, what you did, and the results and conclusions. Keep it brief. This abstract is 121 words long, but yours should not exceed 200 words.

Contents

1. Introduction

The National Hockey League (NHL) is a professional ice hockey league with 32 teams in the United States and Canada. Ice hockey is a sport played on a rink of ice with a small rubber cylindrical puck one inch thick and three inches in diameter. Players use long curved hockey sticks to attempt to score goals. A goal is scored when a player shoots the puck into a net which is 4 feet tall and 6 feet wide (and 40 inches deep). The object of an ice hockey game is to score more goals than the opponent.

1.1. Problem Statement. Given information about the game state, we wish to determine if it is possible to predict whether a shot successfully scores a goal.

1.2. Background. The National Hockey League (NHL) is a professional ice hockey league with 32 teams in the United States and Canada. Teams are divided into two conferences, which are each further divided into two divisions of eight teams each. In a typical season, each team plays 82 regular season games – 41 of these at the team's home venue, 41 of them at other teams' venues. At the end of the season, eight teams from each conference advance to the Stanley Cup playoffs. The top three teams from each division qualify for the playoffs as do the top two remaining teams in the conference. Brackets are fixed before the playoffs begin. In each round of the playoffs, opposing teams play a best-of-seven format with the winner advancing to the next round. The last team remaining is awarded the Stanley Cup at the conclusion of the playoffs.

Each game consists of three periods of twenty minutes. If teams are tied after the third period, play continues. In the regular season, a five minute sudden death overtime is played. If the game remains tied after the overtime, a shootout is held to determine a winner. In the playoffs, the overtime format is adjusted so that each overtime period is twenty minutes in length and play continues until a goal has been scored. No shootouts occur during the playoffs.

Each team is generally allowed to have six players on the ice at any time, with at most one *goaltender* (or *goalie*). Other players are referred to as *skaters*. The goalie typically plays in front of his team's net and attempts to physically block shots taken by the opposing team. The goalie is allowed to wear larger pads and to use a larger stick than other players. Skaters are generally classified into positions as either *forwards* or *defense*, with forwards further subdivided into *centers*, *left wings*, and *right wings*. These classifications are strategic and tactical as opposed to rule-based and there are no limits to the number of any position on the ice besides goaltenders. During the regular season, teams are limited to four players per side during the overtime period. Unlike most sports, teams are allowed to make substitutions while play is in progress. The teams have benches along the side of the rink in the neutral

zone (see the next subsection for definition).

If a player commits a penalty, they are removed from play for a length of time which depends upon the infraction. This is typically two or five minutes, but other lengths are possible (most commonly four minutes, ten minutes, or the remainder of the game). In most cases, teams may not substitute for a penalized player while the penalty is being served, although this is limited to a minimum of four players on the ice (in regular season overtimes, the opposing team is allowed an extra player instead). If this results in uneven numbers, the team with more players on the ice is said to have a *power play* while the other team is *short-handed*. In all other circumstances, the teams are said to be at *even strength*. At the referee's discretion, a *penalty shot* may be awarded in lieu of penalizing a player from the opposing team. This typically happens when the penalty interrupted a play which the referee feels had a high probability of resulting in a goal. During the penalty shot, the clock is stopped and only one player from each team is allowed on the ice. The player from the team awarded the penalty shot has one opportunity to try to score. This is a relatively uncommon circumstance during regular play, however the shootout uses this format.

The game strength is often indicated as m-on-n, where m is the number of skaters for the attacking team and n the skaters for the defending team. The most common game state is 5-on-5, while 5-on-4 is the most common power play state. Since the goalie may be removed from play, any possibility with $3 \leq m, n \leq 6$ may occur.

The NHL makes available play-by-play information on nearly every game dating back to the 2010-11 season. This includes information on all shot attempts during the game. This allows determination of the players on the ice when the shot was taken, the position on the ice that the shot was taken from, as well as general information such as the time remaining in the period and the score at the time of the play.

1.3. Hockey Rink. The NHL plays on ice rinks in the shape of a rounded rectangle 200 feet long by 85 feet wide (see Fig. 1). We overlay the rink with a set of axes with the origin at the center of the rink. We consider the x-axis to correspond with the major axis of the rectangle and the y-axis with the minor axis. Units are taken to be in feet. There are several lines marked on the ice which are parallel to (or coincide with) the y-axis. The *center line* is a red line with $x = 0$. There are two *blue lines* at $x = \pm 25$. Finally, there are two red *goal lines* at $x = \pm 89$.

Nets are placed with the open end of the net on the goal line, centered at $(\pm 89, 0)$. The region between the blue lines is referred to as the *neutral zone*. The other regions are defined relative to the team possessing the puck (the *attacking* team). The region containing the goal that the attacking team is attempting to score into is the *offensive zone* with the remaining region being the *defensive zone*.

Most shots originate within the offensive zone. The largest reason for this is the proximity to the net, however there are also rules-based reasons that this is the case. A play is *offside* if one or more attacking players are in the offensive zone while the puck is still outside the zone. If a shot is taken in this circumstance, any resulting goal will not count. Additionally, the icing rule discourages longer shots even if play is not offside. *Icing* occurs when the puck crosses the center line and the attacking goal line without being touched by any player in between. Icing is negated if the attacking team is short-handed or if the puck enters the net. In all other circumstances, play is stopped and a faceoff occurs in the defensive zone.

2. Datasets

2.1. NHL Statistics API. The NHL makes a statistics API available at https://statsapi.web.nhl.com. There is no official documentation for the API, however there are several unofficial versions available. We used the documentation at https://gitlab.com/dword4/nhlapi for this project. The particular endpoints used were `schedule` (which returns information about all games played and/or scheduled in a given time period) and `feed/live` (returns the live feed for a requested game).

2.1.1. Schedule. We retrieved information about every NHL regular season and postseason game during the 2018-19 season. The NHL also maintains data about preseason games and exhibition games (including all-star games and games involving non-NHL opponents). Since these games don't count in the standings, we ignored them for this project. This particular season was chosen as the most recent season which was not interrupted by the COVID-19 pandemic.

The `schedule` endpoint returns a .json file with information about every game in the requested time span. The file has a `dates` field, which holds a list of entries corresponding to individual dates. For each date entry, there is a `games` field containing a list of entries with one for each game played on that date. Finally, each game entry contains information about the teams involved in the game, the game venue, and other information about the game. We are only interested in the `link` field, which contains a string which provides a relative URL to the live feed for that game. For example, the string `/api/v1/game/2017020659/feed/live` refers to a game played between the Minnesota Wild and the Calgary Flames on January 9th, 2018. The live feed is available at https://statsapi.web.nhl.com/api/v1/game/2017020659/feed/live. The substring `2017020659` represents the ID for this game. The first four characters `2017` denote the season in which the game was played (the 2017-2018 season). The next two characters `02` indicate the type of game (`02` indicates a regular season game; `01` is used for preseason, `03` for postseason, and `04` for other games). The final four characters `0659` represent the particular game. These numbers are sequential for all types of games except for postseason games. In the postseason, the first character of this grouping will always be a `0`. The second character indicates the playoff round (from `1` to `4`, the third character the individual series within that round (`1` to `8` in the first round, `1` to `4` in the second round, `1` to `2` in the third round, and `1` in the final round). The final character indicates the game number within that series (NHL playoff series are best-of-7 games, allowing the final character to range from `1` to `7`).

2.1.2. Live Feeds. We retrieved a live feed for each game ID returned in the schedules. Each live feed contains the information found in the schedules, as well as information about the rosters of both teams and game events. For this project, we required information on shot attempts (coded in the live feed as `"Shot"`, `"Missed Shot"`, `"Blocked Shot"`, or `"Goal"`) as well as faceoffs (coded as `"Faceoff"`). The included data depends

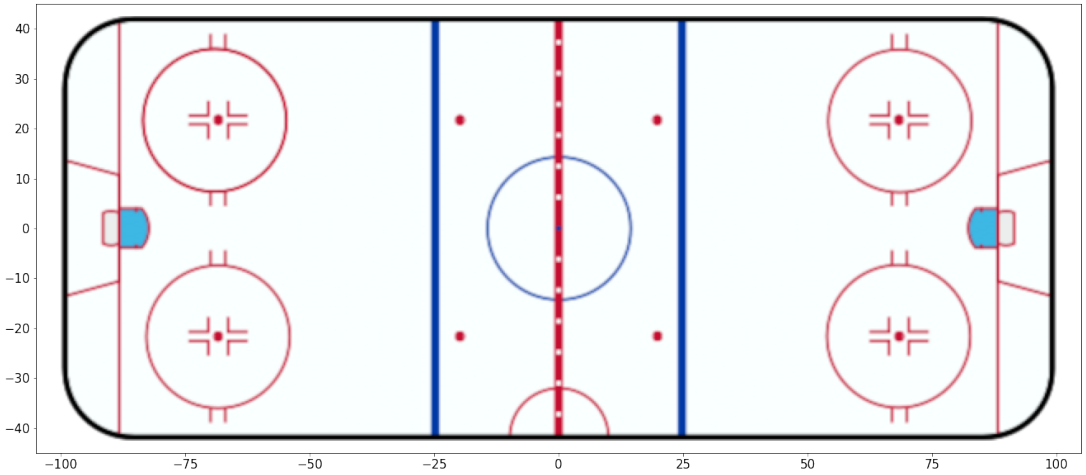

Fig. 1. Standard NHL rink with coordinates

on the event type. For shots (excepting blocked shots), the information present includes the following (see Fig. 2).

- The shooting player
- The goalie
- The result of the play (shot, missed shot, or goal)
- The shot type (e.g., slap shot, wrist shot, etc.)
- The game period
- The time elapsed and time remaining in the period
- The score at the conclusion of the play
- Coordinates of the shot attempt, using the coordinate system described in the previous section
- The event team (the team taking the shot)

2.2. Play-by-Play. The NHL also provides HTML play-by-play reports for each game. These reports contain additional information about plays that is not found in the game live feed. These can be found by using the game ID to construct a URL at the official NHL web site. The game report for the game wtih ID `2017020659` is http://www.nhl.com/scores/htmlreports/20172018/PL020659.HTM. The season characters are separated from the ID and written as an eight-character season ID `20172018`. The remaining part of the ID is given the prefix `PL` (for "play-by-play") and the suffix `.HTM`.

The report provides a row for each play (see Fig. 3 for an example). Unlike the live feed, the HTML report gives the game strength. The report indicates the strength from the point of view of the team with the event, using `EV` for even-strength, `PP` for power play, and `SH` for short-handed.

Additionally, the HTML report provides the sweater numbers of each player on the ice and the position they play. Usually, teams have five skaters and one goalie on the ice. Teams are allowed to remove that goalie and substitute another skater at any time. A team that does this is said to be *pulling their goalie*. This allows an extra attacker but leaves the team's net undefended. Since the HTML report contains information on player positions, it can be used to determine whether a team has done this.

3. Data Cleaning and Wrangling

We obtained all live feeds and play-by-play reports from the 2018-19 season. There were 1,358 games to analyze. For each game, we extracted shot attempts and faceoffs. *Faceoffs* are used to start play and to restart after each stoppage. A game official drops the puck between one player from each team, with those players attempting to gain possession and pass the puck to a teammate. The faceoffs were used in classifying shot attempts as rebounds. Following http://blog.war-on-ice.com/annotated-glossary, a rebound was considered to be any shot taken within three seconds of the previous shot as long as no play stoppage occurred between shots. Play always begins with a faceoff, so the presence of a faceoff attempt was taken to indicate that play had been stopped. Once rebounds were indicated, the faceoff information was discarded. This left us with 160,032 individual shot attempts for the season. This included penalty shot and shootout attempts, which don't occur within the normal game flow. These were discarded, leaving 157,363 attempts. We also decided to remove blocked shot attempts, since the location data for blocked shots indicates where the block occurred and not where the shot occurred. This left 118,280 attempts remaining.

At this point, the data in the frame divides teams by home/away. There are several columns which use the home/away status. These columns will be split roughly evenly between the attacking and defending teams. For consistency, we chose to move the home/away information into a separate column and reclassify any dependent columns on attacking/defending. This meant that no column would refer to both attacking and defending teams.

4. Exploratory Data Analysis

4.1. Initial Findings. As mentioned earlier, teams usually play with five skaters and a goaltender. As a result, most shot attempts should occur at even strength. For remaining shot attempts, teams on the power play have a numerical advantage. Logic would indicate that more shots are attempted on the power play than while short-handed. As indicated in Fig. 4, this is indeed the case. Fig. 5 further divides these counts by considering the number of players on the ice for each shot. The counts shown in this figure don't quite agree with the numbers

```
}, {
  "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"
  },
  "about" : {
    "eventIdx" : 241,
    "eventId" : 525,
    "period" : 3,
    "periodType" : "REGULAR",
    "ordinalNum" : "3rd",
    "periodTime" : "10:05",
    "periodTimeRemaining" : "09:55",
    "dateTime" : "2018-01-10T03:21:10Z",
    "goals" : {
      "away" : 2,
      "home" : 1
    }
  },
  "coordinates" : {
    "x" : 82.0,
    "y" : 15.0
  },
  "team" : {
    "id" : 30,
    "name" : "Minnesota Wild",
    "link" : "/api/v1/teams/30",
    "triCode" : "MIN"
  }
}, {
```


Fig. 2. Shot attempt as recorded in live feed

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

Fig. 3. Shot attempt from Fig. 2 as recorded in HTML play-by-play

used in Fig. 4, although they are close. This can be seen in more detail by looking only at player counts when the play-by-play report indicates that the teams are at even strength. These are shown in Fig. 6. This directly indicates that the game state indicated in the play-by-play report sometimes disagrees with the reporting of the number of players on the ice. When teams are at even strength, all non-diagonal entries should be 0. Impossible situations are also seen in power play shots (Fig. 7) and short-handed shots (Fig. 8). Player counts are also pulled from the play-by-play report. We did find some circumstances in these reports where the player counts were impossible, having more than six or fewer than four players on the ice. Obviously impossible situations were imputed with median values. All other player counts were kept and used to infer the strength from the counts as opposed to directly using the strength from the play-by-play report.

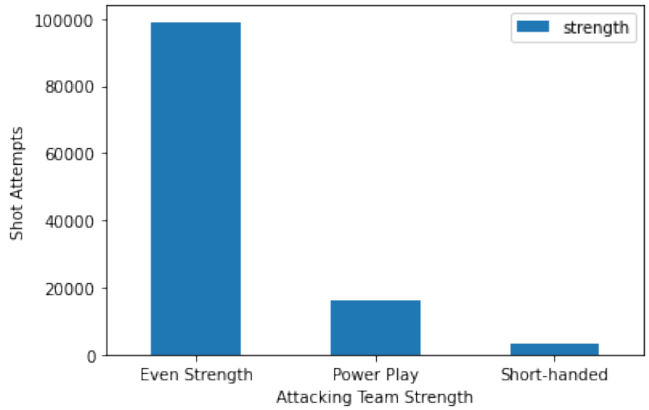

Fig. 4. Shot attempts by strength

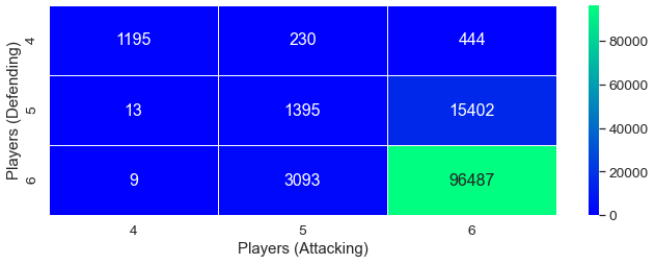

Fig. 5. Shot attempts by players on ice

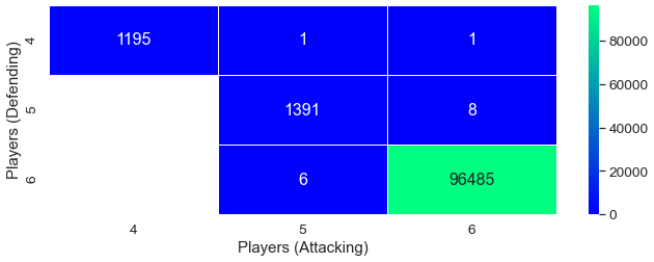

Fig. 6. Even strength shot attempts by players on ice

One would expect that goals are more likely when on the power play and Fig. 9 confirms this. While shots are successful

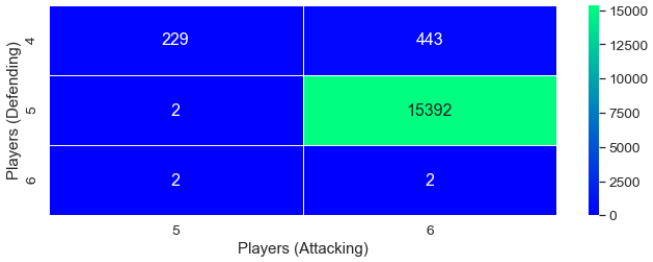

Fig. 7. Power play shot attempts by players on ice

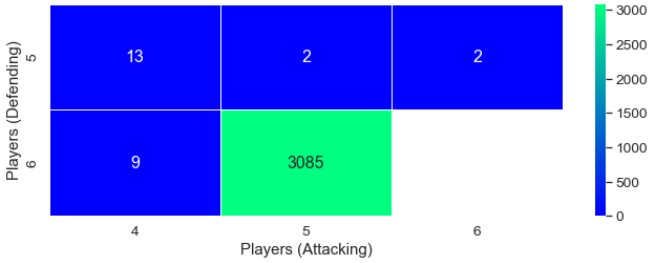

Fig. 8. Short-handed shot attempts by players on ice

slightly more than 6% of the time at 5-on-5, this increases to over 9.5% at 5-on-4 and almost 17% when 5-on-3.

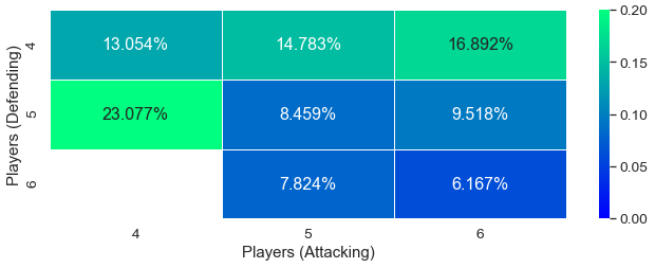

Fig. 9. Goal probability by players on ice

4.2. Discrepancies. Using the coordinates of the shot events, we calculated the distance of each shot and compared this to the official value indicated in the HTML report (Fig. 10). There were a number of shot attempts which seem to have the official distance as calculated using the incorrect goal. This is indicated by the line with negative slope. There is another cluster where either the official or calculated distance of the shot is close to zero. This is likely to have a different cause.

In Fig. 11, we compare the official distance against the absolute difference between the two distances. This appears to confirm our hypothesis that a sizable minority of shot attempts used the incorrect goal to calculate the distance. In particular, a number of shots indicated in the play-by-play as being from the defensive zone are indicated to be impossibly short. No shots less than 60 feet from the goal can be correctly coded as defensive zone. There are a number of shots where coordinates and distance agree where the zone is obviously miscoded. This can be fixed by recoding the zone using the coordinates. The coordinates are more likely to be correct, as the HTML report is posted immediately postgame and not updated, while the live feed can be updated at a later time.

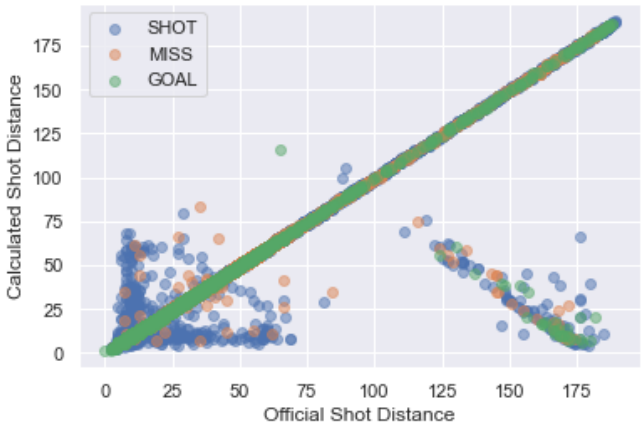

Fig. 10. Official vs. calculated shot distance

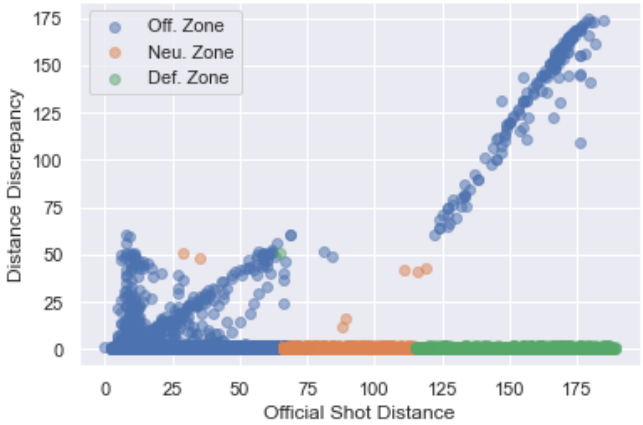

Fig. 11. Official distance vs. discrepancy

There are two other patterns seen in the shots. The first is a line with slope near 1 in the offensive zone plot. The other is a near vertical-line associated with very short distance shots. We initially examined the line with slope approximately 1. For this, we looked at only shots longer than 25 feet within the offensive zone. Interestingly enough, most of the shots considered here seem to have coordinates near $x = 80$ and $y = 0$ (Figs. 12 and 13). This is just in front of the goal, which suggests that these might be shots which were found after-the-fact to have actually been deflected shots. This is a not-uncommon determination in the NHL, which results in an adjustment to the shooting player and distance in the statistics. As above, the live feed is more likely to be correct. We can also resolve this concern by preferring the statistics from the live feed.

Next, we examined short-range shots (Fig. 14). Toward the bottom of this plot, it is possible to see the overlap with the preceding analysis. Consequently, we replotted the data to exclude shots where the coordinates report very short distances, indicating that they are probable deflected shots. This replot is seen in Fig. 15. There was no obvious correction seen when plotting the coordinates (Figs. 16 and 17). We chose to prefer the live feed coordinates here as well.

After recoding zones and recalculating distances where necessary, we see that most of the discrepancies called out for

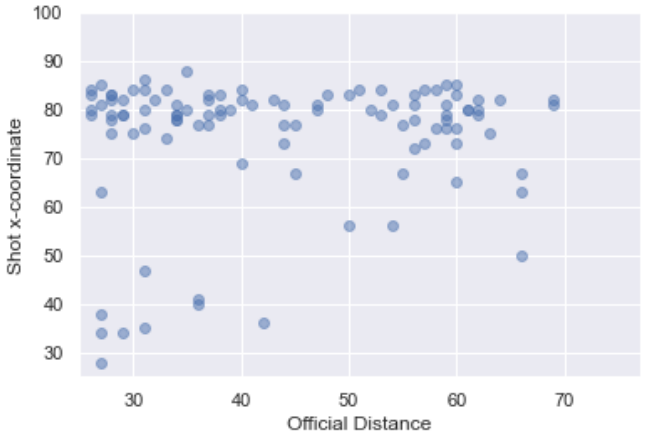

Fig. 12. Official distance vs. x-coordinate (medium range shots)

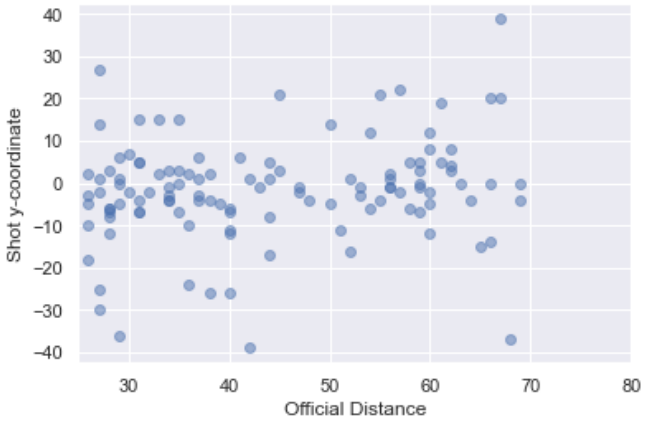

Fig. 13. Official distance vs. y-coordinate (medium range shots)

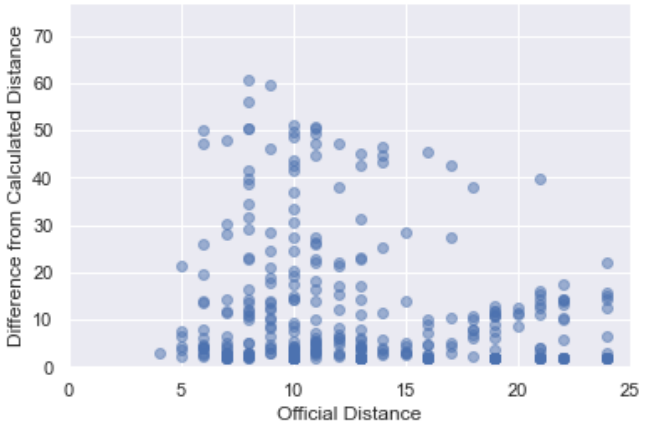

Fig. 14. Official distance vs. discrepancy (short range shots)

correction have disappeared (Figs. 18 and 19). We chose to prefer the shot coordinates and used those to calculate shot distances explicitly from this point on.

4.3. Goal Probability. Figs. 20, 21, and 22 all indicate the shot locations of attempts resulting in goals. Unsurprisingly, most successful shots are taken in a short distance in from of the

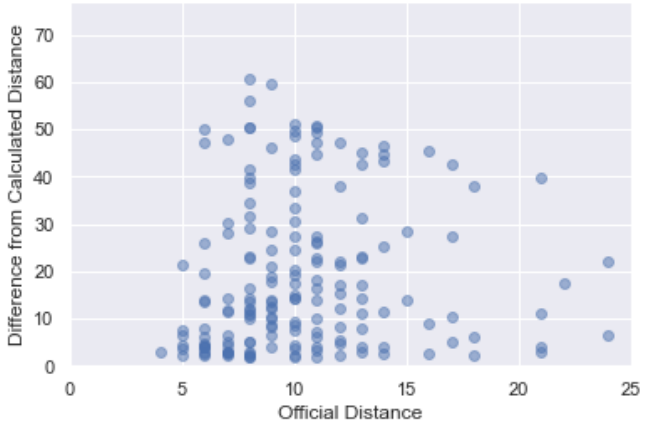

Fig. 15. Official distance vs. discrepancy (short range shots, probable deflections omitted)

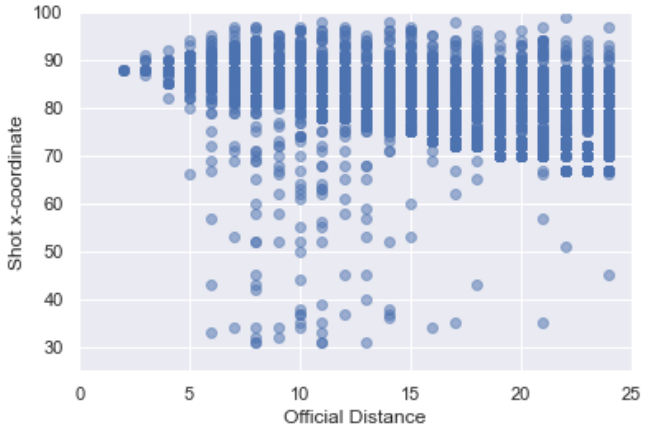

Fig. 16. Official distance vs. x-coordinate (short range shots, probable deflections omitted)

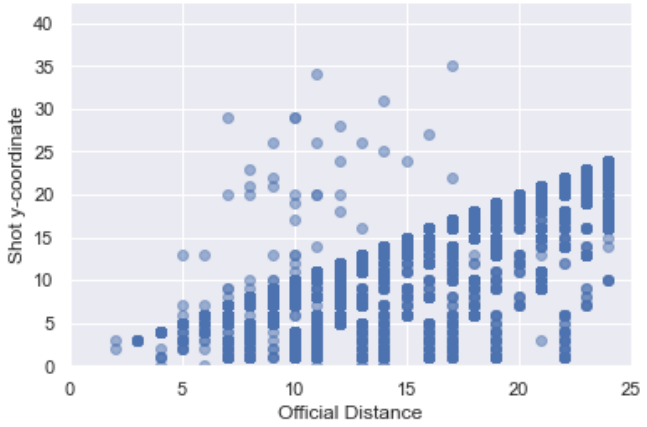

Fig. 17. Official distance vs. y-coordinate (short range shots, probable deflections omitted)

goal. This is confirmed by examining the probability of scoring a goal over all shots in Fig. 23. Shots at extreme angles seem to be more successful as well, as seen in Fig. 24. One interesting sidenote is that 5-on-5 goals are more likely to occur in the

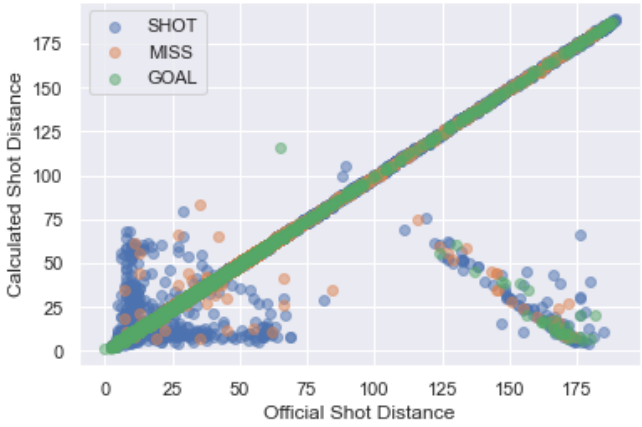

Fig. 18. Official vs. calculated shot distance (corrected)

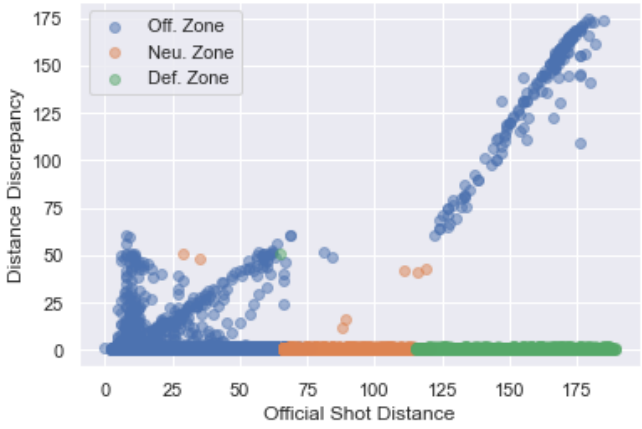

Fig. 19. Official distance vs. discrepancy (corrected

second period than in the first or thirds. The dataset shows a 6.009% success rate in the second period while there are only 5.614% and 5.705% chances in the first and third periods respectively. There are fewer shot attempts in the third period than the other two periods (32,118 in the first, 32,216 in the second, and 29.239 in the third), however the difference in goal likelihood between the first and second periods is statistically significant ($p = 0.027$). The difference between second and third periods was not found to be significant over the course of this season ($p = 0.108$), but there are ruled-based reasons why second-period goals may be more likely. In particular, the teams change sides between periods. In the second period, goaltenders are further from their team's bench than in the first period. This makes it more difficult to make substitutions during play without allowing a scoring chance.

Home goals are also more common than road goals. Home teams saw a success rate of 6.968% overall, while visiting teams converted at a rate of 6.625%. This is statistically significant, indicating that there is an advantage to playing at home. Some of this is rule-based, as there are rules relating to substitutions that favor the home team. Some may be due to the environment as well. Future investigation could examine whether this difference is maintained in neutral ice games, where there are still designated home and visiting teams. Moreover, the 2020-21 NHL season offers a unique

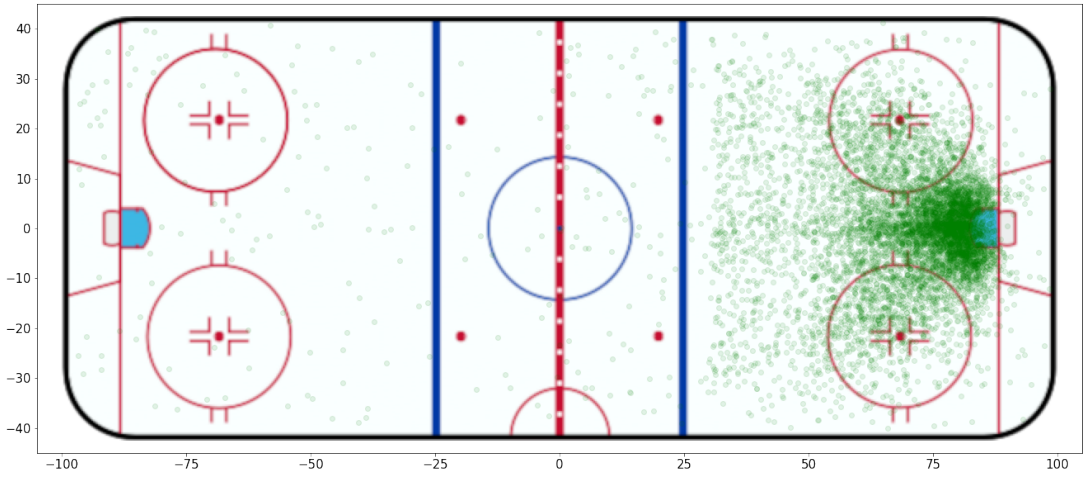

Fig. 20. Location of goals

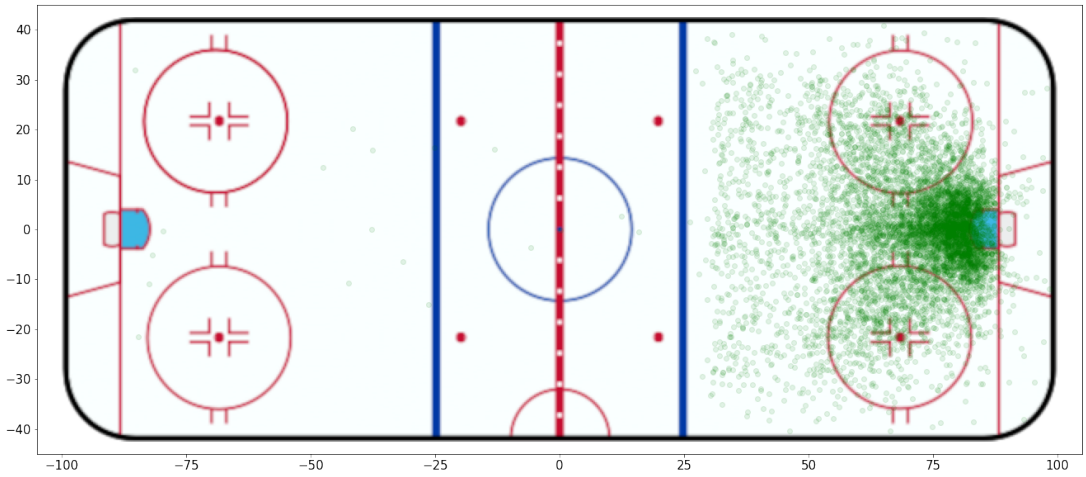

Fig. 21. Location of goals, non-empty-net only

opportunity to explore this, as many games were played in empty or partially-empty arenas.

Finally, we note that rebound shots were far more likely to result in goals (16.844% to 5.993%, $p < 0.001$). This corresponds with expectation.

5. Modeling

5.1. Model Selection. Our ultimate aim was to develop a model to determine whether a particular shot attempt results in a goal. We take the result of the attempt (goal vs. no goal) as the target feature and attempt to use other features to predict whether an attempt was a goal. While the data seems to be appropriate for supervised classification, there are properties to the data that make classification less than ideal. In particular, there is no clear boundary between the two classes in the target feature. Attempts to fit models to the data, even with additional oversampling or undersampling, generally led to models that were not clearly better than a naive random approach. As a result, we abandoned classification and shifted our viewpoint to examine the predicted class probabilities produced by the models.

We tracked two metrics, the log loss and the Brier score loss. The *log loss* (or *cross-entropy loss*) is defined as follows. Let x be a data point in the full dataset X. For any x, let $p_x = p(x)$ be the probability predicted for x by the model and let $y_x = y(x)$ be the actual result of the shot attempt. The result is given by $y = 1$ if the shot resulted in a goal and $y = 0$ otherwise. The log loss L is given by the formula

$$L = -\frac{1}{|X|}\sum_{x\in X} y_x \ln(p_x) + (1 - y_x)\ln(1 - p_x)$$

The log loss is a non-negative value and is minimized precisely when $p_x = y_x$ for all $x \in X$. That is, the minimum value requires the model to correctly predict each data point as well as assign a probability of either 1 or 0 to all points. We wish to minimize the log loss.

The *Brier score loss* B is given by

$$B = \frac{1}{|X|}\sum_{x\in X}(p_x - y_x)^2$$

This is also easily seen to be non-negative and minimized precisely when $p_x = y_x$ for all x. Although we tracked both the log loss and the Brier score loss, we used the log loss for selection purposes. Since the log loss is used to fit the logistic regression model, this provides an obvious baseline for comparison.

5.1.1. Dummy Classifiers. Before examining the logistic regression, we used two dummy classifiers. The first determined the proportion $\hat{p}$ of successful goals in the data set and assigned a

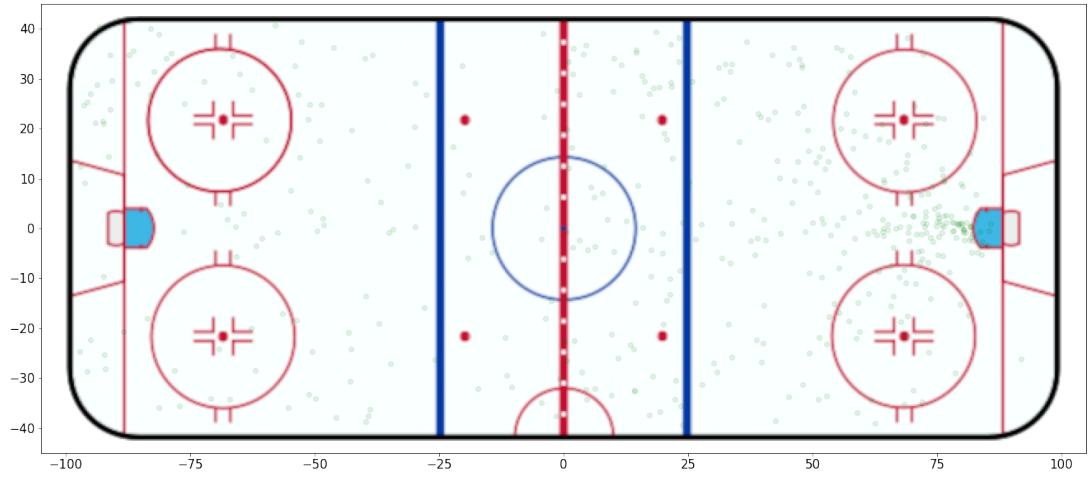

Fig. 22. Location of goals, empty-net only

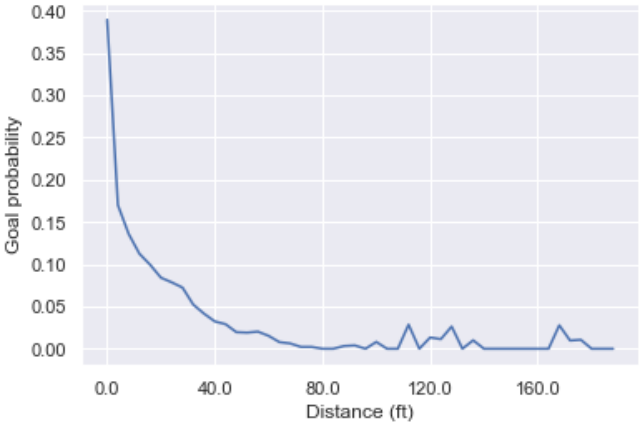

Fig. 23. Probability of goal by distance

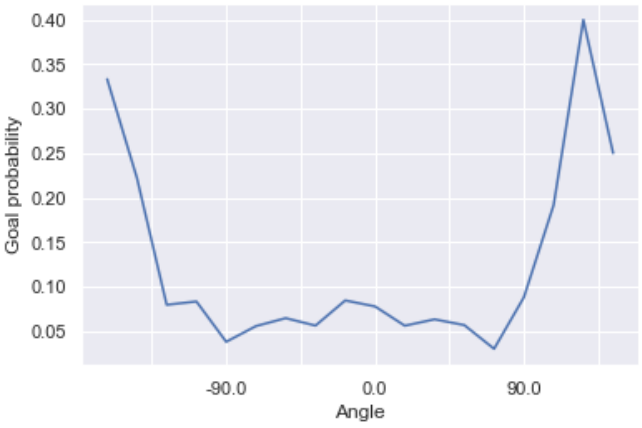

Fig. 24. Probability of goal by angle (in degrees)

probability of $\hat{p}$ to every observation. The second assigned a probability of 1 to a random selection of observations and 0 to the rest. The proportion of the population assigned 1 was equal to $\hat{y}$. The first dummy classifier was distinctly better, giving a log loss of 0.248.

5.1.2. Logistic Regression. We proceeded to examine logistic regression. Using a model with default settings resulted in a log loss of 0.228, improving on the dummy classifier. We examined not only logistic regression by itself, but with various oversampling and undersampling techniques as well. None of the oversampling or undersampling methods provided any noticeable improvements over base logistic regression. This remained true when applying probability calibration methods. Calibration was performed using the `CalibratedClassifierCV` function within scikit-learn. Both isotonic and sigmoid calibration were performed for all models. The results of calibration for the default logistic regression are shown in Fig. 25, while the results with the best performing sampling technique (SMOTE) are in Fig.26.

Because oversampling and undersampling added little to the default models, we dropped them for other models. Some of the models did incorporate sampling techniques into the model algorithm, but no sampling method was explicitly added to any estimator.

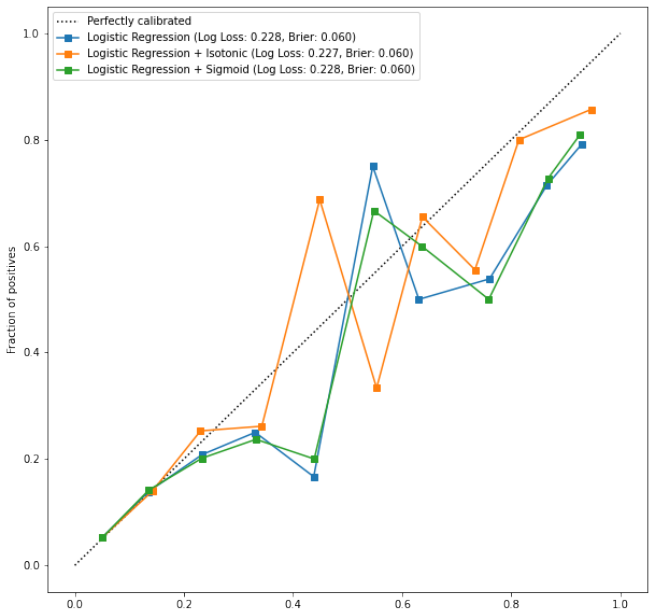

Fig. 25. Probability Calibration - Default Logistic Regression

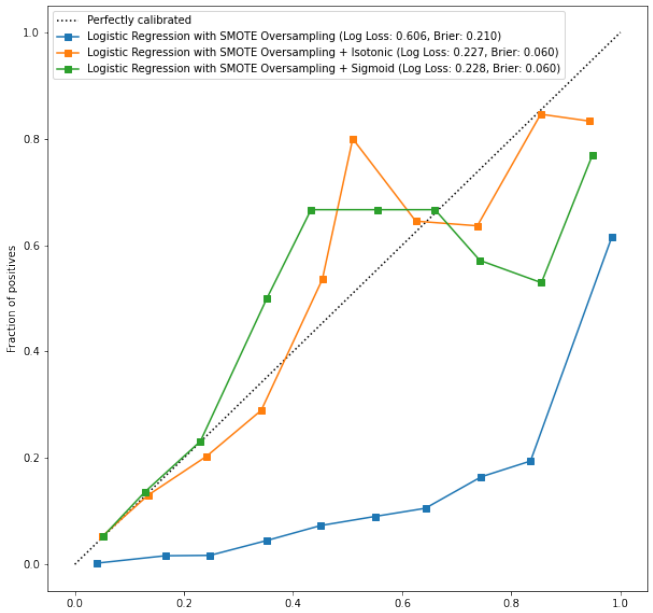

Fig. 26. Probability Calibration - Logistic Regression with SMOTE Oversampling

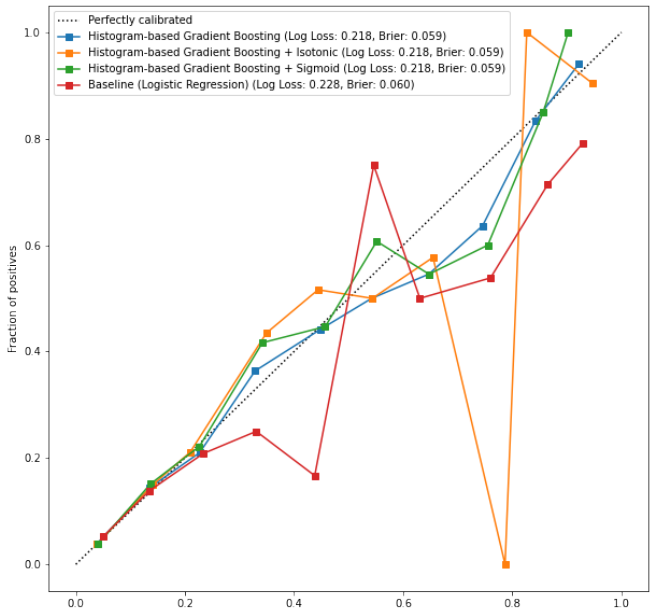

Fig. 27. Probability Calibration - Histogram-based Gradient Boosting

5.1.3. Other Classifiers. We took the default logistic regression as a baseline and directly compared it to each remaining estimator used. A support vector classifier was initially trained, but took calibration would have taken hours. As a result, we removed the model from the collection of estimators assessed.

We examined the following classifiers

- Random forest
  - Default random forest
  - Random forest with balanced class weights
  - Balanced random forest (incorporates random undersampling)
- Histogram-based gradient boosting
- Balanced bagging
  - Default implementation
  - Base estimator of histogram-based gradient boosting
- RUSBoost
- EasyEnsemble
- Gaussian native Bayes
- Nearest neighbors

The best-performing model from these was the default histogram-based gradient boost estimator, with results seen in Fig. 27. The log loss was found to be 0.218. The calculated Brier score loss of 0.059, which also happened to give the best result from all classifiers considered.

5.2. Model Tuning. Having selected a model, we used cross-validation to set hyper parameters. Hyper parameters considered were the learning rate, maximum iterations, maximum depth, and L_2 regularization. The resulting selection scored approximately the same as the default model (see Fig. 28).

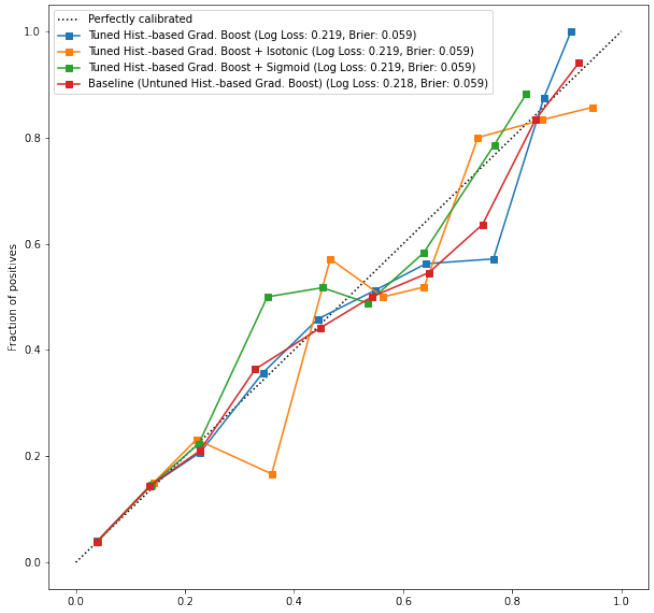

Fig. 28. Probability Calibration - Histogram-based Gradient Boosting after Hyper Parameter Fitting

5.3. Model Interpretations.

5.3.1. Feature Importance. Using a permutation importance algorithm, we determined the most important features for the fitted model. The results are shown in Fig. 29. The relative importance of the most important features were obvious. We found that the coordinate features were the most important features, with the feature `is_empty_net` easily in third place. Other features showing as important are the shooter's shot-type selection (all feature names prefixed with `shot_type`), whether the shot attempt was a rebound and the number of defensive players on the ice. The game situation (time

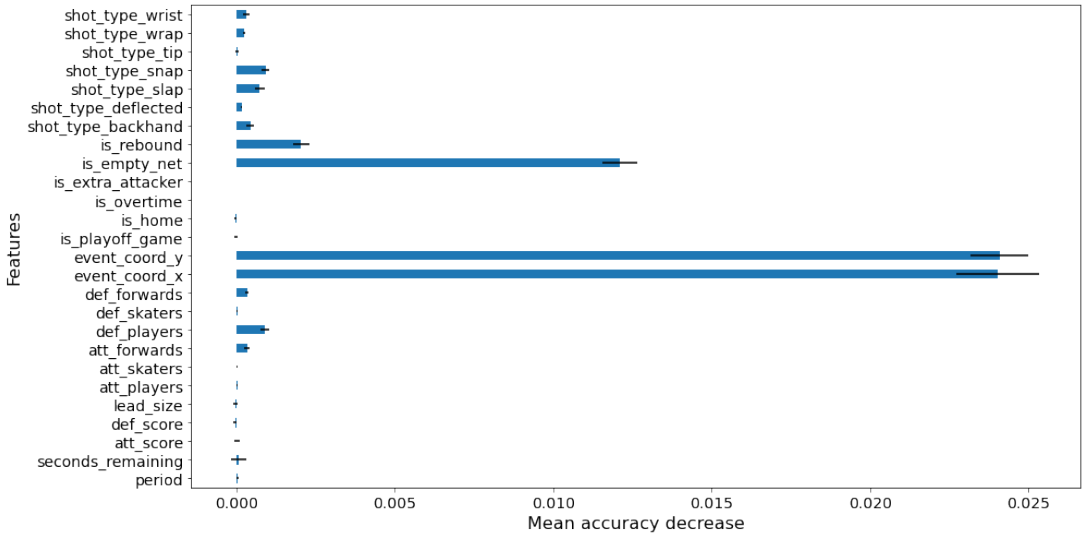

Fig. 29. Feature Importance

remaining, period, score) seemed to have little impact on the model, nor did the number of attacking players on the ice. Additionally, despite the fact that shots taken by the home team are more likely to be successful, the feature `is_home` had little importance in the model.

5.3.2. Prediction Validity. We can use the predicted probabilities to obtain empirical distributions corresponding to the predicted probabilities of shots that actually resulted in goals (X_{goal}) and those that didn't ($X_{\text{no goal}}$). Figs. 30 and 31 provide comparisons of the two distributions. In particular, while both predict most shot attempts to have a low probability (below about 0.3, as shown in the histograms in Figs. 32 and 33), the skew in $X_{\text{no goal}}$ is much more extreme, peaking at the very left.

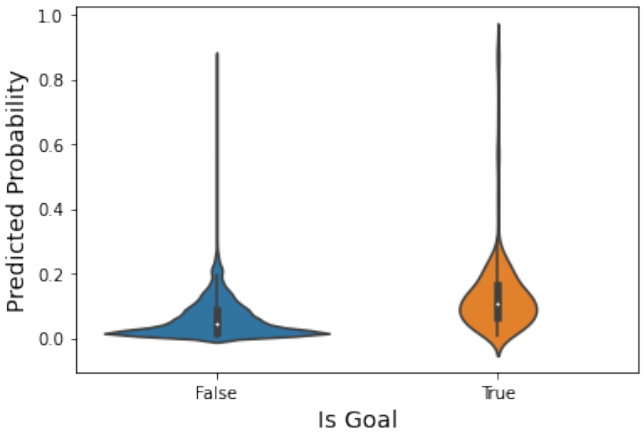

Fig. 30. Comparison of Distributions: $X_{\text{no goal}}$ vs. X_{goal}

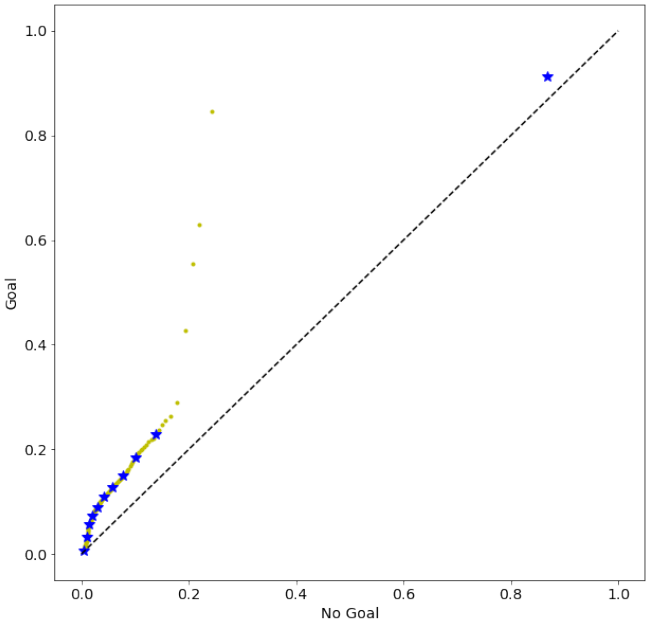

Fig. 31. P-P Plot: $X_{\text{no goal}}$ vs. X_{goal}

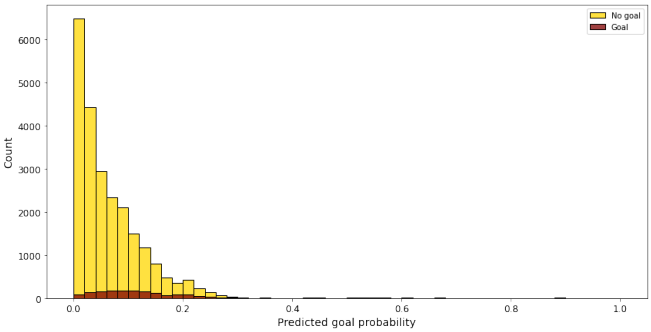

Fig. 32. Histograms of Predictions: $X_{\text{no goal}}$ vs. X_{goal}

6. Future Work

6.1. Home Goals. In Section 4.3, we found that shots taken by home teams are significantly more likely to result in goals than those attempted by visiting teams. There are two obvious possible explanations for this difference. The first of these are the explicit advantages to the home team set into NHL rules. The other possible explanation are the more intangible advantages of playing at home. These would include crowd

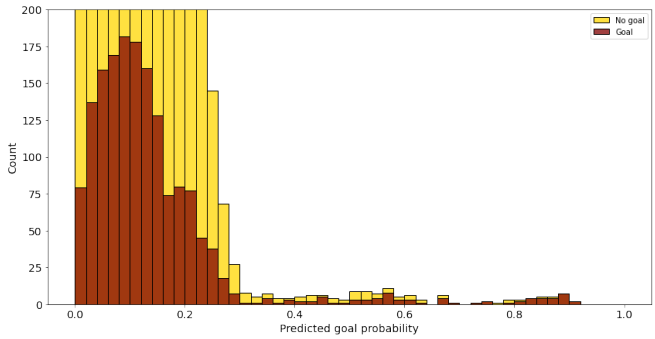

Fig. 33. Histograms of Predictions: $X_{\text{no goal}}$ vs. X_{goal}, truncated

support and familiarity with the venue. If the scoring difference is due to the rules, we would expect this advantage to be maintained in games played at neutral sites and games without a crowd. The COVID pandemic offers a unique opportunity to explore this question, as many games were played at neutral sites and/or without crowds. The playoffs at the end of the 2019-20 season were played in two “bubbles" in Edmonton and Toronto. In the 2020-21 season, games were mostly played at team’s normal venues, however many of these games were played with reduced crowd sizes or no crowds at all. This should allow distinguishing any crowd size effects from other intangible venue-based benefits.

6.2. Feature Engineering. The Fig. 29 gives ideas for possible changes to features for modeling. The first of these is to eliminate some of the game state data that proved to be extraneous, such as features dependent on the current score or the time in the game. Additionally, it may be appropriate to train separate models for game strengths. Since the feature `is_empty_net` has such an outsized importance compared to how often empty-net situations actually appear in game, it may be worthwhile to separate out circumstances when the attacking team is shooting on an empty net. Likewise, even strength, power play, and short-handed opportunities can be considered separately. Finally, it is possible to adjust the coordinate features. One obvious alteration is provided by examination of Fig. 21. In particular, there are three lanes in front of the net where goals are more likely to be scored. One is horizontal heading out to the left of the goal, while two diagonal lanes head out in the direction of the faceoff circles in the offensive zone. Reclassifying ordered pairs of coordinates into a categorical partition of position on the ice may provide improved prediction ability.