

# Lacrosse Expected Goals Model Worksheet

---

The purpose of this worksheet is to compare performance between logistic regression and support vector machines for an imbalanced data set, and how to approach class imbalance scenarios.

Logistic regression is a probabilistic classification model that estimates the likelihood of a binary outcome. For example, goal vs. no goal. It models the log-odds of the outcome as a linear combination of predictors. The output of the model is probabilities between 0 and 1 and a threshold (commonly 0.5) is used to convert the probabilities into class predictions (goal vs. no goal).

Support Vector Machines (SVM) are margin-based classifiers that focus on finding the best boundary between classes. SVM is identifying the hyperplane that separates classes with the maximum margin. Essentially, SVM is finding the best boundary that maximizes separation between classes.

Both models will be used in this worksheet to estimate the likelihood of a goal occurring based on the shot distance and angle as well as if the shot was assisted or unassisted. We will explore Premier Lacrosse League shot location data from the first three years of the league (2019–2021) for our analysis.

## Welcome Message and Introduction

[https://youtube.com/shorts/Tkl0jvzMv\\_0?feature=share](https://youtube.com/shorts/Tkl0jvzMv_0?feature=share)

### Lesson Materials:

Download the dataset and lesson code here:

[GitHub Repository](#)

## Lesson Objectives

- Explain the structure of an Expected Goals (xG) model and evaluate its usefulness in lacrosse analytics.
- Interpret logistic regression models and analyze the meaning of their coefficients in context.
- Compare and contrast Support Vector Machines (SVMs), including their key advantages and limitations.
- Construct and evaluate a baseline (dummy) model for performance comparison.
- Diagnose class imbalance in sports datasets and apply appropriate techniques to address it, justifying their impact on model performance using recall, precision, accuracy, and F1-score.

## A Look into the Data

Table 1: Summary Statistics of PLL Shot Dataset

Statistic	Value
Total Shots	7,745
Total Goals	2,172
Goal Rate	0.280
Assisted Shots	3,781
Assisted Shot Rate	0.488
<b>Shot Outcome Distribution</b>	
Off Target	2,394
Goal	2,172
Messy Save	1,649
Clean Save	786
Blocked	466
Pipe	276

### Key Terms:

1. Off Target: The shot misses the cage.
2. Goal: A ball that crosses the goal line extended in the net is considered a goal.
3. Messy Save: The goalie stops the ball and does not retain control of the ball.
4. Clean Save: The goalie stops the ball and retains control of the ball.
5. Blocked: The ball hits a player before it can reach the net.
6. Pipe: The ball hits one of the goal pipes, ricochets, and does not score a goal

### What percentage of shots resulted in a goal? In a no goal?

28.04% are goals, 71.96% are no goals.

### Why would the proportion of goals vs no goals concern you?

This proportion means that the dataset is imbalanced. Our one class (no goals) appears much more frequently than the other. Therefore, our model can achieve high accuracy without learning anything. Another concern is that the model may be biased towards the majority class.

If you were to build an intercept-only model, what would be the baseline probability of a goal?

The baseline probability of scoring a goal would be 28.04%

If the model always predicted “no goal,” what accuracy would it achieve?

It would achieve an accuracy of 71.96%.

Which metric do you think will be the most informative? Why?

- a. Accuracy
- b. Precision
- c. Recall

The most important metric is recall because this measures out of all the actual goals, how many the model correctly identified. We want to capture as many actual goals as possible.

## Shot Location Should Matter

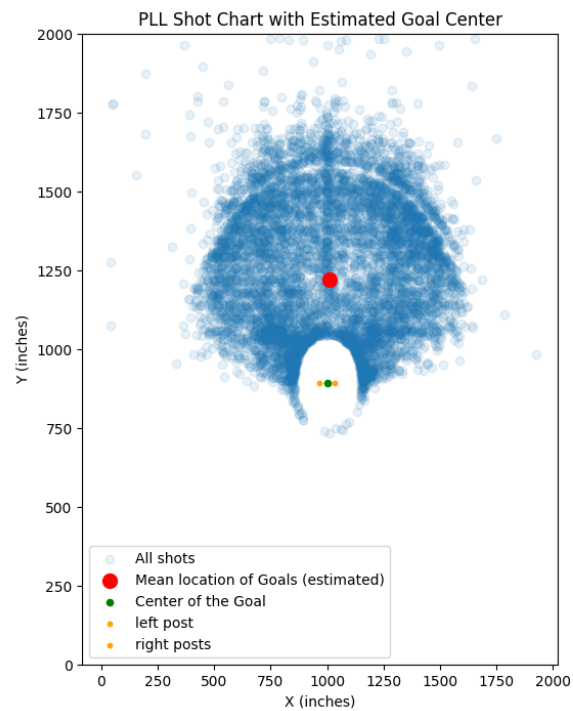


Figure 1: PLL Shot Chart

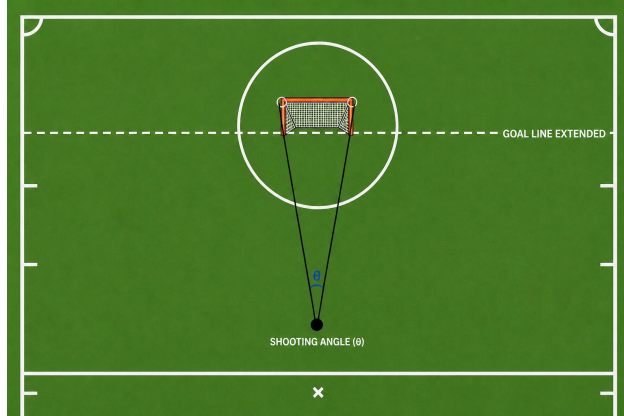


Figure 2: Field Diagram

If the center of the goal is (1000, 892), the left post is (964, 892), and the right post is (1036, 892), how could you re-center the data on (0,0) for the center of the goal (the units are in inches)?

You take the x coordinates and subtract by 1000. Likewise, you take the y coordinates and subtract by 892.

Calculate the distance of a shot (1010, 1219) to the center of the goal.

$$1010 - 1000 = 10$$

$$1219 - 892 = 327$$

$$\text{Distance equation: } \sqrt{x^2 + y^2}$$

$$\sqrt{10^2 + 327^2} = 327.1532 \text{ inches}$$

Which is about 9.08 yards

Calculate the angle of a shot (1010, 1219)? The angle being from the location of the shot to the two posts of the goal as displayed in Figure 2. The posts locations are (36,0) and (-36,0).

$$x_L, y_L = -36, 0$$

$$x_R, y_R = 36, 0$$

$$\text{location of shot: } (10, 327)$$

Next, we make vectors from the shot to posts:

$$SL_x = x_L - dx \text{ (distance x)}$$

$$SL_y = y_L - dy \text{ (distance y)}$$

$$SR_x = x_R - dx$$

$$SR_y = y_R - dy$$

Calculate the magnitudes to each post and the dot product to do the arccos for angle:

$$\text{mag}_{SL} = \sqrt{SL_x^2 + SL_y^2}$$

$$\text{mag}_{SR} = \sqrt{SR_x^2 + SR_y^2}$$

$$\text{Dot product} = SL_x \cdot SR_x + SL_y \cdot SR_y$$

$$\text{Angle} = \cos^{-1} \left( \frac{\text{Dot product}}{\text{mag}_{SL} \text{mag}_{SR}} \right) = 12.6^\circ$$

**Will a shot from (1010, 1219) result in a goal?**

Yes, this shot is more likely than not to result in a goal. This is a shot from 9.08 yards with an angle of 12.6 degrees. The shot is under 10 yards, which is ideal, and is at a high angle to the goal.

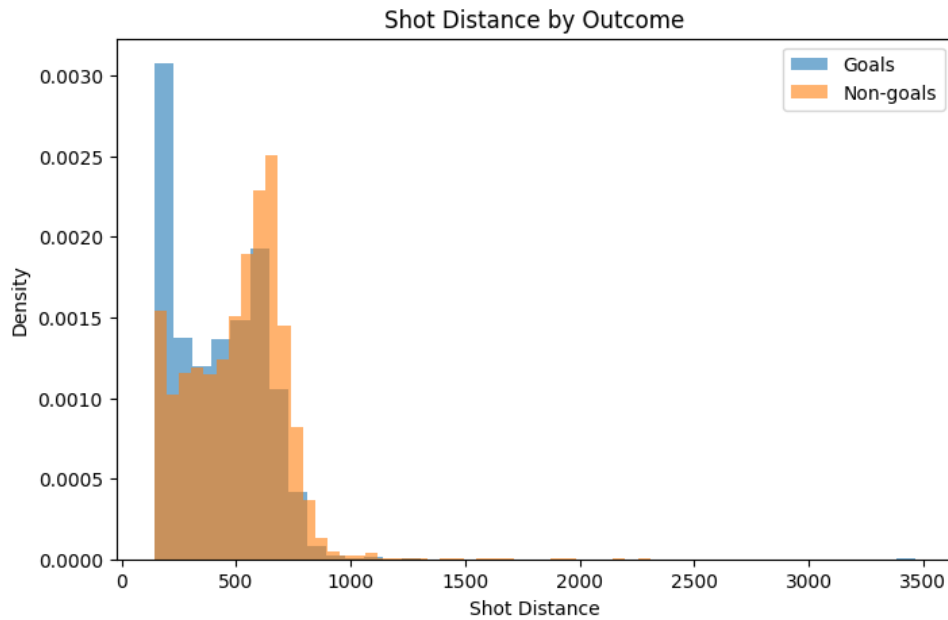


Figure 3: Shot Distance by Outcome

**In Figure 3, how does distance affect goal probability?**

There is a high density of goals from the minimum distance, but then goals dip to a lower density at 500 inches. The goal density then increases again around 700 inches before it drops again.

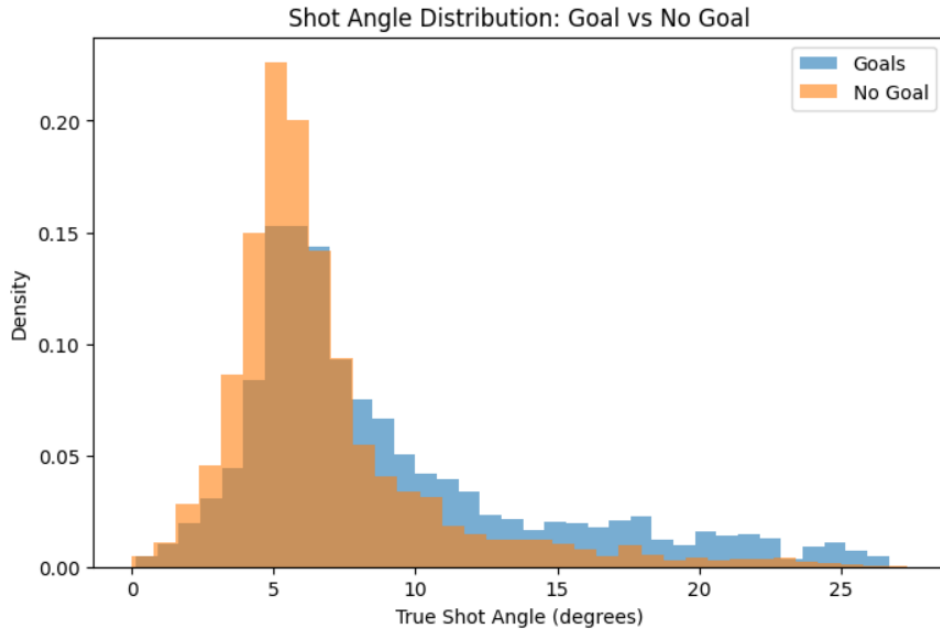


Figure 4: Shot Angle Distribution

In Figure 4, how does shot angle affect goal probability?

The density of goals increases from 0 degrees to 6 degrees and then decreases past 6 degrees.

For a shot at (1000, 1250), would you expect a high or low predicted goal probability?

- a. High
- b. Low



Figure 5: Shot Outcomes

**In Figure 5, why is the shot angle restricted between 0–30? Does this make sense?**

In lacrosse, the maximum angle at which you can shoot is at the top center of the crease; this is where you will have the greatest angle to the posts. A shot from zero degrees is from Goal Line Extended (GLE), the line that runs from post to post. This shot is often referred to as a no-angle shot.” That is why we see the shot angle between 0 and 30 degrees. Refer to Figure 2, to see GLE and the crease.

## Modeling

### Dummy Classifier

Before modeling, let’s recall class balance and Table 1.

**What proportion of shots result in a goal?**

28.04% are goals and 71.96% are not goals.

**Why is this important? What if I create a dummy model that always predicts no goals?**

The data is imbalanced, so a model can achieve high accuracy just by predicting “no goal” most of the time.

In sports analytics and specifically lacrosse, events like goals are rare compared to non-goals. It is common to find that class imbalance exists. Class imbalance is an issue because a model could potentially achieve high accuracy just by predicting “no goal” the whole time. Our model will have high accuracy without learning anything. It can be helpful to create a baseline model before trying logistic regression or SVM. I am going to make a dummy classifier baseline model that only predicts no goals. Even if the dummy model has high accuracy, it is not useful because it never identifies goals. This highlights why accuracy alone can be misleading in imbalanced datasets.

**What is the accuracy of the dummy classifier baseline model?**

71.96%

**Why is recall always going to zero in the dummy classifier baseline model?**

$$\text{Recall} = \frac{TP}{TP + FN}$$

The model is only predicting non-goals so it is going to miss every single goal. The model can't even be right about goals. Therefore, the true positive is going to equal zero. That is why recall is zero.

**Why is precision always going to zero in the dummy classifier baseline model?**

$$\text{Precision} = \frac{TP}{TP + FP}$$

The model is never even trying to predict a goal, so it can't have any correct or false positive predictions. The true positive and false positive both equal zero. Therefore, precision is going to equal zero.

## Logistic Regression and SVM

To address the class imbalance in my modeling, I included "class weights = balanced" in my model, which told it to pay more attention to the minority class (goals) and less to the majority class (no goals) when learning. This is considered as "balancing" the dataset because the model is forced to care about detecting goals rather than cheating by predicting mostly no goals.

Tables 2 and 3 compare the performance of several classification models used to predict whether a lacrosse shot results in a goal. Two different approaches were explored: Standard Logistic Regression and SVM models, and the same models adjusted for class imbalance. The "Dummy Classifier" was included in both tables as a baseline classifier that always predicts "no goal."

The following evaluation metrics provide different perspectives on model performance:

1. Accuracy: measures the overall proportion of correct predictions.
2. Precision: measures how often predicted goals were actually goals.
3. Recall: measures how well the model identifies actual goals.
4. F1-score: balances precision and recall into a single metric.

5. AUC (Area Under the ROC Curve): measures the model's ability to distinguish goals from non-goals across classification thresholds

When working with imbalanced sports data, recall, F1-score, and AUC are often more informative than accuracy alone because a model can achieve high accuracy simply by predicting the majority class.

Table 2: Performance Metrics for Balanced Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Balanced SVM	0.6850	0.4403	0.4585	0.4492	0.6641
Balanced Logistic Regression	0.6675	0.4249	0.5276	0.4707	0.6640
Dummy Classifier	0.7198	0.0000	0.0000	0.0000	0.5000

Table 3: Performance Metrics for Non-Balanced Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.7347	0.7082	0.7347	0.6672	0.6638
SVM	0.7198	0.5181	0.7198	0.6026	0.5898
Dummy Classifier	0.7198	0.0000	0.0000	0.0000	0.5000

Table 4: Logistic Regression Coefficients

Feature	Coefficient
Angle	0.314512
Assisted	0.186663
Distance	-0.229881

What are the logistic regression coefficients written in terms of?

- a. Log-odds
- b. Probabilities

How do we convert the log-odds into odds, and how do we interpret the odds?

Logistic regression coefficients are in log-odds, not probabilities. To interpret them, we exponentiate to get odds.

$$\text{Odds} = e^{\beta}$$

Example: each 1 degree increase in angle multiplies odds by  $e^{0.314512} \approx 1.37$ .

Interpret the logistic regression positive coefficient for angle.

Increasing angle increases the probability of scoring. Each 1 degree increase in angle increases the odds by  $e^{0.314512} \approx 1.37$ . (37%)

Interpret the logistic regression negative coefficient for distance.

Increasing distance decreases the probability of scoring. Each 1 inch increase in distance decreases the odds by  $e^{-0.229881} \approx 0.794$ . (20.6%)

**What is one key conceptual difference between logistic regression and SVM?**

SVM is finding the best separating boundary. It is the max-margin boundary, which is often a more flexible decision boundary. Logistic regression is finding the log-odds which then can be converted into odds.

Recall is the most informative metric due to class imbalance and the importance of correctly identifying goals. While accuracy is inflated by the majority class, recall directly measures the model's ability to capture the minority class of interest. F1-score is also useful as a secondary metric, but recall provides the clearest insight into model effectiveness in this context.

**Which model would you choose if your goal is to identify as many actual goals as possible? Explain your reasoning using the metrics.**

Balanced logistic regression because it has the highest recall. The high recalls for the unbalanced models are misleading because they are heavily influenced by the majority class. Balanced models were trained to care more about the minority class (goals) so they provide a more honest attempt at detecting goals.

In this worksheet, you built and evaluated an expected goals (xG) model using logistic regression and SVM, interpreted model coefficients in context, and compared model performance using appropriate metrics. You also learned the importance of baseline models and how class imbalance affects evaluation, reinforcing that effective modeling requires both technical understanding and thoughtful decision-making. Now that you have this knowledge, what will you explore next?