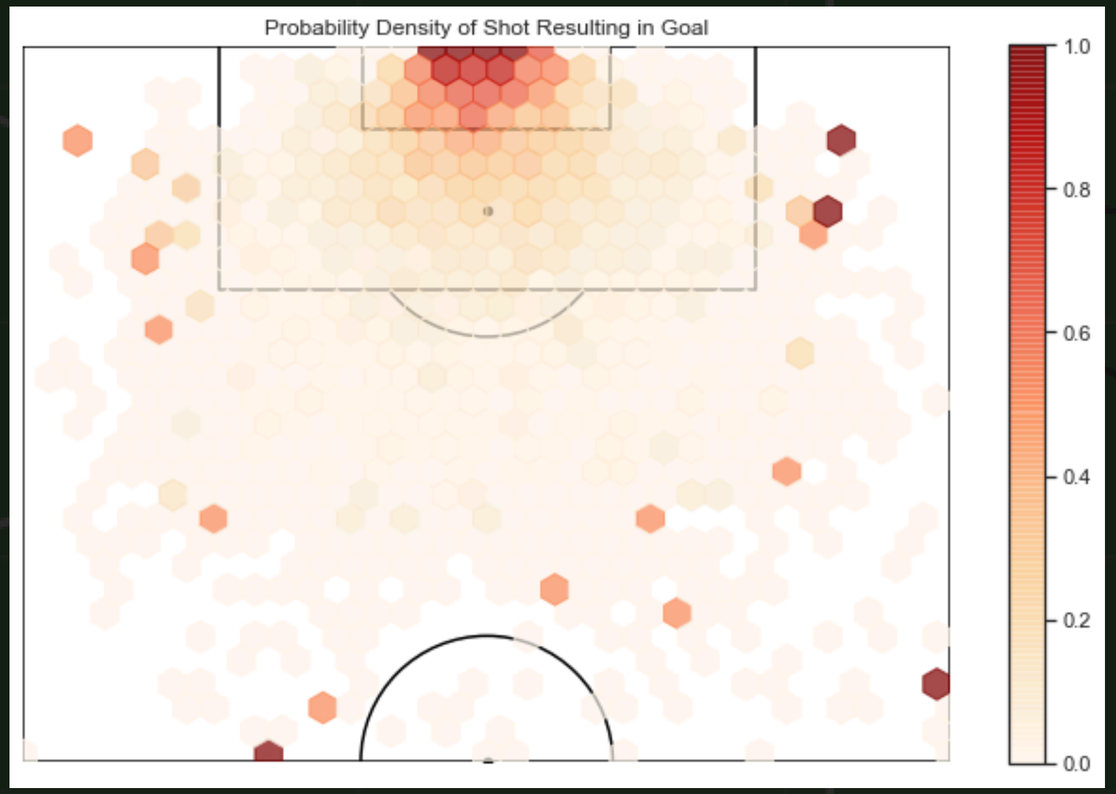
CS-IS-3066-1: Applied Machine Learning in Football

Karnav Popat, Pranav Koka, and Suyog Joshi

What is expected Goals?

xG is a metric first proposed in 2012 which measures the likelihood of a shot translating into a goal.

Today, it is the most widespread metric used in sports analytics to measure player and team impact.



Our goals and contributions

Replicate the industry-leading xG model and create an accessible, reproducible baseline for future contributions	→	A lit review and detailed pipeline of eight Jupyter notebooks which use StatsBomb open data to achieve par-results
Add interpretability to the xG model to facilitate a better understanding of what really matters	→	Replicable visualizations of tree models, feature importance scores, and partial dependence plots
Test our hypotheses based on sports intuition, primarily the influence of temporal patterns, using real data	→	Notebooks and blog-format reports on our hypotheses, lending some support to our initial thought

Our process

Grabbing the data from StatsBomb's API
and compiling it into 11 csv files

Cleaning, formatting and augmenting it
adding 18 features that we calculate ourselves

Exploring and visualizing the data
on the pitch and through conventional charts

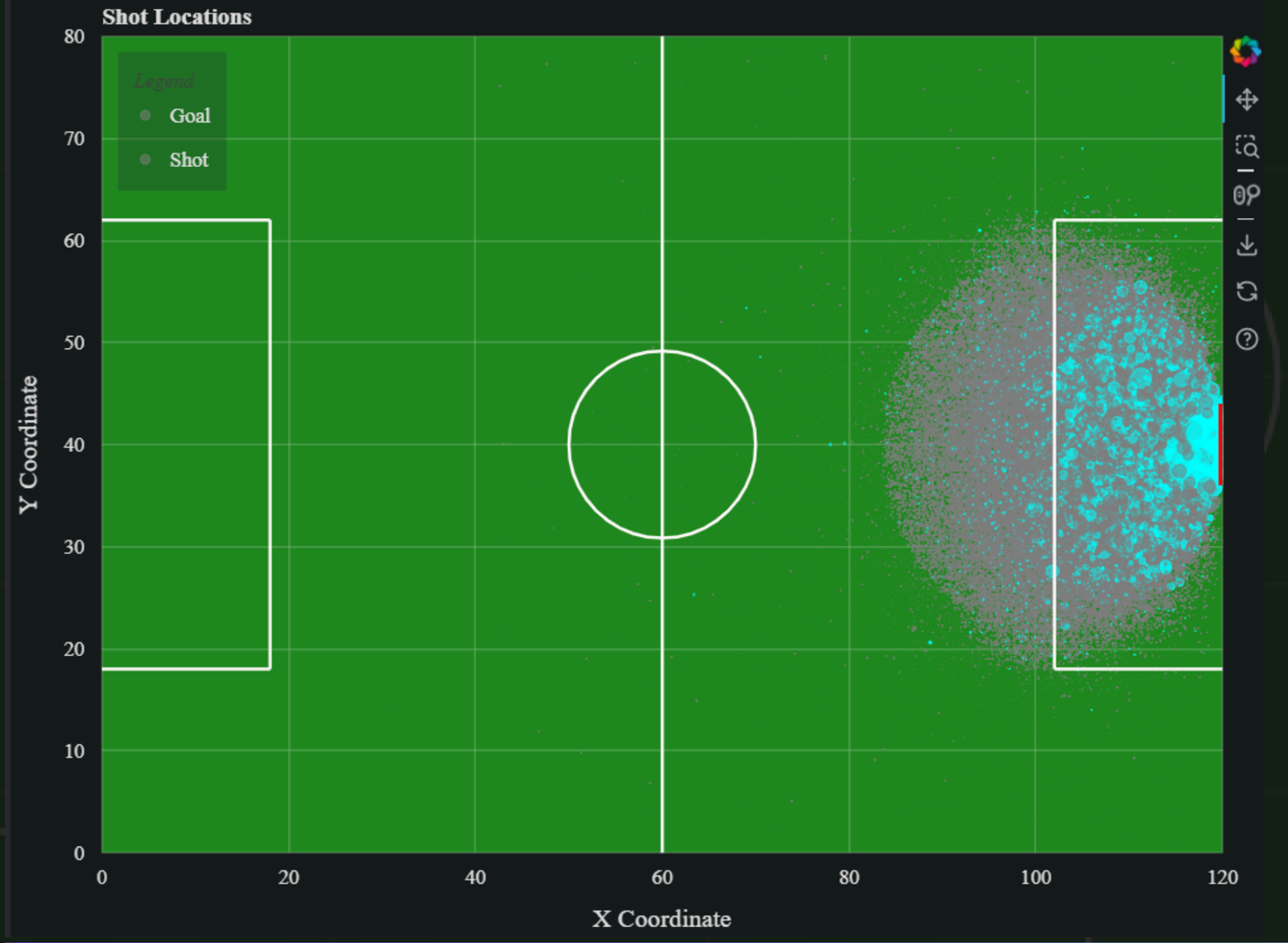
Getting baseline results
with regression, decision trees and ensembles

Improving the models
by oversampling, finetuning, selecting features

Thinking of and finding hypotheses
based on intuition, lit review and visualizations

Explaining results
using feature importance scores, PDPs and a GAM

StatsBomb's Open Data



The dataset:

11.6 million observations

40 seasons

21 competitions

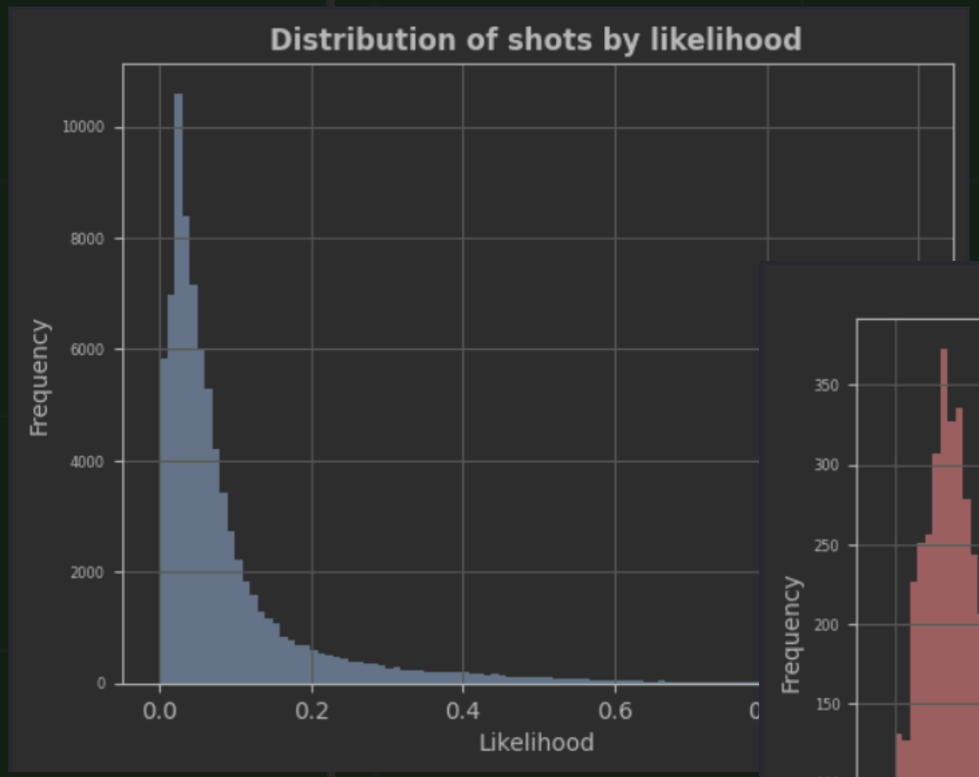
3,500+ matches

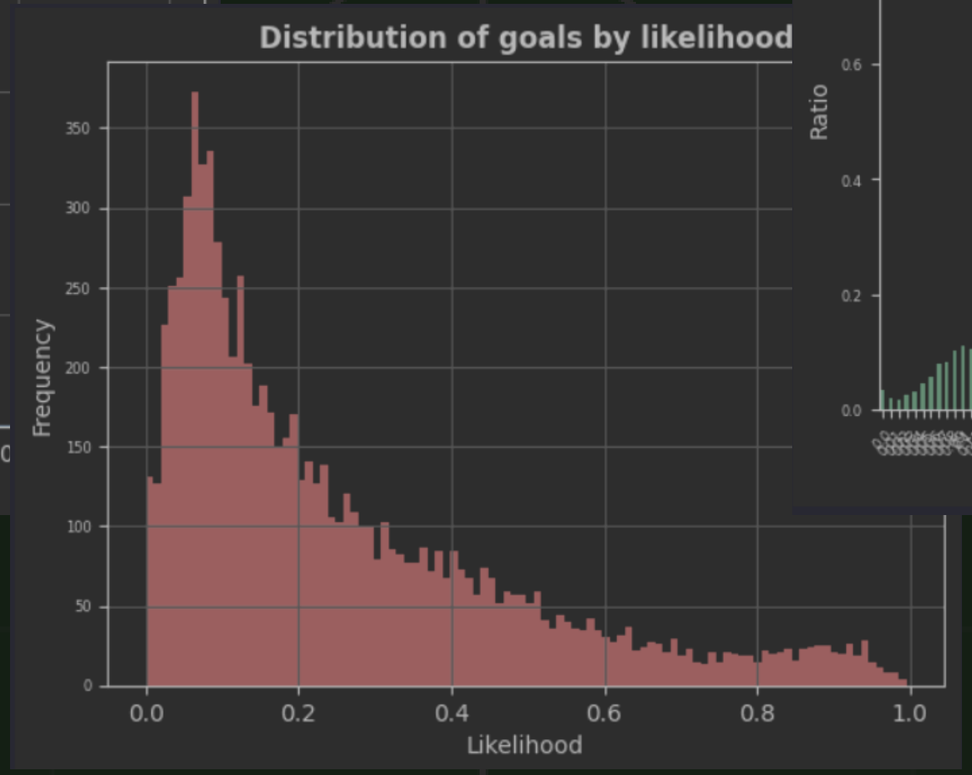
representing

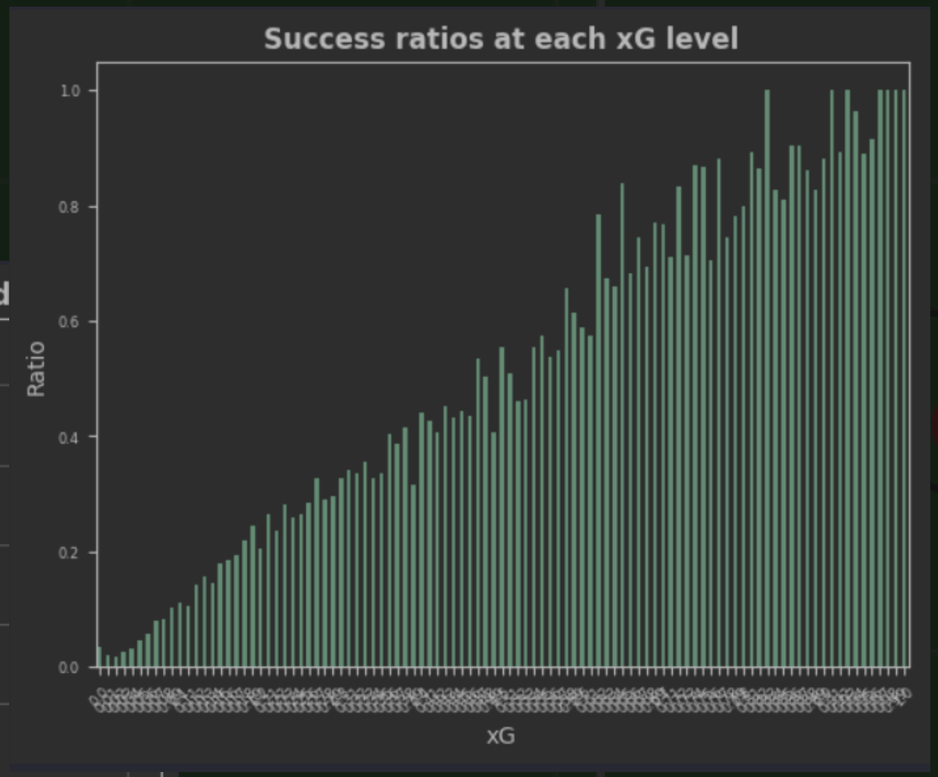
82,000+ shots

8,400+ goals

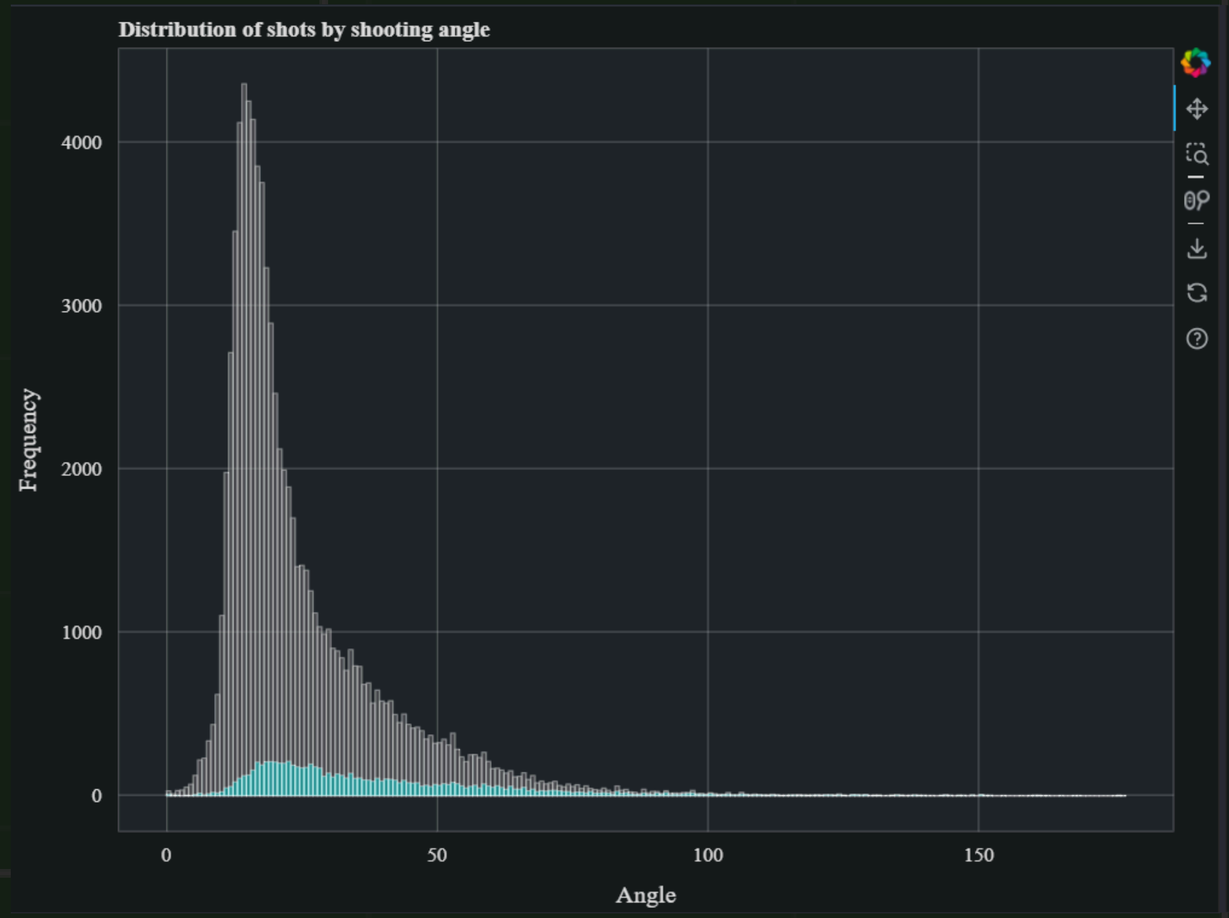
A closer look at the data

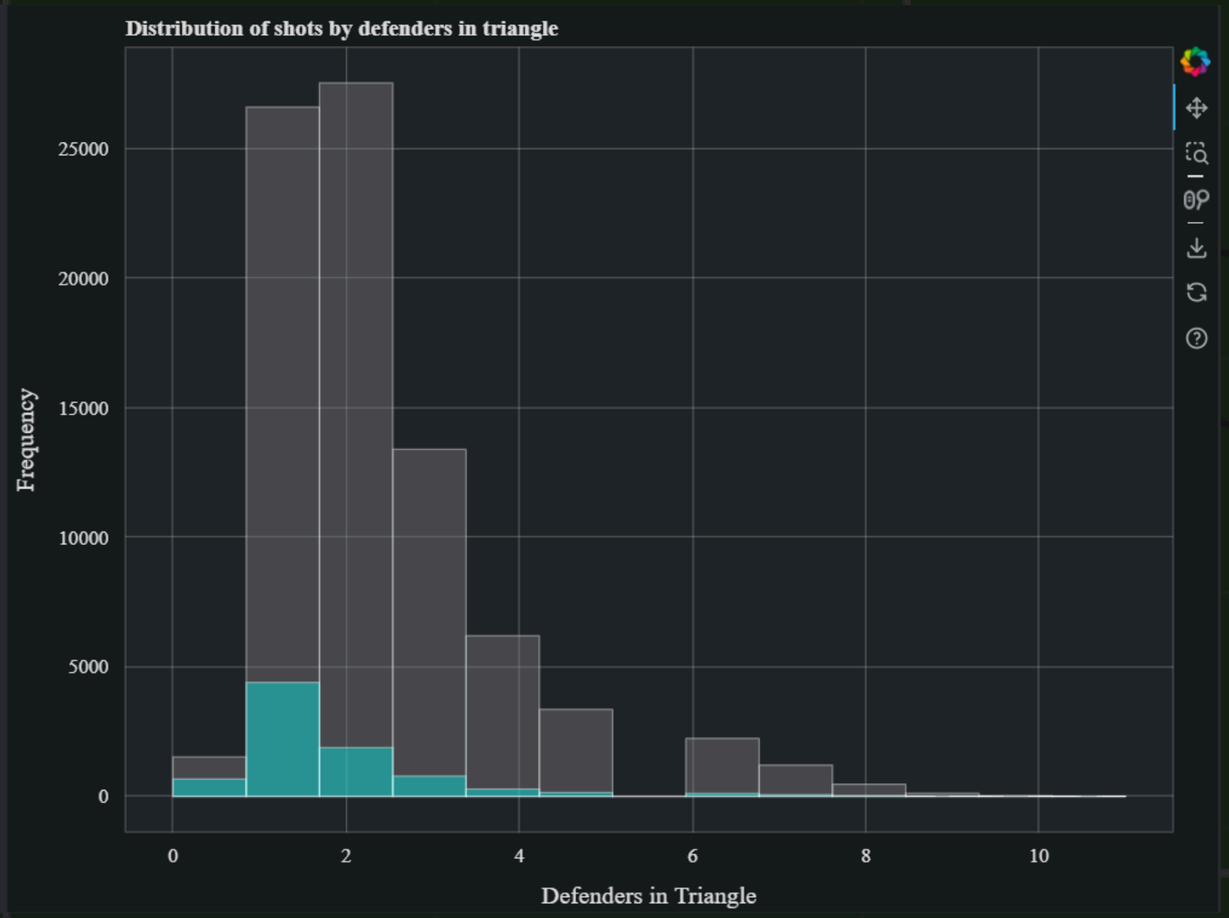




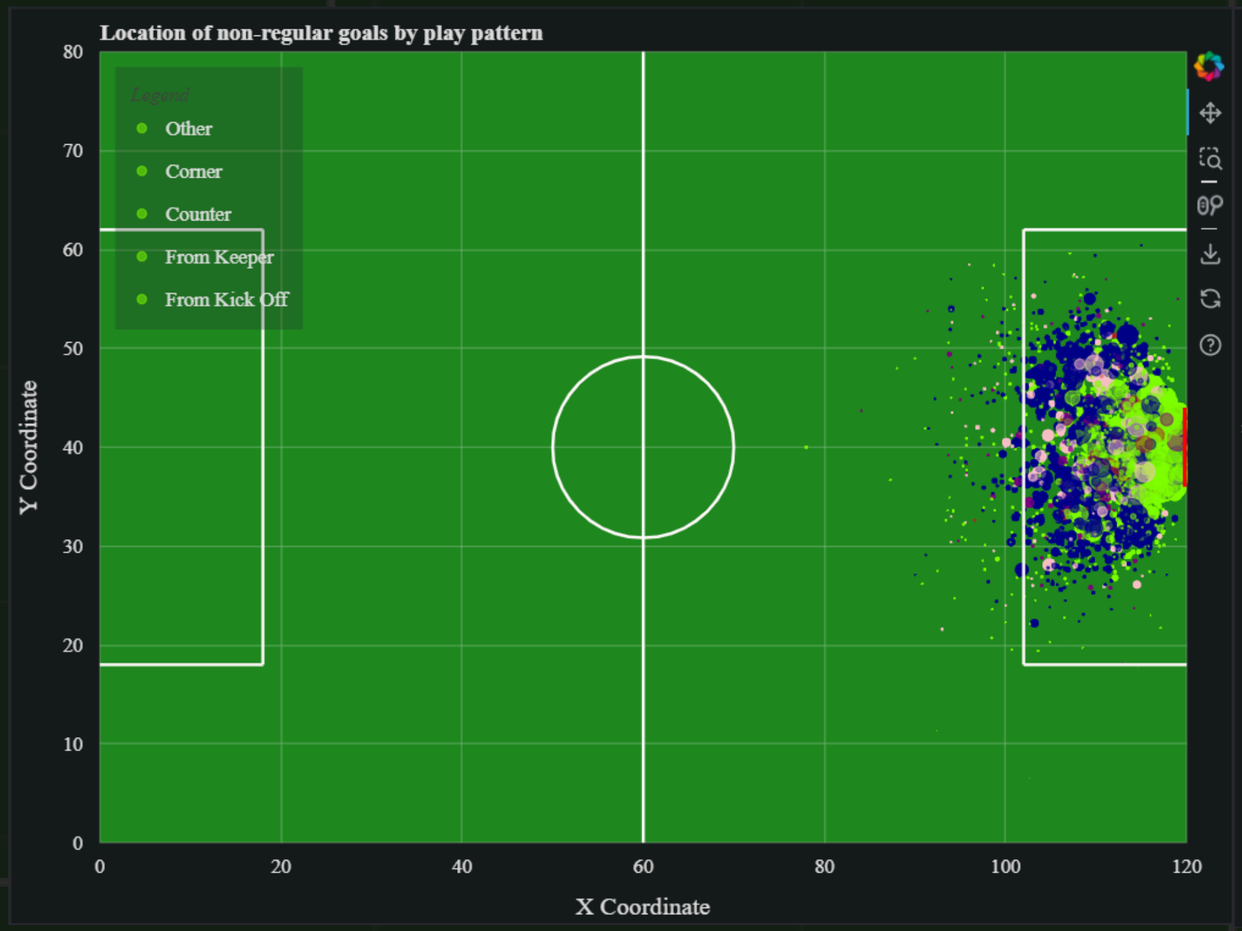


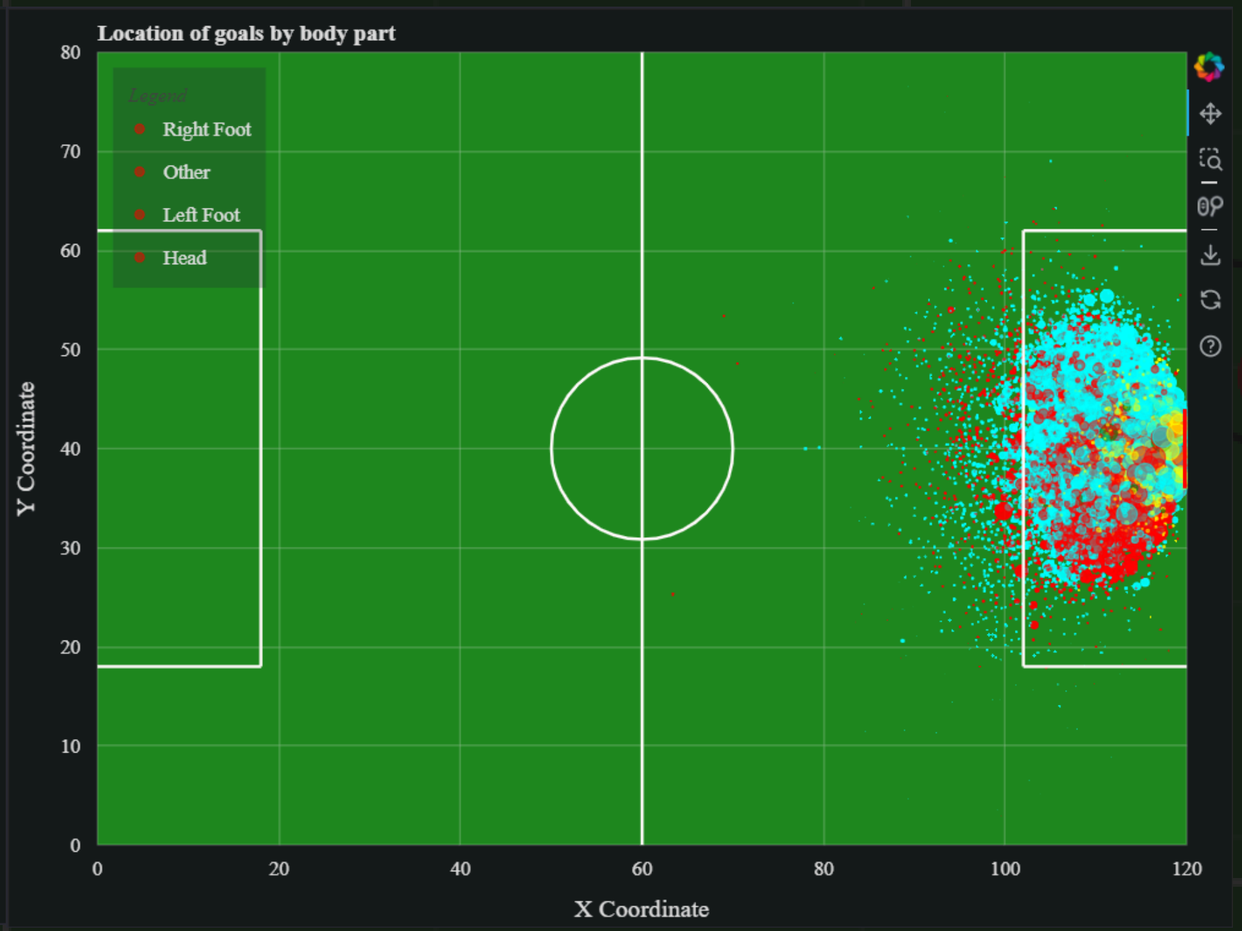
A closer look at the data



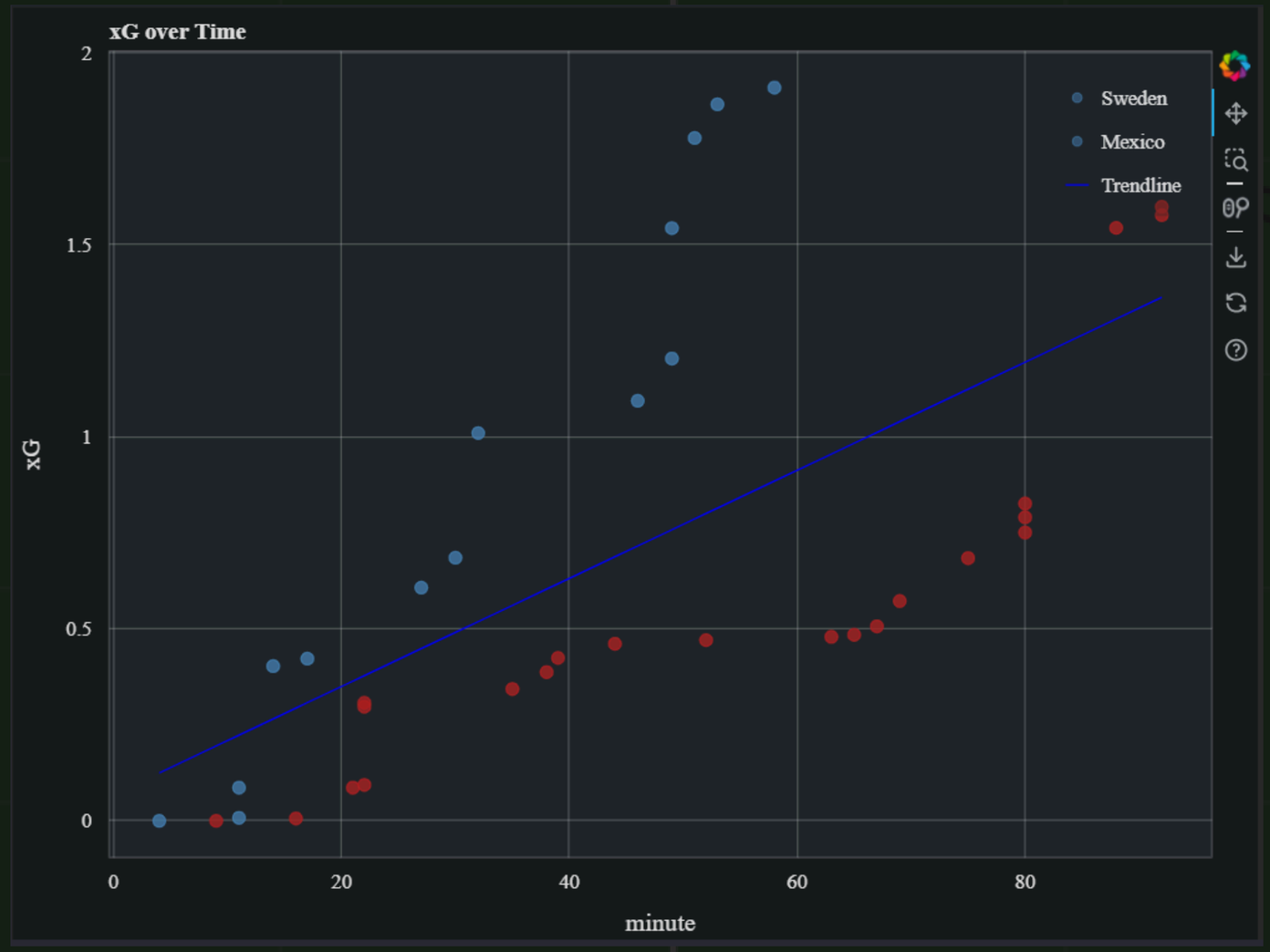


A closer look at the data





A closer look at the data



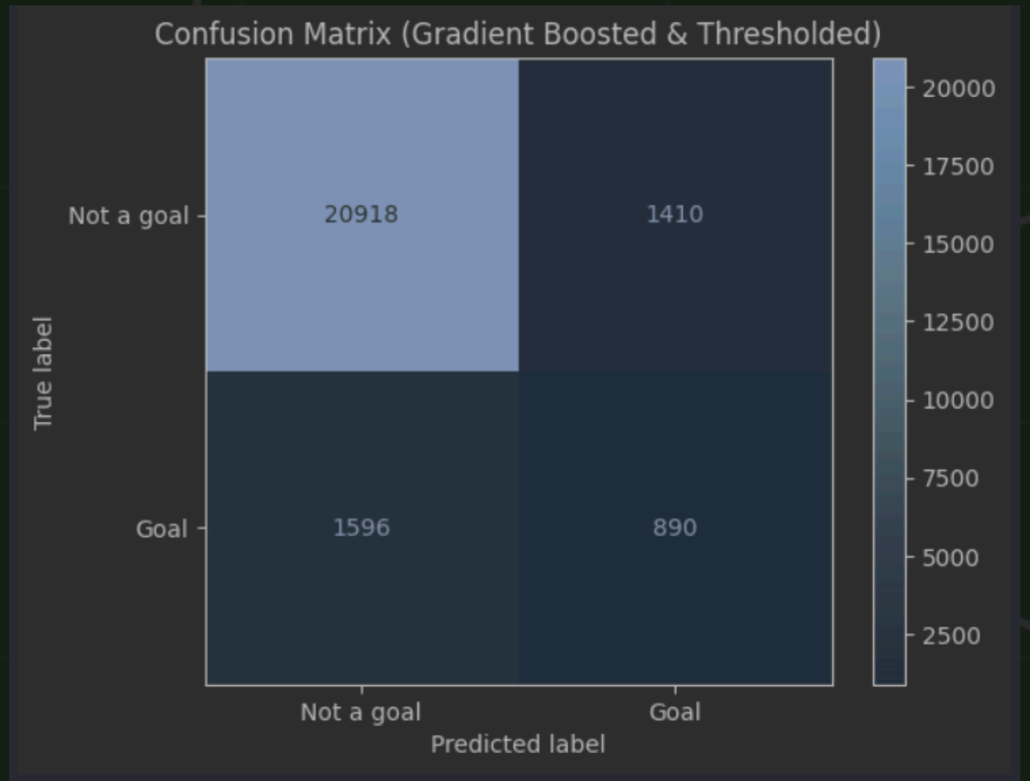
Our selected models

We implemented 5 models:

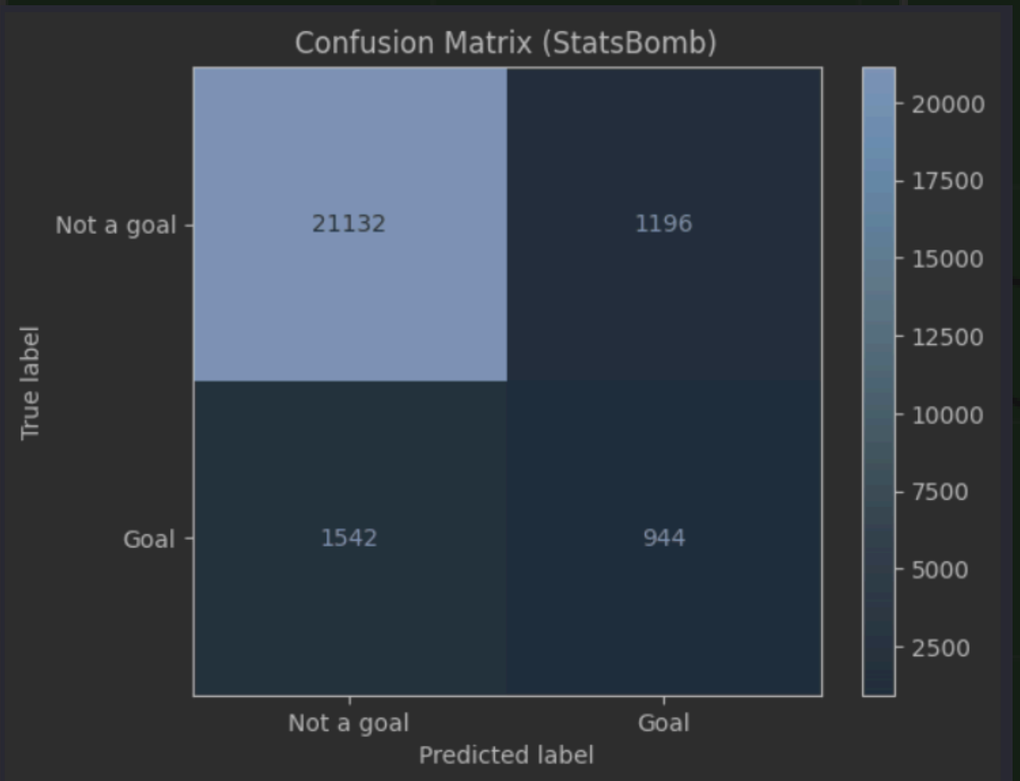
1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. Random Forest
5. Gradient-boosted Tree

Based on our literature review, we expected Gradient-boosted Tree to perform the best, and Decision Tree to be the most interpretable. We wanted to test the performance of all the models as a comparison.

Baseline results

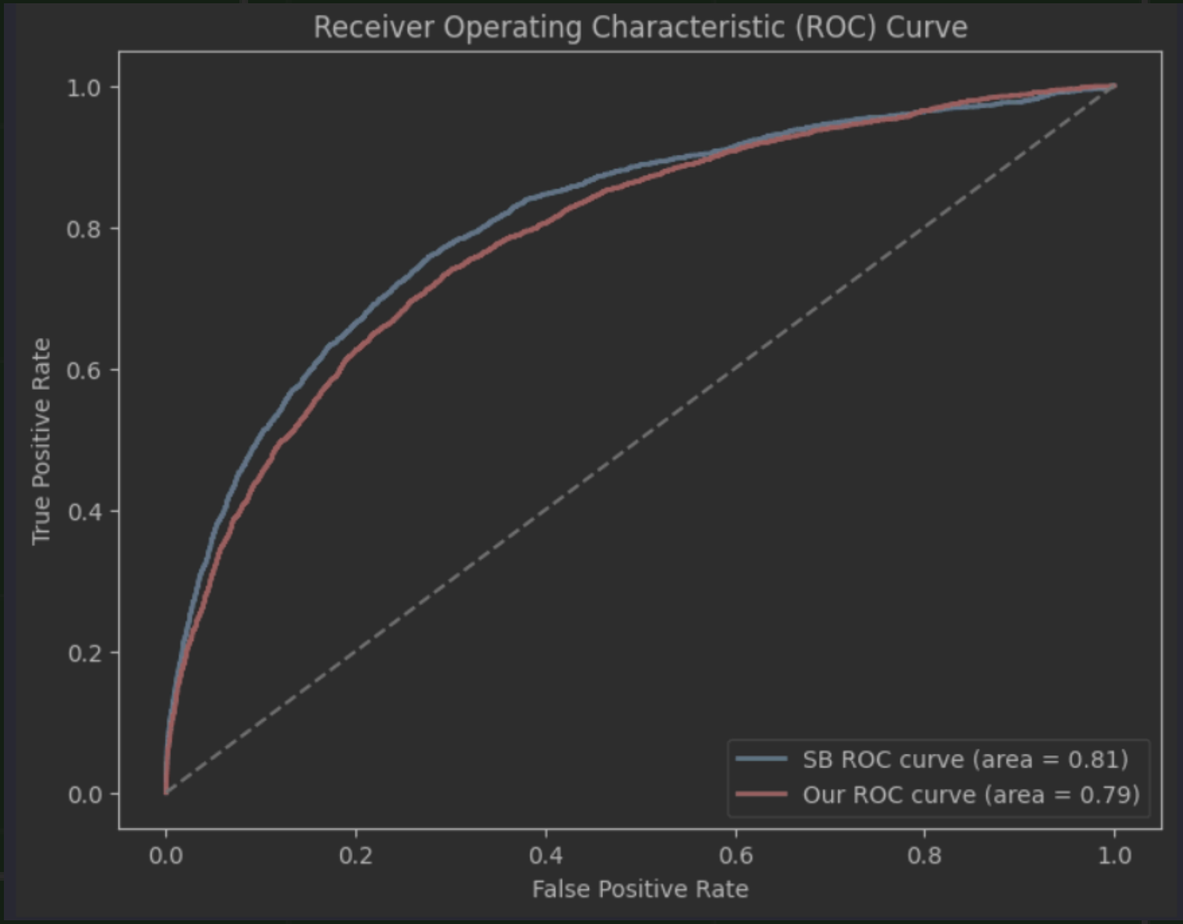


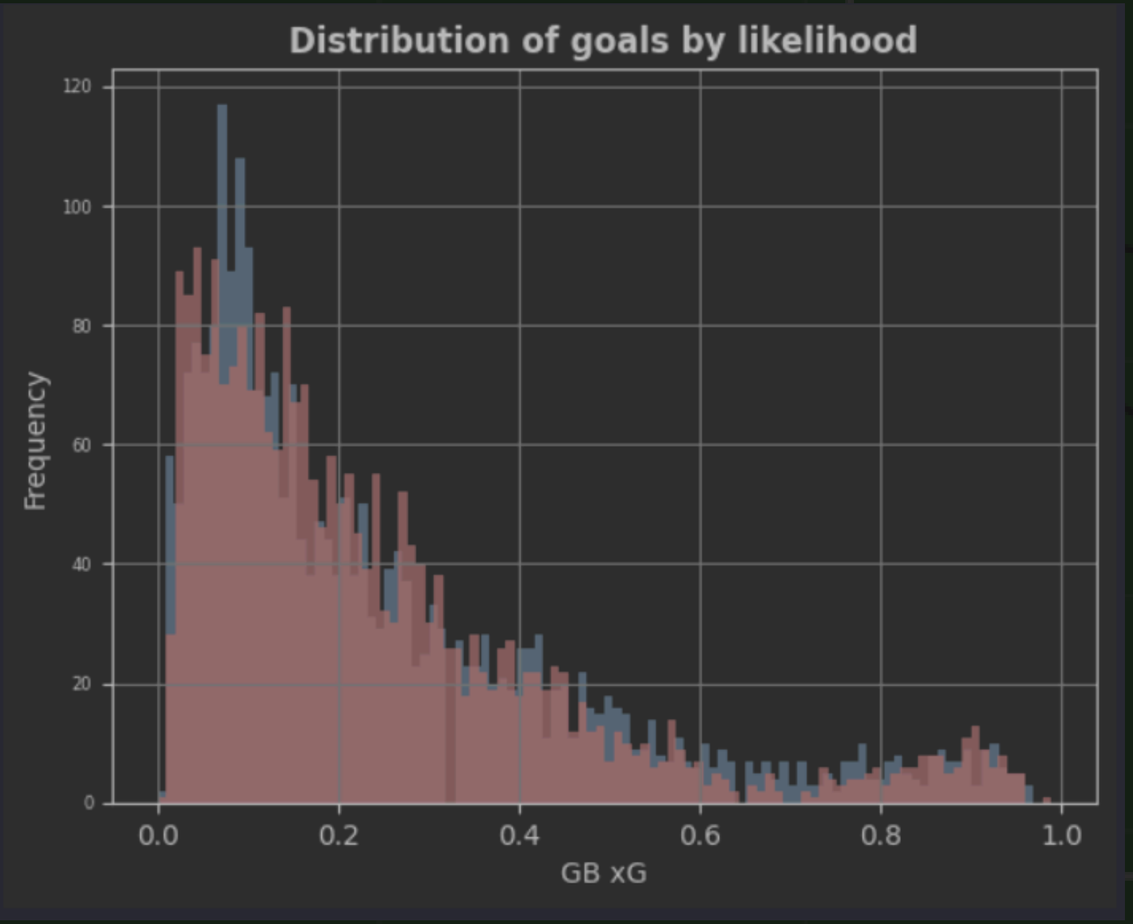
Gradient Boosted & Thresholded
Accuracy: 0.8788587087934231
Precision: 0.3869565217391304
Recall: 0.3580048270313757
F1: 0.37191809444212287



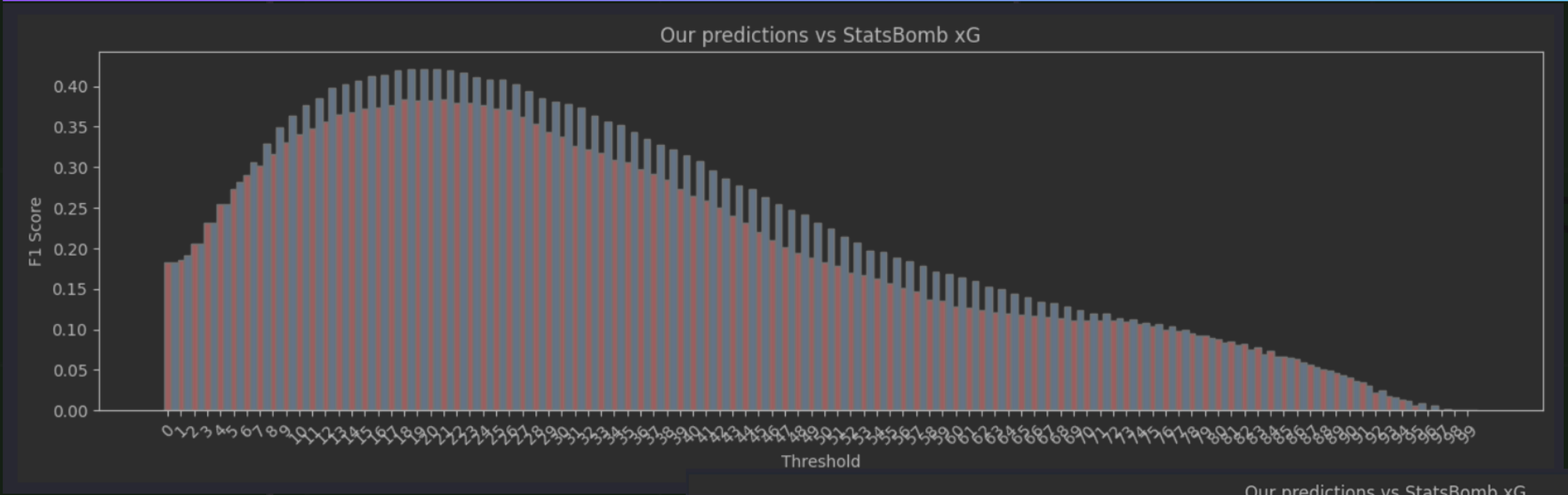
StatsBomb
Accuracy: 0.8896590634319336
Precision: 0.4411214953271028
Recall: 0.3797264682220434
F1: 0.40812797233030695

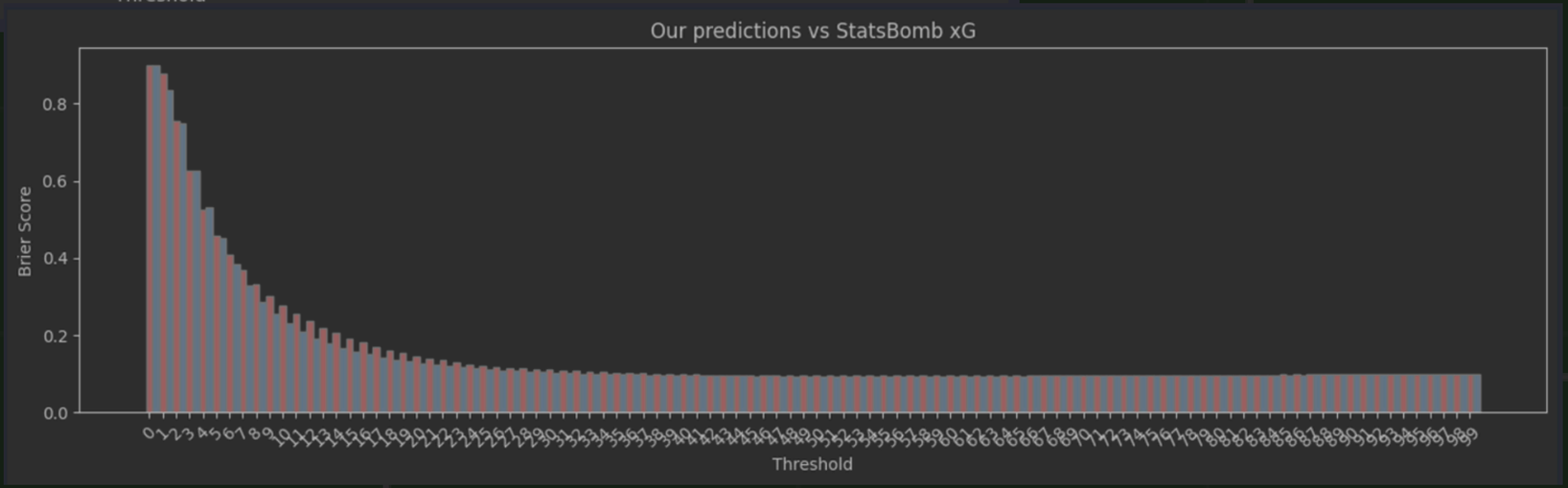
Baseline results





Baseline results





Our feature engineering

Goalkeeper features:

- goalkeeper_x_distance
- goalkeeper_y_distance
- distance_to_goalie

Defensive “pressure” features:

- defenders_in_triangle
- defenders_3m_radius
- influence_in_triangle
- press

Location features:

- shooting_range
- shot_angle
- goal_distance
- best_distance

Hypothesis features:

- good_foot
- shots_so_far & xg_so_far
- game_state

- was_leading
- past_minute & past_15
- is_extra_time

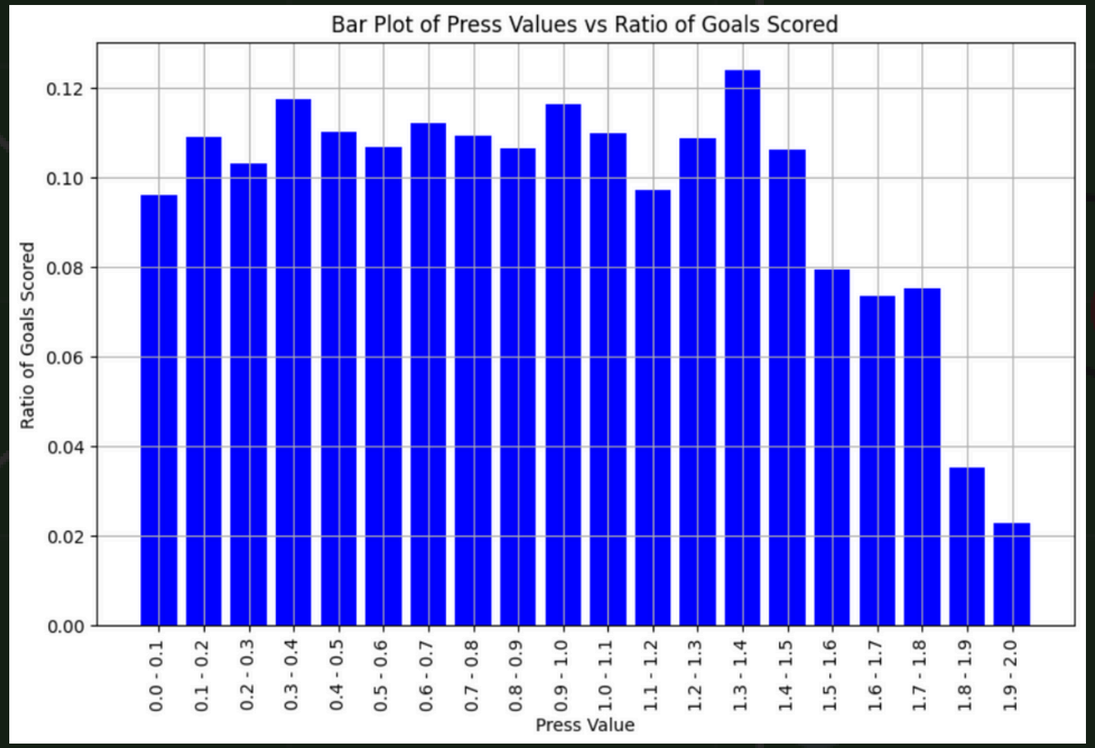
Defenders


Location of players at time of shot
Legend
cyan
lightcoral
paleturquoise
red
Y Coordinate
X Coordinate
0
10
20
30
40
50
60
70
80
0
20
40
60
80
100
120

Press

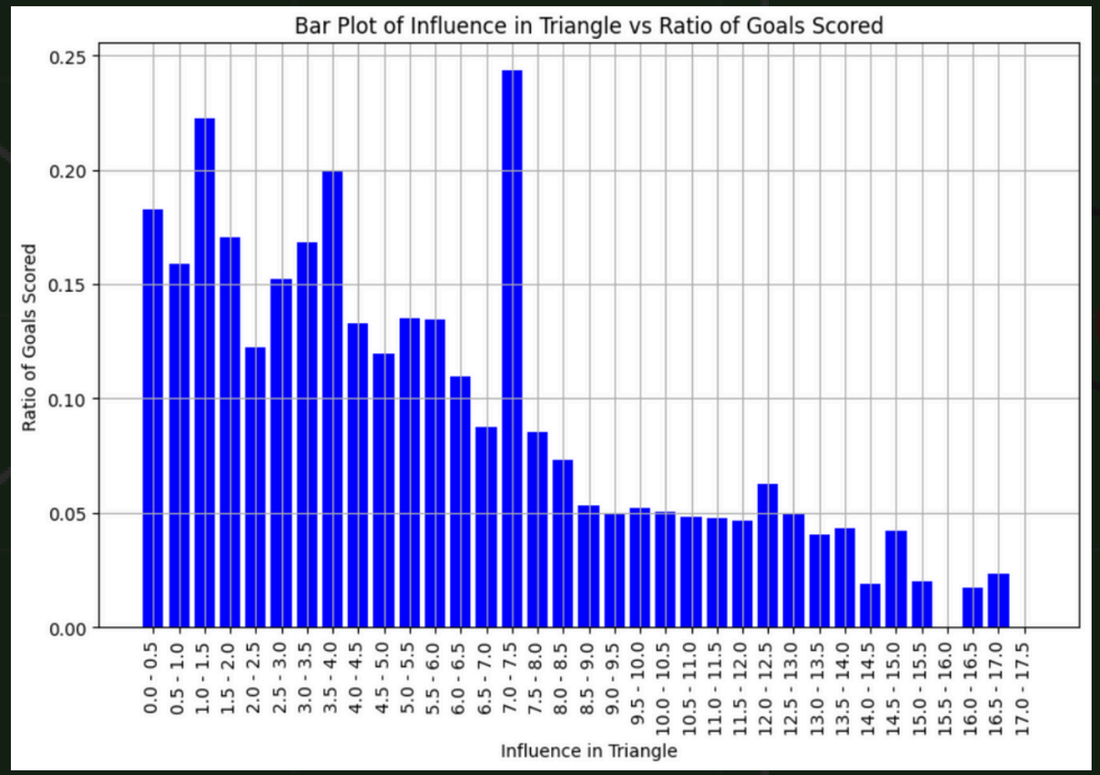
- under_pressure was a feature which had boolean values of whether or not the shot being taken was under pressure.
- The percentage of shots that resulted in goals when the shot was under pressure vs when it wasn't under pressure is 9.2% vs 11.6%.

- We added a feature called press which calculates the amount of pressure there is around a player taking a shot by representing each defender as a 2D gaussian surface.
- As the value decreases from 1.4 to 2.0 there is a decrease in the ratio of shots leading to goals.



Influence in triangle

- As mentioned earlier, we represented each player as a 2-D gaussian surface which meant that each player now had a smoother influence on the pitch.
- We summed up these values for all the players over the 120*80 points on the pitch which gave each point on the pitch a certain influence value.
- We then summed the influence values of all the points inside the triangle made by the shot and the two goalposts.
- As we can see from the graph, the ratio of goals tends to decrease as influence in triangle increases.



Feature selection

We used two methods available with sklearn: SelectFromModel and PermutationFeatureImportance

SELECT FROM MODEL

SFM uses a trained model to select the most important features.
The model's weights and coefficients are used to calculate importance scores.
Features with the highest importance scores are chosen.

LEAGUE TESTING

We ran SFM on LaLiga and Bundesliga data separately to determine if the top features had any differences.
It showed us that there were some minor differences in the rankings of the features

Feature importance methods

DISADVANTAGES OF SFM

- SFM uses a linear model to determine coefficients - it doesn’t use the models we will actually be using.
- The scores might be different depending on the model used for SFM.
- They don’t tell us exactly how the features affect our decision tree model.

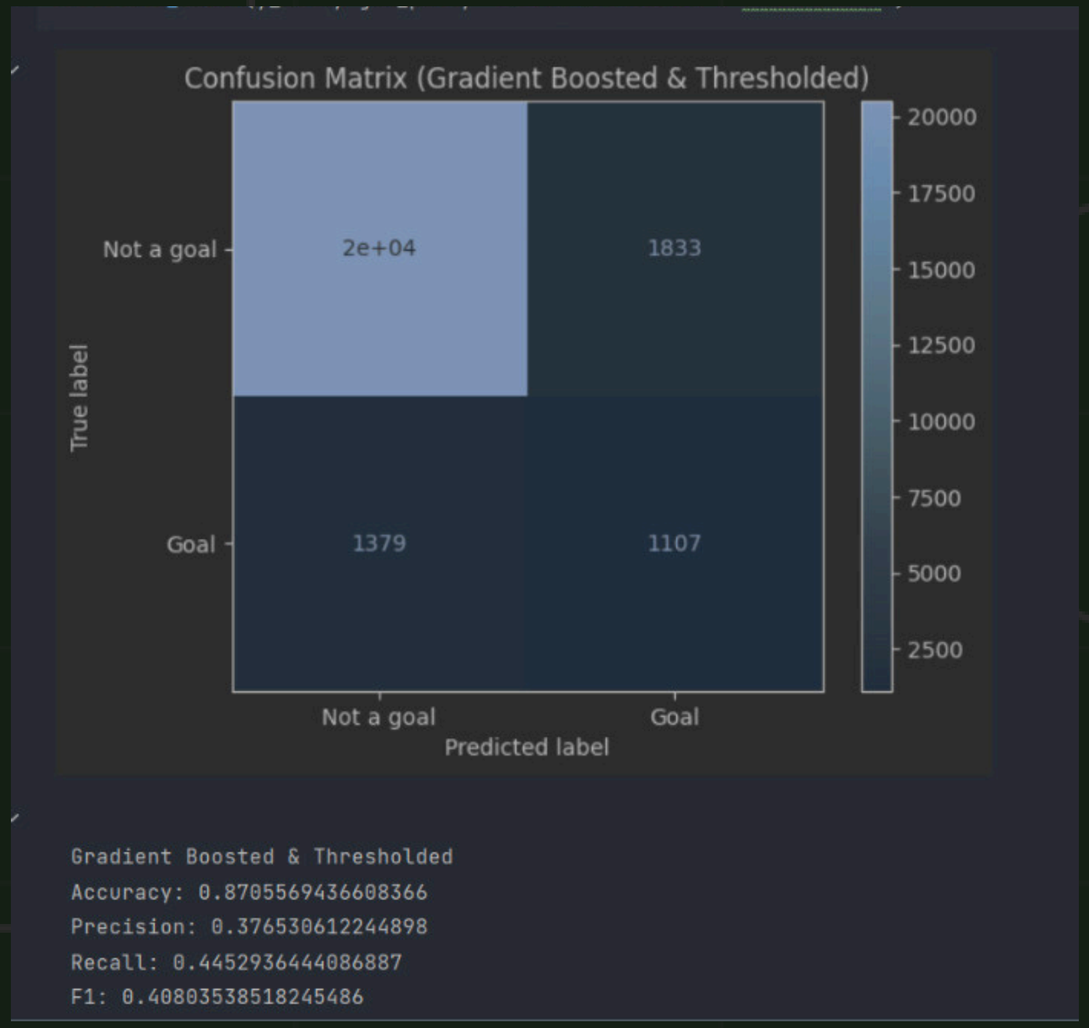
PERMUTATION FEATURE IMPORTANCE

PFI evaluates feature importance by randomly permuting feature values and measuring the change in model performance.

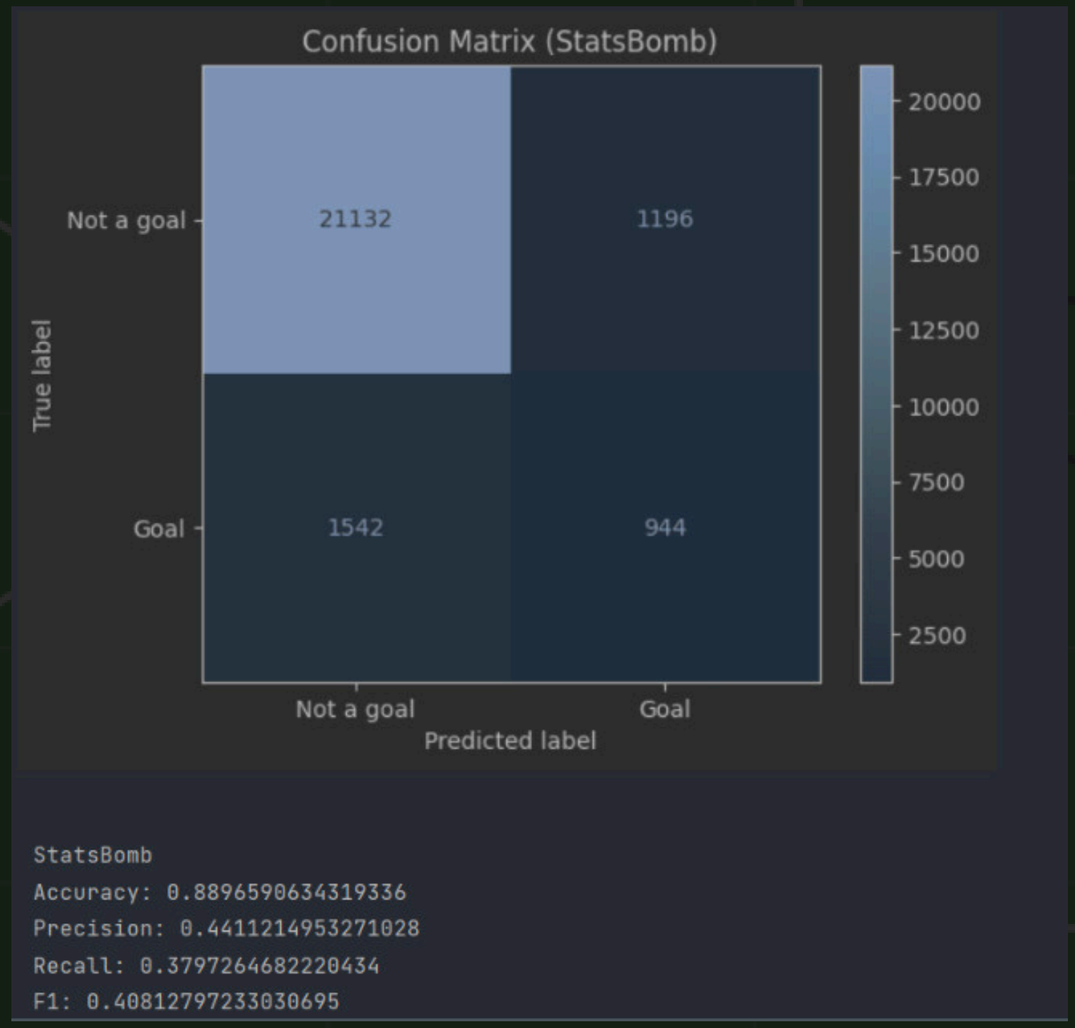
Provides a more accurate estimate of feature importance for the model we are using.

Provides easily interpretable results, as the feature importance scores are directly related to the model's performance.

Augmented results

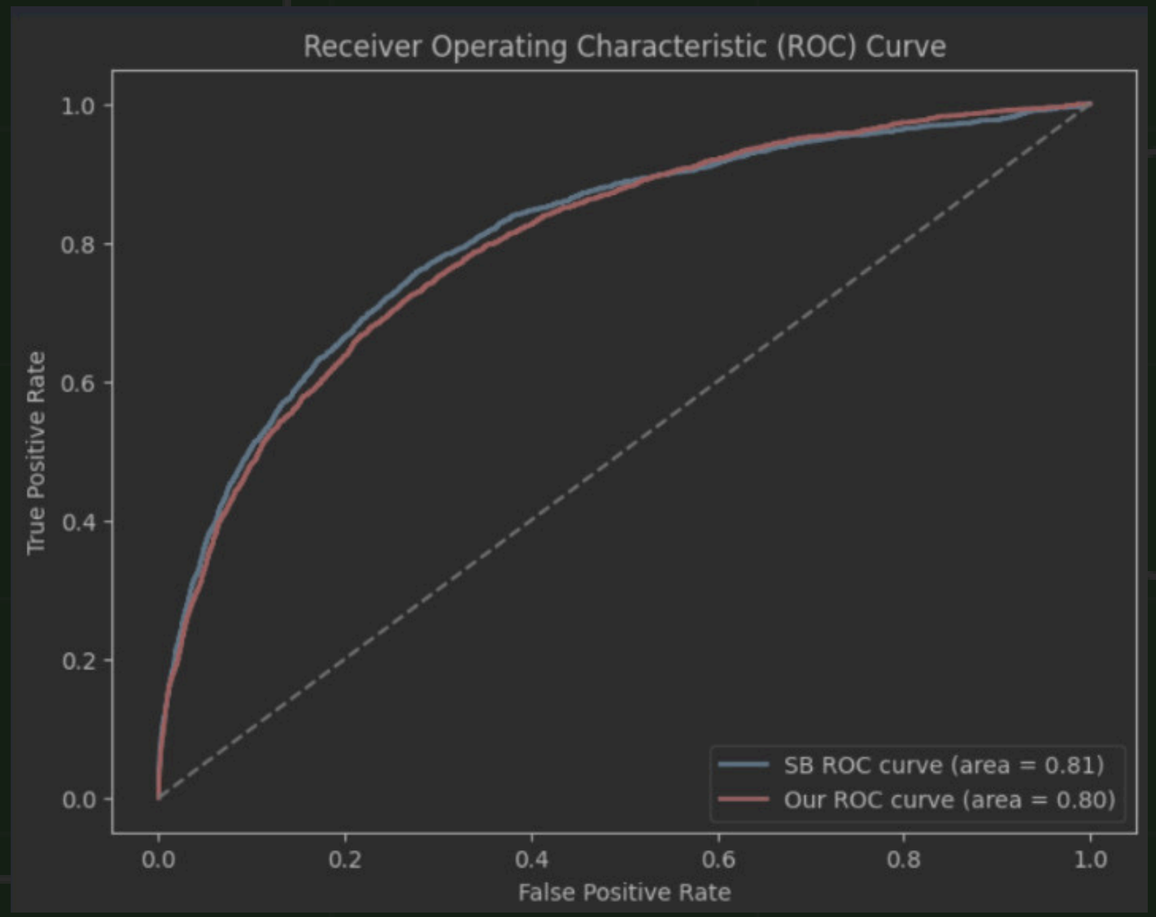


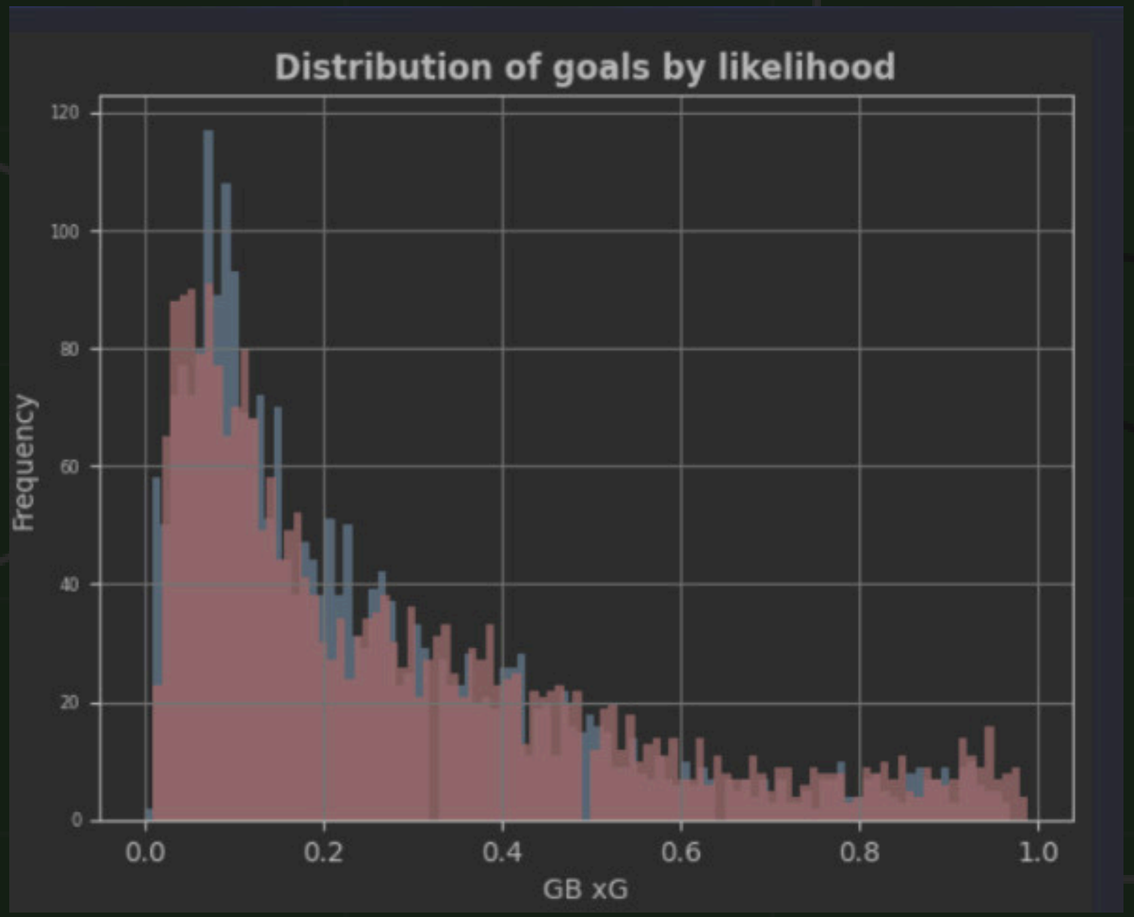
Gradient Boosted & Thresholded
Accuracy: 0.8705569436608366
Precision: 0.376530612244898
Recall: 0.4452936444086887
F1: 0.40803538518245486



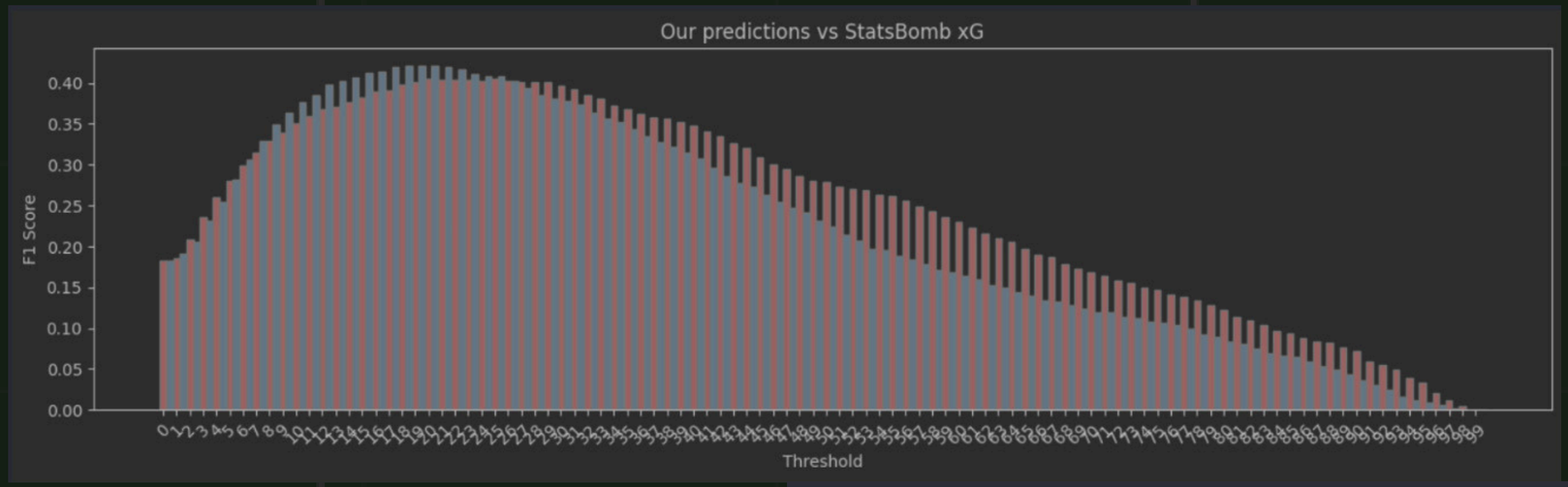
StatsBomb
Accuracy: 0.8896590634319336
Precision: 0.4411214953271028
Recall: 0.3797264682220434
F1: 0.40812797233030695

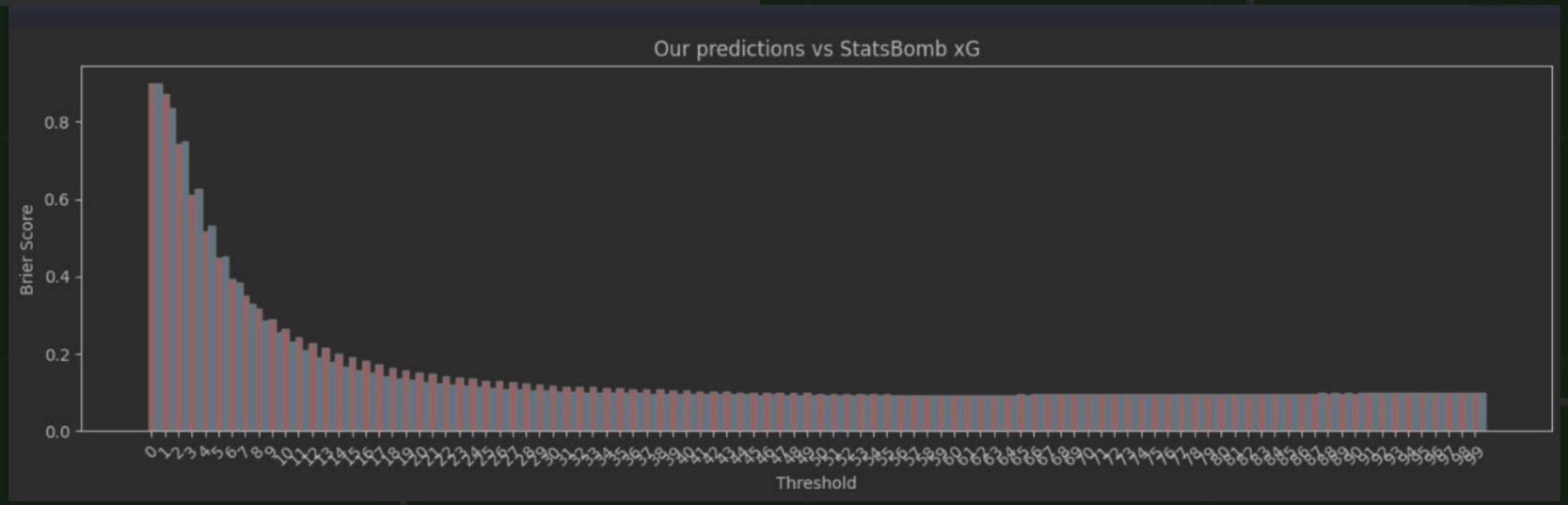
Augmented results



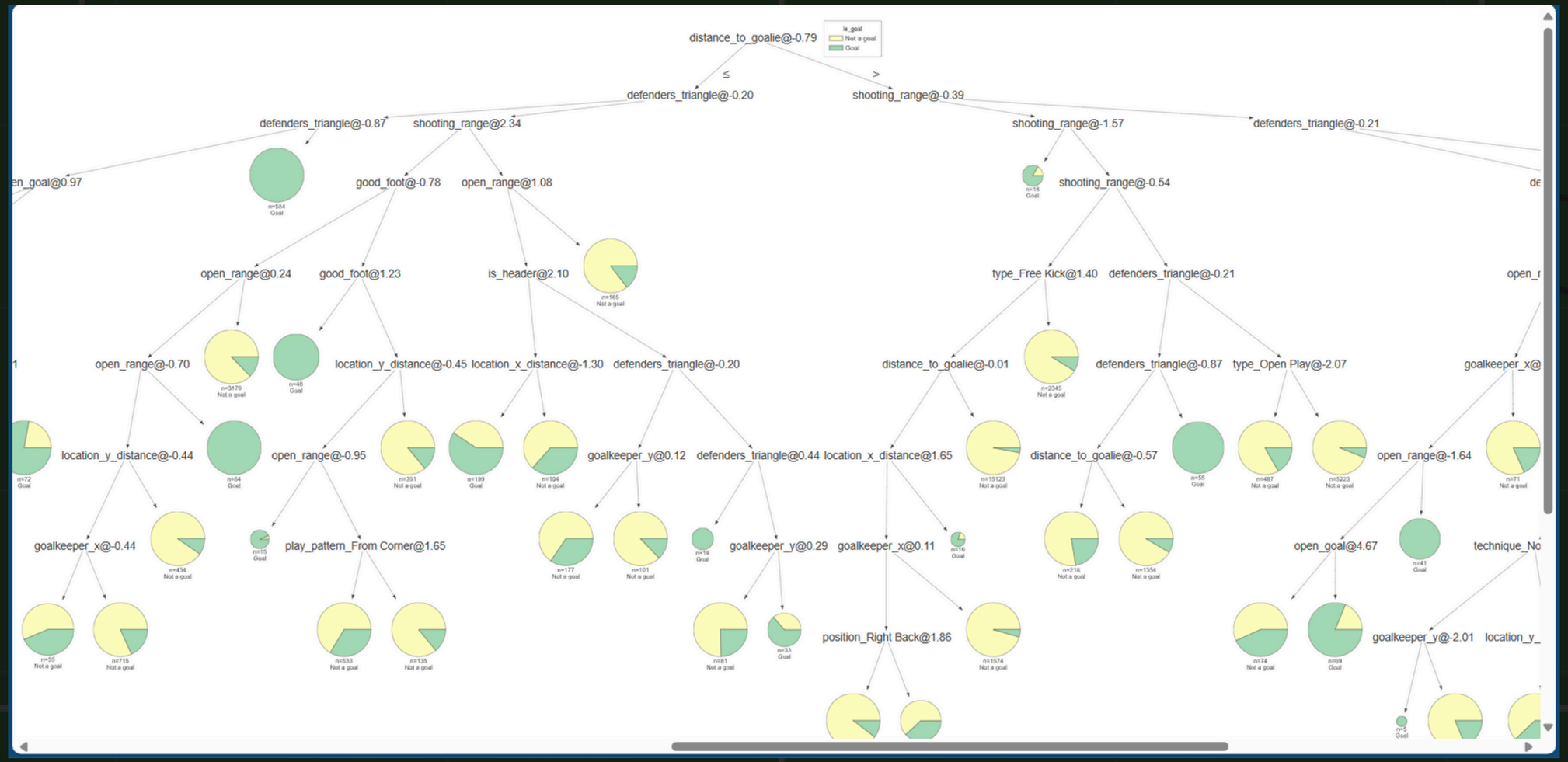


Augmented results





Interpretability



Interpretability (mutual information)

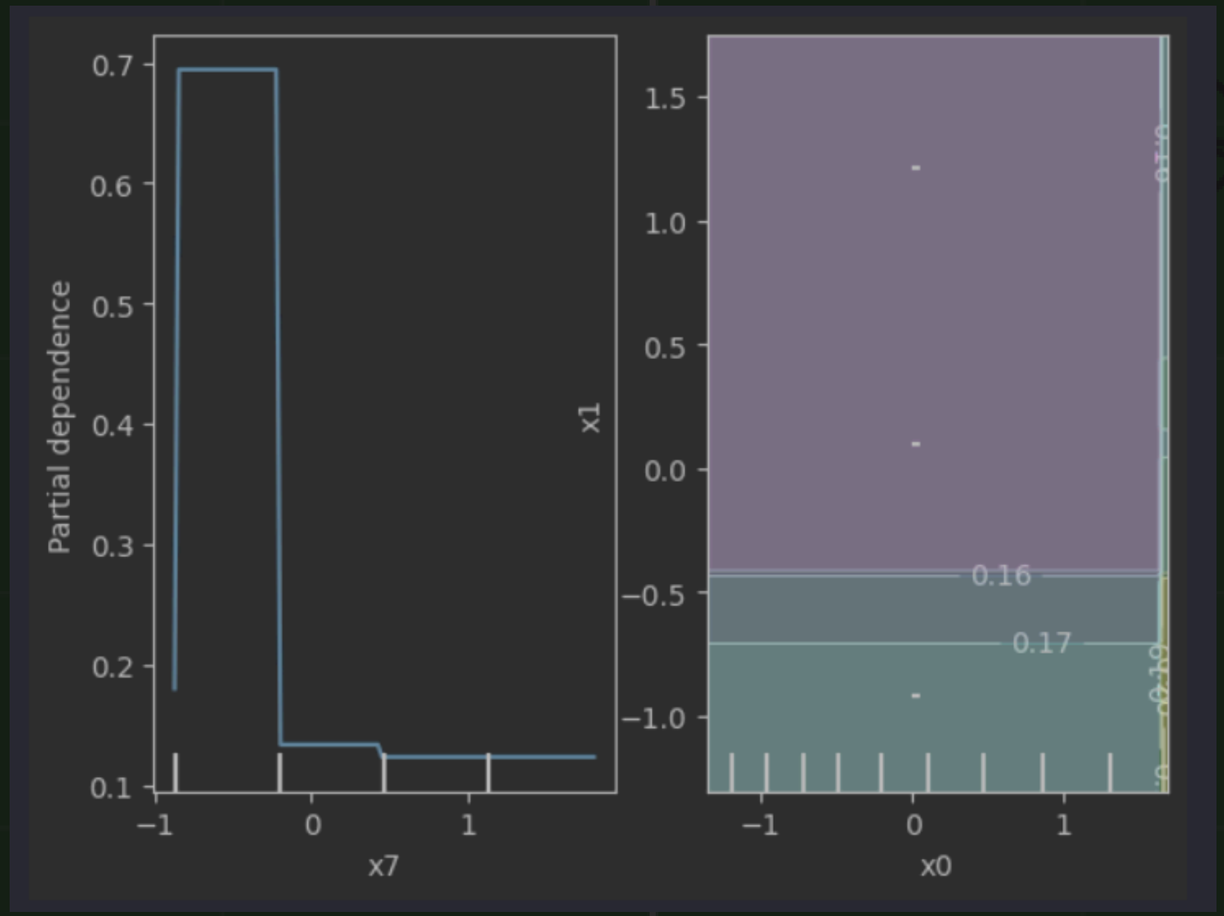
Out 4

location_x	0.033303
goalkeeper_x	0.027605
location_y	0.022635
goalkeeper_y	0.021592
open_goal	0.012058
period	0.010570
one_on_one	0.005486
first_time	0.005285
minute	0.004418
aerial_won	0.001703
possession	0.000323

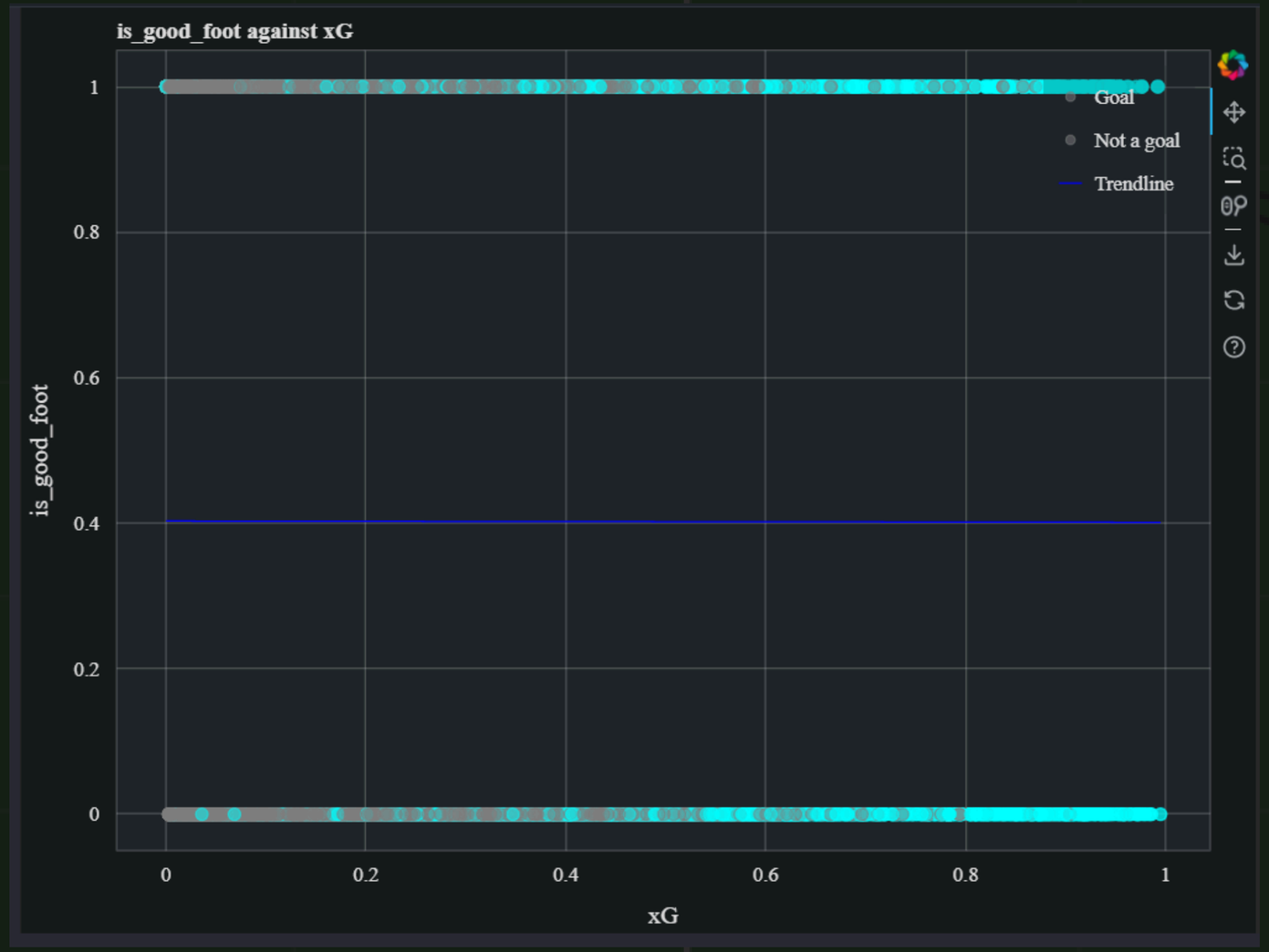
Out 40

	data
distance_to_goalie	0.059088
shooting_range	0.052146
best_distance	0.048842
defenders_triangle	0.041813
open_range	0.039278
location_x_distance	0.033526
goalkeeper_x	0.029003
location_y_distance	0.022313
goalkeeper_y	0.020528
is_penalty	0.016196
pass_length	0.013640
own_past_minute	0.013529
open_goal	0.011265
pass_duration	0.009901
period	0.009733
num_passes	0.007999
good_foot	0.006500
was_leading	0.005723
one_on_one	0.005277
player_type	0.005210
past_minute	0.004777
first_time	0.004634
minute	0.003564
own_past_15	0.003182
past_15	0.002907
is_header	0.002677

Interpretability (partial dependence plots)



Testing Good Foot: a failure and a lesson



Testing generalizability

We wanted to test the ability of the model on out-of-distribution data.

EXPERIMENT DESIGN

- Trained and tested on separate leagues (LaLiga and Bundesliga)
- Compared performance in-distribution (same league) and out of distribution (different league)
- Performed 100 repetitions of each experiment, saved F1 scores

Testing generalizability

RESULTS

- Baseline scores: LaLiga (0.365 ± 0.027), Bundesliga (0.355 ± 0.023)
- Out-of-distribution scores:
  - Train on Bundesliga, test on LaLiga (0.314)
  - Train on LaLiga, test on Bundesliga (0.337)
- Statistical significance testing on distribution of scores using Wilcoxon signed-rank test

Testing generalizability

FINDINGS

- Model performs significantly differently on baseline data
- Performs worse out-of-distribution than in-distribution
- Statistically significant differences found between the baseline and cross-league score for each league.

IMPLICATIONS

- Model generalizability is limited across different leagues
- League-specific factors (e.g. playing style) may affect performance

Feature importance methods

PERMUTATION FEATURE IMPORTANCE - LEAGUES TESTING

Top 5 most important features for LaLiga:

1. shooting_range (0.0927)
2. location_y_distance (0.0138)
3. defenders_triangle (0.0122)
4. goal_distance (0.0044)
5. is_header (0.0028)

Top 5 most important features for Bundesliga:

1. goal_distance (0.0786)
2. distance_to_goalie (0.0357)
3. defenders_triangle (0.0204)
4. location_y_distance (0.0104)
5. shooting_range (0.0056)

Feature importance methods

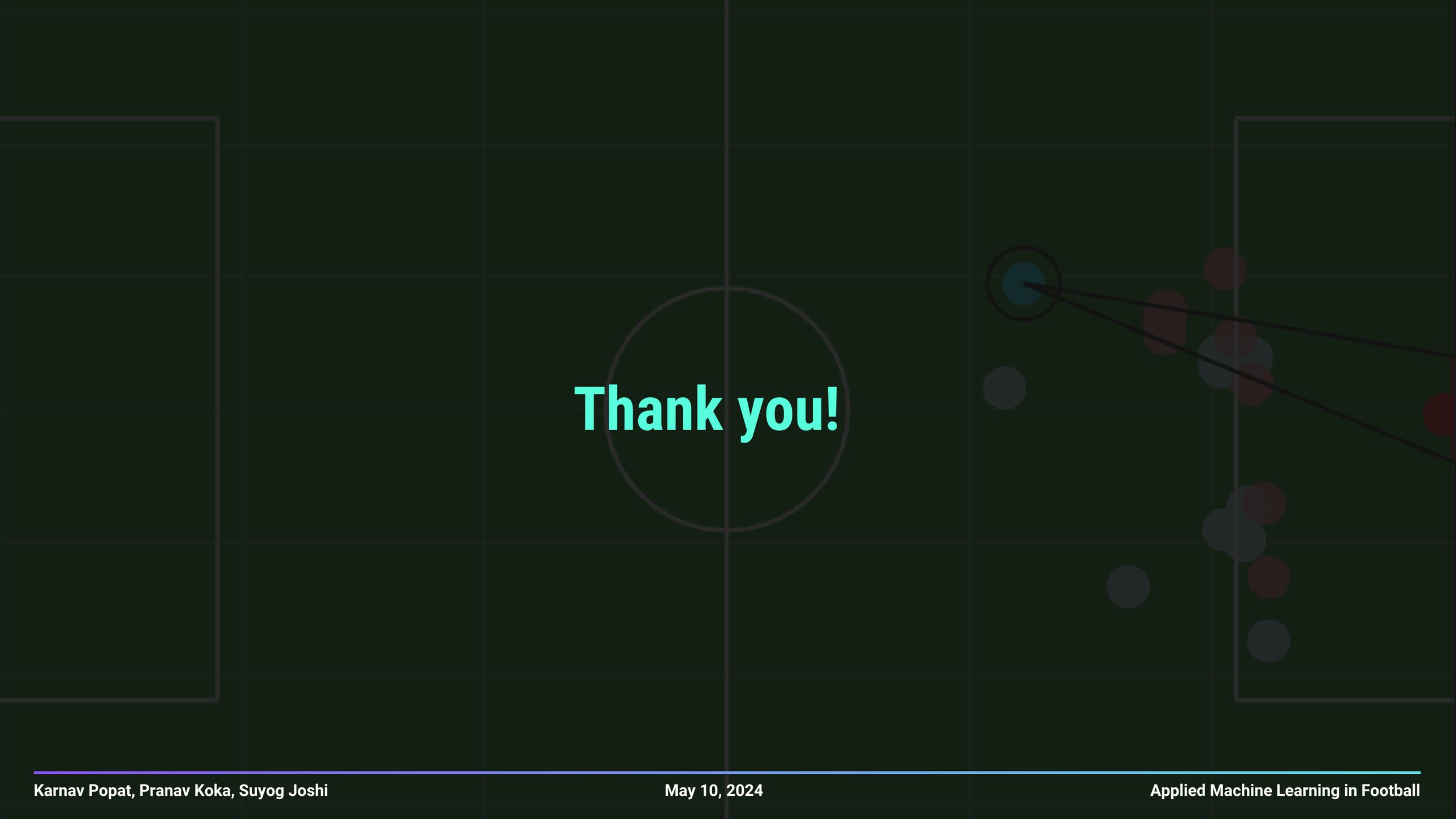
PERMUTATION FEATURE IMPORTANCE - IMPLICATIONS

Most of the top 5 features are common for both leagues, but are ranked differently.

shooting_range is the most important feature in LaLiga, but only ranks 5th in Bundesliga.
distance_to_goalie is present in Bundesliga, but is replaced by is_header in LaLiga.

This suggests differences in play styles among leagues, but more analysis needs to be done to pinpoint the differences, and how exactly each feature affects the model's perfomance.

Thank you!



with immense gratitude for

Our advisor, Professor Raghavendra Singh

Faculty advisors Professors Subhashis Banerjee and Anirban Sen

Aditya Padinjat

Kahaan Shah

Rudransh Mukherjee

Samhith Shankar

Vishnu Prakash

Hanu Trivedi

Aditi Warrier

Santripta Sharma

Saptarishi Dhanuka

Roshni Agarwal

Nandini Bhattacharya