xG: Predicting Shot Outcomes in Soccer

Lucas Kimball

AFC Richmond

Nearly relegated from the premier league last season, AFC Richmond are in need of some firepower if they want to stay up this season.

Business Problem

How can AFC Richmond improve their tactics to score more goals?

Can we get a sense for how potent Richmond's attack is? Are they underperforming?

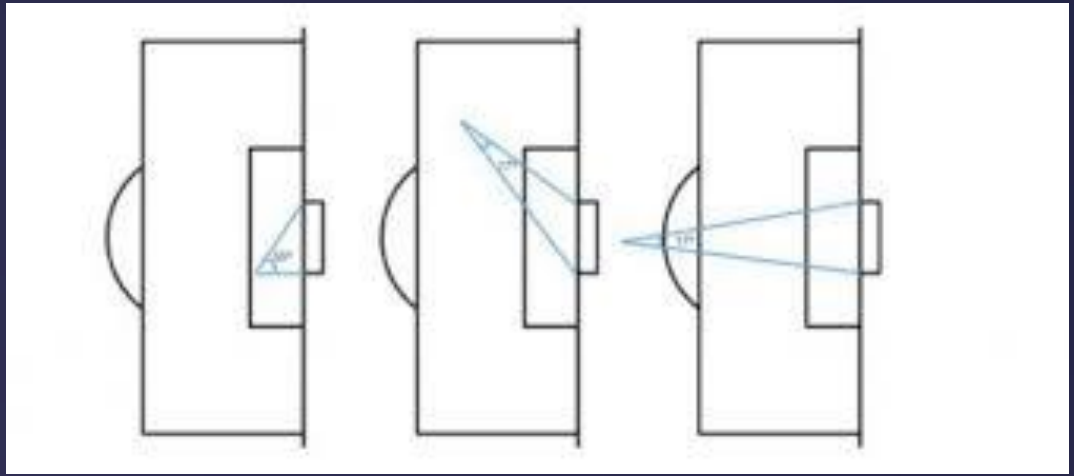
Which players can we target to improve their attack?

The Data

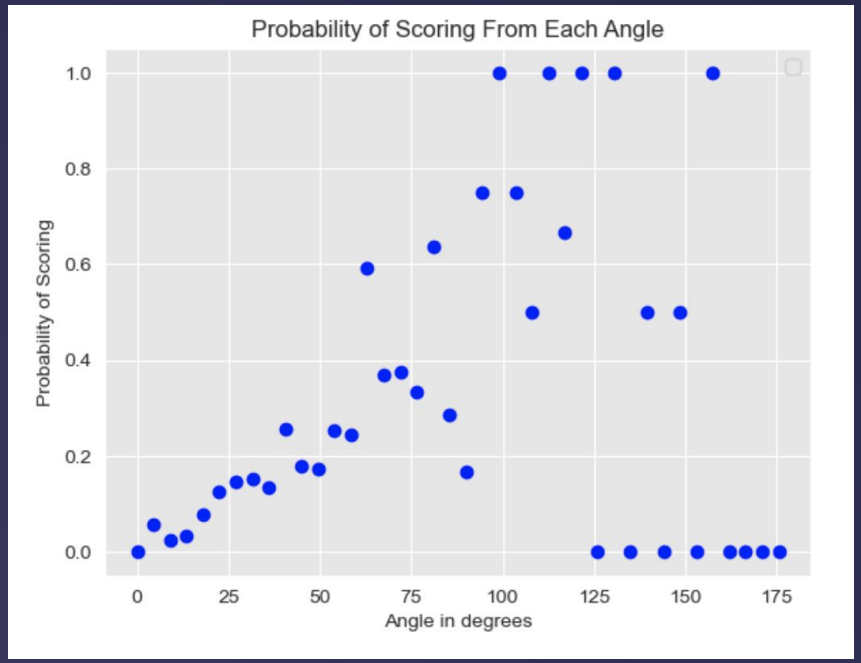
- From Statsbomb's free data repository
- 360 Event Data filtered for shots
- 2020 Euros
- 2020-2021 La Liga
- 2022 World Cup
- 3,348 shots between the three competitions

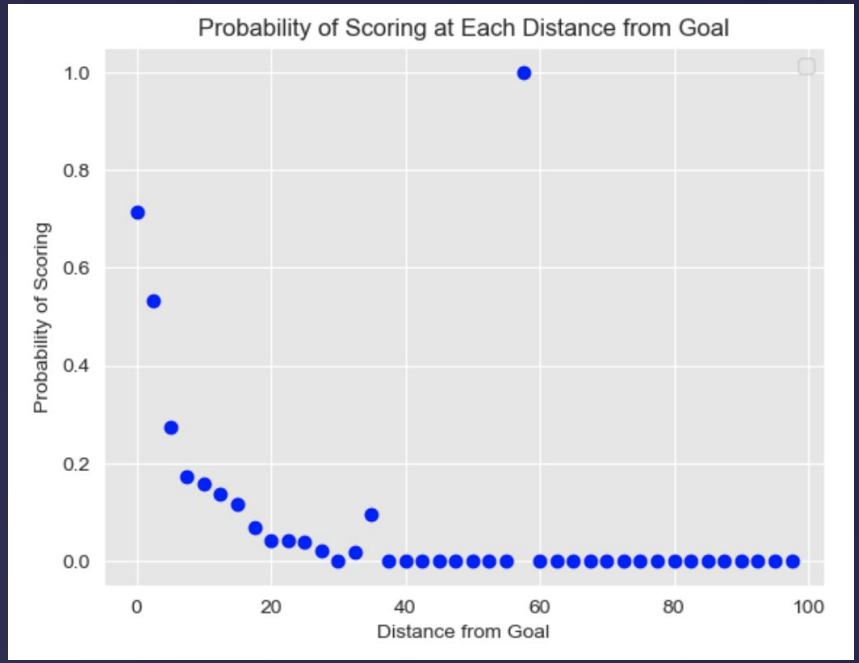
Creating Features

- Calculated the distance to goal for each shot
- Calculated the angle of each shots

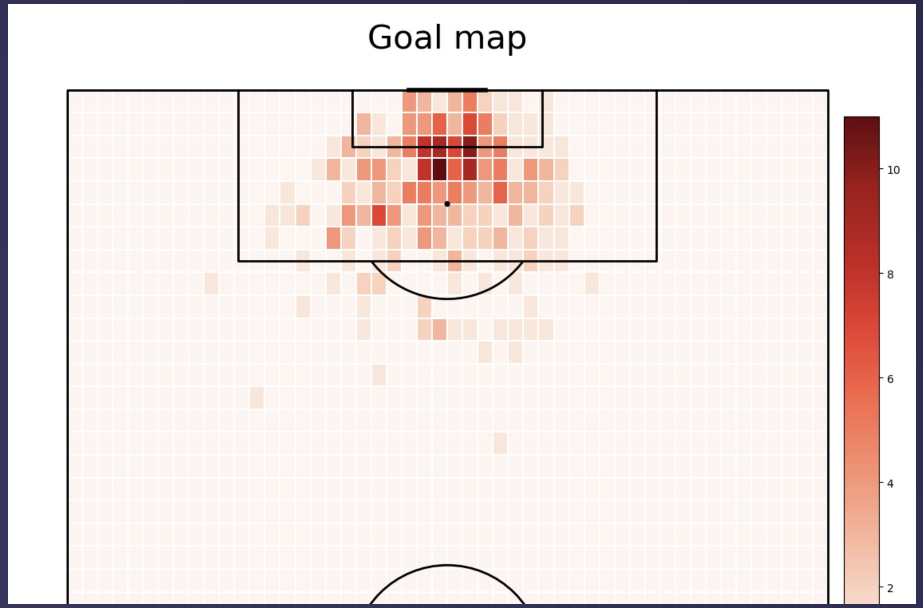


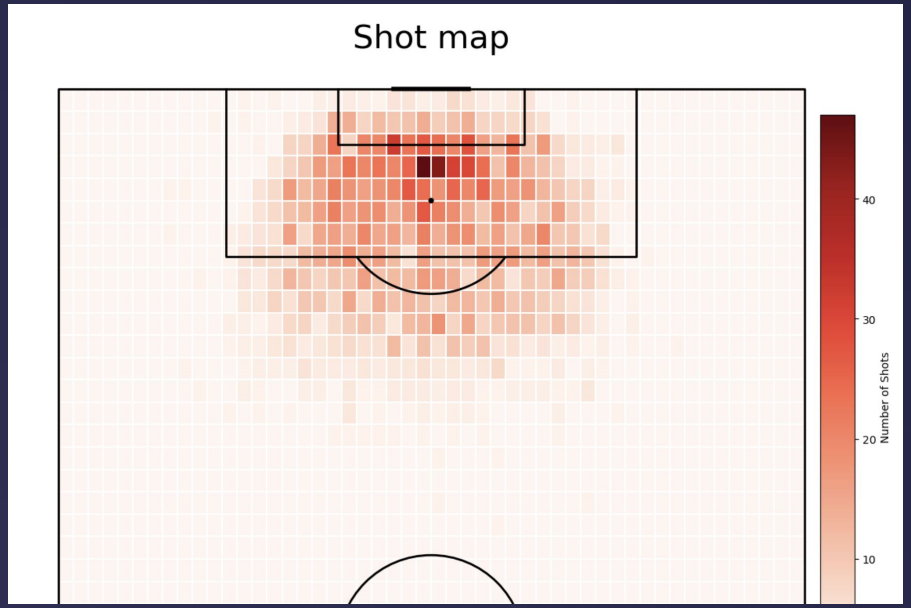
Shot Angle and Distance





Visualizing Goals vs Shots





Modeling Roadmap

1. Import Data
2. Isolate Shots
3. EDA
4. Preprocessing
   a. pd.get_dummies for categorical features
   b. Train-Test Split
   c. Standard Scaling
   d. SMOTE
5. Model: Logistic Regression
6. Model Tuning
   a. L1 + L2 penalties
   b. Threshold optimization
   c. C-value optimization
7. Model Evaluation
   a. ROC AUC

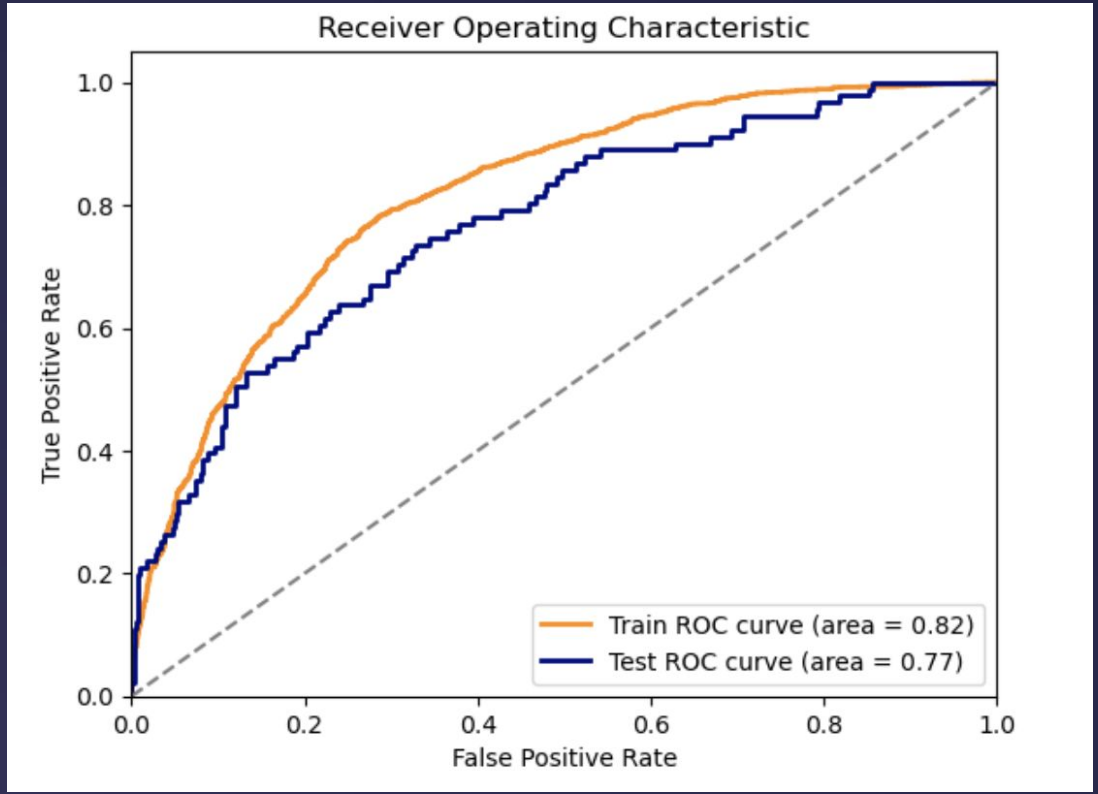
```
Original class distribution:

0    2979
1     369
Name: goal, dtype: int64
----------------------------------------
```

Model Performance

- Why ROC AUC?

- Optimized C parameter
- Optimized Threshold
- Test L1 and L2 penalties



Model Performance

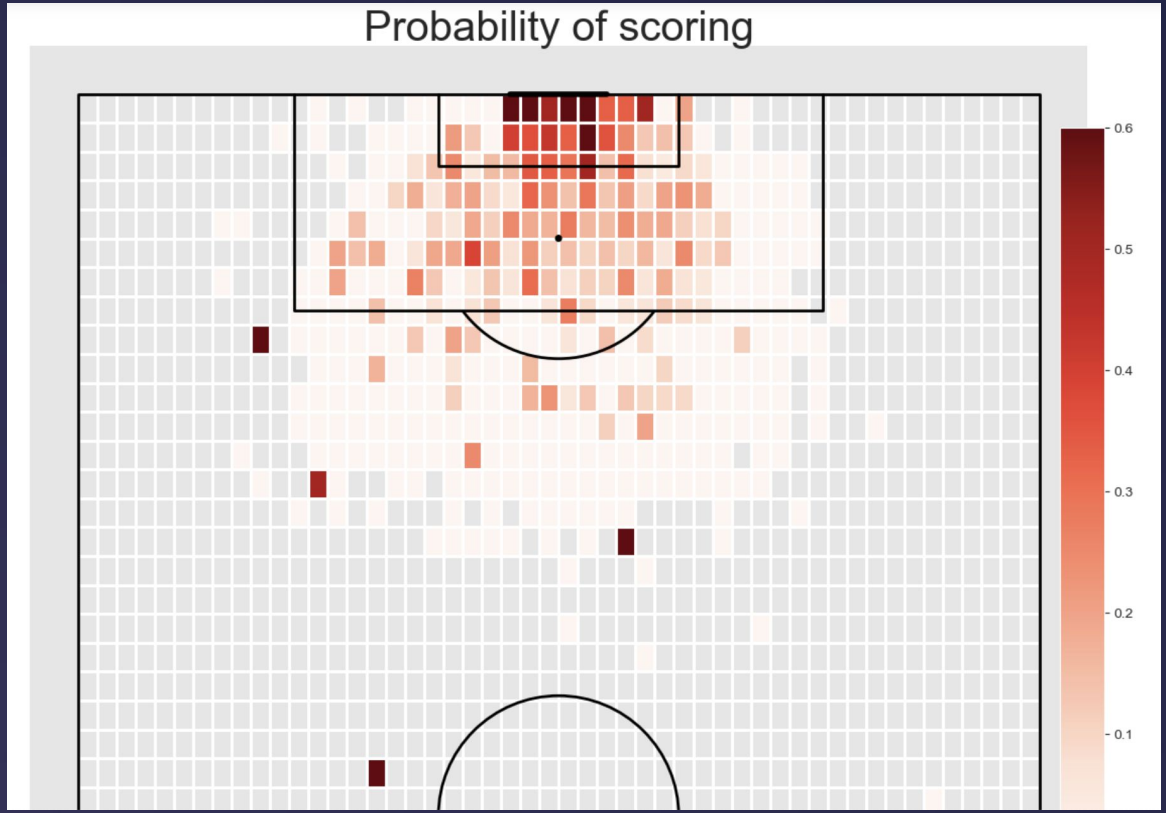
- The Coefficients with the largest magnitude are Angle and Distance

- Headed shots and shots using ‘other’ body parts are also highly correlated with scoring goals

```
Coefficients for each feature:
minute: 0.09914903824089066
shot_deflected: -0.04952079864568503
shot_first_time: -0.04952079864568503
shot_one_on_one: -0.04952079864568503
shot_open_goal: 0.09828824527674639
shot_follows_dribble: -0.04952079864568503
angle: 0.8322171346806971
distance: -0.8821305009785781
play_pattern_From Corner: -0.2919845931426634
play_pattern_From Counter: 0.007786219714834772
play_pattern_From Free Kick: 0.10506668831700669
play_pattern_From Goal Kick: 0.02336506262832272
play_pattern_From Keeper: 0.08046241936615452
play_pattern_From Kick Off: -0.037665030734182495
play_pattern_From Throw In: 0.008772888565810604
play_pattern_Other: -0.05308552530040935
play_pattern_Regular Play: 0.11980666191990992
shot_body_part_Head: -0.5282520952443823
shot_body_part_Left Foot: 0.30965678056963253
shot_body_part_Other: -0.21772989154191508
shot_body_part_Right Foot: 0.14417649227797708
shot_technique_Backheel: -0.08182394772712827
shot_technique_Diving Header: 0.06015816201572259
shot_technique_Half Volley: -0.11651577098515206
shot_technique_Lob: 0.2174188532082075
shot_technique_Normal: 0.15122847216930946
shot_technique_Overhead Kick: -0.09846668877483798
shot_technique_Volley: -0.1503064671251348
```

How AFC Richmond should use this model: Tactics

- Work the ball into the middle of the box
- Focus on getting closer to the goal before shooting



How AFC Richmond should use this model: Self-Diagnosis

- xG is a better predictor of future goals than goals scored itself
- Run the past season's data through this model
- Diagnose whether Richmond was unlucky or poor at creating chances
- Same can be done to determine defensive strength

How AFC Richmond should use this model: Recruitment

- Look to transfer in players with a large amount of expected goals accumulated
- Potential Market Inefficiency with goals scored a more highly regarded statistic in player valuation

Next Steps

- Use model on larger dataset to find specific players to recommend
- Develop model to determine value of non-goalscorers (ie. xThreat, Goal Probability added)
- Incorporate features such as defender and goalkeeper positioning

Questions?

Lucas Kimball

- Email: lucaskimball98@gmail.com
- GitHub: https://github.com/LMK08
- LinkedIn: https://www.linkedin.com/in/lucas-kimball-/