Goal Scoring Data

By: Adelaide Gilley, Walker Oettl, Johann Perera, Dillon Sullivan

Our "Goal"

- Understand and visualize the data!
- Create a model that predicts shot attempt outcome
- Use model to optimize goal scoring efficiency on our team's attack
- Goal-Scoring Efficiency (Tactically)

Methodology

What data we used...

- ❖Opta Goals, Attempts & Build up
  - ❖Match Event Detail
  - ❖x Source
  - ❖y Source
- ❖Match Template- Technical(Gen)
  - ❖Formation Played
- ❖Our Additions
  - ❖Distance from Goalmouth*
  - ❖Zones on Field

Johann

Formation Data Exploration

Formation Played	Goal Count	Percent by Formation
4-2-3-1	78	73.58%
4-4-2	8	7.55%
4-1-4-1	5	4.72%
3-4-2-1	4	3.77%
4-3-3	4	3.77%
4-2-2-2	3	2.83%
5-4-1	3	2.83%
5-3-2	1	0.94%

Johann

Formation Data Exploration

Opponent Formation	Goal Count	Percent by Formation
4-2-3-1	52	49.06%
4-1-4-1	14	13.21%
3-4-2-1	11	10.38%
4-3-3	9	8.49%
4-4-2	7	6.60%
4-4-1-1	5	4.72%
5-4-1	4	3.77%
3-5-2	3	2.83%
3-4-1-2	1	0.94%

Johann

Logistic Regressions

Our Team

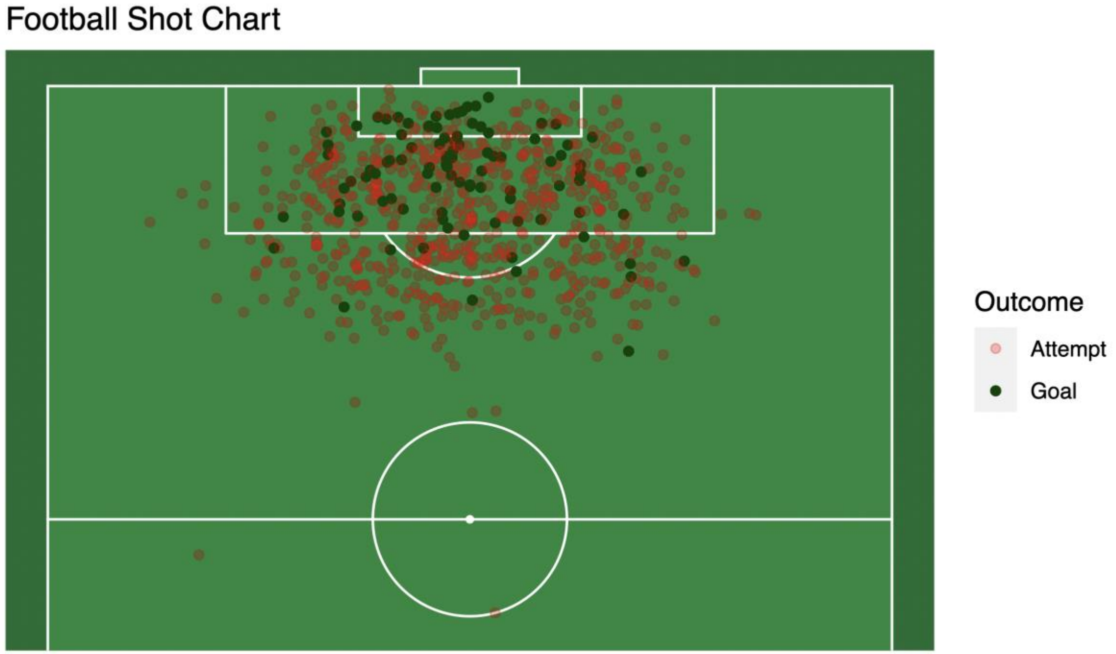
```
Call:
glm(formula = Goal ~ x.Source + y.Source + Form3421 + Form4141 +
    Form4222 + Form4231 + Form433 + Form442 + Form532, family = binomial(link = "logit"),
    data = join)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0128  -0.6074  -0.4713  -0.3544   4.1308

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.934013   1.689513  -5.288 1.24e-07 ***
x.Source     0.079775   0.016868   4.729 2.25e-06 ***
y.Source     0.004150   0.008655   0.480    0.632
Form3421     0.340859   0.858800   0.397    0.691
Form4141    -0.330696   0.801376  -0.413    0.680
Form4222     0.147425   0.909513   0.162    0.871
Form4231    -0.032328   0.651944  -0.050    0.960
Form433      0.363959   0.854654   0.426    0.670
Form442      0.120531   0.750173   0.161    0.872
Form532     -1.310402   1.209077  -1.084    0.278
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 614.09  on 759  degrees of freedom
Residual deviance: 584.60  on 750  degrees of freedom
AIC: 604.6

Number of Fisher Scoring iterations: 5
```

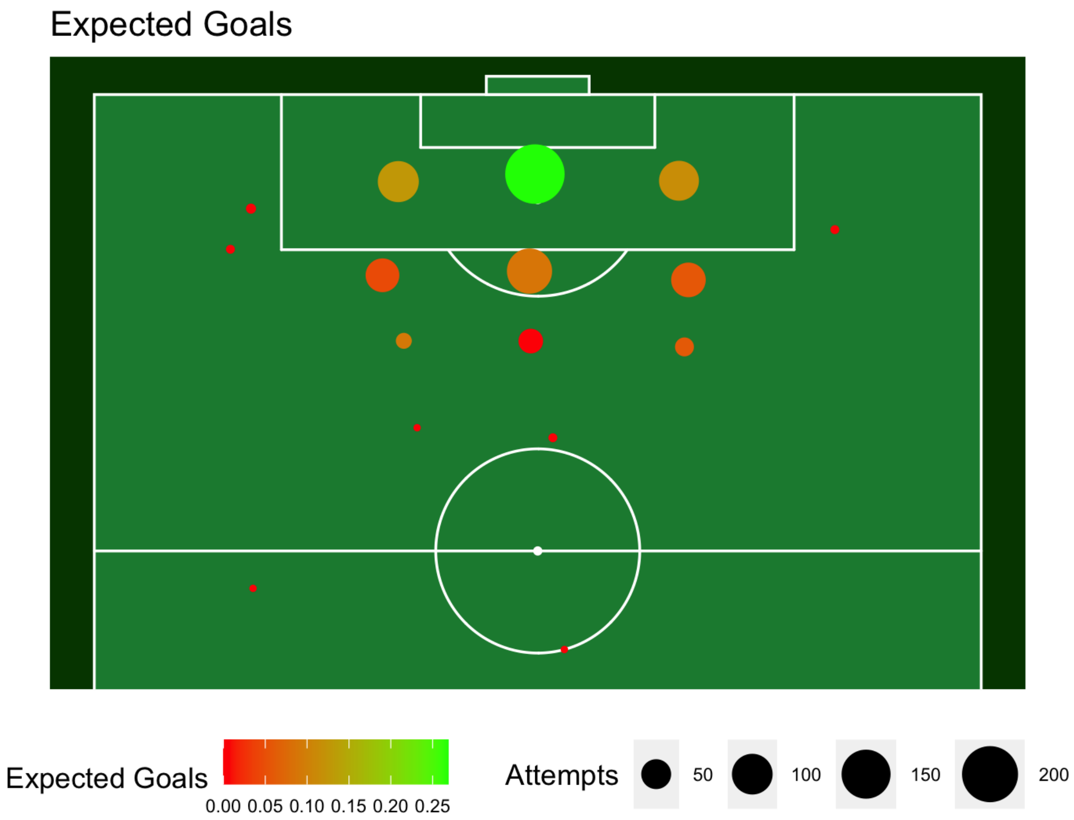
Opponent

```
Call:
glm(formula = Goal ~ x.Source + y.Source + Opp_Form3421 + Opp_Form4141 +
    Opp_Form4231 + Opp_Form433 + Opp_Form442 + Opp_Form3412 +
    Opp_Form352, family = binomial(link = "logit"), data = join)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9452  -0.6088  -0.4802  -0.3533   4.0146

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.693210   1.589761  -5.468 4.54e-08 ***
x.Source      0.079049   0.016927   4.670 3.01e-06 ***
y.Source      0.004401   0.008625   0.510    0.610
Opp_Form3421 -0.224536   0.497976  -0.451    0.652
Opp_Form4141  0.189675   0.481811   0.394    0.694
Opp_Form4231 -0.299201   0.401726  -0.745    0.456
Opp_Form433  -0.305569   0.520913  -0.587    0.557
Opp_Form442  -0.325548   0.553427  -0.588    0.556
Opp_Form3412 -1.064087   1.106209  -0.962    0.336
Opp_Form352  -0.030787   0.742351  -0.041    0.967
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 614.09  on 759  degrees of freedom
Residual deviance: 585.08  on 750  degrees of freedom
AIC: 605.08

Number of Fisher Scoring iterations: 5
```

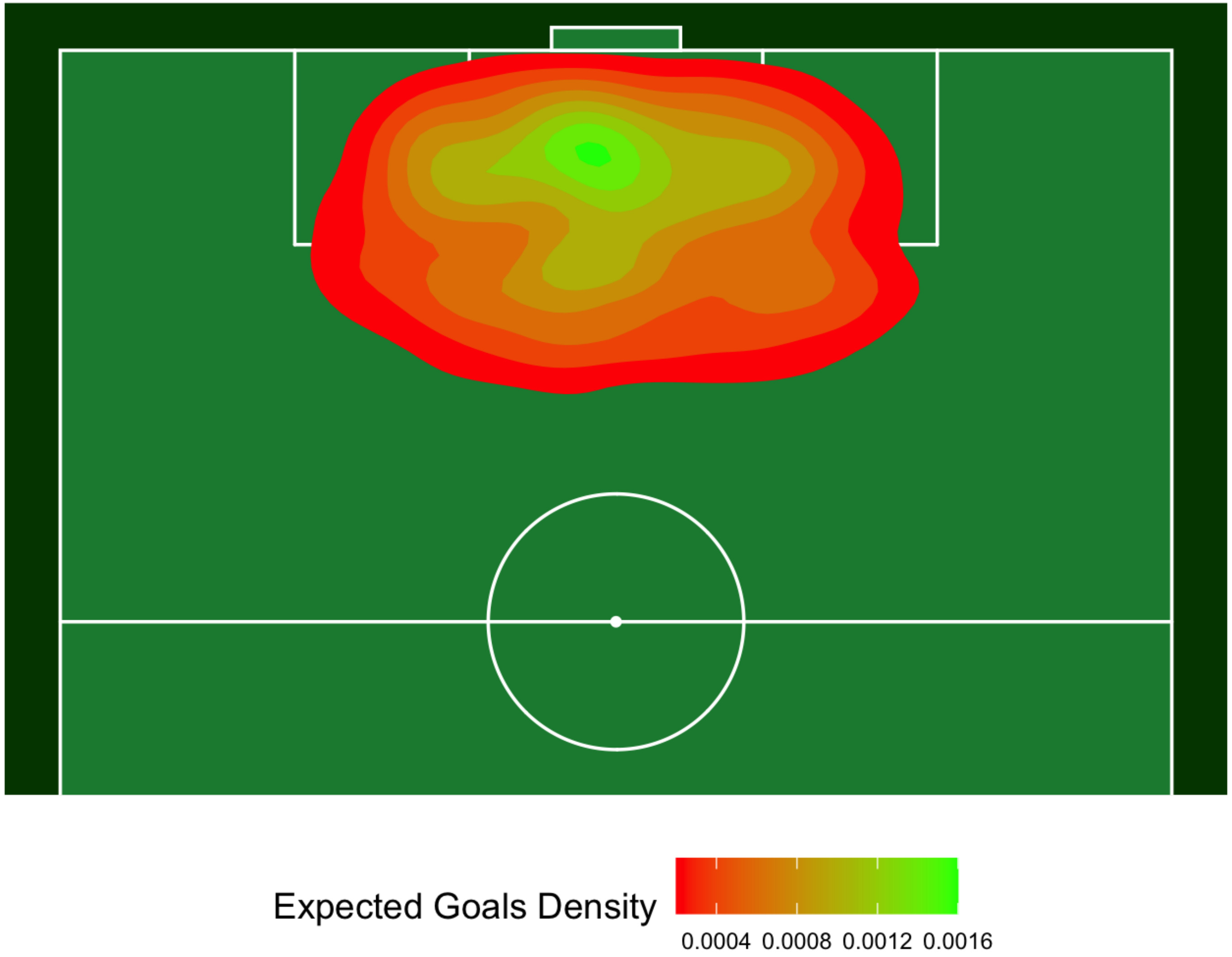
Dillon

Exploring The Goal Data through Visualizations





Expected Goals Heat Map



- This is a heat map showing the density of expected goals based on the location of shots taken during the season.

Adelaide

Our Model

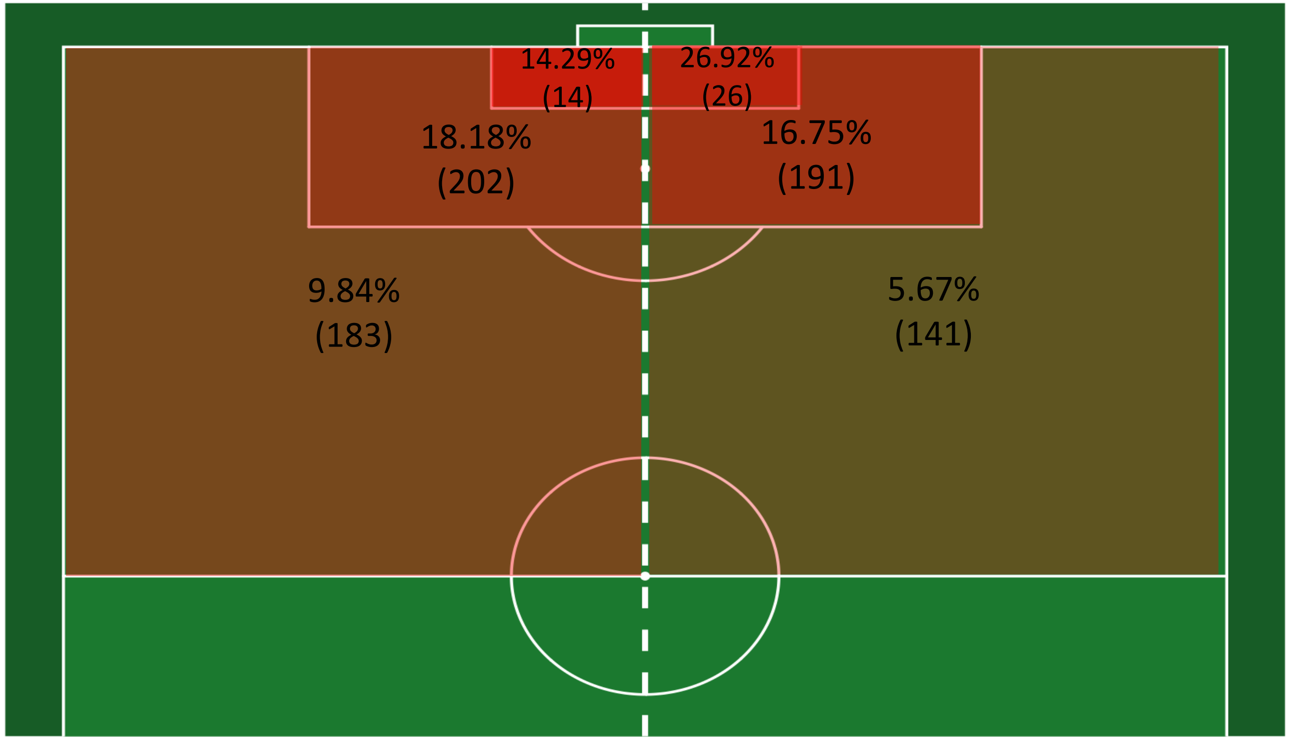
Adelaide

Making The Zones

- Before creating a model to predict goals, we wanted to make our own zones to further analyze where shots are being taken.
- Created a new variable 'zone' for each shot based on its X,Y.
- Used ifelse() function to define each shot zone.

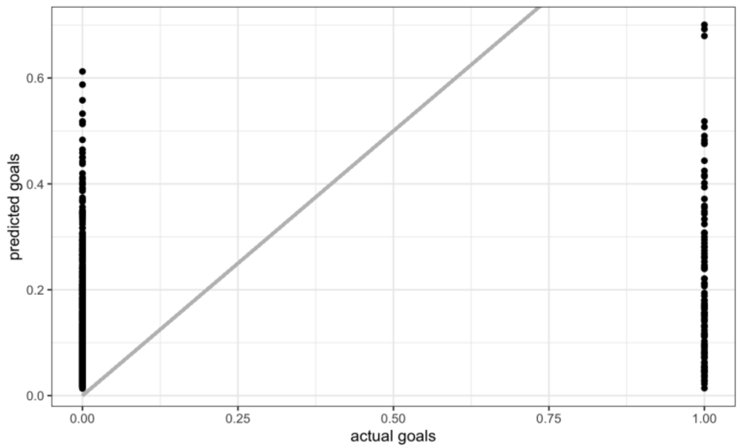
zone	efficiency
18 Yard Left	18.81%
18 Yard Right	16.75%
6 Yard Left	14.29%
6 Yard Right	26.92%
Outside Box Left	9.84%
Outside Box Right	5.67%
Own Half	0.00%

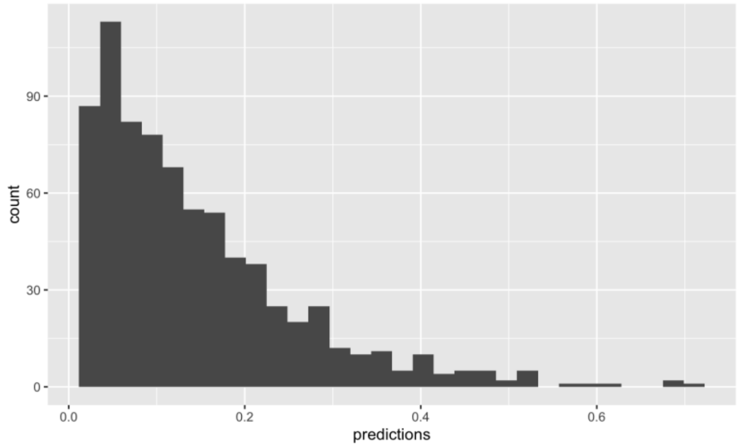
Football Zone Chart

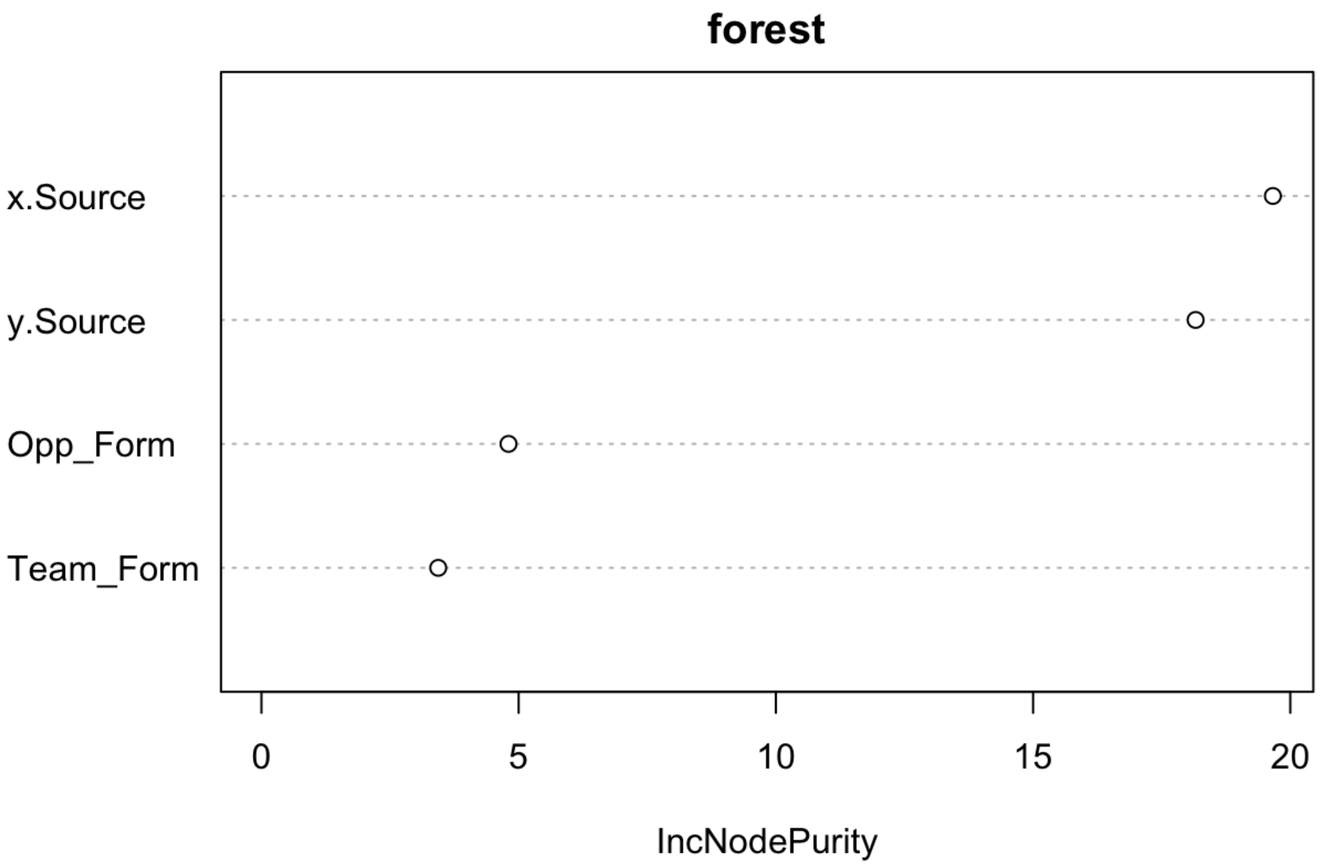


Adelaide

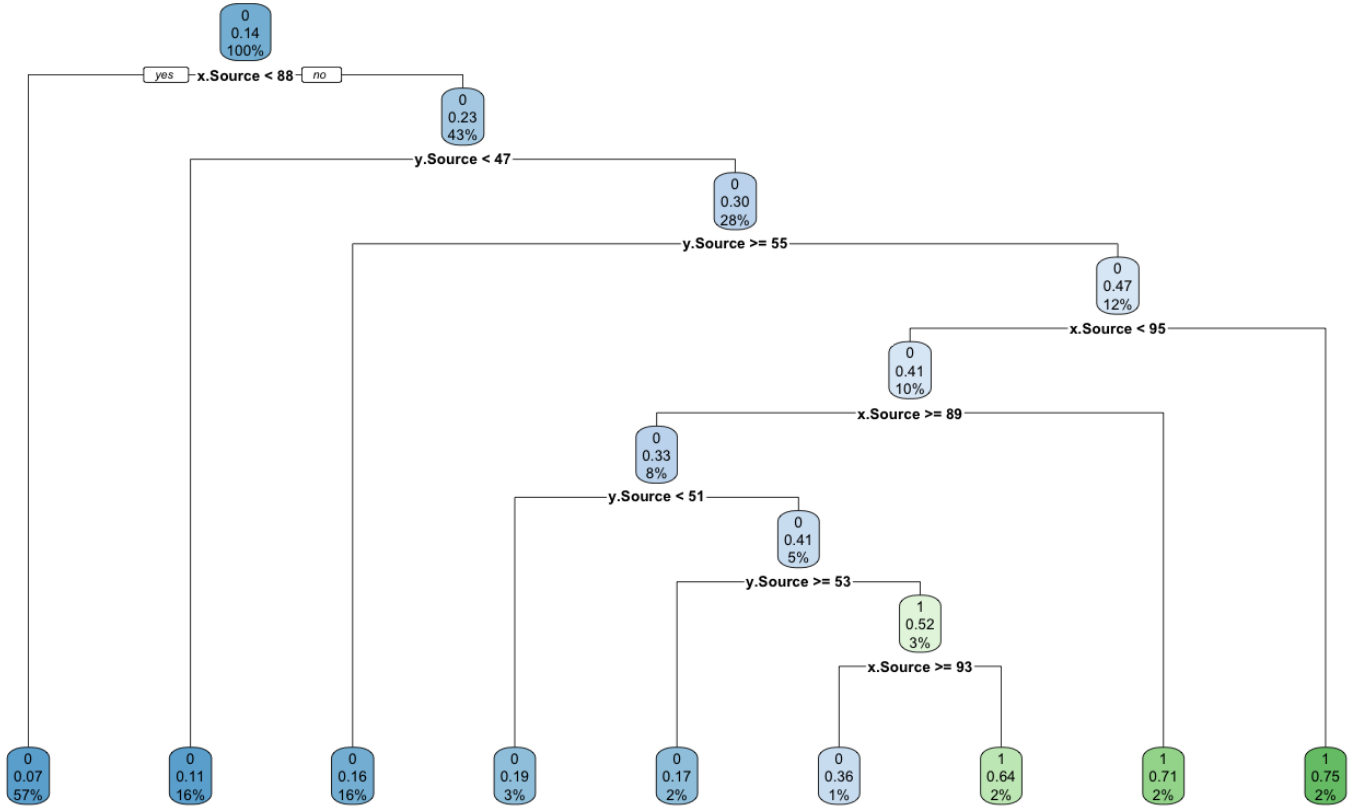
Random Forest

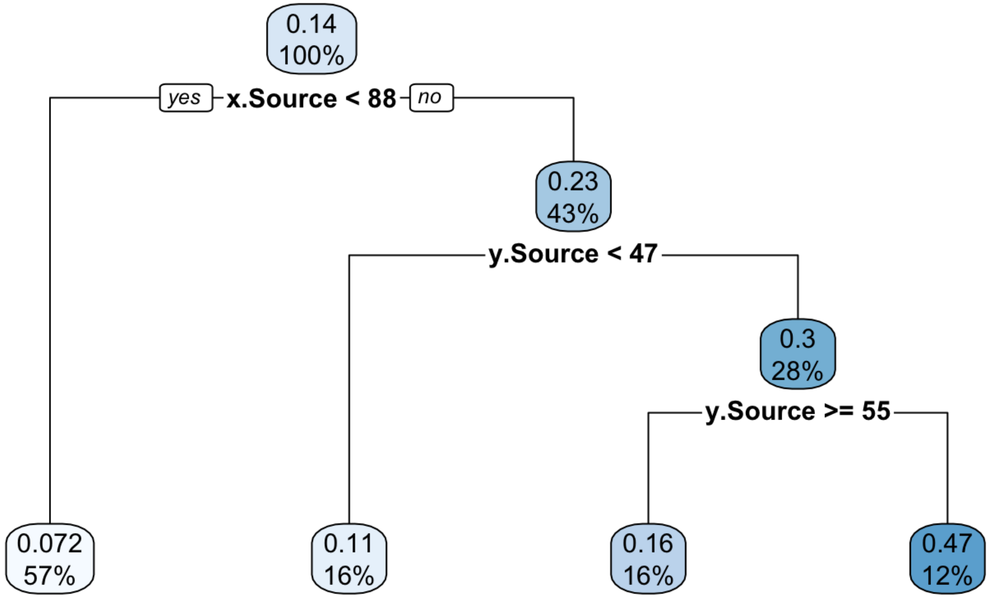






Pre-Optimization Trees





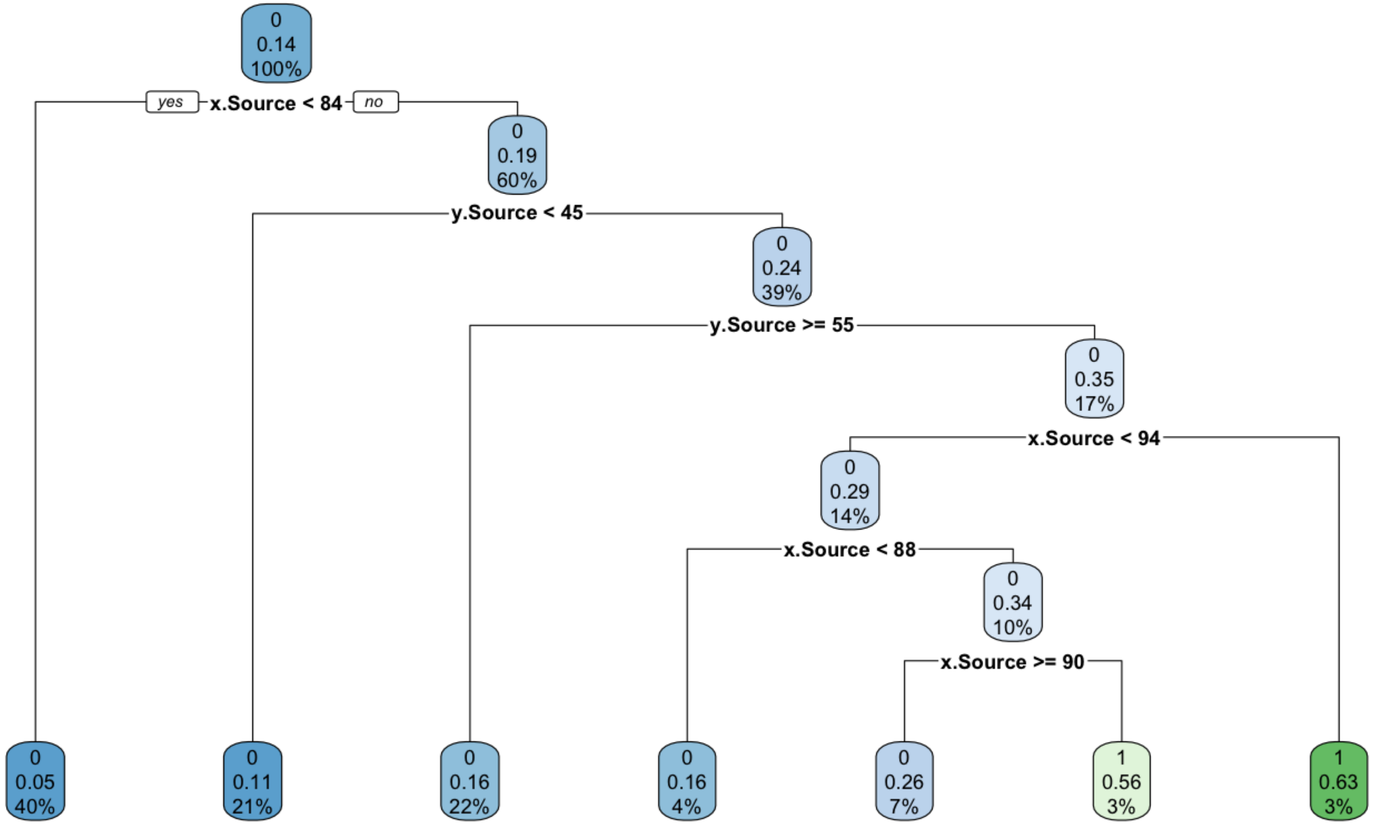
Model of Freedom 

Constrained MaxDOT

Walker

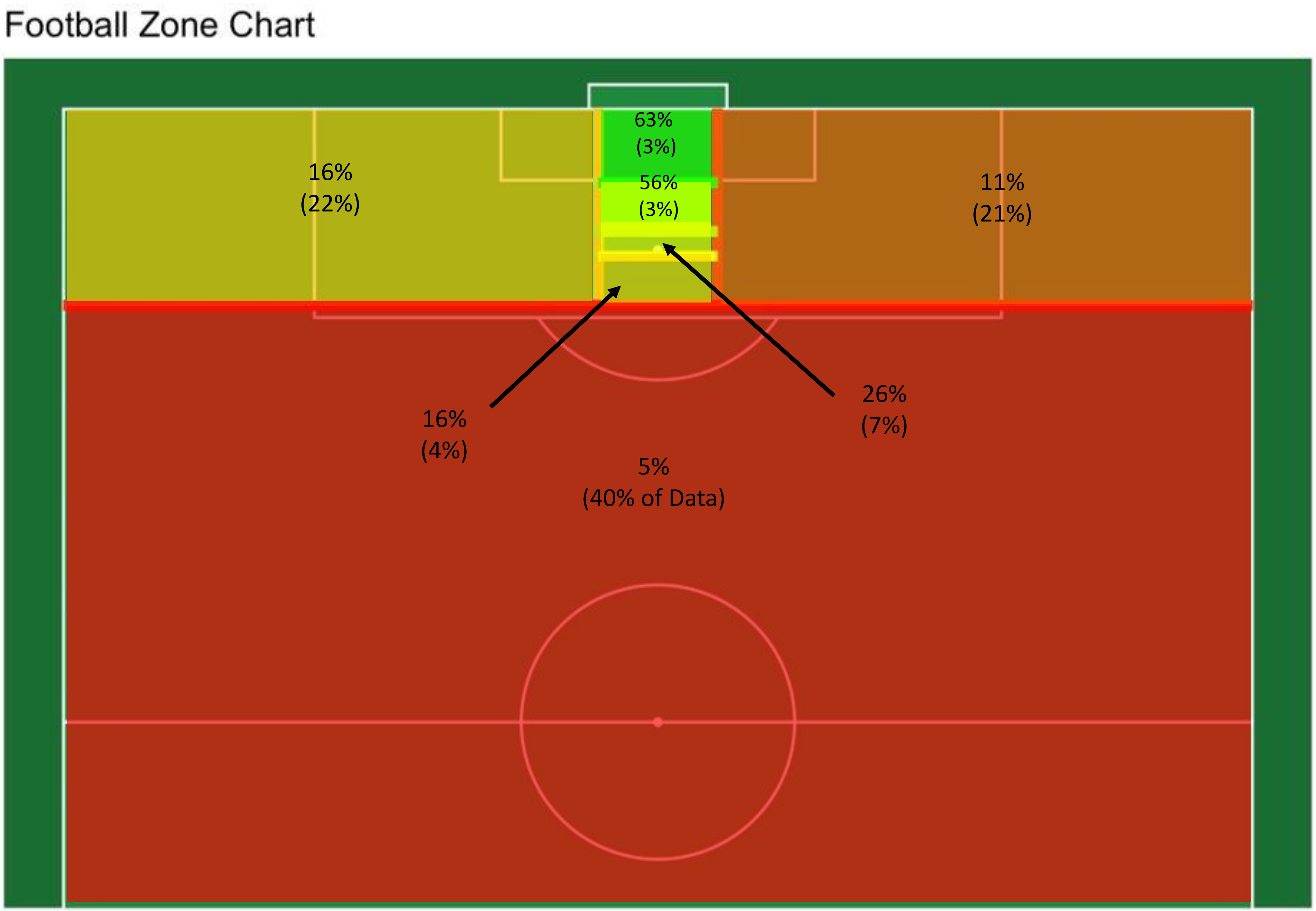
Final Model - Optimized

- Needed to have significant predictive power
- Optimized:
  - Minsplit
  - Minbucket
- Most digestible machine learning model (this class)
- Helped us look for specific zones on pitch to target



Dillon

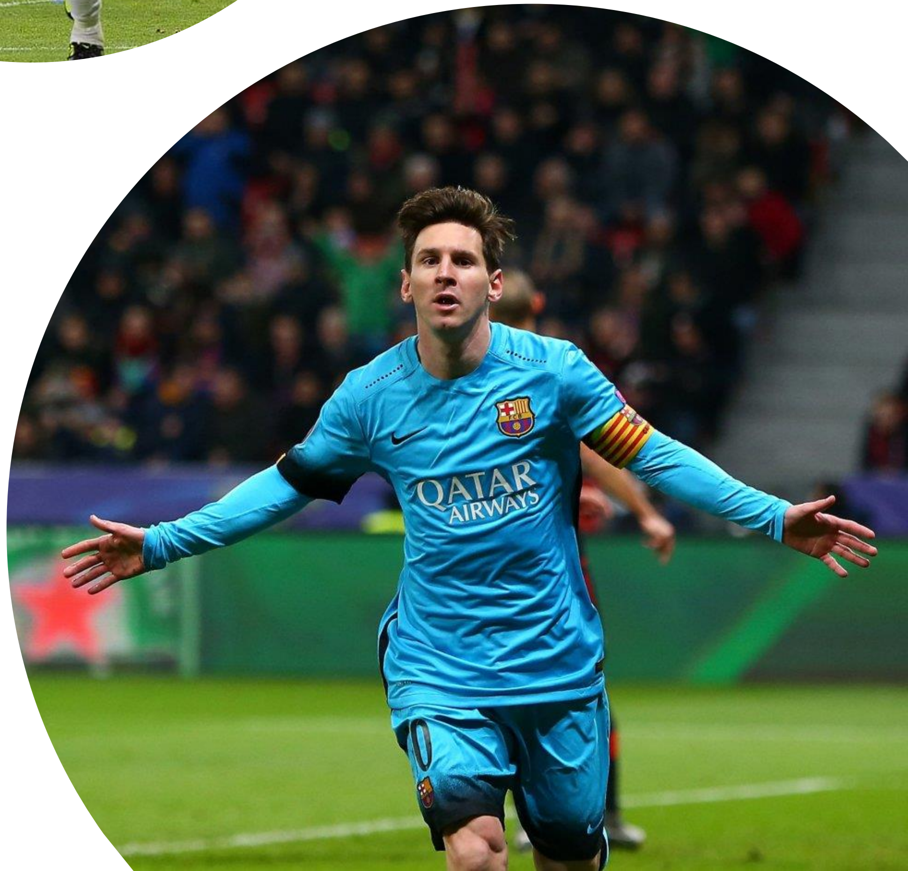
Final Model - Visualized



Dillon

What We "Scored"

- Better scoring efficiency closer to goal
  - **Expected**, but now quantifiable and useful
  - Most important split lied just inside the 18
  - Front/Back splits more than L/R
- No boost for back/near post runs
  - Middle becomes most efficient, even out to 18
  - Our team should look to exploit middle when possible, even if sacrificing distance from net



Walker

Suarez vs. Ghana 2014 WC

Suarez vs. England 2018 WC