Introducing the xG Premium

By Databall

Hpone Myat Khine (A0125002E)

Kwang Hun Lee (A0231958W)

Sae Jin Jang (A0231989M)

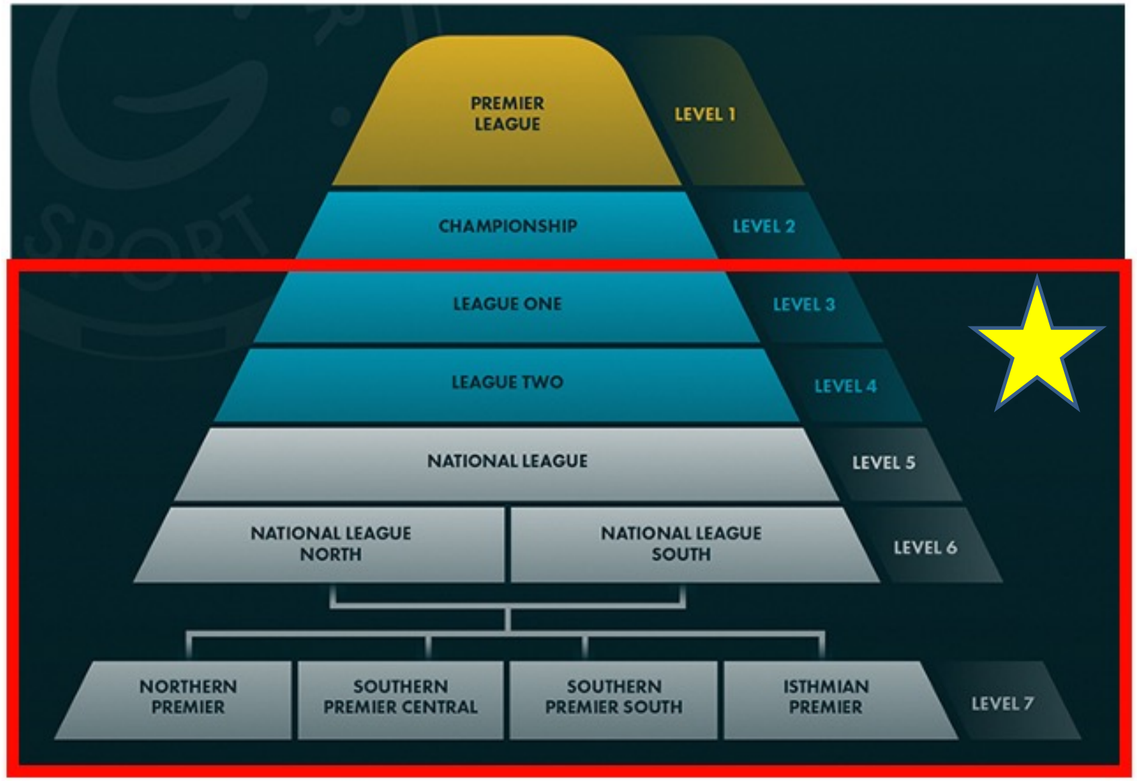
Football Analytics and Clients

StatsBomb work with teams in leagues across the globe...

Expected goals shown on the BBC's Match of the Day

Target Users

Primarily for football teams in lower leagues with limited resources to afford high cost analytics services



1) HIGH COST Football analytics service

2) ONLY the wealthy teams in top leagues can afford

→ Stronger entry barrier to top leagues

3) xG Premium can help to ease that barrier

→ High quality analytics service for FREE

What is Expected Goals (xG)?

Expected goals (xG) measures the quality of goal chances by calculating the likelihood of successfully scoring the goals, factoring in different conditions.

ex) Distance, Shot angle, Play pattern & style, etc.

Customized training system

Tactic Development

Recruitment

xG Example:

High xG value

Close to goal, no goalkeeper, low pressure, ball relatively stable, large angle to goal

Low xG value

Far away from goal, goalkeeper in goal, first-time half volley, small angle to goal

Why is xG useful?

(Gives insight to goal-scoring performance)

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▼
1	Mohamed Salah	EGY	FW	Liverpool	+3.5
2	Reece James	ENG	DF	Chelsea	+3.1
3	Maxwel Cornet	CIV	MF,FW	Burnley	+2.7

Overperformers

Underperformers

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▲
1	Francisco Trincão	POR	FW,MF	Wolves	-2.3
2	Nicolas Pépé	CIV	FW,MF	Arsenal	-2.1
3	Bryan Mbeumo	FRA	FW	Brentford	-2.0

Performing as expected →

Rk	Player	Nation	Pos	Squad	np:G-xG
5	Diogo Jota	POR	FW	Liverpool	-0.3

Source: Fbref.com (statsbomb)

Why is xG useful?

(Gives insight to goal-scoring performance)

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▼
1	Mohamed Salah	EGY	FW	Liverpool	+3.5
2	Reece James	ENG	DF	Chelsea	+3.1
3	Maxwel Cornet	CIV	MF,FW	Burnley	+2.7

Just
Lucky?

Overperformers

Performing as expected →

Rk	Player	Nation	Pos	Squad	np:G-xG
5	Diogo Jota	POR	FW	Liverpool	-0.3

Underperformers

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▲
1	Francisco Trincão	POR	FW,MF	Wolves	-2.3
2	Nicolas Pépé	CIV	FW,MF	Arsenal	-2.1
3	Bryan Mbeumo	FRA	FW	Brentford	-2.0

Just
Not lucky?

Source: Fbref.com (statsbomb)

Why is xG useful?

(Gives insight to goal-scoring performance)

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▼
1	Mohamed Salah	EGY	FW	Liverpool	+3.5
2	Reece James	ENG	DF	Chelsea	+3.1
3	Maxwel Cornet	CIV	MF,FW	Burnley	+2.7

After doing some investigating...

→ **Extraordinary goal scorer**

→ **Scored improbable goals**

→ **Limited sample size**

Overperformers

Performing as expected →

Rk	Player	Nation	Pos	Squad	np:G-xG
5	Diogo Jota	POR	FW	Liverpool	-0.3

Underperformers

Goals - xG
(excluding penalties)

Rk	Player	Nation	Pos	Squad	np:G-xG ▲
1	Francisco Trincão	POR	FW,MF	Wolves	-2.3
2	Nicolas Pépé	CIV	FW,MF	Arsenal	-2.1
3	Bryan Mbeumo	FRA	FW	Brentford	-2.0

After doing some investigating...

→ **Poor finisher**

→ **Poor finisher**

→ **Unlucky**

Source: Fbref.com (statsbomb)

Dataset: Data Source

1) Open data source from StatsBomb (www.statsbomb.com)

2) Data (2003-2020)

- Shots from 39 unique [competition, season]: La Liga 20/21, Women's World Cup 2019, Mens UEFA Euro 2020, Champions League 2019 Final
- All datasets in the statsbomb open source data except 1999/2000 Men's Champions League

	competition_id	season_id	country_name	competition_name	competition_gender	competition_youth	competition_international	season_name	match_upd
0	16	4	Europe	Champions League	male	False	False	2018/2019	2021-06-27T11:26:39.80
1	16	1	Europe	Champions League	male	False	False	2017/2018	2021-06-27T11:26:39.80
2	16	2	Europe	Champions League	male	False	False	2016/2017	2021-06-27T11:26:39.80
3	16	27	Europe	Champions League	male	False	False	2015/2016	2021-06-27T11:26:39.80
4	16	26	Europe	Champions League	male	False	False	2014/2015	2021-06-27T11:26:39.80
5	16	25	Europe	Champions League	male	False	False	2013/2014	2021-06-27T11:26:39.80
6	16	24	Europe	Champions League	male	False	False	2012/2013	2021-06-27T11:26:39.80
7	16	23	Europe	Champions League	male	False	False	2011/2012	2021-06-27T11:26:39.80
8	16	22	Europe	Champions League	male	False	False	2010/2011	2021-06-22T21:17:46
9	16	21	Europe	Champions League	male	False	False	2009/2010	2021-06-22T21:24:20
10	16	41	Europe	Champions League	male	False	False	2008/2009	2021-06-07T14:20:01.69
11	16	39	Europe	Champions League	male	False	False	2006/2007	2021-06-31T04:18:30.43
12	16	37	Europe	Champions League	male	False	False	2004/2005	2021-06-01T06:18:57.45
13	16	44	Europe	Champions League	male	False	False	2003/2004	2021-06-01T00:34:59.47
14	16	76	Europe	Champions League	male	False	False	1999/2000	2020-07-29T0
15	37	90	England	FA Women's Super League	female	False	False	2020/2021	2021-06-01T18:14:40
16	37	42	England	FA Women's Super League	female	False	False	2019/2020	2021-06-01T13:01:18

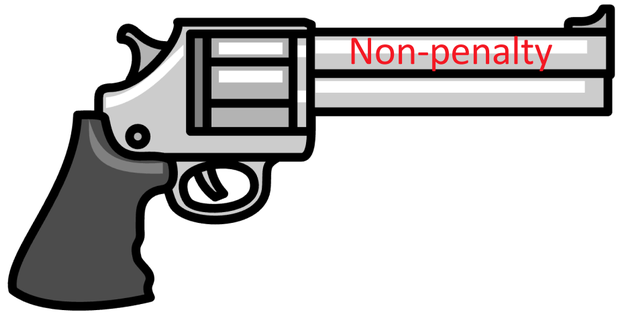
18	43	3	International	FIFA World Cup	male	False	True	2018	2021-06-05T16:04:30
19	11	90	Spain	La Liga	male	False	False	2020/2021	2021-06-24T19:17:07.83
20	11	42	Spain	La Liga	male	False	False	2019/2020	2021-06-15T15:35:02
21	11	4	Spain	La Liga	male	False	False	2018/2019	2021-06-02T17:53:14.52
22	11	1	Spain	La Liga	male	False	False	2017/2018	2021-06-27T11:26:39.80
23	11	2	Spain	La Liga	male	False	False	2016/2017	2021-06-07T22:30:18
24	11	27	Spain	La Liga	male	False	False	2015/2016	2020-07-29T0
25	11	26	Spain	La Liga	male	False	False	2014/2015	2020-07-29T0
26	11	25	Spain	La Liga	male	False	False	2013/2014	2020-07-29T0
27	11	24	Spain	La Liga	male	False	False	2012/2013	2021-06-27T15:44:43.94
28	11	23	Spain	La Liga	male	False	False	2011/2012	2020-07-29T0
29	11	22	Spain	La Liga	male	False	False	2010/2011	2021-06-11T22:57:42.36
30	11	21	Spain	La Liga	male	False	False	2009/2010	2021-06-26T13:56:40.98
31	11	41	Spain	La Liga	male	False	False	2008/2009	2020-07-29T0
32	11	40	Spain	La Liga	male	False	False	2007/2008	2021-06-26T13:13:56.18
33	11	39	Spain	La Liga	male	False	False	2006/2007	2020-07-29T0
34	11	38	Spain	La Liga	male	False	False	2005/2006	2020-07-29T0
35	11	37	Spain	La Liga	male	False	False	2004/2005	2020-07-29T0
36	49	3	United States of America	NWSL	female	False	False	2018	2021-06T05:53:29.43
37	2	44	England	Premier League	male	False	False	2003/2004	2021-06-14T22:29:00.64
38	55	43	Europe	UEFA Euro	male	False	True	2020	2021-06-11T14:00:16.10

Dataset: Cleaning

1) Excluded penalty kicks (no-penalty expected goals)
   - npxG

Total: 27,289 non-penalty shots

	event_type	id	index	period	timestamp	minute	second	possession	possession_team	play_pattern	...	redirect	one_on_one	open_goal	def
1	shot	93adc671-0697-4ad0-8f30-6dea062f3b03	342	1	00:09:04.723	9	4	15	Tottenham Hotspur	Regular Play	...	NaN	NaN	NaN	
2	shot	e18a3a5e-b587-43f3-8dea-a25a14517c14	587	1	00:16:48.573	16	48	25	Liverpool	From Throw In	...	NaN	NaN	NaN	
3	shot	2439947c-b340-48a3-a8fe-b4e217136ccd	758	1	00:20:31.085	20	31	35	Liverpool	From Keeper	...	NaN	NaN	NaN	
4	shot	7c1825a8-ff22-4d87-9a58-b287f0f3104a	768	1	00:21:53.381	21	53	37	Liverpool	From Throw In	...	NaN	NaN	NaN	
5	shot	c678b67c-9658-4d7d-9a9b-d3db9a378829	1347	1	00:37:48.966	37	48	65	Liverpool	Regular Play	...	NaN	NaN	NaN	

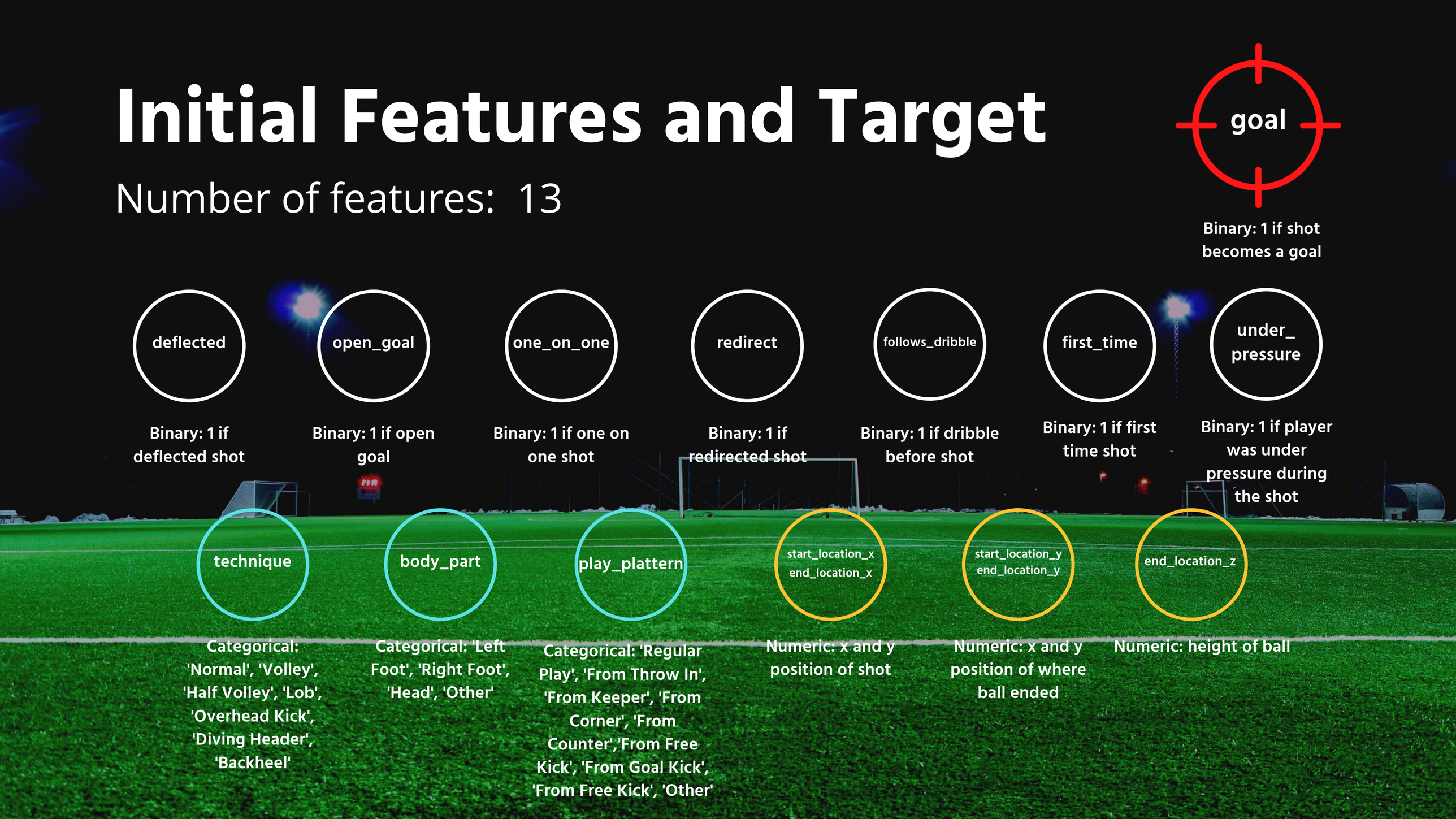


2) Create binary target variable of 'goal'
   where 'outcome' = 'Goal' for 1 and 0 otherwise

3) Create binary variables for features containing True/NaN
where True is 1 and NaN is 0

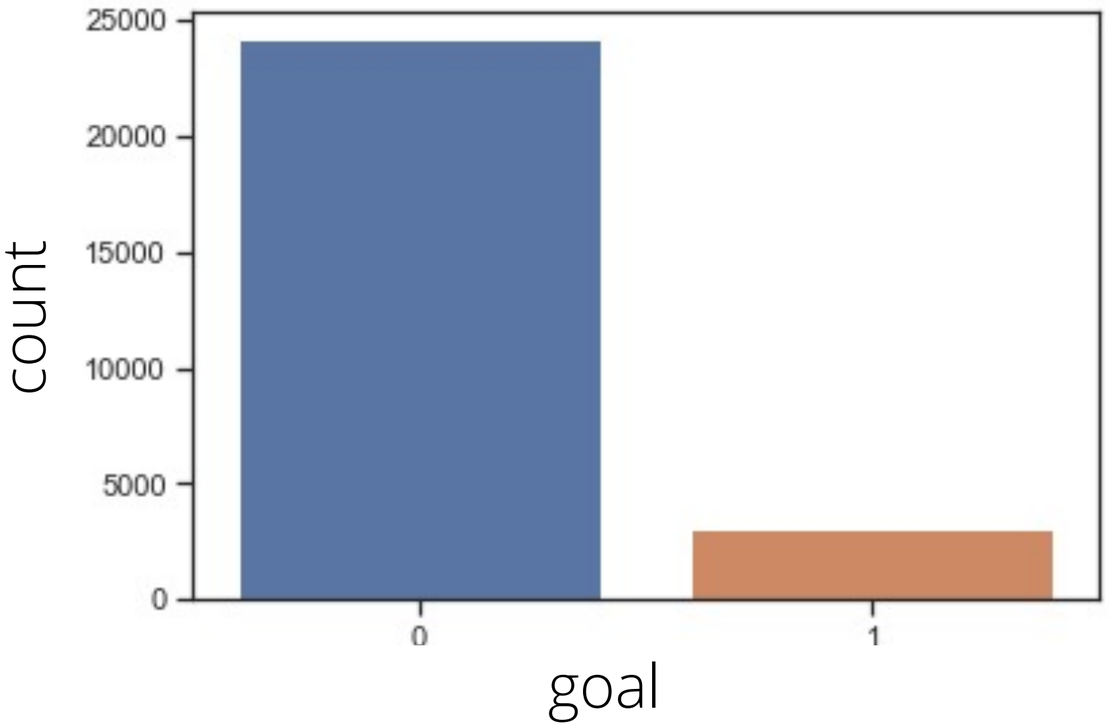
Initial Features and Target

Number of features: 13



Explanatory Analysis

Shots leading to Goal or No Goal

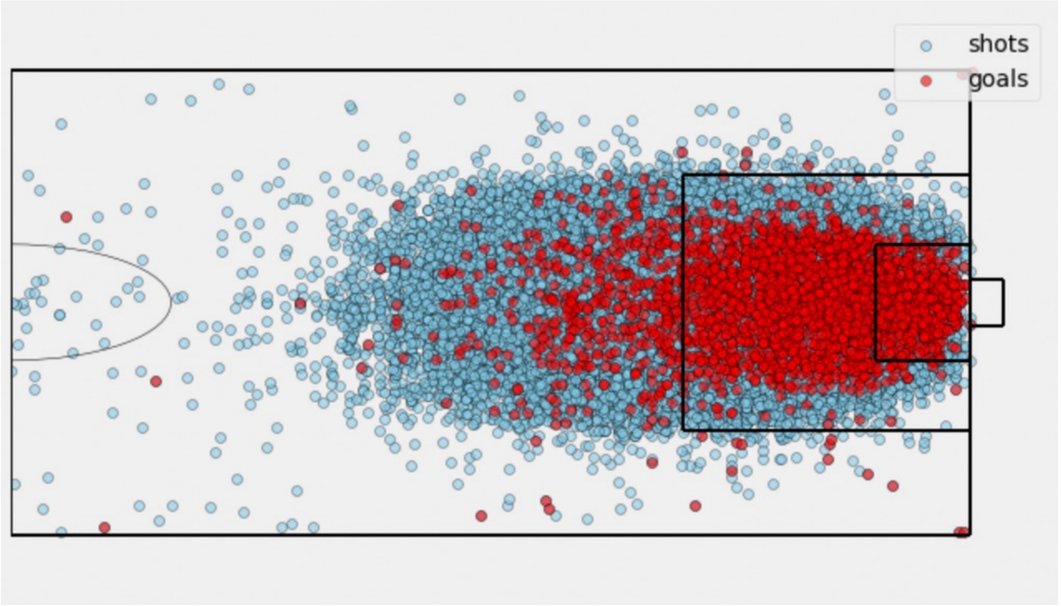


Out of 27,289 shots, only 11.43% (3,119) were goals

- Very low conversion rate

- Implies most shots don't lead to goals

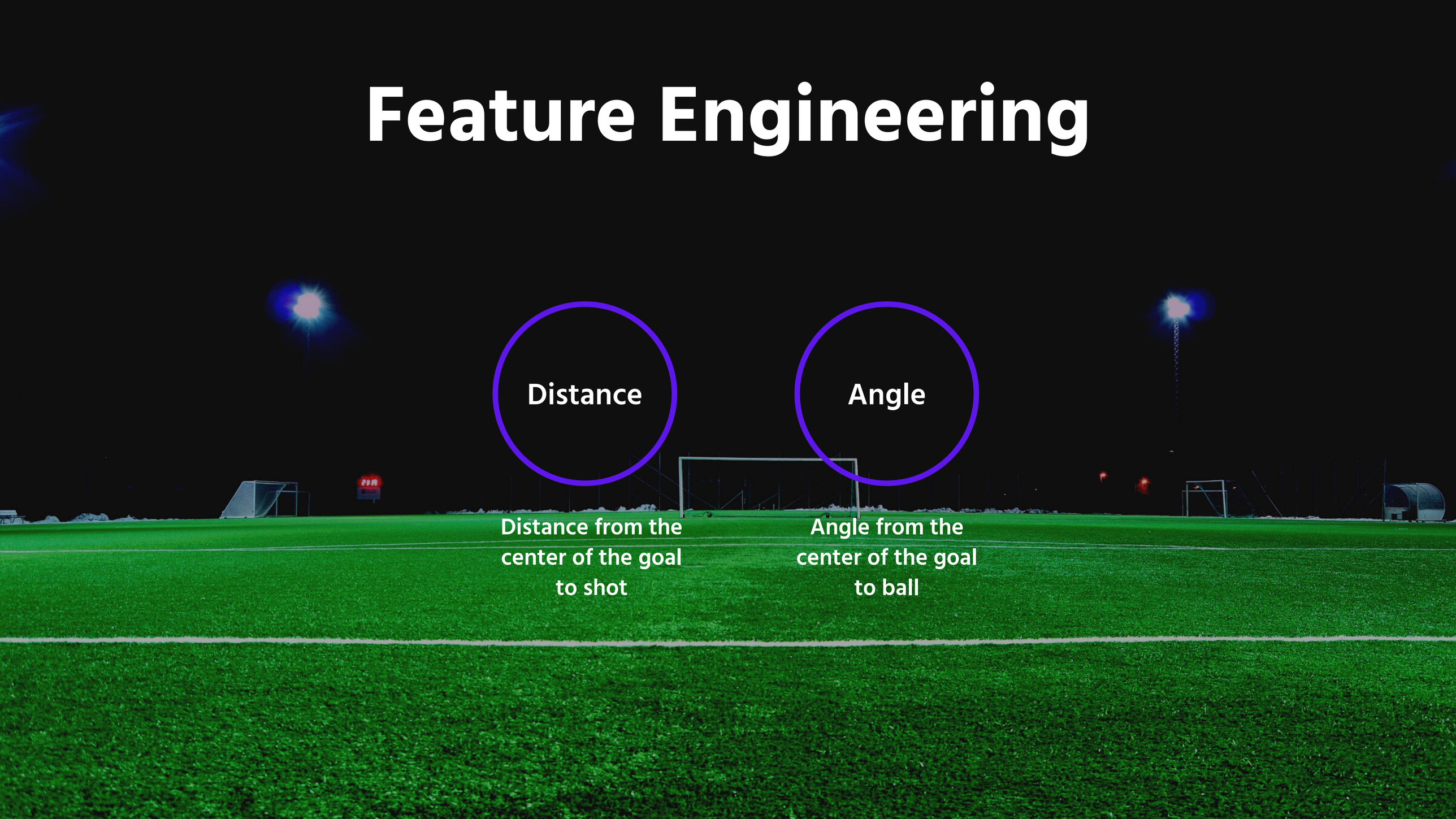
Shot Map



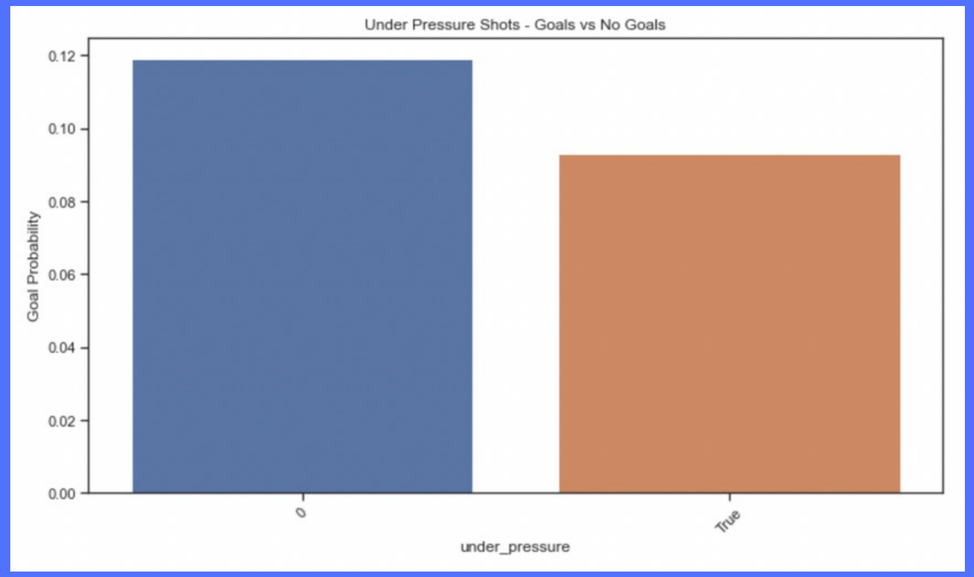
- Distance matters

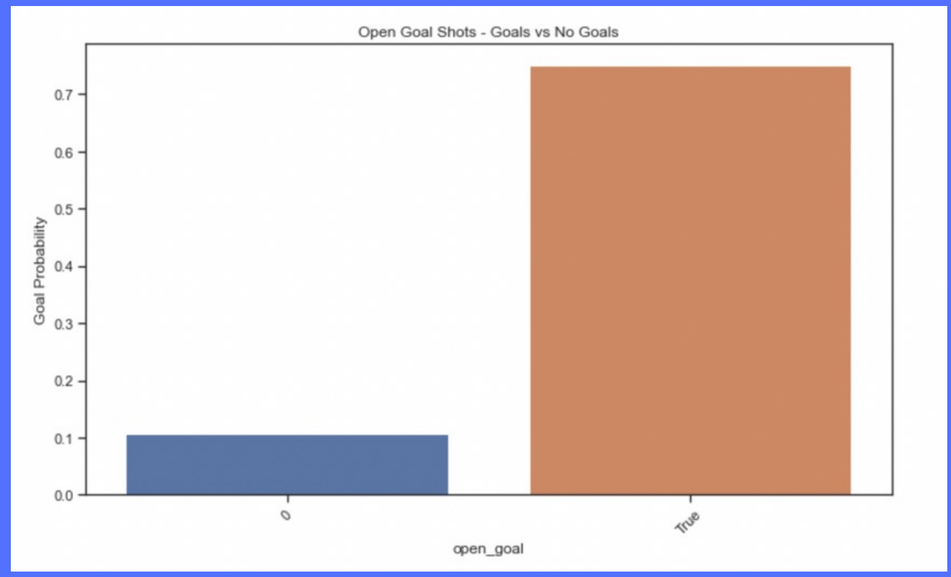
- Angle of the shot matters

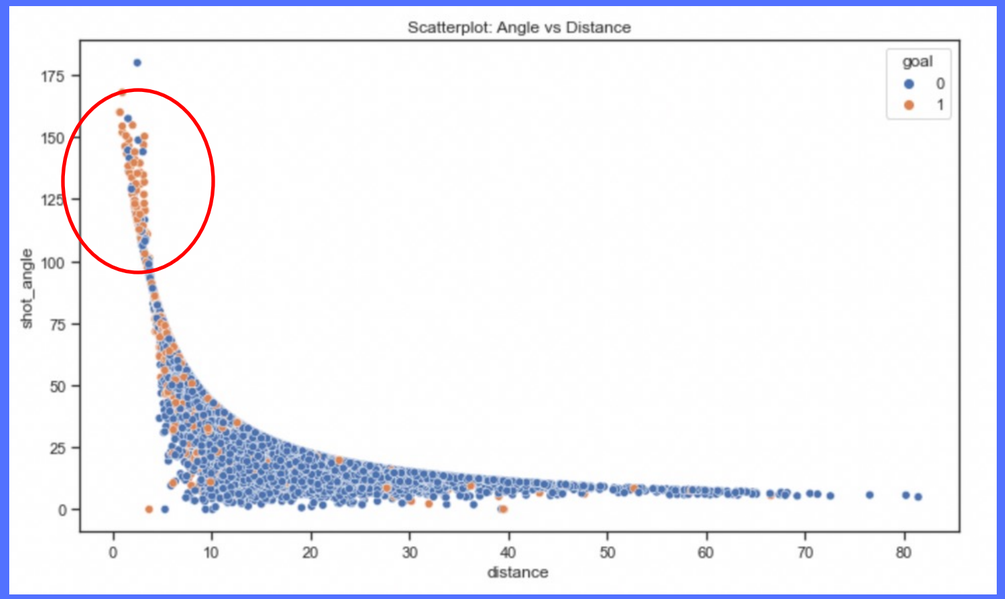
Feature Engineering



Explanatory Data Analysis

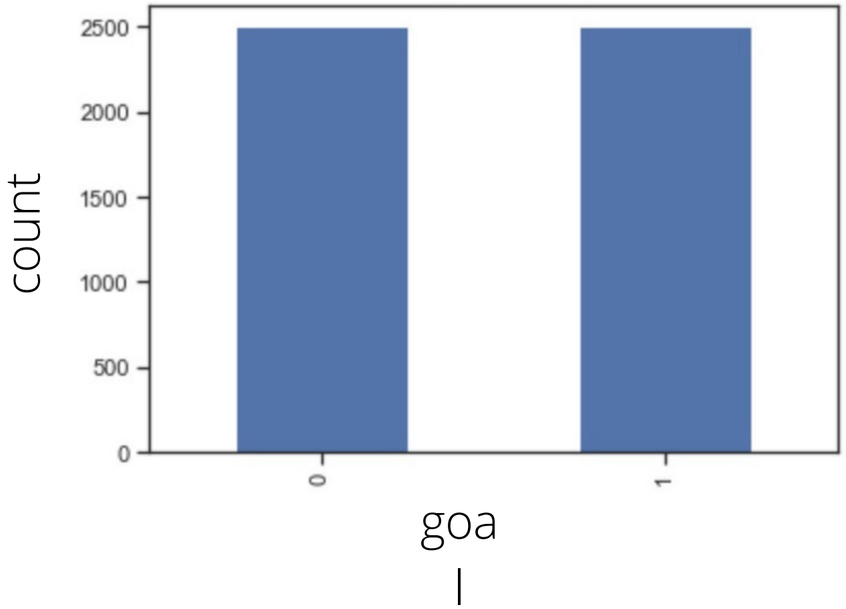






Pre-processing

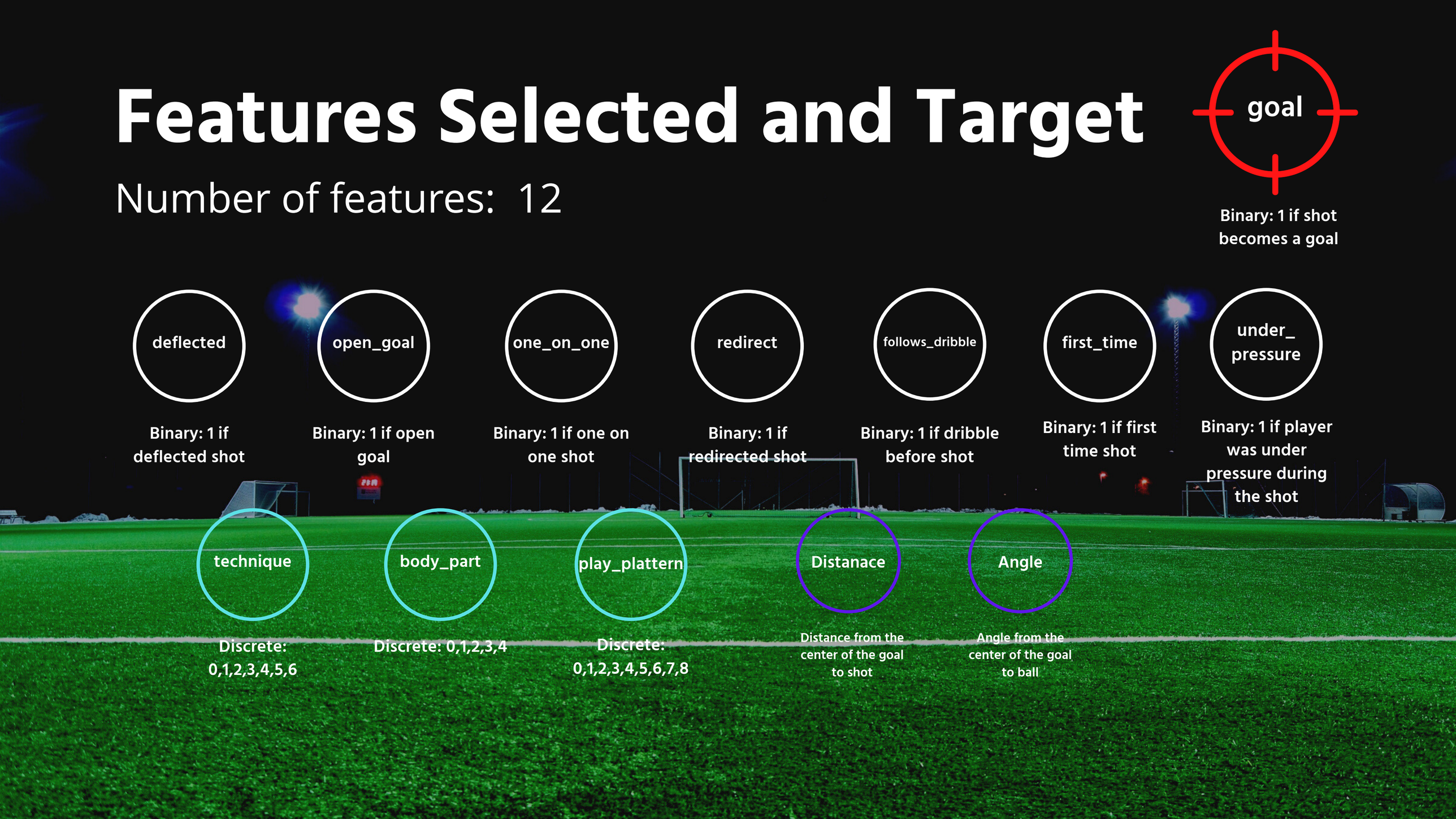
1) 1:1 split undersampling for better modeling



2) Label encoding - convert categorical features to discrete for model training

Features Selected and Target

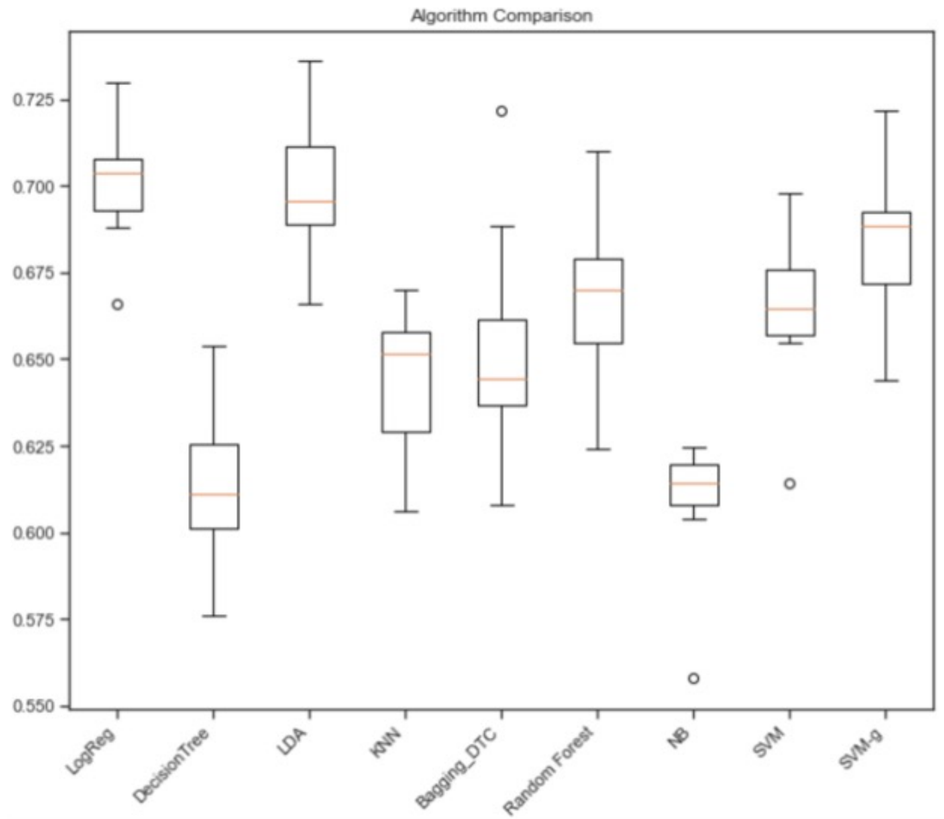
Number of features: 12



Model Selection

- Decision Tree Classifier
- Naïve Bayes
- Bagging
- K-nearest neighbour
- Random Forest
- Support Vector Machine
- Linear Discriminant Analysis
- Logistic Regression

Model Comparison



Model	Mean Accuracy	Accuracy after tuning
Decision Tree	0.613313	Not Done; Model not important
Naive Bayes GuassianNB	0.609306	Not Done; Model not important
Bagging Decision Tree	0.651475	0.684866
K-NearestNeighbor	0.643881	0.7107
Random Forest	0.668661	0.716196
SVM	0.664472	0.71583
SVM (gamma - auto)	0.683255	
Linear Discriminant Analysis	0.699243	0.716196
Logistic Regression	0.700838	0.72151

Stratified K-fold (k=10) & Cross_val_score used to obtain mean accuracy scores to compare different models to identify the most important model.

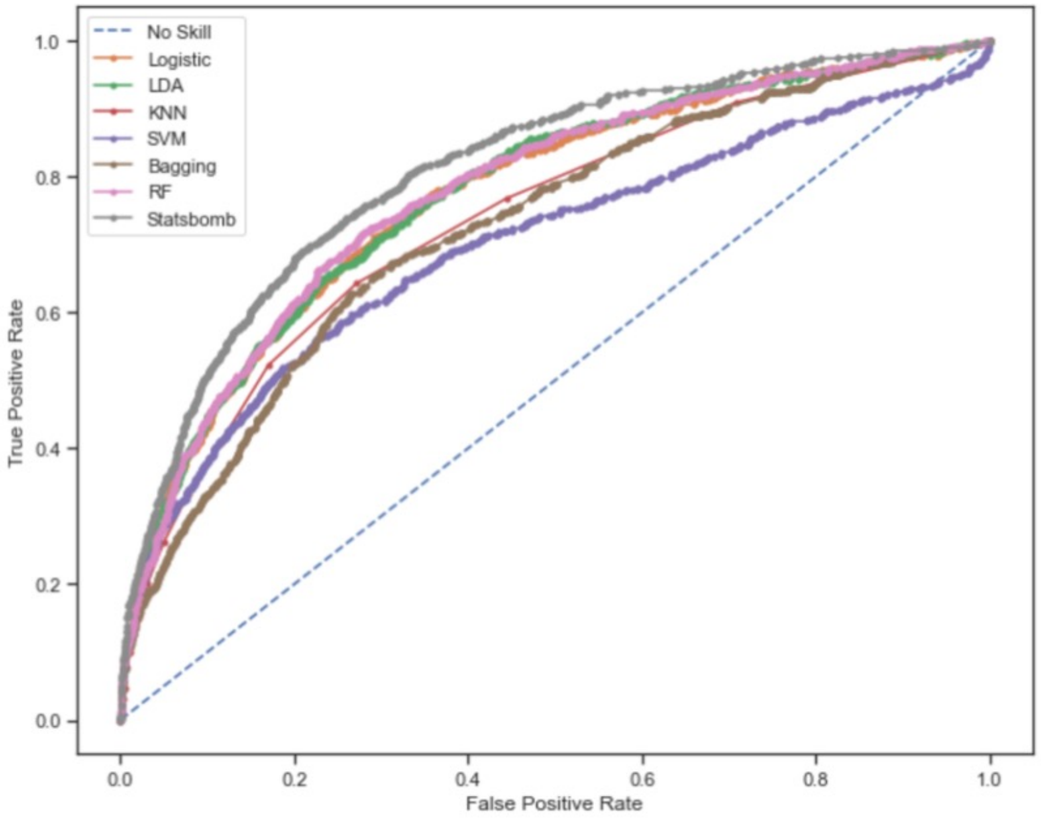
→Tuned hyperparameters and reproduced accuracy score for comparison

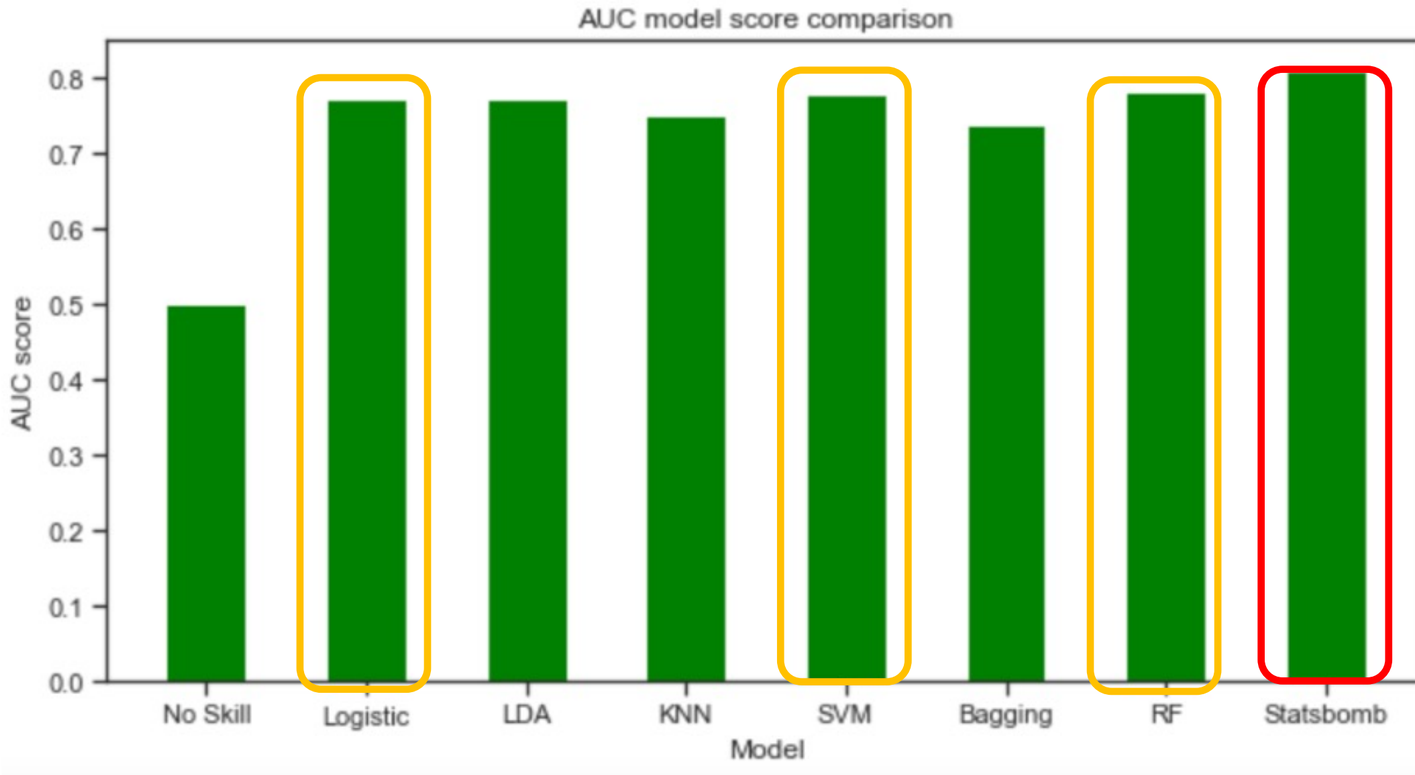
Model Evaluation

Score	Random Forests	SVM	Logistic Regression
Accuracy	0.716	0.716	0.722
Precision	0.24	0.24	0.24
Recall	0.72	0.72	0.69
F-1	0.36	0.36	0.36
AUC	0.782	0.78	0.774

(vs. Statsbomb xG: AUC = 0.810)

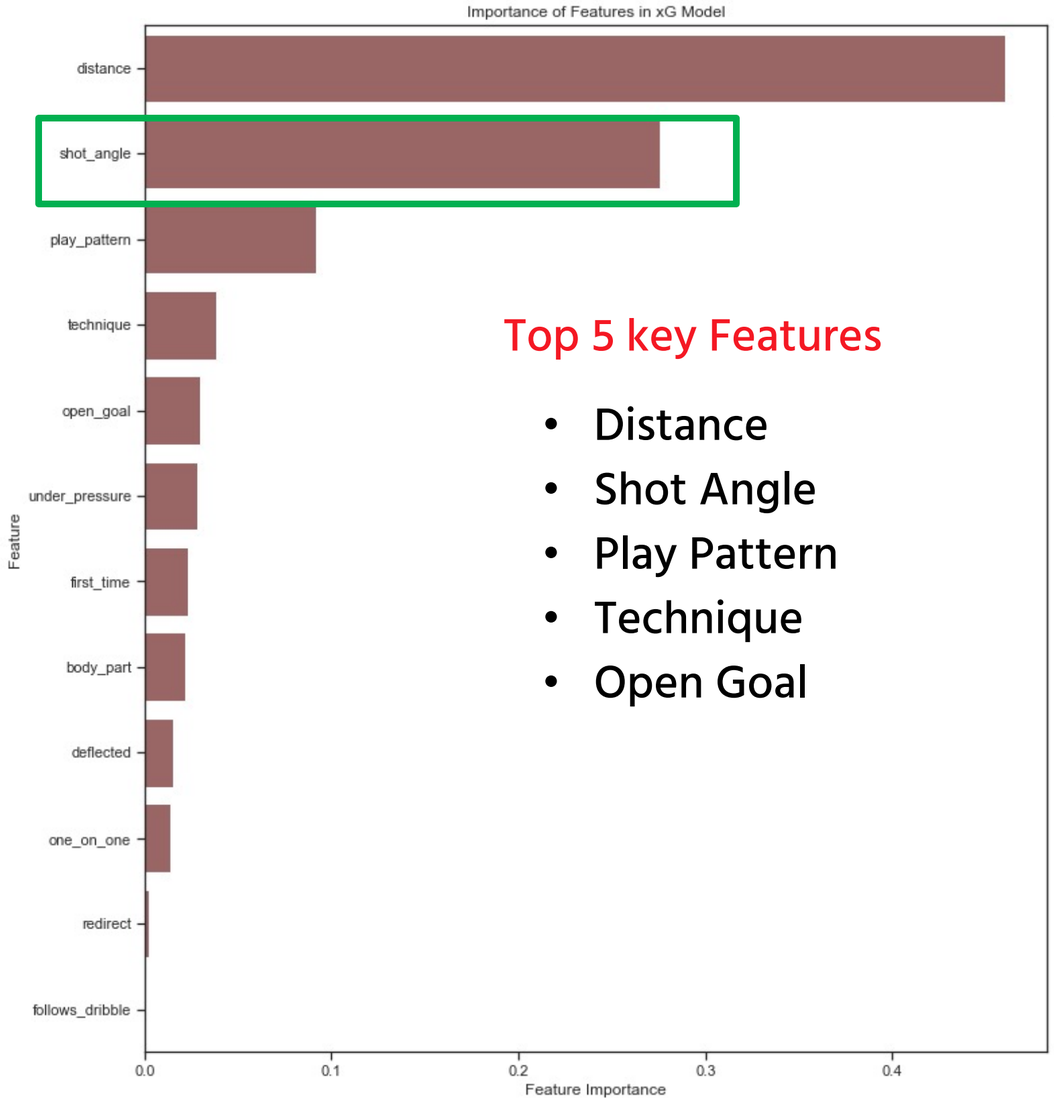
Model Evaluation (ROC Curve & AUC)





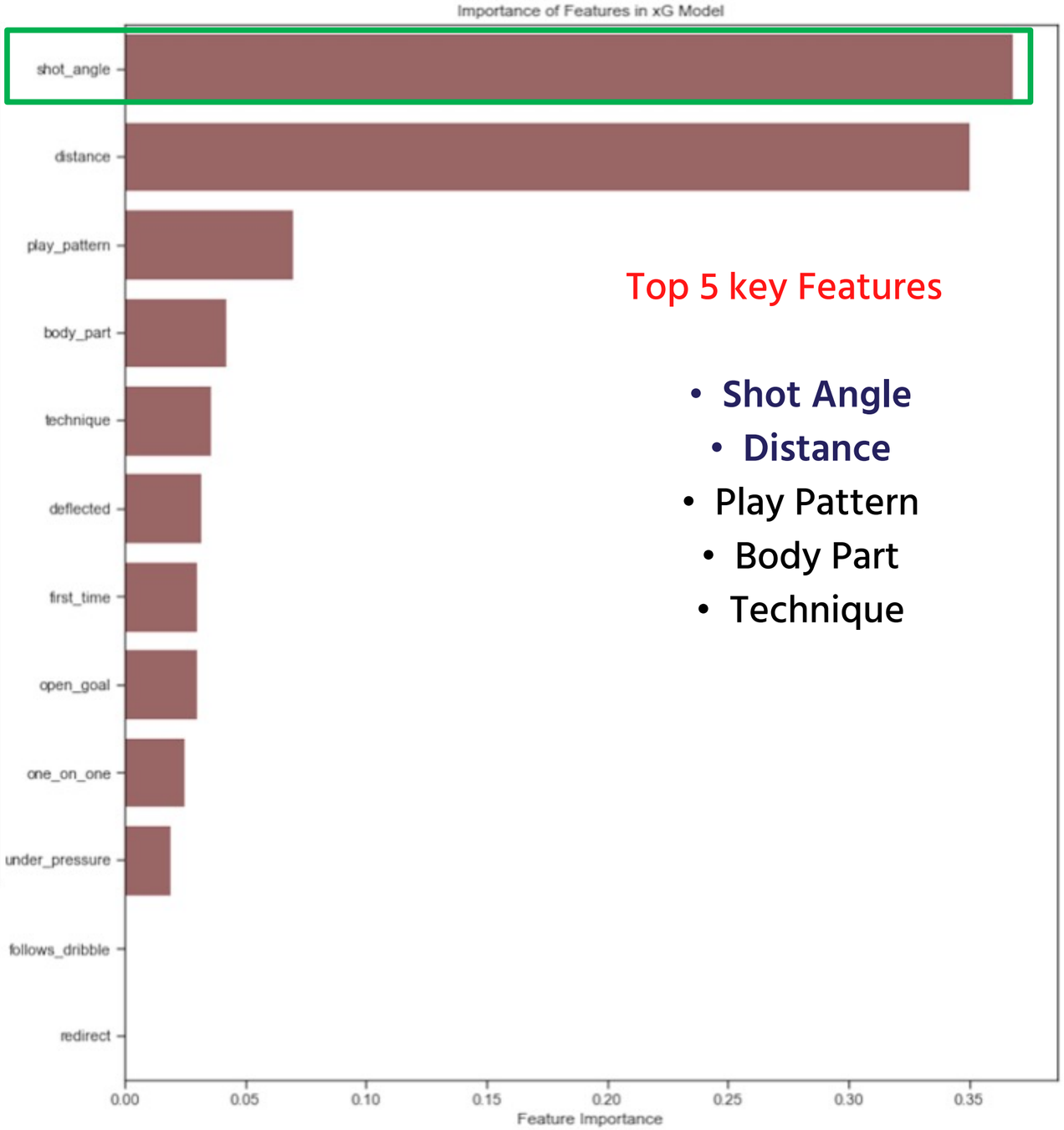
ROC curve for Random Forest (Pink) is the closest to StatsBomb's.
(Largest area under the curve)

Radom Forest Model has best performance with the highest AUC
(vs. StatsBomb Model: -3.5%)



Top 5 key Features

- Distance
- Shot Angle
- Play Pattern
- Technique
- Open Goal



Top 5 key Features

- **Shot Angle**
- **Distance**
- Play Pattern
- Body Part
- Technique

Business Application - Past

VS

Shot attempted

On target

A GUIDE TO

xG Premium By Databall

STEP 1

Collect player performance metrics from team scouts / coaches

STEP 2

Pass values through Databall's Model

STEP 3

Choose Model

STEP 4

Get predicted values to aid you in training players

STEP 5

Use Data to overcome obstacles & contribute to database

Unsure? Feel free to contact the Databall Team!
We're available until September 2023 :)

Conclusion

Open Source! No black box, No secretive models; Collaborative

More you use -> More Contributions -> Model Improves

Similar predictability with xG of World's Leading Analytics Company (Within 5%) !
- Used by Top Flight clubs (EPL / Bundesliga)

The End