UNDERSTANDING SOCCER THROUGH DATA SCIENCE

CONTENT

- SPORTS ANALYTICS METRICS
- PROJECT GOAL
- DATA SPECIFICATION
- TECHNIQUES WE USED

GOALS: A RARE BEAUTY IN SOCCER

Goals are the most important events in soccer, but they are also the most infrequent.

In most leagues, there are only 2.5-3 goals per match.

Goal Difference evaluates teams based on the difference of goals a team scores and concedes.

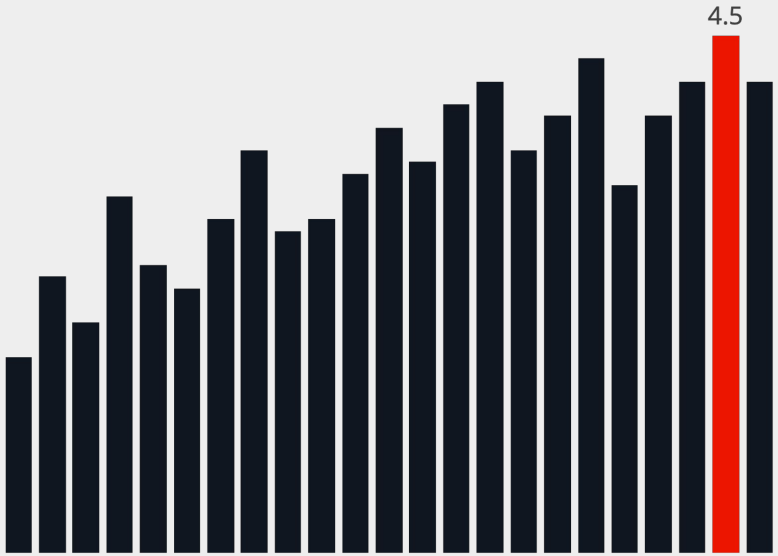
Randomness plays a huge part in game results.

NOT ALL SHOTS ARE CREATED EQUAL

- Goals are a result of shots so it is interesting to see how shots impact the outcome of the game.
- There are on average 25-30 shots per match.
- That gave rise to metrics such as Total Shots Ratio (TSR) that measured team dominance by their share of the shots in a match.
- Interestingly there is no guarantee that a team having high TSR will win the game.

EXPECTED GOAL OR XG

- Expected Goals or XG is a widely used metrics among the sport analytics community to evaluate performances in soccer.
- XG at its core is the probability of whether a given shot results in a goal.
- It looks at the quality of the shots produced by teams and players to gauge performance and evaluate them.



WHY IS XG IMPORTANT ?

- At player level, XG models can help players to understand the relationship between quality chances and goals.
- At a team level, Expected Goals models are more predictive of future performance than both current goal difference and simple shot-count metrics such as Total Shots Ratio (TSR).
- xG models allow us to look beyond current results to get a better idea of the underlying quality of both teams and players.

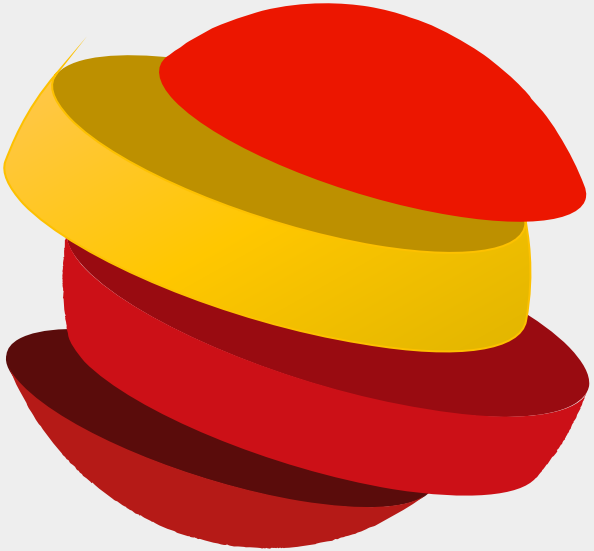
NOTE: XG CANNOT PREDICT RESULTS. BUT RATHER, IT IS A TOOL TO HELP OUR COMPREHENSION OF THE GAME

HOW ARE XG MODELS BUILT?

- Each XG model has its own characteristics. Some of the main factors are:
  - Shot distance to goal
  - Shot angle to goal
  - Body part with which the shot was taken
  - Type of assist or previous action
- Based on historical information of shots with similar characteristics, the xG model then attributes a value between 0 and 1 to each shot that expresses the probability of it producing a goal.

PROJECT GOALS

- The goal of our project is to build an xG model that evaluates the quality of shot based on various factors.
- Using the XG model we will develop probability rings to understand scoring chances from different locations on a soccer pitch.

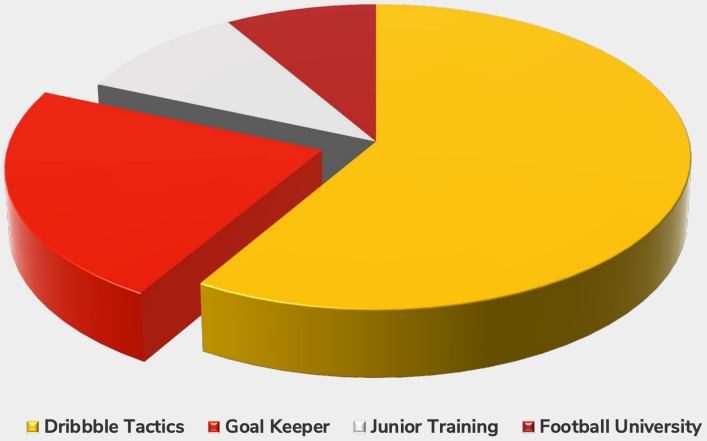


STATSBOMB OPEN DATA

Statsbomb is a leading data provider that provides different kind of soccer data.

Statsbomb open data is provided as JSON files on Statsbomb's github.

Event Data: Records every action on the ball. Actions like, passes, shots, fouls, etc.



Project Dataset

- From the various competitions available we have extracted all matches from La-Liga (17 Seasons)
- From each match we have selected shots data.
- An approx of 12000 shot will be analyzed and used to build a XG model

```
*******La-Liga Analysis******
Total Season: 17
Total Matches:  520
Total Shots:  12838
Total Goal:  1756
```

SHOTS DATA SPECIFICATIONS

- Each row represents a single shot.
- Features for each shot:

  ['id', 'index', 'period', 'timestamp', 'minute', 'second', 'type', ‘possession', 'possession_team', 'play_pattern', 'team', 'player', 'position', 'location', 'duration', 'related_events', 'match_id', 'shot_end_location', 'shot_key_pass_id', 'shot_outcome', 'shot_type', 'shot_body_part', 'shot_technique', 'shot_freeze_frame', 'possession_team_id', 'player_id', 'under_pressure', 'shot_first_time', 'shot_one_on_one', 'shot_aerial_won', 'shot_redirect', 'out', 'shot_deflected', 'shot_open_goal']

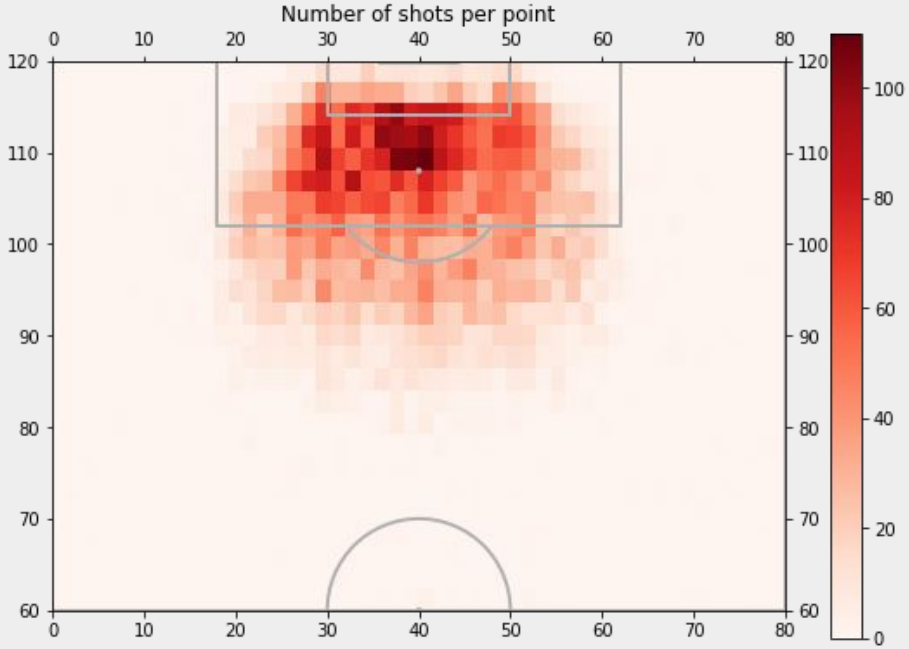
- Target: shot_outcome

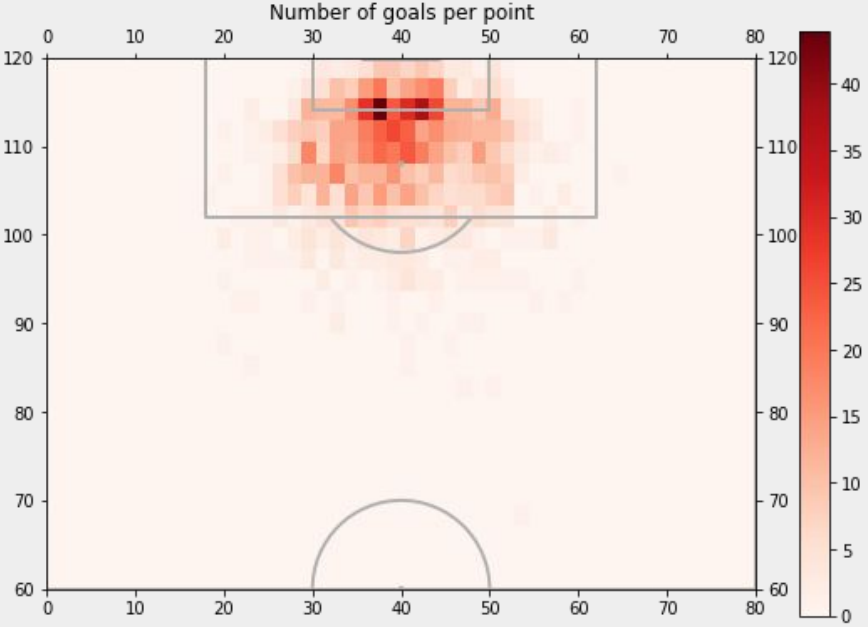
  shot_statsbomb_xg: Statsbomb provides its own XG to every shot. We will use this to validate our model.

Initial Dataset Snippet

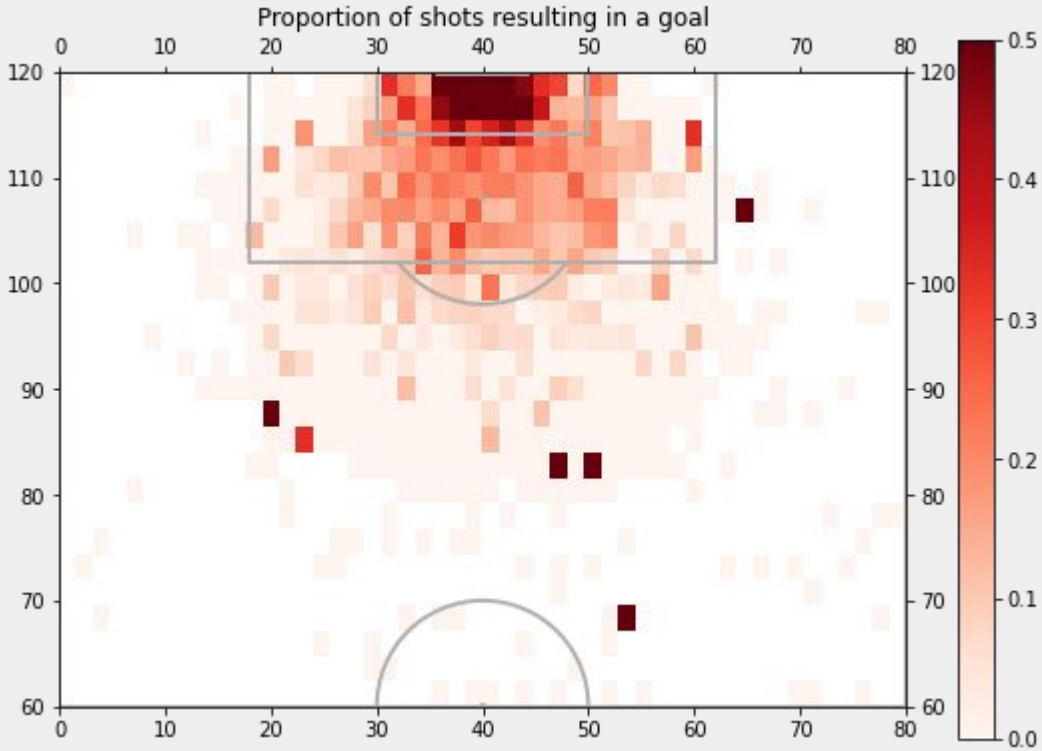
id	index	period	timestamp	minute	second	type	possession	possession_team	team	player	position	location	shot_outcome	shot_type	shot_body_p
bf67c762-daf8-4f83-ae23-fee462132626	1722	1	00:37:48.928	37	48	Shot	65	Real Betis	Real Betis	Borja Iglesias Quintas	Center Forward	[111.7, 42.4]	Goal	Open Play	Left
fa4eeeb9-e077-4dc7-ad71-e54e476c32c3	2682	2	00:13:55.572	58	55	Shot	107	Barcelona	Barcelona	Lionel Andrés Messi Cuccittini	Center Attacking Midfield	[102.5, 52.6]	Goal	Open Play	Left
084f8cb9-c2a4-47d7-b5c2-9c672724b4d5	3312	2	00:29:05.425	74	5	Shot	140	Real Betis	Real Betis	Víctor Ruíz Torre	Left Center Back	[116.2, 42.2]	Goal	Open Play	H
2ac88bc6-9d87-4ac5-9030-83dd624e5dea	3749	2	00:41:20.155	86	20	Shot	160	Barcelona	Barcelona	Francisco António Machado Mota de Castro Trincão	Left Wing	[102.5, 49.9]	Goal	Open Play	Left
b53158b9-64b2-4715-a228-7a1697ebe1d9	1278	1	00:29:41.547	29	41	Shot	62	Barcelona	Barcelona	Jordi Alba Ramos	Left Wing Back	[113.1, 28.2]	Goal	Open Play	Left
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
e05a628b-87f5-49b1-980a-71e3f8d59929	2300	2	00:27:55.175	72	55	Shot	167	Albacete	Albacete	Mark Dennis González Hoffmann	Left Wing	[114.9, 37.3]	Goal	Open Play	H

La-Liga Analysis





Shots Shots Shots (La-Liga Analysis)



TECHNIQUES USED:

- Open Source:
  - Statsbombpy: A python library fetch data from statsbomb github, filters it based on competitions and builds a dataframe.
  - MPL Soccer: A python library to visualize soccer data
- Data Visualization:
  - Matplotlib
  - Seaborn
  - MPL Soccer
- Feature Selection and Creation
  - Domain Knowledge, XGBoost
- Machine Learning
  - Linear Regression
  - Logistic Regression
  - Deep Learning

THANK YOU