UNDERSTANDING SOCCER THROUGH DATA SCIENCE

Project by:

Atif Siddiqui

CONTENT

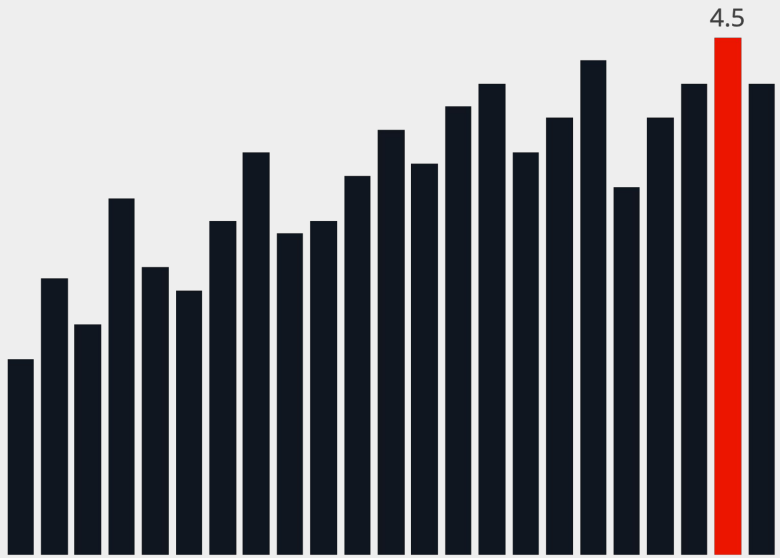
- PROJECT GOAL RECAP
- LA-LIGA ANALYSIS
- LINEAR REGRESSION
- LOGISTIC REGRESSION
- THINGS WE LEARNED

SPORTS ANALYTICS METRICS REVIEW

GOAL DIFFERENCE

TOTAL SHOT RATIO

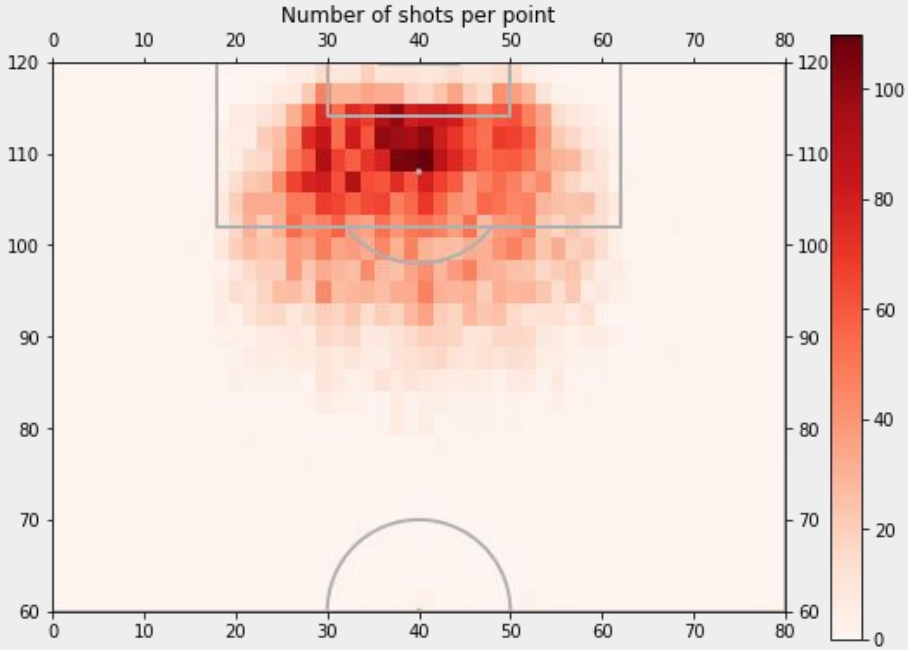
EXPECTED GOALS

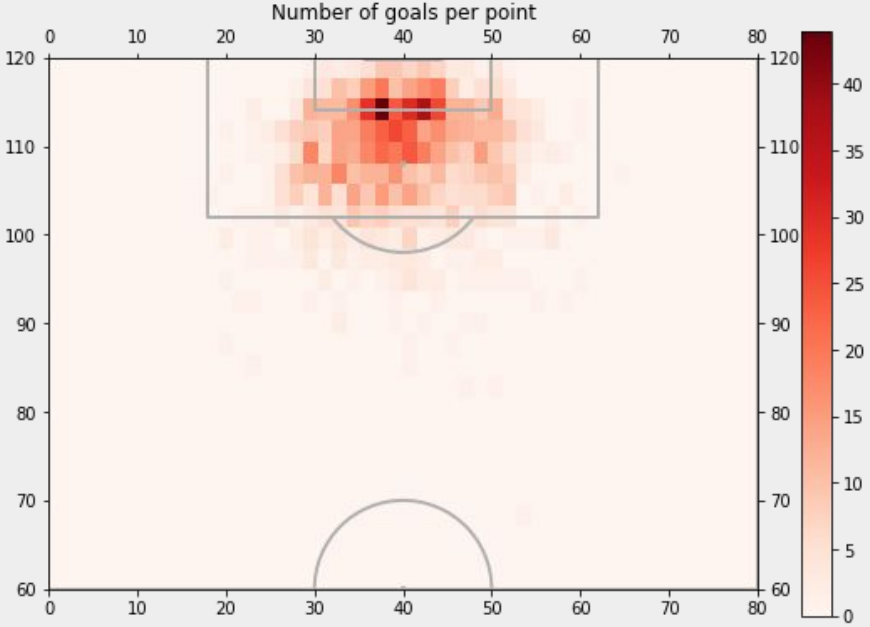


PROJECT GOALS

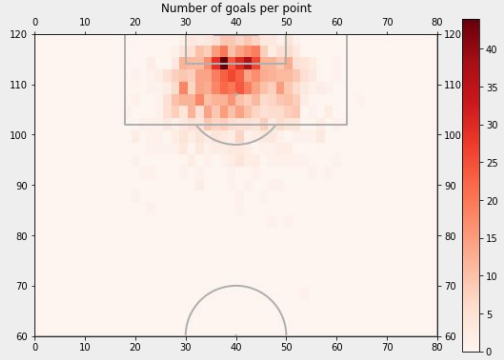
- The goal of our project is to build an xG model that evaluates the quality of shot based on various factors.
- Using the XG model we will develop probability rings to understand scoring chances from different locations on a soccer pitch.

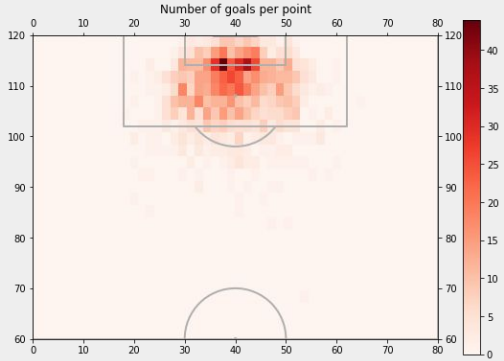
La-Liga Analysis

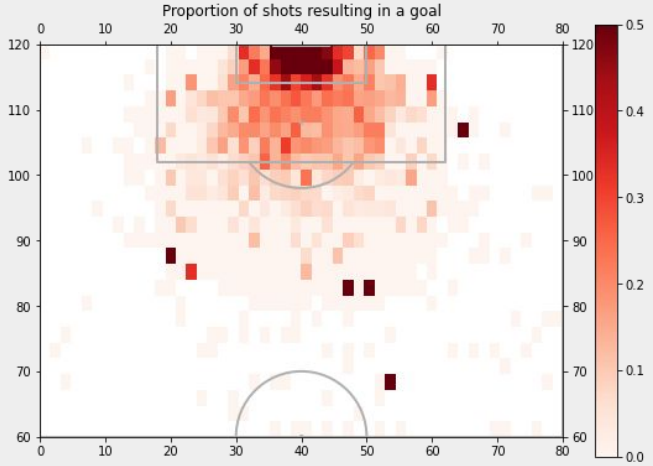




Frequency of Scoring

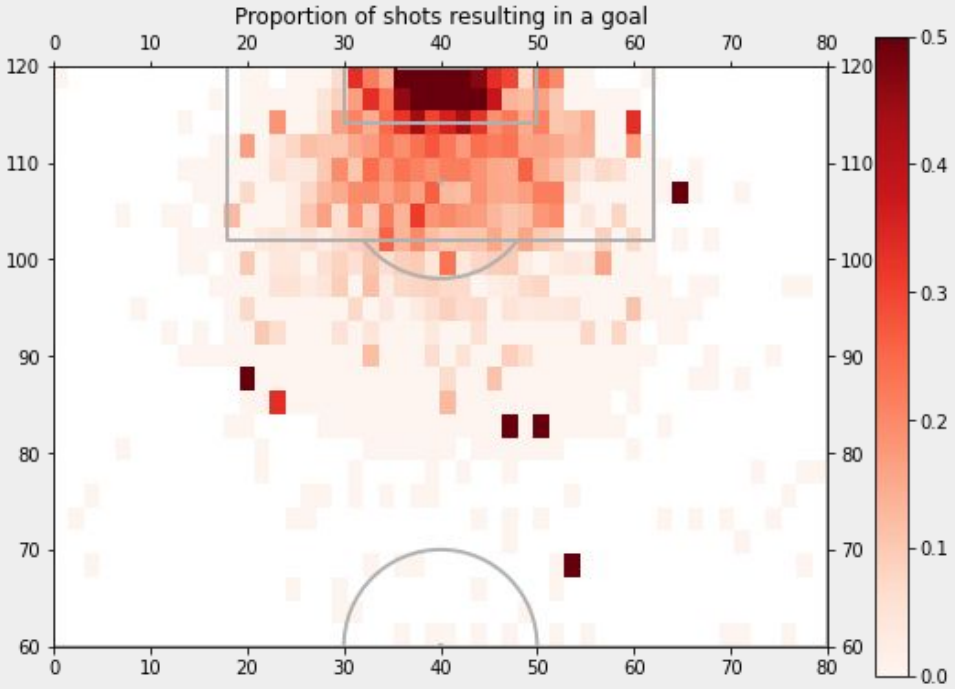




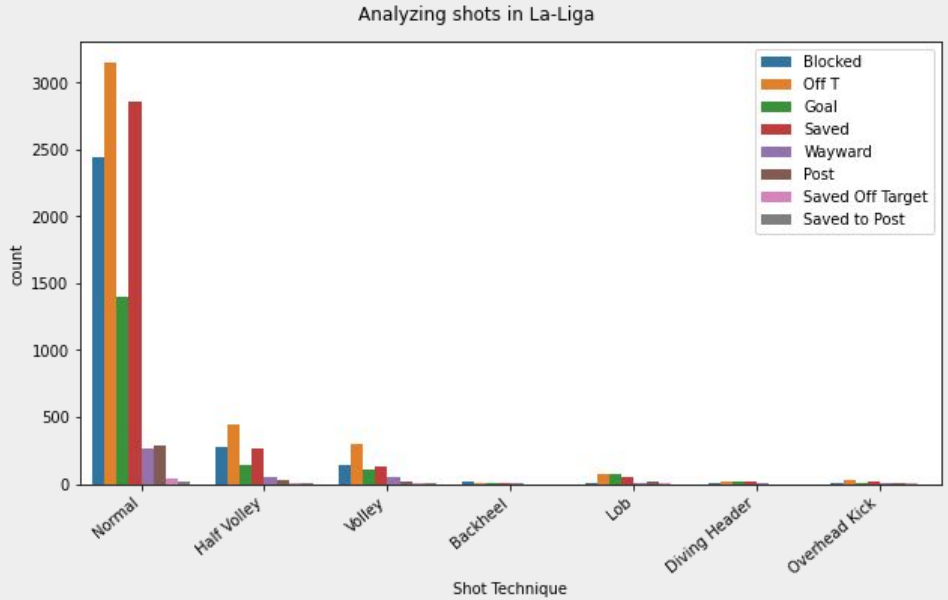


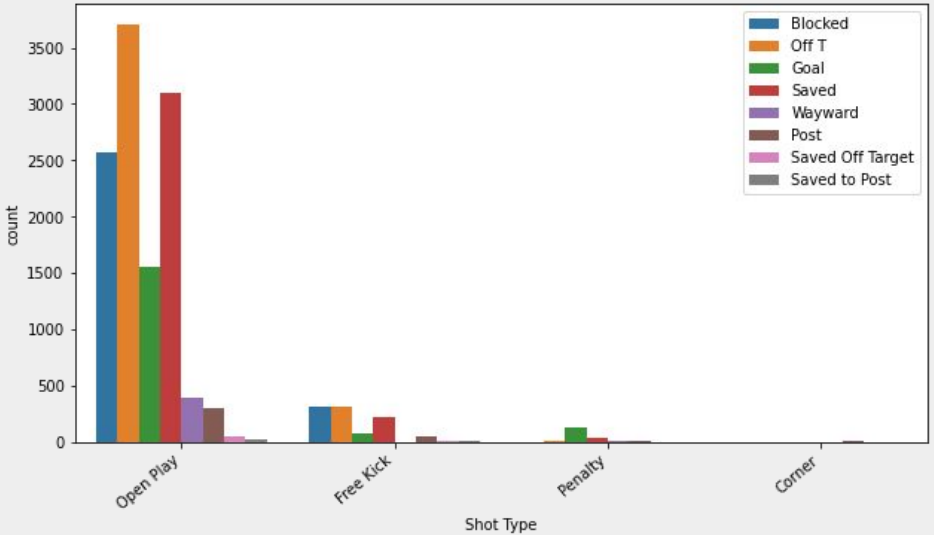
Thinking About XG

- No of shots resulting in a goal decreases as you go farther from the goal line.
- Outliers are points from where few shot were taken from but some of them resulted in a goal

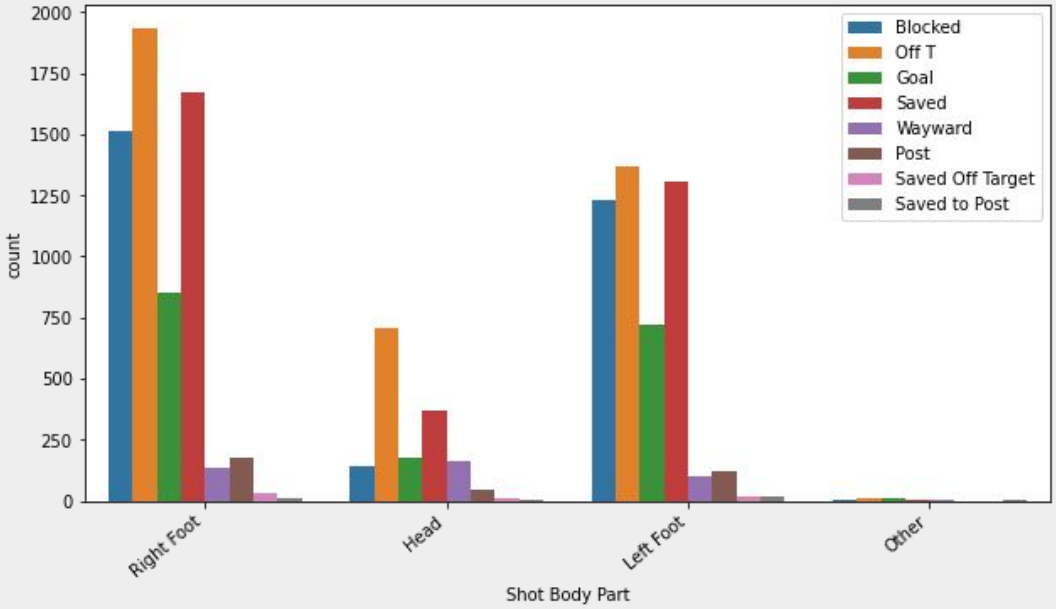


Shots Shots Shots (La-Liga Analysis)





Shots Shots Shots (La-Liga Analysis)

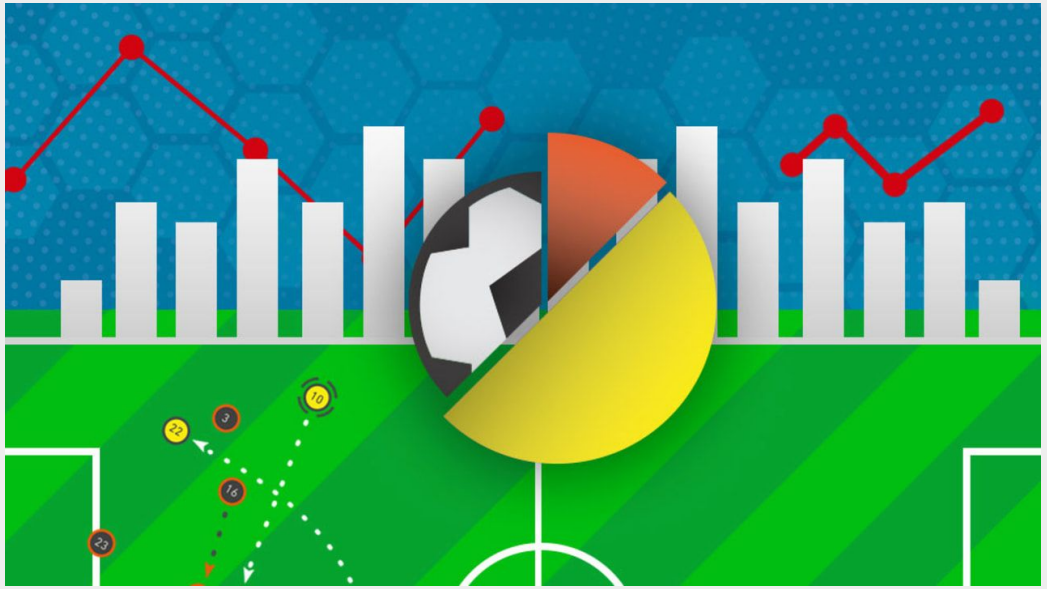


DATA PREPARATION

- Feature Creation
- Feature Selection
- Creating Testing and Training Dataframes

FEATURE CREATION

1. DISTANCE TO GOAL
2. ANGLE TO GOAL
3. PLAYER COUNT

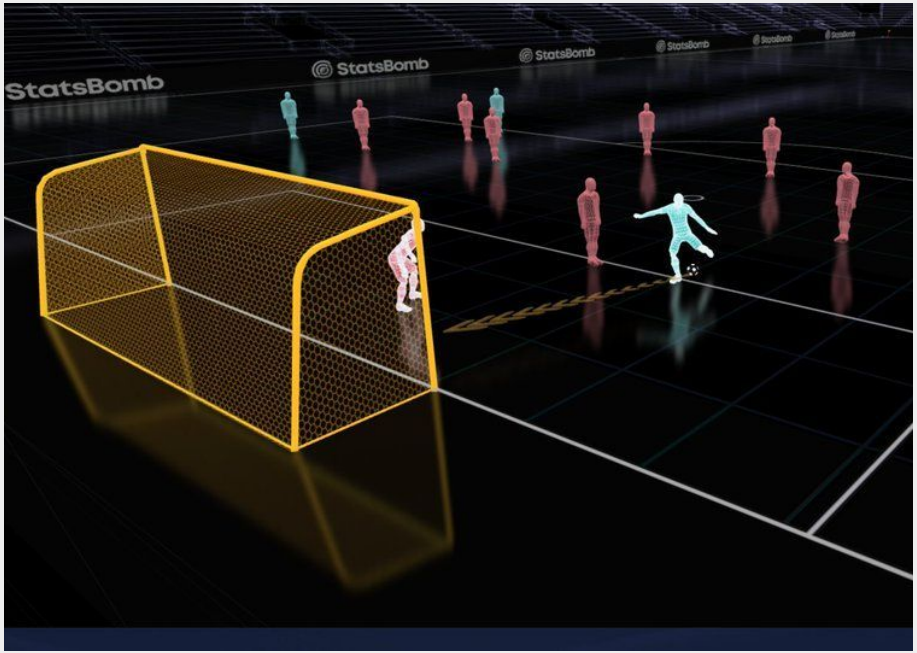


Distance from Goal

Calculated by using the formula

Distance = np.sqrt((x**2) + (y**2))

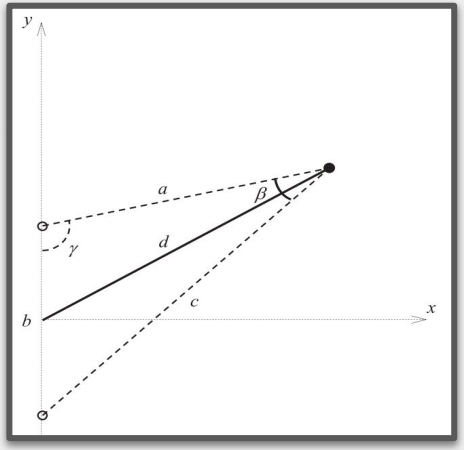
where x and y are the coordinates with respect to goal.



Angle from Goal

Calculated by using the formula

Angle = np.arctan((8.00*x)/(x**2+y**2-((8/2)**2)))

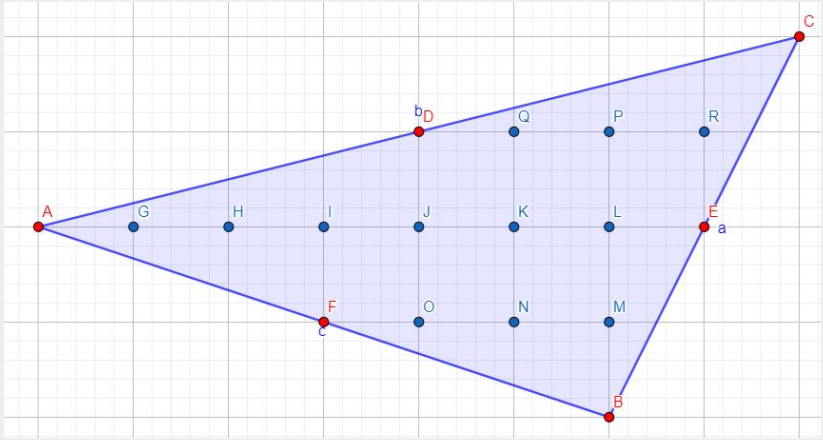


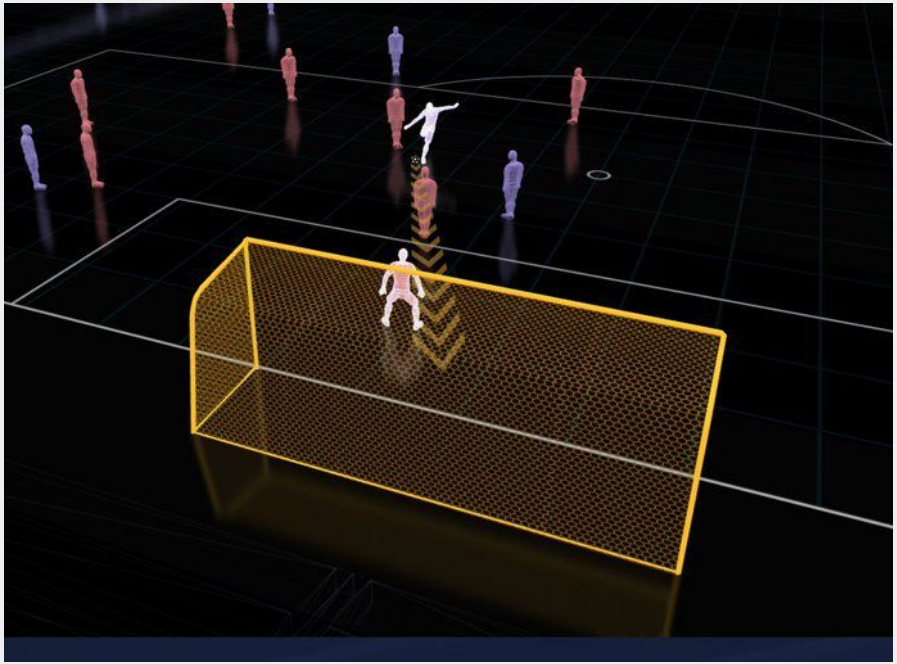
where x and y are the coordinated with respect to goal and,

César A. Morales (2016): A mathematics-based new penalty area in football: tackling diving, Journal of Sports Sciences, DOI: 10.1080/02640414.2016.1177657

Player Count

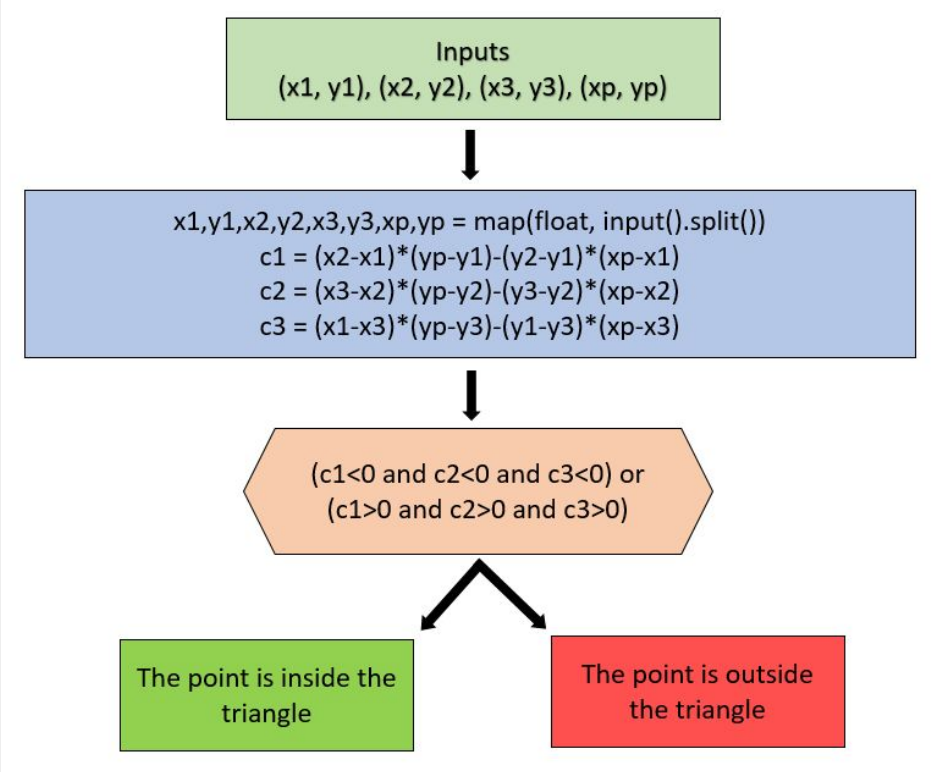
Idea behind calculating the number of players while taking the shot





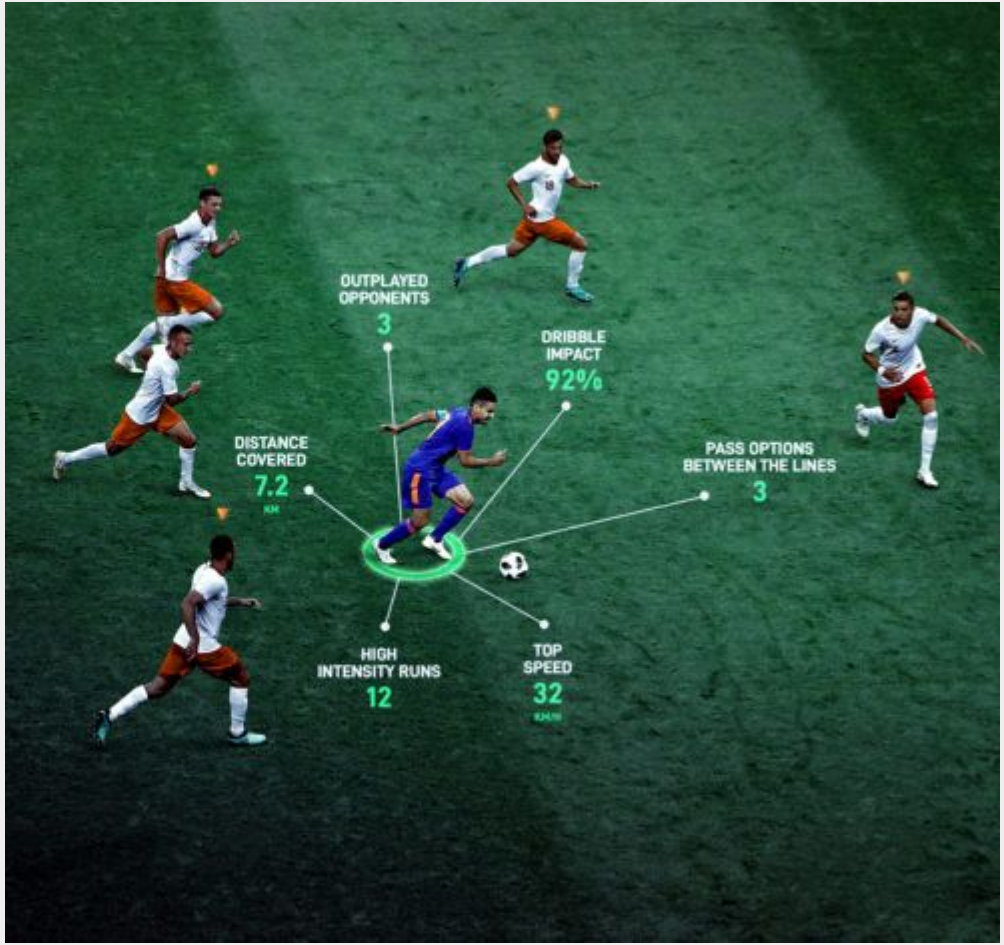
Player Count

Flowchart for calculating the number of players in front of the goal while the player is taking the shot.



FEATURE ENCODING

- SHOT OUTCOME
  - Goal : 1
  - Others: 0
- SHOT BODY PART
  - Left Foot : 1
  - Right Foot : 2
  - Head : 3
  - Other : 4
- SHOT TECHNIQUE
  - Normal : 1
  - Half Volley : 2
  - Volley : 3
  - Backheel : 4
  - Lob: 5
  - Diving Header: 6
  - Overhead Kick : 7



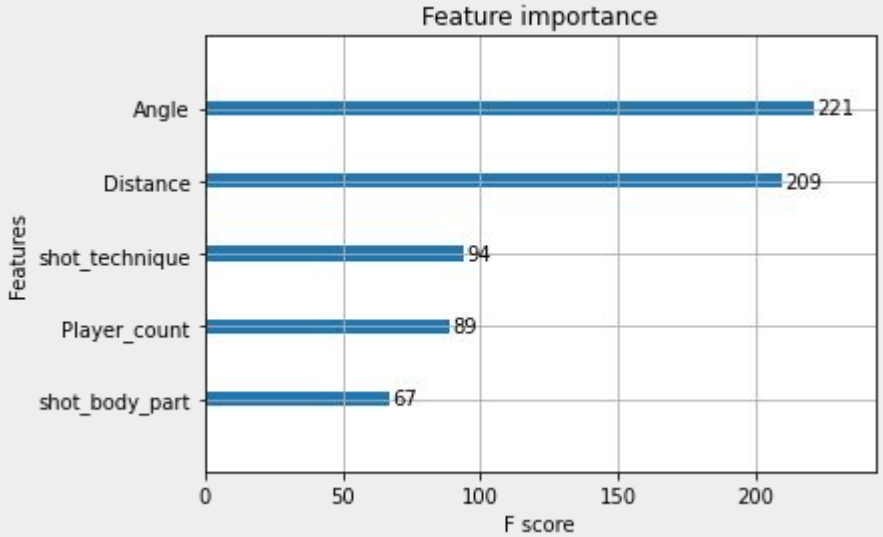
FEATURE SELECTION

Using our domain knowledge and by referring to various resources we have selected these features:

1. Angle to Goal
2. Distance to Goal
3. Shot Type
4. Shot Technique
5. Player Count
6. Shot Outcome (Target)
7. Statsbomb Shot XG (Validation)

Feature Importance

XGBOOST to look at feature importance.



TRAINING DATA

- Training data includes shots from 14 seasons of La-Liga.
- There are a total of 10500 shots

TESTING DATA

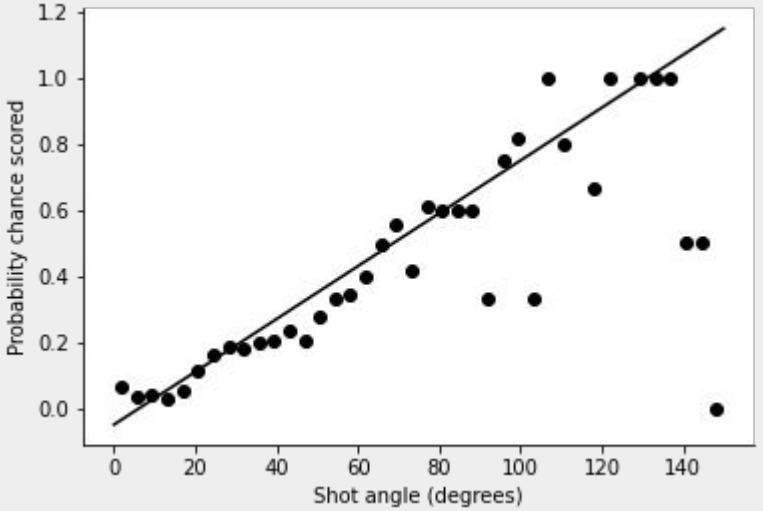
- Our testing consists of 2 different competitions.
- Shots from La-Liga Comp : 1278
- Shots from Champions League: 456

We will test our models on these 2 different datasets.

WHY LINEAR REGRESSION DOESN'T WORK

According to this plot there is a 120% chance of scoring from a 140° angle.

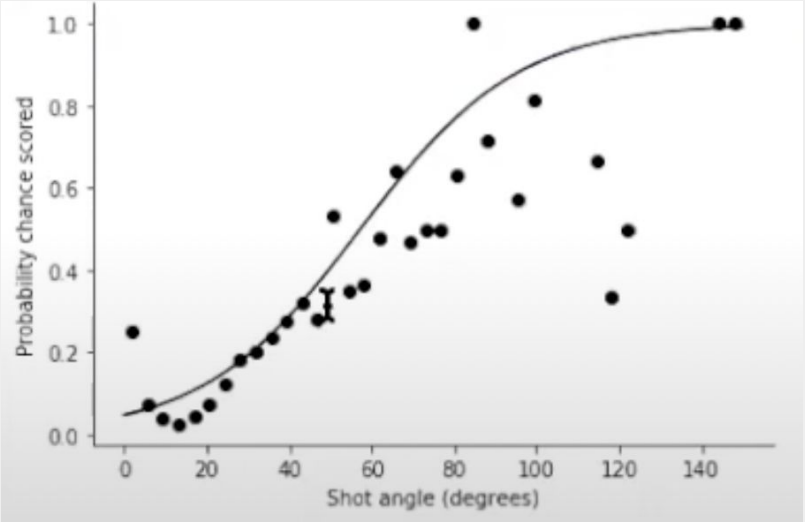
Realistically we can't have a model that predicts more than 100%



LOGISTIC REGRESSION

A Logistic function gives a reasonable fit.

We used Logistic regression to fit different kinds of data to understand what is the best way of predicting a goal.



MODEL BASED ON SHOT DISTANCE AND ANGLE

- Features: Shot Angle, Shot Distance
- Target: Shot Outcome
- Validation Feature: Statsbomb XG

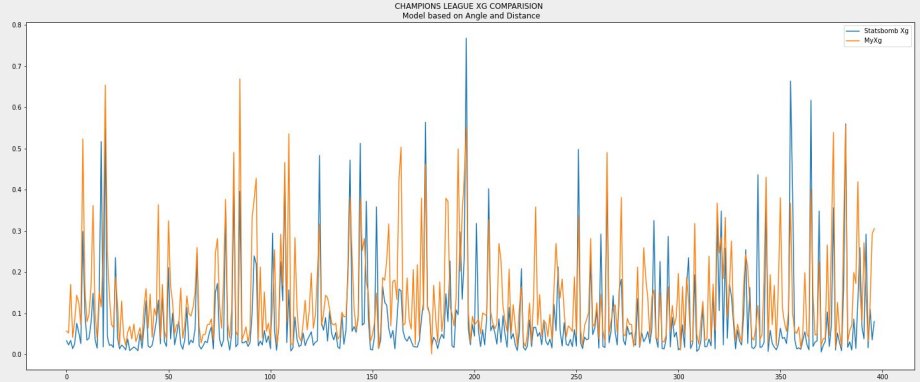
```
**********{GOAL SCORED VS [ANGLE, DISTANCE]**********
Model accuracy on champions league data:  0.8841309823677582
Model accuracy on La-liga test data:  0.8904719501335708
intercept: [-2.16904397]
coef: [[ 0.31546354 -0.72553462]]
```

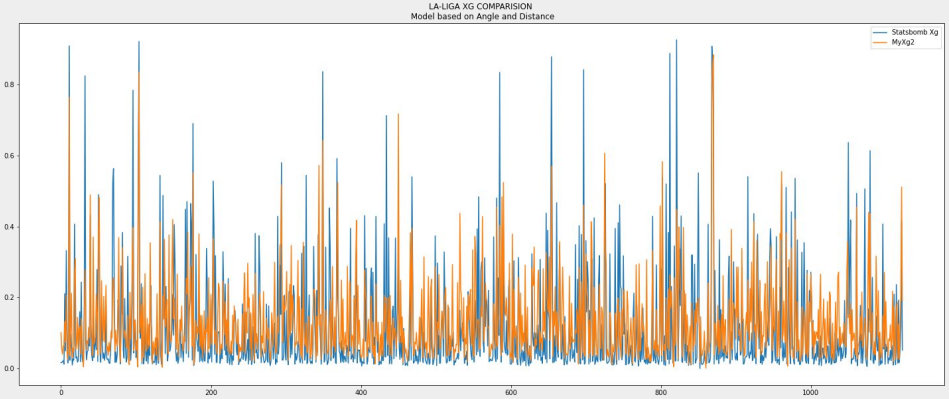
RESULTS AND VALIDATION FOR MODEL 1

- The probabilities we got from our model were plotted against the Statsbomb shot XG's of the testing data.
- Mean Squared Difference between predicted XG and the actual one on our two testing data sets.

```
Mean Squared Difference:  0.011336682445603652

Mean Squared Difference2:  0.011865534485769194
```





MODEL BASED ON SHOT DISTANCE, ANGLE AND BODY PART

- Features: Shot Angle, Shot Distance, Body Part
- Target: Shot Outcome
- Validation Feature: Statsbomb XG

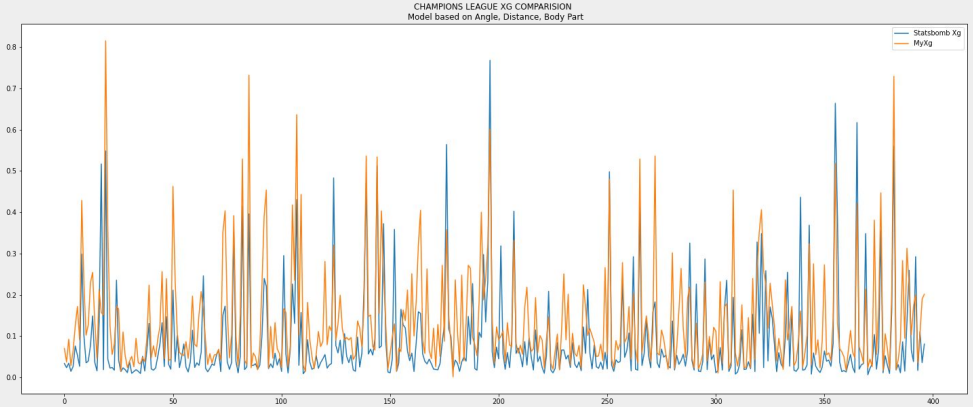
```
************{GOAL SCORED VS [ANGLE, DISTANCE, BODY PART]*************
Model accuracy on champions league data:  0.8967254408060453
Model accuracy on La-liga test data:  0.8833481745325023
intercept: [-1.2803758]
coef: [[ 0.38060027 -0.84356502 -0.56143071]]
```

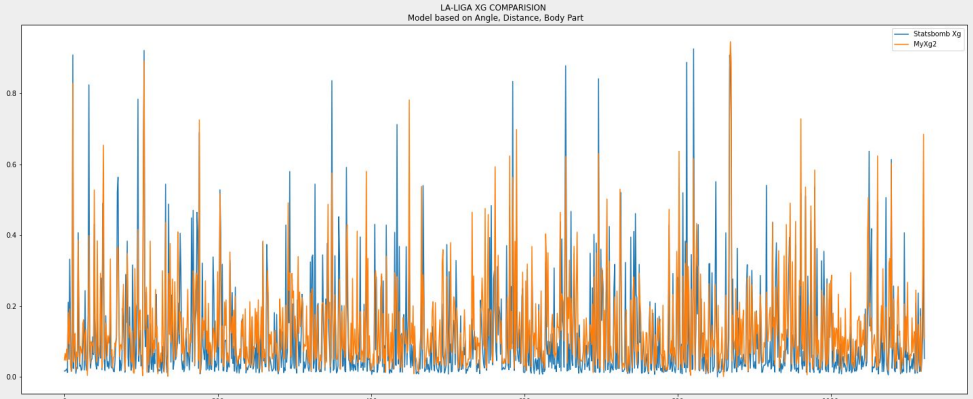
RESULTS AND VALIDATION FOR MODEL 2

Mean Squared Difference:

Champions League: 0.00985

La-Liga: 0.01022





MODEL BASED ON SHOT DISTANCE, ANGLE, BODY PART AND SHOT TECHNIQUE

- Features: Shot Angle, Shot Distance, Body Part and Technique
- Target: Shot Outcome
- Validation Feature: Statsbomb XG

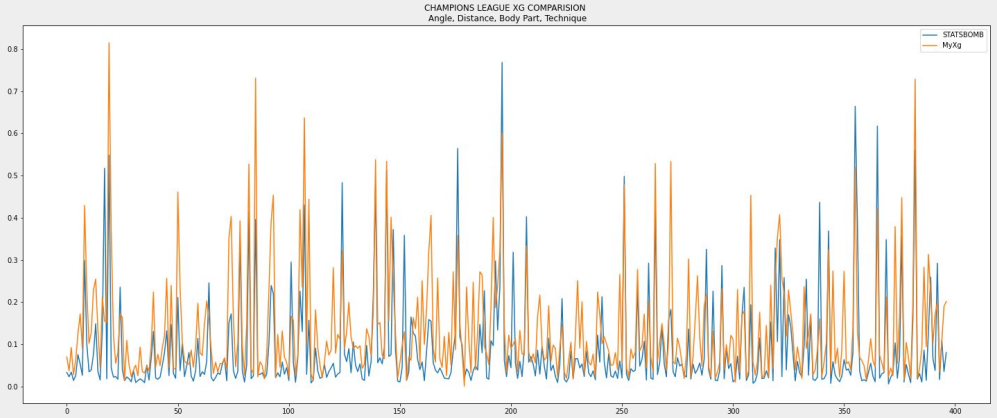
```
************{GOAL SCORED VS [ANGLE, DISTANCE, BODY PART AND TECHNIQUE]**************
Model accuracy on champions league data:  0.8967254408060453
Model accuracy on La-liga test data:  0.8833481745325023
intercept: [-1.27053529]
coef: [[ 0.3807844  -0.84458701 -0.56275679 -0.00558696]]
```

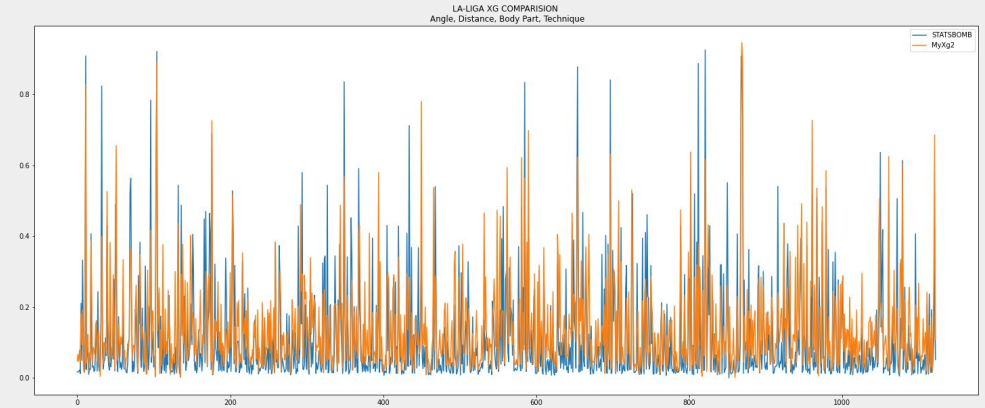
RESULTS AND VALIDATION FOR MODEL 3

Mean Squared Difference:

Champions League: 0.009821

La-Liga: 0.01019





MODEL BASED ON SHOT DISTANCE, ANGLE, BODY PART, SHOT TECHNIQUE AND PLAYER COUNT

- Features: Shot Angle, Shot Distance, Body Part, Player Count
- Target: Shot Outcome
- Validation Feature: Statsbomb XG

```
************{GOAL SCORED VS [ANGLE, DISTANCE, BODY PART, TECHNIQUE AND PLAYER COUNT]**************
Model accuracy on champions league data:  0.8916876574307305
Model accuracy on La-liga test data:  0.8895814781834372
intercept: [-0.65627795]
coef: [[ 0.41373545 -0.67289586 -0.52431084  0.01057816 -0.45607352]]
```

RESULTS AND VALIDATION FOR MODEL 4

Mean Squared Difference:

Champions League: 0.008717

La-Liga: 0.00804

FINAL RESULTS

- Though Distance and Angle are the most important features from our dataset XG model, other features like shot techniques and body part also affect the outcome of a shot.
- The most improvement in the Mean Squared distance occurs when player count is added as a feature.
- Though our XG model doesn’t perfectly align with Statsbomb XG, it can be used to understand what quality chances in soccer are made off.

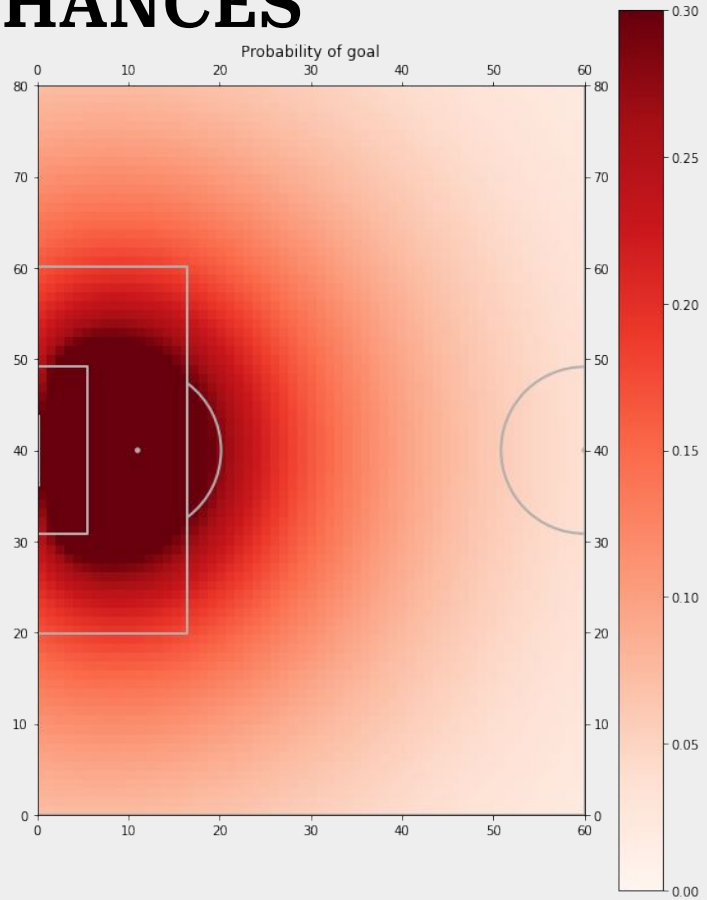
UNDERSTANDING SCORING CHANCES

Using our XG model we can look at the chance of scoring from different point on a pitch.

We can use this to make players understand where their shooting chances are coming from and what quality or XG of chances they are producing.

Two La-Liga strikers have similar finishing skills, we can use the XG model to evaluate them based on often they get in a better scoring chances. Eg how many of their chances are coming from the 30% ring.

We can add the Shot Xg's produced by a team to evaluate them against an opponent.



THINGS WE LEARNED (DATA SCIENCE) .....

- Data Engineering is the one of the important step to every data science project.
- Spend time cleaning and preparing your data for an effective ML model
- Visualization can help us understand the data better.
- Domain knowledge can greatly help with data science projects.


THANK YOU