Building an expected goals model from shot-event data - Capstone Report

The github repo for this project can be found here

Problem Statement:

The aim of this capstone project is to develop an expected goals model from soccer shots data. Expected goals or xG is essentially the probability of a shot to result in a goal.

Dataset Description:

The dataset used for this project is the StatsBomb Women's Soccer Open Data. Only the events data from this dataset is used in this project. Events data describes all the events that occurred in a match. The specific events of interest in this project are shots and any key passes prior to shots

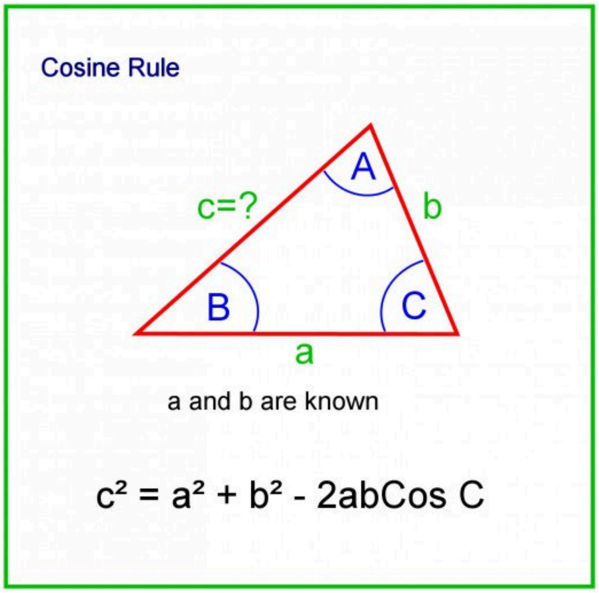
Data Cleaning:

The code for this section can be found here.

The data is in a json format. The data cleaning steps were as follows:

1. Downloaded the data to a local hard drive
    a. The data consisted of separate folders named 'events','lineups' and 'matches'.
    b. There was also a separate json file name 'competitions' which listed all the competitions and their corresponding IDs.
2. From competitions.json, extracted the competition IDs for the women's soccer leagues and tournaments - namely, the NWSL (USA), FA WSL (UK) and the 2019 women's world cup.
3. Using the competitions IDs obtained from competitions.json, the match IDs were extracted from the corresponding matches.json file and appended to a match ID list.
4. Iterating through each entry in the match ID list, the events data was loaded for each match as a dataframe from the corresponding events json file.
5. The events data for each match was comprehensive and documented each event in a particular match. Since the project only required shot and shot-related events, all the other events were dropped from the dataframe.
6. The data-structure for shot and shot-related events contain the following information:
    a. Location
    b. Time
    c. Player information (shooter and non-shooters)
    d. Type of play
    e. Technique used in shot
    f. Body part used for shot
    g. Outcome of shot
    h. Type of key-pass preceding the shot
    i. Preceding event information

    j. Whether the shot was taken first-time
    k. xG predicted by StatsBomb
7. Each of these specific factors were collated into a separate list and once the shot-related data from all the matches were collated, they were joined into a dataframe.
8. Some of the features needed to be engineered. The following features were engineered:
    a. The shot location in x-y coordinates were translated into the Euclidean distance from the center of the goal.
    b. The angle of the shot from the goal face was calculated using the cosine rule as given below



    c. The number of players in the triangle formed by the shot location and the edges of the goal was designated as the packing density. This was calculated by the barycentric algorithm given here.

The total number of shots in the dataframe assimilated was 5929.

Missing values:

1. After assimilating all the features and samples into a dataframe, the following features were found to have missing data
    a. Shot angle
        i. One instance of the shot angle was observed to be NaN. Upon closer inspection it was found that the shot location was on the same vertical coordinates as the edges of the goal. Therefore, for this shot, the missing angle is imputed as 0 degrees.
    b. Preceding pass
        i. Of the 5929 shots in the dataframe, 1845 shots don't have preceding pass information. This could be because some shots are not made in

open play, while some shots did not result from a pass but from a dribble. The missing data was imputed according to the context of the shot

Findings from Exploratory Data Analysis:

The code for this section can be found here and here.

1. Plotting all the shot locations on a pitch-map was very useful to gauge visually where the best shots are taken from. Figure 1 below plots all the shots from this dataset on to a pitch.

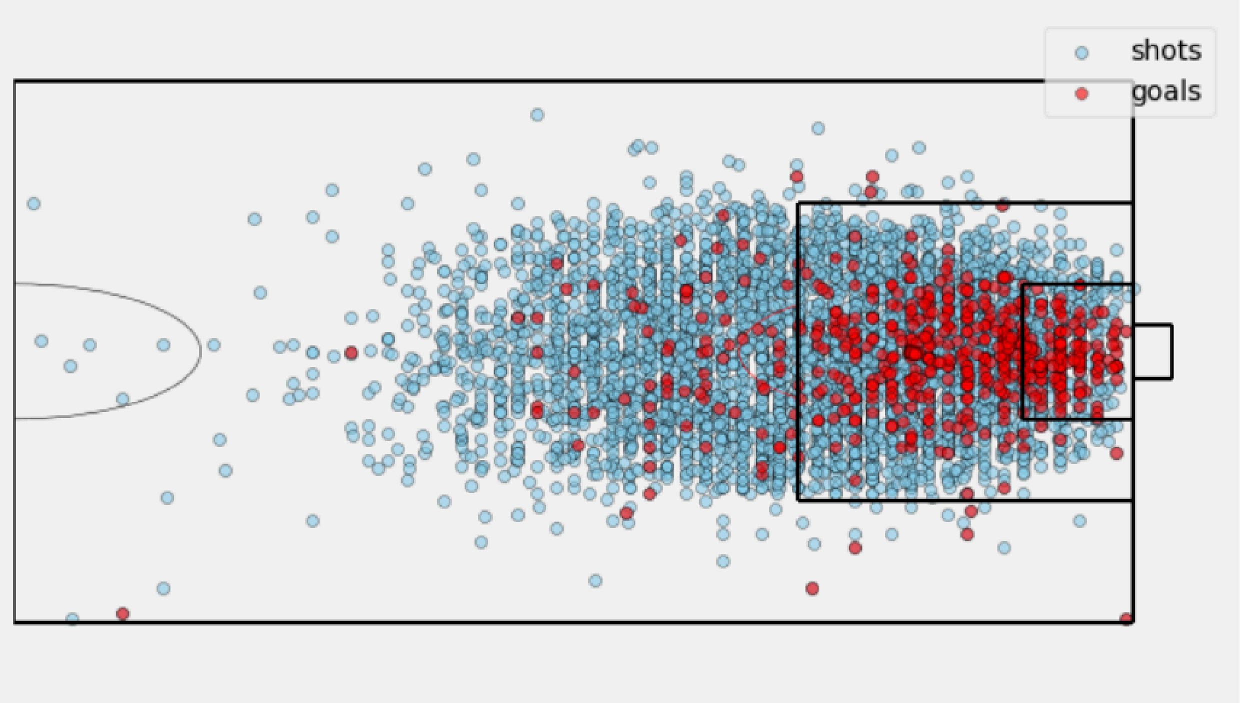


Figure 1 - Shot locations on a pitch

From Figure 1, it was inferred that the best shots are taken closer to goal and are located more centrally.

2. For the rest of the exploratory data analysis, the hypothesis tree presented in the project proposal was used as a guide. This tree has been reproduced below in Figure 2.

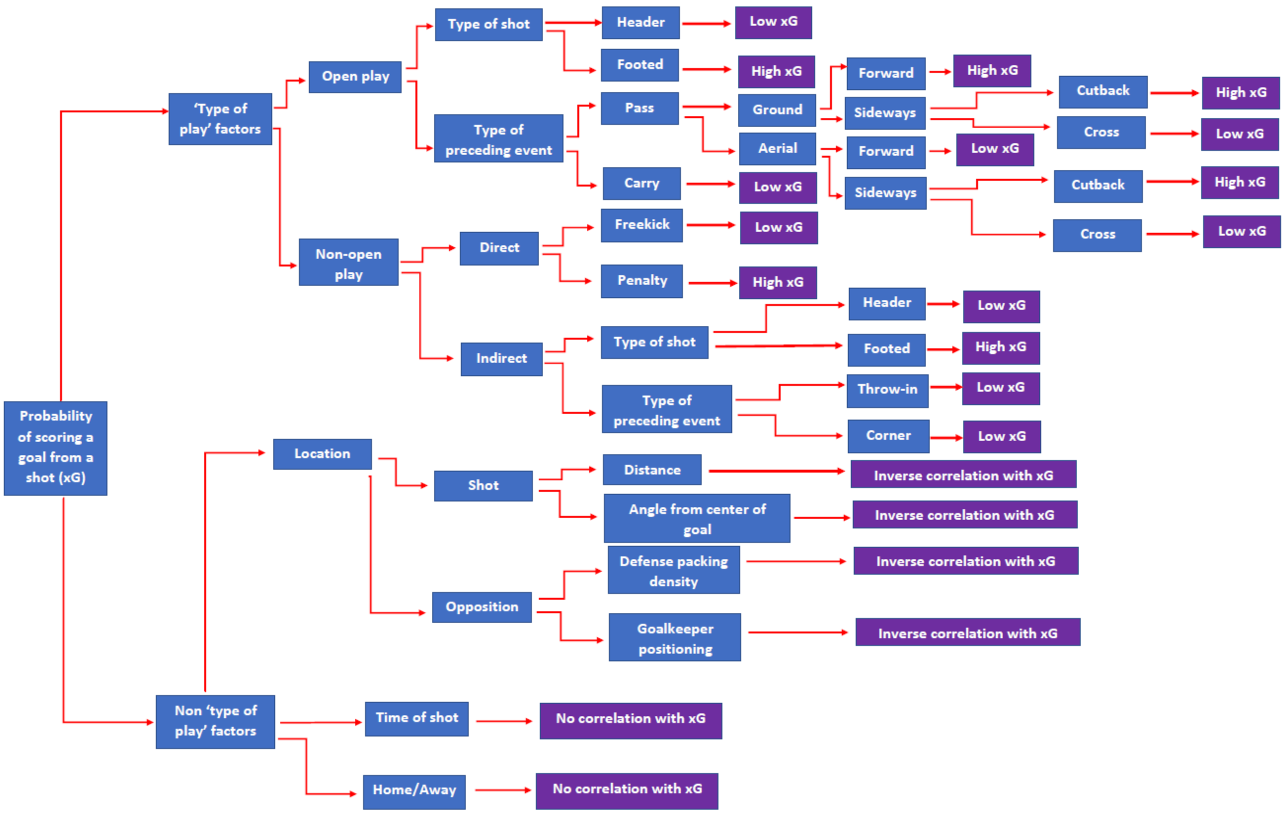


Figure 2 - Hypothesis Tree For Exploratory Data Analysis

3. Analyzing whether headed or footed shots are more valuable, it was necessary to correct for the shot distance (footed shots can be taken from farther out in the pitch). Figure 3 confirms the need to correct for shot distance.

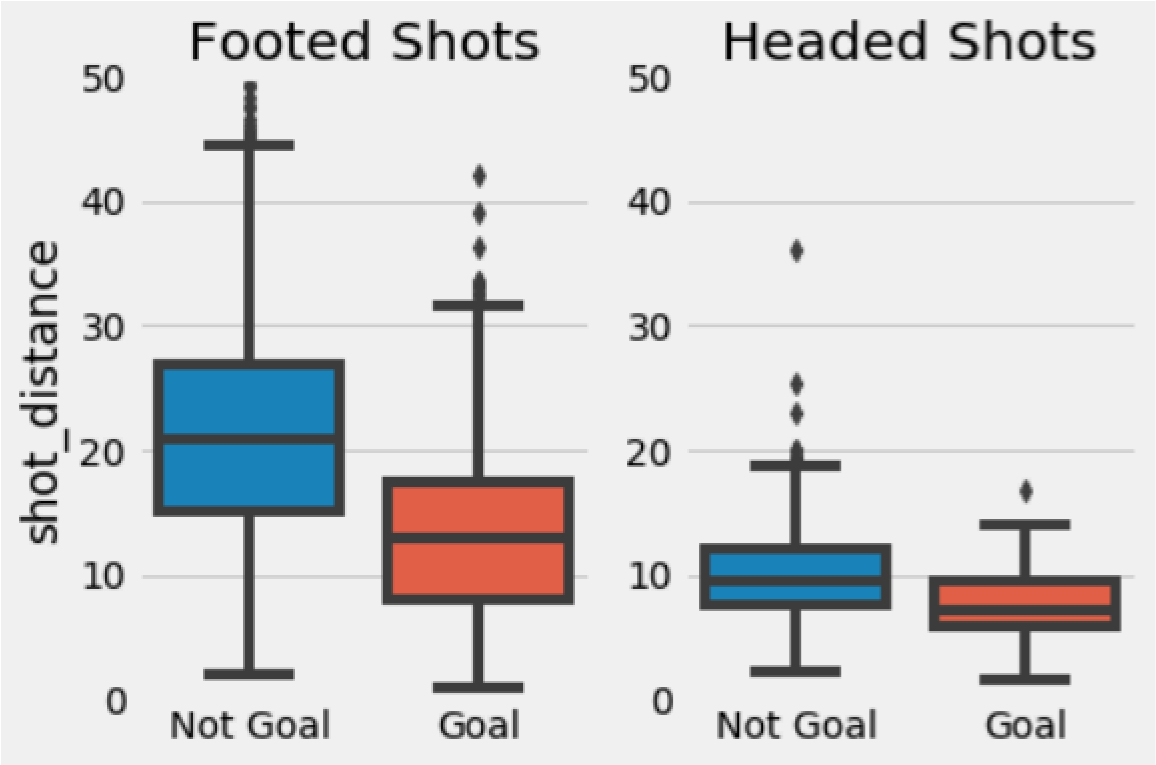


Figure 3 - Average Shot Distance of Headers and Footed Shots

After correcting for distance (all shots taken within the penalty box), the conversion percentage of shots was compared. The conversion percentage of footed shots was calculated to be 17.3% while it was 10.9% for headers. This indicated that footed shots are more likely to be successful. A/B testing confirmed that this observation was statistically significant. (p-value of 0 for null hypothesis) **(Note: During A/B testing in this project, for all features, 10000 permutation replicates were generated)**

4. To check which type of preceding pass creates higher value shots, a count-plot was plotted of all key pass types as shown below in Figure 4.

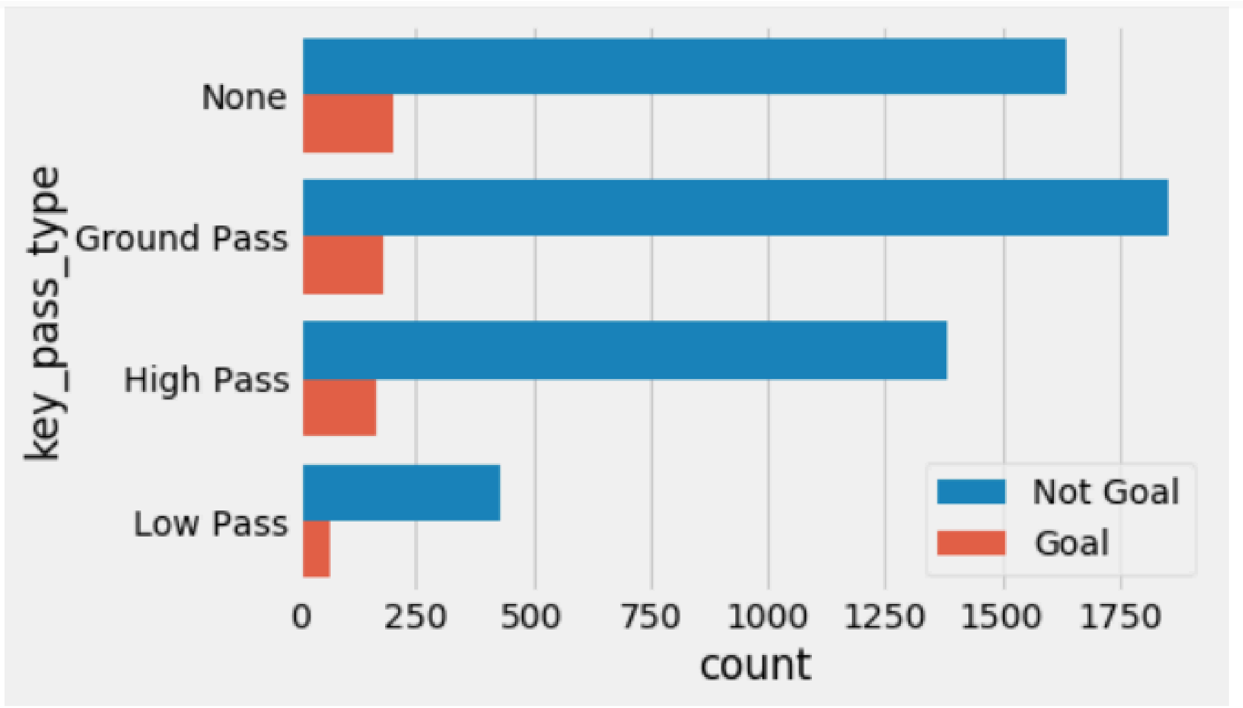


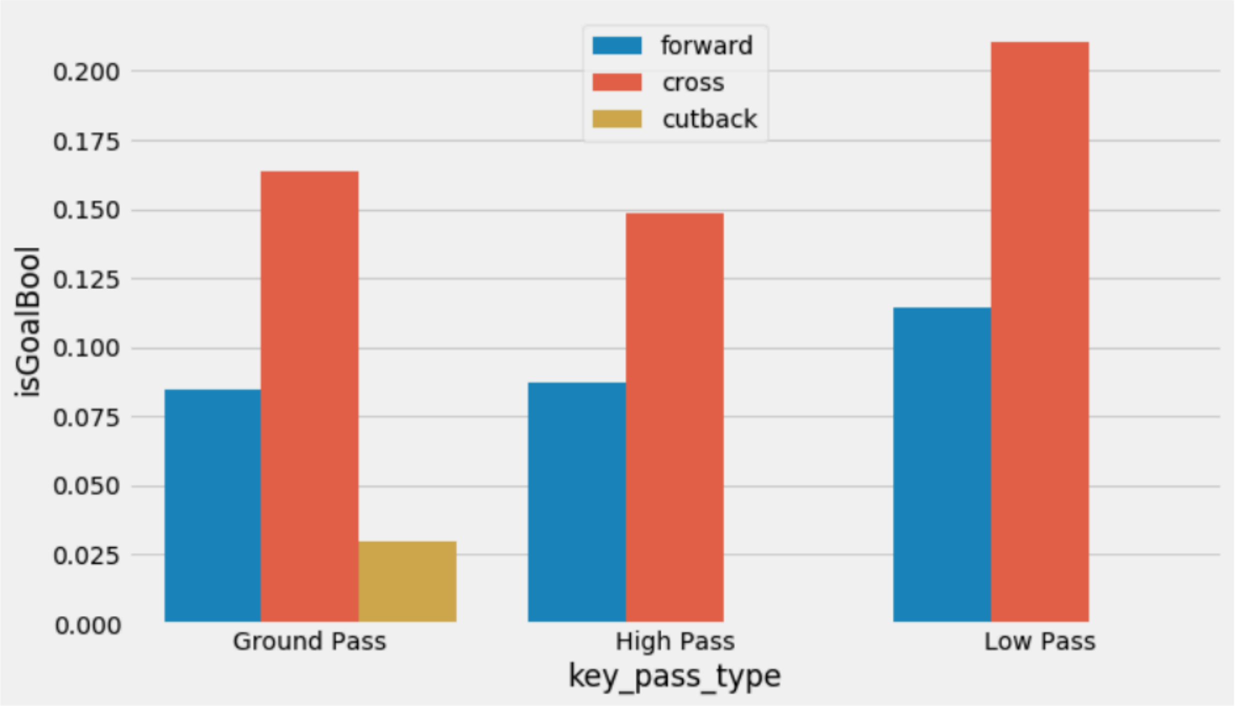
Figure 4 - Count Plot of Key Pass Types

Most shots do have a preceding pass in the dataset. The most common type was 'ground pass' followed by 'high pass' and 'low pass'. A 'low pass' is defined by StatsBomb as a pass where the ball comes off the ground but is under the shoulder level at peak height. A 'high pass' goes above the shoulder level while during a 'ground pass' the ball doesn't come off the ground.

Our hypothesis tree includes checking whether the height of a pass affects the chance of the resulting shot in becoming a goal. Ignoring all other factors conversion percentages of resulting shots was calculated. It was found that low passes create higher value shots with a conversion rate of 13%, followed by high and ground passes with conversion rates of 11 and 9%. While A/B testing confirmed the difference in conversion rates between low and ground passes as statistically significant (p-value of 0.0044). A/B testing couldn’t confirm the difference between ground and high passes (p-value of 0.96) or low and high passes (p-value of 0.088).

5. To check the effect of pass direction on the value of shot created, all key passes were grouped into three categories - crosses, cutbacks and forward passes. The StatsBomb data tags key passes as crosses or cutbacks. The rest of the key passes were tagged during this project as forward passes. Figure 5 below shows the conversion rates of shots grouped by the key pass direction. **(Note: isGoalBool is either 1 or 0 depending**

on whether a shot resulted in a goal. When used in bar plots, the average value was plotted which is in effect, the conversion rate)



<u>Figure 4 - Conversion Rates of Shots Based on Key Pass Direction</u>

It was observed that:

A. Crosses were more likely to create high xG shots followed by passes in the forward direction
B. The highest xG shots were created by low crosses.
C. In general low passes created better shots whether they were in the forward direction or across the field of play
D. Cutbacks were not very common in this dataset. Of the 42 cutbacks in the dataset, only 1 resultant shot was a goal

Because of low sample size of cutbacks, A/B testing was done between cross and forward passes and the difference was found to be statistically significant (p-value of 0)

6. Figure 5 analyzes which types of play patterns resulted in high value key passes (i.e the resultant shots were more likely to be goals)

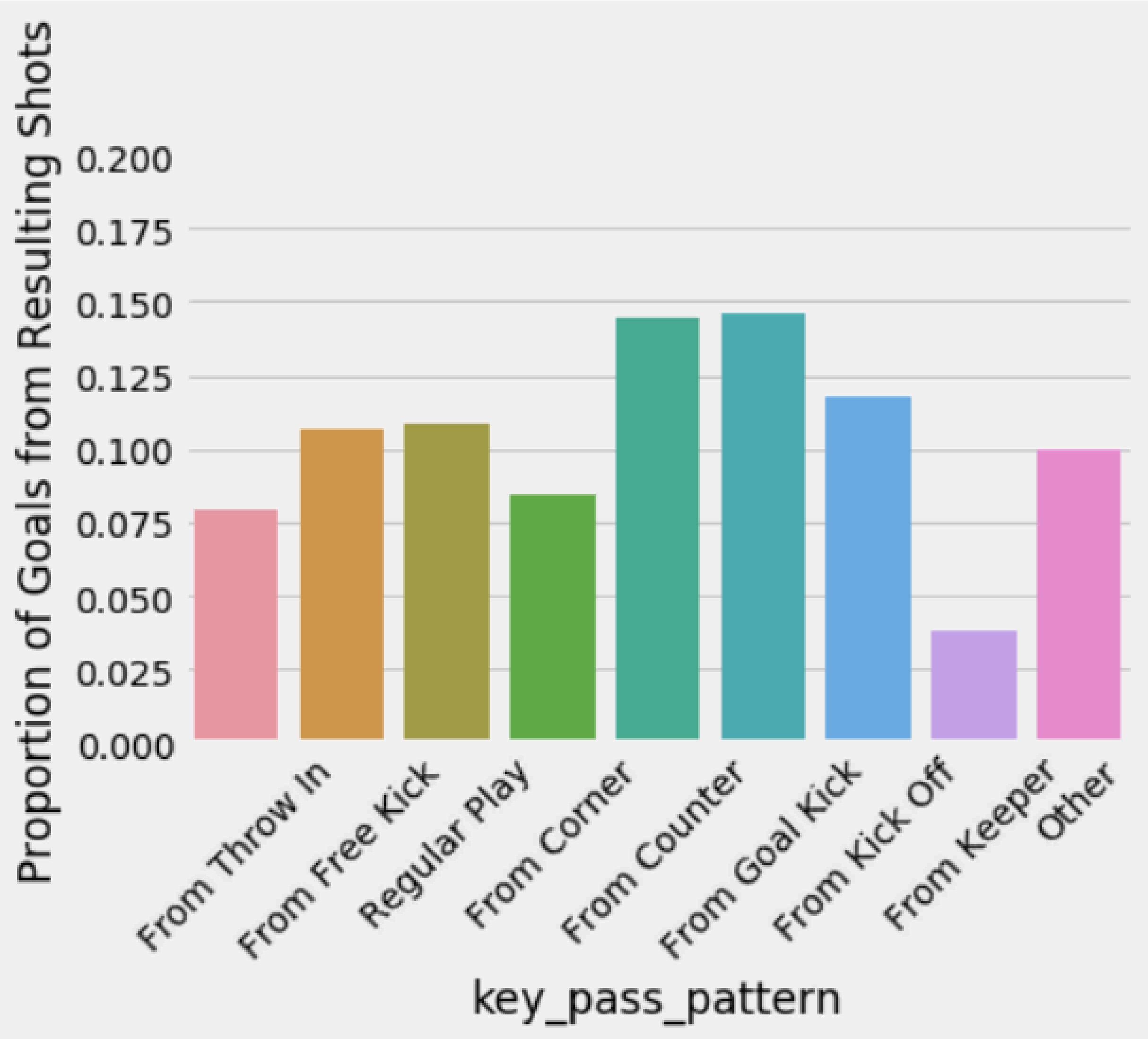


Figure 5 - Conversion Rates of Shots Based on Play Pattern

From the above it appears that best play patterns were goal kicks and counters. But upon checking the sample sizes, as shown below, these play patterns didn't have a significant sample size.

None	1835
Regular Play	1526
From Throw In	842
From Corner	675
From Free Kick	517
From Counter	270
From Goal Kick	157

From Keeper	52
From Kick Off	34
Other	10

However, A/B testing was done to check if regular play and set pieces (free kicks, throw ins and corners) resulted in equally valuable shots. The p-value of 0.0224 did not reject the null hypothesis.

7. Figure 6 compares the conversion rates of first time and shots at the end of a carry (if a shot was not taken first time, it is assumed to be the end product of a carry)

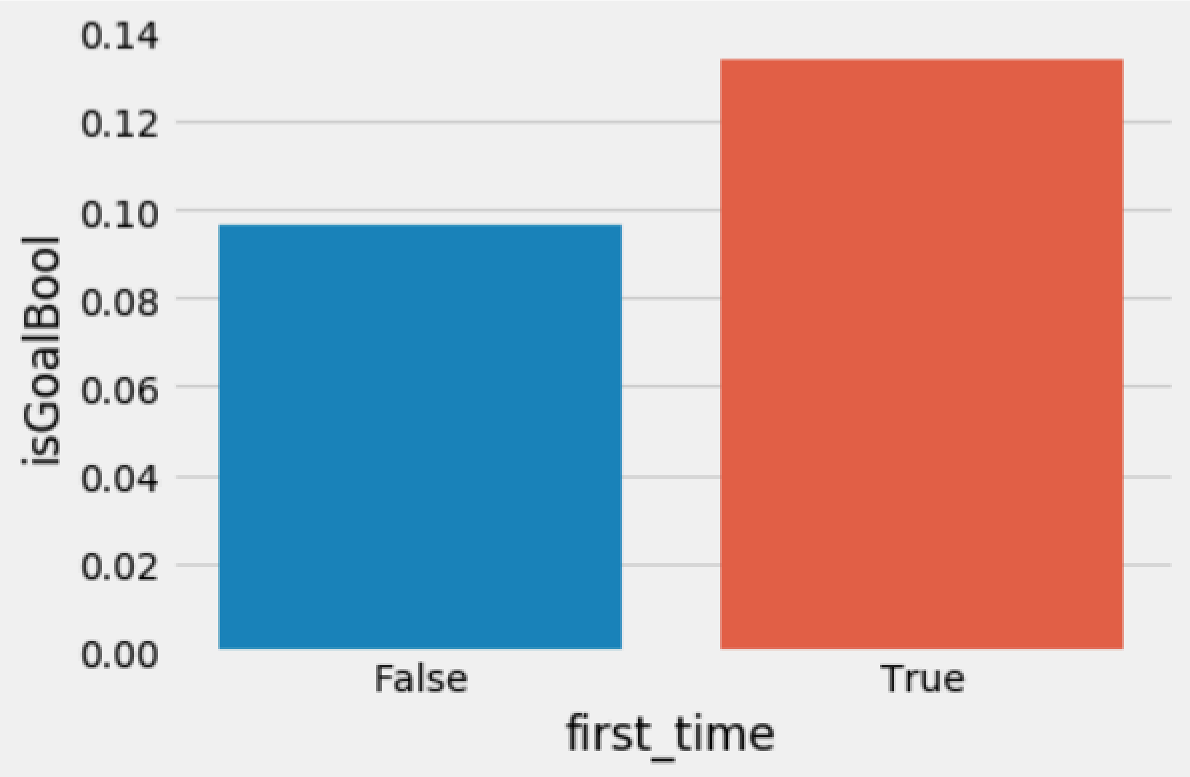


Figure 6 - Conversion Rate Comparison of First Time and Carry Shots

It appears first time shots were more likely to be successful. A/B testing confirmed this with a p-value of 0.0001

8. From the exploratory pitch plot and indirectly from studying other features, it was observed that distance from goal has a clear inverse correlation to conversion rate. Figure 7 below shows that successful shots were usually taken at shorter distances from the goal

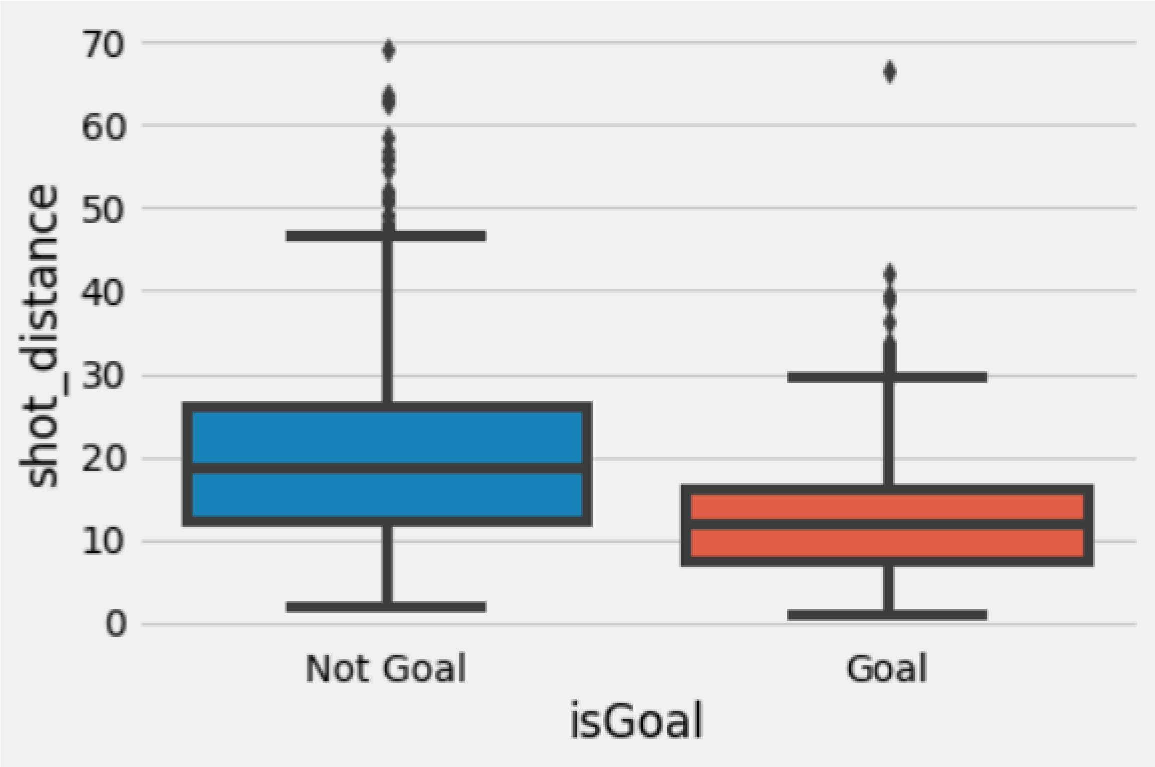


Figure 7 - Boxplots of Shot Distances for Successful and Unsuccessful Shots

A/B testing long and short distances (taking a shot distance of 20 yards as a cut-off between long and short) supported this observation (p-value of 0)

9. Figure 8 shows the boxplots of shot angles for successful and unsuccessful shots. The median shot angle for unsuccessful shots were lower, indicating shots taken from a larger distance. The IQR for the successful shots lied between 25 and 50, while for the unsuccessful shots it lied between 18 and 30. A/B testing actually could not disprove the null hypothesis (p-value of 1)

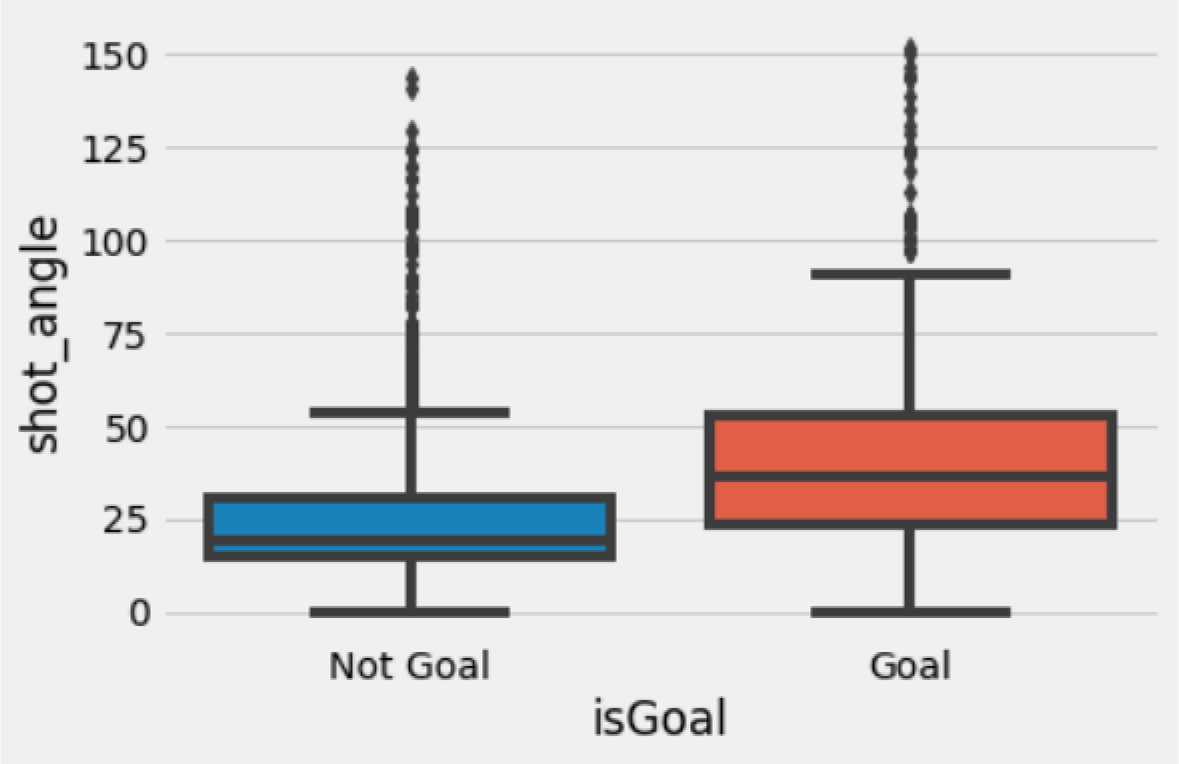


Figure 8 - Boxplots of Shot Angles for Successful and Unsuccessful Shots

10. Comparing the conversion rates between open play, freekicks and penalties showed that penalties had a very high conversion rate of 71% while open play shots had a conversion rate of 10%. Direct free kicks had a conversion rate of only 6%. A/B testing was not done since sample sizes are skewed heavily in favor of open play shots.

11. The next feature studied was the packing density in front of the shot. Figure 9 shows the boxplots of packing densities.

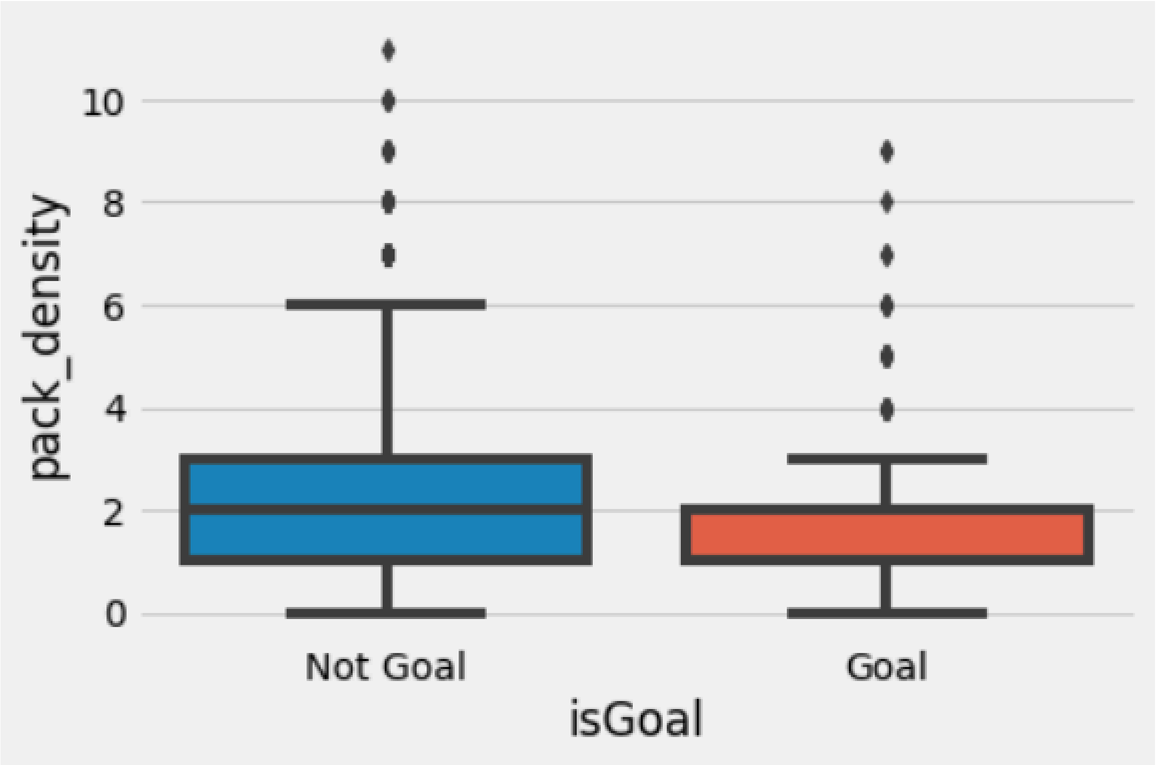


Figure 9 - Boxplots of Pack Density for Successful and Unsuccessful Shots

Most shots, regardless of whether they were goals, were taken when fewer players were directly in between the shot and goal. Unsuccessful shots generally had more players in between the shot and the goal than successful shots. Packing density accounts for whether the goalkeeper was in the way of a shot. It doesn't qualitatively differentiate how good their positioning is. This is hard to quantify as it depends on the physical attributes for each keeper like jumping or diving reach, height etc. For this study, a simplistic approach was adopted wherein packing density was the only metric which considered positioning of other players with respect to the shot. A/B testing packing density with a density of 3 as cut-off resulted in a p-value of 0.

12. Non ‘type of play’ factors such as time of shot and whether the player belonged to the home or away team did not affect the conversion rate of a shot as shown below in Figures 10 and 11. These features were subsequently dropped from further analysis.

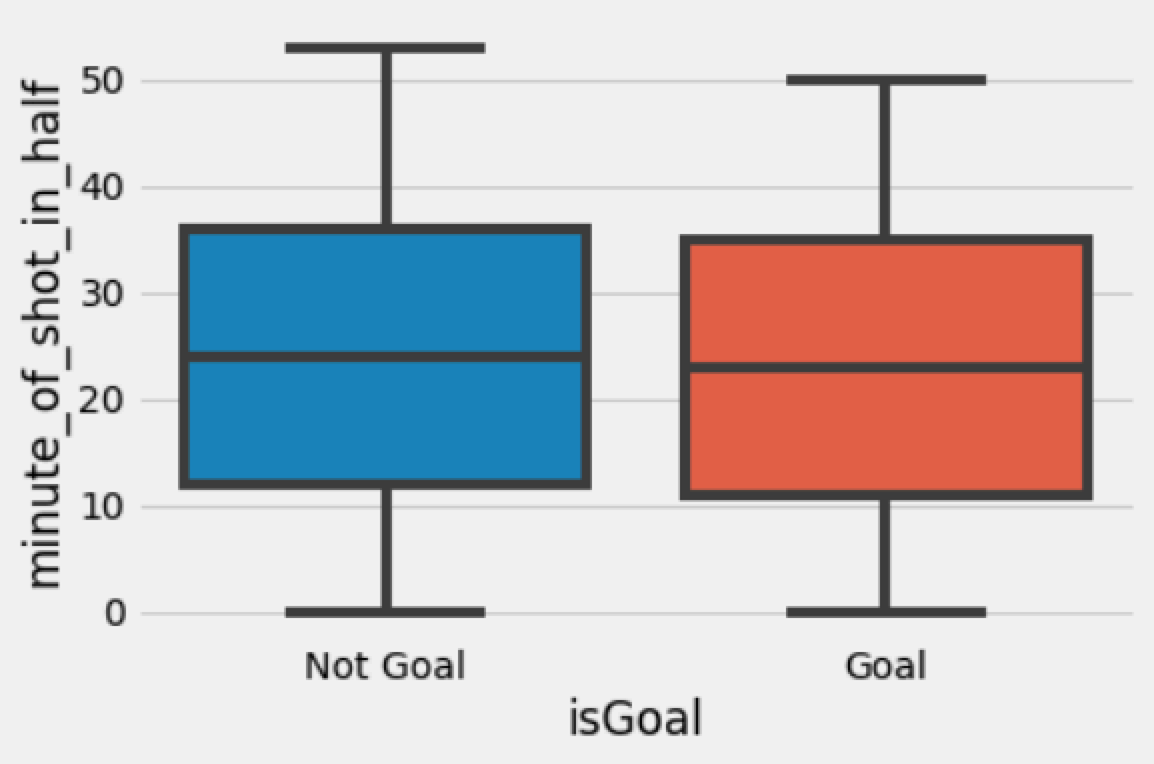


Figure 9 - Boxplots of Time of Shot of Successful and Unsuccessful Shots

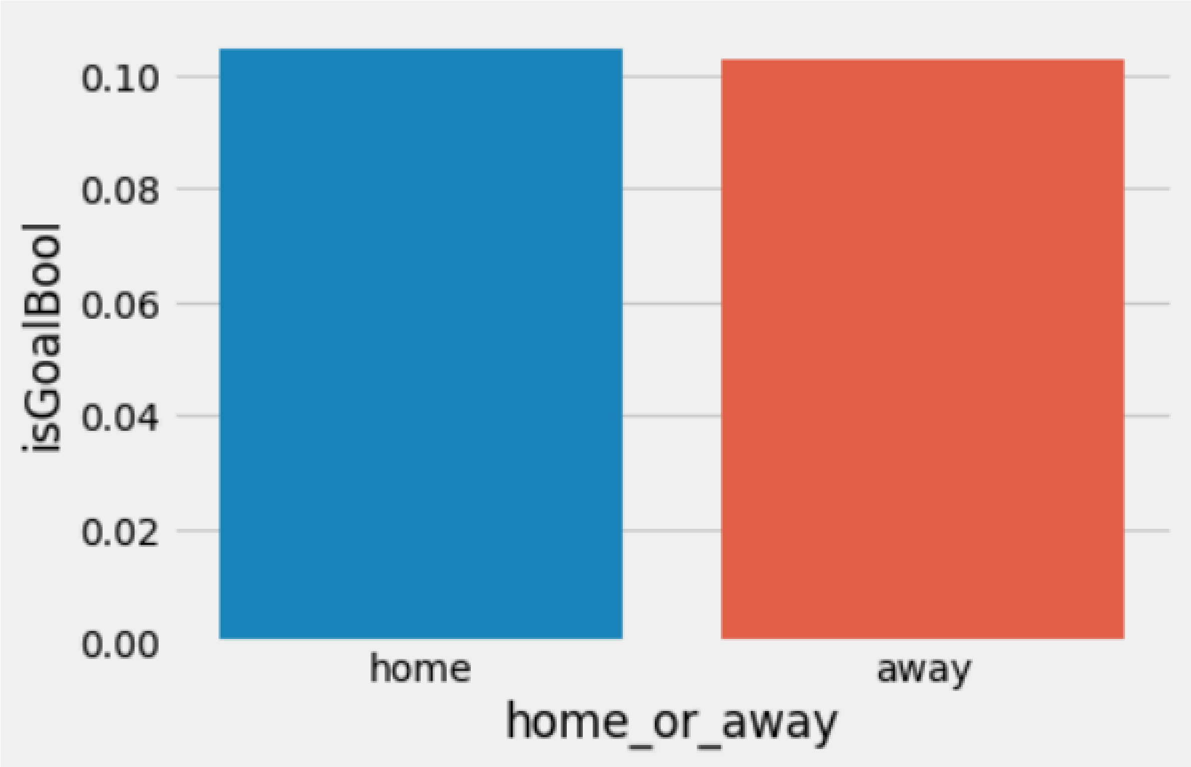


Figure 10 - Conversion Rate of ShotsTaken by Home and Away Players

Machine Learning Pipeline:

Based on the findings from EDA and statistical inference some further data cleaning was required. During this pre-processing, the following steps were followed:

1. Shots resulting from cutbacks were removed as the sample sizes are too small
2. Since sample sizes for the different key pass pattern types such as 'From Kick Off', 'From Keeper' etc. were small, all the key pass patterns were modified to one of three categories - 'None', 'Regular Play' and 'Non Regular Play'
3. Since the actual sample sizes of direct freekicks and penalties were small, they were dropped.
4. The body parts feature was modified to only say 'Foot' or 'Head'. Any shots taken by 'Other' body parts were dropped
5. Additionally, the following features were dropped
    a. Time of shot
    b. Whether shot was taken by a home or away player
6. The final number of shots - 5615
7. Final feature count - 8 (excluding StatsBomb predicted xG which was used for comparisons)

Data pre-processing:

The following features were categorical (text data):

- 'body_part_name'
- 'key_pass_type'
- 'key_pass_pattern'
- 'cross_cutback_forward'
- 'first_time_or_carry'

The following features were numerical:

- 'shot_distance'
- 'shot_angle'
- 'pack_density'

The sklearn ColumnTransformer class was used to simultaneously pre-process both numerical and categorical features. The StandardScaler was used on the numerical features while the OneHotEncoder was used on the categorical features.

Once this pre-processing was done, the train_test_split function was used to split the data into training and test data using a 70-30 split.

Fitting and predicting using Machine Learning models:

Detailed evaluation for each model and the code for this section can be found here. The report will only summarize the machine learning models' performance.

Since our aim was to obtain a probability for each shot, the following classification algorithms were tried:

- Logistic Regression
- Support Vector Machines with Polynomial and RBF kernels
- K Nearest Neighbors
- Bagging Classifier
- Random Forest
- AdaBoost
- GradientBoost
- XGBoost

Table 1 below summarizes the model performance as measured by the AUC scores. The ROC curves and grid search parameters for each model can be found here.

The StatsBomb model gave an AUC score of 0.81 on the test set. The aim of testing different models was to benchmark results against the StatsBomb model and select the best performing models.

Table 1 - AUC Score Ranking for Tuned Machine Learning Models on Test Data

Model	AUC Score	Rank
Logistic Regression	0.771	1
XGBoost	0.769	2
Gradient Boost	0.769	3
SVC - RBF Kernel	0.762	4
Random Forest	0.761	5
AdaBoost	0.761	6
Bagging	0.753	7
KNN	0.752	8
SVC - Polynomial Kernel	0.68	9

The models developed in this project were developed on a significantly smaller sample size than the StatsBomb model, so the difference in performance was expected. However, the best performing models compared very well to the benchmark despite the grid search being an exponential one. Further hyperparameter tuning is expected to increase performance.

Instead of choosing Logistic Regression as our final model of choice, an ensemble was created using the top 3 performing models. The predicted probabilities of our ensemble model were an average of the predicted probabilities of each individual classifier for each sample. The ensemble classifier increased the AUC score t
. Figure 11 below shows the ROC curve for this ensemble classifier.

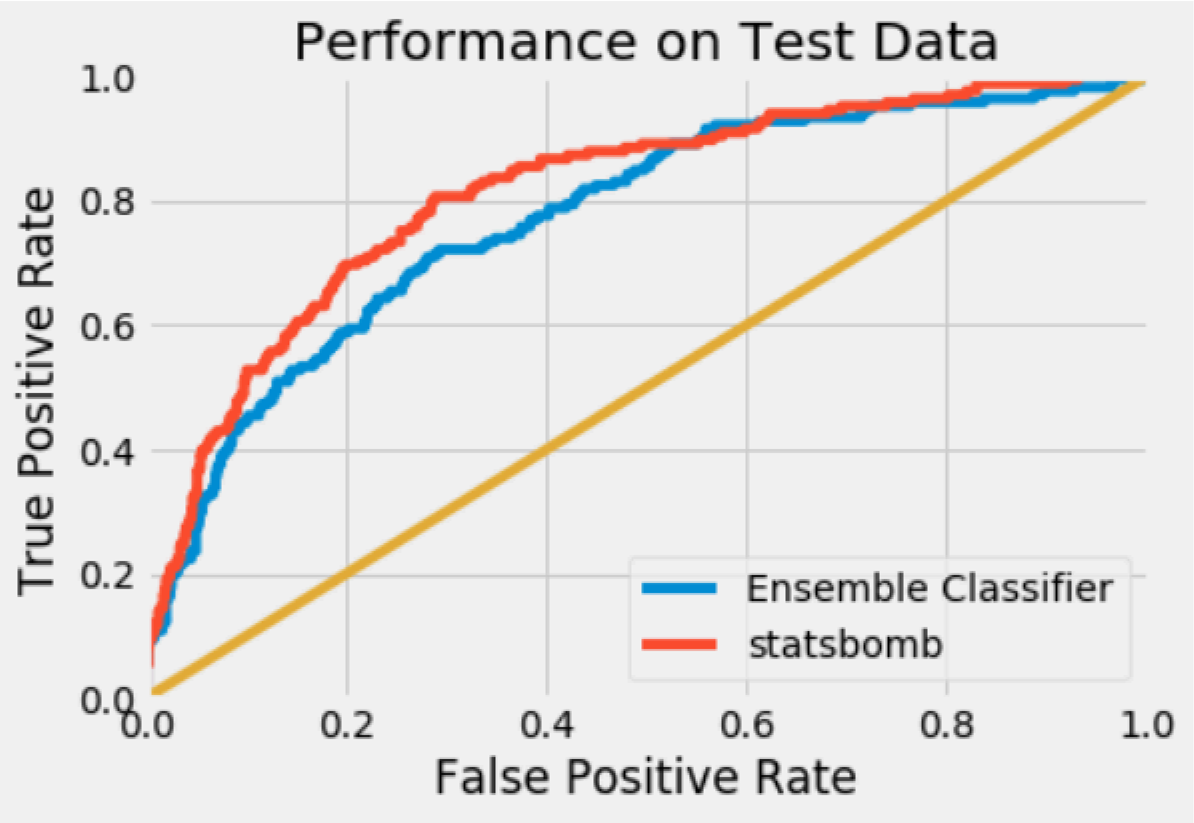


Figure 11 - ROC Curves of Ensemble xG Classifier and StatsBomb xG Model

Feature Importances:

Figure 12 below shows relative feature importances, as predicted by each of our top classifier models. Note, for the logistic regression model, the model coefficients were used as a proxy for feature importance.

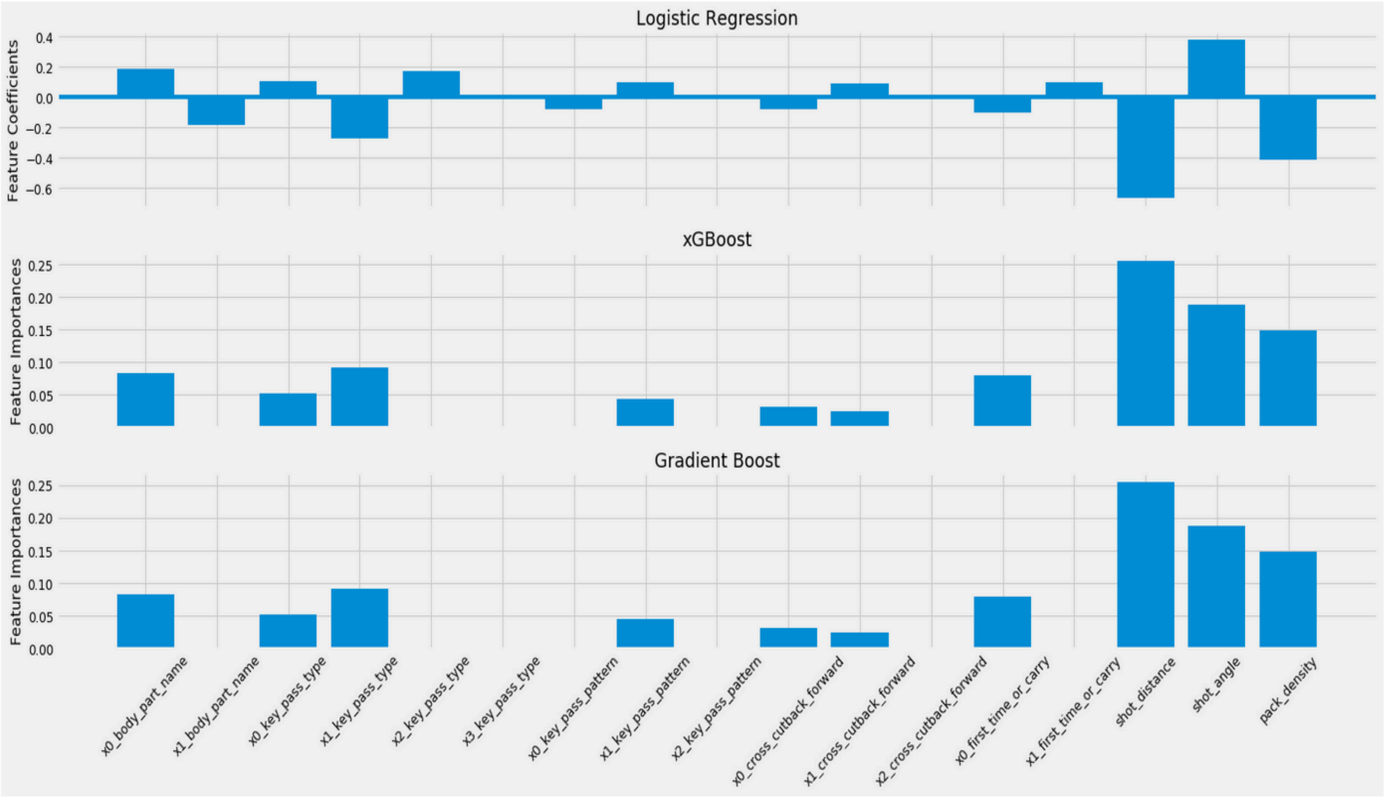


Figure 12 - Feature Importances For Top 3 AUC Score Models

From this figure it can be observed that:

- The most important features were unsurprisingly the shot distance, shot angle and pack density, which made intuitive sense.
- The most important categorical features were the body part used to take the shot, what type of pass preceded the shot and whether it was taken first time. This also made sense when considering the statistical inference and data exploration results.
- XGBoost and Gradient Boost did not consider some of the varieties within each categorical variable important, while Logistic regression gave small coefficients to most of these values.
- Of the features considered, the pattern of play which resulted in the key pass and the pass direction were considered least important by all three models.

Application of xG Model:

The code for this section can be found here.

Our newly developed xG model can be used at three different levels:

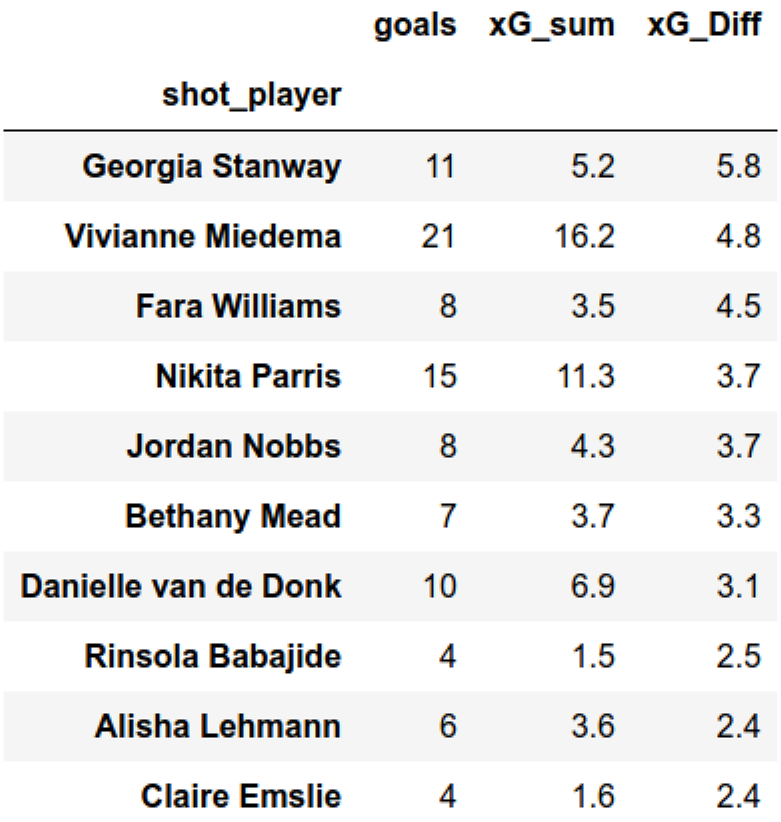
1. Player level - evaluate performance of each individual player from an xG perspective
2. Match level - analyze ebb and flow of a game with xG
3. Team level - analyze how good a team is relative to its competition

The 2018-2019 WSL data was used to illustrate these applications in the following sections since it had a full season's worth of data.

Player Level Analysis

Comparing cumulative xG over a season with actual goals scored is an efficient way to judge a player's goal scoring ability. Table 2 below lists the top 10 players from the 2018-2019 season, ranked by xG difference, i.e. the difference between number of goals scored and the cumulative xG predicted by the model developed in the previous section.

Table 2 - Player Ranking Based on xG Difference - Top 10

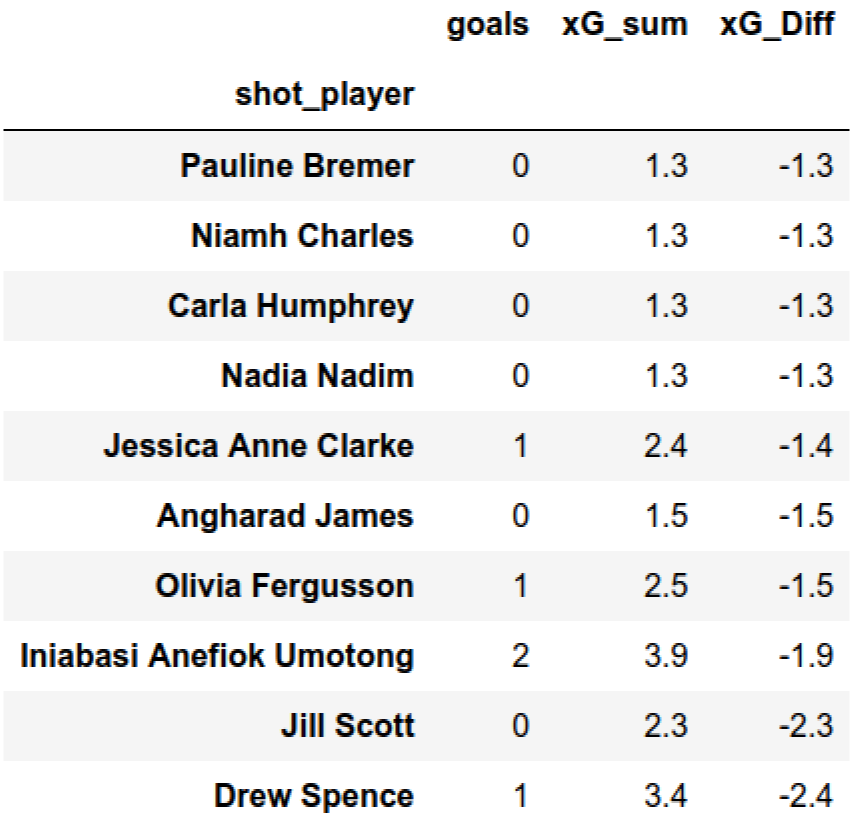


	goals	xG_sum	xG_Diff
shot_player			
Georgia Stanway	11	5.2	5.8
Vivianne Miedema	21	16.2	4.8
Fara Williams	8	3.5	4.5
Nikita Parris	15	11.3	3.7
Jordan Nobbs	8	4.3	3.7
Bethany Mead	7	3.7	3.3
Danielle van de Donk	10	6.9	3.1
Rinsola Babajide	4	1.5	2.5
Alisha Lehmann	6	3.6	2.4
Claire Emslie	4	1.6	2.4

Looking at the top 10 players, we can see that there was a stand-out goal scorer - Vivianne Miedema from Arsenal. She significantly outperformed our xG model, which indicates impeccable finishing. Another impressive goal scorer was Georgia Stanway who had the biggest xG_Diff, who scored 11 when our model only predicted a total of 5 goals based on the features of her shots.

Unsurprisingly, the top 10 is filled with forwards and attack minded midfielders as they tend to shoot more. Now, let's look at the other end of the table at players who underperformed with respect to their cumulative xG as shown in Table 3 below.

Table 3 - Player Ranking Based on xG Difference - Bottom 10



	goals	xG_sum	xG_Diff
shot_player			
Pauline Bremer	0	1.3	-1.3
Niamh Charles	0	1.3	-1.3
Carla Humphrey	0	1.3	-1.3
Nadia Nadim	0	1.3	-1.3
Jessica Anne Clarke	1	2.4	-1.4
Angharad James	0	1.5	-1.5
Olivia Fergusson	1	2.5	-1.5
Iniabasi Anefiok Umotong	2	3.9	-1.9
Jill Scott	0	2.3	-2.3
Drew Spence	1	3.4	-2.4

Drew Spence, a midfielder from Chelsea, didn't have a great finishing season. Jill Scott, a defender, also took some shots but didn't score. Curiously, some of the names in the bottom 10 like Nadia Nadim and Pauline Bremer are forwards. Given how low their cumulative xG was, the reason they were in the bottom 10 could be a lack of playing time. A much better metric would be xG-per-90minutes, so that we normalize xG with the playing time. This was not analyzed as part of this project.

Note that the above stats are for shots which are not direct freekicks or penalties. When comparing the final goal tallies for each player with the actual goals they scored for this season, some variability is expected.

Figure 12 below is a graphical representation of Table 11 and shows how much better Georgia Stanway and Vivianne Meidema were than our xG model predicted them to be.

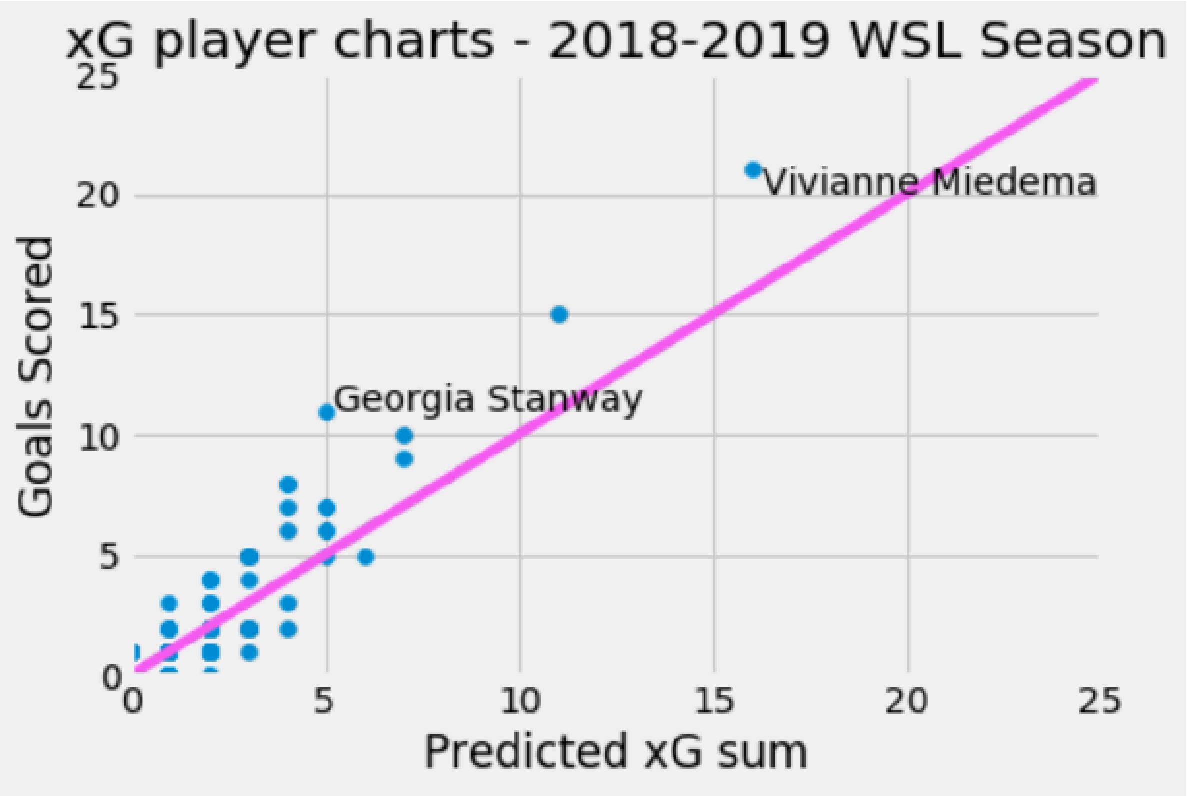


Figure 12 - Player xG Charts from 2018-2019 Season

Match Level Analysis

To illustrate how xG can be used to shape match analysis, the Everton-Brighton game from the 2018-2019 season was selected. This game ended 3-3 and hence seemed a good candidate for data storytelling and visualization. Figure 13 is a trend plot of cumulative xG as predicted by the ensemble model over the whole 90 minutes.

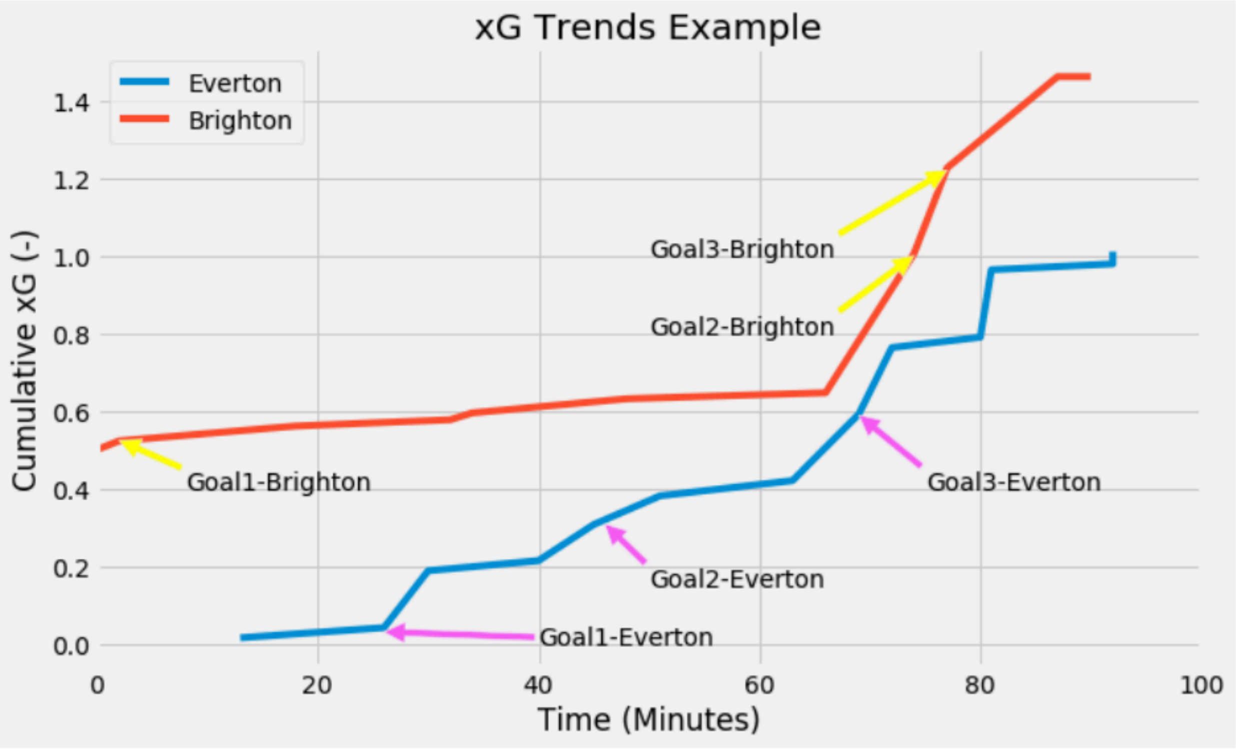


Figure 13 - Cumulative xG Plots Match Level Analysis

Here are some of the observations gleaned from this plot, presented as something a pundit on BBC's Match of the Day (with a statistical edge, albeit without having watched the game) might say:

- Brighton started the game very well and created a very good shot at the very beginning of the game as the xG plot starts at around 0.45. But they didn't score from this shot. They scored shortly after though, so it was a very strong opening to the game - they were ahead within the first 5 minutes! The gain in xG from the opening shot to the scoring shot was not significant, indicating that the scoring shot itself is not a high chance shot. A long range belter, perhaps?
- Everton on the other hand started the game on the backfoot despite being the home team. They didn't have a shot till the 10th minute. Even then their shots were not of the greatest quality as the cumulative xG was below 0.1 till the 25th minute. This indicated that Everton had set up for the game very defensively and on the rare occasion they had a shot in the first 25 minutes, the shots were low xG (either from far out or out wide or into a packed defense etc.)
- After Brighton's great start, they eased the foot off the gas as there was no sharp jump in the cumulative xG plot. This indicated that they might have switched game plans to control the game now that they were in the lead
- Despite the low xG shots Everton were resigned to, they managed to level the game around the 27th minute with what appears to be a wonder goal since the xG was very very low! This seemed to give them more attacking impetus as they started creating much better shots (look at the spikes in the cumulative xG plot). For the rest of the first half, Everton were in control as

Brighton did not create any more good shots. At the end of the first half, Everton capitalized and took the lead. 2-1!
- The second half resumed where the first half finished, with Everton creating better shots while Brighton struggled to create good shots. Everton scored another in the 70th minute and went 3-1 up!
- At this point Brighton made a tactical switch and focussed on attacking as much as possible. This paid off as they scored from two very good shots between minutes 75 and 80. Now the game was tied up!
- From this point on, both teams were trying to win the game as we can see sharp spikes in the xG plots, indicating defensive lapses on both sides. But the game ended 3-3.
- Overall, the cumulative open play xG predicted 1.4 goals for Brighton and 1.0 goal for Everton
- Both teams out-performed their xG either because of superb finishing or defensive lapses
- Despite the final score, Brighton will be disappointed to not have won this game as they created significantly better chances. They had two great chances, one each in the beginning and end of the game, and missing them proved costly

Such detailed match analysis and commentary can be derived from a simple plot. To add more depth to the analysis, pitch maps can be used as shown in Figure 14 below. This is from the same match described above.

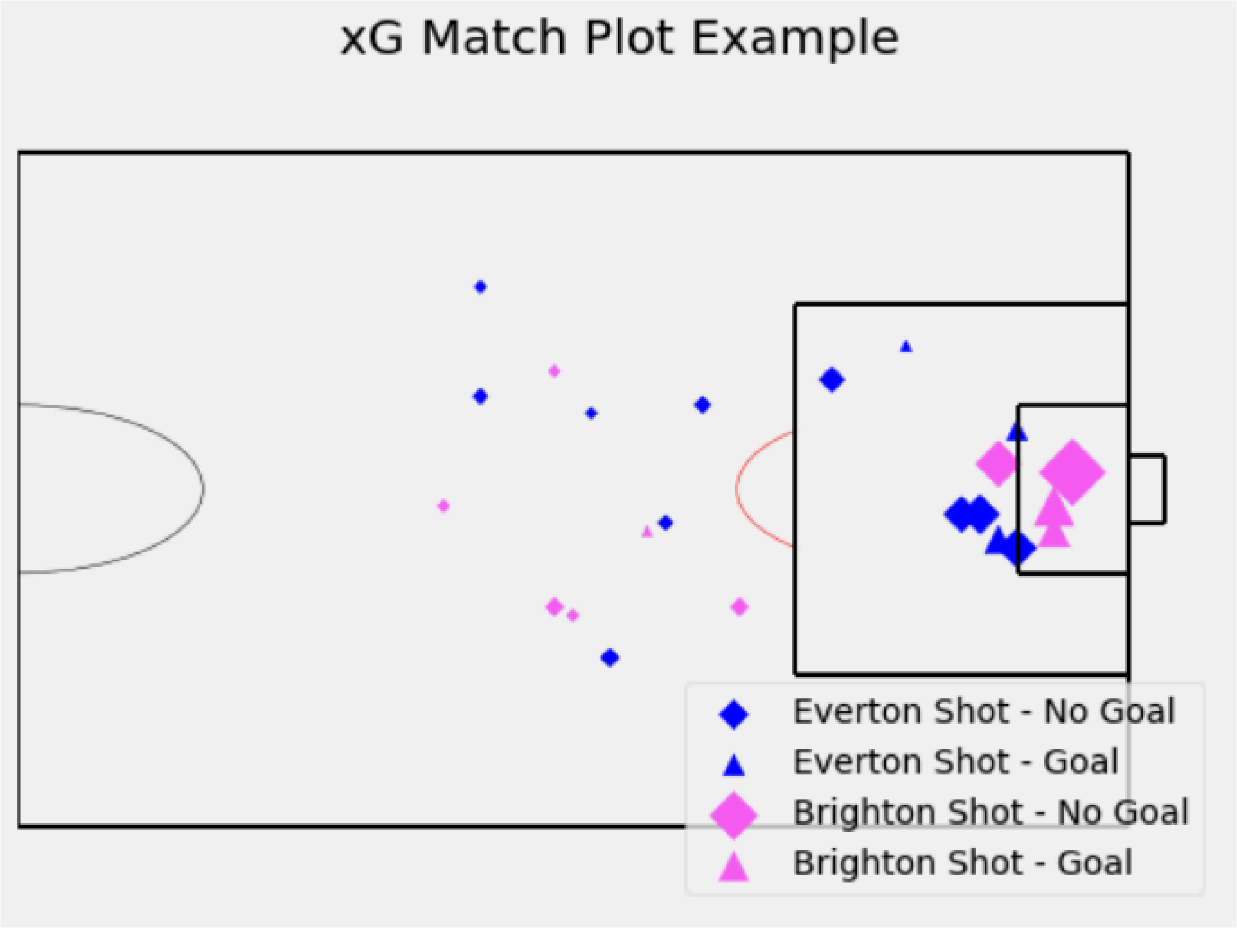


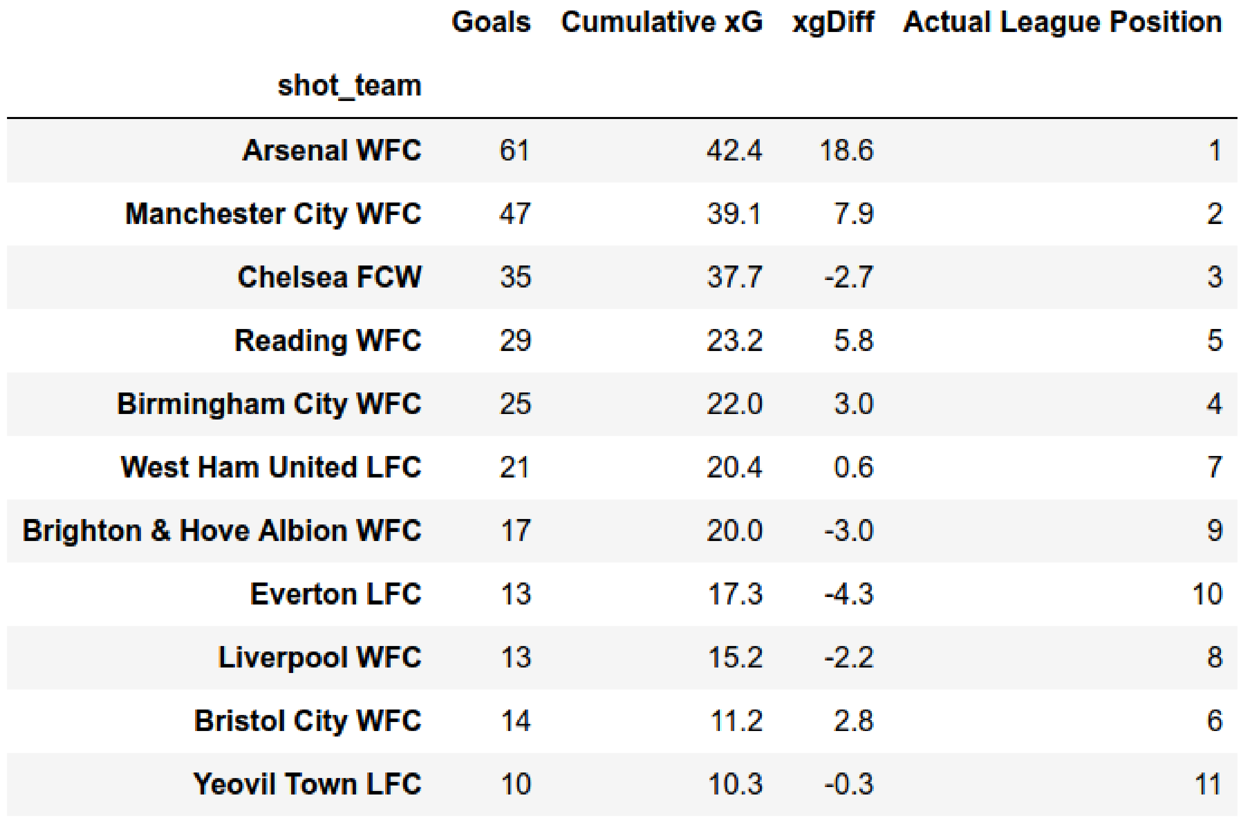
Figure 13 - Match Level Pitch Map Shot Analysis Using xG

The above pitch map indicates xG using marker size. Brighton had some good shots from close to the goal. The biggest xG chance was close to the goal and missed by Brighton. This was the big chance they missed at the beginning of the game. Also notice the Brighton goal scored from outside the box. This was the first goal Brighton scored. The first goal Everton scored was actually inside the penalty box but from a tight angle.

Team Level Analysis

The xG model can also be used to evaluate the performance of teams over a whole season. Table 4 is a season summary of the teams who competed in the 2018-2019 WSL season, sorted by cumulative xG and the actual league positions they finished in.

Table 4 - Seasonal Summary Using Cumulative xG



	Goals	Cumulative xG	xgDiff	Actual League Position
shot_team				
Arsenal WFC	61	42.4	18.6	1
Manchester City WFC	47	39.1	7.9	2
Chelsea FCW	35	37.7	-2.7	3
Reading WFC	29	23.2	5.8	5
Birmingham City WFC	25	22.0	3.0	4
West Ham United LFC	21	20.4	0.6	7
Brighton & Hove Albion WFC	17	20.0	-3.0	9
Everton LFC	13	17.3	-4.3	10
Liverpool WFC	13	15.2	-2.2	8
Bristol City WFC	14	11.2	2.8	6
Yeovil Town LFC	10	10.3	-0.3	11

Looking at the final league positions and comparing it to the xG table, we can see that open play xG was a pretty decent predictor of team performance.

Arsenal significantly out-performed xG numbers and rightfully finished on top while Yeovil were woeful at creating good shots and finished at the very bottom. Everton significantly under-performed their xG and actually finished second-to-last but according to xG they should have finished 8th. Bristol City outperformed their xG numbers and finished 6th while xG had them placed at 10th.

Apart from just league tables, we can also use pitch maps and xG to describe styles of play and quality in attack. Figure 14 below shows pitch maps with shot locations for all the teams in the 2018-2019 WSL season. For reference, Figure 15 shows a pitch map for all the shots from the 2018-2019 season.

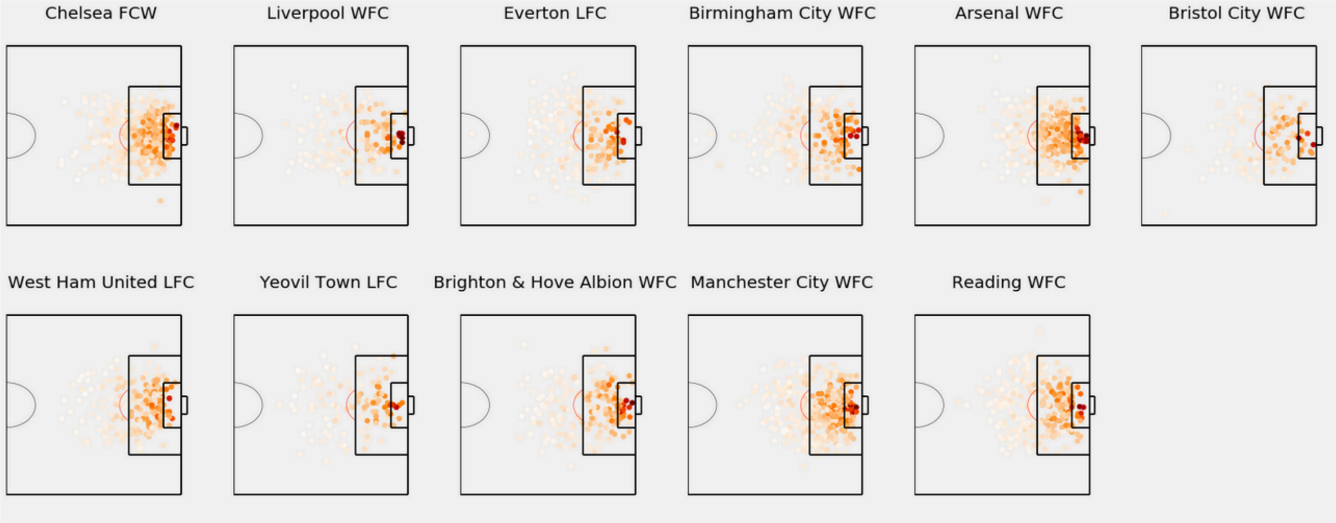


Figure 14 - Team Level Pitch Map Shot Analysis Using xG

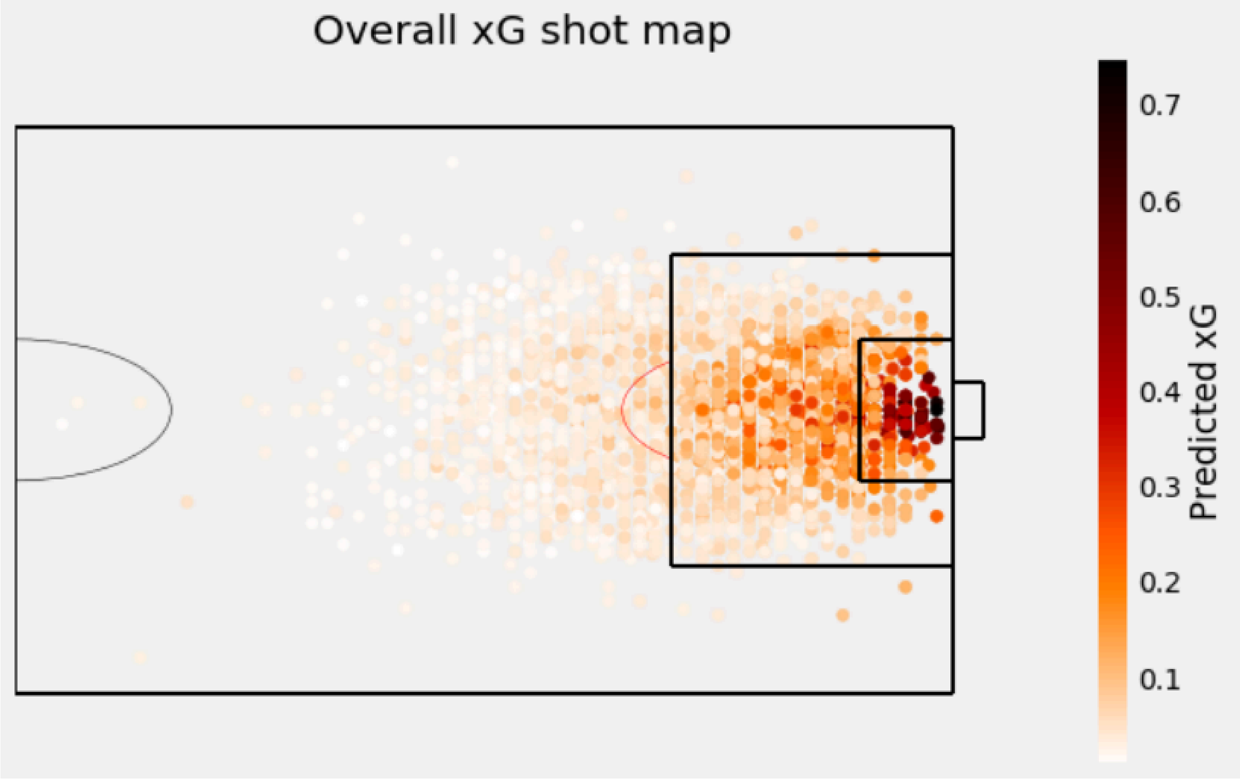


Figure 15 - Overall Pitch Map Shots with xG

Arsenal, Man City and Chelsea created good quality shots from within the penalty box a lot of the time while teams like Yeovil created fewer shots overall and most shots were from outside the box. This indicated a lack of strength in attack. They were playing 'out of their league' and were relegated to a lower division.

There are many other ways to visualize soccer in terms of xG. For e.g: xG can also be used to quantify how good your defense is by looking at the xGA (xG Against). If the xGA is low, it indicates a good defense. But for the purposes of this project, this was not studied.

I hope this was informative and fun and perhaps changed the way you watch soccer at least a tiny bit!

Acknowledgements and References:

- Ankur Verma, Springboard Mentor
- StatsBomb
- FCPython
- Football Hackers by Christoph Biermann