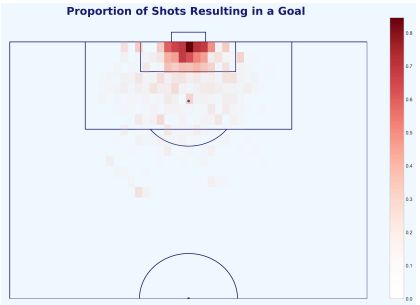
Junior Data Science Challenge

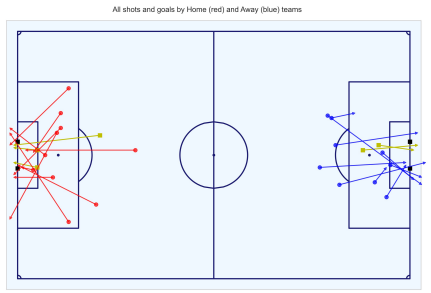
Edward Webster

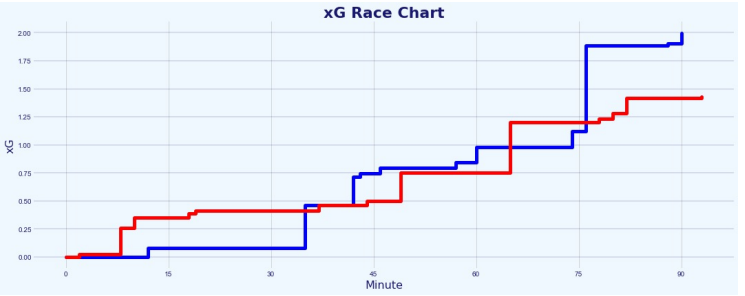
March 2021

Slides: https://docs.google.com/presentation/d/116D0U_ue2sv6hgLBgHnil228cRhph0qfMuOFA4uuxhs/edit?usp=sharing
Code: https://github.com/eddwebster/mcfc_submission
Full pack including all attachments: https://drive.google.com/drive/folders/1Ts_5YOL8JpVpRcaEhqmcTCceLmcBRY3L?usp=sharing

Contents







All code for this challenge: https://github.com/eddwebster/mcfc_submission/tree/master/notebooks/

Introduction and Challenge Brief

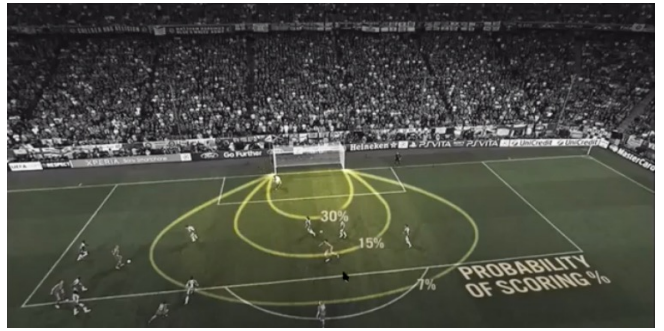
Brief

Part 1 and 2 of the data challenge as outlined in the brief

1

Challenge 1 - Building a Chance Quality Model using shots data

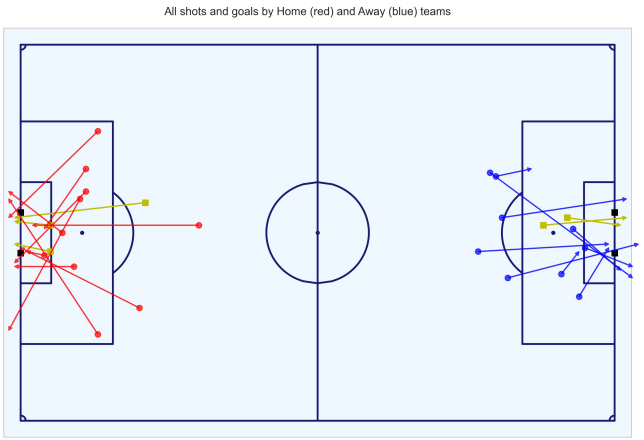
Provided is a sample of just under 11,000 shots (ShotData.csv, and the description in ShotData.txt). Using this data, build a Chance Quality Model (CQM) that calculates the probability of a shot resulting in a goal (i.e. P[goal|shot,situation]) using whichever situational variables in the data deemed informative. Provide a description of the method that you chose, including any metrics and plots that you have used to understand and assess the performance of your model. This description may take the form a slide pack (PowerPoint, Google slides, etc); no more than a total of 10 slides).



2

Challenge 2 - Application of Tracking Data

The second step is to work with the Tracking data for a single game. to analyse the shooting opportunities that each team created. In this GitHub repository*, there is the Tracking data for two matches, along with a description of the data. Using the data for sample game 2 in the repository, identify the shots in this game and write a short report describing the major chances that each team created during the game, making use of the Chance Quality Model developed in Step 1 and any other information that is relevant. Based solely on the quality of chances that each team created, which team deserved to win the game? This report may take the form a slide pack (PowerPoint, Google slides, etc); no more than a total of 10 slides).



Expected Goals images taken from both David Sumpter's Friend of Tracking lesson: 'How to Build An Expected Goals Model 1: Data and Model (https://www.youtube.com/watch?v=bpjLyFyLIXs) and his book Soccermatics.
All code for this challenge: https://github.com/eddwebster/mcfc_submission/tree/master/notebooks/.
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data.

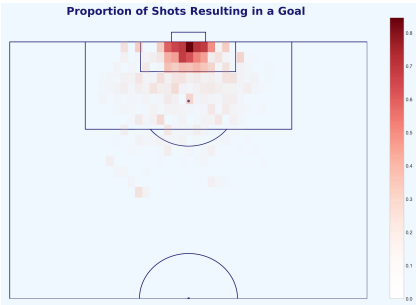
Analysis Process

Explanation of how this data challenge was tackled in 3 parts, explained in a total of 20 slides

1

Define, build and train and Chance Quality Model using Shots data

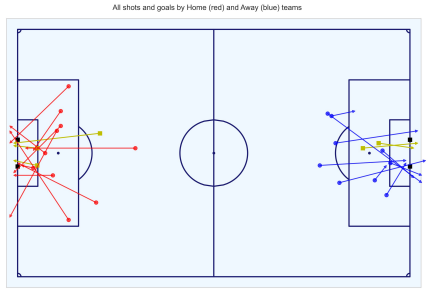
Build and train a basic Logistic Regression model, to determine likelihood of shots resulting in goals from a sample of just under 11,000 shots. Performance of the model observed throughout, through iterations of: performance metric selection and definition, outlier treatment, univariate and multivariate analysis with subsequent feature engineering, production of a final model, measurement of performance, and feature interpretation.



2

Application of Metrica Sports Tracking and Event Data

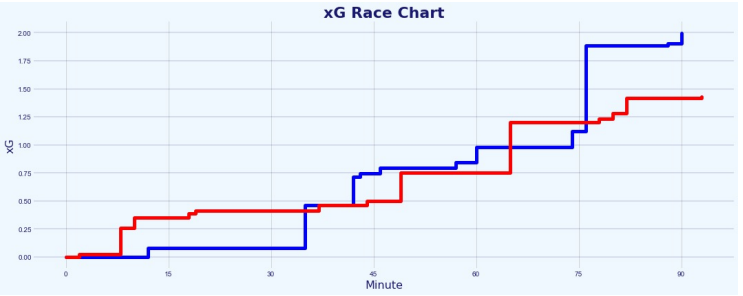
Using the Tracking data and corresponding Event data, determine all the shots for selected match (game 2) and extract features required to determine the Expected Goal value for each shot using the trained Chance Quality Model. These features include: distance to the goal; distance to the centre of the pitch; angle to the goal; number of intervening opponents; number of intervening team mates; interference on the shooter; whether the shot resulted in a goal; whether the shots was a penalty or direct free kick; and whether the shot was taken with the player's foot or head.



3

Assessment of the performance of the teams in game 2 of the sample Metrica Sports data through application of the Chance Quality Model (derived in step 1) and exported Metrica Sports shot data (derived in step 2)

Using the exported Metrica shots data with engineered features, predict the likelihood that the chances that took place in the match resulted in a goal or not using the trained Chance Quality Model. Brief analysis of the match includes a visualisation of an xG race chart for the match and also further discussion about why Expected Goals, when used in isolation for one match, should be treated lightly.



All code for this challenge: https://github.com/eddwebster/mcfc_submission/tree/master/notebooks/.

See definitions table in the appendix for the full list of features and their definitions, as used in the final trained Chance Quality model.

Part 1: Chance Quality Modeling

Creation of a Chance Quality Model (CQM) using a provided sample dataset of shots

Chance Quality Model

Key objectives achieved with this data, explained further in this section

1

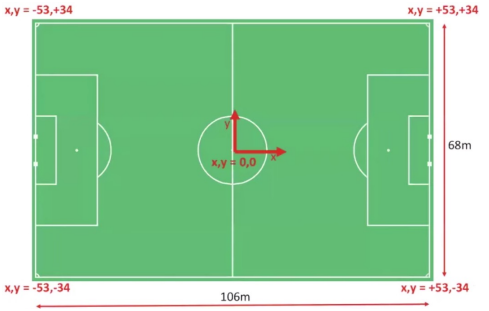
What are Expected Goals (xG)?

Background information about the proxy of choice for the Chance Quality Model. What is this xG metric and why has it made such a huge impact on football data analysis?

2

Data Engineering

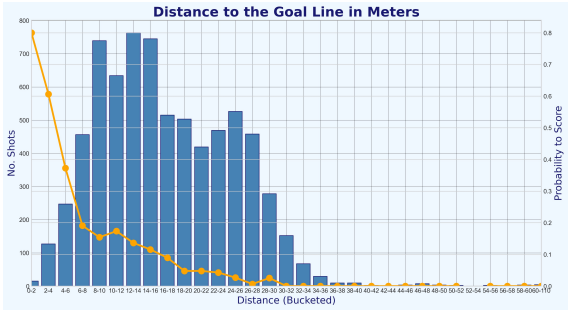
Reworking the dataset into a form ready for modeling including: converting the pitch coordinates to a standardised coordinate system, filtering the data for just Open Play goals, and cleaning attributes.



3

How to build a Chance Quality Model using Expected Goals

A basic model to measure the likelihood of shots resulting in goals from a sample of just under 11,000 shots. The model uses Logistic Regression and the approach is made up of the following steps: performance metric definition, outlier treatment, univariate and multivariate analysis with subsequent feature engineering, production of a final model, measurement of performance, and feature interpretation.



Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

What are Expected Goals and Why are they Important?

Background information of what is arguably the most important metric in football analytics so far – Expected Goals

- 'Expected goals', or 'xG', is a derived football metric that measures the shot quality or probability that a shot with defined position, conditions, and situation, will, on average, result in a goal. Put simply, xG is the probability (from 0 to 1) that a shot taken will be a goal, or not. If a chance is 0.5 xG, it should be scored 50% of the time. The higher the xG , the more likelihood of the opportunity being taken. So

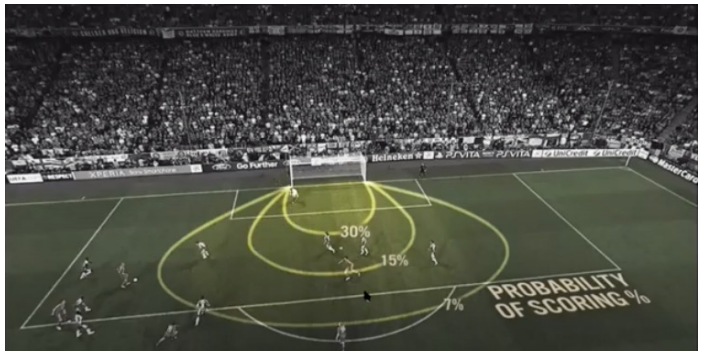
- xG was created by Sam Green at Opta, and the factors that he took into account to calculate the likelihood of an attempt being scored from a specific position on the pitch during a particular phase of play, with features including the distance from goal, the angle of the shot, did the chance fall at the player's feet or was it a header? was chance a one-on-one? what was the assist like? (e.g. long ball, cross, through ball, pull-back), in what passage of play did it happen? (e.g. open play, direct free-kick, corner kick), is the shot a rebound?, and many more.

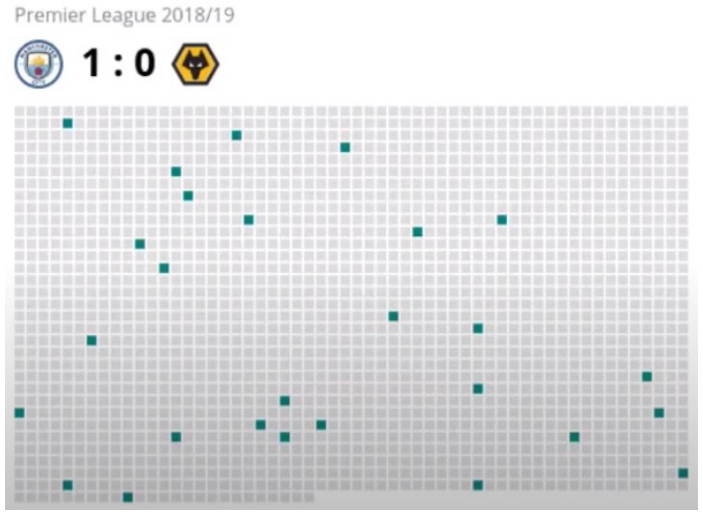
- In a low-scoring game and unpredictable game such as football, final match score often does not provide a clear picture of performance. This is reflected in the success rate of bookmakers, who pick the favourite in football matches less successfully than in other sports. In football, just over 50% of the favourites win, compared with over 60% in top American team sports.

- Goals are also very rate with the average number of goals scored per game is 2.66, with a score every 69 minutes per team. This is also reflected in the number of attempts on goal, with football teams shooting around twelve times per match, around 0.4% of a 3,000 Event dataset. No other sport has so many efforts from a team before anything happens regarding the score line (The Numbers Game, Anderson and Sally). The second visualisation on the right shows the ratio of the number of shots (green) to other events in a match that took place between Manchester City and Wolves during a match in the 18/19 season, where fewer than 1% of the on-the-ball actions in the match resulting in shots.

- Although goals in football are rare and random in isolation, they are predictable over a longer period of time and are closely fitted by a Poisson distribution. The average number of goals can be taken and the Poisson distribution applied (The Numbers Game, Anderson and Sally). Expected Goals allow analysts and statisticians to look an all shots, which happen around ten times as many times as the goals themselves, making for a much better predictor of goals scored by a team in the medium term (How to Build an Expected Goal Model, Sumpter).

- xG can be used to make smarter decisions on recruitment, tactics, and strategy. For example, clubs who recruit strikers can look past the randomness of actual goals scored and identify the underlying shot quality. This can help aid club decrease the risk made by recruitment departments on signing players that are in purple patches, and instead, identify players constantly create enough high-quality chances at the level the scouting team is looking for.



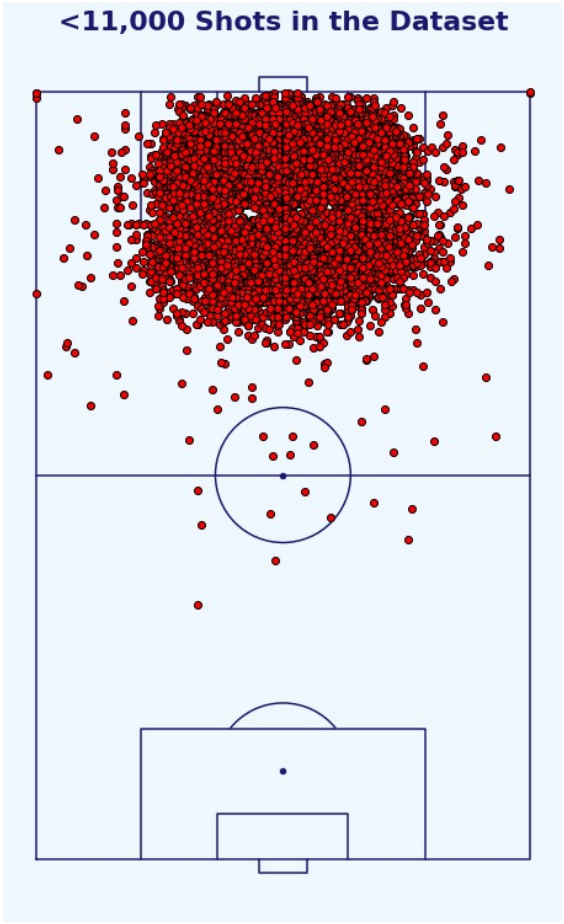


Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb
Expected Goals images taken from both David Sumpter's Friend of Tracking lesson: 'How to Build An Expected Goals Model 1: Data and Model (https://www.youtube.com/watch?v=bpjLyFyLIXs) and his book Soccermatics.
Visualisation of the number of shots in a game taken from Lotte Bransen and Jan van Haaren's talk 'How to find the next Frankie de Jong' for Friends of Tracking: https://www.youtube.com/watch?v=w0LX-2UgyXU.
Statistics taken from chapter two of The Numbers Game by Chris Anderson and David Sally and Soccermatics by David Sumpter.

About the Data

Information about the provided dataset of just under 11,000 shots

- The shots DataFrame contains 10,925 shots and 1,374 goals (12.6%), of which there are:
  - 10,777 Non-Penalty (NP) shots and 1,261 non-penalty goals (11.7% of all NP shots); and
  - 10,269 Open Play (OP) shots and 1,218 Open Play goals (11.9% of all OP shots).
- Each row consists of a single shot event.
- The features available include:
  - *match_minute*: minute of the match in which the shot was taken.
  - *match_second*: second of match_minute in which the shot was taken.
  - *position_x*: position of the shot on the pitch in meters (x-coordinate).
  - *position_y*: position of the shot on the pitch in meters (y-coordinate).
  - *play_type*: game situation in which the shot was taken (open play, penalty, direct free kick, direct from a corner). This feature will be used to filter the dataset for only Open Play shots.
  - *BodyPart*: body part with which shot was taken (left foot, right foot, head, other).
  - *Number_Intervening_Opponents*: The number of opposing players that were obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
  - *Number_Intervening_Teammates*: The number of teammates that are obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
  - *Interference_on_Shooter*: The degree of direct interference exerted on the shot-taker from defenders (Low - no or minimal interference, Medium - a single defender was in close proximity to the shot-taker; High - multiple defenders in close proximity and interfering with the shot).
  - *outcome*: The outcome of the shot (blocked, missed, goal frame (post or bar), saved, goal or own goal).



Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb
Full details of the dataset can be found in the provided documentation, ShotData.txt.

Initial Data Engineering

Required adjustment made to the data before modeling

Pitch Dimension Conversion

- The x-coordinate measures perpendicular distance from the goal line while the y-coordinate measures perpendicular distance from the line that joins the centre spot and the centre of the goal. The full dimensions of the pitch are not confirmed in the documentation but for this analysis, they have been assumed to be 106 meters by 67.84 meters (Fig. A). The min and max of the y coordinates are -33.92 and +33.92 respectively (y = 67.84) and the x coordinates have been assumed be the length of a standard pitch [link], with the penalty spot at (10.97, 0) meters.
- The coordinate system was standardised to match those as used by Laurie Shaw when working with the Metrica Sports data, for simplicity in the following stages (Fig. B). See Laurie's repository [link].

Data cleaning

- Dataset relatively clean but certain attributes required some attention to detail before modelling.
- NULL values for 'Interference_on_Shooter replaced with 'Unknown' string.
- Values in the 'play_type' column tidied including 'Direct Corner' and 'Direct corner'.
- Values for the Body Part assigned the values (Left – 1, Right- 2, Head – 3, Other – 4), as 'BodyPartCode'.
- Attribute created for 'isGoal' – used as the target attribute during the Logistic Regression modeling. 1 is goal, 0 is no goal.

Filtering

- The data was subsetted to only include shots taken from Open Play (OP), removing all shots from Penalties or Free Kicks. There is potential for a separate Expected Goals model to be created for these play types. However, due to time constraints, this was not carried out (see next steps and conclusion).
- Upon investigation, where 'Interference_on_Shooter is 'Unknown', these 43 shots had 100% probability of resulting in a goal. As it is not known what this Interference is and how these could be related in the Metrica Sports data, these were dropped from the dataset.
- This leaves an OP DataFrame with 10,269 shots and 1,218 goals (11.9%), removing 656 shots and 156 goals.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb
Diagram of the pitch post-conversion credit to Laurie Shaw.
Dimension of a standard football pitch: https://en.wikipedia.org/wiki/Football_pitch

A

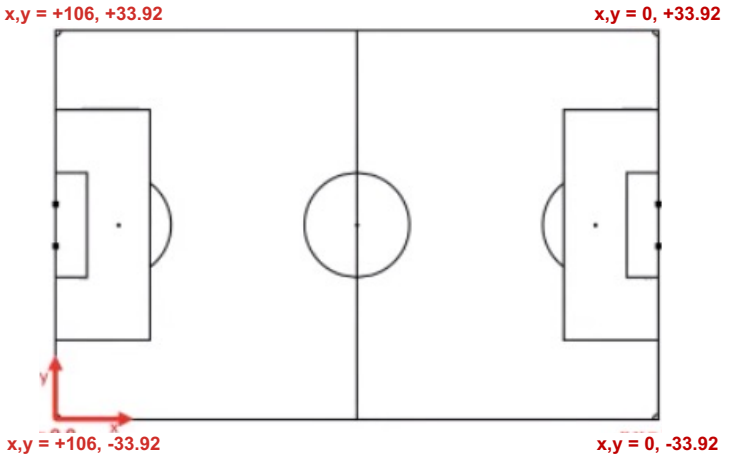


Fig. A: Pre-conversion of coordinates

B

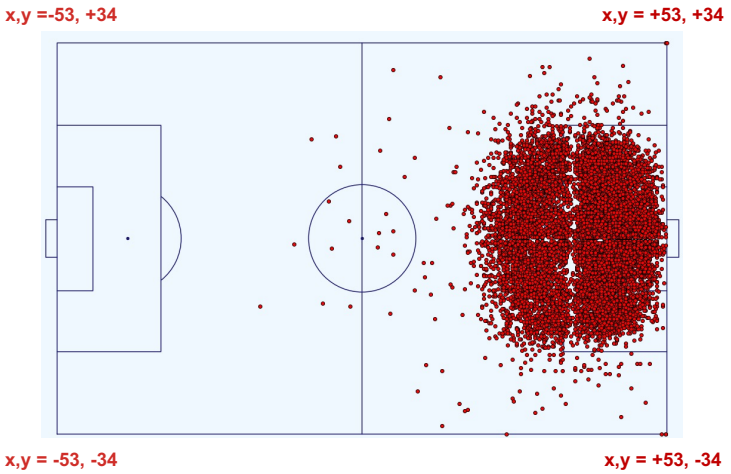


Fig. B: Post-conversion of coordinates including shots

Initial Modelling and Selected Metrics

Creation of a baseline model for which future performance of iterated models was compared

- The first model created as a baseline uses the features first included in the dataset. Future iterations include engineered features and one-hot encoded categorical features.

- The two that are of most interest to start are:
  - The standardised position of the shot was taken on the pitch, 'position_xM' and 'position_yM'
  - The body part the shot was taken with, i.e. left foot, right foot or head, encoded in BodyPartCode

- The dataset was split into a training and test set, with ratios 0.7 and 0.3 respectively. The training set is used to build the model and the test set is used to evaluate the model's predictions.

- The model was the trained using 'position_xM', 'position_yM', and 'BodyPartCode' features, with the newly derived 'isGoal' feature as the target variable.

- The performance of the initial model was the following:
  - Accuracy: 88.5%
  - Log Loss: 0.33182
  - ROC AUC: 72.8%

- As this Chance Quality Model is required to calculate the probability of a shot resulting in a goal (i.e. P[goal|shot,situation]) given a certain state, this model will be assessed by reducing the Log Loss (Binary Cross Entropy). More details about why this is a much more appropriate metric than Accuracy and why the ROC AUC is also not appropriate but is still reported, can be found in section 6 (Metric Definition) of the Chance Quality Model notebook [link].

- At the end of the first basic model, no considerations had been made for: the bias in the underlying data, outlier treatment, feature transformation (one-hot encoding), feature engineering, or univariate and multivariate analysis. Using the first model as a benchmark, the next stages work to improve the model through iteration of each of these steps.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Treating Outliers

Finding and handling unlikely and irregular shots in the dataset that could have an impact on the Chance Quality model

- The provided data included a number of examples of Open Play shots taken and subsequently scored from improbable locations (Fig. A).
- This are most likely from uncharacteristic errors from the goalkeeper and/or poor positioning from the keeper in the current moment in the match i.e. keeper up for a last minute corner and conceding a goal from 50+ meters into an empty goal (see example of Xavi Alonso against Luton [link]).
- These shots took place, but we do not want the model to learn that there is a chance to score from this area in our limited dataset (can be approached using more detailed Event and/or Tracking data).
- For now, these shots are excluded changing the result of all shots with distance >35m and all goals outside the 18-yard-box with shot angle >35 degrees removed from the data.
- This results in 33 Open Play shots values changing from goal (1) to no goal (0) (0.27% of the Open Play shots changed). This is quite an arbitrary rule, with full Event data, more sophisticated outlier removal can take place such as setting the outcome of all shots taking place from areas of the pitch where shots take place >1% of the time, setting these to no goal (Fig. B). More information for such methods can be found in Treating Outliers section (section 7) of the Chance Quality Model notebook [link] and in the appendix.

A

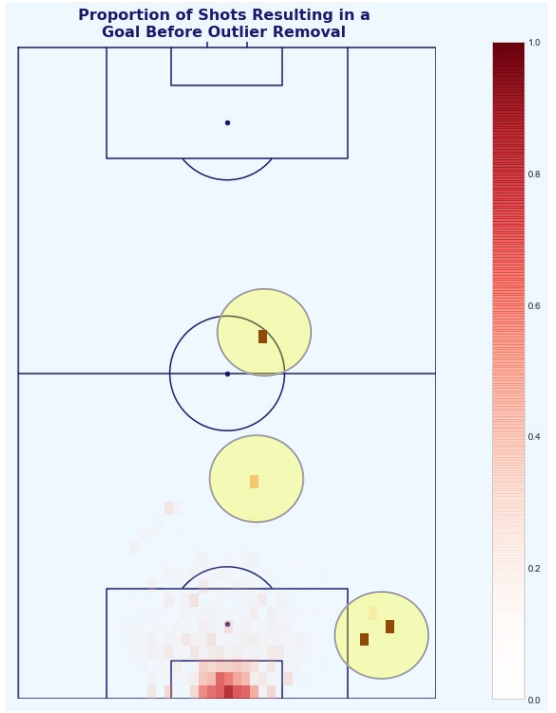


Fig. A: Heatmap of the proportions of shots to goals in the dataset (goals divided by shots). This visualisation flags some of the outlier shots scored from inside the attacking teams half, and at acute angles to the left of the keepers left post.

B

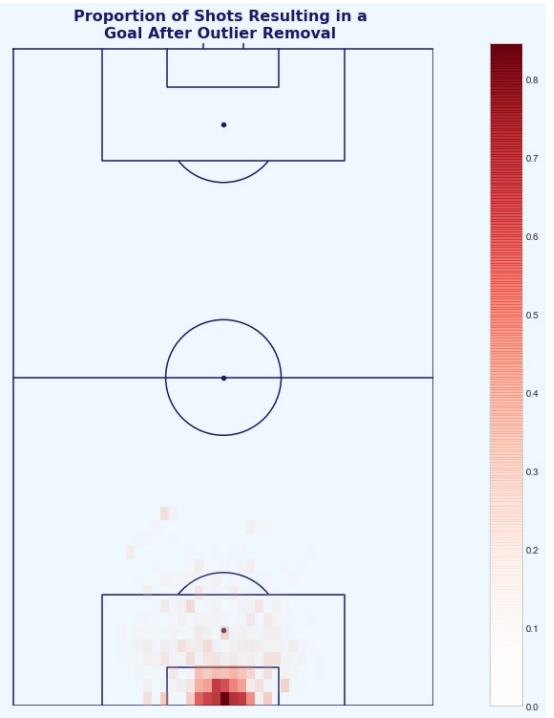


Fig. B: Heatmap of the proportions of shots to goals in the dataset post-outlier treatment. These outliers have now been removed to prevent this affecting the Chance Quality model.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb
Xavi Alonso's 70-yard goal against Luton in the FA Cup is a great example of a player scoring an unlikely goal from their own half with the goalkeeper out of position: https://youtu.be/4OTQwuAc4HU.

Feature Engineering and Univariate Analysis

Creating and analysing the features used in the improved Chance Quality Model.

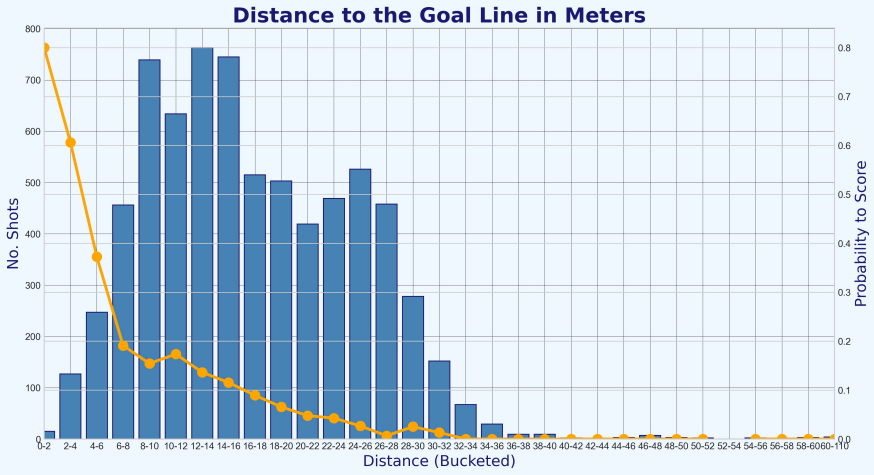
- When selecting features to use for the Chance Quality model, a number of considerations needed to be taken, including:
  1. The feature can be found or engineered in both the Shot data provided and the Metrica Event/Tracking data for which the predictions of chance created are to be made;
  2. The continuous and discrete variables selected are monotonic in nature (e.g. distance to the goal line, number of intervening opponents); and
  3. Categorical features are one-hot (dummy) encoded (e.g. interference on the shooter).

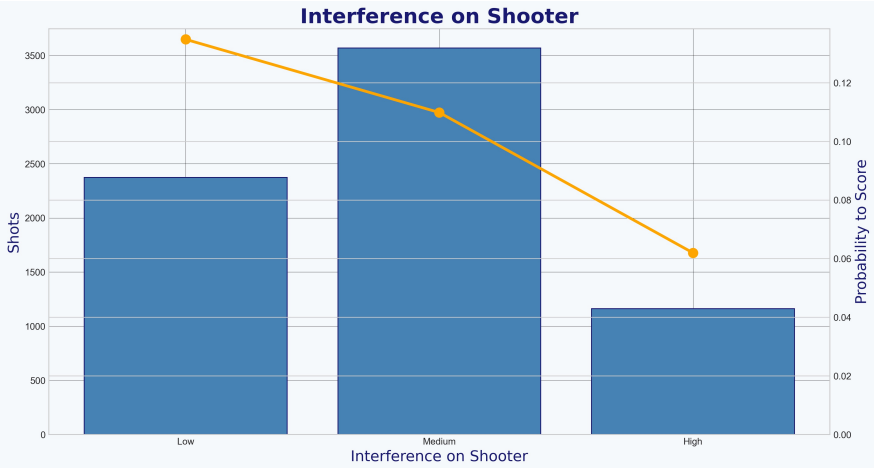
- The features in the data that require no engineering including:
  - Number of Intervening Opponents (taking values between 0 and 11); and
  - Number Intervening Teammates (taking values between 0 and 7).

- Features engineered from the data available to be Logistic Regression friendly:
  - Distance from goal – engineered from both the x and y coordinates coordinate;
  - Distance from the centre of the pitch – engineered from the y coordinate;
  - Angle to the goal – determined using the distance from the goal and centre of the pitch
  - Dummy encoding of categorical features including Body Part to is_foot and is_head, and Interference on the Shooter to Low, Medium and High.

- The visualisations on the right show features that demonstrate monotonic relationships. The further a shot is away from the goal line, the less likely the shot taker is to score. (Fig. A) and the lower the interference level on the shooter, the higher the probability that they are to score from the shot (Fig. B).

- Much more information about analysis each feature for their monotonic relationship can be found in Univariate Analysis (section 8) of the Chance Quality Model notebook [link]. Further details of the all the features engineered with the available data, and features that I would lhypothetically engineer from a full Event dataset can be found in the appendix.





Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Multivariate Analysis

Dealing with highly correlated features

- At this stage in the analysis, it's important to not just look at each feature individually, but how they interact with each other. Multivariate analysis is all about finding highly correlated features and thinking about how to deal with them. This is important for two reasons:
  - Highly correlated features might lead to overfitting and have a negative effect on results; and
  - Highly correlated features can very easily mess with the feature interpretation and importance.

- The visualisation on the right is a final heat map correlation matrix, with the features that were used in the final model. Highly correlated features such as angle and distance_to_centerM, and isHead and header_distance_to_goalM were treated appropriately, remove one of each. Medium_Infererence_on_the_Shooter was also dropped from the final model, as when one-hot encoding a categorical column, you should always delete one of the created columns.

- The Number of Intervening Opponents and the Number of Intervening Teammates are highly correlated features, but are features that were provided in the raw dataset that I can probably safely assume are not derivatives of each other, and will therefore be left untouched.

- The performance of the final model was the following:
  - Log Loss: 0.289 (from 0.332)
  - ROC AUC: 80.7% (from 72.8%)

- More information about the multivariate analysis can be found in section 9 of the Chance Quality Model notebook [link].

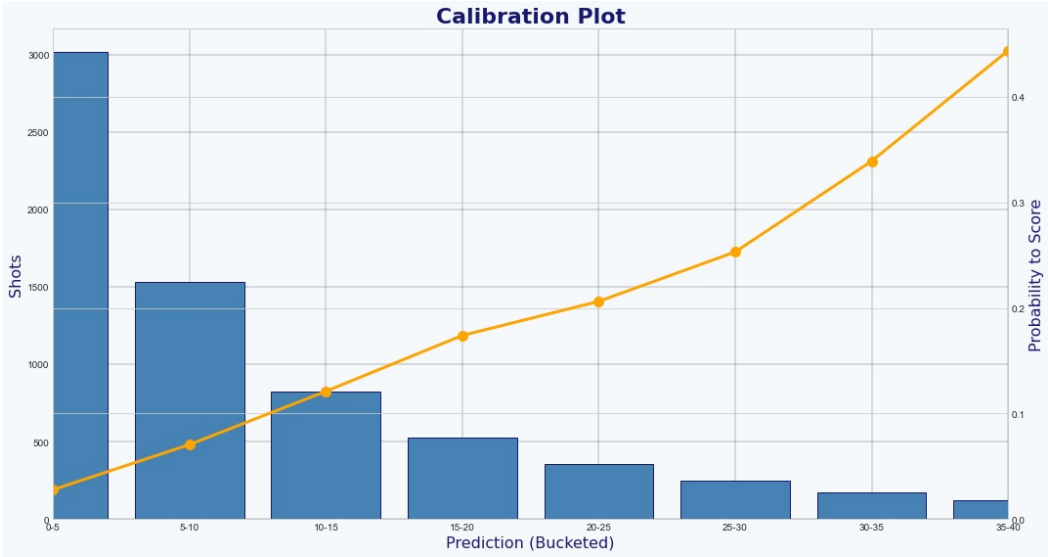
Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Final Model Evaluation and Feature Interpretation

Analysis of the final model before being used for prediction on the Metrica Sports data

- The purpose of this Chance Quality Model is to calculate probability of a shot resulting in a goal (i.e. P[goal|shot,situation]). From this, it is possible to determine the probability that a shot results in a goal.
- The main focus of optimising the model was to focus on minimising the Log Loss, but the important outcome is to see if the model probabilities correctly. This can be visualised through Calibration Plots (Fig. A).
- The observed likelihood of the shots in the dataset resulting in goals is between the 0% and 5% bucket. The model is well-calibrated and ready to use with the Metrica Sports shot data.
- Through the Feature Interpretation, it is possible to analyse the coefficients of the features, by both sign and magnitude (Fig. B). The take home messages from the model are:
  - Position on the field is extremely important when shooting. The further away the shot, the less likely it is to result in a goal;
  - This distance to the goal line is super important, and even more so when taking a header; and
  - The further you are from the centre of the field, the more unlikely to score.
- More information about the Final Model and Feature Interpretation can be found in sections 11 and 12 respectively of the Chance Quality Model notebook.

A



B

```
Coefficients
----------------------------------------
distance_to_goalM: -1.072
angle: -0.220
Number_Intervening_Opponents: -0.512
Number_Intervening_Teammates: 0.048
isFoot: -0.131
High: -0.328
Low: 0.082
header_distance_to_goalM: -0.638
```

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Part 2: Application of Tracking Data

Using Metrica Sports Tracking and corresponding Event data with sample game 2

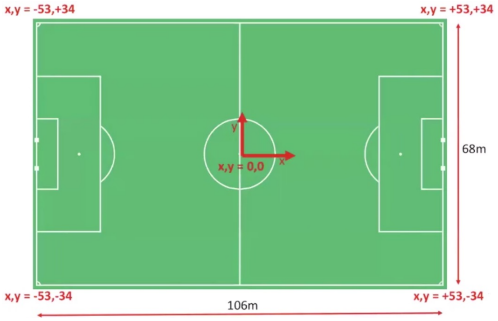
Metrica Sports Tracking and Event Data

Key objectives achieved with this data, explained further in this section

1

Data Engineering

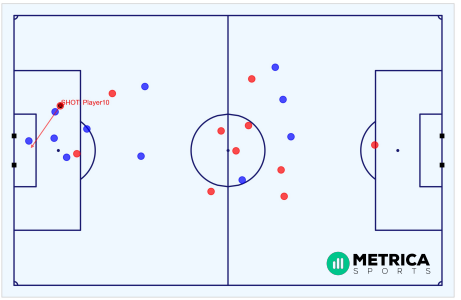
Including converting the pitch coordinates to a standardised coordinate system, subsetting home and away DataFrames, reversing the direction of players to ensure each team is in the same direction for the full 90 minutes, and determining each of the player's positions, speed, acceleration, and movement at a defined moment, using the timestep and distance travelled to calculate the the relative positions, speed, and acceleration of players for a given moment/sequence in play.



2

Exploratory Data Analysis of the match

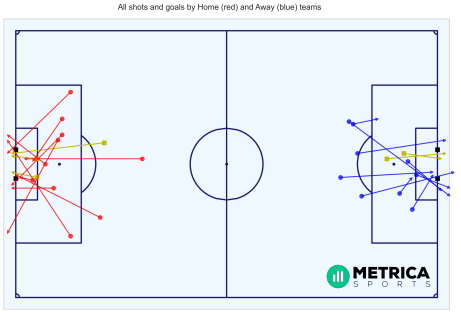
Visualisation of the 24 shots taken during the match by the home and away teams, including Tracking data visualisation to conduct an eye test of the key chances and goals.



3

Create Features required for predictions to be made for the data using the trained Chance Quality model

These features include: distance to the goal; distance to the centre of the pitch; angle to the goal; number of intervening opponents; number of intervening team mates; interference on the shooter; whether the shot resulted in a goal; whether the shots was a penalty or direct free kick; and whether the shot was taken with the player's foot or head. This data is then exported



4

Export data

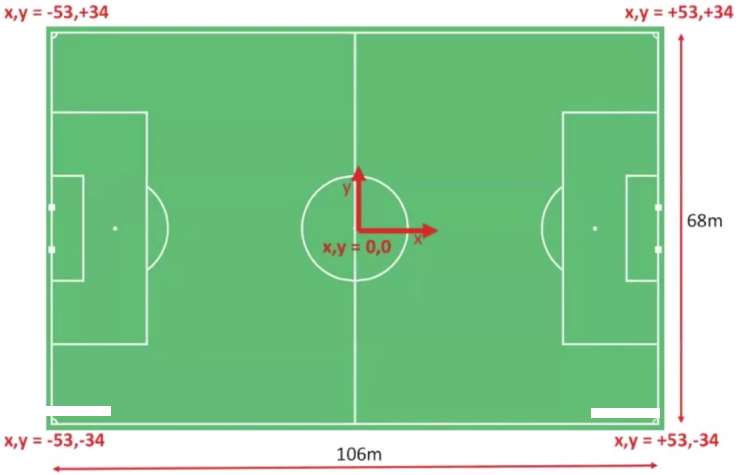
Data converted to a format that matches the provided Shots data and exported, ready to be analysed and predicted upon for an assessment of the attacking chances during the match.

Notebook to work with the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb

Initial Data Engineering

Required adjustment made to the data before modeling

As per the shots Data, the Metrica Sports Event and Tracking data required some steps of Data Engineering before being in a state ready for analysis.



Pitch dimension conversion

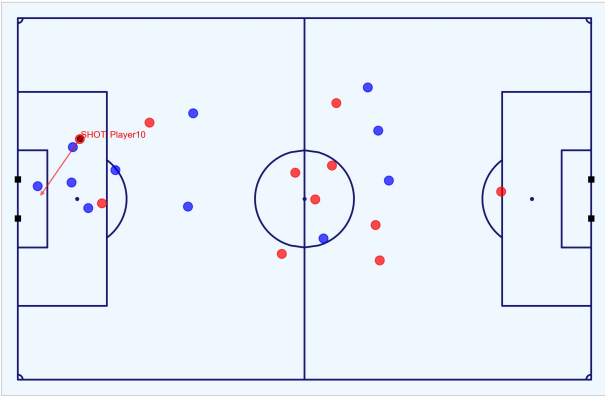
- Converted Metrica's 0-1 coordinate system to a standardised 106 by 86 coordinate system, as previously conducted with the Shots data, using the to_metric_coordinates function from the Metrica IO (mio) library created by Laurie Shaw [link]. This is done for both the Event and Tracking data.

Reverse player direction

- Reversal of the direction of the home and away players to ensure that each team players in the same direction for the full 90 minutes,

Determine each of the player's positions, speed, acceleration, and movement

- With very little data engineering, it is possible to determine the position, speed, acceleration, and direction for all of the players on the field at any given moment.
- Metrica Sports Tracking data is collected at twenty five frames s^{-1}. Therefore, a player's speed can be calculated by dividing the distance a player has covered between two frames by 0.04. The acceleration is calculated as the 2nd order derivative i.e. speed divided by dt = 0.04 again.
- Velocities and accelerations determined using the Metrica Velocity (mvel) library by Laurie Shaw [link].



Subset DataFrames

- Separate home and away DataFrames and also create a separate shots DataFrame.

Notebook to work with the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data
Diagram credit to Laurie Shaw.

Exploratory Data Analysis of the Match

Determine the outcome of the match from the Event data and visualise the chances in the Tracking data

- For the Tracking data, the following shots and goals can be observed:
  - Fig. A visualises all the shots and goals for both the Home (red) and Away (blue teams). The shots are in the respective team colours and chances that resulted in a goal are in yellow.
  - Fig. B visualises the first goal scored by the Home team at 8m8s (1-0).
  - Fig. C visualises the second goal scored by the Away team at 35m22s (1-1).
  - Fig. D visualises the third goal (header) scored by the Home team at 49m19s (2-1).
  - Fig. E visualises the forth goal (penalty) scored by the Away team at 76m40s (2-2).
  - Fig. F visualises the fifth goal scored by the Home team at 80m41s to win the game (3-2).
- From the data, we can see that the Home team won the match 3-2.

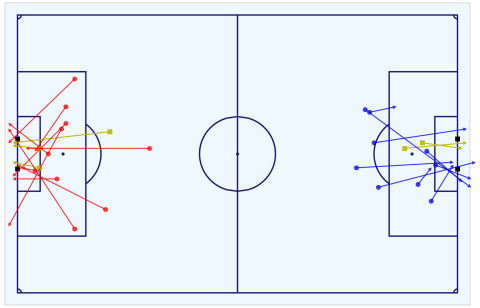


Fig. A: All shots and goals

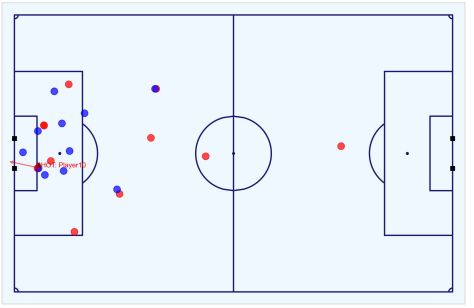


Fig. B: 1-0 Home goal (8m8s)

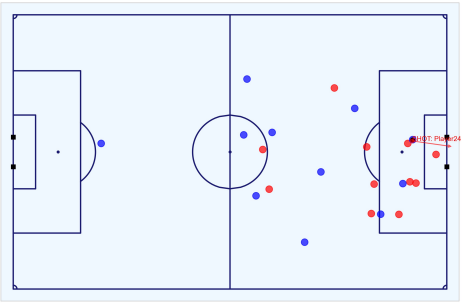


Fig. C: 1-1 Away goal (35m22s)

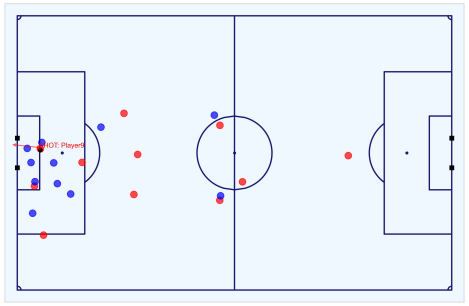


Fig. D: 2-1 Home goal (49m19s)

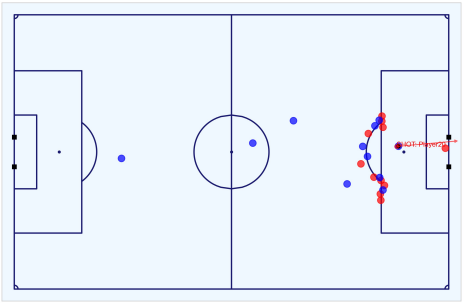


Fig. E: 2-2 Away penalty goal (76m40s)

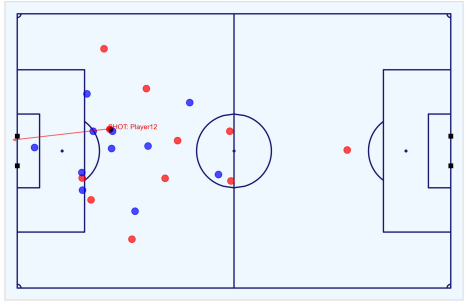


Fig. F: 3-2 Home goal (80m41s)

Notebook to work with the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb.
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data.
All PNG figures produced from Tracking data: https://github.com/eddwebster/mcfc_submission/tree/master/img/fig/metrica-sports.
All MP4 videos produced from Tracking data: https://github.com/eddwebster/mcfc_submission/tree/main/video/fig/metrica-sports.

Creating Additional Features Required for Compatibility with CQM

Features derived from the Tracking data

- The be compatible with the trained Expected Goals model from the Shots data, a number of key features were required to be derived from the Metrica Sports Event and Tracking data including: distance to the goal; distance to the centre of the pitch; angle to the goal; whether the shot resulted in a goal; whether the shots was a penalty or direct free kick; whether the shot was taken with the player's foot or head; the interference on the shooter; and the number of intervening opponents and team mates.

- From the Tracking data, it was possible to derive both the interference on the shooter and the number of intervening opponents and team mates:
  - The interference on the shooter can be determined by calculating the the distance between the player and the ball and whether or not this is less than the defined radius. For this analysis, a radius of 5-yards (4.572m) as is used by StatsBomb to define pressure events [link].

  - The number of intervening opponents and team mates can be determined by calculating the area of the four triangles alternating between the player, the ball, and each post (Fig. B). When the area of the triangle between both posts and the ball is equal the the triangles between the player-ball-left post, player-ball-right post and player-left post-right post, the observed player lies within this triangle and is therefore intervening in play. The number of players can then be summated.

- The number of players determined as interfering or interfering is then summated at a team level and included in the subsetted Metrica Sports Shots DataFrame.

Notebook to work with the Metrica Sports Event and Tracking data:
https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data
How StatsBomb Data Helps Measure Counter-Pressing: https://statsbomb.com/2018/05/how-statsbomb-data-helps-measure-counter-pressing/

A

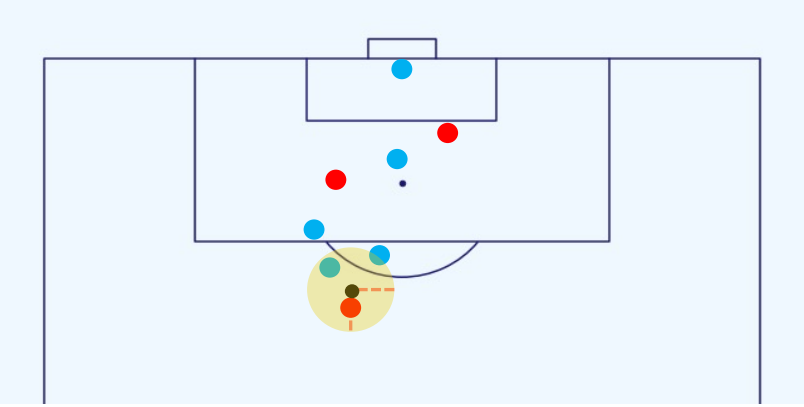


Fig. A: A player is defined as interfering with the shooter if they are within a defined radius of the ball, in this case 5m (not to scale). This can be determined mathematical by observing whether the distance between the player and the ball is less than the defined radius, or not.

B

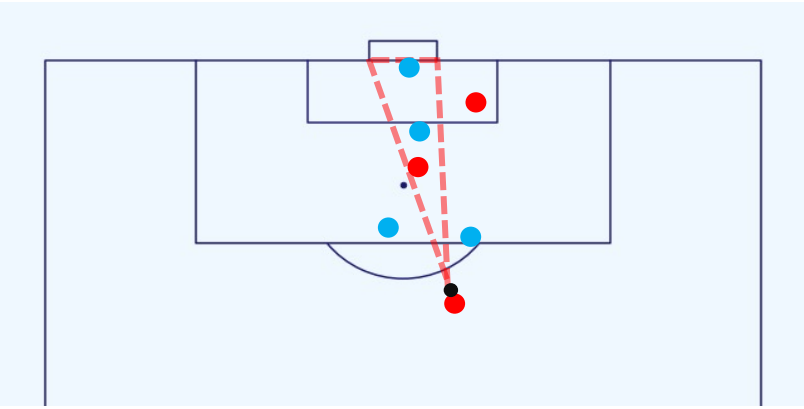


Fig. B: The number of intervening team mates and opponents can be visualised as a triangle between the ball and the posts of the goal at the moment the shot is taken. This interference can be determined mathematical by summating the areas of each triangle.

Part 3: Application of the Trained Chance Quality Model with the Metrica Sports Shot Data

Finalised dataset ready for analysis

Which Team Deserved to Win the Game?

Assessment of the chances in the match through the application of the Chance Quality Model

- As was discussed in the Data Engineering section, the shots in this model were filtered for only Open Play (OP) shots. However, game 2 of the Metrica Sports data includes a penalty.
- To deal with this, any penalty shots in the Metrica Sports data are assigned the value of 0.76 xG. as per the value used by StatsBomb/FBref [link]. StatsBomb models penalty kicks as all sharing the same characteristics and this assumption is made in this model. This amendment changes what is a very tight contest regarding cumulative xG (Fig. A) into a clearer projected win for the Away team (blue), once assigned this new xG value for the penalty.
- When taking all the shots and sorting by the highest xG value, for the top five shots, only three were scored. The two goals outside this top five are shots by the Home team, with xG values of 0.235 and 0.054 respectively.
- It is the winning goal scored by the Home team in the 80th minute (shot 20) with the xG value of just 0.054 which is probably the biggest contributing reason why the Home team won the game, despite having less accumulated xG than the Away team. The Home team did however miss the biggest non-penalty chance in the game, a shot with an xG value of 0.450 (shot 16).
- The analysis summarises to say that arguable, the Away team were more deserving to win the game when considering only the quality of chances that each team created when applying the Chance Quality Model, despite losing the game three goals to two.

A

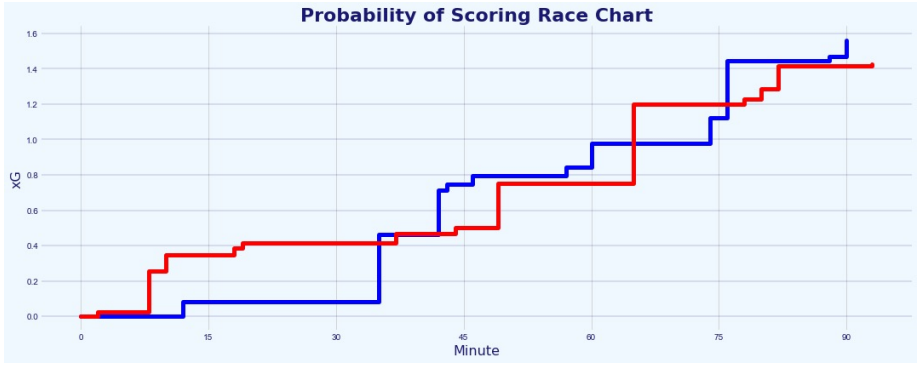


Fig. A: Probability of Scoring race chart before any amendments to xG values

B

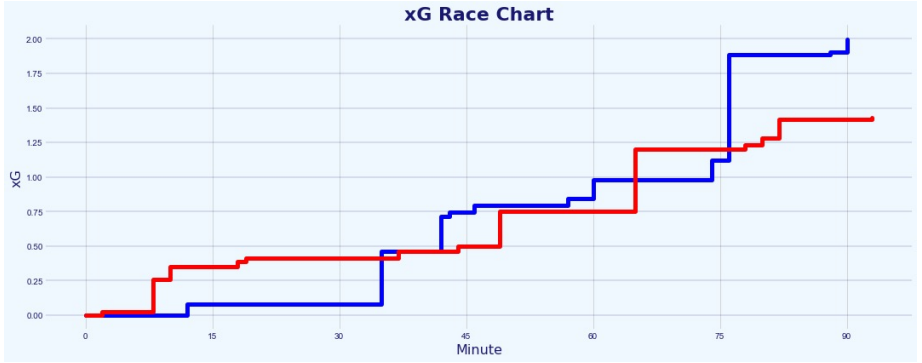


Fig. B: xG race chart with amended xG value for the penalty

Analysis of game 2 of the Metrica Sports sample data can be found in section 14 and 15 of the Chance Quality Model notebook: https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

For more examples of visualising xG Race Charts in Tableau, please see slide 23 the Football Intelligence pack: https://docs.google.com/presentation/d/1uT6vT1J_FS10hW3hvWQx9ofqHUq8mT-kUbDRMrrUqH8/edit?usp=sharing

How to Create xG Flow Charts in Python by McKay Johns: https://www.youtube.com/watch?v=bvoOOYMQkac&list=PL10a1_q15HwqVEcnqt3tXs1bgvawjsQNW

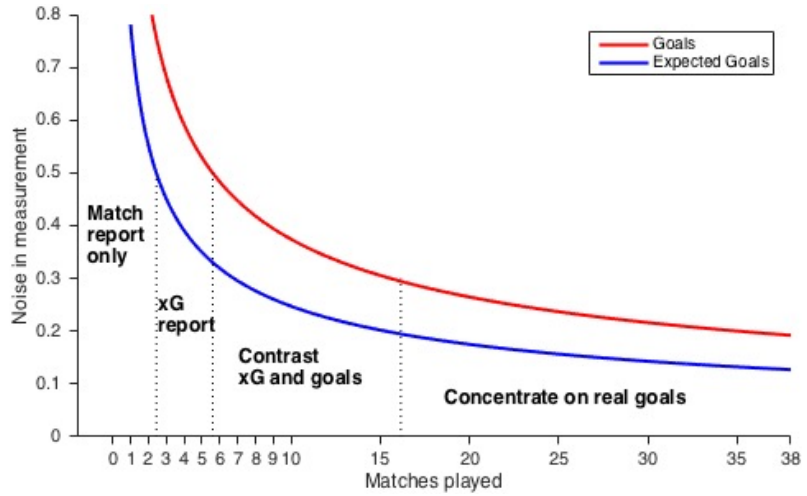
xG explained by FBref (include penalty xG of 0.76): https://fbref.com/en/expected-goals-model-explained/

With Great Power Comes Great Responsibility

Expected Goals are useful, but when used incorrectly or in isolation, can also be useless

- In his piece 'Should you write about real goals or expected goals?', David Sumpter discusses how regarding Expected Goals, how the magnitude of the noise in the measurement performance decreases with the number of matches played in season (Fig. A).
- Between 1 and 2 matches, the noise is high for both Expected Goals and real goals. Between 3 to 6 matches, the noise is less than 0.5 goals per match, making an xG report quite functional. Between 7 to 16 matches, both goals and xG have only between 0.3-0.5 goals of noise which allows for comparison. After 16 matches, the difference between the noise in xG and the noise in real goals is only 0.1 goal per match. At this point, real goals are better than Expected Goals.
- For this reason, even though the Chance Quality Model suggests that the Away team is a more deserving winner of the match by proxy of a greater accumulation of xG, any analytical insights and conclusions based over a single 90 minute period should be made with caution.
- This poor usage of Expected Goals can be found in a variety of forms of media (Fig B, C, and D), where there has been an unfortunate tendency to misuse the metric over individual matches, instead of over several games. I would therefore conclude this analysis to say that predictions of any one team's performance over a single 90 minutes is difficult and will be in most cases, inconclusive at best.

A



B

C

D

Analysis of game 2 of the Metrica Sports sample data can be found in section 14 and 15 of the Chance Quality Model notebook: https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

See David Sumpter's article: Should you write about real goals or expected goals? A guide for journalists: https://soccermatics.medium.com/should-you-write-about-real-goals-or-expected-goals-a-guide-for-journalists-2cf0c7ec6bb6

Caley graphic for Liverpool vs. Athletico Madrid (12/03/2020): https://twitter.com/Caley_graphics/status/1237870631024095234

The xG Philosophy tweet for Juventus vs. Porto (09/03/2021): https://twitter.com/xGPhilosophy/status/1369418380642549762

Next Steps and Conclusion

Next Steps to Improve the Chance Quality Model

Some of the steps I plan/wish to take in the near future

- The focus of my approach to answer this data challenge was to not build the absolute best performing ML model, with the best performance metrics and fanciest algorithm. My objective was to conduct and end-to-end process for building a model, including all the key stages such as feature engineering, univariate and multivariate analysis, and iterated performance assessment and improvement – for this reason, Logistic Regression was chosen. However, Gradient Boosting algorithms, such as XGBoost, LightGBM, and CatBoost, would all most likely lead to a slightly improved performance and reduced Log Loss. This are algorithms that I intend to work with after the initial submission of this analysis. More detail about these modeling approaches can be found in my the Data Science pack that I submitted as part of my initial application (see: https://docs.google.com/presentation/d/16stYbJoI8aYqtn_grJHSdTbMA1xwo-Iy9g9JJruMOBQ/edit?usp=sharing).

- Application of a full Event data set, such as those from StatsBomb and Wyscout, to create an Expected Goals model with more features. Such features that were not possible to include in this model but that could be added with Event and/or Tracking data include: strong/weak foot, flag for counter attack, flag for smart pass, determine whether a shot had been immediately taken before, whether the shot was from a cross. This is discussed in more detail in the Feature Engineering section (section 9) of the Chance Quality Model notebook [link]. A comparison of the features used in the respected xG models by Sam Green and Michael Caley can be found in the index.

- To see really test the performance of this model, it would be great to quality test performance and xG prediction with those of other providers such as StatsBomb and observe the level of variance in predictions between this basic model and a professional one created using a much larger dataset with much more features.

- Creation of separate Expected Goals models for Direct Free Kicks and Corners. Currently, only Open-Play shots considered and a xG value for penalties was taken from StatsBomb/FBref [link].

- Add fake shots to the shots data – see David Sumpter's tweet for the benefits of including fake data in an Expected Goals model [link].

All code for this challenge: https://github.com/eddwebster/mcfc_submission/tree/master/notebooks/.
Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Next Steps to Improve the Tracking Data Analysis

Some of the steps I plan/wish to take in the near future

- This current analysis only focuses on the shots taken by the teams. The next stage of this analysis would be to apply this Tracking data, not to just a Shots dataset, but to a full Events dataset, taking the basic concepts of analysis and feature extraction observed in this submission and then start to apply more sophisticated modeling approaches such as Pitch Control or Expected Possession Value (EPV) models such as the VAEP model by SciSports and KULeuven, or the Expected Threat (xT) model by Karun Singh. This can be taken on further, by combining these two modelling approaches to analyse value that certain actions of interest brought to the team during a particular play in the match and determine the Expected Value-Added. This was unfortunately not possible to do in this analysis as the Event data provided only included Shot data, but it would be something I would like to take on and do in the future, using publicly available Event data from StatsBomb and Wyscout, with the sample Tracking data from Metrica Sports. More detail about these models can be found in my the Data Science pack that I submitted in my initial application (see: https://docs.google.com/presentation/d/16stYbJoI8aYqtn_grJHSdTbMA1xwo-Iy9g9JJruMOBQ/edit?usp=sharing).

- Further enrich the Event data through Tracking data, adding further detail and specificity, which again can be used to further improve the Expected Goals model. This was observed in this analysis with addition of the Intervening and Interfering teammates and opponents. Features that were not considered in this analysis, include aspects such as the goalkeeper and defender positions in the moment of the shot e.g.: how much of the goal was covered by the goalkeeper? are the defenders in position? These are attributes that can be derived from the Tracking data to gain additional insight previously not possible.

All code for this challenge: https://github.com/eddwebster/mcfc_submission/tree/master/notebooks/.
Notebook to work with the Metrica Sports Event and Tracking data: https://nbviewer.jupyter.org/github/eddwebster/mcfc_submission/blob/main/notebooks/metrica-sports/Metrica%20Sports.ipynb
Sample Tracking data provided by Metrica Sports: https://github.com/metrica-sports/sample-data

Conclusion

Summary of this data challenge submission

1. Defined, built, trained, and evaluated a Chance Quality Model with Shots data, using Expected Goals and Logistic Regression.

2. Determined the shots from game 2 of the sample Metrica Sports Tracking and Event data and made these available for predictions using the Chance Quality Model.

3. Assessed the teams in game 2 of the sample Metrica Sports data through application of the Chance Quality Model (derived in step 1) and exported Metrica Sports shot data (derived in step 2), to determine that based on chances created and the Expected Goals predicted, the Away team was more deserving of being the winner of the match, despite losing the game three goals to two to the Home team. However, as this is over just a 90 minute period, this analysis should be treated with caution.

For More Information

If you would like to find out more…

- **GitHub:**
  - General: github.com/eddwebster
    - Manchester City Junior Data Scientist submission: github.com/eddwebster/mcfc_submission
    - Football analytics: github.com/eddwebster/football_analytics
- **Google Drive of all code, data, visualisations, and analysis in this pack:** https://drive.google.com/drive/folders/1Ts_5YOL8JpVpRcaEhqmcTCceLmcBRY3L?usp=sharing
- **Slide decks:**
  - Junior Data Science Challenge: https://docs.google.com/presentation/d/116D0U_ue2sv6hgLBgHnil228cRhph0qfMuOFA4uuxhs/edit?usp=sharing
  - Data Science pack (initial submission for Junior Data Scientist position): https://docs.google.com/presentation/d/16stYbJoI8aYqtn_grJHSdTbMA1xwo-Iy9g9JJruMOBQ/edit?usp=sharing
  - Football Intelligence pack (previously submitted for Football Intelligence Analyst position): https://docs.google.com/presentation/d/1uT6vT1J_FS10hW3hvWQx9ofqHUq8mT-kUbDRMrrUqH8/edit?usp=sharing
- **Tableau Public profile:** public.tableau.com/profile/edd.webster.
- **Website:** eddwebster.com
- **LinkedIn:** linkedin.com/in/eddwebster
- **Twitter:** @eddwebster
- **Email:** edward.webster@cityfootball.com and edd.j.webster@gmail.com

Appendix

Treatment of Outliers

Full progression of handling unlikely and irregular shots in the dataset

A

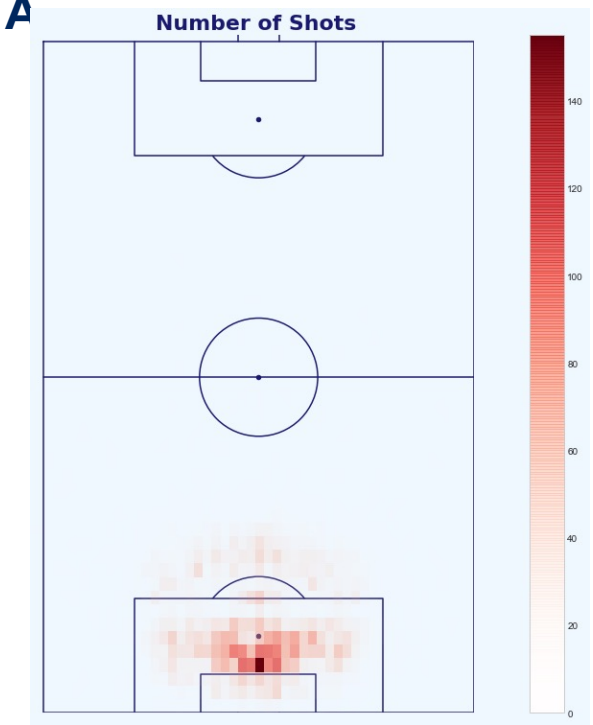


Fig. A: Heatmap of shots in the dataset, the center of the goal outside the 6-yard-box the most frequent.

B

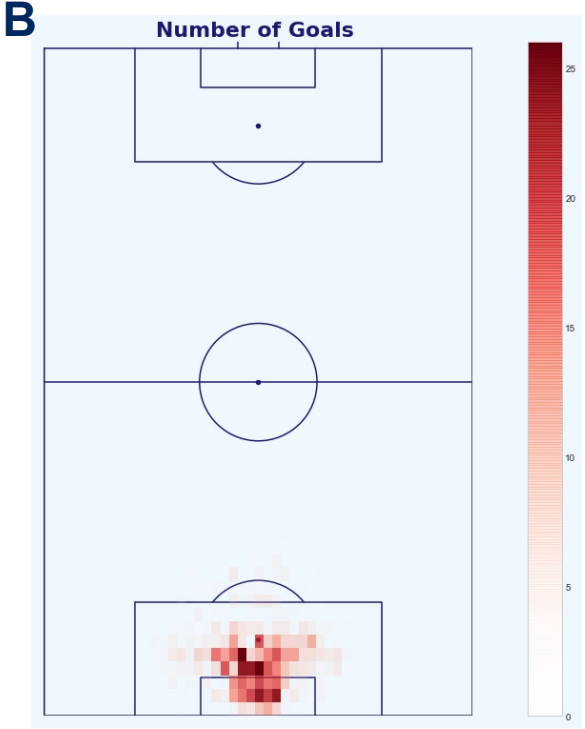


Fig. B: Heatmap of goals in the dataset, most scored within the 6-yard/18-yard box.

C

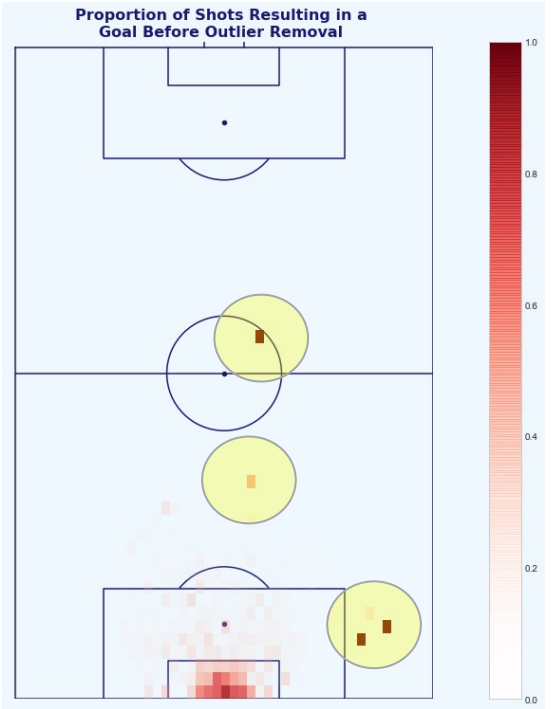


Fig. C: Heatmap of the proportions of shots to goals in the dataset (goals divided by shots). This visualisation flags some of the outlier shots scored from inside the attacking teams half, and at acute angles to the left of the keepers left post.

D

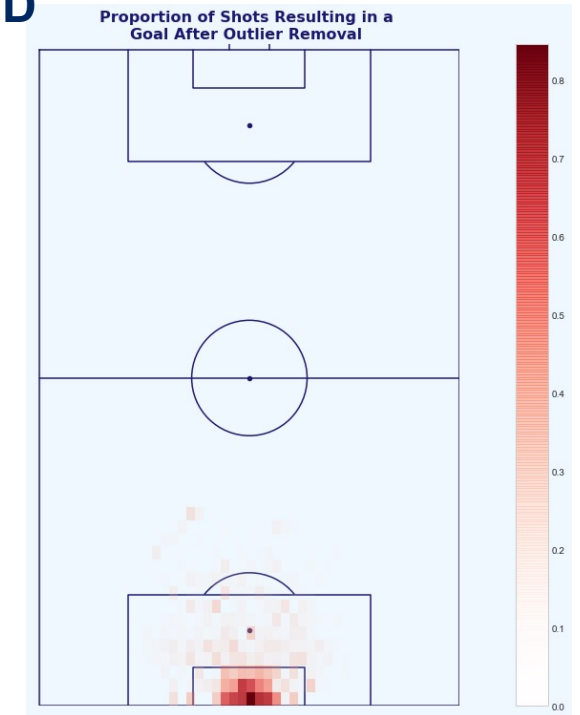


Fig. D: Heatmap of the proportions of shots to goals in the dataset post-outlier treatment. These outliers have now been removed to prevent this affecting the Chance Quality model.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Features of Final Model

Description and status of usage of the Chance Quality model when applied to the Metrica Shot data

Feature	Used in Final Model	Description
distance_to_goalM	Y	Continuous feature that determines the distance the shot was taken along the y-axis in relation to the goal, in meters.
distance_to_centerM	N	Continuous feature that determines the distance the shot was taken along the x-axis in relation to the center of the pitch, in meters.
angle	Y	Continuous feature that determines the angle in which the shot was taken to the goal.
number_intervening_opponents	Y	Continuous feature that determines the number of opposing players that were obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
number_intervening_teammates	Y	Continuous feature that determines the number of teammates that are obscuring the goal at the instant of the shot (from the perspective of the shot-taker).
is_foot	Y	Boolean feature that indicates whether the shot was taken with the player's foot, or not.
is_head	N	Boolean feature that indicates whether the shot was taken with the player's head, or not.
is_high_interference	Y	Boolean feature that indicates whether the shooter experience a High level of interference (multiple defenders in close proximity and interfering with the shot).
is_medium_interference	N	Boolean feature that indicates whether the shooter experience a Medium level of interference (a single defender was in close proximity to the shot-taker).
is_low_interference	Y	Boolean feature that indicates whether the shooter experience a High level of interference (no or minimal interference).
header_distance_to_goalM	Y	Continuous feature that determines the distance that a headed chance was taken from.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Wishlist Features with Complete Event Data

The features I would like to include in an a Chance Quality Model given a full Event dataset

Feature	Description
is_strong_foot	Categorical feature that determines whether or not the shot was taken with the player's preferred foot. This can be determined from player bio data, or, it by taking a full set of Event data, analysing how many actions were done per foot in a match, per player, and assigning the most used foot as the preferred foot. From the data, we can then compare that player's preferred foot to the foot with which each shot was taken and determine whether it was taken with their strongest foot, using a Boolean is_strong_foot attribute, and also a is_weak_foot attribute.
is_counter_attack	Boolean feature to indicate if the shoot was part of a counter attack or not. This can be determined using the position on the pitch from which a ball was won and the number of opposing defenders behind the ball, relative to the attack team.
is_smart_pass	Boolean feature to indicate whether the assist for the shot broken through the opponents line.
is_from_cross	Boolean feature to indicate whether a goal was scored from a cross.
time_from_previous_shot	Time in seconds from the last shot of the same team in the same half of the same game. This can be taken on further to determine whether a shot was taken from before and is likely a rebound i.e. is_shot_before or is_rebound. This is of interest as it can determine whether the goalkeeper is out of position and making a reflex save, with the subsequent shot being in a state different to that of a normal shot in that position.

Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Features Used in Notable Expected Goals Models

Key features as used in the highly regarded xG models by Sam Green and Michael Caley

Sam Green's original model with Opta:

- Distance - distance to the middle of the goal (the mid-point between the goalposts).
- Visible angle of the goal - the angle formed between the shot location and the two goal posts.
- Passage of play - one of open play, direct free kick, set play, corner kick, assisted, and throw-in.
- Assist type - one of a long ball, cross, through ball, danger-zone pass, and pull-back.
- Post take-on/ dribble - whether the shot follows a previous attempt to beat a player.
- Rebound - whether the shot follows a previous shot that has rebounded.
- Header - whether the shot came off the attacking player's head.
- 1 versus 1 - a shot where there is just one defensive player to score past.
- Big chance - a situation where a player should reasonably be expected to score, usually in a one on one scenario or from very close range when the ball has a clear path to goal and there is low to moderate pressure on the shooter (Opta specific).

Caley's model:

- Fast break - an attempt created after the defensive quickly turn defence into attack winning the ball in their own half.
- Counterattack – an engineered feature to capture counterattacks that are not marked as fast breaks by Opta's coders. Defined as actions that begin with an open play turnover of possession, in which the attacking team moves steadily forward to the goal without recirculating the ball.
- Established possession - an engineered feature that is defined as "an attack that involves at least five completed passes in the attacking half without the ball being forced back into the defensive zone."
- Relative angle to the goal - the angle to the nearest post. If a player is in a central position, the angle is 1. If a player is at a 45-degree angle to the nearest post, the angle is 0.5.
- Interaction between the distance and angle - an interaction that captures interactions between distance and angle to the goal. The distance to the goal multiplied by the relative angle to the goal.
- Dribble distance – the distance a player has dribbled before taking the shot.
- Error - whether the shot follows an error by another player.
- Body part - the body part used to take the shot.
- Game state – the game state is a feature that describes whether the team taking the shot is losing, drawing, or winning the match at the time of the shot.
- League - a feature for the league, for example, the Bundesliga or the English Premier League.

Sam Green's xG model: https://www.optasportspro.com/news-analysis/assessing-the-performance-of- premier-league-goalscorers/
Michael Caley's xG model: https://cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methoderology

Performance of the Model

Iterations of the model during the modeling process

Model	Log Loss	ROC AUC	Accuracy*
Initial	0.33182	72.8%	88.5%
Outlier Removal	0.32670	72.8%	-
Univariate Analysis	0.28979	80.7%	-
Multivariate Analysis	0.28892	80.7%	-
Final Model	0.28924	80.6%	-

* The Accuracy metric not included most the initial model as it is explained in further detail in the Chance Quality Model notebook, that this is not an appropriate metric to measure performance of a probability model.
Notebook to create the Chance Quality Model from the shots data:
https://github.com/eddwebster/mcfc_submission/blob/main/notebooks/chance_quality_modelling/Creating%20a%20Chance%20Quality%20Model%20from%20Shots%20Data.ipynb

Exported Metrica Sports data

Extract of the 24 shots exported from game 2 of the Metrica Sports data, including added features

match_minute	match_second	outcome	position_xM	position_yM	distance	angle	distance_to_centerM	isGoal	isFoot	isHeader	team	period	start_frame	end_frame	Number_Intervening_Opponents	Number_Intervening_Teammates	Interference_on_Shooter	interference_on_shooter_number_players	subtype	isPen	isDirectFK
2	57	Missed	39.22	18.36	22.9560014	53.1101352	18.36	0	1	0	Home	1	4419	4443	3	0	Medium	1	OFF TARGET-OUT	0	0
8	8	Goal	47.7	-3.4	6.2968246	32.6805547	3.4	1	1	0	Home	1	12202	12212	1	0	High	2	ON TARGET-GOAL	0	0
10	59	Saved	41.34	11.56	16.4191717	44.7532497	11.56	0	1	0	Home	1	16484	16499	2	0	Medium	1	ON TARGET-SAVED	0	0
12	21	Blocked	31.8	10.2	23.5261557	25.6936997	10.2	0	1	0	Away	1	18515	18520	1	1	Low	0	BLOCKED	0	0
18	14	Missed	43.46	-6.12	11.3342843	32.6805547	6.12	0	0	1	Home	1	27345	27360	1	1	High	2	BLOCKED	0	0
19	50	Saved	31.8	-13.6	25.1872984	32.6805547	13.6	0	1	0	Home	1	29754	29777	3	1	Low	0	HEAD-OFF TARGET-OUT	0	0
35	22	Goal	44.52	2.72	8.90554883	17.7838884	2.72	1	1	0	Away	1	53049	53075	1	0	Medium	1	ON TARGET-GOAL	0	0
37	23	Missed	42.4	6.12	12.2398693	30.0003533	6.12	0	0	1	Home	1	56079	56131	2	0	Low	0	HEAD-OFF TARGET-OUT	0	0
42	14	Missed	45.58	0.68	7.45109388	5.23619982	0.68	0	0	1	Away	1	63362	63372	1	0	Medium	1	OFF TARGET-HEAD-OUT	0	0
43	11	Saved	28.62	-3.4	24.6159379	7.93918296	3.4	0	1	0	Away	1	64772	64799	3	1	Medium	1	ON TARGET-SAVED	0	0
44	43	Missed	39.22	-18.36	22.9560014	53.1101352	18.36	0	1	0	Home	1	67067	67107	2	0	Medium	1	OFF TARGET-OUT	0	0
46	35	Missed	30.74	10.88	24.7766422	26.0479553	10.88	0	1	0	Away	2	69887	69927	2	1	Low	0	OFF TARGET-OUT	0	0
49	19	Goal	47.7	1.36	5.47170906	14.3917978	1.36	1	0	1	Home	2	73983	73995	2	0	Medium	1	HEAD-ON TARGET-GOAL	0	0
57	28	Missed	32.86	2.72	20.3228443	7.6915211	2.72	0	1	0	Away	2	86191	86219	4	2	Low	0	OFF TARGET-OUT	0	0
60	7	Saved	46.64	-11.56	13.1940593	61.1815963	11.56	0	1	0	Away	2	90165	90176	2	0	Low	0	ON TARGET-SAVED	0	0
65	55	Saved	48.76	-4.08	5.88421618	43.8982939	4.08	0	1	0	Home	2	98880	98896	1	0	Low	0	ON TARGET-SAVED	0	0
74	30	Blocked	43.46	-7.48	12.1227885	38.0988265	7.48	0	1	0	Away	2	111758	111763	3	1	Low	0	BLOCKED	0	0
76	40	Goal	40.28	1.36	12.7924978	6.10277965	1.36	1	1	0	Away	2	115009	115024	1	0	Low	0	ON TARGET-GOAL	1	0
78	9	Missed	41.34	7.48	13.8530141	32.6805547	7.48	0	0	1	Home	2	117218	117245	2	0	Medium	1	HEAD-OFF TARGET-OUT	0	0
80	41	Goal	30.74	5.44	22.9150867	13.7330299	5.44	1	1	0	Home	2	121027	121055	2	0	Medium	1	ON TARGET-GOAL	0	0
82	53	Missed	45.58	0	7.42	0	0	0	0	1	Home	2	124336	124365	3	0	Medium	1	HEAD-OFF TARGET-OUT	0	0
88	23	Missed	33.92	-8.16	20.7516746	23.1550858	8.16	0	1	0	Away	2	132570	132597	6	4	Low	0	OFF TARGET-OUT	0	0
90	42	Missed	47.7	-2.72	5.95721411	27.1672369	2.72	0	0	1	Away	2	136060	136078	2	0	High	2	HEAD-OFF TARGET-OUT	0	0
93	16	Saved	21.2	1.36	31.8290685	2.44889322	1.36	0	1	0	Home	2	139891	139925	4	0	Low	0	ON TARGET-SAVED	0	1

References and Further Reading 1/3

Tracking data

Data Sources:
- Metrica Sports Tracking and correspond Event data: github.com/metrica-sports/sample-data

Vender Documentation
- Metrica Sports documentation: github.com/metrica-sports/sample-data/blob/master/documentation/events-definitions.pdf

Tutorials
- Friends of Tracking Tracking data tutorials by Laurie Shaw:
  - Part 1 – Introduction to analysing tracking data in Python: https://www.youtube.com/watch?v=8TrleFkIEsE
  - Part 2 - Measuring the physical performance of players: https://www.youtube.com/watch?v=VX3T-4IB2o0
  - Part 3 – Advanced metrics: Pitch Control: https://www.youtube.com/watch?v=5X1cSehLg6s
  - Part 4 – Evaluating player actions and passing options: https://www.youtube.com/watch?v=KXSLKwADXKI

GitHub Repositories
- LaurieOnTracking by Laurie Shaw for Tracking data implementation

Seminars and Videos
- How Tracking Data is Used in Football and What are the Future Challenges with Javier Fernández, Sudarshan 'Suds' Gopaladesikan, Laurie Shaw, Will Spearman and David Sumpter for Friends of Tracking: https://www.youtube.com/watch?v=kHTq9cwdkGA
- 'Demystifying Tracking Data' by Sam Gregory (Inter Miami) and Devin Pleuler (Toronto FC): https://www.youtube.com/watch?v=miEWHSTYvX4
- 'Classifying and Analysing Team Strategy in Professional Soccer Matches' by Laurie Shaw (City Football Group) at the 2019 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 3rd October 2019: https://www.youtube.com/watch?v=VU4BOu6VfbU. Paper: https://static.capabiliaserver.com/frontend/clients/barca/wp_prod/wp-content/uploads/2020/01/56ce723e-barca-conference-paper-laurie-shaw.pdf. Blog: https://eightyfivepoints.blogspot.com/2019/11/using-data-to-analyse-team-formations.html.
- 'Routine Inspection: Measuring Playbooks for Corner Kicks' by Laurie Shaw at the 2020 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 23rd October 2020: https://www.youtube.com/watch?v=yfPC1O_g-I8. Paper: https://www.springerprofessional.de/en/routine-inspection-a-playbook-for-corner-kicks/18671052.
- Masterclass in Pitch Control by Will Spearman for Friends of Tracking: https://www.youtube.com/watch?v=X9PrwPyolyU.
- 'A framework for tactical analysis and individual offensive production assessment in soccer using Markov chains' by Sarah Rudd (Arsenal FC) at the 2011 Sports Analytics Lab at Harvard University/American Statistical Association Section on Statistics in Sports on 3rd October 2019 [link].

References and Further Reading 2/3

Expected Goals models

Seminars and videos

- The Ultimate Guide to Expected Goals by David Sumpter (@Soccermatics) (Hammarby) for Friends of Tracking: https://www.youtube.com/watch?v=310_eW0hUgQ
- How to explain Expected Goals to a football player by David Sumpter (@Soccermatics): https://www.youtube.com/watch?v=Xc6IG9-Dt18
- What is xG? by Alex Stewart for Tifo Football: https://www.youtube.com/watch?v=zSaeaFcm1SY
- Opta Expected Goals presented by Duncan Alexander (@oilysailor): https://www.youtube.com/watch?v=w7zPZsLGK18
- Sam Green OptaPro Interview: https://www.youtube.com/watch?v=gHIY-MgDh_o
- Anatomy of a Goal (with Sam Green) for Numberphile: https://www.youtube.com/watch?v=YJuHC7xXsGA

Tutorials

- Friends of Tracking Expected Goals tutorials by David Sumpter (@Soccermatics):
  - Part 1 – How to build an Expected Goals model 1 – Data and model: https://www.youtube.com/watch?v=bpjLyFyLIXs. See GitHub: xG model: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/3xGModel.py
  - Part 2 – How to build an Expected Goals model 2 – Statistical fitting: https://www.youtube.com/watch?v=wHOgINJ5g54. See GitHub: Linear regression: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/4LinearRegression.py, xG model fit: github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython/blob/master/5xGModelFit.py

Notable xG models

- Sam Green: https://www.optasportspro.com/news-analysis/assessing-the-performance-of- premier-league-goalscorers/
- Michael Caley: https://cartilagefreecaptain.sbnation.com/2014/9/11/6131661/premier-league-projections-2014#methoderology

Professional and Fanalyst examples:

- An xG Model for Everyone in 20 minutes (ish) by Paul Riley (@footballfactman): https://differentgame.wordpress.com/2017/04/29/an-xg-model-for-everyone-in-20-minutes-ish/
- Tech how-to: build your own Expected Goals model by Jan Van Haaren and SciSports: https://www.scisports.com/tech-how-to-build-your-own-expected-goals-model/. For code, see: https://bitbucket.org/scisports/ssda-how-to-expected-goals/src/master/
- soccer_analytics repository by Kraus Clemens (@CleRraus): github.com/CleKraus/soccer_analytics/
  - Expected goal model using Logistic Regression: github.com/CleKraus/soccer_analytics/blob/master/notebooks/expected_goal_model_lr.ipynb
  - Challenges using Gradient Boosters: github.com/CleKraus/soccer_analytics/blob/master/notebooks/challenges_with_gradient_boosters.ipynb
- Expected Goals thesis by Andrew Rowlinson (@numberstorm): github.com/andrewRowlinson/expected-goals-thesis
- Expected Goals deep dive by Andrew Puopolo: github.com/andrewsimplebet/expected_goals_deep_dive
- Fitting your own football xG model by Ismael Gomez: https://www.datofutbol.cl/xg-model/
- Python for Fantasy Football by Fantasy Futopia (Thomas Whelan): http://www.fantasyfutopia.com/python-for-fantasy-football-introduction-to-machine-learning/

Articles:

- xG explained by FBref: https://fbref.com/en/expected-goals-model-explained/
- Should you write about real goals or expected goals? A guide for journalists by David Sumpter (@Soccermatics): https://soccermatics.medium.com/should-you-write-about-real-goals-or-expected-goals-a-guide-for-journalists-2cf0c7ec6bb6
- How data availability affects the ability to learn good xG models by Jesse Davis and Pieter Robberechts: https://dtai.cs.kuleuven.be/sports/blog/how-data-availability-affects-the-ability-to-learn-good-xg-models
- Expected Goals and Unexpected Goals by Garry Gelade: https://web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/expected-goals-and-unexpected-goals/
- Assessing Expected Goals Models. Part 1: Shots by Garry Gelade: https://web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/evaluating-expected-goals-models/
- Assessing Expected Goals Models. Part 2: Anatomy of a Big Chance by Garry Gelade: https://web.archive.org/web/20200724125157/http://business-analytic.co.uk/blog/assessing-expected-goals-models-part-2-anatomy-of-a-big-chance/

Literature:

- Expected Goals literature: https://docs.google.com/document/d/1OY0dxqXIBgncj0UDgb97zOtczC-b6JUknPFWgD77ng4/edit

References and Further Reading 3/3

Miscellaneous

Data visualisation

- How to Draw a Football Pitch by Peter McKeever: http://petermckeever.com/2020/10/how-to-draw-a-football-pitch/
- How To Create xG Flow Charts in Python by McKay Johns: https://www.youtube.com/watch?v=bvoOOYMQkac
- Dimension of a standard football pitch: https://en.wikipedia.org/wiki/Football_pitch

Official documentation

- scikit-learn documentation for Logistic Regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- scikit-learn documentation for Decision Trees: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- scikit-learn documentation for Random Forests: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=random%20forest#sklearn.ensemble.RandomForestClassifier

Libraries:

- mplsoccer by Andrew Rowlinson: https://github.com/andrewRowlinson/mplsoccer
- SoccermaticsForPython by David Sumpter: https://github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython
- LaurieOnTracking by Laurie Shaw: https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

Books:

- The Numbers Game by Chris Anderson and David Sally
- Soccermatics by David Sumpter

Vender Documentation

- Metrica Sports documentation: github.com/metrica-sports/sample-data/blob/master/documentation/events-definitions.pdf

Key Python libraries used:

- pandas
- Matplotlib
- scikit-learn

Resources:

- Concise list of publicly available Football Analytics resources by Edd Webster: github.com/eddwebster/football_analytics
- YouTube playlists by Edd Webster:
  - Sports Analytics / Data Science: https://www.youtube.com/playlist?list=PL38nJNjpNpH9OSeTgnnVeKkzHsQUJDb70
  - xG: https://www.youtube.com/playlist?list=PL38nJNjpNpH_VPRZJrkaPZOJfyuIaZHUY
  - Tracking data: https://www.youtube.com/playlist?list=PL38nJNjpNpH-UX0YVNu7oN5gAWQc2hq8F